



**HAL**  
open science

# Studying the genetic architecture of complex traits in a population isolate

Anthony Francis Herzig

► **To cite this version:**

Anthony Francis Herzig. Studying the genetic architecture of complex traits in a population isolate. Populations and Evolution [q-bio.PE]. Université Sorbonne Paris Cité, 2019. English. NNT : 2019USPCC110 . tel-03177047

**HAL Id: tel-03177047**

**<https://theses.hal.science/tel-03177047>**

Submitted on 22 Mar 2021

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Thèse de doctorat de l'Université Sorbonne Paris Cité

Préparée à l'Université Paris Diderot

**ÉCOLE DOCTORALE PIERRE LOUIS DE SANTÉ PUBLIQUE À PARIS**

**ÉPIDÉMIOLOGIE ET SCIENCES DE L'INFORMATION BIOMÉDICALE (ED 393)**

*Unité de Recherche : Inserm U1141 - NeuroDiderot, Equipe GenMedStroke*

# Studying the Genetic Architecture of Complex Traits in a Population Isolate

Par Anthony Francis Herzig

**Thèse de doctorat de Santé Publique**

*Spécialité: Epidémiologie Génétique*

Dirigée par Anne-Louise Leutenegger

*Présentée et soutenue publiquement à Paris le 26 Mars 2019*

<b>Mme Catherine André</b>	Directrice de Recherche, CNRS	Rapportrice
<b>Mr Tristan Mary-Huard</b>	Chargé de Recherche, INRA	Rapporteur
<b>Mr Frédéric Austerlitz</b>	Directeur de Recherche, CNRS	Président du jury
<b>Mme Nabila Bouatia-Naji</b>	Directrice de Recherche, INSERM	Examinatrice
<b>Mme Emmanuelle Génin</b>	Directrice de Recherche, INSERM	Examinatrice
<b>Mme Anne-Louise Leutenegger</b>	Chargée de Recherche, INSERM	Directrice de thèse

## Abstract

### **Etude de la composante génétique de caractères complexes dans une population isolée**

Mon projet de thèse vise à exploiter le potentiel des isolats de population pour étudier la composante génétique des maladies multifactorielles. En effet, les isolats peuvent faciliter l'identification des facteurs génétiques habituellement trop rares en population générale. Cette thèse est composée de deux études principalement : l'imputation génétique et l'analyse de l'héritabilité. Chacune de ces études ont été abordée sous deux angles : l'un théorique, s'appuyant sur une vaste étude de simulations basée sur les caractéristiques de la population isolée du Cilento, permettant d'évaluer des stratégies d'analyse et de déterminer la plus adéquate ; l'autre appliqué, s'appuyant sur l'analyse de données génétiques réelles issues de la même population.

L'imputation génétique est une étape cruciale pour effectuer des analyses d'association dans un isolat et représente une méthode peu coûteuse pour obtenir les séquences complètes du génome ou de l'exome des individus de la population. L'efficacité de cette approche dépend de la précision de l'imputation; nous avons donc étudié plusieurs stratégies pour obtenir une précision d'imputation maximale dans un isolat. Nous avons montré que les logiciels utilisant des algorithmes qui s'appuient sur les caractéristiques particulières des isolats n'étaient pas, de façon inattendue, aussi performants que ceux conçus pour les populations générales. De plus, malgré la disponibilité de panels de référence publics contenant plusieurs milliers de chromosomes, nous avons confirmé qu'un panel de référence spécifique de la population d'étude, même de taille très réduite, était essentiel pour la qualité de l'imputation. Ceci était d'autant plus vrai pour les variantes rares.

Pour de nombreux traits, il existe des discordances entre les estimations de l'héritabilité obtenues à partir d'individus apparentés et à partir d'individus non apparentés. En particulier, la plupart des chercheurs considère que les effets dominants (non additifs) ne jouent pas un rôle majeur malgré les résultats contrastés des études sur les isolats. Notre deuxième analyse a révélé des mécanismes possibles pour expliquer la disparité de ces estimations publiées entre populations isolées et populations générales. Cela nous a permis de faire des déductions intéressantes pour nos propres analyses dans le Cilento. En particulier, nous avons identifié la possibilité d'une composante de dominance non nulle pour les niveaux de lipoprotéines de basse densité (LDL). Cela nous a amenés à effectuer des analyses d'association pan-génomique des composantes additives et non-additives pour LDL dans le Cilento et nous avons pu identifier des gènes qui avaient déjà été liés au trait dans d'autres études.

Dans le contexte de nos deux études, nous avons observé l'importance de conserver l'incertitude génotypique (dosage pour l'imputation, vraisemblance des génotypes pour les données de séquençage). Dans la perspective de cette thèse, nous avons proposé des moyens d'incorporer cette incertitude à certaines méthodes utilisées dans ce projet.

Nos résultats concernant les stratégies d'imputation et l'analyse de l'héritabilité seront très utiles pour la poursuite de l'étude de l'isolat de Cilento. Mais, ils seront également instructifs pour les chercheurs travaillant sur d'autres populations isolées et également applicables plus généralement à l'étude des maladies complexes.

**Mots-clés : Maladie multifactorielle, population isolée, génétique, statistique, identité-par-descendance, phasage, imputation, héritabilité, génétique dominance, déséquilibre de liaison, effet fondateur**

## **Studying the Architecture of Complex Traits in a Population Isolate**

My thesis project is concerned with tapping the potential of population isolates for the dissection of complex trait architecture. Specifically, isolates can aid the identification of variants that are usually rare in other populations. This thesis principally contains in depth investigations into genetic imputation and heritability analysis in isolates. We approached both of these studies from two main angles; first from a methodological standpoint where we created extensive simulation datasets in order to investigate how the specificities of an isolate should determine strategies for analyses. Secondly, we demonstrated such concepts through analysis of genetic data in the known isolate of Cilento.

Imputation is a crucial step to performing association analyses in an isolate and represents a cost-efficient method for gaining dense genetic data for the population. The effectiveness of imputation is of course dependent on its accuracy. Hence, we investigated the wide range of possible strategies to gain maximal imputation accuracy in an isolate. We showed that software using algorithms which specifically evoke known characteristics of isolates were, unexpectedly, not as successful as those designed for general populations. We also demonstrated a very small study specific imputation reference panel performing very strongly in an isolate; particularly for rare variants.

For many complex traits, there exist discordances between estimates of heritabilities from studies in closely related individuals and from studies on unrelated individuals. In particular, we noted that most researchers consider dominant (non-additive) genetic effects as unlikely to play a significant role despite contrasting results from previous studies on isolates. Our second analysis revealed possible mechanisms to explain such disparate published heritability estimates between isolated populations and general populations. This allowed us to make interesting deductions from our own heritability analyses of the Cilento dataset, including an indication of a non-null dominance component involved in the distribution of low-density lipoprotein level measurements (LDL). This led us to perform genome-wide association analyses of additive and non-additive components for LDL in Cilento and we were able to identify genes that had been previously linked to the trait in other studies.

In the contexts of both of our studies, we observed the importance of retaining genotype uncertainty (genotype dosage following imputation or genotype likelihoods from sequencing data). As a prospective of this thesis, we have proposed ways to incorporate this uncertainty into certain methods used in this project.

Our findings for imputation strategies and heritability analysis will be highly valuable for the continued study of the isolate of Cilento but will also be instructive to researchers working on other isolated populations and also applicable to the study of complex diseases in general.

**Keywords: Multifactorial Disease, Population Isolates, Genetics, Statistics, Identity-By-Descent, Phasing, Imputation, Heritability, Genetic Dominance, Linkage Disequilibrium, Founder Effect**

## Acknowledgments

I would like to thank, first and foremost, Anne-Louise Leutenegger for her support during these last 3½ years. It was incredible to get the chance to come to France and work with you on this project. Throughout, you have given me a lot of freedom to explore different ideas whilst always making sure I stayed on track and got through to the end. It has been a highly enriching and enjoyable experience thanks to your guidance, encouragement, and teaching. I cannot really imagine having a better advisor for my doctoral studies.

I am also very grateful to all of whom who have collaborated with me on this project.

I would like to thank Hervé Perdry, who has a great ability for giving elegant explanations of complicated ideas; and has always been happy to explain things twice when I have needed. It has also been great to attend many of Hervé's excellent lectures.

I would like to thank Céline Bellenguez for our many detailed discussions, for always pushing me towards deeper questioning and understanding of our results, and for always finding time to give detailed responses to all of my messages.

I would like to thank Marina Ciullo for sharing with us the invaluable resource of the Cilento population for this project, for answering my frequent questions, and for her many insights into our investigation. It was a pleasure to visit Naples twice and thank you again for the warm welcomes that you extended to me each time.

I would like to thank Marie-Claude Babron for all of her kind suggestions and assistance. I was very fortunate that my first scientific publication was in fact Marie-Claude's 100<sup>th</sup> publication and so I could benefit from her considerable experience.

I would like to thank Teresa Nutile and Daniela Ruggiero for our many communications and for their enormous help with the data used in this project. Indeed, I would like to thank all the participants of the Cilento cohort and all of those who contributed to the creation and organisation of the "Genetic Park of Cilento and Vallo di Diano" project.

I would like to thank everyone at the CEPH and members of U946 for welcoming me during the first three years of my time in Paris. I made many good friends here who I look forward to staying in touch with. Thank you to Florence Demenais, Emmanuelle Bouzigon, Marie-Hélène Dizier, Hamida Mohamdi, Karine Chantrel-Groussard, Martine Bothua, Patricia Jeannin, Simone Benhamou, Mourad Sahbatou, Victor Renault, Habib Zouali, Amaury Vaysse, Pierre-Emmanuel Sugier, Myriam Brossard, Chloé Sarnowski, Yuanlong Liu, Raphaël Vernet and everyone else for the time we spent together.

I also thank Elisabeth Tournier-Lasserre and the members of the GenMedStroke team for welcoming me to their lab during the final months of my doctoral studies.

I am also thankful to the help I have had from Université Paris Diderot and Doctoral School 393 during my time in Paris; particularly for the help from France Mentré and for her advice and encouragement in all of the organisational aspects of my studies. My studies in Paris were also made possible by the funding I received through an international Ph.D. fellowship from Sorbonne Paris Cité (convention HERZI15RDXMTSPC1LIETUE) and through the Fondation Recherche Médicale (convention FRM FDT201805005384).

Sophie Duthil, who became Sophie Herzig on a beautiful day in June 2018, has been by my side for just over 7 years now. Certainly, I would never have dared to undertake my doctoral studies in France (and probably not even at all) without knowing I could count on her companionship and moral support at every step along the way. Thank you so much for the huge part you have played in getting to the end of this adventure and for joining me on the next.

Being (somewhat) far from home is never easy, so I am very grateful for all of the times my parents Frank and Petra and my brother Robbie came to visit. And for all of the encouragement from the wider Herzig clan and my ever supportive friends who were also regular visitors to Paris. It has also been great to spend more and more time chez les Duthils and to feel very much a part of my new French family.

Merci à toutes et à tous.

## Contents Page

Abstract.....	1
Acknowledgments.....	3
Contents Page .....	5
Publications.....	7
Communications.....	7
Opening Remarks.....	8
Chapter 1: Introduction .....	11
1.1 Principals of Genetic Epidemiology.....	11
1.1.1 Genome Structure.....	11
1.1.2 Genetic Data .....	15
1.1.3 Inheritance .....	16
1.1.4 Genes and Mendelian Traits .....	18
1.1.5 Complex Traits and Causal Variants.....	22
1.2 Principals of Population Genetics .....	26
1.2.1 The Theoretical Population.....	26
1.2.2 Frequencies of Alleles .....	27
1.2.3 Genetic Distance and Linkage Disequilibrium.....	28
1.2.4 Identity-By-Descent .....	31
1.2.5 Population Isolates.....	34
1.2.6 Studying Complex Traits in Isolated Populations: Track Record.....	38
1.2.7 Studying Complex Traits in Isolated Populations: Prospective.....	44
1.2.8 Applying Knowledge of the Genetics of Complex Traits.....	50
Chapter 2: The Cilento Isolate, real and simulated versions .....	53
2.1 Gioi, Cardile, and Campora .....	53
2.2 Integral Simulation Study.....	57
Chapter 3: Phasing and Imputation in Isolated Populations .....	61
3.1 Experimental Design to Investigate Phasing an Imputation Accuracy .....	61
3.2 Phasing.....	65
3.2.1 Review of Haplotype Phasing Methods.....	65
3.2.2 Accuracy of Phasing Software - Results from our Publication.....	73
3.2.3 Genotyping Errors in Phasing.....	80
3.3 Imputation .....	82

3.3.1 Accuracy of Different Imputation Software - Results from our Publication.....	82
3.3.2 Reference Panels: Local and Global - Results from our Publication.....	84
3.3.3 HapGen+Pedigree vs. Pedigree.....	87
3.3.4 Info and RSQ.....	88
3.3.5 Imputation with an Exome panel.....	92
3.4 Prospective for Phasing and Imputation in Isolated Populations.....	99
3.5 Conclusions on Phasing and Imputation.....	102
Chapter 4: Heritability.....	103
4.1 Introduction of Concepts.....	103
4.2 Published Results for Non-Additive Heritability Estimation.....	111
4.2.1 Interplay of Trait Architecture and Variance Components Estimates.....	111
4.2.2 How to Estimate Relatedness Matrices in an Isolate.....	115
4.3 Confounding with Shared Environmental Factors.....	120
4.4 Analysis of the Cilento Isolates.....	124
4.4.1 Heritability Analyses.....	124
4.4.2 GWAS of LDL in Cilento.....	127
4.5 Prospective: Estimating K and D with Uncertain Genotypes.....	137
4.6 Conclusions on Heritability.....	147
Chapter 5: Discussion.....	149
References.....	156
Glossary of Terms.....	170
Annexes.....	173



## Publications

**Herzig, AF**, Nutile, T, Babron, M-C, Ciullo, M, Bellenguez, C, & Leutenegger, A-L. (2018). Strategies for phasing and imputation in a population isolate. *Genet Epidemiol* 42,201-213.

**Herzig, AF**, Nutile, T, Ruggiero, D, Ciullo, M, Perdry, H, & Leutenegger, A-L. (2018). Detecting the dominance component of heritability in isolated and outbred human populations. *Scientific Reports* 8, 18048

Nutile, T, Ruggiero, D, **Herzig, AF**, Tirozzi, A, Sorice, R, Marangio, F, Bellenguez, C, Leutenegger A-L, & Ciullo, M. Whole-Exome Sequencing in the Isolated Populations of Cilento from South Italy. *Scientific Report, Accepted February 2019*.

## Communications

**Herzig AF**, Nutile T, Babron M-C, Ciullo M, Bellenguez C, Leutenegger A-L. Comparison of phasing and imputation algorithms on simulated sequence data in a population isolate. EMGM 11-12 May 2016, Newcastle. [Poster Presentation. Gained prize for best student poster.](#)

**Herzig AF**, Nutile T, Babron M-C, Ciullo M, Bellenguez C, Leutenegger A-L. Genetic Analysis of Complex Traits in an Isolated Population. YRLS 18-20 May 2016, Paris. [Poster presentation.](#)

**Herzig AF**, Nutile T, Babron M-C, Ciullo M, Bellenguez C, Leutenegger A-L. Impact of genotyping errors and missing genotypes on phasing and imputation in a population isolate. IGES 24-26 October 2016, Toronto. [Poster presentation. Gained prize for third best student poster.](#)

**Herzig AF**, Nutile T, Babron M-C, Ciullo M, Bellenguez C, Leutenegger A-L. Human Population Isolates: Challenges in Phasing and Imputation. DNA Polymorphisms in Human Populations, 7-9 December 2016, Musee de l'Homme Paris. [Poster and Oral Presentation.](#)

**Herzig AF**, Nutile T, Ciullo M, Perdry H, Leutenegger A-L. Revisiting Broad-Sense Heritability Estimation in a Population Isolate, 9-11 September 2017, IGES 2017 Queens' College Cambridge. [Poster Presentation.](#)

**Herzig AF**, Nutile T, Babron M-C, Ciullo M, Bellenguez C, Leutenegger A-L. Populations humaines isolées: les enjeux du phasage et de l'imputation, Assises de Genetique Humane et Medicale, 24-26 January 2018, Nantes. [Poster Presentation.](#)

**Herzig AF**, Nutile T, Ciullo M, Perdry H, Leutenegger A-L. Estimations de l'héritabilité au sens large dans une population isolée, Assises de Genetique Humane et Medicale, 24-26 January 2018, Nantes. [Poster presentation. Awarded le Prix Josué Feingold de la Société Française de Génétique Humaine.](#)

**Herzig, AF**, Nutile, T, Ruggiero, D, Ciullo, M, Perdry, H, & Leutenegger, A-L. What insights can be gained through comparisons of broad-sense heritability estimates in isolated and outbred populations? EMGM 18-20 April 2018, Cagliari. [Oral Presentation.](#)

**Herzig, AF**, Nutile, T, Ruggiero, D, Ciullo, M, Perdry, H, & Leutenegger, A-L. Exploring broad-sense heritability estimation in isolated and outbred populations. Quantitative Genomics 2018 14 June 2018, London. [Poster and Oral Presentation.](#)

## Opening Remarks

There exists great variation in the human genome in the modern-day global population, and describing the source of such variation is a theme that has intrigued those working in many different fields of research; from biochemistry to applied mathematics. In population genetics, current consensus following years of analysis of human genetic diversity has arrived at the “out of Africa” model. This describes a genesis and subsequent long period of human activity in East Africa followed by multiple waves of gradual migration into surrounding continents. The study of human genetic variation has revealed many moments where new populations expanded in new territories following the migration of relatively few individuals from a large ancestral population. We will loosely term this effect on population expansion as a ‘bottleneck’; capturing the idea that the new population is connected through ancestry to its population of origin through a relatively small number of lines of inheritance. The occurrence of bottlenecks in this demographic expansion of human beings across the planet is easily conceivable when considering the geography of habitable regions and the distances and routes between continental landmasses. From Figure 0.1 taken from Liu, et al. <sup>1</sup> we can well imagine why there exists less genetic variation in the native populations of South America than in Europe or Asia; which in turn harbour less genetic variation than in Africa<sup>2</sup>. These principal prehistoric bottlenecks have broadly shaped the world’s population.

The study of isolated populations is concerned with the same phenomenon that have governed human genetic variation but on a smaller scale. Various geographic, historical, or cultural selections can lead to a small group of individuals founding a new community that grows in relative seclusion from the wider population from which the founding individuals originated. For an example, the Icelandic population gives a good illustration of these principals. The earliest record of human settlement in Iceland was in the 9<sup>th</sup> or 10<sup>th</sup> Century, an event that was documented in the *Landnámabók*<sup>3</sup>, written originally by Ari Thorgilsson the Learned, born in 1067, a historic text detailing family history and important events in the first few centuries of Icelandic history.

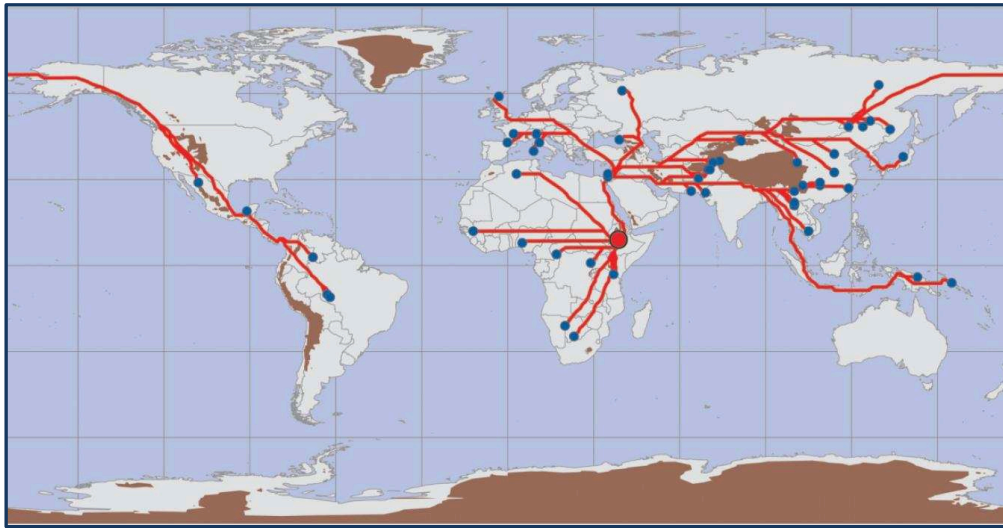


Figure 0.1 - Taken from Liu, et al. <sup>1</sup>. Possible principal routes of human population expansion based on shortest possible geographic distances; taking terrain characteristics into account. Blue dots represent populations found in the CEPH human genetic-diversity panel<sup>4,5</sup>; a landmark study of human genetic variation.

Excerpt 1: Page 8

**CHAPTER. VI.** That summer when Ingolf set out with his companions to settle Iceland, Harald Fairhair had had been for twelve years King over Norway. There had elapsed from the creation of the world six thousand and seventy three winters, and from the Incarnation of our Lord eight hundred and seventy four years. They held together until they sighted Iceland, then they seperated. When Ingolf sighted Iceland he cast overboard his high seat pillars for an omen, and he made the vow that he would settle there wherever his high seat pillar came ashore.

Excerpt 1: Page 13

**Of Helgi Bjola.** Helgi Bjola, the son of Ketil Flatnose, went to Iceland from Sodor=The Hebrides. He was with Ingolf the first winter, and settled under his advice the whole of Kjalarness, between Mogil's river and Mydal's river. He dwelt at Hof. His sons were Slaught-er Hrapp, and Kollsvein, father of Thorgerd, the mother of Thord, the mother of Ogmund, the father of Bishop John, the Holy.

*Excerpts from the Landnámabók.* The text gives highly detailed accounts of familial relationships, along with atmospheric passages.

This represents an invaluable resource for Icelandic geneticists. The genetic origins of the modern day Icelandic population have been widely studied<sup>6-8</sup> over recent decades and continues to drive research including recent analyses of ancient genomes<sup>9</sup>.

The genetic properties of isolated populations have long been of interest to epidemiologists. Many isolates have unusually low or high prevalence of certain diseases. As a case study we could look briefly to the example of Tay-Sachs disease which was observed as having a 10 times greater prevalence in Ashkenazi Jews<sup>10,11</sup>; a population extensively studied and established to be a religious or cultural isolate<sup>12-14</sup>. Isolated populations may well be characterised having a high proportion of consanguinity and so present an opportunity for uncovering genes that cause rare recessive disorders. Indeed, for the Hutterite population of North America<sup>15</sup>, one of the most widely investigated genetic isolates, over 30 such recessive disorders have been recognised in the population<sup>16</sup> and one can even find an extensive clinical database of genetic disorders associated with the Hutterites and similar populations<sup>17</sup>.

In the '90s and '00s, isolated populations were championed as a powerful tool for unpicking the genetic architecture of complex traits<sup>18-24</sup>. We refer to a trait as 'complex' when: (a) we observe the following high variability in the presentation of the trait; and (b) we anticipate that multiple factors, both genetic and environmental, will be necessary to explain the trait's distribution without any one factor being sufficient to do so. Indeed, many studies of complex traits in isolates have already led to several notable discoveries<sup>21,25</sup>. Some particular illustrations include the extensive study of Alzheimer's disease in Iceland, Schizophrenia in Finnish cohorts, and Asthma in the Hutterites, which all will be discussed in detail in Section 1.2.6.

This thesis is concerned with the study of complex traits in isolated populations, specifically in the setting of recent advances in the gathering of genetic data and recent trends in study design in genetic epidemiology. This will both involve theoretical discussion of the statistical methods involved using detailed simulation data as well as analysis of a known isolate; the three villages of the Cilento population from Southern Italy.

## Chapter 1: Introduction

### 1.1 Principals of Genetic Epidemiology

#### 1.1.1 Genome Structure

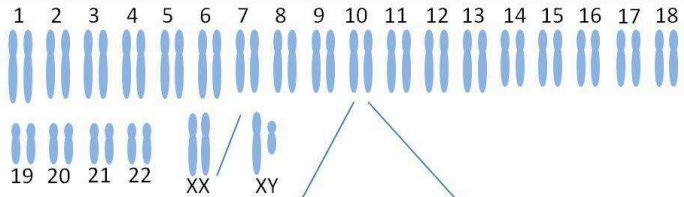
The human genome is a set of complex double-helical chains of deoxynucleic acid (DNA). This structure of this chain provides a biochemical code for the production of all necessary proteins required for the development and sustainment of the human body. Each of the two linked strands that comprise a section of DNA is a sequence of nucleotides composed of one of four chemical bases: adenine (A), guanine (G), cytosine (C), or thymine (T). The two stands of the double helix are bound together at each nucleotide position, with adenine binding exclusivity to thymine and cytosine exclusively to guanine. Within each human cell nucleus, 23 pairs of long DNA molecules are found. These 23 different sections of DNA are known as chromosomes, with the first 22 existing in homologous pairs (the autosomal chromosomes) and one pair of sex chromosomes where males have one X and one Y chromosome (which differ greatly in size and in composition), and females have two X chromosomes. As each chromosome presents in a pair, humans are 'diploid' organisms. Outside of the cell nucleus, there is also a comparatively small amount of mitochondrial DNA that is haploid; meaning there is only one copy unlike the 22 diploid autosomal chromosomes. The ensemble of the 23 chromosome pairs and the mitochondrial DNA is the human genome (see Box 1.1.1).

Grand scale scientific projects and collaborations gave the firsts complete maps of the human genome in the 1990's<sup>26-28</sup> before the human genome would be for the first time sequenced comprehensively in the early 2000's<sup>29-31</sup>. From hence, one could describe each position of each chromosome by the exact number of pairs of base nucleotides from the beginning of the chromosome. Given that within the double helical structure, adenine binds exclusivity to thymine and cytosine exclusively to guanine, it is only necessary to keep track of one of the two strands in each chromosome.

For each position, the two bases (one from each chromosome) make up the genotype which we write as  $A_1A_2$ ; examples could be GG, CC, TA, etc. For most positions on the human genome, the values of  $A_1$  and  $A_2$  will be equal and invariant across all individuals. However, for a fraction of positions, multiple nucleotides can be observed in different individuals and indeed within the genotype of a single individual. Positions such as these are ‘genetic variants’, and this is indeed the simplest case of a genetic variant: a single position where two or more nucleotides can be observed in a population. We term the different possible values that a genetic variant can take as ‘alleles’.

**Box 1.1.1 – Genome Structure**

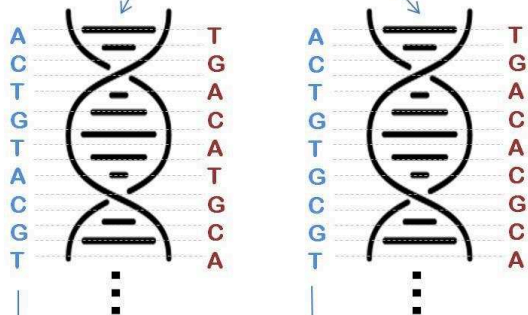
The genome of an individual with 23 pairs of chromosomes:



Note that we order chromosomes by physical size.

If we zoom in on a short section from two homologous autosomal chromosomes:

Sequences of base pairs on the ‘Up’ strand (blue) and the ‘Down’ strand (red) are displayed.



We keep track of only the Up strand (without loss of generality) and record the genotypes as follows:

... A/A C/C T/T G/G T/T A/G C/C G/G T/T ...

In particular, we notice that at the 6<sup>th</sup> position of the short example sequence, this individual has different base-pairs.

*“This Individual has the genotype A/G at position 6 on chromosome 10”!*

*Box 1.1.1 – A view of the human genome, and a view as to how to describe particular genomic positions.*

There exist many more complex forms of genetic variation. Often genetic variants span many positions and have multiple possible alleles. Explicitly, a genetic variant could describe the different possibilities for a short sequence of base pairs as well as occurrences such as: (a) ‘insertions’ where additional stretches of DNA appear to be inserted into a chromosome; (b) ‘deletions’ where a stretch

of base pairs can appear to disappear in certain individual; (c) ‘repeats’ where a short sequence (such as ‘CT’) can be repeated (i.e. CTCTCT...) a different number of times; (d) ‘inversions’ where a short sequence can appear but in reverse order; or (e) translocations where a short sequence can appear in an unexpected position (or even chromosome). The example of a genetic variant in Box 1.1.1 is a single nucleotide polymorphism (SNP) and in general, when any genetic variant is discussed in this thesis, it is assumed to be a SNP unless otherwise stated. Furthermore, in this work our discussion will focus on genetic variants that have only two possible values within a population and which lie on one of the 22 autosomal chromosomes (unless otherwise indicated). Going forward, when discussing a single variant in generalised terms, we will name these two alleles as ‘*A*’ and ‘*a*’. ‘*A*’ will denote the ‘reference’ allele, or the most commonly observed allele, and ‘*a*’ will be the ‘alternative’ allele that is less commonly observed, apparently arising from mutation. Several alternative alleles may occur at a given position, but in general we will assume only one possible alternative allele. Therefore, there will generally be three possible genotypes: *AA*, *Aa*, and *aa* as each individual will have two alleles at every position on the 22 autosomal chromosomes. In practical terms we can encode the genotypes *AA*, *Aa*, and *aa* as 0, 1, and 2 (describing the count of minor alleles in the genotype). One would only require a dictionary to look up the nucleotide values of the reference allele *A* and the alternative allele *a*. The terms ‘major’ and ‘minor’ will also be used for the reference and alternate alleles.

To give a concrete example of how our DNA actually acts as well as the effects of genetic variation, we can focus in on the APP gene found on chromosome 21. In section 1.1.4, we discuss the definition of a gene in detail, for now we can simply picture a section of DNA or long string of nucleotides (APP is approximately 290,000 base pairs long) that work together giving a certain biological functionality. APP is a gene responsible for the production of Amyloid beta peptides ( $A\beta$ ). Subtle variations in the proportions of different  $A\beta$  compositions can lead to these molecules to fold upon themselves and congregate in the brain, and their presence has been widely linked to neurodegenerative effects, i.e. they may be harmful to nerve cells. APP is a gene of ancient origins<sup>32</sup>

(back to when vertebrates were first developing) and  $A\beta$  seems to now be a multi-functioning tool in the human body (in particular the brain) including sealing leaks in the blood-brain barrier, aiding brain recover after injury and regulating synaptic function relating to memory consolidation<sup>33</sup>. The list of mutations or genetic variants within the APP gene is extensive; an excellent and detailed resource can be found at <https://www.alzforum.org/mutations/app>.

Two examples of SNPs that have been studied within the APP gene are 'rs63750264' (at position 27,264,096 of chromosome 21 with reference allele C and potential alternative alleles A, G, or T) and 'rs63750847' (at position 27,269,932 and with reference allele C and alternative allele T). The occurrence of the rs63750264 mutation was initially shown to be associated with greater risk on early onset of Alzheimer's disease<sup>34</sup> from a study of affected British families and later investigation showed this mutation to result in higher production of  $A\beta_{42}$ <sup>35</sup> (an apparently particularly unfortunate variation of  $A\beta$ ). Mutations at this position are very rare worldwide, but they have been observed in families from many different ancestral origins. Our other example, rs63750847, was discovered in a study of 1,795 Icelandic individuals<sup>36</sup> and is extremely rare outside of Scandinavia. This variant was shown to decrease the production of  $A\beta$  after having been found associated with a significant decrease in risk for developing Alzheimer's disease. These two example mutations (whose actions we have rather simplified) only scratch the surface of the multitude of results that have been carefully gathered to describe possible relationships between APP and Alzheimer's disease; and APP itself is by no means the only gene that has been linked to the condition. However, these examples give an insight into the interplay of genome and physiology, the complexity of the task in hand of describing the genetics of complex traits, as well as hinting at the potential of studying isolated populations.



### 1.1.2 Genetic Data

In the last 20 years, the size (in terms of number of measured genomic positions and also in the size of cohort studies) of genetic data has drastically increased. Technological advancements have allowed the cost of genotyping entire human genomes to drop sufficiently such that studies involving thousands of fully sequenced individuals have become widely prevalent. We will discuss three types of genetic data in this thesis: (a) genotyping array data, to be referred to as “Array”; (b) whole-exome sequencing data, to be referred to as WES data, and (c) whole-genome sequencing data, to be referred to as WGS data. Array data consist of hundreds of thousands of genotyped SNPs, known to be polymorphic in multiple worldwide populations. Such data is capable of characterising uniquely the genome of any individuals, of observing relatedness between individuals, and for estimating the ancestry of a given individual. Furthermore, through Linkage Disequilibrium (a description of statistical correlation between pairs of nearby genetic variants, discussed fully in section 1.2.3), Array type data can represent the variation of the entire genome. This is a consequence of the fact that the many thousands of genotyped SNPs span the 23 chromosomes in an even manner (when viewed macroscopically). These Array positions are aiming to ‘mark’ as much genetic variation as possible, through the fact that any given genomic position should not be too far away from a genotyped Array position. The human exome describes the fraction (~2%) of the entire genome that is directly responsible for protein coding. Simply put; it is the ensemble of genome segments (exons) that have the most direct actionable effects. This is not to say that the remaining ~98% of non-coding genome is not important, indeed it is from these regions that the majority of significant associations with complex traits have been found<sup>37,38</sup>. WES data is the full sequence of these exonic regions, which will comprise a large set of small packets of closely grouped, if not even sequential, genetic variants. WGS data is a record of every position on the human genome, though typically only the polymorphic markers will be analysed for obvious reasons. This will usually represent many millions of genetic variants.

Figure 1.1.2 gives a visualisation of WGS, WES, and Array positions that are present in the datasets of the Cilento isolate (discussed fully in section 2.1) that fall within the first million base pairs of chromosome 10. We can observe here the density of the WGS data, the regions with clustered WES data and the spread of Array positions. The plot contains 3,918 WGS positions, 149 WES positions and 52 array positions.

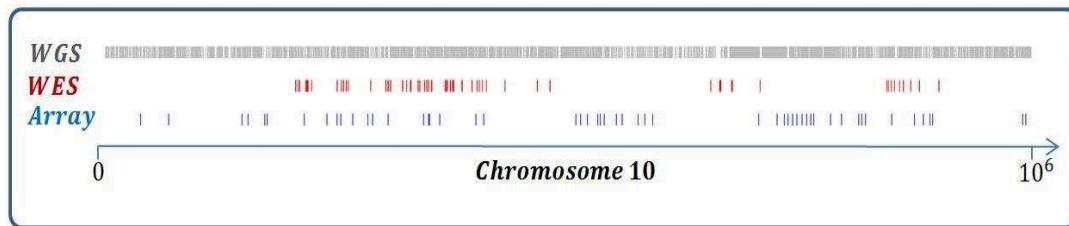


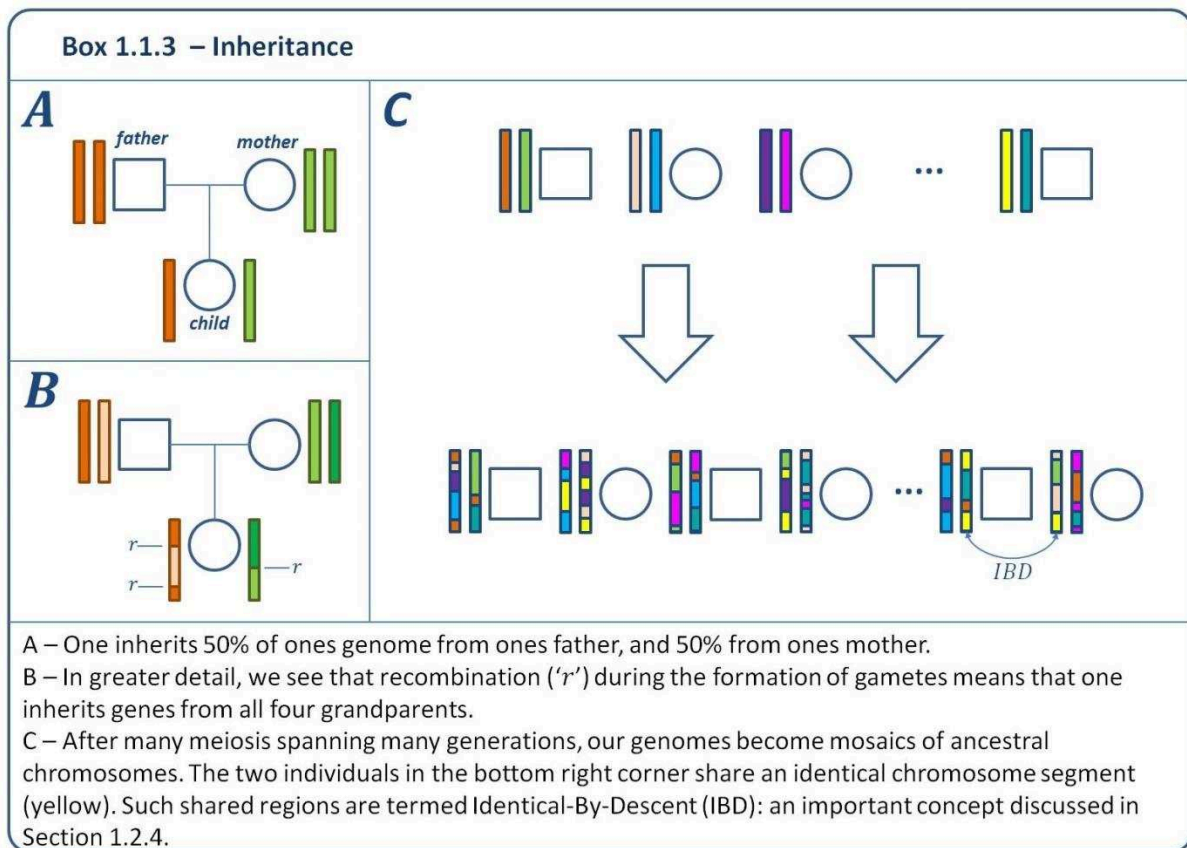
Figure 1.1.2 – Different types of genetic data, with different densities of positions. A short example of positions from Whole-Genome Sequencing (WGS) data, Whole-Exome Sequencing data (WES) and genotype array data (Array) on the first million base pairs of chromosome 10.

### 1.1.3 Inheritance

There exist two notable exceptions to the previous statement that each human cell contains 46 chromosomes – the male sperm cells and female ovule cell contain only 23 chromosomes; one copy of each autosomal chromosome and one sex chromosome (ova will carry one X-chromosome, and spermatozoa will either carry either an X or a Y chromosome. Such cells carrying half of an individual's DNA are known as gametes, and the joining of two gametes during sexual reproduction is the mechanism by which genetic data is passed from parents to offspring (Box 1.1.3, section A).

The creation of gametes (a process known as meiosis) allows for the transmission of material from both parental chromosomes to their offspring. Hence, each chromosome within each new gamete is composed of fused fragments of the two parent chromosomes (Box 1.1.3, section B). The process that recombines fragments from each individual's two pairs of chromosomes and then reconstructs new chromosomes in gamete formation is known as recombination. In section B, of Box 1.1.3, we can see that two recombination events have occurred during the creation of the orange gamete passed from

father to child, and one recombination event has occurred during the maternal transmission. These recombination events are marked in the figure by 'r'. In section C of Box 1.1.3, we see this transmission of genetic material within a population over many generations, involving many recombination events. This gives the idea that our genomes are mosaic like, built of small sections of ancestral genomes. We also touch here on the idea of identity-By-Descent (IBD), the situation where multiple identical copies of a portion of a chromosome, inherited from a common ancestor, are observed in a population.



Box 1.1.3 - The concepts of inherited genetics, recombination, and the transmission of chromosome segments.

### 1.1.4 Genes and Mendelian Traits

Before continuing, we should establish what is meant by a ‘gene’. This is a term that has evolved alongside the centuries of progression of scientific understanding of genetics<sup>39</sup>. In what is a widely celebrated example of scientific curiosity and endeavour, Augustinian friar Gregor Mendel’s investigation into pea plant characteristics<sup>40</sup> in 1866 marked the beginnings of the concept of physical entities (Mendel referred to these as ‘elements’) that are transferred between generations. We now know that he had discovered what we would still call genes in the genome of *Pisum sativum* that affected observable characteristics such as the colour of the flowers or the exterior texture of the peas. After the turn of the century, the link between chromosomes and inheritance was developed<sup>41,42</sup> and hitherto, to a large extent, a gene was considered as a single-point entity, lying on a chromosome. When the double-helical structure of DNA was co-discovered by James Watson & Francis Crick<sup>43</sup>, Rosalind Franklin<sup>44</sup>, and Maurice Wilkins<sup>45</sup> (the accreditation of this discovery being an extensive subject in its own right), we arrive at the concept of a gene being a string of nucleotides that give a code for the production of a specific protein; a molecule of messenger ribonucleic acid (mRNA).

To make a brief aside, this is the definition that I would tend to hold inside my head. However, often it is still most intuitive not to reflect at all on the physical description of DNA, and to consider each chromosome to be an enormously long string of random variables from a probabilistic point of view. Or, in an even more practical sense, as a long series of zeros, ones, and twos stored digitally as the scale of genetic data can simply only now be dealt with advanced computational algorithms.

Yet, the coding string of DNA definition is still an over simplification. There is great variation in mechanisms for producing mRNA, and the model where one gene codes one specific molecule has multiple counter examples. What is more, the boundaries between pairs of genes and between genes and non-coding regions are often practically indeterminable. The scale of interaction between DNA elements has also been shown to be highly complex and much study has gone into the action of

entire networks of genes, the role of genes within diverse biological pathways as well as the activity of genes in different parts of the body. A further layer of complication comes from the fact that chromosomes are tightly folded in on themselves, with regions segregating into 'topologically associating domains' whose distributions are known to play a part in gene regulation. One can also make reference to the studies of epigenetics, to endogenous viral elements within the genome, of chromosome X inactivation. Indeed, the list of topics not considered in this thesis grows worryingly large. Here, one should highlight both the huge challenges as well as rich opportunities for genetic studies. It is a subject that is well disposed to intrigue biologists, chemists, mathematicians, statisticians, computer scientists as well as other disciplines. There is a great importance for collaborative efforts across these disciplines in order to effectively unlock the functionality of the human genome.

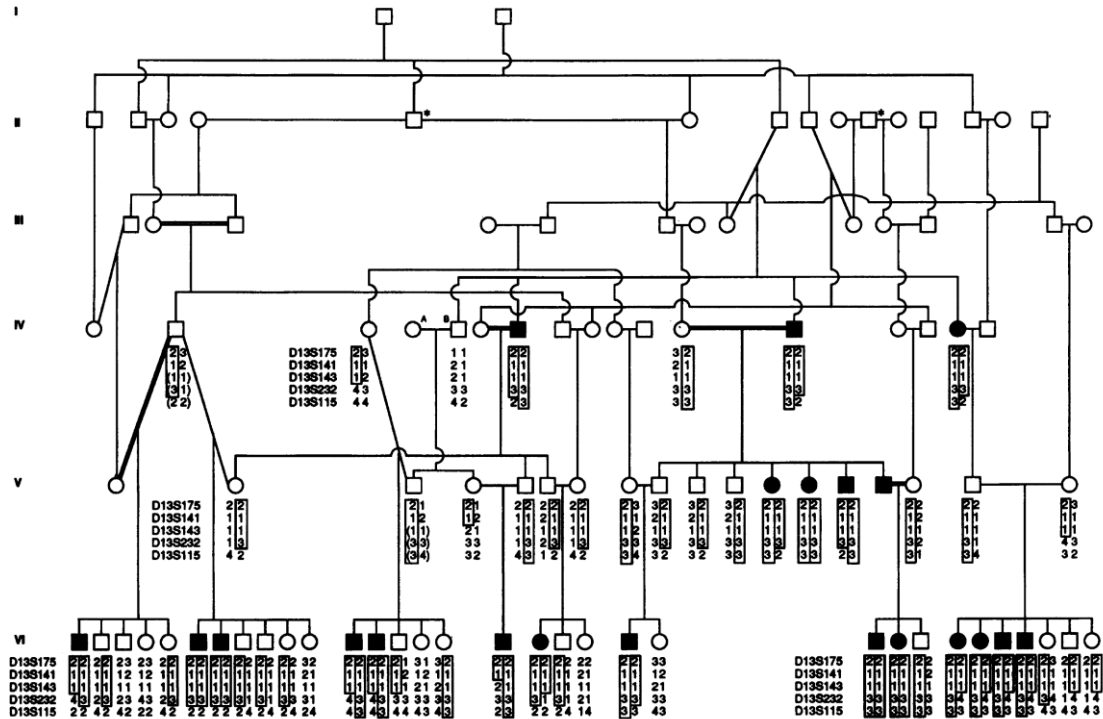
The behaviour of many genetic disorders is described as Mendelian or monogenic. This is the case when an individual's genotypes across variants in a single gene are the sole genetic variation that governs an observable phenotype, or disease. Different possible mutation at different positions in the gene may be implicated in the prevalence of the disorder. A classic example of such as disorder is Cystic Fibrosis<sup>46</sup>, a genetic disorder that affects multiple systems in the body due to mutations in the CFTR gene on chromosome 7. Around two thirds of all cases are caused by the presence of a three-nucleotide deletion called  $\Delta F508$ <sup>47</sup>, Affected individuals will this deletion on both copies of their 7<sup>th</sup> chromosomes. The remaining third of the cases are caused by neighbouring mutations, of which over 1,000 have been identified, many of which have been studied in isolated populations. One notable example being the mutation MR1101K (in gene CFTR) that was discovered in the Hutterite population<sup>48</sup> where the mutation is unusually frequent.

Mendelian traits are characteristically rare (infrequent within a population) and the relationship between the mutation and the disorder is usually clear-cut. To be precise, there is a high or complete penetrance. The penetrance of a genetic variant involved in a Mendelian disorder is the

proportion of carriers of the mutation who also develop the disorder. The higher this proportion, the more likely the genotype of the variant will determine the distribution of the trait.

Family based studies involving linkage analysis have proved very successful in identifying for many such disorders as well as disorders that are oligogenic (dependent on a few genes), each acting in a Mendelian way. Study designs for linkage analysis do not require whole genome sequencing, but only a dense enough set of genetic markers in order that any conceivable variant in a gene that affects the trait will be in 'close' enough (concepts of genetic distance are discussed in Section 1.2.3) with one or several of the genotyped markers. These markers are chosen to be highly variable so that the transmission of chromosome stretches (haplotypes) can readily be tracked through meiosis in the family structure. Linkage analysis is usually based on calculating Morton's LOD scores<sup>49</sup>, which describes the logarithmic odds of whether genetic markers are close or not to a causal variant. Such methods require the ability to calculate the likelihood of given genotypes in a family based on a known pedigree; notably using either the Elson-Stewart<sup>50</sup> algorithm or the Lander-Green-Kruglyak algorithm<sup>51,52</sup>.

Figure 1.1.4 gives an example of a highly complex pedigree of five generations from a Bedouin tribe which includes high rates of inbreeding; also given is an extract from the corresponding study describing the population. This example is taken from Scott, et al.<sup>53</sup>, who were able to locate the gene that would be shown to be responsible for the numerous individuals within the family affected with Nonsyndromic Autosomal Recessive Deafness. The gene they found was DFNB1 which is located on chromosome 13, which subsequently was shown to carry certain mutations in this Bedouin family<sup>54</sup>. In the figure, haplotypes for five variants are tracked through the pedigree, helping to pinpoint the linkage signal.



The Bedouin family described in the present study belongs to a tribe founded ~200 years ago by an Arab-Bedouin male who emigrated from Egypt to the southern region of what was then Palestine. He married a local woman and had seven children, five of whom survived to adulthood. Consanguineous marriage has been the rule in the tribe since its third generation. The tribe is presently in its seventh generation and consists of some 3,000 people, all of whom reside in a single geographic area in Israel that is separated from other Bedouin communities. Birth rates within the tribe are high, and polygamy is common.

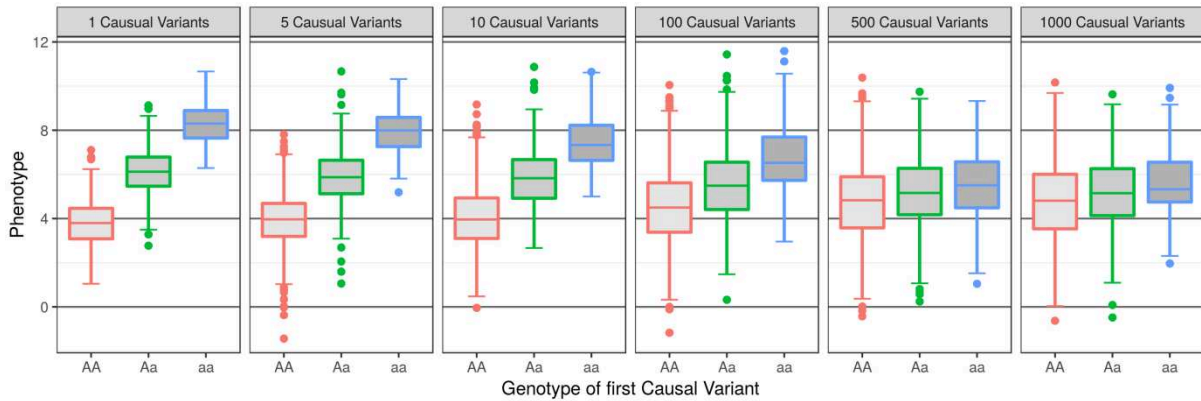
Figure 1.1.4 - Pedigree used for linkage analysis of Nonsyndromic Autosomal Recessive Deafness in a Bedouin population along with excerpt from Scott, et al.<sup>53</sup>

### 1.1.5 Complex Traits and Causal Variants

This thesis is concerned with the behaviour of complex traits. This is the study of continuous human characteristics, and diseases and disorders with varied phenotypic presentations. Complex traits may also be binary in nature, and in this case, it is the risk of having the disease that is thought of as continuous. Such traits are demonstrably non-Mendelian, yet there is clear evidence from the analysis of family data that they can be (partially) inherited. The idea is that there is correlation between the phenotypes of parents and their offspring: tall parents are more likely to have tall children. However, a gene for tallness is not forthcoming. The heights of adult males and female are normally distributed, with many non-genetic factors affecting the measurement. R.A. Fisher was the first to formally describe how the multiple genetic variants, each acting in a Mendelian way could combine together to give continuous distributions of complex traits observed across a population<sup>55</sup>. The central idea being the polygenic Model: some large number ( $j = 1, \dots, M$ ) of genetic variants each contribute to the trait with values of  $u_{j0}$ ,  $u_{j1}$ , and  $u_{j2}$ , dependent on their three possible genotypes  $AA$ ,  $Aa$ , and  $aa$ .  $M = 1$  describes Mendelian inheritance, as might be seen within a study of an affected family,  $M < 10$  was the level of polygeny that was expected to be found for most human complex traits until recent decades, and  $M \gg 10$  approaches Fisher's infinitesimal model which gives a mathematical explanation that bridged the concepts of Mendelian inheritance and the observed continuous distributions of human traits.

If we can assume that these 'causal' variants will act independently on the phenotype, then we can assume that each causal variant will act in a somewhat Mendelian way on the trait, and if we look at the marginal effect of each of these variants over a large cohort of individuals we should observe something akin to Figure 1.1.5a. However, the more polygenic the trait architecture, the less observable the marginal effects will be.





*Figure 1.1.5a – Relationship between the total number of causal variants and the marginal effect of a single causal variant (under the assumption that the ensemble of all causal variants contribute equally to the trait variance in each case).*

If the variance of the trait is fixed, then the polygenic model leads to this relationship between the number of causal variants ( $M$ ) and the effect sizes of such a variant. Hence, for a highly polygenic trait, it becomes very difficult to detect the effects of causal variants and may require a large number of individuals to notice these patterns. Currently, the most prevalent method is the Genome-Wide Association Study (GWAS). Profiting from recent sequencing technology advances and the availability of WGS data, this method exhaustively tests the statistical association between each sequenced variant and a given trait. In effect, no prior knowledge of which genes may play a role in the distribution of the trait is needed, as the dense sequencing data allows for a near-comprehensive sweep of the genome. This method is not fully comprehensive as there are still types of genomic variation that are very difficult to successfully sequence which are therefore not always included<sup>56</sup>. To date, GWAS have had great success<sup>56</sup> in uncovering associations between phenotypic and genetic variation. What is more, the breadth of such studies has gone through an explosive increase in the last 10 years in terms of sample sizes and numbers of traits studied<sup>57</sup>. Figure 1.1.5b shows the numbers of GWAS ‘hits’ or associations currently archived in the online resource the GWAS catalogue.



Figure 1.1.5b - All published GWAS hits to date. A plot that many will recognise; we now see that after many decades of study that the GWAS hit flags have begun to dominate the chromosomes. Generated from <https://www.ebi.ac.uk/gwas/> (December 3<sup>rd</sup> 2018) where the plot can be remade to show the published results at any date desired.

Here I will introduce statistical methods pertaining to a basic GWAS of a complex trait and some notation relevant for the thesis in general. Assume, that for  $N$  individuals we have a vector of phenotypes  $Y$ . If we need to denote a single individual's phenotype, we write  $Y_i$ , where the index  $i$  runs from 1 to  $N$ . After applying transforming the trait; lets also assume that the expectation of the trait  $E[Y] = 0$  and the variance of the trait  $var[Y] = 1$ . Furthermore; we will assume a model where  $Y$  follows a normal distribution. This allows us to fit a linear regression model for  $Y$  and test for an association with each genetic variant. A GWAS represents a collection of tests of the following model:

$$Y \sim MVN(\beta_j X_j + U\beta_0, \sigma_E^2 I_N)$$

$X_j$  is the vector of the genotypes of genetic variant  $j$  across the sample. Writing a single entry of  $X_j$  as  $X_{ij}$ , the genotype of variant  $j$  for individual  $i$ , where:

$$X_{ij} = \begin{cases} 0, & \text{genotype of variant } j \text{ for individual } i = AA \\ 1, & \text{genotype of variant } j \text{ for individual } i = Aa \\ 2, & \text{genotype of variant } j \text{ for individual } i = aa \end{cases}$$

This represents an ‘additive’ model for the genetic effect.  $U$  is a matrix of other non-genetic covariates,  $\beta_j$  is a constant,  $\beta_0$  is a vector of constants and  $\sigma_E^2$  describes variance in the trait coming from environmental effects.  $I_N$  is an  $N \times N$  identity matrix. Our statistical test will be against the null hypothesis:  $H_0: \beta_j = 0$ .

When dealing with a sample which include related individuals, such as in an isolated population, we can expect the phenotypes of closely related individuals to be correlated. In order to be able to separate these patterns from the effects of a single variant requires a more complex variance structure and so a Linear Mixed Model (LMM) is fitted. Now we assume that the phenotype follows a multi-variate normal distribution:

$$Y \sim MVN(\beta_j X_j + U\beta_0, \tau K + \sigma_E^2 I_N)$$

This variance structure is comprised of the environmental variance  $\sigma_E^2 I_N$  that is independent across individuals and a genetic component  $\tau K$ . Here,  $K$  is a variance-covariance matrix representing the correlation between phenotypes of pairs of individuals (usually based on their relatedness);  $\tau$  is a constant. This LMM and the variance-covariance structure  $K$  will be the main focus of Chapter 4 where we will revisit these models in much greater detail.

For GWAS, we are discussing a model which involves a continuous distribution of a trait and which involves very subtle patterns that will not present themselves clearly in data collected from a family. This leads to the study of genetics over wide cross sections of populations rather than orientating one’s ideas around the genetics of a single individual. In the following chapter, we will introduce concepts of population genetics, including isolated populations, before exploring further the study of complex traits.

## 1.2 Principals of Population Genetics

It should be noted that in animal/plant studies, the ability to select and oversee crosses between individual organisms facilitates many potential analytical approaches. This not being possible in humans, large existent human populations are studied. In a sufficiently large population, the global genetic characteristics will be maintained between different generations, under certain conditions. However, external forces and circumstances may lead to the natural evolution of population characteristics and the study of population genetics is the study of the possible mechanisms that interplay with the genetics of the population as a whole. In this section we will give only a brief summary of concepts that will be of importance to the analyses we have performed in this thesis (Chapters 3 and 4). The following materials will give much richer descriptions: ‘The Genetic Structure of Populations’ – A. Jacquard (English Translation)<sup>58</sup>, ‘Introduction to Quantitative Genetics’ – D. S. Falconer<sup>59</sup>, and ‘Génétique des Populations’ – J-L. Serre<sup>60</sup>.

### 1.2.1 The Theoretical Population

The context of all models that will be introduced depends on the assumption of large underlying populations of individuals from which our study individuals have emerged or have been sampled. The population of the entire world seems a logical realisation but, as described in the opening remarks of this thesis, there exist too many structures in the world’s population due to the history of migration and settlement across the planet. Instead, models are based on a more homogenous theoretical population with a sufficiency large sampling size such that there is approximate temporal stability - i.e. from one generation to the next, the population characteristics should not significantly change. When mathematically minded, we can simply describe this population as infinite, though we will instead usually describe this general population as having finite size  $N$  where  $N$  is very large. As described by Falconer<sup>59</sup>, the most significant assumption that we need for this populations is that the each generation transmits gametes to subsequent generations completely at random. In effect,

that all individuals have equal fertility and all potential mating pairs have equal probability of occurring. In detail, this requires that the formation of gametes should not engender any biases as to which alleles are transmitted to subsequent generations. Further assumptions about this theoretical population that are usually made include: (a) that it does not change in size between generations; (b) alleles are transmitted immaculately (without mutations); and (c) there is no migration of new individuals into the population.

### 1.2.2 Frequencies of Alleles

We have already introduced the notion of a genetic variant with (in the simplest case) two possible alleles ( $A$  and  $a$ ) and three possible genotypes ( $AA, Aa, aa$ ). Within a population, if we have access to genotypes of all members, we can then calculate the frequencies of each genotype, which we shall denote as  $p_{AA}, p_{Aa}$ , and  $p_{aa}$ . We could equally think of the distribution of the genotype coming from  $N$  individuals but from  $2N$  chromosomes (two from each individual). We can write down the frequencies of each of the two alleles which we denote as  $p_A$ , and  $p_a$ . By simple counting, we have the following:  $p_A = p_{AA} + \frac{1}{2}p_{Aa}$  and  $p_a = \frac{1}{2}p_{Aa} + p_{aa}$ . When we discuss a single variant in general, we will use the following ubiquitous notation of  $p = p_A$  and  $q = p_a$ . As  $q$  is the frequency of the minor (or mutant) allele, it is more often the quantity of interest and will often be referred to as the Minor Allele Frequency (MAF).

If we then consider a new individual being born within this theoretical population, one can attempt to predict their genotype for this variant. The probability of receiving allele  $A$  from a parent (who has genotype  $G_p$ ) will be equal to:

$$1 \times \text{Prob}(G_p = AA) + \frac{1}{2} \times \text{Prob}(G_p = Aa) + 0 \times \text{Prob}(G_p = aa) = p_{AA} + \frac{1}{2}p_{Aa} = p$$

Similarly, the probability of receiving allele  $a$  will be equal to  $q$  by symmetry. Indeed, it is clear that this new individual's two alleles are effectively random draws from the pool of alleles in the parent population. This uses the conditions outlined in Section 1.2.1 specifying that this new individual has

two parents drawn completely at random, and each allele in the population is equally likely to be one of the two passed from these two parents to the child. Denoting the new individual's genotype as  $G_o$ , with the random mating assumption, the inheritance of each allele is an independent event and the probability of the three possible genotypes are  $Prob(G_o = AA) = p^2$ ,  $Prob(G_o = Aa) = 2pq$ , and  $Prob(G_o = aa) = q^2$ . If we then consider the birth of many individuals, or in effect the arrival of an  $(n + 1)^{th}$  generation arising from random mating in the  $n^{th}$  generation, we should expect to see genotype frequencies in this  $(n + 1)^{th}$  generation approximately equal to  $p^2$ ,  $2pq$ , and  $q^2$ . These frequencies are known as the proportions of Hardy-Weinberg. If a genetic variant presents in these proportion in a population respecting the structural assumptions laid out in Section 1.2.1 then it is said to be in Hardy-Weinberg Equilibrium (HWE). Often, a Pearson's  $\chi^2$ -test is employed to test for significant departures from Hardy-Weinberg proportions (HW) for a single variant. Small departures from HW can indicate structures within the population such as assortative mating, non-homogenous populations, or inbreeding. However, finding an extreme departure for a variant can be an indication of genotyping errors and so such variants are often excluded from subsequent analyses.

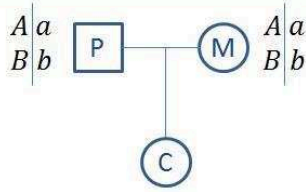
### 1.2.3 Genetic Distance and Linkage Disequilibrium

Genetic variants are linked together in chromosomes, and inheritance works through the transmission of long stretches of these chromosomes (Box 1.1.1). If we know that an individual has inherited a particular allele at a particular position from one of their two parents, we can infer that in fact they have also inherited a substantial chain of alleles (or haplotype) from that parent. Whilst the possibility of recombination events will suggest this inherited haplotype may be limited in size (i.e. will not span the whole chromosome for example), this inference can be highly informative and this linking of inherited alleles will drive many statistical models for population genetics. Hence it is of great importance to be able to describe the genetic distance between two markers on a chromosome in order to know how likely it is for haplotypes containing both markers to be

transmitted from parent to child. Two alleles on completely different chromosomes (e.g. on chromosome 1 and chromosome 19) each have a 50% chance of passing on an offspring, and as they are on separate chromosomes, there is a 25% chance of them both being passed on. This cannot be said of two alleles nearby on the same chromosome where two alleles on the same haplotype will either both be transmitted or neither will be; unless one (or possibly several) recombination event(s) occur between the two positions.

In Box 1.2.3, we describe how genetic distance can be derived from the frequency of recombination events during the production of gametes via the Haldane mapping function<sup>61</sup>. Box 1.2.3 also gives a brief explanation of Gametic Disequilibrium and Linkage Disequilibrium (LD). LD is the persistence of Gametic Disequilibrium in a population (see Box 1.2.3 or for a full description see '*Génétique des populations*' by Jean-Louis Serre<sup>60</sup>). The LD-matrix plot in Box 1.2.3 was generated using R-package 'Gaston'<sup>62</sup>. This plot is a graphical representation of the lower-triangular portion of a square matrix that gives the correlation between each pair of a set of 150 genetic variants; with the rows ordered by the physical position on the genome. The darker coloured entries represent higher values of correlation. Two distinct regions are clearly seen, wherein variants are correlated with each other, but with very little correlation between variants in different regions. Such regions are referred to as LD-blocks.

**Box 1.2.3 – Genetic distance, Gametic Disequilibrium, and Linkage Disequilibrium**



**Example scenario:** two genetic variants on the same chromosome, each parent carries two copies, and both carry heterozygous genotypes  $A/a$  and  $B/b$ .

For both, the genetic ‘phase’ of these variants is  $AB | ab$ . This implies that both parents have one chromosome with alleles  $A$  and  $B$  and a second chromosome with alleles  $a$  and  $b$ .

The key idea is that the offspring individual  $C$  will either inherit an  $AB$  or  $ab$  ‘haplotypes’ (chromosome segment), unless there is at least one recombination event between the two variants during either paternal or maternal transmission. Thus, the probability of a recombination event between the two loci can be used to describe their proximity.

If we analyzed a large set of  $N_g$  gametes produced by either parent, we should see the following distribution for some positive value of  $\theta$ .

Gamete	$AB$	$ab$	$Ab$	$Ba$
Frequency	$N_g(1 - \theta)/2$	$N_g(1 - \theta)/2$	$N_g\theta/2$	$N_g\theta/2$

$\theta$  is the rate of recombination between the two positions if we see recombinant gametes ( $Ab$  and  $Ba$ ) with a frequency of  $\theta$ . If the loci are completely un-linked then  $\theta = 0.5$ , otherwise if  $\theta < 0.5$ , the two positions are in **Genetic Linkage**.

**The Haldane mapping function:**  $d_H = -\frac{1}{2} \log(1 - 2\theta)$ .

The transformation gives a useable metric in order to describe distance on a chromosome. The units of this distance are Morgans (M), equivalent to the expected number of recombinations per meiosis that are expected between two loci. As the values are usually very small, one will always tend to describe distances in centi-Morgans (cM) calculated as  $100d_H$ .

If we consider these two loci across a population, then our expected frequency of observing the haplotype  $ab$  ( $p_{ab}$ ) is equal to  $p_a p_b$  if we assume these alternate alleles occur independently. However, this expectation may not be met. If we write  $D = p_{ab} - p_a p_b$ , then  $D$  measures the departure from this assumption. If  $D = 0$  then the two loci are in **Gametic Equilibrium**, and if  $D \neq 0$  then they are in **Gametic Disequilibrium (GD)**.

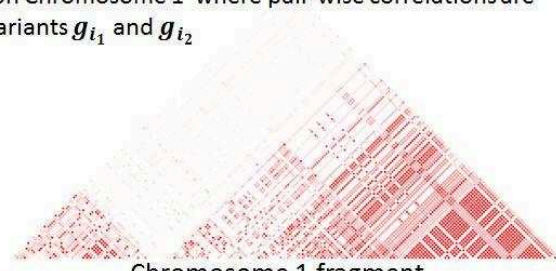
Over many generation, genetic recombination will cause **GD** to dissipate, but when **GD** persists in a population, this is referred to as **Linkage Disequilibrium (LD)**.

**Potential driving forces of LD:**

- Genetic selection and stochastic genetic drift can favour specific haplotypes
- Migration or movements of populations causing mixtures of different haplotype patterns of different origins
- Mutations in the population

LD between two loci is most usually quantified as  $r^2 = \frac{D^2}{p_A p_B p_a p_b}$ . This can be estimated directly as the statistical correlation between the two loci with genotypes coded as 0,1, and 2.

Below is an example of an LD-plots of 150 variants on Chromosome 1 where pair-wise correlations are presented. The magnitude of the LD between two variants  $g_{i_1}$  and  $g_{i_2}$  is shown by the redness of the corresponding pixel:

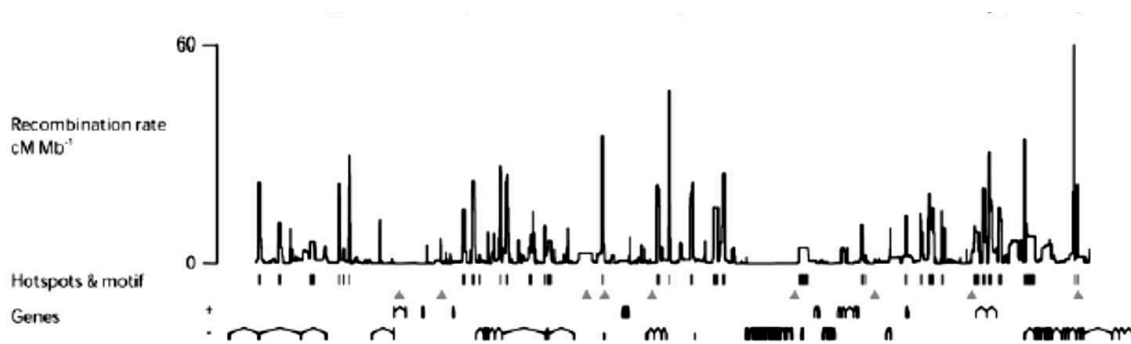


LD example plot from R-Package Gaston.

*Box 1.2.3 – Key concepts of genetic distance and linkage disequilibrium.*



Recombination events do not occur completely at random across the genome; there are particular places that are far more susceptible to cleave during meiosis. These regions are known as recombination hotspots, usually between 1Kb and 2Kb in length<sup>63</sup> (Kb=Kilo-base, 1000 base pairs, similarly we will also refer to length in Mb – Mega-base, 1,000,000 base pairs). Around 30,000 hotspots have been identified on the human genome<sup>64</sup> with an average hotspot resulting in recombination slightly less than once every 1,000 meioses (Myers, et al. <sup>64</sup> suggested that the average probability of recombination at a hotspot is roughly  $0.00075 \approx 1/1300$ ). In Figure 1.2.3, the recombination hotspots on chromosome 21 are displayed.



*Figure 1.2.3 - Taken from Myers, et al. <sup>64</sup> shows estimates of recombination rates on chromosome 21. A high peak indicates a high ratio between the distance between loci in cM compared to distance in base pairs. These peaks indicate particular high frequencies of recombination events (recombination hotspots). Of interest, the distribution of genes (on both the Up (+) and Down (-) strands) are also presented.*

#### 1.2.4 Identity-By-Descent

IBD was briefly introduced in Box 1.1.3, this is a key concept in this thesis. Simply, if we observe two alleles, then our assumption is that this pair will fall in one or three formations: (a) the two are different (e.g. *A* and *a*), (b) they are identical (e.g. *A* and *A*), or (c) they are identical and what is more both alleles originate from a recent common ancestor. Alleles that are identical are said to be Identical-By-State (IBS) and alleles that are identical and share a common ancestral origin are said to be Identical-By-Descent (IBD). IBD-sharing is a particular example of IBS-sharing. In Box 1.2.4, a simple example of IBD is given from a nuclear family (section A), then from a pedigree (family tree diagram) with an example of a consanguineous marriage (section B), and finally we present a well-

known and useful method of notation to describe possible IBD configurations between two individuals (Section C). For example:  $\Delta_1$  is the probability of all four alleles being inherited IBD;  $\Delta_4$  is the probability that individual  $i$  has two alleles which are IBD, and individual  $k$ . The probabilities of the nine IBD states for the two siblings in section A of Box 1.2.4 are  $(0, 0, 0, 0, 0, 0, \frac{1}{4}, \frac{2}{4}, \frac{1}{4})$ . For the two siblings in section B, the corresponding values would be  $(\frac{1}{64}, 0, \frac{2}{64}, \frac{1}{64}, \frac{2}{64}, \frac{1}{64}, \frac{15}{64}, \frac{30}{64}, \frac{12}{64})$ .

**Box 1.2.4 – Identity-By-Descent**

**A**

1. The two siblings share two alleles Identical-by-State (IBS).  
 2. Furthermore, the shared allele  $a$  is Identical-By-Descent (IBD) as both siblings inherited it from the father.  
 3. The allele  $A$  could also be shared in IBD, but this cannot be determined without looking back at previous generations.

**B**

In this pedigree with a consanguineous marriage, individual  $k$  inherited two alleles IBD where the common ancestor is grandparent  $\alpha$ .

---

**C Identity by descent sharing states of two individuals,  $i$  &  $k$**

For a single variant, the four alleles are displayed as points (right) and IBD-sharing as connecting lines between points.

$\Delta_1$     $\Delta_2$     $\Delta_3$     $\Delta_4$     $\Delta_5$     $\Delta_6$     $\Delta_7$     $\Delta_8$     $\Delta_9$

Alleles of  $i$  ● ●

---

Alleles of  $k$  ● ●

Note: If the order of the alleles within each individual is also recorded, then there are in fact 15 possible states (so that a difference is recognized between states such as & ). Hence,  $\Delta_{1,\dots,9}$  are referred to as Jacquard's 'condensed' identity coefficients.

*Box 1.2.4 – Key concepts of Identity-By-State (IBS) sharing and Identity-By-Descent (IBD) sharing.*

The expected values of nine IBD states, first described by Jacquard<sup>58</sup>, can be estimated from a given pedigree. This can be done using recursion algorithms<sup>65</sup>, with the idea being that the probabilities of each IBD state between two individuals can be derived explicitly if we know the probabilities of each IBD state between every pair of the two individuals' four parents. If this principal is followed, within a finite pedigree, eventually one arrives at the 'founding' individuals, where the founders of a

pedigree are simply those for whom no information about their parents has been recorded. It is often cumbersome to work with all nine probabilities. If inbreeding (IBD-sharing between two alleles within an individual's genotype) is expected to be very low, it is often assumed that only the last three coefficients ( $\Delta_7, \Delta_8, \Delta_9$ ) are non-zero and the notation  $(k_2, k_1, k_0)$  is often used (or similar), which indicates the proportions of two alleles shared IBD, one allele shared IBD, and zero alleles shared IBD. For brevity, we will write IBD=2, IBD=1, and IBD=0. Knowledge of IBD-sharing between individuals and identifying shared regions can be used to infer degrees of relatedness between individuals and has many applications such as in linkage analyses<sup>66</sup>, in methods for establishing genetic phase and for genetic imputation (focus of Chapter 3), and in heritability studies (focus of Chapter 4).

Before continuing, one further concept that is necessary to introduce is the kinship coefficient between two individuals  $i$  and  $k$ . This is the probability that two alleles from the same position,  $a_i$  &  $a_k$ , that are randomly sampled from two individuals  $i$  &  $k$  will be IBD. We can write this as  $P(a_i =_{IBD} a_k)$ . Given that the IBD state of this position will necessarily be one of the nine states described above, we can condition on these possible states and write:

$$P(a_i =_{IBD} a_k) = \sum_l P(a_i =_{IBD} a_k | IBD \text{ state } l) \Delta_l$$

Going through the possible pictograms in Box 1.2.4, it is clear for example that if the two alleles are drawn from a position in IBD state 1, then they will certainly be IBD. If the position has IBD state 2, 4, 6, or 9, then as there are no IBD connections between  $i$ 's alleles and  $k$ 's alleles then the two randomly drawn alleles cannot be in IBD. For states 3, 5, 7, and 8, the two alleles may be in IBD, depending on the random draws. The probabilities of selecting two alleles IBD from these states are  $\frac{1}{2}$ ,  $\frac{1}{2}$ ,  $\frac{1}{2}$ , and  $\frac{1}{4}$ , respectively. In this way we arrive at the following:

$$P(a_i =_{IBD} a_k) = \Delta_1 + \frac{1}{2}(\Delta_3 + \Delta_5 + \Delta_7) + \frac{1}{4}\Delta_8$$

The usual notation for this kinship coefficient is  $\varphi_{ik}$ . The ‘self-kinship’ comes from the same idea of randomly drawing two alleles from the same individual,  $\varphi_{ii}$  in effect. By considering each possible IBD case, we can see that:

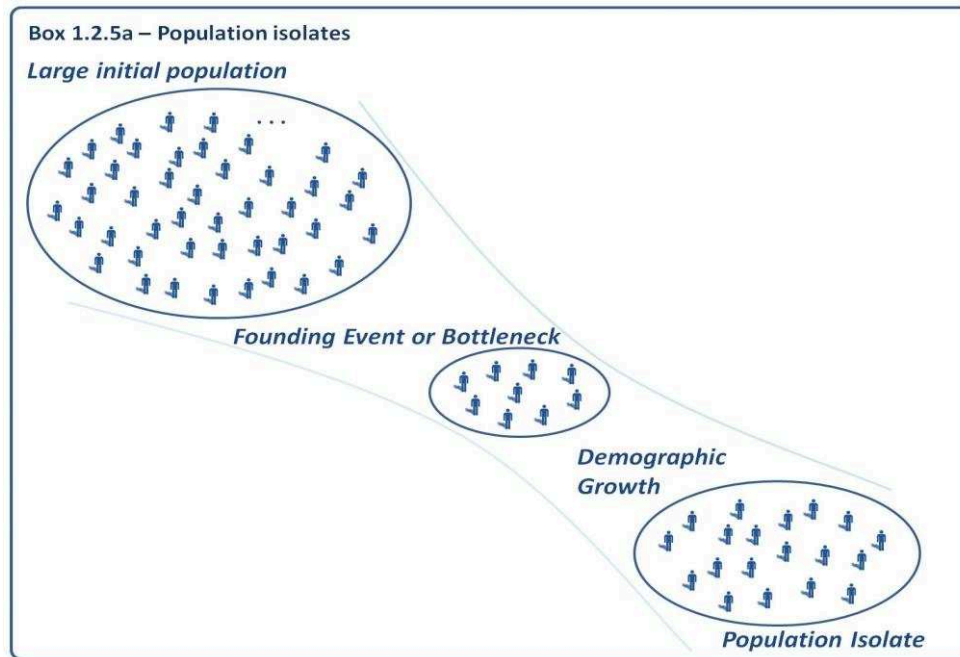
$$\varphi_{ii} = \Delta_1 + \Delta_2 + \Delta_3 + \Delta_4 + \frac{1}{2}\Delta_5 + \frac{1}{2}\Delta_6 + \frac{1}{2}\Delta_7 + \frac{1}{2}\Delta_8 + \frac{1}{2}\Delta_9$$

This we write as  $\frac{1}{2}(1 + f_i)$  where  $f_i$  is the inbreeding coefficient of individual  $i$  and is equal to  $\Delta_1 + \Delta_2 + \Delta_3 + \Delta_4$ ; this describes the probability of the individual having two alleles inherited from a single ancestor at a given position. The inbreeding coefficient of individual  $i$  is also observed to be the kinship coefficient of the two parents of individual  $i$ .

### 1.2.5 Population Isolates

There is no strict definition of an isolated population; here we depict an isolate by evoking first large population, followed by a bottleneck or founding event, and a subsequent demographic growth in isolation due to geographical or cultural factors (Box 1.2.5a). The attributes of the isolate will depend on the makeup of the population when it was at its smallest size or at the founding event and indeed the initial period of growth after the founding event that may involve a high percentage of marriages between related individuals. These attributes will likely represent a drift away from the initial large population.

In the opening remarks of this thesis, I gave the Icelandic population as an example of an isolated population. The key elements that characterise this particular population are: (a) the founding event - the period of settlement during which the first arrivals from Scandinavia and the British Isles began to reside permanently on Iceland; and (b) the expansion of the population in isolation from the rest of Europe. This is an over simplified description of the population history of Iceland as indeed subsequent bottleneck events linked to famine and epidemics<sup>6</sup> have increased the drift in the characteristics of this isolate.

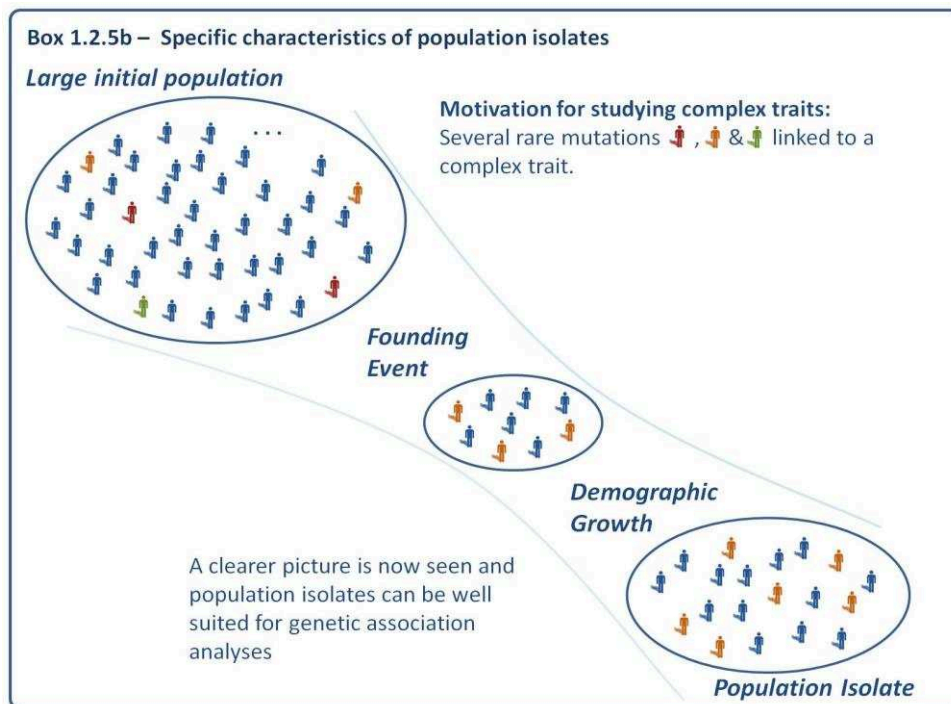


*Box 1.2.5a – Characterising an isolated population.*

The study of genetic isolate data began as a logical extension of studies of large families with many individuals affected by (mainly) Mendelian disorders<sup>67,68</sup>. The example of non-syndromic deafness in Bedouins (as given in Section 1.1.4) being a classic example. Advances in sequencing technology have brought the opportunity of performing association analyses with more complex traits on large numbers of individuals from within an isolate. Indeed, isolates hold many intrinsic advantages for association analysis.

The bottleneck and the subsequent period of growth will result in changes in allelic frequencies. This phenomenon is a particular example of ‘genetic drift’ first described by S. Wright<sup>69</sup>. This means that one should expect the frequencies of genetic variants in an isolate to be different from the population of origin. For a complex trait with a polygenic architecture, the divergent frequencies of causal variants in the isolate may facilitate the identification of said variants. Box 1.2.5b shows an isolate where the founding event has led to certain causal variants to disappear (as in no carriers of the alternative allele remain); and furthermore, some causal variants exhibit a higher minor allele frequency in the isolated population. Shortly after the founding event, the small size of the population will further drive this perturbation of allele frequencies.

Drift caused by a bottleneck can lead to specific presentations of the trait in question within the population. For example, the studies of the isolate Pima Indian populations have found unusually high rates of type 2 diabetes but also near null rates of type I diabetes<sup>70-72</sup>. Furthermore, background genetic variation in a population isolate will also be reduced due to the relatively small gene pool (or founding pool) from which all individuals are descendant. Therefore, there is less risk of potential false positive results from association analyses due to the homogeneous nature of the population<sup>20</sup>. In a general population, genetic variation can easily be confounded with environmental factors that affect a given trait. This can happen when individuals from multiple ethnicities are studied together, whose differing cultures and lifestyles may be associated with a particular trait. Stratification within a population is a common source of bias, but it is one that should be largely avoided when studying an isolate.



*Box 1.2.5 – Motivation for studying complex traits in isolates due to unique trait architecture.*

A further benefit of studying isolated populations is that all individuals can be assumed to share a similar environment. This is not a completely safe assumption; isolated populations will still contain a mix of smokers and non-smokers for example and so data for potentially important non-genetic

explanatory variables must still be collected. However, the fact that all members share a broadly similar lifestyle makes isolated populations appealing for complex trait analyses as it suggests that a larger proportion of the phenotypic variance should be genetic based and that there will be less noise affecting the trait. A final unwanted potential source of variation in the trait can come from different diagnosticians or measurement methods. Peltonen, et al.<sup>21</sup> give the example of studies in Finland where all clinicians are trained in one of five medical schools, all of which share academic traditions. In practical terms, studying an isolate can again be beneficial due to the practicality of having all individuals living nearby one another and with data collection being performed at perhaps a single location. The contained (geographical or cultural) nature of an isolate also facilitates the collection of genealogical data. For many isolates it has been possible to gather extensive pedigree information; the deCode project in Iceland and the Hutterite population being important examples.

Isolated populations will also have specific haplotype structures. From the point of view of population genetics, this is manifest as patterns of LD that are unique to the isolate due to the particularity of the founding individuals and the particular stochastic realisation of recombination and pairings in the population<sup>73-75</sup>. Notably, population isolates will often exhibit particularly long blocks of LD. From the point of view of family/pedigree based genetics; this specific haplotype structure can be described by relatedness between all individuals; with the expectation being that even pairs of individuals who are not closely related will still share very long haplotypes, or long stretches of IBD=1<sup>76-79</sup>. This property enables the use of many statistical methods based on IBD sharing in the study of isolated populations.

The degree to which isolates will display such noticeably specific characteristics will depend on their age, size and the nature of their bottlenecks or founding events<sup>20,21</sup>. Efforts have been made to classify and compare genetic isolates; by estimating the number of generations since the founding event<sup>80</sup>, levels of inbreeding<sup>81</sup>, and the (effective) size of the isolate<sup>82</sup>. The effective population size is a concept from the field of population genetics. For an isolate describes the size of a hypothetical

idealised population that would produce similar results when analysed equivalently to the true observed data from the isolate. Recently a formal index of isolation has been proposed by E. Zeggini's group<sup>83</sup>. Their measure, *I<sub>sx</sub>*, combines information regarding the time since the founding event, the level of genetic drift, and the effective population size. Such characteristics are important as they may determine the particular strength or weaknesses of an isolated population for gene-mapping (finding associations between genetic variants and complex traits). Younger isolates that still present relatively high inbreeding (following consanguineous pairings after the bottleneck) and extended LD may be particularly useful for identifying new genomic regions associated with a trait<sup>84,85</sup>. Older isolated populations carry the same advantages (if not to the same extent), but here larger sample sizes can be gathered. This can allow searches for rarer genetic variation and more precise localisations of disease variants can be found<sup>80</sup>.

### 1.2.6 Studying Complex Traits in Isolated Populations: Track Record

A myriad of multifactorial traits have been studied by genetic epidemiologists, and data from isolated populations have played an important role. Indeed, the current largest combined panel of human reference genomes, the Haplotype Reference Consortium<sup>86</sup> (HRC), contains many datasets arising from cohort studies of small founder populations including the MANOLIS<sup>87</sup> cohort, the Val Borbera<sup>73</sup> cohort, and cohorts from Sardinia<sup>88</sup>, Finland<sup>89</sup>, and the Orkney Islands<sup>90</sup>. I will give three case studies of three different traits analysed in three different isolates to demonstrate the concepts and potential of studying complex traits in isolated populations.

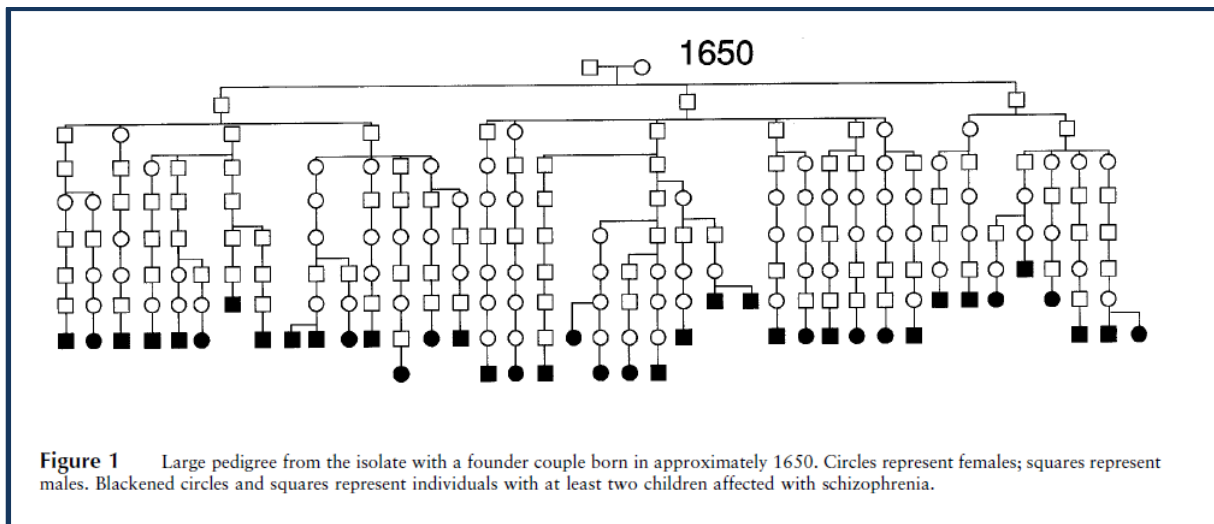
The first example is the study of Alzheimer's Disease (AD) in the Icelandic population. Iceland represents a near extreme example of an isolate in that (a) the modern day population is now of very large size, (b) it is an isolate derived from a large number of founding individuals, and thus (c) the expected IBD sharing probabilities between individuals is quite low compared to other isolates. This population has produced discoveries of genetic variation that affect the development of AD. I have already made reference to the discovery of a protective allele against AD found within the APP



gene in section 1.1.4. Other significant discoveries coming from Iceland include variants associated with increased risk of AD in the genes *TREM2*<sup>91</sup>, *TM2D3*<sup>92</sup>, and *ABCA7*<sup>93</sup>. The specific variant found in *TM2D3* had a MAF 10-fold greater (0.005 vs 0.0005) in the Icelandic study sample than the MAF observed in other European populations. This genetic drift leading to this ‘enrichment’ of a mutant allele, facilitating its own discovery through association analyses, highlights one of the aforementioned advantages of studying isolates for complex traits. Similarly, the variant found in *TREM2* had a MAF of 0.062 in Iceland compared to MAFs between 0.001 and 0.002 in the replication studies carried out in other European populations.

Note that these variants were still very rare in Iceland and that large sample sizes are required in order to spot associations with such variants and further characteristics of the population are also crucial to the success of such studies. In Steinberg, et al.<sup>93</sup>, we see that the associating testing that lead to the discovery of a rare variant in the *ABCA7* gene involved 3,419 individuals with AD (the cases) and 151,805 control individuals! There are two points of interest here, firstly all of the cases were diagnosed with AD using one of only two possible strict sets of established clinical criteria and were all enrolled at the same university hospital in Landspítali. Whilst we cannot vouch first hand for the quality of the study, it seems reasonable to presume an advantageous homogeneity in the phenotypic data. More impressive is the number of controls involved in the study, however the majority of this data was not directly sequenced but inferred from a set of 2,636 individuals with WGS data. Most individuals only had Array type data<sup>76</sup> but dense genetic data could be inferred by methods based on sharing of long haplotypes IBD between individuals and the presence of closely related individuals. Hence, the haplotypic structure of the population and opportunities in data collection made feasible the statistical analyses that led to these discoveries. Subsequently, these discoveries have been replicated in other populations and have improved biological understanding of AD<sup>94</sup>.

A second example is studies of Schizophrenia in Finland. Studies in Finland have been undertaken on the general Finnish population (which has similarities to Iceland) but also on certain small internal founder populations arising from settlements in the North and East of Finland. A description of the demographic history is given in Hovatta, et al.<sup>95</sup>; the first evidence of inhabitation in Finland dates from 11,000 years ago, but with the largest expansion in the population following a bottleneck around 4,000 years ago. The general population of Finland remained isolated for geographic and cultural reasons and in 16<sup>th</sup> and 17<sup>th</sup>; new settlements in the North and East resulted in small internal isolates. Rates for schizophrenia and related conditions had been observed to be above those measured in other parts of the world, particular in rural areas including the small internal founder populations<sup>95</sup>. An early analysis of families from one such isolate uncovered multiple possible loci associated with the trait<sup>96</sup>. Below, the genealogical connections between individuals with two children who have schizophrenia are shown (Figure 1.2.6a) for a large family from within the internal isolate study in Finland.



*Figure 1.2.6a - Pedigree structure from the north-east internal isolate in Finland; taken from Hovatta, et al.<sup>96</sup>*

One such region that was found was on chromosome 1, the signal would later be identified as coming from the gene DISC1. Originally, translocations in this gene were associated with Schizophrenia from family-based studies in Scotland<sup>97,98</sup>. However, the translocation in this family is

incredibly rare in general populations. It was from continued study of the Finnish populations (both the general population as well as further families from the small internal isolates) that the importance of the gene began to emerge<sup>99-102</sup>. Indeed, the presence of very specific haplotypes in the internal Finnish isolate proved to be particularly informative<sup>103,104</sup>. It is now a gene that has been studied extensively and is known to be of significant importance for many cognitive functions<sup>105</sup>. Interestingly, other known genes associated with schizophrenia have been shown to not play significant roles in the development of the trait in Finland<sup>106</sup>. Again, many of the advantages of studying isolated populations helped to aid this discovery. Schizophrenia is difficult to diagnose, but the clustering of cases in the small internal isolate was recognised and the study of which could be carried out uniformly between patients. The availability of genealogical data, and the haplotypic specificity of the population were also very important; plus, the advantage of a natural decrease in other possible genetic risk factors related to the affliction.

In a last example, we turn to studies of allergic Asthma in the Hutterite population. This population is an exemplum of a small founder population. A full description and detailed history of this population was given by A.P. Mange<sup>107</sup>. Originating from the 16<sup>th</sup> Century from communities in the Tyrolian Alps (modern day border zone between Italy and Austria), this religious sect moved East across Europe, before a group of about 800 individuals made the journey to the New World, settling in South Dakota. This population has existed in isolation due to resolute beliefs and desire to self-govern; to quote directly from A.P. Mange (who has a nice turn of phrase).

“Firm believers in God and Jesus Christ, in absolute pacifism, adult baptism, and in the avoidance of birth control measures, this group has successfully remained aloof from the rest of its species for over 400 years.”

“Since its origin in the Tyrol, the sect was subjected to attempts to change its religious beliefs or to enlist its aid in militaristic or other national goals. The usual response to these pressures (as is their response to pressures in twentieth century America) was to move.”

*Excerpts from ‘Growth and Inbreeding of a Human Isolate’, Mange<sup>107</sup>.*

In Figure 1.2.6b, the movements of the Hutterites across Europe are depicted. Importantly, there were many bottlenecks in the population size, the most well documented being shortly before the move from Wallachia to Russia where the size of the population dropped below 100. It is estimated that the modern day population can be traced back to just 64 founding individuals<sup>108</sup>. A pedigree structure of over 12,903 individuals over 13 generations, connecting 1,415 modern day individuals with genetic data to this small set of founders, has been assembled<sup>109-111</sup>. This has proved to be a very powerful dataset for genetic epidemiologists and for research into the theory of population genetics.

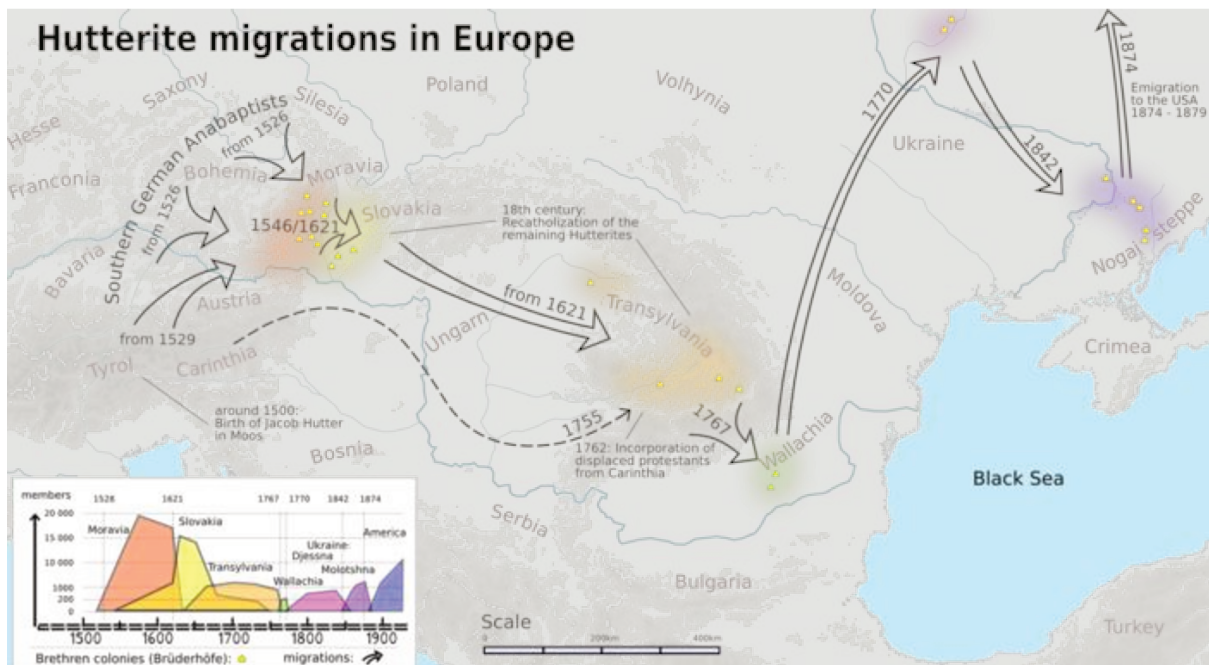


Figure 1.2.6b - Artistic interpretation of Hutterite movements throughout Europe. This is an isolate of many bottlenecks even before the small number of individuals founded the modern day colonies in South Dakota, USA. Source: <https://en.wikipedia.org/wiki/Hutterites>.

Here we discuss investigations of Asthma in the Hutterite population. This is a particularly complex trait to study due to its phenotypic variance and strong interplay with environmental factors. For a review of Asthma epidemiology, touching both on genetic but also non-genetic risk factors, see Ober and Yao<sup>112</sup> and references within. Studies of the Hutterite population for Asthma have gone from whole-genome linkage analysis<sup>113-115</sup>, through to GWAS<sup>116</sup> and beyond to studies looking at rare genetic variation<sup>117</sup>. Discoveries that bear up to scrutiny in Asthma genetics have often required

Meta-analyses across many populations, of which data from the Hutterites invariably plays a role. To highlight one in particular, investigation of the link between the trait and the HLA gene on chromosome 6 was aided greatly by the analysis of haplotypes in the Hutterite population<sup>118</sup> as this is a region that has high variability and is notoriously difficult to analyse. Further findings coming principally from studies of the Hutterites include the link with the genes CH13L1<sup>116</sup> and NEDD4L<sup>117</sup>.



Figure 1.2.6c – Artistic impression of GWAS hits for Asthma, taken from Ober<sup>119</sup>. The strongest GWAS signals are characterised as the fruit that hang lowest, which are thus easiest to capture.

In general however, the overall picture of genetic aetiology of Asthma still holds many mysteries and further discoveries will require more advanced analytical techniques and richer datasets (containing both genetic and non-genetic information)<sup>119</sup>. In Figure 1.2.6c, the ‘low hanging-fruit’ hypothesis is displayed, taken from C. Ober’s review “Asthma Genetics in the Post-GWAS Era”<sup>119</sup>. This shows the genes that have been discovered to be associated with Asthma as those that have descended most from the tree’s top; where this distance reflects the strength of association signals that are observed in GWAS. Different coloured vertical sections represent the 22 autosomal chromosomes. The idea

being that the techniques that have been applied so far are fine for finding the more obvious sources of Asthma variation in the genome; while new approaches will be required to explore the higher parts of the canopy. In the next section, I will explore the current role of isolated populations in a hyper-modern context.

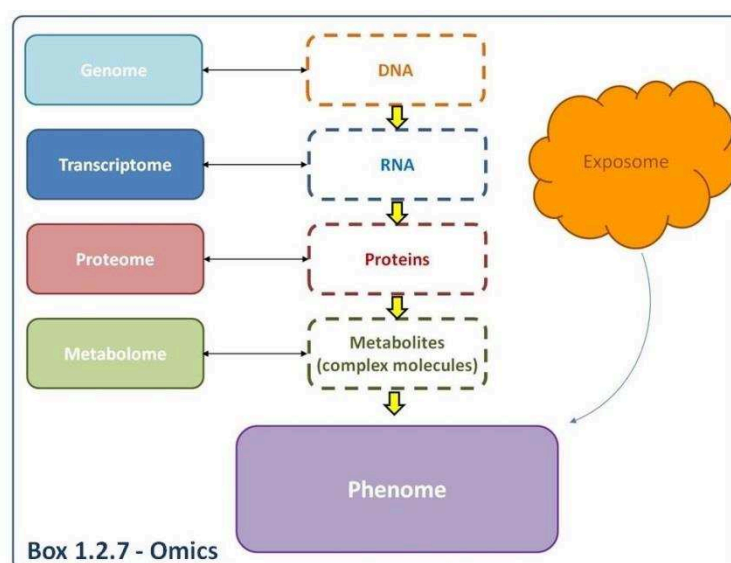
### 1.2.7 Studying Complex Traits in Isolated Populations: Prospective

Directly sequencing whole genomes in isolates can be motivated by a desire to gain high confidence in the genotyping and to find types of rare genetic variation that would otherwise not be observed. I will give a quick aside as to why the search for genetic variants that drive quantitative trait architecture has increasingly turned on to more complex and/or more unusual genetic variation. In the study of many traits, progress has encountered a hurdle in the well-known phenomenon of ‘missing heritability’. Heritability will be the central theme of Chapter 4, for now a simple definition can be given. When modelling such traits, either the quantitative phenotype itself or the odds of developing a disease may be transformed in order to follow a Gaussian distribution. The variance of this distribution can be written as  $V_Y$  and a general assumption is that this variance can be split into two components, one underpinned by genetics, the other by environmental factors. Therefore, we can write  $V_Y = V_G + V_E$ , the sum of the genetic variance  $V_G$  and the environmental variance  $V_E$ . The heritability of the trait is then the quotient  $V_G/V_Y$ . This describes the proportion of the observed variance of the trait that can be ascribed to inherited genetic factors. How can this heritability be ‘missing’ exactly? This refers to the observation that estimates of heritability using data from families are very often considerably higher than estimates from cohorts of unrelated individual based on observed genotypes<sup>120-123</sup>. Thus, we can have an estimate of the role of one’s whole genome, but this value cannot be attained when attempting to list and observe the particular elements that comprise this role. Furthermore, current lists of confirmed or putative causal variants can often only explain a fraction of the heritability estimated from family data. The meaningfulness of these gaps between estimations of heritability has often been overstated or wrongly

interpreted<sup>124</sup>. Missing heritability has, nevertheless, motivated the scientific community to explore approaches that are more sophisticated in order to make new discoveries, often concentrating on rare genetic variation.

The sequencing of genetic isolates in recent years has already supplied contributions to this end. Sequencing studies have recently been carried out in Greek isolates<sup>125,126</sup>; in Croatian isolates<sup>127</sup>; the Hutterites<sup>117</sup>; Sardinia<sup>88</sup>; Finland<sup>128</sup>; the Faroe Islands<sup>129</sup>; Ashkenazi Jews<sup>130</sup>, Qatar<sup>131</sup>, and French Canadians<sup>132</sup>. We have also already mentioned sequencing studies in Iceland<sup>133</sup>, Orkney<sup>90</sup>, and in a North Italian isolate (Val Borbera)<sup>73</sup>. Without exception, all studies listed above revealed numerous novel genetic variants and many new association signals. A notable example of a finding that involved very rare variants specific to the population is found in Gilly, et al.<sup>125</sup> who studied the MANOLIS isolate (mountainous villages from the island of Crete). Here, association between aggregated rare mutations in the FAM189B gene with measured levels of triglycerides were found; with the variants driving the signal only being uncovered through deep (high quality) whole-genome sequencing in the population. A concept that was developed in recent publications of the same group is that the small population sizes in isolated populations limit the action of purifying selection – the diminishing of deleterious mutations in the population<sup>83</sup>. In an isolate, certain variants (which are usually highly rare in general populations) will occur with elevated frequency. By considering the hypothesis of low effectiveness of selection, it can also be expected that these variants with higher allele frequencies found in isolates may also have high functionality. In two studies of recently attained WGS data (one on MANOLIS<sup>126</sup>, and one on data from many isolates<sup>83</sup>) enrichments of such rare variants with predicted high functionality. Similar enrichments were found in WES studies in Finland<sup>134</sup> and in the Vis island in Croatia<sup>127</sup>. This strengthens the idea that isolated populations should continue to be good sources of new association signals for complex traits and the recent increase in WGS data creation in isolates will likely soon produce further discoveries.

Sequencing technology and genetic imputation methods have greatly improved the precision of genome data. Looking more closely at the biological chain of events that links DNA and complex traits, there are clearly many other possible types of ‘omics’ data that can be gathered and analysed. There has been recent interest in adding similar precision for measurements of the transcriptome (levels of RNA transcribed from DNA), the proteome (levels of proteins generated by RNA), and the metabolome (the complex molecules such as lipids, sugars and amino acids with direct biological actions, (see Box 1.2.7). There are complex interactions between these different layers, and while DNA remains mostly constant, other omics data vary greatly between different cell types in the body. As of yet, few studies of isolated populations have begun to explore such data, though as described above, the functional classes of variants found in isolated populations are now often being considered. In Benton, et al. <sup>135</sup>, a study of the Norfolk Island isolate, a region on chromosome 1 was identified to be associated with multiple cardiovascular related traits and an exploration of gene expression data revealed that the locus was associated with transcripts of 55 genes. An investigation of these genes indicated a possible connection to a biological pathway controlling purine metabolism. Without plunging into too many details, the important message is that in this isolate, the transcriptome data for a range of genes had been previously shown to be heritable in the



*Box 1.2.7 – The different levels of ‘omics’ data.*



isolate<sup>136</sup>, and a connection was made between a significant GWAS hit in a non-coding (intragenic) region and the expression levels (amount of RNA produced) in distant and potentially relevant genes. This might suggest that the advantages of studying isolated populations could carry through to more complex analyses involving omics data.

What is also apparent when looking at Box 1.2.7, is the importance of gathering far more precise data regarding phenotypes (the phenome) and even environmental exposures (the exposome). As described above, population isolates have proven beneficial, as it may be the case that very specific forms of phenotypes can be observed and that data collection can be well regulated. Similarly, high quality data relating to environmental exposures can be obtained in isolates for precisely analogous reasons. Again, in the MANOLIS population, recent analysis has found both genetic and environmental explanations for rates of cardiovascular disease. A variant in the APOC3 gene, which occurs with an unusually high MAF in the cohort, was shown to be protective against heart disease as it was associated with low levels of certain lipids in the blood<sup>137</sup>. However, in this population, particularly poor dietary habits are quite common and augment the risks of such traits<sup>138</sup>. This highlights a potential for studies in isolated populations to be able to study both genetic and non-genetic risk factors with precision. Studies in isolates involving statistical interactions between genes and environment may also prove to be successful due to the lack of statistical noise in both factors. Again, by looking at the extreme examples of human genetics found in isolates, it may be possible to enhance biological understanding that can be applied to treatments and initiatives in wider populations.

However, the main research directions in genetic epidemiology which explore the architecture of complex traits (that often appears increasingly convoluted) are relying on methods that leverage information from huge numbers of individuals. The unavoidable weakness of studying population isolates is the limitations of sample size (outside of examples such as the Icelandic population), and hence statistical power for many analyses. The polygenic model, where 100s or even 1000s of genes

are implicit in a traits distribution, has been suggested as insufficient. The new ‘omnigenic’ model<sup>139,140</sup> has provided a working hypothesis for the observation that for many traits, it would appear that most regions of the genome are contributing to the heritability of a trait<sup>141,142</sup> and that pleiotropy is widespread<sup>143</sup>. Furthermore, this model highlights the involvement of genomic regions that regulate the expressions of ‘core’ genes, rather than focusing on the roles of these ‘core’ genes themselves. Note that debate continues as to the relevance of this recently proposed model<sup>144</sup>. The role of gene expression levels in different tissues in the body is also now being studied<sup>145</sup>. Assumptions involving huge numbers of variants, interacting in complex networks, each contributing tiny marginal effects, are hard to test. For many traits, GWAS are now operated by large consortiums performing meta-analyses on data recruited from many cohorts. In 2018, several meta-analyses were published involving over 1 million genomes<sup>146-148</sup>. What is more, a vogue has emerged for statistical methods that are based on exchanged summary test statistics from individual GWAS studies. Specifically, the test statistics and estimated effect sizes relating to each association test with each genetic variant in a GWAS have been shown to hold sufficient information to carry out further analysis without observing the underlying individual genotype data<sup>149</sup>.

It can therefore be hard to see what role studies of small genetic isolates will continue to play in the face of studies that involve huge sample sizes and the appearance of resources such as the UK Biobank<sup>150</sup> ( $N \sim 500,000$ ) in complex trait genetics. As has been discussed, the characteristics of isolates may give increased power for detecting signals, but hyper-rare variants or variation/patterns with very subtle effects may be beyond the scope of studies in population isolates.

Therefore, a necessary future direction for population isolates might be to combine resources. While each isolate will contain its own unique genetic characteristics, the signals found in one isolate have often been replicated in general populations. To give an example, a variant found in the gene *NDST3* strongly associated with schizophrenia in the Ashkenazi Jewish population was also shown to be associated with the trait in a range of further cosmopolitan samples<sup>151</sup>. Furthermore, replicated

results can occur between multiple isolates. I earlier referred to the heart-disease protective variant in APOC3 in a Greek isolate; this was indeed a replication result as this variant had already been found in an Amish founder population<sup>152</sup>. In Xue, et al.<sup>83</sup>, by looking at WGS data from several isolates, an illustration is given of how rare genetic variation with important predicted functions can be shared across different isolates. This provides hope of the potential of ensemble data from different isolates; indeed as far back as 2010, a successful meta-analysis was carried out on five isolates of different European origins<sup>153</sup>; finding associations between three genes and levels of creatinine serum (linked to kidney function).

Another area where data across multiple isolates has been studied is when looking at the effects of inbreeding. Runs of Homozygosity (ROH) describe sections of the genome where an individual harbours two haplotypes in IBD, and is therefore homozygous at every position across this IBD region. A recent review looked at the relationship between proportions of occurring ROH and complex traits<sup>154</sup>. Many studies that have so far looked at ROH are based on isolated populations, including the study of Joshi, et al.<sup>155</sup> which found associations between ROH levels and traits such as height and cognition. Isolated populations are suitable for such studies due to the presence of inbreeding.

So far when discussing the genetics of complex traits and GWAS in isolated populations, this has referred to tests under 'additive' models as this is by far the most common practice. The additive model (given in Section 1.1.5) assumes that the effect of a causal variant is linear in the number of minor alleles in the genotype. However, departures from this model are often observed, for example many Mendelian disorders follow recessive or dominant models. In a recessive model for example, the disorder will only occur in individuals with genotype containing two mutant alleles ( $aa$ ), and a dominant model would be where the individuals with genotypes  $Aa$  or  $aa$  will have the disorder. Isolated populations with high consanguinity such as the Hutterites have proved useful in detecting rare recessive autosomal disorders. Genetic variants involved in complex traits can also act (in part

or completely) in such non-additive ways; this will be fully explored in Chapter 4. When performing additive GWAS, there is far less power to detect variants with non-additive effects, particularly for those that act recessively<sup>156</sup>. By considering isolated populations, where both inbreeding and pairs of individuals sharing regions IBD=2 may be present, it may well be possible to investigate such genetic variation far more effectively than in general populations.

To summarise, the important future directions for isolated populations will involve attaining high quality WGS data in isolates, looking at more creative hypothesis and models for complex trait gene-mapping, and gathering data from multiple isolates and meta-analyses.

### **1.2.8 Applying Knowledge of the Genetics of Complex Traits**

Finally, what is the motivation of the search for associations between genetic variants and complex traits? Common multifactorial diseases place an enormous burden on modern health-care resources as well as debilitating vast numbers of individuals. If each individual's susceptibility to a certain trait depends on what is possibly a huge number of genetic variants, it can be difficult to imagine what solutions could be laid out from the increasing lists of GWAS signals. However, finding such association signals has led to many successes. Identifying a locus associated with a trait can be the first step in understanding a particular biological functionality and how one's DNA can affect such a function. This in turn can lead to possible therapies and the development of new medications. This is by no means the case for every GWAS hit discovered but a small number have been involved directly in new treatments<sup>157</sup>. Another important benefit of studying complex trait genetics is in diagnosis. It has become clear that traits such as Asthma are really in fact umbrella descriptions of many phenotypes<sup>158</sup>. Understanding underlying genetic effects leads to the possibility of far more pertinent descriptions of each individual's affliction.

This leads to the concept of 'precision medicine', where an individual's treatment is tailored to their genome and this is already an active area of modern medicine. One important example is the now

routine tests for a particular genetic variant in the HLA-B gene that was shown to be linked with extreme sensitivity to the drug Abacavir used in the treatment of HIV<sup>159</sup>. Individuals carrying the rare allele are prescribed alternate treatment. In the field of oncogenomics, there have long existed treatment protocols that depend on the genetic makeup of an individual's tumour cells as well as genetic scans to assess pre-symptomatic risk<sup>160-162</sup>. For complex traits, there may be currently less examples but many studies have begun to assess the relationship between genetic variation and possible treatments. For example, individuals carrying mutations in the DRD2 gene have been shown to respond particularly well to certain treatments for Parkinson's Disease<sup>163,164</sup>. It is likely that in the coming decades, more and more of the findings from research into complex trait genetics will find their way into clinicians' and diagnosticians' toolboxes. There is also the possibility of predicting the development of phenotypes from genotypes; this being relevant mostly for traits with a late age of onset. This approach that has been explored extensively, though often with limited success, and debate continues as to the utility and practicality of such prediction for many traits<sup>165</sup>.

As alluded to in previous sections, the genetic architectures of complex traits have proven to fully live up its billing – they are highly complex. However, this should not discourage. Whilst the measureable overall impact of genetic epidemiological studies into complex traits may have not met certain expectations, the proven successes so far and the current velocity of research indicate that the progress in subsequent decades should be substantial.



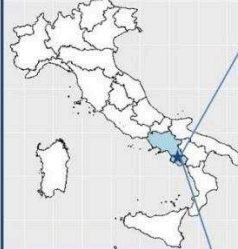
## Chapter 2: The Cilento Isolate, real and simulated versions

### 2.1 Gioi, Cardile, and Campora

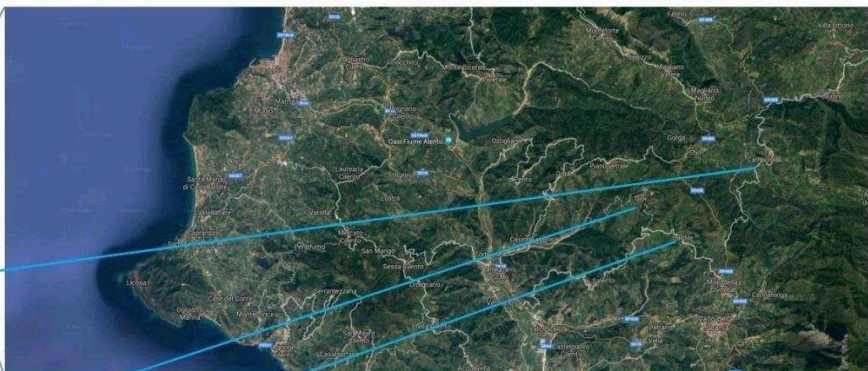
For this project, we have access to a very valuable resource - genetic data from the genetic isolates of Cilento. We have access to this data thanks to our continued collaboration with Dr. Marina Ciullo, Dr. Teresa Nutile, and Dr. Daniela Ruggiero based at the Institute of Genetics and Biophysics, A. Buzzati-Traveso – CNR in Naples, Italy. This dataset comes from three remote hill villages, Campora, Cardile and Gioi located in the National Park of Cilento in Southern Italy (Box 2.1).


**Box 2.1 The Cilento Isolates**

Genetic Park of Cilento and Vallo di Diano Project




In this thesis, we will extensively study data from three known genetic isolates in Southern Italy.






CAMPORA



Cardile

- Genetic data has been collected for 1,617 individuals from three remote villages in the Cilento region.
- Genealogical data from 7,585 individuals dating back to the 16<sup>th</sup> Century have been gathered from marriage records in local parishes.



Gioi

#### *Box 2.1 - The Cilento Isolates*

The area was likely to have been first settled by individuals of Greek origin, in the 8<sup>th</sup> Century BC. The region was subsequently conquered and then reconquered by the Lucanians (Italic tribe of Southern Italy) and the Greeks, respectively. The grouping of inhabitants into significant villages in this area occurred later on in the 10<sup>th</sup> and 11<sup>th</sup> centuries. This change was driving by the presence of monks from the Byzantine Empire who had fled the coastline and the frequent Saracen raids. Records of a

settlement in Campora do date back to the around the 5<sup>th</sup> Century BC, before the arrival of the monastics around the 11<sup>th</sup> century AD. Gioi first appeared in the 9<sup>th</sup> Century AD, and Cardile seems to have (at least in part) been founded from individuals coming from Gioi in the 11<sup>th</sup> Century<sup>81,166,167</sup>.

Importantly, the region suffered dramatically from an outbreak of the Plague in the 17<sup>th</sup> Century AD which greatly reduced the population size. This caused a bottleneck in the populations of the three villages. The region then experienced isolation until the end of the Second World War. Since, the populations of all the three villages have decreased in size due to outward migration. Hence, the three villages represent an example of a small and young genetic isolate. Genetic studies of the three villages began early in the new millennium<sup>168</sup> establishing the isolation of the region<sup>81,166</sup> and leading to publications relating to obesity<sup>169</sup>, behaviour of smokers<sup>170</sup>, vascular endothelial growth factor (VEGF) serum levels<sup>171</sup>, Placental Growth Factor serum levels (PGF)<sup>172</sup>, and CRIPTO serum levels<sup>173</sup>. Very recently, evidence has been found of individuals from Cilento who either suffer from CADASIL disease or Pseudoxanthoma Elasticum (both are Mendelian disorders). This was based on the detection of known disease alleles in the population through whole-exome sequencing<sup>174</sup>.

The study sample is composed of 2,304 individuals with deep phenotyping (anthropometric, cardiometabolic, and haematological traits) and detailed health status information (structured questionnaires and clinical records). For 1,617 individuals we have dense marker genotyping data (~600,000 mostly common variants spread across the genome) and in addition we have deep exome sequencing data (~400,000 mostly rare variants in protein coding genes) for 247 of these individuals which were generated at the Sanger Institute, UK. These 247 individuals form a study specific panel for imputation. A subset of 19 individuals (with some overlap to the 247 WES individuals) has also been sequenced at the Sanger Institute across the whole genome. The 1,617 individuals with Array data were not all sequenced on the same chip: individuals from Campora and Cardile have been genotyped on an Illumina 370 K array (370K), whilst individuals from Gioi have been genotyped on an Illumina HumanOmniExpress array (OMNI). The WES and WGS datasets are



both of high quality; having mean sequencing depths (a measure of the thoroughness of the sequencing) of  $\sim 75x$  and  $\sim 50x$ , respectively.

Genetic imputation has been carried out in Cilento, a process that imputes genetic data at positions not found on genotyping arrays. This imputation was achieved using the publically available panel of cosmopolitan reference haplotypes from the 1000 Genomes Project<sup>175,176</sup>. This panel consists of 2,504 individuals with high quality WGS data from a wide range of global populations. Methods for genetic imputation will be a focus of Chapter 3 of this thesis. The imputation in Cilento provided the possibility to approximate WGS data for all 1,617 individuals in Cilento. This has enabled more in depth analysis of the Cilento dataset and has made possible collaborative meta-analyses that have combined data from Cilento with data from other populations. In 2016, a meta-analysis led by our collaborators in Naples using data from Cilento uncovered six novel loci associated with VEGF serum levels<sup>177</sup>. This meta-analysis incorporated results from diverse sources including the Framingham Heart Study, Icelandic and Sardinian cohorts, and the Val Borbera isolate of Northern Italy. Cilento has also recently participated on a large meta-analysis on the effects of homozygosity and inbreeding which including many European Isolates<sup>155</sup>.

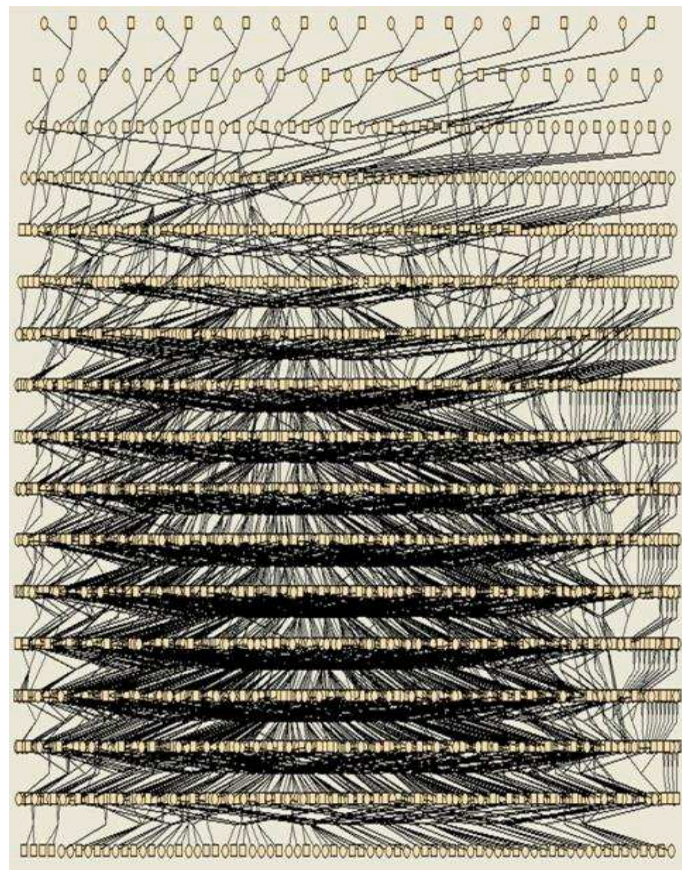
Previous analysis of the Cilento isolates has examined the average inbreeding and kinship coefficients between pairs of individuals<sup>81,166</sup>. The three villages have lower coefficients than the Hutterite populations. Levels of relatedness for Cilento fell slightly below estimates from previous studies of isolates from Sardinia<sup>178</sup> and slightly above estimates from the Icelandic population<sup>179</sup>. The most important bottleneck involved in the history of these villages occurred quite recently (17<sup>th</sup> Century). The Cilento isolates can be characterised as young (< 20 generations) with mild inbreeding. In this thesis we will mostly treat Cilento as a single isolate, though the three villages can be thought of as representing three distinct isolates. Campora is the least similar to the other two villages. Gioi and Cardile have a shared origin, being estimated to have separated approximately 1000 years ago<sup>166</sup>. In a recent study, the WES data from Cilento has been explored where many

novel variants, variants with allele frequencies that appear to have drifted, as well as rare, disease-causing variants have been found<sup>174</sup>. This analysis of the WES data has further explored the consequences and the evidence of isolation of the three villages of Cilento.

The aforementioned successes from analyses of the Cilento isolate and the recent efforts to gain sequencing data in Cilento indicate that continued studies of the genetics of this isolate will lead to further important results.

## 2.2 Integral Simulation Study

In this thesis, we will analyse both the Cilento dataset and simulated datasets with the same structure as Cilento. In order to create realistic simulated data with similar properties to Cilento, we used the software Genedrop<sup>180</sup>. This method uses the pedigree structure of Cilento. Below, in Figure 2.2, the pedigree structure for part of the Cilento population is given.



*Figure 2.2 – Pedigree structure of Campora village, 17 generations are represented. This structure will be used to simulate data representative of isolated populations.*

To carry out gene-dropping, first, one provides phased genotypes for each founding individual in the pedigree. Therefore, for each chromosome, each founder receives two complete haplotypes. Second, transmission at each meiosis and stochastic recombination events, based on genetic distance in cM, are simulated. Working through all lineages, mosaics of the initial founding haplotypes will be giving to every member of the pedigree. We used publically available WGS haplotype data from the UK10K imputation reference panel<sup>181</sup> (UK10K) as our source of founding

haplotypes. The combined pedigree of Cilento (incorporating family histories of the three villages, Gioi, Cardile and Campora) has 1470 founding individuals. We used a similar application of the program Genedrop to Gazal, et al.<sup>182</sup> in order to simulated whole-genome sequence data. Explicitly, we only applied Genedrop on a sparse set (or grid) of genetic markers, thus only allowing recombination events to occur at a limited number of positions. Then, in a second step, more precise recombination locations were simulated by selecting random WGS positions in between adjacent grid markers. Care was taken to ensure that inherited haplotypes composed of previously recombined haplotypes in the pedigree would also inherit the same precise recombination locations. This elaborate process was necessary due to the internal limitations of Genedrop regarding number of genetic variants, but we were able to build this complementary routine to Genedrop in order to simulate WGS data.

Using this technique, we simulated WGS data for 1,444 individuals in the connected pedigree of the three villages. We performed this simulation six times using different random draws of founding haplotypes with independent realisations of gene-dropping on to the pedigree. We simulated 22,989,093 genetic variants in this was across the 22 autosomal chromosomes. These simulation datasets which have the population structure of Cilento will be called upon to test different statistical methods and hypothesis.

The required number of founding haplotypes to run the simulation was  $2 \times 1470 = 2940$ . However, a previous study of the village of Campora which analysed data from Y chromosomes and mitochondrial DNA<sup>81</sup> suggested that the true founding pool of Cilento should be far smaller. Y-chromosomes are transmitted directly from father to son without recombination (aside from a very small region which is homologous to the X chromosome) and mitochondrial DNA is passed directly from mother to offspring. Hence, analyses of the observed variability of these two unique regions of the genome can give a good estimation of the number of patrilineal and matrilineal ancestors – i.e. the number of founders. It was estimated that 97% of the genetic diversity in the population of

Campora originated from 17 men and 20 women. We extrapolated from this finding and assumed that every member of the Cilento dataset (three villages) has descended from approximately 200 founding haplotypes. We used the software HapGen2<sup>183</sup> to account for this problem. In each simulation iteration, we randomly sampled 200 haplotypes from the UK10K panel, and using HapGen2, we created a pool of mosaic of size 2,940 as to be able to draw the number of haplotypes required to gene-drop onto the Cilento pedigree. The idea of including this HapGen2 step was to simulate the unrecorded links between the known pedigree structure and the founding individuals.

In addition to these six complete simulated populations, we also simulated 200 versions of WGS data for chromosome 10 for 477 individuals from Campora which will be used for testing phasing and imputation software (Chapter 3). Note that here, for the large set of versions of chromosome 10 that were simulated, for 100 iterations we included this HapGen2 step in the simulation, but in the other 100 iterations we skipped HapGen2 and sampled founding haplotypes directly from the UK10K panel. Further details of our data simulation are found in Annex A.



## Chapter 3: Phasing and Imputation in Isolated Populations

### 3.1 Experimental Design to Investigate Phasing and Imputation Accuracy

The first main discussion of the thesis will centre on one practical aspect of genetic studies - the obtaining of genetic data. There is something very alluring (to statisticians at least) about the fact that we each have our own precise genetic coding - that our biology is so readily described by a large ensemble of discrete values. Furthermore, the fact that each genetic variant is hidden from view and is inherited completely at random gives strength to causative hypothesis that link genetic values to observable trait characteristics. Whilst this seems ideal, our DNA is largely inaccessible and has required years of dedicated technological innovation in order to collect this precious data.

The cost of such full genetic sequencing is still relatively high. Even if a full human genome has come below 1000 US dollars, the hypothetical cost of whole-genome sequencing the ~2000 member of the Cilento cohort is still considerable. WGS, WES, and array data were introduced in Section 1.1.2 and these formats would appear in the same order if ranked by descending monetary cost. What is more, the quality of WGS data depends primarily on the thoroughness or 'depth' of the sequencing. Simply put, the depth describes the number of times each genetic marker will be read by the sequencing machine and the higher the depth the more accurate and rich the final dataset will be. Hence, a researcher will be faced by a difficult choice when deciding how to best allocate funds for gathering genetic data due to the large menu of different platforms available as well as the desire to attain genotypes for as many individuals as possible.

However, in many ways the composition of the human genome behaves in a reasonably predictable way. Most variants lie within a set of inter-correlated variants; termed a block of LD<sup>184,185</sup> (see Section 1.2.3). Such structure can allow for the inference of unobserved genetic variants from observed neighbouring variants.

Missing genotype imputation was first achieved by considering the inheritance patterns between pairs of related individuals and genetic imputation was first applied to family data<sup>186-188</sup>. In certain pedigrees, exact genotype imputation can be straight forward. In other cases, an expected genotype can be calculated by calculating likelihoods of genotypes over the pedigree – i.e. using the Elston-Stewart or Lander-Green-Kruglyak algorithm. By expected genotype we mean the expected count of minor alleles and so software (e.g. MERLIN<sup>189,190</sup> and MENDEL<sup>191,192</sup>) were adapted to work with genotypes that no longer had to strictly take a value from the set  $\{0,1,2\}$  (the minor allele counts of genotypes  $AA, Aa$  and  $aa$ , respectively) but could be continuous in the interval  $[0,2]$ . The importance of imputation is both intuitive in that it gives the researcher more information to play with as well as demonstrable in that in terms of increase in statistical power of subsequent analyses. Methods that worked with pedigree information were overtaken as genetic imputation was extended to non-related pairs of individuals. New methods were developed based on the structures of LD in the genome, and the fore-runners of such methods were fastPHASE<sup>193</sup>, MaCH<sup>194</sup>, BEAGLE<sup>195</sup>, and IMPUTE<sup>196</sup>. Over the last 10 years, three groups have continued to improve their algorithms; at time of writing, the latest available software are MINIMAC4<sup>197</sup>, BEAGLE (version 5.0)<sup>198</sup>, and IMPUTE4<sup>150</sup>.

Our first large analysis was to examine different strategies for genetic imputation in isolated populations. We considered state of the art software as well as various informatics pipelines that could be applied to the Cilento data, seeking to find the approach that would lead to the most accurate genome-wide imputation in our dataset. We focussed on the most common and computationally tractable pipeline for imputation which is a two stage approach; first haplotype phase is estimated for the SNP array data for all individuals, and then haplotype imputation is performed using a haplotype reference panel. This two stage strategy was introduced in Howie, et al.<sup>199</sup> and has been shown to be an effective approach. Our analysis focused on the impact of different software choices for the phasing and imputation steps, as well as the choice of reference panel for the imputation step.



In order to do so, a large scale simulation study was carried out. Here we used the multiple simulated versions of chromosome 10 as described in section 2.2; using the two simulation strategies that we termed ‘Pedigree’ and ‘HapGen+Pedigree’ in Annex A, and will also do so here (see Figure 3.1). Having simulated WGS data from the structure to the village of Campora, we obscured the positions not present on the SNP genotyping arrays, re-imputed as many missing positions as possible, and then compared imputed genotypes to the true simulated genotypes. The true haplotype phase was also known from the simulation, but would be jumbled and then re-estimated. We tested different methods for haplotype phasing across the entirety of chromosome 10 and tested different imputation software on the 20Mb telomeric region of the short arm of chromosome 10 (the first 20 million base pairs of chromosome 10, reading from the top).

In Figure 3.1, a schematic of the whole study is given, including the two different methods of simulating founder haplotypes, the gene-dropping process, and the phasing and imputation pipelines used. For genetic imputation, a bank of reference haplotypes is required. In our study we used principally the latest version of the 1000 Genomes panel<sup>176</sup> (1000G). We also tested the HRC panel as well as small Study Specific Panels (SSPs) created from our simulated data.

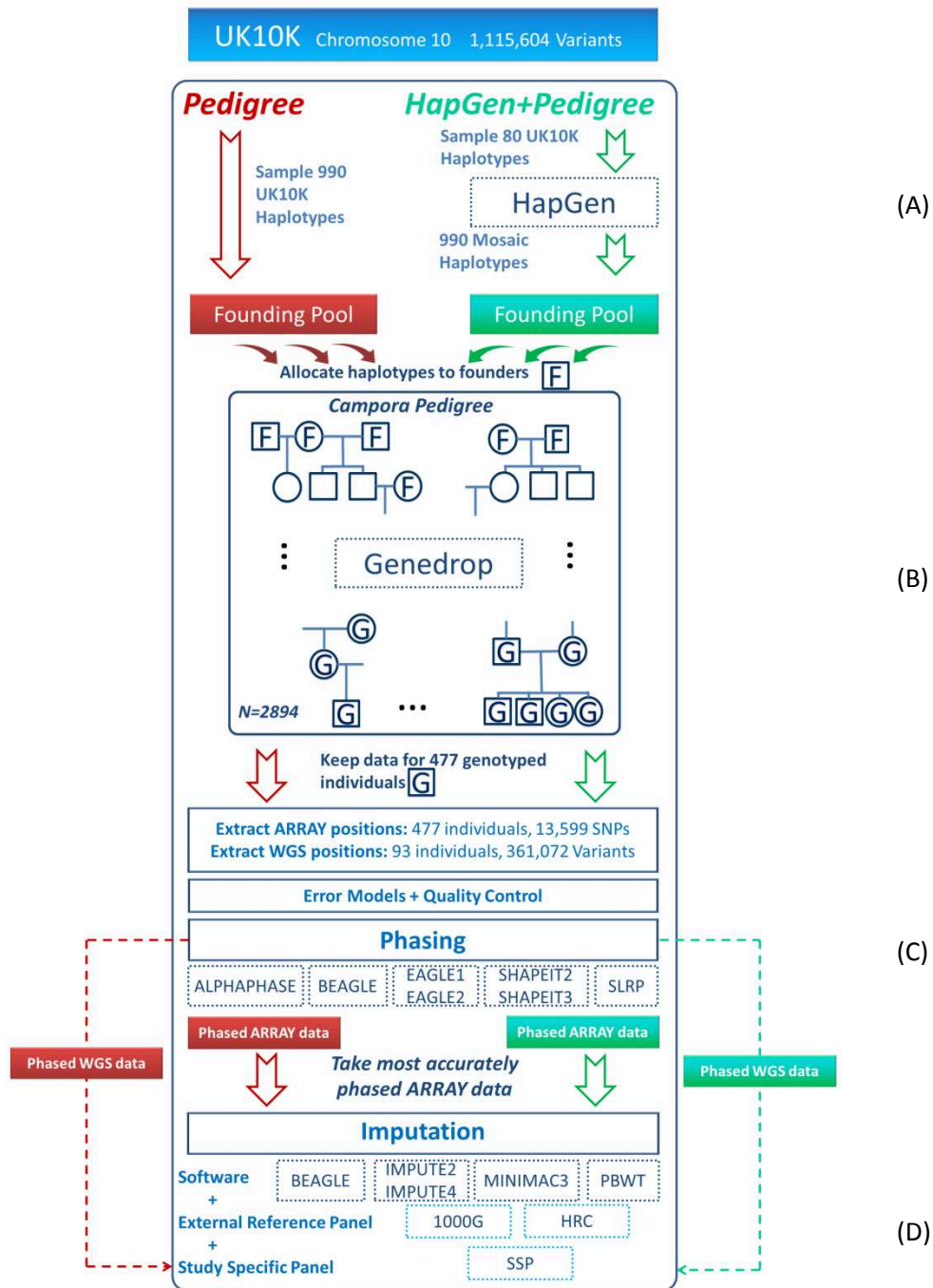


Figure 3.1. Schematic of Campora simulation study. Taken from supplementary materials of Annex A.

- (A) Founding event simulated in two different ways; either direct draws from the UK10K, or from a pool of mosaic haplotypes created by HapGen2 from 80 UK10K haplotypes.
- (B) Gene-dropping through the Campora pedigree.
- (C) Phasing is performed on the simulated array data using a range of different software.
- (D) Taking the most accurately phased data, imputation is performed using a range of different software and reference panels; including a local study specific panel built from simulated WGS data for 93 individuals.

## 3.2 Phasing

### 3.2.1 Review of Haplotype Phasing Methods

First we will focus on the different possible phasing algorithms that could be used. Table 3.2.1 details the different software that we analysed and the methodological approaches behind their algorithms.

<b>Table 3.2.1</b> <b>Software</b>	<b>References for software(s)</b>	<b>Methodology</b>	<b>Algorithms based on:</b>	<b>References for algorithm(s)</b>
<b>SHAPEIT2/SHAPEIT3</b>	200 201	LD-based	Li-Stephens model	202
<b>EAGLE 1/EAGLE2</b>	203 204	IBD- and LD-based	Long Range phasing and/or haplotype clustering	203 204 205
<b>BEAGLE</b>	195 206	LD-based	Haplotype clusters	207
<b>SLRP</b>	208	IBD-based	Long range phasing	76 209
<b>ALPHAPHASE</b>	210	IBD- and pedigree based.	Long range phasing and Pedigree based.	76 210

*Table 3.2.1 – Phasing algorithms compared in our study.*

We chose to test methods that were either recently published, had been shown to give accurate phasing results in previous comparisons, and which were based on algorithms we perceived to be potentially appropriate for isolated populations. The only previous comparison of phasing algorithms when applied to data from isolated populations is found in O'Connell, et al.<sup>211</sup> Here, the best performing software was SHAPEIT2; what our study would add to these findings was to perform a more rigorous test of phasing algorithms via repeated simulations; and to include the new software EAGLE whose combination of IBD-based and LD-based methods appeared to be very promising. All above methods require a substantial number of individuals to work with when phasing; in our simulation we tested these software on samples of 477 individuals (which is sufficient, though higher sample sizes have been widely shown to increase phasing accuracy; e.g. Loh et al.<sup>204</sup>).

First I will give further details of two of the involved algorithms, (1) Long Range Phasing (LRP) which characterises all of the IBD-based phasing methods, and (2) the Li-Stephens model<sup>202</sup> as SHAPEIT2

was the best performing software and most LD-based phasing and imputation software use a form of this underlying model. The model of BEAGLE for example, is conceptually similar as the Li-Stephens model (broadly speaking), though the modelling method is quite different<sup>212</sup>.

*(1) IBD-based methods and Long Range Phasing*

LRP was first introduced in Kong, et al.<sup>76</sup> and used on the Icelandic population by the deCode group. This was then mathematically formalised and implemented in an open source software SLRP<sup>208</sup> by a team from the Sanger Institute, UK. LRP focuses on uncovering long stretches of IBD between individuals and indeed across chains of individuals. In essence, LRP seeks to infer phase in an individual  $i$  (and also to impute un-typed genotypes) by regarding the haplotypes of individuals who share a haplotype IBD with  $i$ . This group of IBD sharers with  $i$  are termed as the surrogate parents of individual  $i$ .

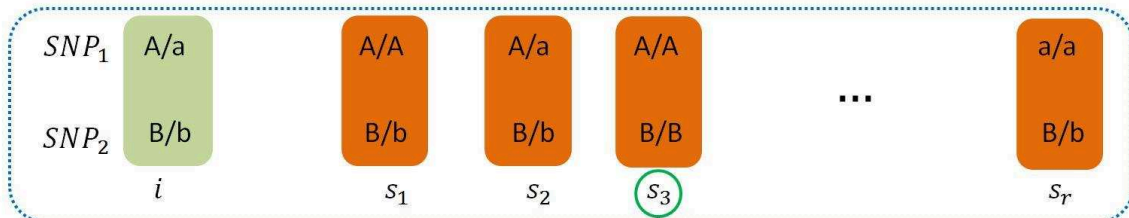
**Box 3.2.1a – Long Range Phasing**

An individual  $i$  (green) is heterozygous at two nearby positions, SNP1 and SNP2; the phase is unknown.

**Step1.** Search through the rest of the sample for individuals who share one haplotype IBD with individual  $i$ . Here the set is enumerated  $s_1, s_2, s_3, \dots, s_r$  (orange).

Long IBD matches can be found in various ways, a plausible starting assumption is that IBD stretches should span a minimum distance (e.g. 4cM). Finding such regions where  $IBS \geq 1$  can infer an IBD match.

**Step 2.** Search through the IBD matches (often referred to as surrogate parents) for phase informative matches.



Individual  $s_3$  is phase informative for  $i$ . Individual  $s_3$  has two copies of the A-B haplotype, one of which is shared with individual  $i$ . Hence, individual  $i$  must have the following haplotypes: A-B and a-b.

*Box 3.2.1a – Central ideas of the Long Range Phasing algorithm taken from Kong, et al.<sup>76</sup>*

In the genomic region where IBD has been found and phase is being estimated, this group who share a haplotype IBD are locally as closely related as parents to individual  $i$ . LRP was designed to be effective in samples of both closely related and distantly related (separated by  $\sim 1$ -20 meiosis) pairs. Thus, avoiding potential limitations of family based methods requiring close family members and LD-based methods for unrelated individuals. An outline of the method of long range phasing is given in Box 3.2.1a.

To search for IBD between two individuals, LRP will first look for evidence against IBD. Whenever individuals have opposite homozygous genotypes ( $AA$  vs.  $aa$  which can be described as  $IBS=0$ ), then there can be no IBD-sharing. If over a large region (e.g. across  $4cM^{204}$  or 1000 genotyping array SNPs<sup>76</sup>), there is no instance of  $IBS=0$ , then the algorithm will take this as potential evidence of  $IBD>1$ . After these searches have been made across all pairs and across the whole chromosome, the phase between pairs of heterozygous sites can be inferred. In Box 3.2.1a, in a region containing two SNPs, individual  $i$  was matched to individuals  $s_1, \dots, s_r$ ; an internal test having established this set of surrogates for individual  $i$  to indeed be a set of individuals who share at least one haplotype with  $i$  in IBD. As individual  $s_3$  is trivially phased, this informs the phase of individual  $i$ ; note that once the phase of individual  $i$  is known, in the example given, the phase of individuals  $s_1, s_2$ , and  $s_r$  can also be derived. Working through the network of connections, LRP will try to estimate phase for every position; however where IBD sharing cannot be assumed, or phase informative matches cannot be found, phase cannot be estimated. Hence, the original LRP algorithm outlined by Kong, et al. <sup>76</sup>, when first tested on 35,528 Icelandic individuals, was able to phase approximately 95% of all heterozygous sites. SLRP, a phasing software that further developed the original LRP algorithm, when presented in Palin, et al. <sup>208</sup> was able to phase up to 93% of all sites on simulated isolate type data; though in O'Connell, et al. <sup>211</sup> (who tested SLRP across populations with a range of characteristics), the best yield of SLRP that they observed was 88% for data from the Orkney Islands isolate.

(2) LD-based methods and the Li-Stephens model

The majority of LD-based methods use Hidden Markov Modelling. A brief description of a Hidden Markov Model is given in Box 3.2.1b. A chain of hidden (unobserved) events, following a Markovian process, emit observable values at each step. From these observations, inference can be made about the hidden path as well as the mechanisms that govern the transition between hidden states and emission of observed elements.

**Box 3.2.1b – Hidden Markov Models**

Let's take an unobserved ordered sequence of events:  $U_j, j = 1, 2, \dots, M$   
 and a sequence of associated observed events:  $F_j, j = 1, 2, \dots, M$

<b>Unobserved</b>	$U_1$	→	$U_2$	→	$U_3$	→	$U_4$	...
	↓		↓		↓		↓	
<b>Observed</b>	$F_1$		$F_2$		$F_3$		$F_4$	...

➤ How can we build a model for this situation?

1. Initial Probabilities:  $Prob(U_1 = u_r), i \in \{1, 2, 3, \dots, R\}$
2. Transition Probabilities:  $T_{rs} = Prob(U_j = u_s | U_{j-1} = u_r), r, s \in \{1, 2, 3, \dots, R\}$

**Markov Property: The sequence of hidden states,  $U_j$ , is completely connected and memoryless:  $P(U_j | U_{j-1}, \dots, U_1) = P(U_j | U_{j-1})$ .**

3. Emission Probabilities:  $E_{rv} = Prob(F_j = f_v | U_j = u_r), j \in \{1, 2, 3, \dots, n\}, v \in \{1, 2, 3, \dots, V\}$

By observing the sequence,  $F$ , we may either wish to estimate the hidden path ( $U_i$ ) or the probabilities  $T_{rs}$  or  $E_{rv}$ .

*Box 3.2.1b – The concept of the Hidden Markov Model. An unobserved Markovian sequence emitting an observable sequence from which inference can be made. The Hidden Markov Models described in this chapter will describe unobserved sequences taking values at every position on a chromosome. Hence, we re-use the subscript  $j$  running from 1 to  $M$  which we have usually used to describe a long list of genetic variants.*

Hidden Markov Models (HMMs) are attractive tools for modelling genetic data as it puts at one's disposal many existing algorithms that make inference from HMMs that can give solutions to a many interesting questions. To give examples: (a) the Forward algorithm for estimating the likelihoods of the observed states<sup>213</sup>; (b) the Forward-Backward algorithm for estimating posterior probabilities of hidden states<sup>214</sup>; and (c) the Viterbi algorithm for estimating the most likely hidden path<sup>215,216</sup>. For

further elaboration on HMMs, we refer the reader to work by L. R. Rabiner<sup>217</sup>, and to useful introductory level summaries by V. A. Petrushin<sup>218,219</sup>.

We will explore in detail the Li-Stephens HMM<sup>193,202</sup> which is used by SHAPEIT2 as an example of an LD-based method in order to contrast against the long-range phasing method already given above.

The philosophy of the Li-Stephens model is given in Box 3.2.1c, where we can see that the model describes a process where a new haplotype is built from segments of already observed haplotypes. As the model allows for small discrepancies within the segments that are ‘copied’ from known haplotypes, this is often described as a process of imperfect mosaics. The hidden chain in this HMM is realised at each sequential position on a chromosome. The hidden states in this HMM are elements of the set of known haplotypes from which the new haplotype is being copied – often referred to as ‘copying states’.

The transition states describe a Poisson process relating to the number of recombination events between two adjacent positions. Either there is no recombination event, so the copying state remains the same; or if there is at least one recombination event, it is assumed that a new copying state is drawn at random (including the possibility of returning to the current state) which is why we see the factor  $\left(\frac{1}{N}\right)$  associated with the Poisson probability of at least one event:  $1 - e^{-\lambda_j}$ . This recombination rate  $\lambda_j$  is based on genomic distance in cM. The observable states are the alleles of the new haplotype and the emission states are nearly trivial (in that alleles are likely to be copied directly) though there is provision for both mutation and genotyping errors.

### Box 3.2.1c – The Li-Stephens model

A central idea will be that if we look at a single chromosomal haplotype of an individual, that should be an ‘imperfect mosaic’ of haplotypes of other individuals.

AGTGACTCTCAGACTGGGAACTCTTGAT0GATGGCATTAGGCGCGATTAGGCTCTCGATCTAGCTGCGCTATGAGCACGCCG ...  
 Target haplotype                      Missing value                      Mutation or Error

$h_1$  AGTGACTCTCAGACTAGGCACTGTTGATGGATGGTATTACGCCGATTAGGCTCTTAATCTAGTTGTGCTACAAACGCCCG ...  
 $h_2$  AATGGCTCTCAGACTGGGAACTCTTGATCGGTGGCTCTAGACGCGATTGGGTTCTAAATCTAGGTGTGCTATGGGCGCACCG ...  
 $h_3$  AATGACATTCAGACTGAAGACTCCTGATCGATGGCATCTAGGCGCGATTAGACTCTCGGTCTGGCCACGCGATGAGCGAGCCG ...  
 $h_4$  AATGGCTCCCAGACAGGAGACACCTGATCTATCGCACTTGGGCGCGAATAGCCTCTCAATCTAGCTACGTATGAGCGCGCCG ...  
 $h_5$  AGTAACCCTCAGACTGAGAGCTCTAAATCGATGGCATTAGGCGCGACTAAGCTCTCGATCTAGCTGCGCTATGAGCGAGCCG ...  
 Reference haplotypes

We can find a suitable reference haplotype from which to copy at each position in order to construct our **Target Haplotype**.

111111111111222222222222222233333333333355555511111111111155555555555554444444444444444 ...  
 Copying States

This leads to a **Hidden Markov Model (HMM)**, with the copying indices of reference haplotypes as hidden states, and the observed alleles of the target haplotype as the observed states.

If we denote the hidden copying states as  $u_1, u_2, u_3, \dots \in \{h_1, h_2, h_3, \dots, h_R\}$  and the observed alleles as  $a_1, a_2, a_3, \dots$  then an HMM is built as follows:

**Initial Probabilities:**  $P(u_1 = h_r) = \frac{1}{R}, \forall r$ .

**Transition Probabilities:**

$$P(u_j = h_r | u_{j-1} = h_s) = \begin{cases} \exp(-\lambda_j) + (1 - \exp(-\lambda_j))(1/R), & \text{if } r = s \\ (1 - \exp(-\lambda_j))(1/R), & \text{otherwise.} \end{cases}$$

$\lambda_j$  is the expected number of recombinations between the two positions  $j$  and  $j - 1$ :  
 $\lambda_j = 4N_e d_{Hj} / R$  where  $N_e$  is the theoretical population size and  $d_{Hj}$  is the genetic distance between genetic markers  $j$  and  $j - 1$ .

**Emission Probabilities:**

$$P(a_j = x | u_j = h_r) = \begin{cases} \frac{R}{R + \theta} + \left(\frac{1}{2}\right) \frac{\theta}{R + \theta}, & x = h_{rj} \\ \left(\frac{1}{2}\right) \frac{\theta}{R + \theta}, & x \neq h_{rj} \end{cases}$$

Where  $\theta$  can incorporate the probabilities of mutation or genotyping errors at marker  $j$ .  
 $h_{rj}$  is the allele at position  $j$  in reference haplotype  $h_r$ .

Box 3.2.1c - The Li-Stephens HMM model for haplotype mosaicism in a population.



In a sense, this model appears to invoke ideas of IBD-sharing, but in fact it is only modelling IBS-sharing. This model does not make any assumption of recent shared ancestry. The idea is that haplotypes are likely to be locally identical across individuals in a population due to the enduring LD structure in a population. High correlations between nearby variants will imply that pairs of individuals from the same population are likely to have identical haplotypes over short distances.

Thus, the model argues that after observing  $N$  haplotypes from a population, the  $N + 1^{th}$  observed haplotype will be an imperfect mosaic haplotype built of short segments that have already been observed. This may implicitly imply some ancient common ancestry between two individuals sharing a short haplotype segment. However, this is not the same as IBD, which in a sense implies a recent common ancestor. What recent means in the definition of IBD is ambiguous; a common definition would be that this is an individual arising after the population (from which our sample is drawn) was founded. In any case, the important distinction is that the Li-Stephens model relies on IBS-sharing without bothering to consider whether shared haplotypes are long enough or specific enough as to be considered as IBD.

SHAPEIT2 employs a version of the Li-Stephens model to estimate phase. To estimate phase for a given individual, SHAPEIT2 must draw two paths through the Li-Stephens HMM, giving two haplotypes, and these must be compatible with the observed genotypes. Specifically, when individual  $i$  is being phased, SHAPEIT2 takes the current estimated phased haplotypes of all other individuals in the sample as the reference haplotypes; and searches for two new haplotypes compatible with individual  $i$ 's genotype. The algorithm loops through individuals repeatedly, with each new estimation of phase being re-supplied to the model in order to facilitate the phasing of subsequent individuals. SHAPEIT2 can also leverage information from an external reference panel of haplotypes, by adding these to the pool of reference haplotypes from which new haplotypes can be copied from. A further option for SHAPEIT2 is the 'duohmm' option<sup>211</sup> which is designed to search the pedigree information in the sample for instances of nuclear families, and to refine final phase

estimates for these individuals. Full details of the SHAPEIT2 algorithms, including different versions and options are found in Delaneau, et al.<sup>220</sup>, Delaneau, et al.<sup>221</sup>, and Delaneau, et al.<sup>222</sup>

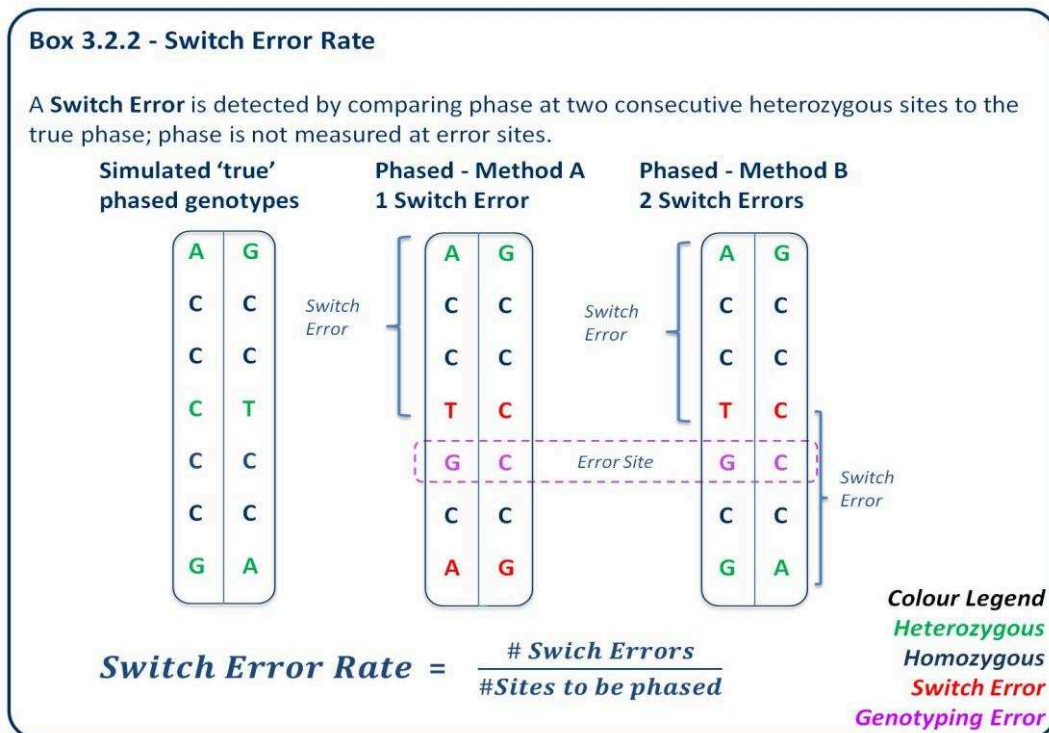
For an isolated population, we might initially expect to see an advantage for IBD-based phasing methods over LD-based phasing methods. LD information informs us only of the probability of recombination events between variants, allowing us to infer the probability of combinations of short haplotype matches between individuals and hence one can infer probabilities of haplotype phase. IBD-based methods concentrate on finding longer haplotype matches. If IBD-sharing can be confidently established, phase can be directly inferred, and when there remains some uncertainty about IBD-sharing, haplotype phase probabilities can then be derived. This is the approach of SLRP, which uses an HMM to model IBD-sharing and to tie together the more easily identifiable long haplotype matches, before calling phase using the concepts of LRP. SLRP incorporates the pairwise HMM of Genovese, et al.<sup>209</sup>, which models the IBD-sharing states along the four chromosomes of a pair on individuals<sup>223</sup>. The prevalence of IBD in an isolate would intuitively suggest using a method such as SLRP.

The most recent assessment of phasing algorithms for isolate type data was in O'Connell, et al.<sup>211</sup> Here SHAPEIT2 was the strongest performer and was shown to outperform SLRP. Prior to our study, the EAGLE software had not been tested for isolated populations; it was our hypothesis that it might be the most appropriate method as it combines an initial LRP algorithm, with a second LD-based method. Thus, EAGLE could theoretically take advantage of long stretches of IBD in an isolate, without suffering (as SLRP has been shown to do) from being unable to phase areas outside of IBD sharing. EAGLE1 combines Long Range Phasing with a secondary pseudo-HMM. EAGLE2 however will act differently depending on whether or not external reference haplotypes are provided. If no external reference panel is provided to EAGLE2, it performs LRP followed by an application of the Positional Burrows-Wheeler Transform (PBWT)<sup>205</sup> for haplotype phasing, an indexing algorithm that facilitates phasing by finding nearest local neighbours and again invokes ideas of IBS-sharing and

imperfect mosaic haplotypes. If however, an external panel is supplied, the LRP algorithm is not used, and only the second PBWT algorithm is used. It is important to note that EAGLE was designed to phase very large samples of individuals without incurring large computational burden. Indeed, the LRP aspect of EAGLE was motivated primarily as a time saving device.

### 3.2.2 Accuracy of Phasing Software - Results from our Publication

To evaluate phasing software, we measure the Switch Error Rate (SER) (Box 3.2.2). Note that as our simulation contained missing data and error sites, we only measured switch error rates at truly heterozygous sites. In Figure 3.2.2a, we have calculated the SERs of phasing on chromosome 10 using different software and indeed different strategies. In Figure 3.2.2b, the results restricting to sites successfully phased by SLRP is shown.



Box 3.2.2 – Switch Error Rate, a metric for phasing accuracy used throughout our study. Switch errors were measured between pairs of heterozygous sites, excluding positions that were known simulated genotyping errors.

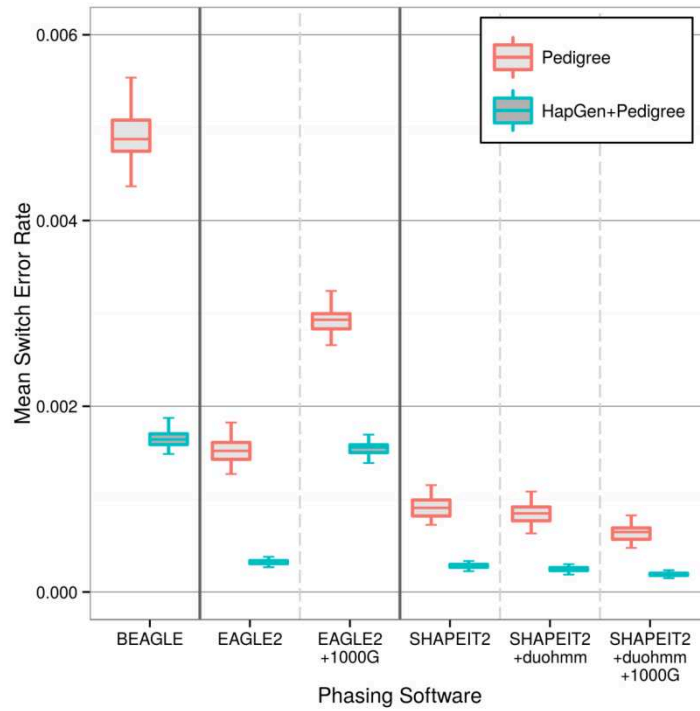


Figure 3.2.2a - Mean switch error rates (SERs) across the multiple iterations of our simulation. Results are split between the two simulation strategies (described in Section 2.2). Note that a lower SER indicates a better accuracy of phasing.

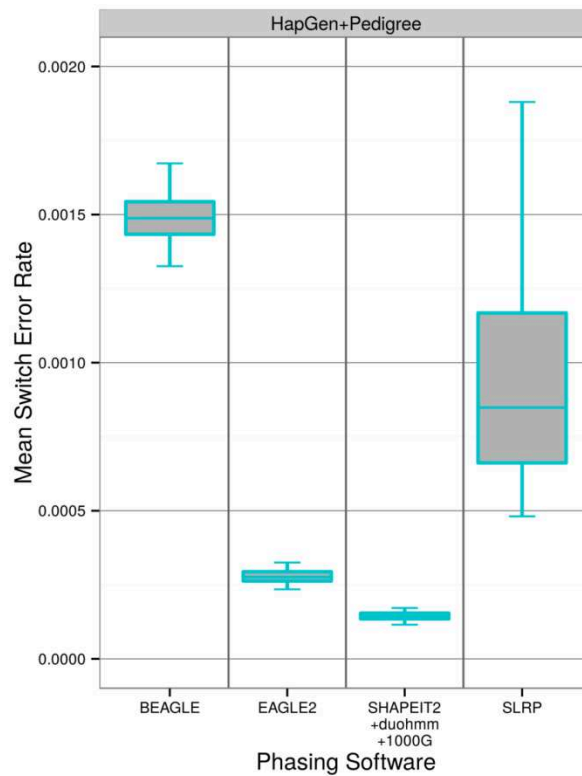
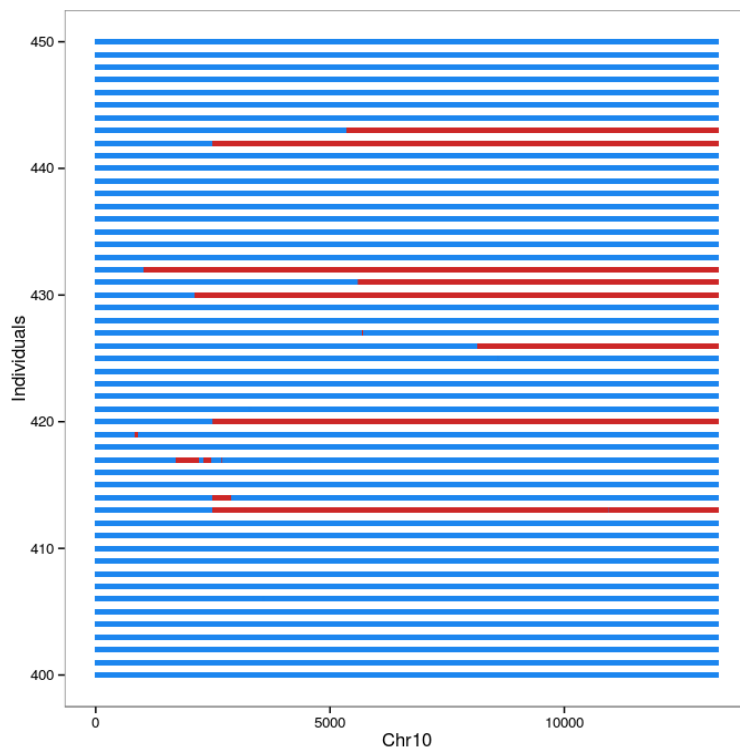


Figure 3.2.2b - Mean switch error rates (SERs) on the 100 iterations of the HapGen+Pedigree simulation strategy including results from SLRP. In this figure, the calculation of SER is restricted to the set of variants that could be successfully phased by SLRP.

The best strategy was found to be “SHAPEIT2+1000G+duoHMM”, full details of all the phasing strategies are given in Annex A. This strategy refers to running SHAPEIT2 with the 1000G as a reference panel and the “duohmm” option in operation.

In Figures 3.2.2c-e, we give visual interpretations of the phasing performance of three algorithms by plotting a small sub-section of the phased output of a single simulation iteration. In these pictures, changes in colour represent switch errors. For 50 individuals, their phased chromosome 10 data are stacked in these graphs. A complete blue line shows a perfectly phased individual; a line that changes once from blue to red shows an individual with one switch error; a line that changes often between blue and red shows a poorly phased individual with many switch errors.



*Figure 3.2.2c - Zoom in on phasing performance of SHAPEIT2+duohmm+1000G on the HapGen+Pedigree simulation.*

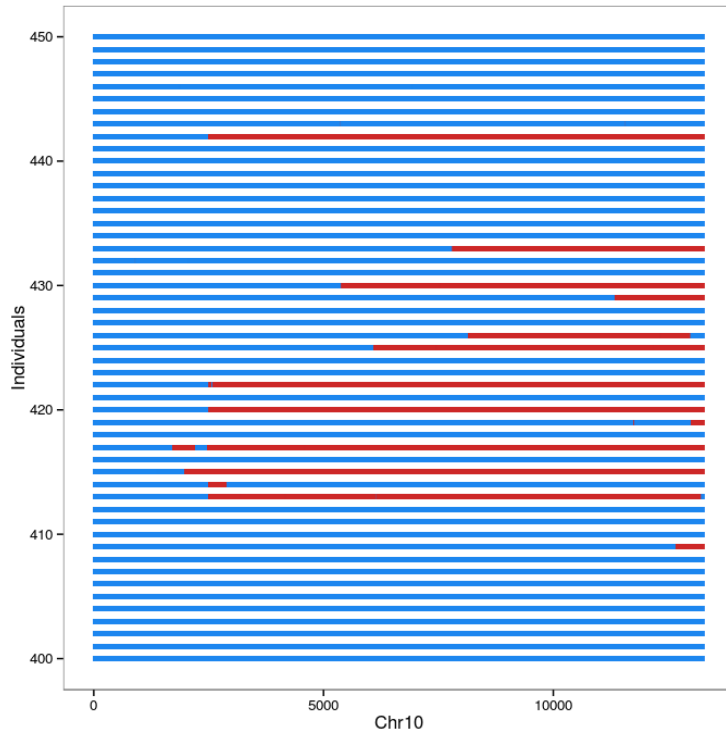


Figure 3.2.2d - Zoom in on phasing performance of EAGLE2 on the HapGen+Pedigree simulation.

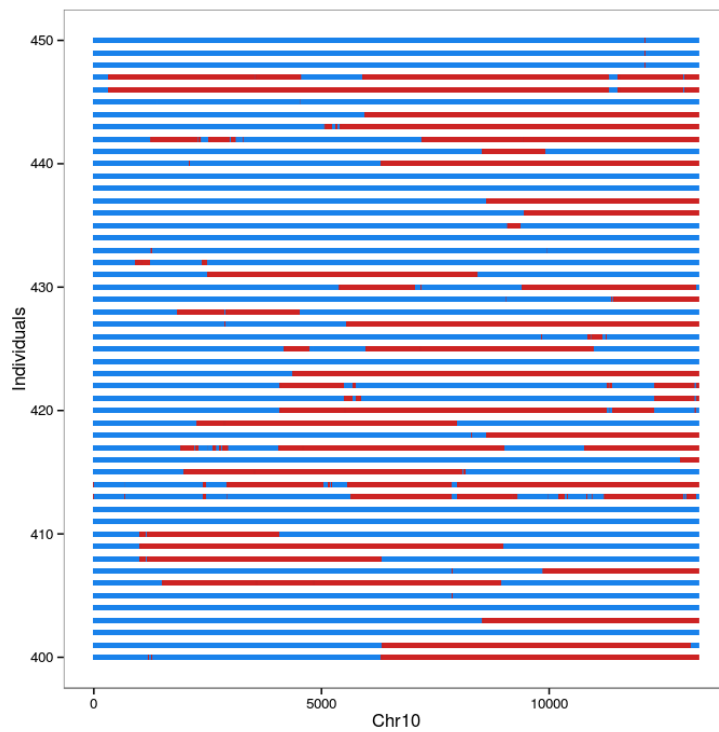


Figure 3.2.2e - Zoom in on phasing performance of BEAGLE on the HapGen+Pedigree simulation.

In these three diagrams (Figures 3.2.2c-e), we can see the strong and often similar performances of SHAPEIT2 and EAGLE2; with the less successful software BEAGLE as a counterpoint.

SHAPEIT2 facilitates investigation into the certainty of the phase estimates that it provides. In Figure 3.3.3f, we have produced a detailed look at the phasing performance of SHAPEIT2. At this resolution, we again have each individual represented horizontally and now individual SNPs can be plotted. For the first panel (right), the colour scheme now represents homozygous (and so not important for phasing) sites in grey, correctly phased SNPs are plotted in blue, and switch errors are plotted in red. This right panel represents the most likely paths ascertained by the SHAPEIT2 HMM, and indeed only a few switch errors are found when comparing the phase estimated to the true simulated phase.

We can then demand that SHAPEIT2 to make random draws from its HMM giving different possible realisations of phase. In the left box, 100 random draws were taken and now the number of times the phase changed with respect to the right box is given. This represents the certainty or uncertainty of SHAPEIT2's phase. We can extrapolate that when SHAPEIT2's phase estimate at a site changes five times out of 100 random draws, then SHAPEIT2 is around 95% certain of the phase at the site.



*Figure 3.2.2 - Phasing uncertainty with SHAPEIT2. Left panel: Rows represent individual genomes; each small vertical bar is a variant. Blue bars are correctly phased heterozygous sites; red bars are incorrectly phased with respect to the preceding (reading left to right) heterozygous site. Grey bars are homozygous. Right panel: The number of times SHAPEIT2 returned the same phase when asked to make 100 random draws from its final HMMs.*

This functionality of SHAPEIT2 is rather onerous to perform genome wide but could be an important tool for establishing the quality of phase in a region of interest.

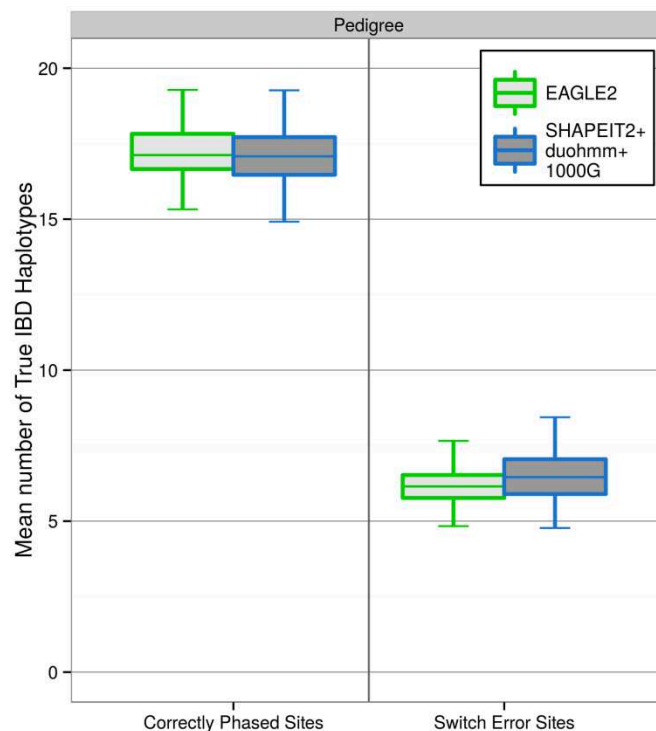
Going back to the overall results of Figure 3.2.3a, an interesting finding was that EAGLE2 outperformed EAGLE2+1000G, this suggested that LRP was indeed an accurate method for our simulated data as better results were obtained when EAGLE2 was obliged to perform it. In regards to SLRP, the results also reflect positively on LRP, we were able to phase very high percentages of chromosome 10 and with high accuracy. In Kong, et al.<sup>76</sup>, LRP was demonstrated to be more effective with more individuals, specifically when a large fraction of the total population (in their case this was the population of Iceland whose sample size was given as 316,000 individuals) was genotyped. They showed that if only 1% of the population was genotyped, useful results could still be obtained, and that going up to 10% would yield very accurate results. In Cilento, it was estimated at the genesis of the study that the set of genotyped individuals covered 80% of the total eligible populations of Campora, 86% of Gioi, and 64% of Cardile. At time of writing, the current number of inhabitants of the three villages are in the region of 400 for Campora, 800 for Gioi and 500 for Cardile (personal communication with Teresa Nutile). Effective population size estimates of the three villages are in the region of a few thousand individuals<sup>174</sup>. It is difficult to precisely estimate the fraction of the population that has been genotyped for the Cilento isolates, yet we feel confident that we surpass the 10% mark that should make Cilento a perfect candidate for LRP. Indeed, when tested on the true data of Campora, SLRP was able to phase ~99% of heterozygous sites (Supplementary Figure 8a of Annex A), both this and the SERs for SLRP were more positive results for SLRP than any we found in the literature.

Perhaps the most noteworthy result was the observation that all algorithms gave highly accurately phase data compared to previous published SERs<sup>201,203,204,224</sup>. In particular, on the HapGen+Pedigree simulation, where IBD sharing in the simulated individuals was boosted even higher by the tighter



bottleneck, the SERs observed were very low as all methods appear to benefit from the increased IBD-sharing and lowered genetic variation.

O’Connell et al<sup>211</sup> also observed SHAPEIT2 outperforming IBD-based phasing algorithms, and in their discussion they suggest that the Li-Stephens model was perfectly capable to utilising long IBD. To test this idea, we explored the location of switch errors in conjunction with the location of shared IBD sections which we were able to track through the simulation of the data. This result was described in the Supplementary Materials Annex A and presented in Supplementary Figure 11. This is also given here in figure 3.2.2g.



*Figure 3.2.2g – The number of available surrogate parents (true IBD haplotypes) in the sample for random picks of Switch Error Sites and correctly phased sites. The same numbers of sites were chosen for the two sets of boxplots in the plot, from similar individuals and with the sites specified to have similar MAFs. This analysis was completed on the Pedigree simulation, as it was in this scenario that we could record IBD-sharing perfectly (See Annex A).*

By comparing the locations of switch errors to the locations of correctly phased sites, it was clear that both EAGLE2’s and SHAPEIT2’s performed in similar way depending on the number of existing long IBD matches (surrogate parents) in the sample at that position. This is an intuitive result for a

software using LRP where the number of surrogate parents that can be found will naturally affect the ability to phase. For an LD-based method, it is less obvious why one would observe a pattern involving the number of IBD-sharing partners. The fact that EAGLE2 and SHAPEIT2 seemed to behave so similarly in this respect strengthens the idea of O'Connell, et al.<sup>211</sup> in that SHAPEIT2's approach based on very local LD patterns can (through communications in the graphical HMM) benefit from long shared IBD segments. This idea is also discussed in Howie, et al.<sup>225</sup> This would be an interesting area for continued exploration. We know that HMM based methods have been widely used to locate segments of HBD and IBD<sup>79,226-230</sup>. Some of these methods indeed could even facilitate the simultaneous detection of IBD sharing between multiple pairs<sup>231,232</sup>. Most of these algorithms come up against limitations of computational complexity, though recently the ultra-fast method of PBWT has been reengineered to estimate IBD in large sample<sup>233</sup> using similar ideas as to when PBWT had been adapted for phasing and imputation<sup>205</sup>. If we know that the phasing algorithm SHAPEIT2 gives high on perfect phase in an isolate, it would suggest that the final HMM models that this software produces (which described haplotype sharing across a whole sample of individuals) could be ideal for identifying almost all IBD matches in a sample.

### 3.2.3 Genotyping Errors in Phasing

In this study we also simulated genotyping errors in the Array data and observed how the presence of such errors affected the accuracy of phasing. These results were given in Supplementary Figures 8a, 8b, and 10 in Annex A, with errors particularly affecting SLRP and ALPHAPHASE. Here we give one further analysis that we carried out but that was not included in the publication. Here we were able to show how errors simulated at a particular site would engender phasing errors at the same site in other individuals. This analysis is presented in Figure 3.2.3a. Genotyping errors will clearly disrupt IBD sharing patterns and also LD estimations, yet it is interesting to see that EAGLE and SHAPEIT2 had very similar increases in SER at the locations where errors were present.

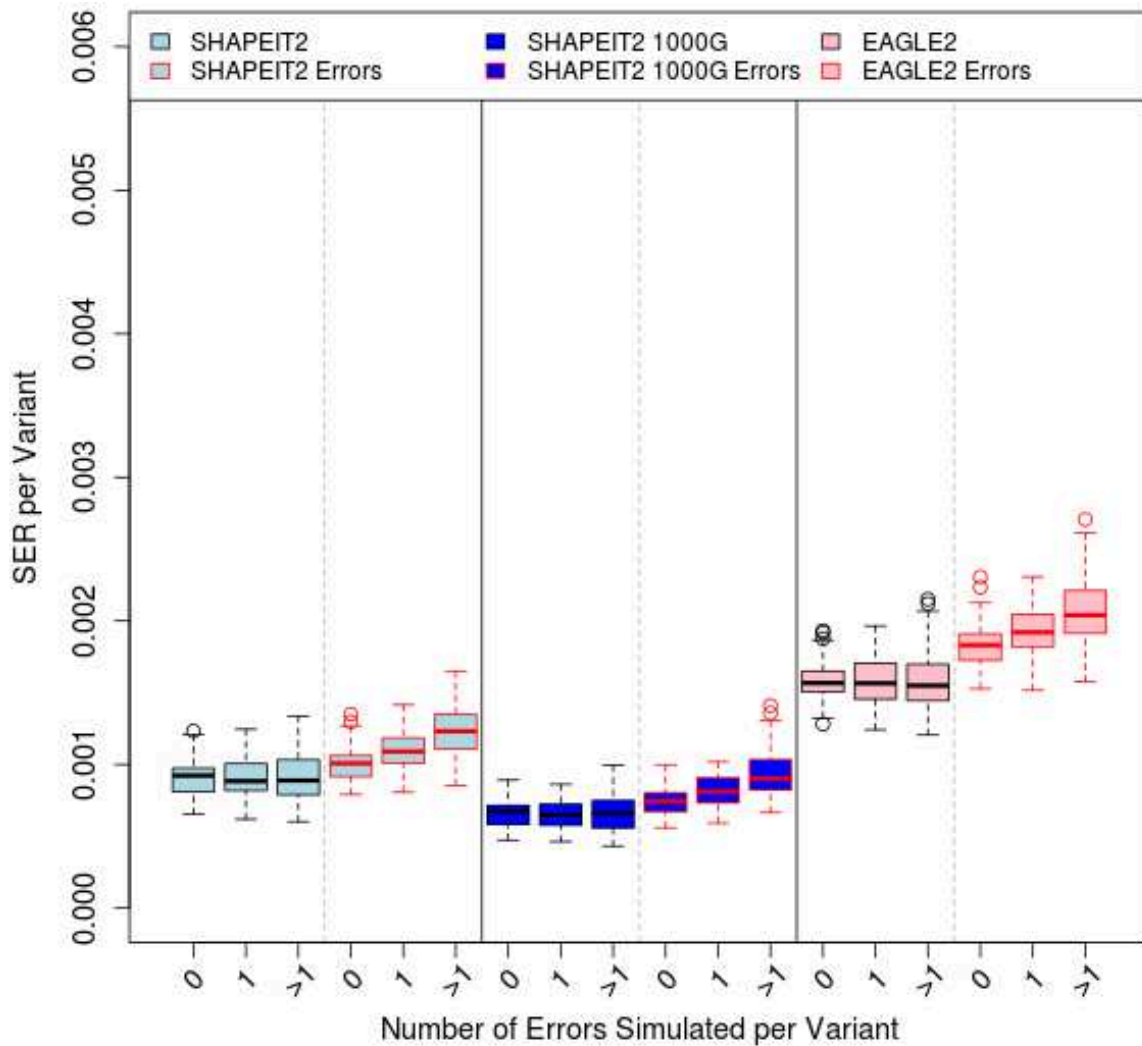


Figure 3.2.3a – We selected sites with either 0, 1, or >1 simulated genotyping errors. Boxplots of SER are labelled by the phasing pipeline used and whether or not the data had simulated genotyping errors.

We did not measure for switch errors at the exact individual genotypes with simulated genotyping errors, but we do measure the SER at the position across all other individuals in the sample. The pink-bordered boxplots show the increase in SERs at the positions with more and more genotyping errors. The black bordered boxplots give a counterpoint, showing the trend (of course) disappearing when no genotyping errors were simulated in the first place.

This analysis was on the Pedigree simulation strategy only.

## 3.3 Imputation

### 3.3.1 Accuracy of Different Imputation Software - Results from our Publication

In each iteration of the simulation, we retained the phased haplotypes from the SHAPEIT2+1000G+duoHMM strategy and proceeded to test a range of imputation software and imputation reference panels. We tested IMPUTE2<sup>225</sup>, IMPUTE4<sup>150</sup>, MINIMAC3<sup>197</sup>, BEAGLE (version 4.1)<sup>206</sup>, and PBWT<sup>205</sup>. We tested public imputation reference panels: the 1000G and the HRC. We also tested the use of a WGS SSP.

All imputation software tested used LD-based methods; again with central concepts of haplotype mosaicism described by the Li-Stephens model (Box 3.3.1); interpretations of which are used in IMPUTE2, MINIMAC3, BEAGLE, and PBWT. MINIMAC3 and IMPUTE2 are very similar; differing most in their methods for calculating model parameters relating to transition probabilities and in the initial selections of potential copying states<sup>234,235</sup>. Previous versions of BEAGLE used a similar haplotype clustering model as used for their phasing algorithm but the latest version of BEAGLE that we tested relied again on the Li-Stephens model. The PBWT algorithm allows for a highly efficient graphical description of a large haplotype reference panel, from which the estimated haplotypic frequencies allow for new haplotypes to be matched to the reference under the Li-Stephens model<sup>236</sup>.

Figure 3.3.1a displays the mean imputation accuracy (split my MAF) for different software when using the 1000G as a reference panel; MINIMAC3 was the best performer. Imputation Accuracy was measured as the correlation between true genotype and imputed dosage genotypes (Box 3.3.1) on the first 20Mb of chromosome 10.

### Box 3.3.1 – The Li-Stephens model for imputation

????A????????????G????T????????G????G????G????T????C????G????C? ...

Target haplotype with many missing entries. For example, from phased array data outputted by SHAPEIT2.

- $h_1$  AGTGACTCTCAGACTAGGCAGCTGTTGATGGATGGTATTACGCCGATTAGGCTCTTAATCTAGTTGTGCTACAAACGCCCG ...
- $h_2$  AATGGCTCTCAGACTGGGAAGCTTGTATCGGTGGCTCCTAGACGCGATTGGGTTCTAAATCTAGGTGTGCTATGGGCGCACCG ...
- $h_3$  AATGACATTGAGACTGAAGACTCCTGATCGATGGCATCTAGGCGCGATTAGACTCTCGGTCTGGCCACGCGATGAGCGAGCCG ...
- $h_4$  AATGGCTCCCAGACAGGAGACACCTGATCTATCGCACTGGGCGCGAATAGCCTCTCAATCTAGCTACGCTATGAGCGCGCCG ...
- $h_5$  AGTAACCCTCAGACTGAGAGCTCTAAATCGATGGCATTAGGCGCGACTAAGCTCTCGATCTAGCTGCGCTATCAGCGAGCCG ...

Reference haplotypes

111111111112222222222222222222222233333333333333335555555111111111111555555555555555544444444444444 ...

Copying States: again the same principal applies, if we can find the hidden copying states (denoted here by colour) then the missing entries in the target haplotype can be easily filled in.

If we again denote the hidden copying states as  $u_1, u_2, u_3, \dots \in \{h_1, h_2, h_3, \dots, h_R\}$  and the observed sequence of alleles as  $A_{1,\dots,R}: a_1, a_2, a_3, \dots, a_R$  then by applying the HMM mechanics (Forward-Backward algorithm) we calculate the posterior probabilities of each hidden state:

$$P(u_j = h_r | A_{1,\dots,R}) = \rho_{jr}$$

The HMM must now accept missing values, so that  $\rho_{jr}$  can be calculated at every position, even when  $a_j$  is missing.

Lets say, for some  $j = t$ ,  $a_t$  is missing. The possible values for  $a_t$  are 0 and 1 (corresponding to major and minor allele,  $A$  and  $a$  respectively).

We then have the following posterior probabilities of the possible values of  $a_t$ :

$$P(a_t = x) = \sum_r P(a_t = x | u_t = h_r) \rho_{tr}$$

Imputation programs typically can operate on haplotype data (already phased) as in this example, or there must be additional (and time-consuming) summations over different potential realisations of haplotype phase.

From these probabilities, expected values of the genotypes can be calculated and imputation programs will usually return posterior probabilities of each genotype and the expected value of the genotype which is referred to as the imputed dosage:

<p><b>True Genotypes</b></p> <table border="1" style="border-collapse: collapse; text-align: center;"> <tr><td>AA</td></tr> <tr><td>Aa</td></tr> <tr><td>aa</td></tr> <tr><td>AA</td></tr> <tr><td>...</td></tr> </table>	AA	Aa	aa	AA	...	<p><b>True Genotypes coded additively</b></p> <table border="1" style="border-collapse: collapse; text-align: center;"> <tr><td>0</td></tr> <tr><td>1</td></tr> <tr><td>2</td></tr> <tr><td>0</td></tr> <tr><td>...</td></tr> </table>	0	1	2	0	...	<p><b>Posterior probabilities</b> <small><math>P(AA), P(Aa), P(aa)</math></small></p> <table border="1" style="border-collapse: collapse; text-align: center;"> <tr><td>0.88, 0.10, 0.02</td></tr> <tr><td>0.40, 0.55, 0.05</td></tr> <tr><td>0.15, 0.32, 0.53</td></tr> <tr><td>0.98, 0.02, 0.00</td></tr> <tr><td>...</td></tr> </table>	0.88, 0.10, 0.02	0.40, 0.55, 0.05	0.15, 0.32, 0.53	0.98, 0.02, 0.00	...	<p><b>Dosage Genotypes</b> <small><math>P(Aa)+2P(aa)</math></small></p> <table border="1" style="border-collapse: collapse; text-align: center;"> <tr><td>0.14</td></tr> <tr><td>0.65</td></tr> <tr><td>1.68</td></tr> <tr><td>0.02</td></tr> <tr><td>...</td></tr> </table>	0.14	0.65	1.68	0.02	...
AA																							
Aa																							
aa																							
AA																							
...																							
0																							
1																							
2																							
0																							
...																							
0.88, 0.10, 0.02																							
0.40, 0.55, 0.05																							
0.15, 0.32, 0.53																							
0.98, 0.02, 0.00																							
...																							
0.14																							
0.65																							
1.68																							
0.02																							
...																							

Box 3.3.1 – Concepts of imputation of missing genotypes using the Li-Stephens model.

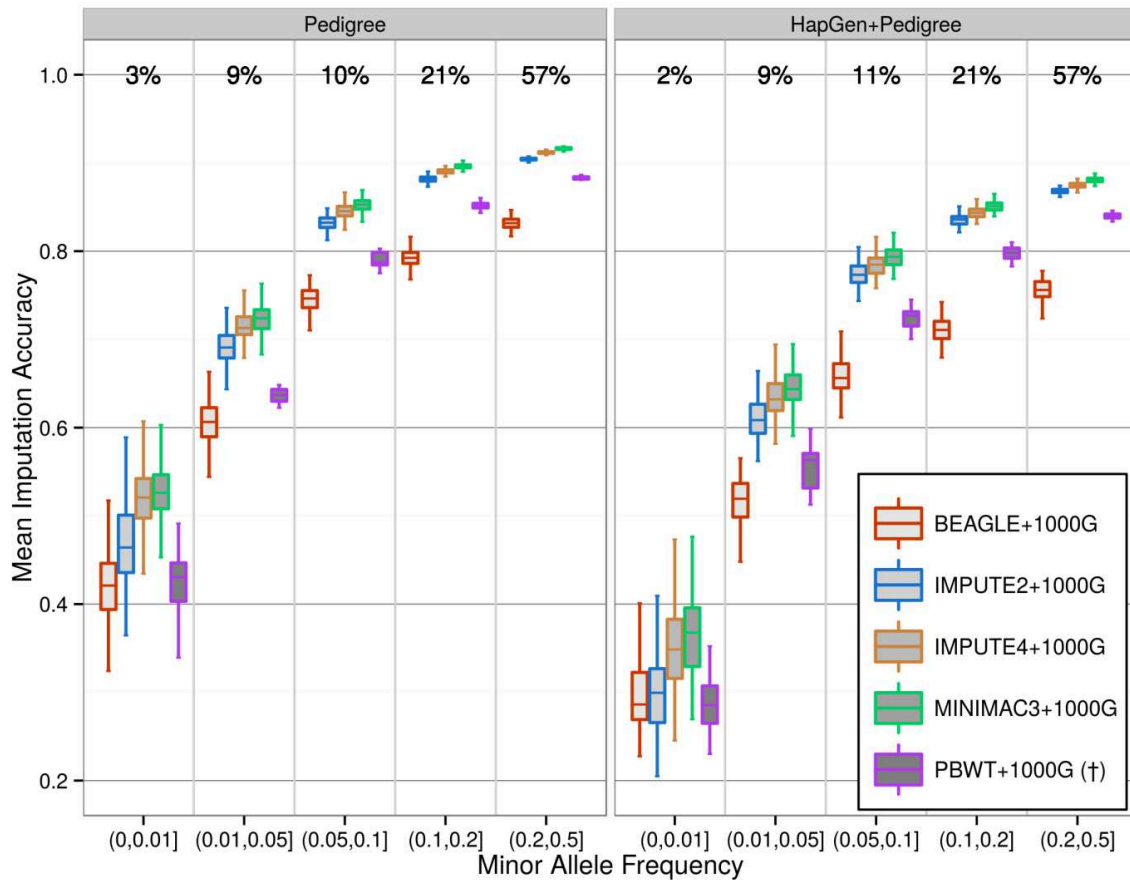


Figure 3.3.1a - Imputation accuracies from different imputation software, split by MAF and simulation strategy. Imputation was more accurate on the Pedigree simulation; this finding is discussed in section 3.3.3. (†) PBWT results are only from 25 simulation runs.

### 3.3.2 Reference Panels: Local and Global - Results from our Publication

In Cilento, 247 individuals have exome sequencing data, and these individuals were selected to give a good representation of the whole sample in order to serve as an SSP. Explicitly, kinship was calculated between all pairs of individuals. Then, in an iterative manner, individuals were selected to join the SSP by picking those with the highest mean kinship to all over individuals, without selecting an individual with a kinship coefficient above 0.1 with any other individual already selected for the SSP. This method was based on Urrichio et al.<sup>77</sup> In this simulation study, we only considered Campora, where 93 of the 477 individuals have Exome data. We considered hypothetical situations where these 93 individuals constituted either a study specific WES panel or a study specific WGS panel. In our publication, we only gave results for the WGS panel. This choice was based on the fact

that WGS panels are more often used and are more similar in nature to the public reference panels such as the 1000G. Furthermore, the results for the WES SSP were more complex than we had envisaged and so we chose not to take these results through to the publication as they were perhaps overly specific to the scenario in Campora (see section 3.3.5) and so of less interest to the wider scientific community. In agreement with many other studies that looked into the use of study specific imputation panels in human populations<sup>88,224,237-243</sup> and indeed in animal populations<sup>244</sup>, the 93 WGS individuals were very successful as an SSP. We felt that it was a significant contribution to the known literature to show that an imputation panel of only 93 individuals (even when admitting that they were closely related to the target individuals) could outperform a public panel of the size of the HRC.

In the Figure 3.3.2a and 3.3.2b, we compared different choices and combinations of reference panels for two software - IMPUTE2 and MINIMAC3. These two were chosen because IMPUTE2 had the unique ability to combine reference panels internally and because MINIMAC3 had given the most accurate results in our initial test. For both software, the inclusion of an SSP was highly effective. For MINIMAC3, we were able to test the hyper-diverse and hyper-large public reference panel, the HRC, which further improved the imputation accuracy. However, using the SSP alone was also very effective.

Our final recommendation for choice of imputation software depended on the number of reference panels available to the researcher. When imputing with a single panel, we observed an advantage for MINIMAC3 but the best imputation involved combining an external and a study specific panel. When an SSP is available, we felt that certain options of IMPUTE2 made it a more attractive choice; particular due to the possibility of merging two reference panels (this merging procedure is discussed in detail in section 3.3.5). Further details of our recommendations are given in the discussion of Annex A.

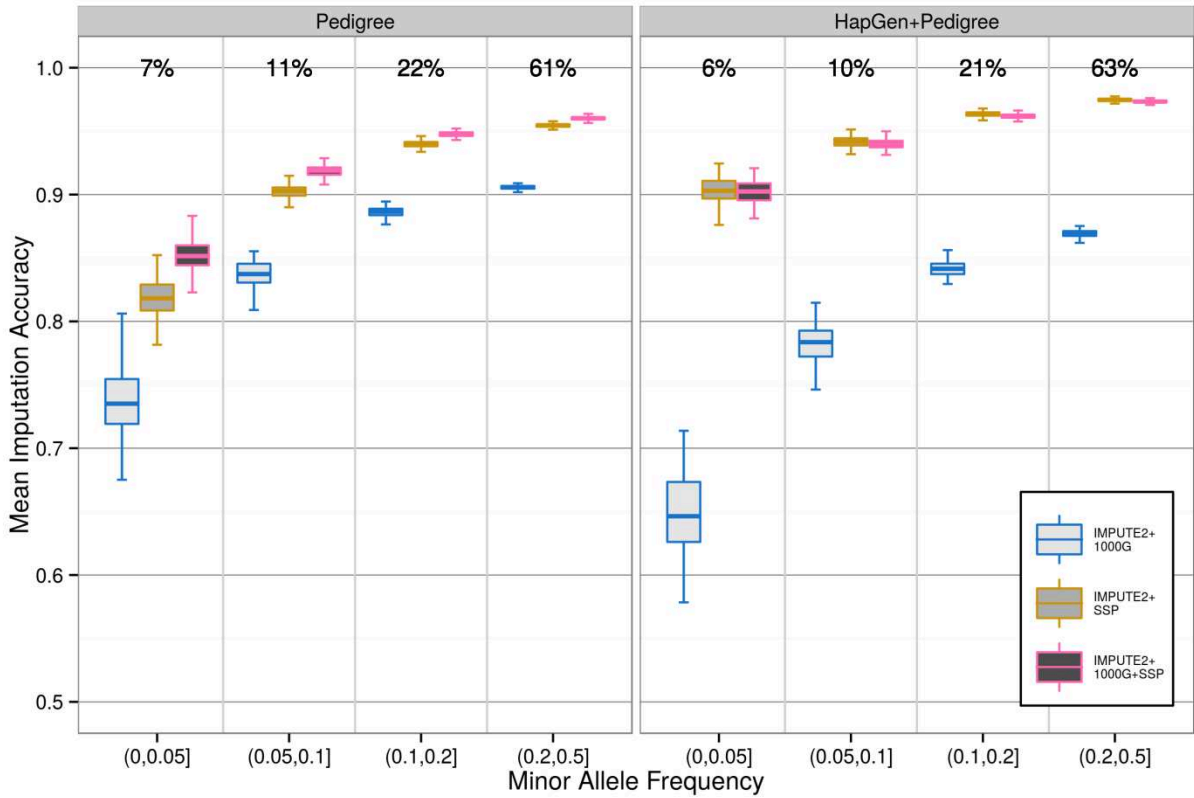


Figure 3.3.2a – Comparison of Imputation Accuracy for IMPUTE2 with different reference panel choices.

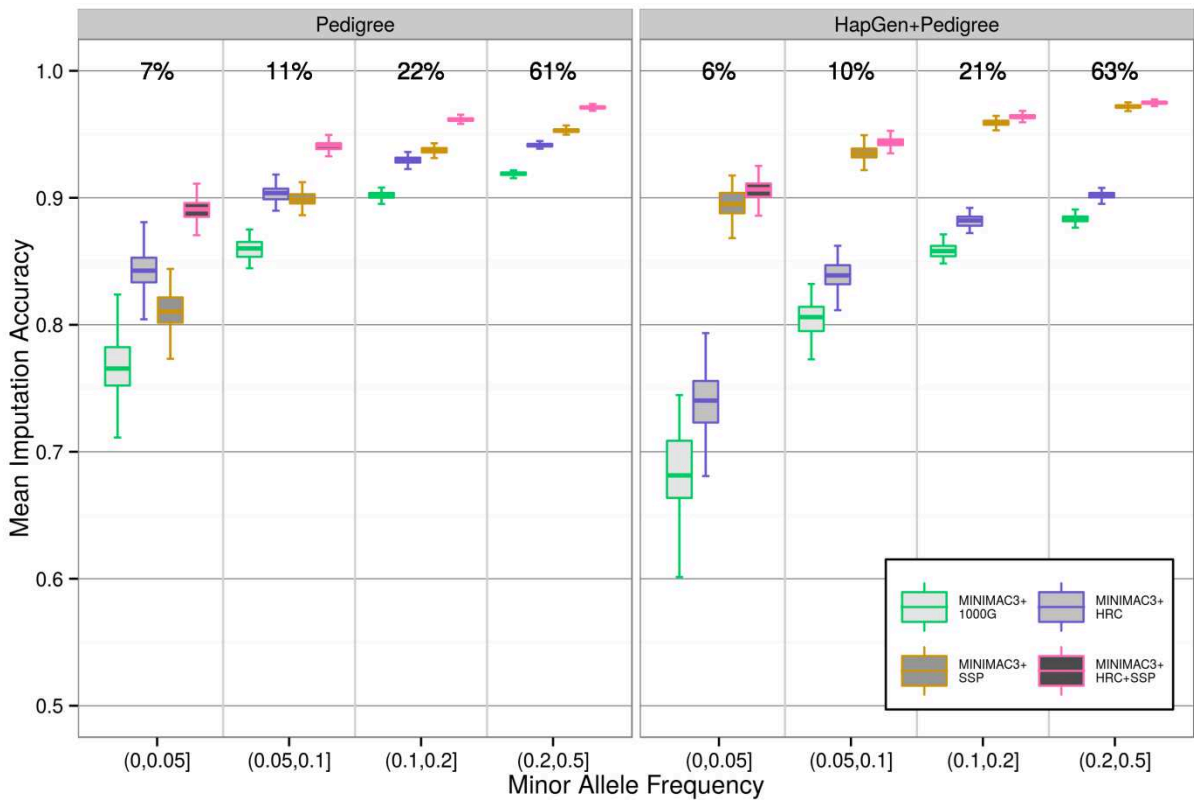


Figure 3.3.2b - Comparison of Imputation Accuracy for MINIMAC3 with different reference panel choices.



To run MINIMAC3 with both the HRC and the SSP we had to combine these two panels ourselves, thus having to restrict to sites present in both panels and so excluding the possibility of imputing variants that are specific to the sequencing of the study population.

### 3.3.3 HapGen+Pedigree vs. Pedigree

One of the most interesting results that we observed in this simulation study was the contrast between the two simulation strategies. For phasing, accuracy was much higher in the HapGen+Pedigree simulation as compared to the Pedigree simulation (Figure 3.2.2a). This trend was however reversed when calculating imputation accuracies (Figure 3.3.1a) when we were not using an SSP. To reconcile this phenomenon, we must remember that at the heart of each phasing algorithm, similarities between members of the sample are being leveraged. However, in the case of imputation using only the 1000G, the algorithm is now reliant only on similarities between each individual member of the sample and the set of external reference panel haplotypes. The HapGen+Pedigree strategy simulated a sharp bottleneck; thus creating a population with more specific characteristics and therefore one that was less well represented by the 1000G panel; hence imputation was poorer. We explored the similarity between the target population and the reference populations by looking at differences in MAFs between the two. Here we were able to show that the HapGen+Pedigree simulation created greater departures from public reference panels in terms of MAFs. Furthermore, when we used an SSP, the order of the results flipped - i.e. the imputation was more accurate on the HapGen+Pedigree simulation. In this scenario, the reference panel now contained haplotypes specific to the population and so the results took on a similar pattern to our results for phasing. In our publication we confirmed two logical suppositions regarding this result: (a) that individuals most closely related to the SSP were imputed more accurately; and (b) that variants which had a higher frequency in the target populations than in the external reference panel benefited most from the inclusion of an SSP. The results of these investigations are described in Supplementary Figures 15, 16a-d, and 17 of Annex A.

### 3.3.4 Info and RSQ

Our study also contained an investigation of two measures of imputation accuracy: the ‘info’ score and the ‘RSQ’ score provided by IMPUTE2 and MINIMAC3, respectively. Most imputation software will report a metric per-variant that describes the imputation quality. In the Supplementary Materials of J. Marchini’s and B. Howie’s review of genetic imputation<sup>245</sup>, a review of these quality scores and their differing calculations is given. Here it was shown that the different scores were all highly correlated with each other and that the two scores cited above were the best performers. The RSQ score describes directly the uncertainty of the imputed data; it is the quotient of the empirical variance of the imputed dosages over the ‘true’ variance of the hidden genotypes. This true variance is estimated as  $2\bar{q}(1 - \bar{q})$ , where  $\bar{q}$  is an estimator of the MAF from the dosage data. The ‘info’ score of IMPUTE2 is a ratio of observed and complete statistical information regarding the MAF; again one calculated from dosage data and one calculated using  $\bar{q}$ , the estimate of the MAF from the imputed dosages.

In Figure 3.3.4a, the utility of these metrics is evaluated by plotting them against the true imputation accuracy. The graph is generated from imputation during one iteration of the Pedigree simulation using the 1000G as a reference panel. Similar figures can be found in the supplementary information of Pistis, et al.<sup>238</sup> The imprecision of such metrics is well known, a detailed investigation of these metrics is found across publications by N.R. Roshyara and collaborators<sup>240,246,247</sup>. Commonly applied post-imputation quality thresholds have been to exclude variants with  $\text{info} < 0.4$  or  $\text{RSQ} < 0.3$ ; though different thresholds have been suggested when using an SSP.<sup>238</sup>

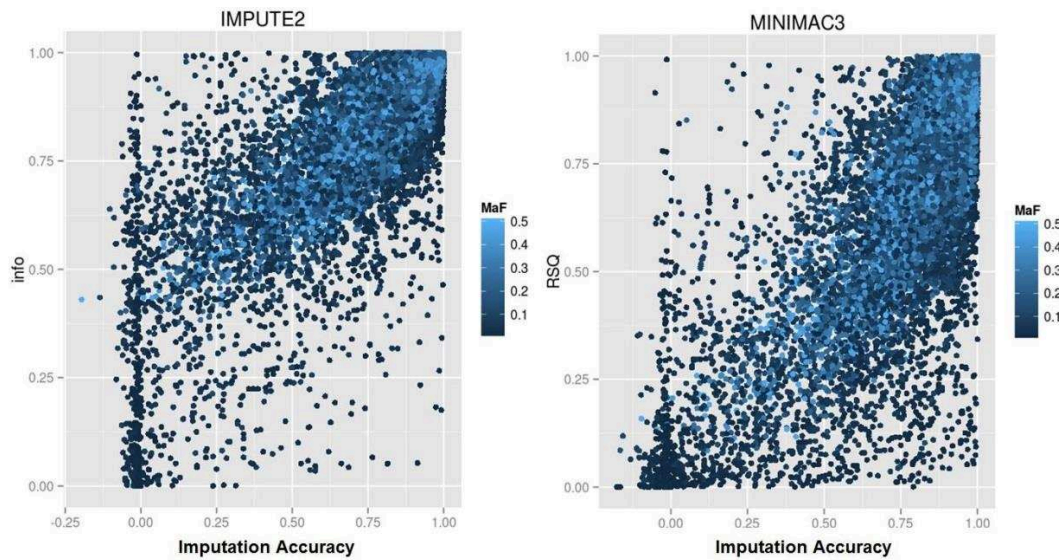


Figure 3.3.4a - Imputation Accuracy against 'info' and 'RSQ' scores based on the imputation assays IMPUTE2+1000G and MINIMAC3+1000G on one iteration of the Pedigree simulation. The blue shading represents the MAF of each variant. The darkest blue variants have the lowest MAF.

For rare variants (darkest blue in Figure 3.3.4a) the correlation between imputation metrics and true accuracy is notably quite poor. Hence, higher thresholds are often applied for the low end of the MAF spectrum<sup>248</sup>. In our study, we showed exactly how the choice of threshold would affect the overall accuracy of remaining variants as well as discussing that the choice of such a threshold is a balancing act between keeping as many imputed variants as possible, whilst removing those with poor imputation accuracy (Figure 3.3.4b, Supplementary Figure 18b in Annex A, and discussion in Supplementary materials of Annex A).

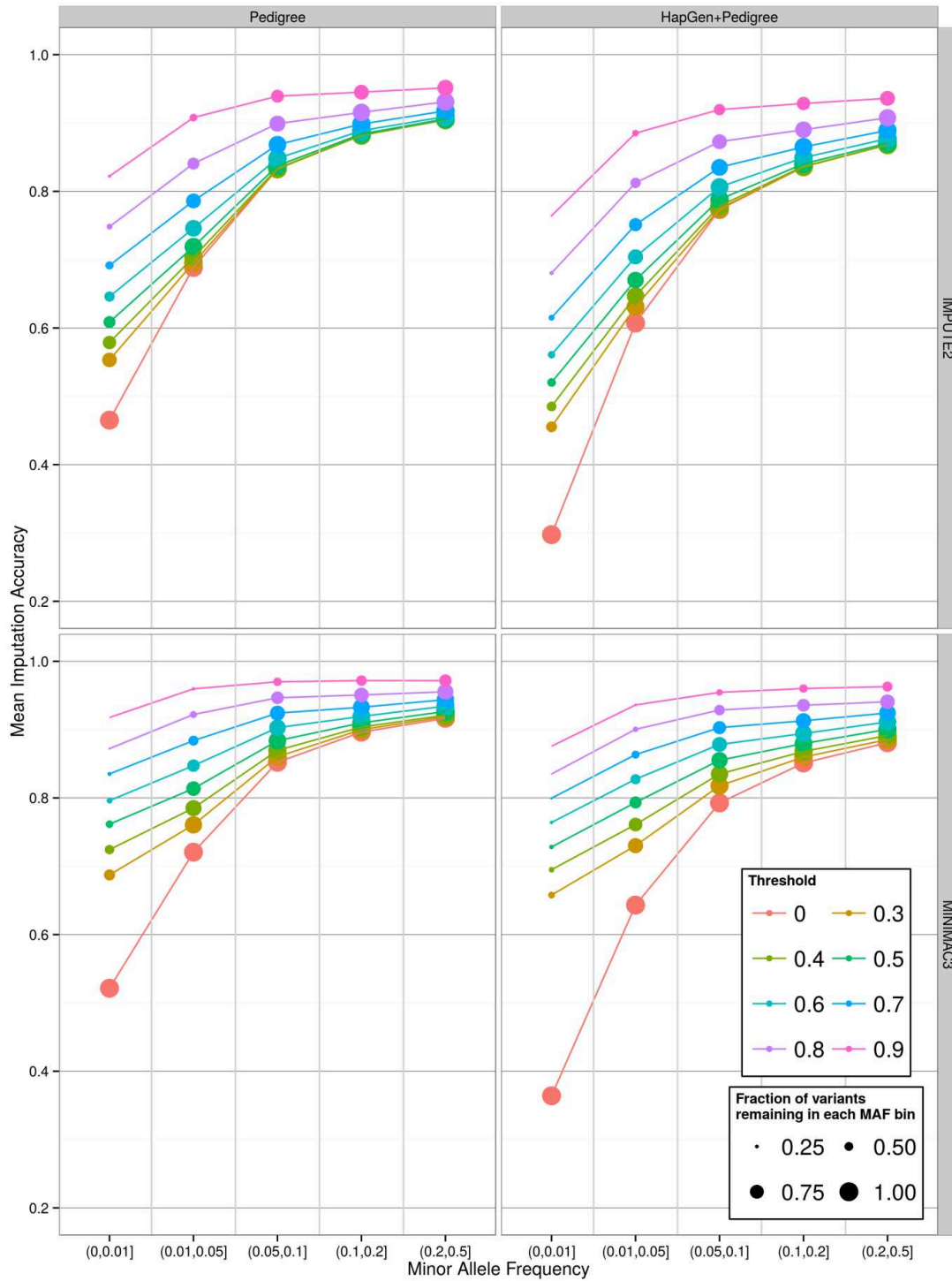


Figure 3.3.4b. Imputation accuracy across all MAFs following post imputation quality control based on either ‘info’ scores for IMPUTE2 or ‘RSQ’ scores for MINIMAC3. Imputation accuracy and imputation quality scores are derived from IMPUTE2+1000G or MINIMAC3+1000G imputation. The size of circles details the fraction of remaining variants after a given threshold.

We also briefly looked at exactly how the choice of threshold could be made. By changing the threshold gradually and measuring the numbers of well imputed variants (accuracy above 0.8)

against the number of poorly imputed variants (accuracy below 0.8), an ‘optimal’ threshold could be described as the point where one begins to remove more well imputed variants than poorly imputed ones. These analyses are given in Figure 3.3.4c and suggest that across the whole MAF range, threshold values of about 0.8 for info and 0.6 for RSQ would be appropriate. This is because in each test, as we increased the info threshold, we were removing more poorly imputed variants than well imputed variants up until tipping points where we began to remove more and more well imputed variants. Identifying the best thresholds will always be problematic as one first has to decide what constitutes ‘well’ imputed; and furthermore, the distribution of such quality scores will depend on many factors pertaining to the imputation strategy and in particular the similarity between target and reference haplotypes<sup>240</sup>.

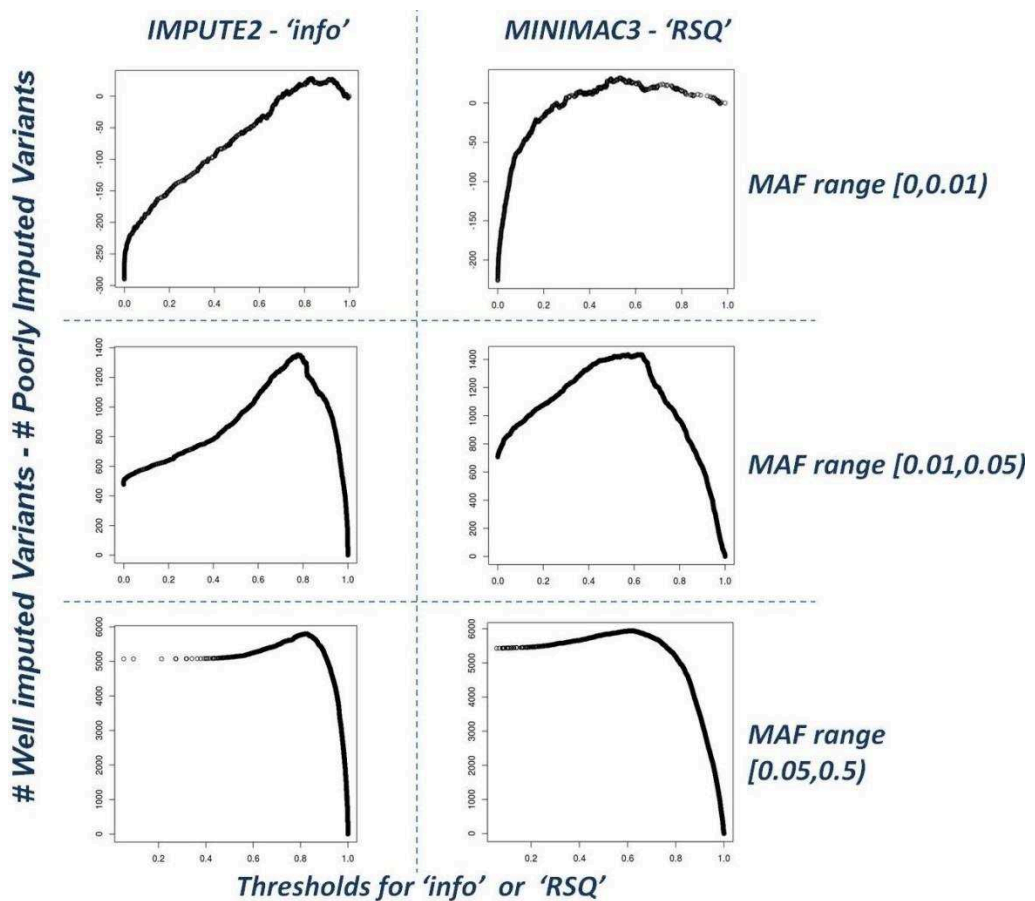


Figure 3.3.4c - Here the ‘info’ and ‘RSQ’ thresholds are varied (x-axes) and plotted against the difference of the numbers of remaining ‘well’ imputed variants and ‘poorly’ imputed variants. As the thresholds increase, we see this difference initially change as the majority of removed variants are ‘poorly’ imputed. In each case, a tipping point is reached and as the threshold increases further, more ‘well’ imputed variants are being removed than ‘poorly’ imputed variants.

### 3.3.5 Imputation with an Exome panel

In Cilento, our SSP is actually WES sequenced, not WGS sequenced which was the scenario we explored in our publication. Initially we planned to present results for both types of SSP; running the simulations with WES panels as well as WGS panels in each iteration. However, it became clear that the results for the WES panel were far harder to interpret than in the case of the WGS panel. In particular, our summarising statistics that measured imputation quality consistently showed that the strategy IMPUTE2+1000G+WES was outperformed by IMPUTE2+1000G, contrary to our expectations and to previous publications<sup>241</sup>.

By investigating different tweaks to the imputation pipeline involving a WES SSP and expanding our analysis to include all three villages; we observed three things in particular: (a) using an exome panel only improved imputation in exonic regions; (b) it would be best to impute individuals from Cilento separately based on their genotyping arrays; and (c) that the methods involving imputation with multiple reference panels had substantial impact on the final accuracy. Two choices that played a big role were the following: whether available Array positions were added to the WES panel and whether the merge-ref-panel option was engaged in IMPUTE2.

The 247 individuals with WES data in Cilento also have Array data (both in the simulation and in reality) and so when forming an SSP of haplotypes, it was possible to add these Array positions to the WES data. This was suggested in Joshi, et al.<sup>241</sup> and is an immediately logical decision. The extra Array position will aid the phasing of the SSP, particularly as they add data in between exons. The extra Array positions in the SSP will also facilitate imputation programs to match the target individuals with the SSP haplotypes across all Array positions.

The other important choice requires more explanation regarding the internal workings of IMPUTE2 pertaining to the merge-ref-panel option. In Figure 3.3.5a we give the relevant schema describing the merge-ref-panel option taken from the IMPUTE2 website.

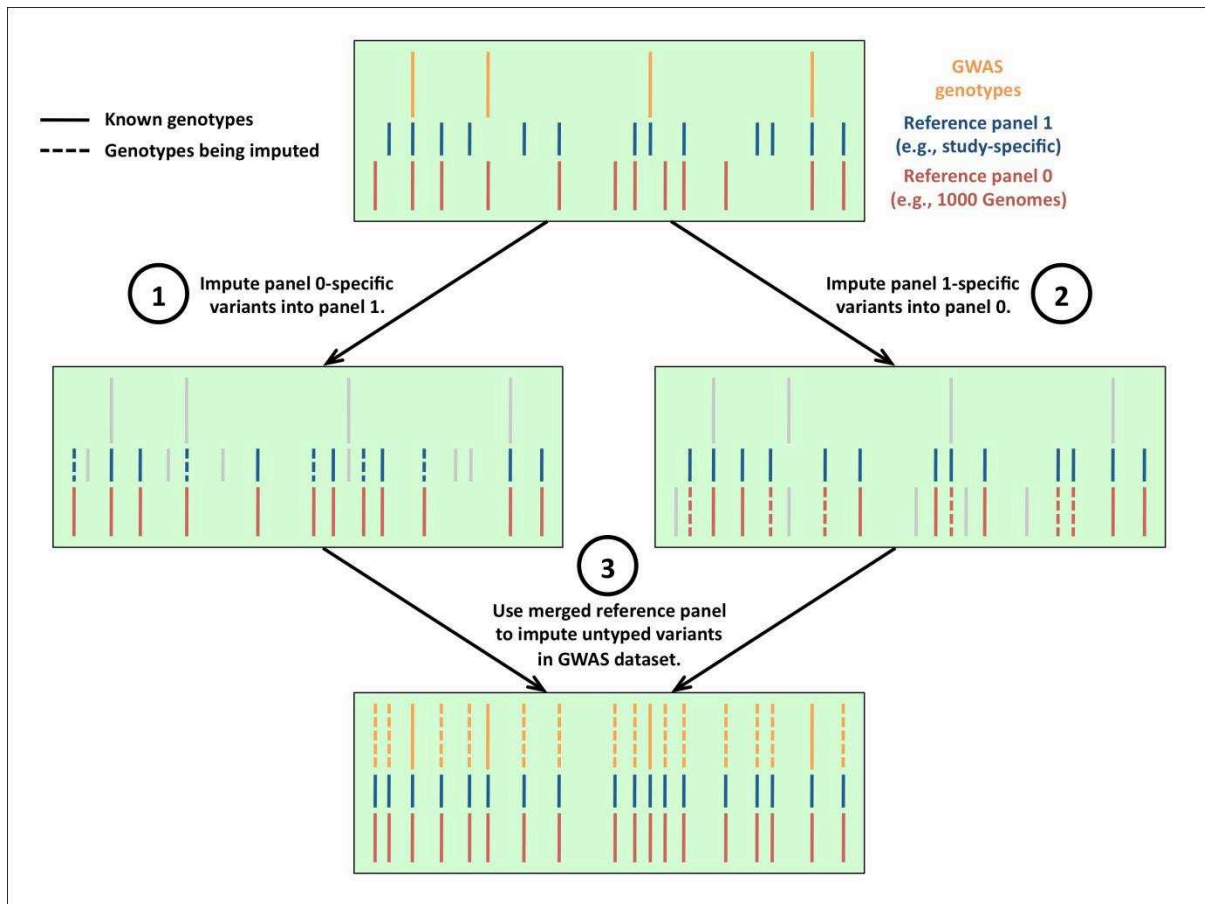


Figure 3.3.5a – Merging reference panels in IMPUTE2. First, sites that are missing in either one of the two reference panels are cross-imputed to form a single combined reference panel.

Source: [https://mathgen.stats.ox.ac.uk/impute/impute\\_v2.html#merging\\_panels](https://mathgen.stats.ox.ac.uk/impute/impute_v2.html#merging_panels)

In order to merge the two panels, any sites that are not present in both panels must be filled in via imputation in order so that both panels will contain identical lists of variants. This cross-imputation is performed in much the same way as the general algorithm for imputation in IMPUTE2 and crucially, ‘best guess’ or ‘hard called’ genotypes are then taken for these cross-imputed sites. In short, the genotype with the highest posterior probability is imputed as a certain genotype during cross-imputation.

The fact that IMPUTE2 does not later account for the uncertainty in the cross imputation appears to be the main reason for the downstream loss of accuracy when using a WES panel. When analysing various strategies for imputation with the WES SSP we took the opportunity to test whether it would be better to combine all array data from all three villages (the previous approach of our

collaborators in Naples) or to impute separately. Bear in mind that Campora and Cardile were genotyped on a different chip to Gioi and combining these datasets would mean restricting to common sites between the two arrays. These two arrays were described in Section 2.1 and we will note them as ‘370K’ and ‘OMNI’. In Table 3.3.5, the differences in imputation accuracy (against the baseline IMPUTE2+1000G strategy) are given for a variety of slightly different imputation pipelines when tested on our simulation datasets.

Imputation Strategy	Imputation Accuracy of Non-Exonic Variants		Imputation Accuracy of Exonic Variants	
	MAF < 5%	MAF ≥ 5%	MAF < 5%	MAF ≥ 5%
<b>IMPUTE2+1000G</b> ( <i>baseline</i> )	0.470	0.781	0.510	0.797
<b>IMPUTE2+1000G+WES</b> (A1) <i>Reference panels are merged</i>	0.489 <b>+4.3%</b>	0.789 <b>+1.2%</b>	0.561 <b>+10%</b>	0.820 <b>+2.9%</b>
<b>IMPUTE2+1000G+WES</b> (A2) <i>Array positions added to SSP Reference panels are merged</i>	0.464 <b>-1.2%</b>	0.770 <b>-1.4%</b>	0.863 <b>+69%</b>	0.943 <b>+18%</b>
<b>IMPUTE2+1000G+WES</b> (A3) <i>Array positions added to SSP Reference panels not merged</i>	0.470 <b>-0.005%</b>	0.781 <b>-0.009%</b>	0.851 <b>+67%</b>	0.939 <b>+18%</b>
<b>IMPUTE2+1000G+WES</b> (A4) <i>Array positions added to SSP Reference panels are merged No relatedness to SSP</i>	0.487 <b>+3.8%</b>	0.786 <b>+0.7%</b>	0.520 <b>+1.9%</b>	0.804 <b>+0.9%</b>
<b>IMPUTE2+1000G+WES</b> (B1) <i>Array positions added to SSP Reference panels are merged Separate imputation on two genotyping arrays</i>	0.643 <b>+37%</b>	0.894 <b>+14%</b>	0.868 <b>+70%</b>	0.952 <b>+20%</b>

*Table 3.3.5 - On our simulated data, we performed imputation using the WES SSP and a variety of different pipelines. The baseline imputation was performed on all three villages together, using only variants found on both of the genotyping arrays in Cilento and using IMPUTE2 and the 1000G as a reference panel. In the middle section of the table we tested different ways of including a WES SSP. In the last section of the table we testing imputation on the two genotyping arrays separately.*

The clearest conclusion was that the biggest improvement to imputation that one will gain from a WES SSP was on exonic variants. Also it was clear that it would be best to perform our imputation separately on the two genotyping arrays (row B1). To go into details of Table 3.3.5, we observe that the only instance when imputation accuracy decreased substantially was for the non-exonic positions and when array position were included in the WES panel and when the merge-ref-panel option was used (row A2). We had applied this strategy for the WGS panel in our simulation study.



We can rationalise these observation in the following way: for a given target haplotype, the most likely copying states depend on the ability to match the array positions of the target haplotype to the array positions in the reference panel. Therefore, adding the array positions to the WES panel will facilitate matching between target haplotypes in the isolated population and study specific reference panel individuals (they are of course closely related in many cases). Hence, imputed genotypes become more likely to be inferred from the 93 individuals - not the 1000G individuals. This explains the increased imputation accuracy for WES positions when the array positions are added (rows A2 and A3 against row A1 for Exonic Variants).

However, as the Array positions are distributed relatively evenly, the same enhancement for matching will take place outside of exons as well. Hence, the imputed genotypes from non-exonic regions are also most likely to be inferred from the 93 WES individuals. Reflecting on this, it is evident that if the reference data for imputation outside of exons comes from the 247 SSP individuals, it will contain mostly variants that will have been already imputed from the 1000G through cross-imputation during the merging step. However, the quality of the imputation during the merging may be poor and so the hard calls made during cross imputation with create erroneous genotypes that will be subsequently imputed again. Therefore, it makes sense that the quality of imputation outside of exons will not be better when we include the SSP, and may in fact be worse (row A2). If the merge-ref-panel option is not activated, then this problem is avoided (row A3). When the option is not activated, imputation is performed in two steps, first from the SSP onto the target, second from the 1000G onto the target.

Furthermore, if the merge-ref-panel option is retained but the relatedness between the SSP and the target haplotypes is removed, again this problem does not occur (row A4 against row A2). We removed this relatedness artificially by using an SSP from one iteration of our simulation as an imputation panel for simulated target individuals from a different iteration. Therefore, we were

adding a WES panel of 247 individuals who were not related to our target haplotypes, but could nonetheless prove useful for imputation as they still consist of imperfect UK10K mosaic haplotypes.

In Figure 3.3.5, a visual representation of some of these changes in imputation accuracy was plotted. In this plot we focus in on a single exon and show the changes in imputation accuracy against the baseline accuracy of IMPUTE2+1000G for each variant against its physical position on the chromosome.

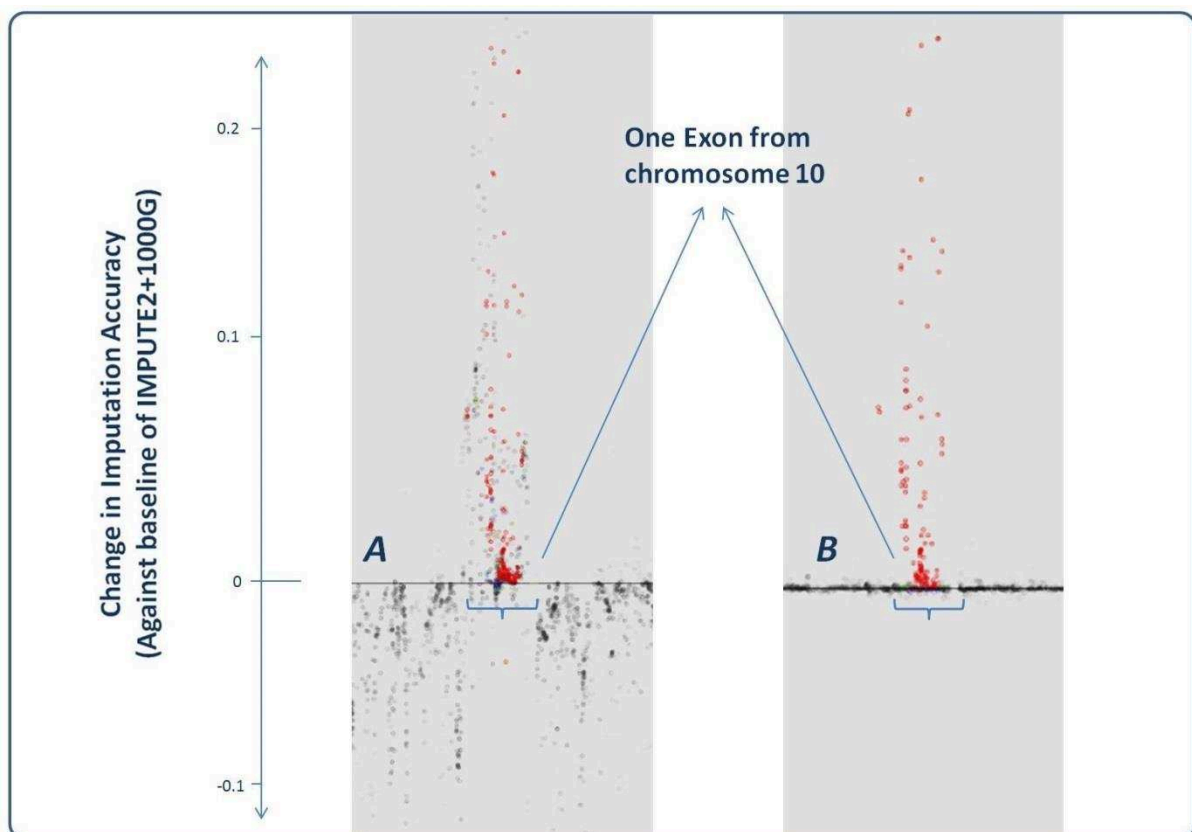


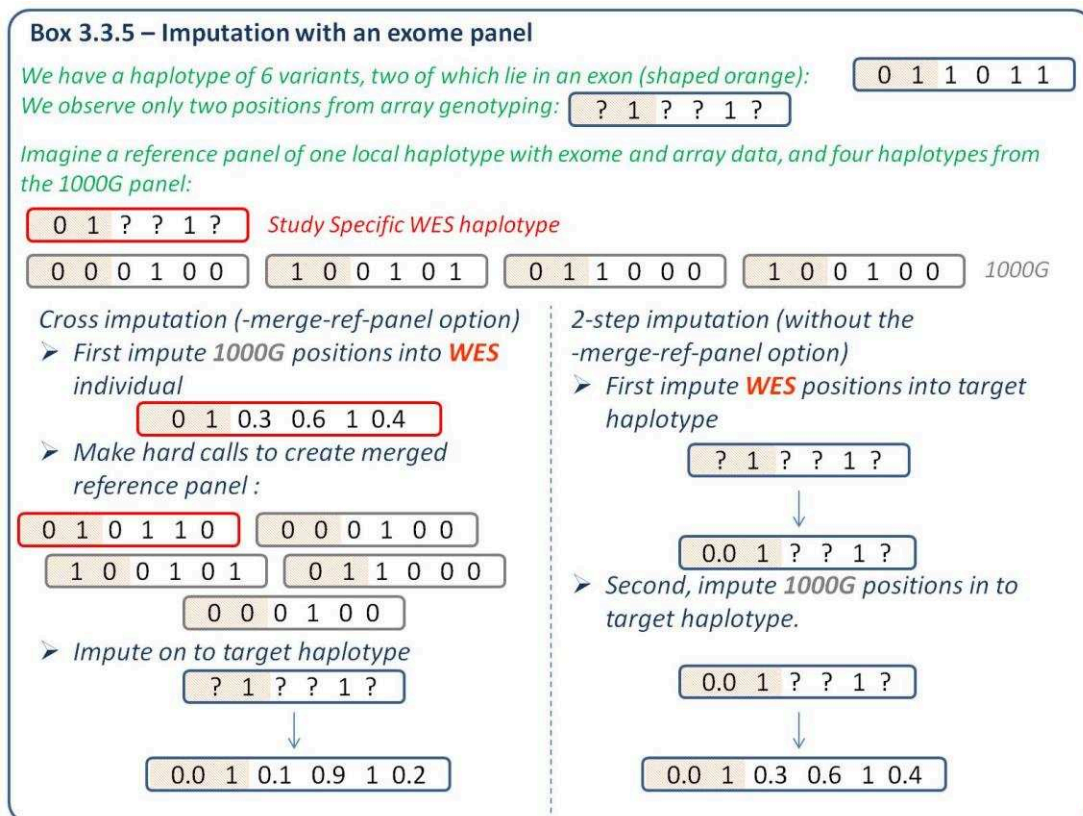
Figure 3.3.5. The change in imputation accuracy (against baseline imputation quality from the strategy IMPUTE2+1000G) is plotted across a chromosome 10 exon and surrounding regions. The variants present on the WES panel are highlighted in red and one exon of chromosome 10 is clearly visible, flanked by regions populated by variants present on the 1000G panel (black).

**Section A: IMPUTE2+1000G+WES. Array positions added to SSP. Reference panels are merged (row A2 in Table 3.3.5).**

**Section B: IMPUTE2+1000G+WES. Array positions added to SSP. Reference panels NOT merged (row A3 in Table 3.3.5).**

In Section A of Figure 3.3.5, the SSP with both exome and array positions is used and the merge-ref-panel option is activated. Here, we observe that the exonic variants (red) have increased imputation

accuracy; 1000G variants (black) within and very close to the exon also have increased accuracy. Outside the exon the accuracy decreases. In Section B, there is only one difference to section A: the panels are not merged. Here, we see that the exome positions in red increase in accuracy but all 1000G positions remain unchanged. In Box 3.3.5, an elucidation of how accuracy can be lost in in this specific case is given.



Box 3.3.5 - An example target haplotype with two positions in an exon and four positions outside is given. The imputation outside this exon from the 1000G is not very accurate, giving dosages of {0.3, 0.6, 0.4} for the three unobserved sites, whose true values are {1, 0, 1}. When the WES panel is used, during the merging stage, the values {0.3, 0.6, 0.4} are imputed into the WES panel (left section) and are then called as {0,1,0}. This exacerbates the inaccuracies of IMPUTE2 with final imputed dosages in the target of {0.1, 0.9, 0.2}, even further from the true haplotype. This issue is avoided if we do not merge the panels (right section).

Further exploration may be required to fully interpret these results and to understand what characteristics of the HMM imputation models have been demonstrated. The fact that the relatedness between target and reference individuals has led to problems in this instance perhaps indicates that when dealing with population isolate samples with complex structures, methods that do specifically model for IBD sharing can potentially suffer. Furthermore, this analysis shows that for

human genetics, imputation that involves WES reference data should be handled carefully due to the large gaps between exons. However, in animal models where LD extends much further and WES SNPs can better tag non-exonic regions<sup>249</sup>, these difficulties associated with using a WES SSP may well not have arisen.

To carry out imputation in Cilento, we decided to use the strategy IMPUTE2+1000G+WES with Array positions added to the WES SSP and using the merge-ref-panel option; despite knowing that this would lead to some loss of accuracy outside of exons. Our reasoning was that: (a) this strategy would improve the imputation in and nearby exons greatly and as our sequencing data is only on exons this should be a priority; (b) if we did not apply the merge-ref-panel option, any variants specific to the WES SPP would be removed by IMPUTE2 as without this option, only positions in the largest panel (1000G) can be imputed. Finally, we also carried out the imputation separately on the two genotyping arrays present in Cilento as this gave clear improvement in the simulation. We used the 1000G as the external reference panel as it contained many more variants than the version of the HRC that was available to us. We were able to confirm our simulation results by showing increases in the ‘info’ score when including the WES SSP for the imputation of the real data of Cilento<sup>174</sup>.

To use the full HRC panel would have required us to upload our data on to an external imputation server and thus preclude the merging step with our local WES panel. It is well known that imputation benefits from using the largest and most diverse imputation panel possible and it was not a surprise to see that the HRC performed better than the 1000G in our simulation. The idea being that as imputation algorithms are able to pick the most appropriate haplotypes for Copying States, making available additional haplotypes can only improve imputation<sup>225</sup>. Indeed, similar ideas about increasing reference panel diversity also appear in the literature of imputation in animal models<sup>250,251</sup>. It would have been potentially of great interest to use the full HRC for Cilento as it includes many haplotypes of Greek and Italian origin. However, our priority was to pursue the use of our WES SSP.

### 3.4 Prospective for Phasing and Imputation in Isolated Populations

We have established that haplotype phasing of array type data is exceptionally accurate in isolated populations. This means there remains little room for improving phasing accuracy, though leveraging larger external reference panels such as the HRC for phasing might provide another small improvement. The current main prospective for haplotype phasing in general is to be able to harness the phase information of read data. As WGS data is constructed from large ensembles of short genotyped reads (small haplotypes that are ‘read’ by the sequencing machinery), often the haplotype phase has actually already been measured. Indeed, a version of SHAPEIT2 is already in place to infer phase directly from read data<sup>222</sup> which stands alongside previous read-based phasing methods such as WhatsHap<sup>252,253</sup> and HapCut<sup>254</sup>. Furthermore, short distance haplotype calling (determining phase between very closely located variants) is currently available in the commonly used calling algorithm GATK<sup>255</sup>. The field of read-based phasing has now been joined by new methods whose development is tied directly into innovations in whole-genome sequencing methods<sup>256</sup>. A discussion of the phasing of WGS data including some of these methods was published in 2018<sup>257</sup> which showed SHAPEIT2 and EAGLE2 to still be very competitive with what is referred to as ‘laboratory-based’ methods; particularly when also considering the difference in cost of such approaches. However, as future studies of population isolates may rely more on direct WGS data, potentially using technology that can read longer haplotype fragments, it might become standard practice to adopt genotyping protocols that directly generate phased data.

One of the negative points of our first publication was that the rather limited choices of imputation software that we compared. All software tested were LD-based, whereas it would have been of interest to include software based on different methodologies as we had done so with the phasing analysis. We considered the use of imputation software PRIMAL<sup>111</sup>. PRIMAL was a perfect candidate for this study as it presents a phasing and imputation method specifically designed for isolated populations. It combines ideas of LRP and pedigree based imputation but, like EAGLE, utilises an LD-

based method (IMPUTE2) to complete the imputation in areas of the genome where IBD-based methods are not informative. We initially encountered difficulties with implementation, and despite continued efforts and discussions with the authors M. Abney and D. Nicolae of PRIMAL, we were never able to satisfactorily run the software on our simulated dataset. Indeed, we were advised to wait for the release of a new version. We also considered pedigree based imputation methods and chose to try and test GIGI<sup>258</sup> on the simulated dataset; this too was problematic due to the size of the Cilento Pedigree; finally we decided to not pursue matters further and submitted our work without including an imputation method based on IBD sharing.

We returned to the subject of phasing and imputation recently (after the publication of the article) to once again try to use GIGI. This imputation software directly uses the known pedigree structure along with array genotypes to infer the patterns of the inheritance. With this information, if some of the individuals (an SSP) have sequencing data, their genotypes may be imputed in to closely related individuals based on their connections in the pedigree. GIGI is indeed a development of the first genotype imputation methods in families discussed in section 3.1. We found that GIGI alone was not able to operate efficiently in Cilento for the following reasons: (a) the pedigree had to be split and the sub-pedigrees of genotyped members were too small; (b) the members of the SSP in Cilento were too few and too widely spread for this method. However, the ‘ped-pop’ method<sup>259,260</sup> that combines output from GIGI and IMPUTE2/MINIMAC3, gave hope of being able to attain imputation accuracy higher than from IMPUTE2/MINIMAC3 alone. This method selected the most confidently imputed genotype between the two output files and as we had observed that GIGI could impute well from SSP members to their siblings it was possible that the results of IMPUTE2 could be improved upon. However, when using ‘ped-pop’ we still observed a drop in accuracy of approximately 10% compared to IMPUTE2.

Very recently, a new software known as Kinpute<sup>261</sup> was released which takes the posterior genotype probabilities from IMPUTE2 as initial estimates (or as prior probabilities) and then re-estimate the

imputed genotypes via estimations of IBD sharing coefficients between the target individuals and individuals in the SSP. This method was shown to produce highly accurate results on the Hutterite data, but initial tests on our simulation data are showing a decrease in imputation accuracy (roughly 1% decrease on common variants though rare variants had equal accuracy) when compared to IMPUTE2.

The possible idea of combining IBD- and LD-based methods for genotype imputation that have been developed by PRIMAL, GIGI, and KINPUTE seem ideal for isolated populations but as of yet we have not been able to demonstrate their efficacy. This will be a continued area of future investigation using our simulation datasets.

During this thesis, I have exclusively concentrated on phasing and imputation literature relating to human genetics; but these are techniques that are also heavily used in the study of livestock populations. One software that could have been included in our study was FIMPUTE<sup>262</sup>, a long range phasing/imputation method first implemented for cattle datasets. However, this method had been previously shown to be slightly inferior to methods such as, IMPUTE2, and MINIMAC3<sup>263-265</sup>, though it may be able to outperform BEAGLE<sup>266</sup> and continues to be used in animal populations where detailed pedigree data is available<sup>267</sup>. We also did not test ALPHAIMPUTE<sup>268</sup> (after observing that ALPHAPHASE had very poor performance in our simulation); though this method was updated in 2017<sup>269</sup> and so should be re-examined. Now, inference from ALPHAIMPUTE can be combined with that of MaCH (a precursor to MINIMAC3), again bringing forth an idea of combining IBD- and LD-based methods. The idea that methodologies developed for studies of animals could be borrowed for analysing isolated populations (and vice-versa) is driven by the fact that in both cases, it may be possible to attain detailed pedigree structures whose information could be taken advantage of.

### 3.5 Conclusions on Phasing and Imputation

From this first study, we were able to establish the most appropriate phasing and imputation strategy for ourselves in Cilento, whilst also being able to provide useful recommendations to other researchers studying isolated populations. We had felt that combinations of IBD- and LD-based methods could unlock the highest levels of phasing accuracy; but SHAPEIT2 continued to demonstrate the highest performance levels. For imputation, we demonstrated the strengths of SSPs in a thorough simulation setting for the first time as well as showing that even a very small SSP can be highly effective in a population such as Campora.

Our unpublished results on using a study specific WES panel showed a potential weakness of IMPUTE2. The issue that we have highlighted regarding imputation with a study-specific WES panel could be easily rectified by the developers of the IMPUTE software, and could even be circumvented by ourselves – though it would take a rather lengthy and convoluted pipeline. Alternatively, there is no reason why genotype uncertainty could not be retained within a haplotype reference panel. A method allowing for this scenario would solve the issue we observed with the cross-imputed WES SSP. Previously, on a somewhat similar note, a specific issue had been brought up with the Li-Stephens model for haplotype phasing. C. Nettelblad<sup>270</sup> described how the presence of individuals sharing two haplotypes IBD can lead to phase estimates fixating on an incorrect outcome. Allowing the algorithm to escape from this local maximum was one of the many adjustments between SHAPEIT2 and the original incarnation SHAPE-IT<sup>220</sup>. The recent priority of imputation software has been to accommodate enormous sample sizes; for example, IMPUTE4 has far less user specified options than IMPUTE2 and so researchers focused on samples of related individuals may find that LD-based methods will be less and less likely cater to their specific needs and perhaps more bespoke methods that specifically model for IBD could be developed and provide advantages.



## Chapter 4: Heritability

### 4.1 Introduction of Concepts

For complex quantitative traits, there is often interest in estimating the proportion of the phenotypic variance that can be ascribed to each individual's inherited genome. This proportion of variance coming from the genome is referred to as heritability. Classically, the approach was to decompose the variance of a phenotype by calculating correlations of phenotypes between pairs of close relatives<sup>55</sup>. Analyses commonly focused on contrasting monozygotic and dizygotic twins, or by studying sets of sibling pairs of parent-offspring pairs<sup>110</sup>. Subsequently, linear mixed modelling has enabled the estimation of trait variance components from samples including pairs of individuals with various degrees of relatedness, such as studies of population isolates<sup>271</sup>. Such approaches either use extended pedigree information that gives the expected relatedness coefficients between each pair or individuals, or use individual level genotypes for directly estimating relatedness on the genome. It is intuitive that exact estimations of relatedness should be more informative than the expected values given by the pedigree due to the large potential stochastic variability of genome sharing between relatives<sup>228,272</sup>. Such techniques initially relied on finding shared sections of the genome identical-by-descent (IBD)<sup>79,226,229</sup> but more recently, it has been shown that correlations between genotypes allow for heritability estimation from any sample of individuals, including large samples of unrelated individuals<sup>273,274</sup>.

Firstly, I will fully explain the concepts of additive and non-additive heritability in a (non-inbred) population. The initial assumption is that our phenotype  $Y$  can be modelled as follows:

$$Y_i = \sum_{j=1}^M g_{ij} + \varepsilon_i$$

$M$  is the total number of causal variants. The index  $i$  indicates individuals and  $j$  indicates genetic variants,  $\varepsilon_i$  is the environmental component, assumed to be normally distributed under  $N(0, \sigma_E^2)$

and is independent of  $g_{ij}$ .  $g_{ij}$  is the genetic value of variant  $j$  of individual  $i$  which is dependent on the genotype which we denote as  $G_{ij}$ :

$$g_{ij} = \begin{cases} u_{j0}, & G_{ij} = AA \\ u_{j1}, & G_{ij} = Aa \\ u_{j2}, & G_{ij} = aa \end{cases}$$

As we assume that the genetic values are invariant across individuals, the subscript  $i$  will sometimes be dropped for brevity.

Considering the genetic value  $g_j$  as a random variable (as the genotypes are here considered as random), then the expected value of  $g_j$  is  $E[g_j] = p_j^2 u_{j0} + 2p_j q_j u_{j1} + q_j^2 u_{j2}$  where  $q_j$  is the minor allele frequency of variant  $j$  and  $p_j = 1 - q_j$ . We define  $\mu_j = E[g_j]$ .

The random variable  $g_j$  is defined on a probability space, mapping one of the three events  $\{AA, Aa, aa\}$  which occur with probabilities  $\{p_j^2, 2p_j q_j, q_j^2\}$  to the set of outcomes  $\{u_{j0}, u_{j1}, u_{j2}\}$ . It can be shown that  $g_j$  lies within a vector space of three dimensions governed by the values  $u_{j0}, u_{j1}, u_{j2}$  and with the following inner product:

$$\langle g_j^1, g_j^2 \rangle = E[g_j^1 g_j^2] = p_j^2 u_{j0}^1 u_{j0}^2 + 2p_j q_j u_{j1}^1 u_{j1}^2 + q_j^2 u_{j2}^1 u_{j2}^2 \quad \text{Eq. 4.1a}$$

The superscripts indicate different realisations of  $g_j$  coming from different values of  $\{u_{j0}, u_{j1}, u_{j2}\}$ . We can denote this vector space as  $\mathcal{V}$  where each random variable contained in  $\mathcal{V}$  is characterized by a tuple  $u = (u_{j0}, u_{j1}, u_{j2})$ .

We will first define  $X_j = g_j - \mu_j X_1$ , where  $X_1$  is a particular member of  $\mathcal{V}$  characterised by  $(1,1,1)$ .

$X_j$  is characterized by the tuple  $(u_{j0} - \mu_j, u_{j1} - \mu_j, u_{j2} - \mu_j)$ .

$X_j$  has an expected value of zero and is orthogonal to the variable  $X_1$ . This is seen as  $\mu_j = \langle g_j, X_1 \rangle$

and hence  $\langle X_j, X_1 \rangle = \langle g_j, X_1 \rangle - \mu_j \langle X_1, X_1 \rangle = 0$ .

We can decompose  $X_j$  into two further orthogonal components  $X_j^a$  and  $X_j^d$  so that for some values  $a_j$  and  $d_j$ :

$$X_j = a_j X_j^a + d_j X_j^d$$

Here,  $X_j^a$  and  $X_j^d$  are specified to have expectations of zero and  $\langle X_j^a, X_j^a \rangle = \langle X_j^d, X_j^d \rangle = 1$ . As with linear 3-dimensional space, there are countless ways to partition a vector into orthogonal components as there are countless different orthogonal bases that can be chosen, but we will choose two directions relevant to the variance of a genetic trait.

In particular we choose  $X_j^a$  to relate to the number of minor alleles in the genotype. Intuitively, this direction must be  $X_{012} = (0,1,2)$  so we specify  $X_j^a = \alpha(X_{012} - \beta X_1)$ . And solve for the constants  $\alpha$  and  $\beta$  by seeing that  $\langle X_j^a, X_j^a \rangle = 1$  and  $\langle X_j^a, X_1 \rangle = 0$ , this gives:

$$X_j^a = \frac{X_{012} - 2q_j X_1}{\sqrt{2p_j q_j}} = \left( \frac{-2q_j}{\sqrt{2p_j q_j}}, \frac{1 - 2q_j}{\sqrt{2p_j q_j}}, \frac{2 - 2q_j}{\sqrt{2p_j q_j}} \right)$$

This is recognizable as the normalized additively coded genotypes of the individual. Hence, as  $X_j^a$  describes the contribution of additive effects of the variant, we name the remaining orthogonal component  $X_j^d$  as the ‘non-additive’ component.

By setting  $X_j^d = (\delta_1, \delta_2, \delta_3)$  and by observing that by construction  $\langle X_j^d, X_j^d \rangle = 1$ ,  $\langle X_j^d, X_1 \rangle = 0$ , and  $\langle X_j^d, X_j^a \rangle = 0$ , it is straightforward to solve simultaneously to give the non-additive coding for the genotypes:

$$X_j^d = \left( \frac{q_j}{p_j}, -1, \frac{p_j}{q_j} \right).$$

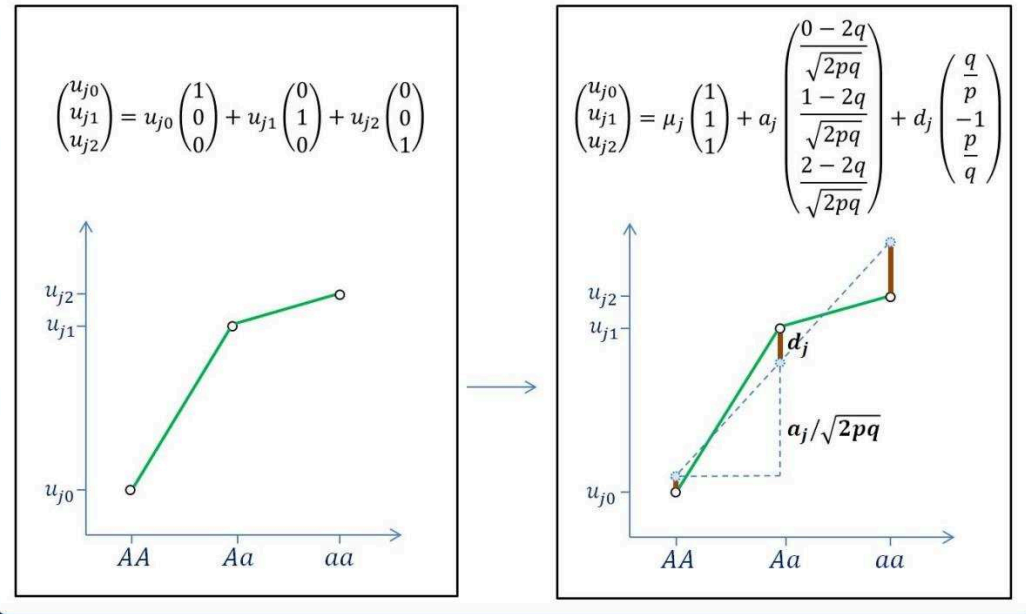
Hence:

$$Y_i = \sum_j \mu_j + \sum_j (a_j X_{ij}^a + d_j X_{ij}^d) + \varepsilon_i \quad \text{Eq. 4.1b}$$

In Box 4.1, the separation of additive and non-additive effects is depicted to give a clearer image of what is being described by each component.

**Box 4.1 Additive and Non-Additive components**

Here, a graphical representation of the decomposition of additive and non-additive components is given, invoking the idea of a change of orthogonal basis in 3D space.



Box 4.1 – The decomposition of a genetic value into additive and non-additive components.

In matrix form:  $\mathbf{Y} = \boldsymbol{\mu} + \mathbf{X}_A \mathbf{a} + \mathbf{X}_D \mathbf{d} + \boldsymbol{\varepsilon}$ , where  $\mathbf{X}_A$  and  $\mathbf{X}_D$  are  $N \times M$  matrices that represent the additively coded and non-additive coded genotypes of all individuals;  $\mathbf{a} = (a_1, \dots, a_M)$  and  $\mathbf{d} = (d_1, \dots, d_M)$ ; and  $\boldsymbol{\mu}$  is a vector of length  $N$  with all values equal to  $\sum_j \mu_j$ .

### 1) Classical interpretation

To reach this point, our stance has always been that the effects  $\mathbf{a}$  and  $\mathbf{d}$  are fixed and the genetic components  $\mathbf{X}_A$  and  $\mathbf{X}_D$  are random. This interpretation is the classical version introduced by R. A. Fisher<sup>55</sup>. We will later turn to the modern stance introduced by P. Visscher. The differences between these two stand-points is discussed in detail by Dandine-Roulland and Perdry<sup>275</sup>.

Consider the covariance of two individuals' phenotypic values,  $Y_i$  and  $Y_k$ :

$$\text{cov}(Y_i, Y_k) = \text{cov} \left( \sum_j \mu_j + \sum_j (a_j X_{ij}^a + d_j X_{ij}^d) + \varepsilon_i, \sum_j \mu_j + \sum_j (a_j X_{kj}^a + d_j X_{kj}^d) + \varepsilon_k \right)$$

We assume that all variants are in linkage equilibrium. Hence, all cross terms of the form  $cov(X_{ij_1}^a, X_{kj_2}^a)$  and  $cov(X_{ij_1}^d, X_{kj_2}^d)$  are null. As are the cross terms involving environmental effects as we assume there to be no covariance between genetic and environmental terms. In non-inbred populations, the cross terms  $cov(X_{ij}^a, X_{kj}^d)$  are also null. As a side note, while such terms do theoretically contribute in populations with inbreeding, there is very little consanguinity in Cilento and we chose not to consider the more elaborate covariance structures outlined in Abney, McPeck, & Ober<sup>110</sup>. Hence we arrive at the following:

$$cov(Y_i, Y_k) = \sum_j \left( a_j^2 cov(X_{ij}^a, X_{kj}^a) + d_j^2 cov(X_{ij}^d, X_{kj}^d) \right) + cov(\varepsilon_i, \varepsilon_k).$$

Therefore, we must evaluate the covariance of pairs of additive and pairs of non-additive components between pairs of individuals. To do this requires a decomposition over the different possible IBD sharing states between the two individuals:

$$\begin{aligned} cov(X_{ij}^a, X_{kj}^a) &= \sum_{l=1}^9 \Delta_l^{(ik)} cov(X_{ij}^a, X_{kj}^a | \Delta_l^{(ik)}), \\ cov(X_{ij}^d, X_{kj}^d) &= \sum_{l=1}^9 \Delta_l^{(ik)} cov(X_{ij}^d, X_{kj}^d | \Delta_l^{(ik)}). \end{aligned}$$

$\Delta_l^{(ik)}$  is the probability of individuals  $i$  and  $k$  being in IBD state  $l$ .

The workings through the possible states can be found in Jacquard<sup>58</sup>, and we arrive at the following well known results for a non-inbred population:  $cov(X_{ij}^a, X_{kj}^a) = 2\varphi_{ik}$ , where  $2\varphi_{ii} = 1$ ; and  $cov(X_{ij}^d, X_{kj}^d) = \psi_{ik}$ , where:

$$\psi_{ik} = \begin{cases} 1, & i = k \\ \Delta_7^{(ik)}, & i \neq k \end{cases}.$$

Importantly,  $cov(X_{ij}^a, X_{kj}^a)$  and  $cov(X_{ij}^d, X_{kj}^d)$  do not depend on the variant  $j$ . Therefore,

$$cov(Y_i, Y_k) = 2\varphi_{ik} \left( \sum_j a_j^2 \right) + \psi_{ik} \left( \sum_j d_j^2 \right) + cov(\varepsilon_i, \varepsilon_k).$$

The environmental components are assumed to be pairwise independent; i.e.  $cov(\varepsilon_i, \varepsilon_k)$  is equal to 1 if  $i = k$ , but is zero otherwise. We now introduce  $N \times N$  matrices  $\mathbf{K}$  and  $\mathbf{D}$  where  $K_{ik} = 2\varphi_{ik}$  and  $D_{ik} = \psi_{ik}$ .

Hence, in matrix form, the variance-covariance structure of  $\mathbf{Y}$ , written as  $\text{var}(\mathbf{Y})$ , is as follows:

$$\text{var}(\mathbf{Y}) = \left( \sum_j a_j^2 \right) \mathbf{K} + \left( \sum_j d_j^2 \right) \mathbf{D} + \sigma_E^2 \mathbf{I}_N$$

We then write  $\tau_a = \sum_j a_j^2$  and  $\tau_d = \sum_j d_j^2$  and define broad- and narrow-sense heritability, respectively, as:

$$H^2 = \frac{\tau_a + \tau_d}{\tau_a + \tau_d + \sigma_E^2} \quad \text{and} \quad h^2 = \frac{\tau_a}{\tau_a + \tau_d + \sigma_E^2}$$

Furthermore, we will write the additive and non-additive heritabilities respectively as:  $h_a^2 = h^2$  and  $h_d^2 = H^2 - h^2$ . If the trait is scaled by dividing through by  $(\tau_a + \tau_d + \sigma_E^2)$ , we have:

$$\text{var}(\mathbf{Y}) = h_a^2 \mathbf{K} + h_d^2 \mathbf{D} + (1 - h_a^2 - h_d^2) \mathbf{I}_N \quad \text{Eq. 4.1c}$$

## 2) Modern interpretation

An alternative interpretation was introduced by P. Visscher and first demonstrated in a landmark study of heritability in a sample of unrelated individuals<sup>273</sup>. Here, the important distinction is that the effects  $\mathbf{a}$  and  $\mathbf{d}$  are now considered as random and the genetic components  $\mathbf{X}_A$  and  $\mathbf{X}_D$  are considered as fixed. We assume that entries of the vectors  $\mathbf{a}$  and  $\mathbf{d}$  are draws of independent and identically distributed normal variable with means equal to 0 and variances equal to  $\sigma_a^2$  and  $\sigma_d^2$ , respectively. In this setting, the variance-covariance matrix of  $\mathbf{Y}$  becomes:

$$\text{var}(\mathbf{Y}) = \sigma_a^2 \mathbf{X}_A \mathbf{X}_A^T + \sigma_d^2 \mathbf{X}_D \mathbf{X}_D^T + \sigma_E^2 \mathbf{I}_N$$

Which we rewrite arbitrarily as:

$$\text{var}(\mathbf{Y}) = \tau_a \frac{1}{M} \mathbf{X}_A \mathbf{X}_A^T + \tau_d \frac{1}{M} \mathbf{X}_D \mathbf{X}_D^T + \sigma_E^2 \mathbf{I}_N$$

The link between this model and equation 4.1c is clear once we notice that as  $a_j \sim N(0, \sigma_a^2)$ ,  $(\sum a_j^2)$  is a moment estimator of  $M\sigma_a^2 = \tau_a$ , and likewise for  $\tau_d$  and  $(\sum d_j^2)$ . Again, if the trait is scaled by dividing through by  $(\tau_a + \tau_d + \sigma_E^2)$ , we have:

$$\text{var}(\mathbf{Y}) = h_a^2 \frac{1}{M} \mathbf{X}_A \mathbf{X}_A^T + h_d^2 \frac{1}{M} \mathbf{X}_D \mathbf{X}_D^T + (1 - h_a^2 - h_d^2) \mathbf{I}_N$$

Consider single entries of the two matrices  $\mathbf{X}_A \mathbf{X}_A^T$  and  $\mathbf{X}_D \mathbf{X}_D^T$ :

$$\frac{1}{M} \{\mathbf{X}_A \mathbf{X}_A^T\}_{ik} = \frac{1}{M} \sum_{j=1}^M X_{ij}^a \times X_{kj}^a$$

$$\frac{1}{M} \{\mathbf{X}_D \mathbf{X}_D^T\}_{ik} = \frac{1}{M} \sum_{j=1}^M X_{ij}^d \times X_{kj}^d$$

For any given variant  $j$ ,  $X_{ij}^a \times X_{kj}^a$  and  $X_{ij}^d \times X_{kj}^d$  are moment estimators of  $2\varphi_{ik}$  and  $\psi_{ik}$ , respectively. By averaging over all causal variants;  $\frac{1}{M} \sum_{j=1}^M X_{ij}^a \times X_{kj}^a$  is a moment estimator of  $K_{ik}$  and likewise  $\frac{1}{M} \sum_{j=1}^M X_{ij}^d \times X_{kj}^d$  is a moment estimator of  $D_{ik}$ .

Once the matrices  $K$  and  $D$  (which we will no longer write in bold from here on for ease of reading as we are always clearly referring to matrices) have been calculated, either from IBD probabilities or as moment estimates from the individual level genotypes,  $\tau_a$ ,  $\tau_d$ , and  $\sigma_E^2$  can be estimated via a linear mixed model (LMM). In our study, we use average information maximum likelihood estimation (AIREML)<sup>276</sup> to do so. When  $K$  and  $D$  are calculated as moment estimators from genotype data, we will refer to them as Genetic Relatedness Matrices (GRMs). This particular LMM will be referred to as Model KD:

$$\text{Model KD: } Y \sim MVN(U\beta_0, \tau_A K + \tau_D D + \sigma_E^2 I_N)$$

The term  $U\beta_0$  describes any potential addition covariates (such as sex or age) that may be added to the model.

An important distinction between the classical and modern derivations is that there is a difference in the interpretation of the relative sizes of the genetic effects. In the classical interpretation, there is no restriction or prior expectation on the values that  $u_{j0}$ ,  $u_{j1}$ , and  $u_{j2}$  can take. In the modern interpretation, the genetic effects  $a_j$  and  $d_j$  are assumed to be draws of normal distributions and contribute to the trait via the formula previously given above:

$$Y_i = \sum_j \mu_j + \sum_j (a_j X_{ij}^a + d_j X_{ij}^d) + \varepsilon_i$$

As the genotypes have been normalised to form  $X_{ij}^a$  and  $X_{ij}^d$ , this essentially leads to a relationship between the contribution of a genetic variant to the trait and the frequency of the variant. Indeed, this model of P. Visscher predicts that rarer variants will, on average, have larger genetic effect sizes.

In either case, it is typical for only narrow-sense heritability (the contribution of only the additive effect of the number of minor alleles in the genotype) to be calculated and presented. The non-additive component is not fitted. Results for non-additive heritability for complex traits have only rarely been estimated and results often vary greatly between different studies, this led us to begin a second simulation study primarily investigating non-additive heritability.



## 4.2 Published Results for Non-Additive Heritability Estimation

### 4.2.1 Interplay of Trait Architecture and Variance Components Estimates

In our second study we demonstrated some possible advantages of studying isolated populations for broad-sense heritability estimation. Through simulation and the study of the Cilento isolates we explained to an extent why large differences in estimations can be observed in the literature. We used the simulated data for all three villages of Cilento, using the combination of HapGen2 and gene-dropping as described in Section 2.2. Data from six different runs of this simulation were used in this study. We also simulated ‘outbred’ populations - simply by running HapGen2 on the UK10K to produce mosaic haplotypes. In fact, for this simulation study we did not actually use HapGen2 directly, we coded their algorithm ourselves so that we could record the mosaic segments used in the simulation; this would allow us to know the ‘true’ IBD sharing between all pairs of individuals in the simulated populations. For the majority of results we concentrated on four populations: (a) – “Isolated(1444)”, one realisation of the simulated isolated population with 1,444 individuals and the structure of the three Cilento villages; (b) “Outbred(1444)” – an outbred population of 1,444 individuals; (c) “Outbred(4332)”; and (d) “Outbred(8664)” – larger simulated outbred populations of size 4,332 and 8,664, respectively.

The first section of the study concentrated on estimations of  $h_a^2$  and  $h_d^2$  from running linear mixed models on phenotypes simulated under simple polygenic models. Details of phenotype simulation are given in the Method section of Annex B. Below, in Table 4.2.1, a collation of previous estimates of both heritability components from commonly studied traits are given. For two traits, LDL and BMI, I have plotted these estimates to demonstrate their diversity (Figure 4.2.1a).

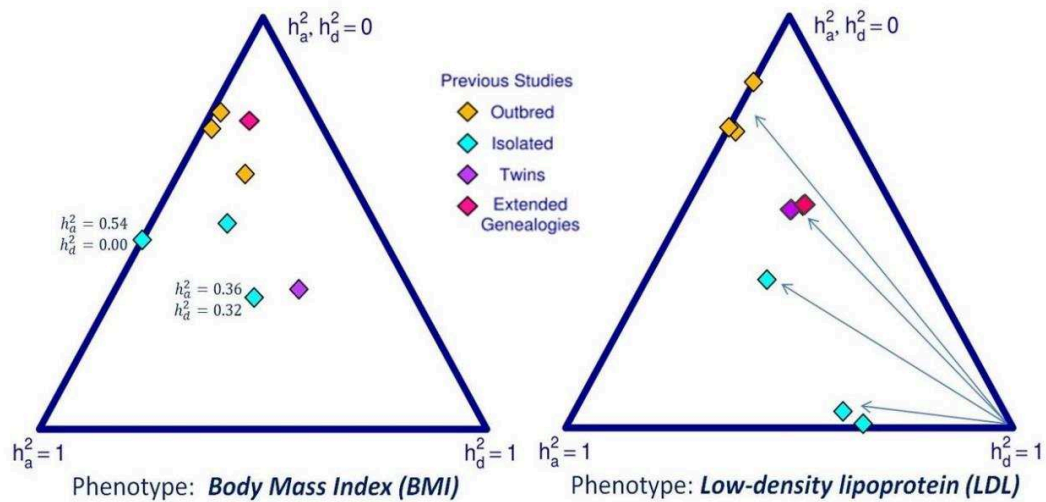


Figure 4.2.1a - We decided to plot these joint heritability component estimates on triangular figures to represent how the LMM distributes the phenotypic variance between the three variance components from the three variance-covariance matrices,  $K$ ,  $D$ , and  $I_n$ . This results in a slight bending of 2D space; essentially an isosceles triangle with vertices at  $(0,0)$ ,  $(1,0)$ , and  $(0,1)$  has been mapped on to an equilateral triangle to give equal prominence to all three components and a better aesthetic. Two specific values of BMI estimates have been added to aid comprehension of how values of  $h_a^2$  and  $h_d^2$  vary in the simplex. The construction of the graph means that one can roughly gauge the magnitude of the two estimates by observing the distances from a point to each vertex. The variety of estimates for  $h_d^2$  for the trait LDL have been indicated by their distance to the  $h_d^2$  vertex.

The difference in estimations from different study designs, we reasoned, could relate to either specific trait architectures, or to the different methods used for calculating matrices  $K$  and  $D$  (IBD-based in isolated populations, or moment based estimates in outbred populations). Our first analysis was concerned with the former conjecture – different trait architectures. Hence, we decided to analyse simulated populations in the same way: we simulated a myriad of phenotypes using random draws of random causal variants. For all populations, we calculated the variance-covariance matrices as GRMs and fitted the same linear mixed models (LMMs). We varied the polygenic trait architecture by changing the number and frequencies of causal variants.

Table 4.2.1	Abney, McPeck, & Ober. 2001. <sup>109</sup> , <i>N</i> = 806, Isolate (1)		Pilia et al., 2006 <sup>277</sup> , <i>N</i> = 6,148, Isolate (1) (2)		Traglia et al., 2009 <sup>278</sup> , <i>N</i> = 1,803, Isolate (1) (2)		Zaitlen et al., 2013 <sup>279</sup> , <i>N</i> ≈ 15,000, Extended Genealogies (3)		van Dongen et al., <sup>280</sup> , 2013, <i>N</i> ≈ 7,500, Twin Study (4)		Chen et al., 2015 <sup>281</sup> , <i>N</i> = 7,740, Twin Study (5)		Chen et al., 2015 <sup>281</sup> , <i>N</i> = 5,779, Outbred (5) (6)		Zhu et al., 2015 <sup>282</sup> , <i>N</i> = 8,682, Outbred (6)		Nolte et al., 2017 <sup>122</sup> , <i>N</i> = 13,436, Outbred (6)	
Phenotype	$h_a^2$	$h_d^2$	$h_a^2$	$h_d^2$	$h_a^2$	$h_d^2$	$h_a^2$	$h_d^2$	$h_a^2$	$h_d^2$	$h_a^2$	$h_d^2$	$h_a^2$	$h_d^2$	$h_a^2$	$h_d^2$	$h_a^2$	$h_d^2$
Height	-	-	0.77	0.23 *	0.78	0.22 *	-	-	0.81	0.09	0.77	0.09*	0.62	0.00	0.48	0.02	0.49	0.00
BMI	0.54	0.00	0.36	0.32 *	0.33	0.17	0.16	0.09	0.41	0.37	0.28	0.41*	0.21	0.02	0.23	0.15*	0.25	0.02
TGLY	0.37	0.00	0.30	0.42 *	0.39	0.35 *	-	-	0.33	0.25	0.42	0.14	0.31	0.28*	-	-	0.19	0.01
HDL	0.63	0.00	0.47	0.11	0.62	0.00	0.42	0.14*	0.40	0.27	0.66	0.00	0.24	0.01	0.25	0.07	0.19	0.00
Total Chol	-	-	0.38	0.29 *	0.23	0.77 *	-	-	0.51	0.16	0.28	0.19*	0.15	0.00	0.21	0.01	0.23	0.00
LDL	0.36	0.60 *	0.37	0.27 *	0.33	0.66 *	0.20	0.26*	0.51	0.18	0.23	0.24*	0.16	0.00	0.26	0.02	0.27	0.00

Published results for additive and dominant genetic variability from various study designs.

\* Estimates of  $h_d^2$  presented as statistically significant at the 5% level.

- Trait not studied for dominance in the article.

(1) Estimates based on estimating  $K$  and  $D$  from expected proportions of identity-by-descent (IBD) sharing coming from pedigree information.

(2) The depth of pedigree information in these studies did not allow the differentiation between a dominance model (including non-additive genetic variation) and a household model (including an effect of shared environment between siblings).

(3) The authors of this study analysed a large sample from the Icelandic population for whom extensive pedigree data was available, Matrices  $K$  and  $D$  were estimated by locating and counting stretches of IBD between pairs of individuals.

(4) This study analyses a large cohort of monozygotic and dizygotic adult twins. Standard errors are only presented for broad-sense heritability, though it is likely that the estimates for  $h_d^2$  for all traits other than height were significantly different to zero.

(5) The authors of this study performed separate analysis, firstly a twin based study using structural equation methods with adjustments for reported levels of time spent in a shared environment between twins, and secondly a study of a large sample of unrelated which included one individual out of most twin pairs in the first analysis.

(6) Estimates based on calculating correlations between additively and non-additively coded genotypes to compute matrices  $K$  and  $D$ .

**Abbreviations:** BMI: Body-mass index; TGLY: Triglycerides; HDL: High-density lipoproteins; Total Chol: Total cholesterol; LDL: Low-density lipoproteins;  $N$ : Sample size.

Having then estimated  $h_a^2$  and  $h_d^2$ , we could notice patterns across multiple realisations of the phenotype simulation. To give a full explanation of how the plots in the publication were designed, we include here Figure 4.2.1b (Figure 1 from Annex B) with a full description.

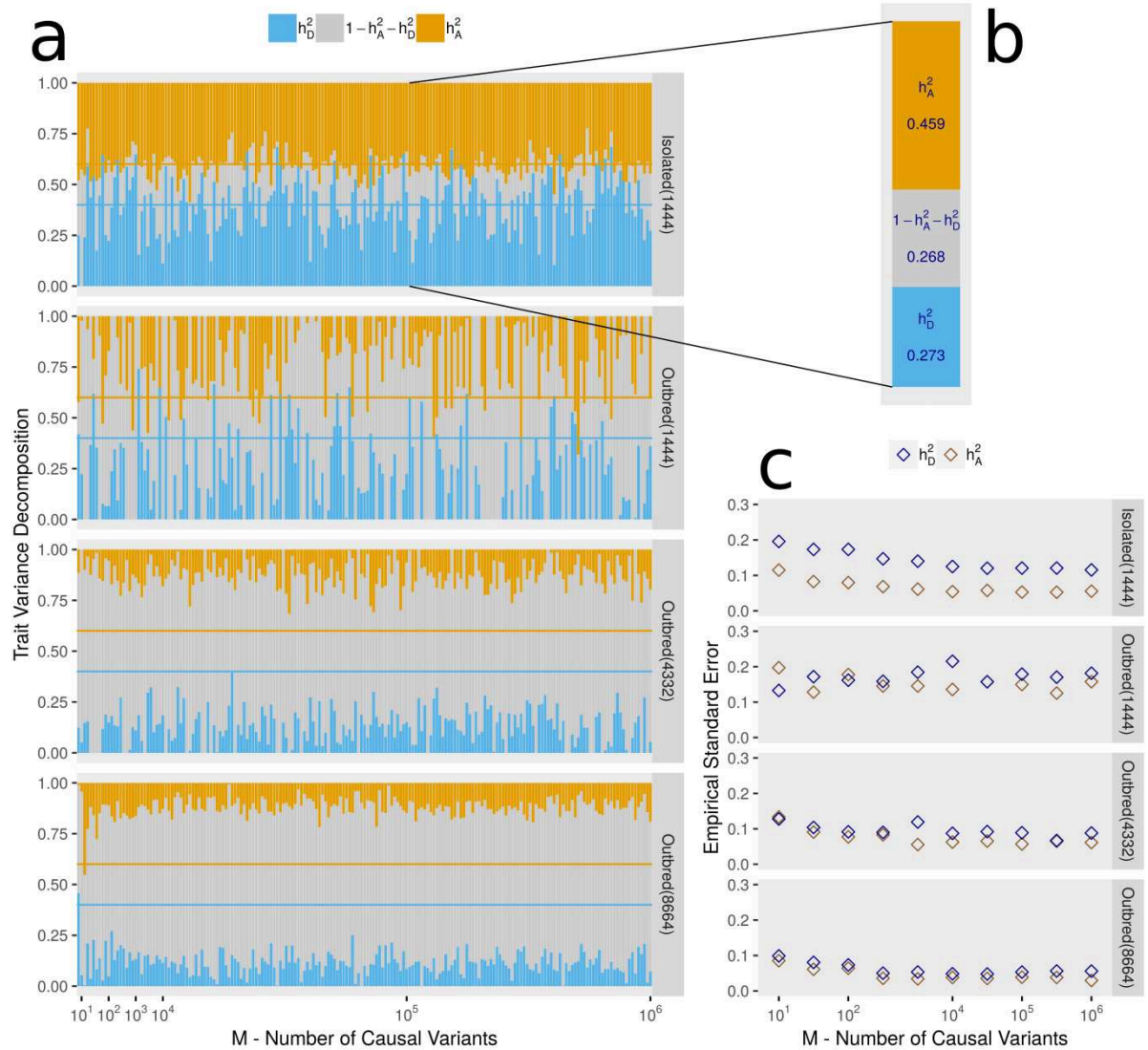


Figure 4.2.1b – For each of our simulated populations, we repeatedly simulated phenotypes by taking random draws of causal variants with either additive effects, non-additive effects, or both; and random draws of genetic effects ( $a_j$  and  $d_j$  from section 4.1); scaled so that  $h_a^2$  and  $h_d^2$  were both equal to 0.4; the horizontal lines in each sub-plot of section (a) denote these true values. In section (a), we vary the number of causal variants ( $M$ ) and present the maximum likelihood estimates (MLEs) for  $h_a^2$  (gold),  $h_d^2$  (blue), and indeed  $1 - h_a^2 - h_d^2$  (grey). Each simulated phenotype results in one thin vertical bar on the plot. Section (b) gives one example of such a set of MLEs. For a few particular values of  $M$ , we repeated the simulation 500 times in order to compute empirical variances of the estimates of  $h_a^2$  and  $h_d^2$  which we present in section (c), observing that there was less precision for phenotypes simulated using very small numbers of causal variants.

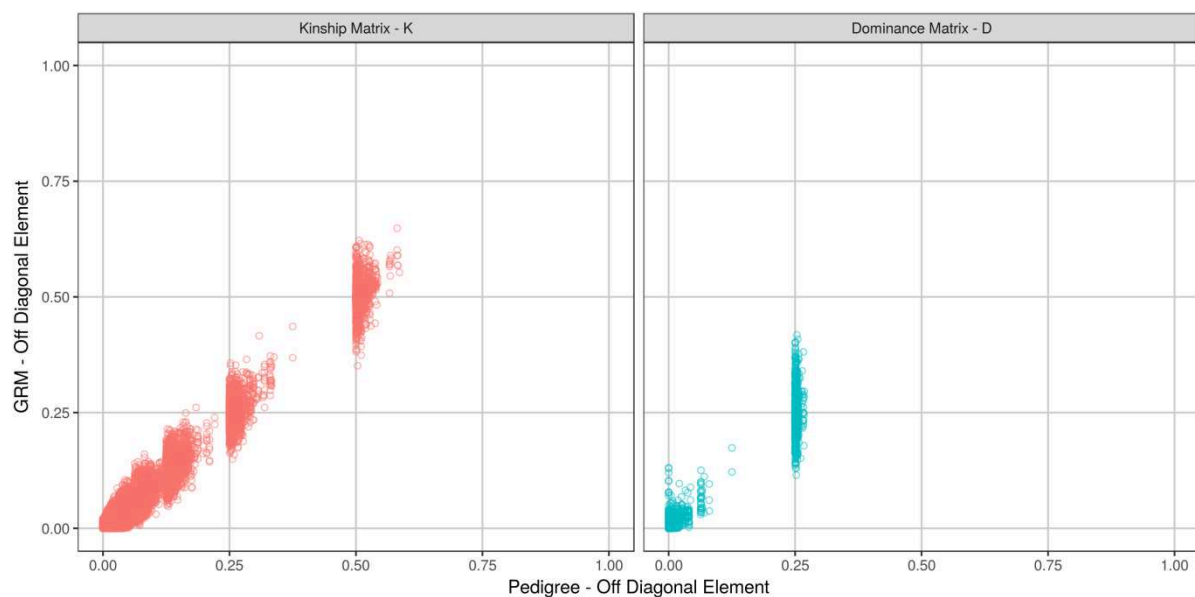
Two things were immediately obvious, that there would be very little precision in estimates of  $h_a^2$  in all populations and that estimates from outbred populations were downwardly biased; particularly when causal variants were rare. Estimates from isolated populations were however robust to changes in causal variant frequencies and to the numbers of variants made available for calculating GRMs (Figure 2 and Supplementary Figures 1 and 2 of Annex B). By increasing sample sizes of both isolated and outbred populations, we could substantially increase the precisions of the estimates of  $h_a^2$  and  $h_d^2$  (Figures 1,2, and 3 in Annex B). For the additive component, it is well known and well documented that analyses of unrelated individuals will underestimate heritability<sup>120,121</sup>. It was not surprising to see in our analysis a mirroring of this well documented phenomenon for non-additive heritability. This somewhat contradicts the conclusions of Zhu, et al.<sup>282</sup> who postulated that non-additive heritability was largely unimportant based upon interpretations of estimates of unrelated individuals when using very similar methods to estimate  $h_a^2$  as we have used here. However, our simulation was very simple and potentially naïve as our simulated phenotypes followed idealised polygenic models. It is unlikely that any human trait is a simple amalgamation of hundreds of thousands of small independent effects.

Furthermore, estimates from isolated populations may also be problematic and the rest of our study concentrated specifically on these challenges.

#### 4.2.2 How to Estimate Relatedness Matrices in an Isolate

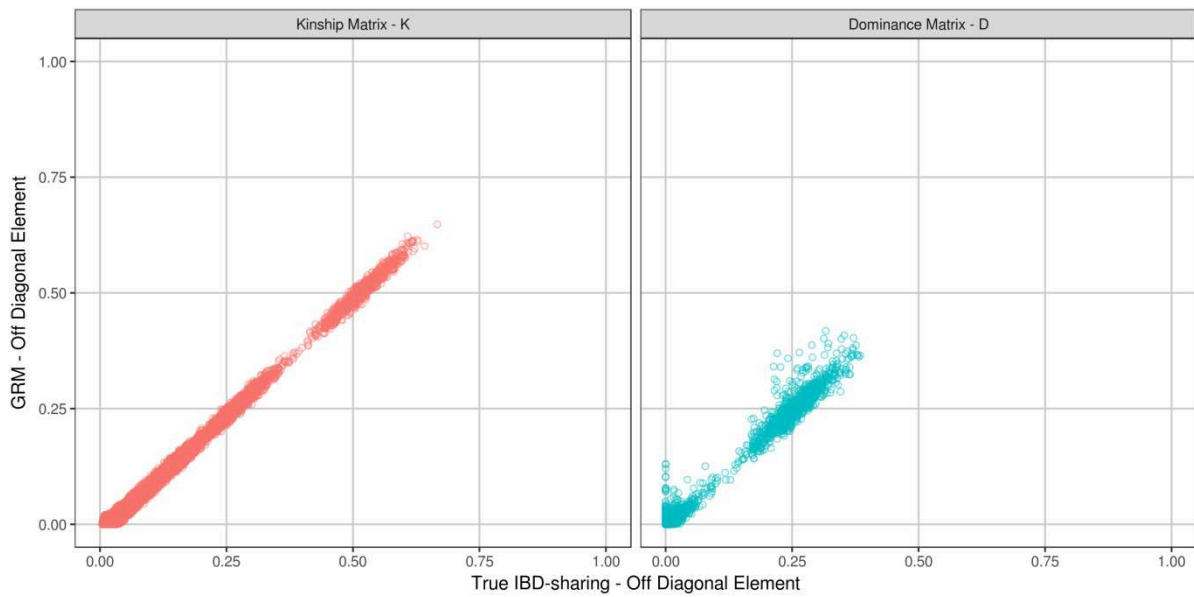
We investigated different methods for estimating the matrices  $K$  and  $D$  in isolated populations. We noted that in our comparison of previously published joint estimates of  $h_a^2$  and  $h_d^2$  that the studies of isolated populations used the expected IBD coefficients based on known pedigree structures. Conversely, the studies of outbred populations, not having evidence of close relationships, used GRM moment estimators to calculate  $K$  and  $D$ .

We compared three methods for estimating  $K$  and  $D$  in isolated populations: (a) pedigree based estimates; (b) true IBD-sharing based estimates (as we could either track every simulated allele back to a particular founding haplotype); and (c) using GRMs. In Figures 4.2.2a and 4.2.2b (taken from the Supplementary Figures 4b and 4c of Annex B) we compare the off-diagonal entries of these matrices. The off-diagonal elements from pedigree-based matrices are less continuous than other methods. From a pedigree, only the expected values of IBD sharing can be calculated. Hence, for example, all of the sibling pairs have an expected value for their non-additive covariance roughly equal to 0.25 (Figure 4.2.2a). Whereas the estimated values for these covariance values from GRMs vary greatly around 0.25. The idea being that, for example, we expect two siblings to have a kinship coefficient of 0.5 and a coefficient of IBD=2 sharing equal to 0.25. However, the actual values will depend on the stochastic transmission from the two parents. It is intuitive therefore that using coefficients that capture the exact IBD-sharing between each pair will better reflect the assumptions of the LMM.



Figures 4.2.2a – Comparison of off diagonal elements of matrices  $K$  and  $D$  calculated either from the pedigree of Cilento or as GRMs.

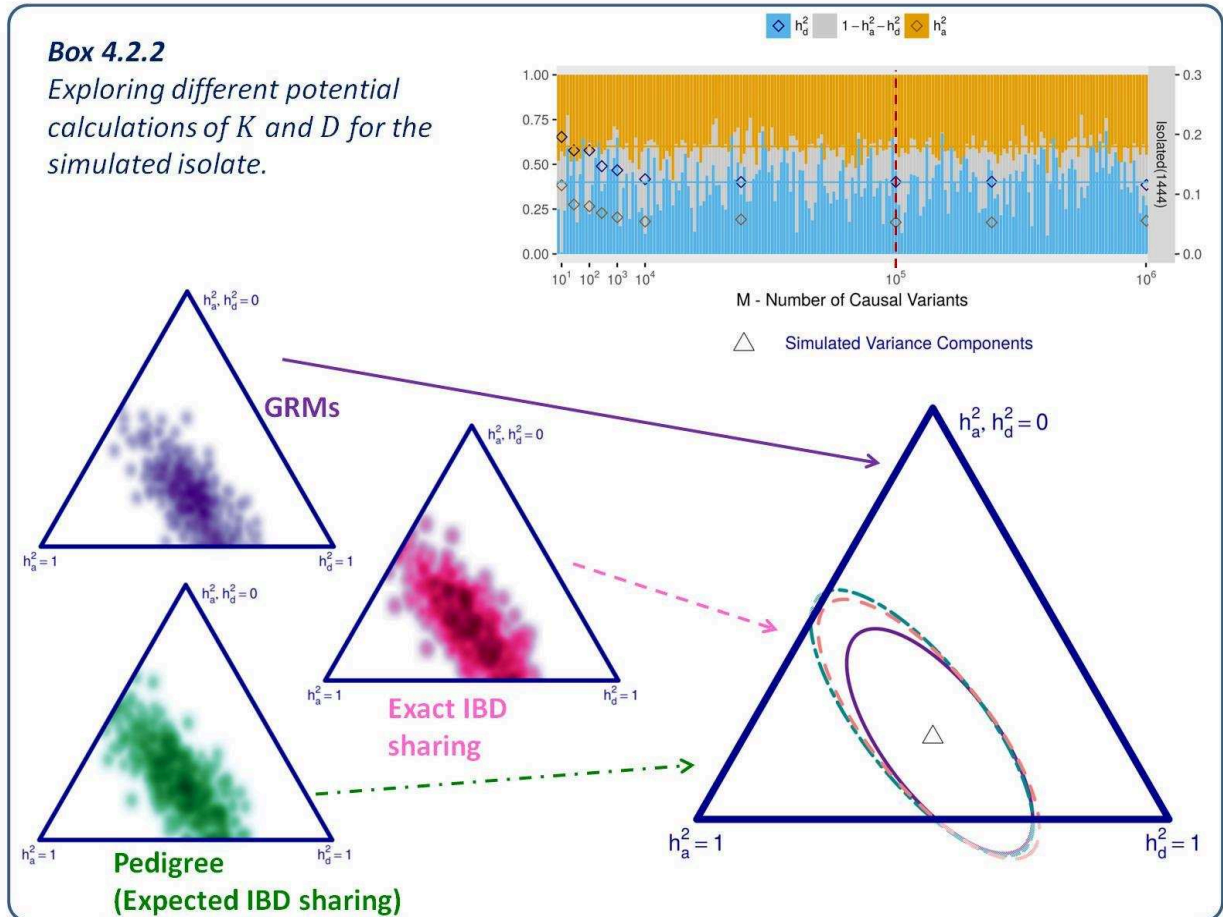
We estimated heritabilities  $h_a^2$  and  $h_d^2$  for previously simulated phenotypes, using these different methods for estimating  $K$  and  $D$ .



*Figures 4.2.2b – Comparison of off diagonal elements of matrices  $K$  and  $D$  calculated either from the true IBD-sharing (recorded during the simulation) or as GRMs.*

The most interesting result was to see that the ‘best’ estimates came from GRMs, and this was a universal result across many different scenarios (Figure 4 in Annex B and Supplementary Figure 5 in Annex B). One particular result is summarised in Box 4.2.2.

The ellipses in Box 4.2.2 reflect the variance in the estimation of the additive and non-additive components, with their forms showing the large lack of precision in the estimate of  $h_a^2$  which could generally vary between 0 and close to 1. If we concentrate on the estimation of the additive components, we can see that all methods performed quite similarly as the ellipse have similar minor axes. For the major axis which describes the non-additive component, there were some differences and the GRMs appear to give more precise results. We were somewhat surprised that the GRMs outperformed the matrices derived from the true simulated IBD-sharing. We presented the following conclusions from this particular analysis: firstly, using GRMs in isolated populations is appropriate (as this was not an approach generally used before though has been previously explored for related individuals<sup>283</sup>). Secondly, that using GRMs may hold an advantage over methods that specifically model tracts of IBD-sharing; IBD-sharing estimation methods have been applied for isolated populations, e.g. Vitart, et al.<sup>284</sup>



*Box 4.2.2. Here, minimal ellipses containing at least 95% of the 500 maximum likelihood estimates from repeated phenotype simulation are given. A section of Figure 4.2.1b is given to put these new results in context. We set the number of causal variants to  $10^5$  and replicated the phenotype simulation 500 times. The three small LMM triangles on the left of the box show heat maps of all of the 500 MLEs resulting from fitting out LMM using different versions of matrices  $K$  and  $D$ .*

The reason that GRMs could hold an advantage is that they are able to capture similarities between genotypes that go beyond the IBD-sharing that we recorded in our simulation. In our simulation, we assume that our original 200 founding haplotypes, which were drawn to simulate our data, were all completely unrelated. But these haplotypes were from the UK10K and so could easily share regions IBD. Certainly they will share many short regions shared IBS. This could seep down through our simulation and lead to similarities between individuals that we do not record as IBD.

Which approach is ‘better’ may in fact not a useful question to ask. First, all of our interpretations have been based on the assumption of the polygenic model. If we do accept this model, depending on how the actual causal variants are distributed, it may be arbitrary as to which method will give an



estimate that happens to better represent the set of variants that truly have an effect. GRMs were calculated using different sets of variants, one large set, and one much smaller set, and results were largely unaffected. This is shown in the closeness of Figure 4 in Annex B and Supplementary Figure 5 in Annex B. However, as causal variants were always selected at random, in some cases many of them may have been sampled from the sets of variants used to estimate the GRMs. Indeed, we observed during early explorations of our simulation that if the exact set of causal variants were used to calculate the GRMs, we would of course get highly accurate heritability estimates.

### 4.3 Confounding with Shared Environmental Factors

From the initial results regarding the interplay of trait architecture and heritability estimation, it could be tempting to conclude that the low estimates of  $h_d^2$  from previous studies of outbred populations indicate trait architectures involving many rare variants; or possibly variants that are not well covered by LD. Equivalent hypotheses are currently being explored in great detail for narrow sense heritability for many complex traits<sup>274,285,286</sup>.

Our simulated phenotypes followed polygenic models precisely. Two important assumptions that our polygenic model makes and that we know to not be good assumptions to make are that: (a) there is no contribution to the covariance structure from interaction effects between different genetic variants; (b) there is similarly no contribution from interactions between genetic and environmental effects. We did not address point (a) in our study, though interactions between genetic variants (epistasis) have been shown to significantly affect heritability estimation<sup>120</sup>. Neither did we explore explicit non-independence of genetic and environmental effects. We did however broach this topic by simulating phenotypes using models that describe covariance between environmental effects between certain pairs of individuals: siblings. We showed that this gives a confounding between the non-additive genetic and environmental components. Phenotypes were simulated under the following model:

$$\text{Model KDS: } Y \sim MVN(U\beta_0, \tau_A K + \tau_D D + \sigma_S^2 S + \sigma_E^2 I_N)$$

The additional matrix,  $S$ , has values of 1 on the diagonal and at every off-diagonal element corresponding to pairs of siblings in the sample; all other entries are zero. A confounding arises because the non-additive effect as the matrices  $D$  and  $S$  are so similar, making it difficult to simultaneously estimate  $\tau_d$  and  $\sigma_S^2$ . The matrices are similar because the siblings are the pairs of individuals with by far the highest proportions of IBD=2 sharing.

The proportion of variance assigned to the shared environmental component between siblings will be written as  $h_S^2$  (defined below) and is sometimes referred to as the ‘Household’ effect.

$$h_S^2 = \frac{\sigma_S^2}{\tau_a + \tau_d + \sigma_S^2 + \sigma_E^2}$$

Indeed, in Traglia, et al.<sup>278</sup> and in Pilia, et al.<sup>277</sup> there was complete compounding of these effects as they studied populations where the only pairings with expected values of IBD=2 sharing different from zero were siblings. Note that this was a result of using pedigree based estimates for the relatedness matrices. This highlights another motivation for using matrices based on estimates of exact IBD-sharing or GRMs rather than the expected values from pedigree structure. If a true continuum of IBD-sharing values make up the off-diagonal elements, this is unlikely to be matched and hence confounded with shared environmental effects<sup>109,110</sup>. From results presented in Abney, et al.<sup>110</sup>, we could see that the distribution of estimates of  $\Delta_7$  in the Hutterite population, estimated from their pedigree, was less polarised than the equivalent distribution in Cilento. To elaborate, in the Hutterite population, the amount of loops in the pedigree and the sharpness of the founding bottleneck predict a wide range of potential IBD=2 sharing probabilities. In Cilento, conversely, we can see in Figure 4.2.2a that there are only sufficient loops in the pedigree to give expected values of off-diagonal elements of  $D$  that are only slightly different from either 0 (unrelated) and 0.25 (siblings). These coefficients are then more continuous when estimated from GRM matrices yet values are still clearly divided into two groups (sibling pairs and non-sibling pairs). Hence, analyses may be susceptible to confounding with sibling status.

We showed through our simulation that large confounding was possible even when there was just a very small Household effect. For example we simulated phenotypes with the following variance parameters:  $h_A^2 = 0.4$ ,  $h_D^2 = 0.4 - h_S^2$ , where  $h_S^2$  took the following values: 0.00, 0.02, 0.05, 0.10, 0.20, and, 0.40. The results from these simulations are given in Supplementary Figures 7a-f in Annex

B. The results where  $h_S^2 = 0.2$  are presented here in Figure 4.3 and also in Figure 5 in the main text of Annex B.

These analyses showed that when the shared environmental effect increased - it was very likely that LMMs would give a broad-sense heritability estimate of 1, a phenomenon that will be interesting to explore further and that can in fact be observed in previous estimates from isolated populations (see Table 4.2.2). When pedigree information was used to estimate the matrices  $K$  and  $D$ , there appeared to be slightly more sensitivity to the presence of both non-additive effects and shared environmental effect; probably because in this case the matrices  $D$  and  $S$  had fewer differences.

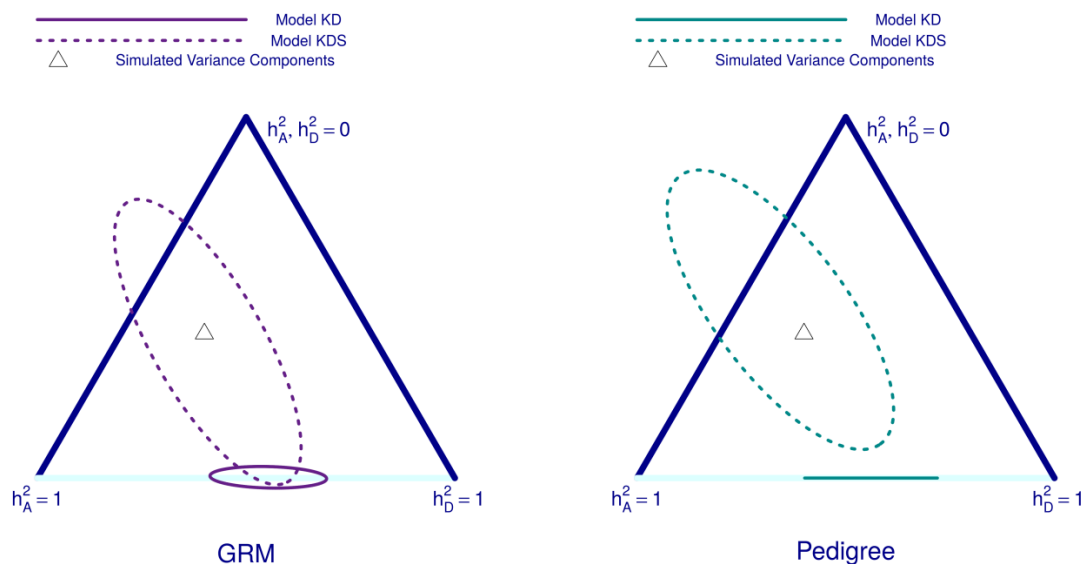


Figure 4.3 - Contrasts between model KD and model KDS, illustrating the confounding between non-additive effects and shared environmental effects between siblings in the simulated population isolate. Results are split dependent on whether GRMs or Pedigree estimates were used to form matrices  $K$  and  $D$ . The plot follows the same form as in Box 4.2.2

In our simulation study we were able to show that one way around this problem would be to exclude one member of each sibling pair from variance component analysis. However, in an isolate with the size and characteristics of Cilento, this would decrease the sample size greatly. We presented results from a simulated population that we called “Isolated(5136)\_nosibs”, a congregation of simulated individuals across the six simulated versions of Cilento without any sibling pairs. We showed that a reasonable estimates of  $h_a^2$  could be found in an isolated population with no siblings and just from the information carried by more distantly related pairs who nonetheless

share regions with IBD=2. Results from this analysis are given in Supplementary Figures 8a-d of Annex B.

In both this analysis, which excluded siblings, and the analysis involving different sample sizes in a population isolate (Figure 3 Annex B), we had combined multiple simulation datasets based on Cilento. Without larger pedigree structures at hand for other isolates, this was our choice of method to explore isolate type data that was not limited to 1,444 individuals or less. These analyses raised the question of whether it would ever be possible to combine data from several population isolates in order to perform such heritability analyses. Our results suggested that there would be clear advantages in the precision of the estimates as well as allowing for analysis of subgroups of pairs (e.g. non-sibling pairs). However, would it be reasonable to do so? Given that in each population, we should expect the values of  $h_a^2$ ,  $h_d^2$ , and  $1 - h_a^2 - h_d^2$  to be different. The amount of environmental variation will be different between isolates as will the genetic components depending on the frequencies and distribution of causal variants in each population. Nevertheless, whilst the values of heritability estimates across different populations might be nonsensical, it could still be a powerful approach for establishing whether certain components are at least different from zero.

## 4.4 Analysis of the Cilento Isolates

### 4.4.1 Heritability Analyses

We analysed six traits for heritability in Cilento, with our findings slotting in nicely into the collection of previous results from the literature that we had gathered in our study. We focused particularly on two traits, BMI and LDL. By comparing estimations based on the Model KD and Model KDS and by comparing our results to those of the simulation study involving shared environmental factors we tried to decipher our heritability estimates from Cilento and to see what conclusions could be drawn. In Figure 4.4.1a, we present the likelihood profiles of the two variables  $h_A^2$  and  $h_D^2$  from the Model KD for the two traits BMI and LDL. We include the MLE for  $h_A^2$  and  $h_D^2$  from the Model KDS (green peak) and we can see that in both cases that changing from Model KD to Model KDS results in a reduction in the estimate of the dominance component. In Table 4.4.1, all heritability estimates from the study are given (taken from Table 2 in Annex B), where various different models were tested and further contrasts between using GRMs or pedigree information are given.

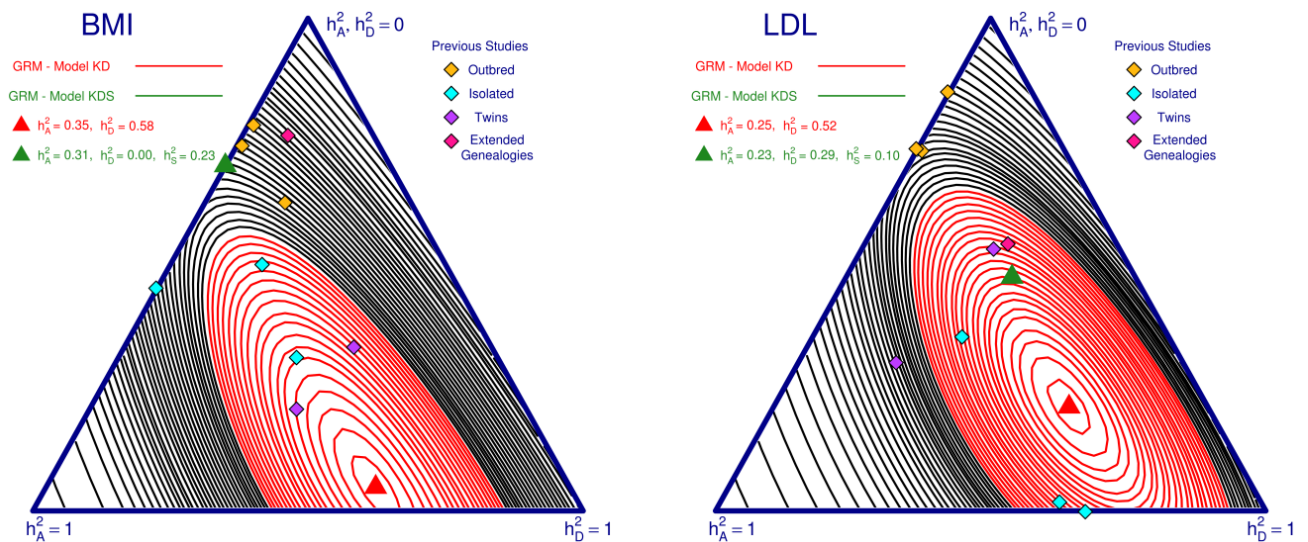


Figure 4.4.1a – Likelihood profiles from the model KD for BMI and LDL. The red region represents the 95% confidence interval of the MLE estimates for heritability (red peaks). We overlay the previous results from the literature as well as the MLEs from the model KDS (green peaks).

The initial estimate of BMI from Model KD gave  $H^2$  as 0.93 (close to 1) and a high value for  $h_a^2$  of 0.58. If pedigree information was used rather than GRMs, then  $H^2$  was estimated as exactly 1 and  $h_a^2$  as 0.65 (Table 4.4.1). The red confidence interval for the MLE suggested that the estimate for  $h_a^2$  was significantly different from zero. However, when the sibling matrix was included (Model KDS) the estimate for  $h_a^2$  fell to zero, with most of the variance that had previously been ascribed to  $h_a^2$  switching to the environmental component ( $1 - h_a^2 - h_s^2$ ). Adding the matrix  $S$  to the covariance structure in the model did not affect estimates of  $h_a^2$  in Cilento, which was in accordance with our simulations that compared models KD and KDS. If we compare the results for BMI to our simulation results that compared Model KD with Model KDS (described in Section 4.3), they are probably most compatible with the scenarios where  $h_a^2$  was small or even zero and  $h_s^2$ , the Household effect, was large.

For the trait LDL, the estimate for  $h_a^2$  was also significantly different from zero under the Model KD based on the presented 95% confidence interval. The MLE under Model KD is also quite far from the bottom edge of the figure. Also, the shift of the MLE when we moved to Model KDS was comparatively small when compared to the other traits studied such as BMI. Coupling these observations suggests that the initial estimation of  $h_a^2$  was probably not being completely driven by an unaccounted for Household effect. This allowed us to draw a tentative conclusion that our analyses gave a good indication of a non-zero non-additive (or dominance) component for LDL in Cilento. Based on these results we decided to continue to analyse the trait LDL in detail, including analysis of a recently completed imputation dataset of Cilento.

<b>Table 4.4.1</b>	GRM Model K	GRM Model KD		GRM Model KS		GRM Model KDS			Pedigree Model K	Pedigree Model KD		Pedigree Model KS		Pedigree Model KDS		
<b>Phenotype</b>	$h_a^2$	$h_a^2$	$h_d^2$	$h_a^2$	$h_s^2$	$h_a^2$	$h_d^2$	$h_s^2$	$h_a^2$	$h_a^2$	$h_d^2$	$h_a^2$	$h_s^2$	$h_a^2$	$h_d^2$	$h_s^2$
Height	0.76	0.74	0.13	0.74	0.04	0.74	0.12	0.01	0.75	0.74	0.15	0.74	0.04	0.74	0.15	0.00
BMI	0.40	0.35	0.58	0.31	0.23	0.31	0.00	0.23	0.44	0.35	0.65	0.35	0.21	0.35	0.00	0.21
TGLY	0.27	0.24	0.26	0.21	0.11	0.21	0.00	0.11	0.28	0.23	0.45	0.23	0.11	0.23	0.41	0.01
HDL	0.49	0.49	0.00	0.44	0.02	0.44	0.00	0.02	0.48	0.49	0.00	0.48	0.01	0.48	0.00	0.01
Total Chol	0.29	0.23	0.55	0.23	0.18	0.22	0.27	0.12	0.29	0.21	0.72	0.22	0.18	0.21	0.47	0.06
LDL	0.32	0.25	0.52	0.24	0.17	0.23	0.29	0.10	0.33	0.24	0.66	0.24	0.16	0.24	0.45	0.06

Maximum likelihood estimates for the contribution of each variance components considered in a Linear Mixed Model. Model names refer to the set of variance components included.  $K$  denotes the additive genetic component,  $D$  the non-additive or dominant genetic component, and  $S$  the component accounting for shared environmental effects between siblings. Matrices  $K$  and  $D$  are calculated either as genetic relationship matrices (GRMs) or from pedigree information.



#### 4.4.2 GWAS of LDL in Cilento

To complement our heritability results, we set out to perform both additive and non-additive GWAS of the trait LDL in Cilento. We tested these components separately, knowing the tests to be orthogonal by construction, to see if evidence of our polygenic model comprising additive and non-additive effects (equation 4.1b) could be found. Note, that testing both of these orthogonal components and then combining their test-statistics would be equivalent to a two degrees of freedom test under a general model where each of the three genotypes is assumed to have a specific effect (*AA* vs. *Aa* vs. *aa*).

This section will touch on our previous investigation into genotype imputation in Cilento as it involves our completed genome wide imputation dataset in Cilento, using the whole-exome sequenced study specific panel (the WES SSP). The imputed dataset has been passed back to our collaborators in Naples and will continue to be analysed for multiple traits and as we have harnessed local sequencing data, we have hope of finding new associations in Cilento. Through our detailed simulation studies, we believe that we have created a highly accurate imputation dataset. Here, it will be instructive to trial this dataset by performing a GWAS for a widely studied trait such as LDL. Hence, the following analysis is both a confirmation of our imputation strategy and a follow-up of our heritability results. We performed our GWAS on hard-called imputed genotypes because, as of yet, we have not implemented methods necessary for analysing non-additive components from imputed dosage data (this is discussed in detail in Section 4.5). Using hard-calls is of course problematic when imputed genotypes are uncertain. Hence, we put a threshold of 0.8 for the top posterior genotype probability and set genotypes failing this threshold to zero. This will detract from our results, but for the moment this was primarily an initial test of the imputation dataset and a continued look at non-additive effects.

The trait LDL describes the level of low-density lipoprotein in blood plasma. This has been widely studied in genetic epidemiological studies, finding associations with many genes and loci<sup>287-290</sup>, many

of whose functionalities have previously been well described<sup>291</sup>. 789 GWAS hits for LDL from 65 studies are currently (December 2018) listed in the GWAS catalogue. Particular genes that could well be found include: APOE, APOC1, APOC2, SORT1, LDLR, APOB, PCSK9, LDLRAP1, and NPC1L1<sup>289,292</sup>.

The distribution of the trait across Cilento and within each of the three villages is given in Figure 4.4.2a. The trait was judged to not require a transformation though it was adjusted appropriately for the intake of certain lipid lowering medications. We fitted LMMs including additional explanatory variables for the membership of each village, age, sex, and the interaction age $\times$ sex. The covariance structure used in the model included the two relationship matrices  $K$  and  $D$ , both estimated as GRMs from dense, hard-called, high quality (IMPUTE2 ‘info’ score > 0.7) imputed data.

To give a brief aside, we are in effect applying Model KD with additive/non-additive components of each variant being included as explanatory variables in the model in turn. This was our chosen initial approach but discussion continues with our collaborators as to what should be the optimal procedure before they embark on further association tests of the many phenotypes available in Cilento. For example, when testing a variant on a certain chromosome, it may have been possible to make gains in power by using estimators of  $K$  and  $D$  based on markers from all other chromosomes<sup>293</sup>. Furthermore, we must also scrutinise the decision of whether or not to include the dominance matrix  $D$  in our association model, as we should assess how the precision (or lack of precision) of our estimates of this matrix will affect our association models<sup>294</sup>.

We proceeded to perform GWAS of the additive and non-additive genetic components for LDL in Cilento; here 9,633,547 variants were tested and in Figure 4.2.2b I give QQ-plots of p-values in these two GWAS and corresponding Manhattan plots in Figure 4.2.2c.

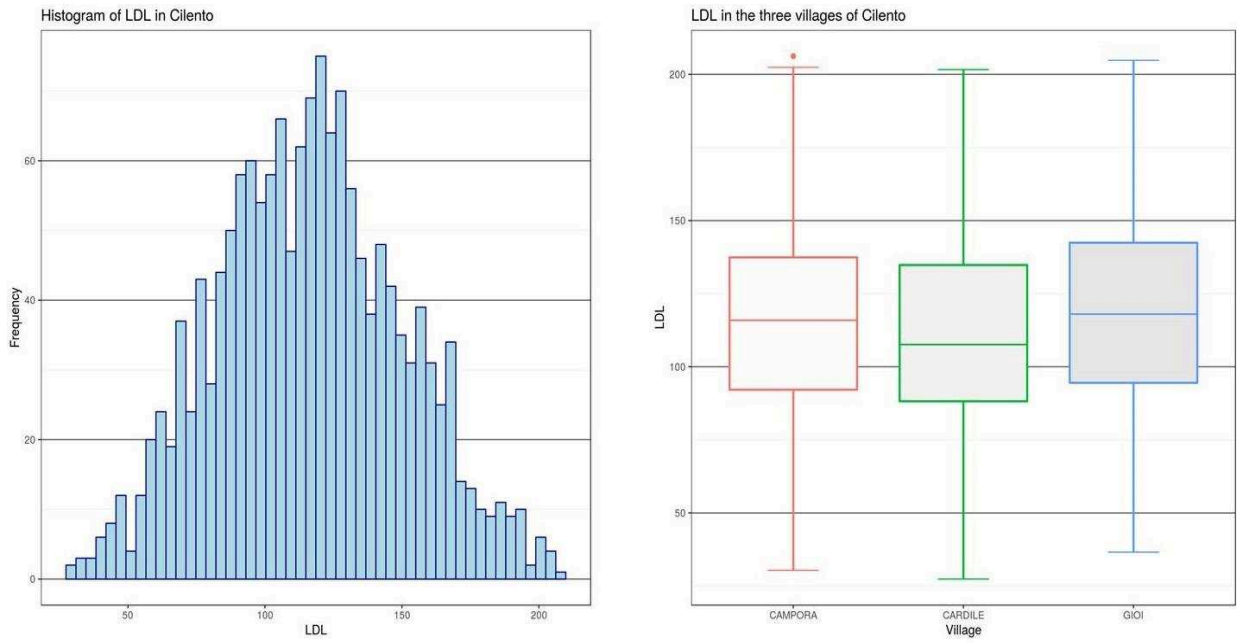


Figure 4.4.2a – The distribution of LDL in Cilento and in each village of Cilento.

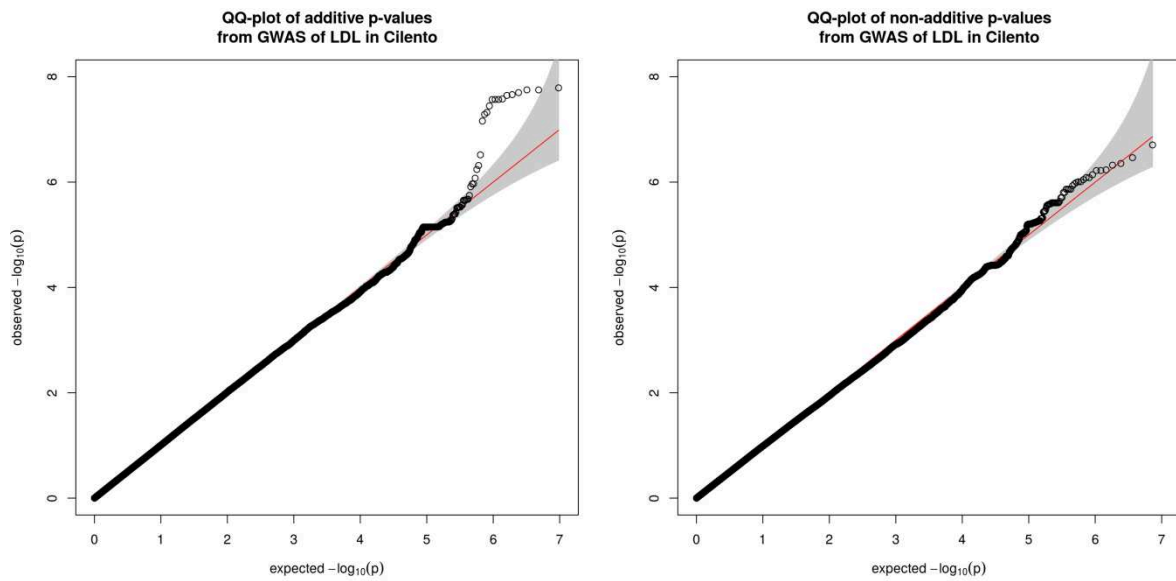


Figure 4.4.2b – QQ plots of the p-values from additive and non-additive GWASs of LDL in Cilento. The red line and grey zone describe the expected distribution of p-values under the null distribution.

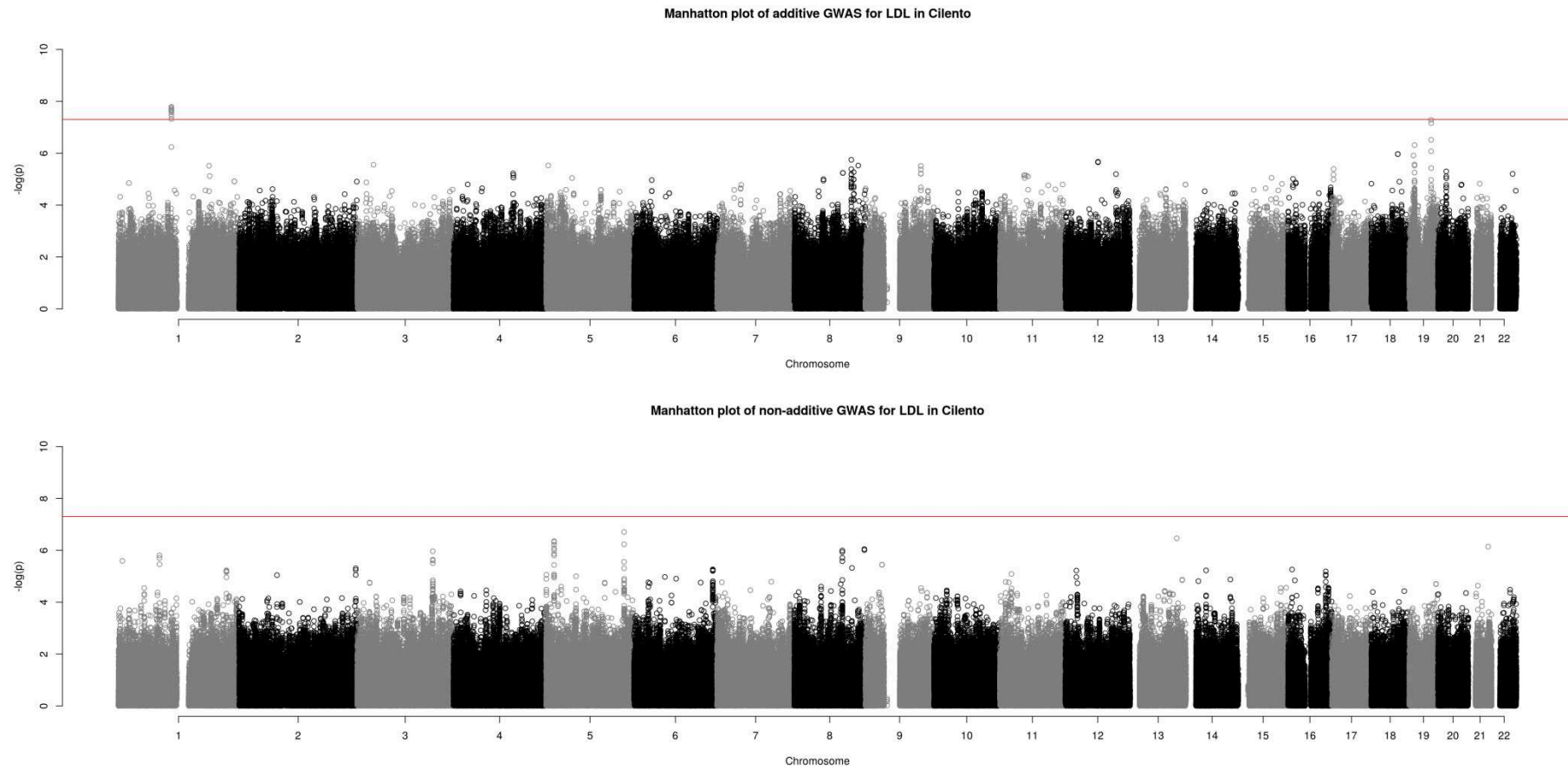


Figure 4.4.1c – Manhattan plots of p-values from *additive (top) and non-additive (bottom) GWASs of LDL in Cilento*. The p-values are plotted against their position on the genome, with the 22 autosomal chromosomes clearly visible. Horizontal red lines represent a common threshold for genome wide significance of  $5 \times 10^{-8}$ .

First we will discuss the results of the additive GWAS. In Table 4.2.2, the 30 variants with the lowest p-values on the additive GWAS are presented. The table notes also the gene for each variant where appropriate, minor allele frequencies, and ‘info’ scores from the two imputation runs. We also detail the ‘source’ of each variant, i.e. whether it is a variant from the Array data, from the WES panel, from both Array and WES data, or from the 1000G.

<b>Table 4.4.2a</b>	<b>CHR</b>	<b>MAF Cilento</b>	<b>MAF 1000G ALL</b>	<b>MAF 1000G EUR</b>	<b>info 370K</b>	<b>info OMNI</b>	<b>Data Origin</b>	<b>p-value (additive)</b>	<b>Gene</b>
<b>Variant id</b>									
rs583104	1	0.22	0.36	0.22	0.97	1.00	1000G	$1.6 \times 10^{-8}$	<i><b>SORT1</b></i>
rs7528419	1	0.22	0.20	0.21	1.00	-	OMNI	$1.8 \times 10^{-8}$	<i><b>SORT1</b></i>
rs12740374	1	0.22	0.20	0.21	1.00	1.00	1000G	$1.8 \times 10^{-8}$	<i><b>SORT1</b></i>
rs629301	1	0.22	0.24	0.21	1.00	1.00	1000G	$2.0 \times 10^{-8}$	<i><b>SORT1</b></i>
rs1277930	1	0.22	0.36	0.22	0.96	1.00	1000G	$2.2 \times 10^{-8}$	<i><b>SORT1</b></i>
rs56246620	1	0.18	0.12	0.18	0.99	1.00	1000G	$2.3 \times 10^{-8}$	<i><b>SORT1</b></i>
rs646776	1	0.22	0.24	0.21	-	-	370K, OMNI	$2.6 \times 10^{-8}$	<i><b>SORT1</b></i>
rs660240	1	0.21	0.23	0.20	1.00	-	OMNI	$2.7 \times 10^{-8}$	<i><b>SORT1</b></i>
rs57677983	1	0.21	0.25	0.20	1.00	1.00	1000G	$2.7 \times 10^{-8}$	<i><b>SORT1</b></i>
rs599839	1	0.22	0.36	0.22	0.96	-	OMNI	$2.7 \times 10^{-8}$	<i><b>SORT1</b></i>
rs4970836	1	0.21	0.35	0.22	0.97	1.00	1000G	$3.6 \times 10^{-8}$	<i><b>SORT1</b></i>
rs602633	1	0.21	0.35	0.21	0.97	1.00	1000G	$4.8 \times 10^{-8}$	<i><b>SORT1</b></i>
rs1065853	19	0.06	0.08	0.06	0.99	0.99	1000G	$5.2 \times 10^{-8}$	<i><b>APOE</b></i>
rs7412	19	0.06	0.08	0.06	0.97	-	OMNI, WES	$6.9 \times 10^{-8}$	<i><b>APOE</b></i>
rs7254892	19	0.04	0.08	0.03	0.96	0.93	1000G	$3.0 \times 10^{-7}$	<i><b>APOE</b></i>
rs111867267	19	0.40	0.28	0.44	0.99	0.99	1000G	$4.8 \times 10^{-7}$	<i><b>LDLR</b></i>
rs4970834	1	0.18	0.17	0.19	0.98	-	OMNI, WES	$5.6 \times 10^{-7}$	<i><b>SORT1</b></i>
rs61679753	19	0.04	0.07	0.03	0.98	0.99	1000G	$8.4 \times 10^{-7}$	<i><b>APOE</b></i>
18:54736162:T:C	18	0.001	0.0002	0	0.96	0.96	1000G	$1.1 \times 10^{-6}$	<i><b>LINC-ROR</b></i>
18:54714904:G:A	18	0.001	0.0006	0.001	0.95	0.89	1000G	$1.1 \times 10^{-6}$	<i><b>LINC-ROR</b></i>
rs117535884	19	0.02	0.007	0.02	0.91	0.87	1000G	$1.2 \times 10^{-6}$	<i><b>MUC16</b></i>
rs6469644	8	0.34	0.22	0.40	0.95	0.97	1000G	$1.8 \times 10^{-6}$	-
12:67605169:G:A	12	0.004	0.0006	0	0.78	0.83	1000G	$2.1 \times 10^{-6}$	-
rs370812494	12	0.004	0.002	0	0.78	0.84	1000G	$2.1 \times 10^{-6}$	-
12:67609274:T:C	12	0.004	0.001	0	0.78	0.81	1000G	$2.2 \times 10^{-6}$	-
12:67457528:C:A	12	0.004	0.0002	0	0.79	0.80	1000G	$2.2 \times 10^{-6}$	-
12:67679011:TC:T	12	0.004	0.001	0	0.91	0.88	1000G	$2.2 \times 10^{-6}$	-
rs72658867	19	0.02	0.003	0.01	0.98	0.85	WES	$2.7 \times 10^{-6}$	<i><b>LDLR</b></i>
rs202056691	3	0.04	0.01	0.02	0.85	0.92	1000G	$2.8 \times 10^{-6}$	<i><b>CLASP2</b></i>
rs72643748	5	0.05	0.20	0.05	0.82	0.81	1000G	$3.0 \times 10^{-6}$	-

*Table 4.4.2a – Lowest 30 p-values from additive GWAS of LDL in Cilento. The minor allele frequencies of each variant in Cilento, in all 1000G populations, and in all European 1000G populations are given. As imputation was performed separately on the two genotyping arrays in Cilento (370K and OMNI), we give both of the imputation quality scores in the table. The Data Origin column describes whether variants originated in either of the genotyping arrays, the WES data of Cilento, or from the 1000G. Where possible, if a variant lies within a gene (or very close to a gene) this information is given. Rows in the table are marked as grey if there is a known or conceivable relationship between the gene and the trait LDL; orange otherwise.*

The results from the additive GWAS show that the null-model was largely followed without any clear and apparent presence of many false-positive results. By looking through the lowest p-values, some familiar genes known to be associated with LDL have been identified. SORT1, APOE, and LDLR have been detected, with a few variants on SORT1 actually passing genome-wide significance. Beyond these genes, and beyond the first 30 lowest p-values, variants do not resonate with any other previous results. The association with the gene SORT1 was the strongest, and interestingly the top variant rs583104 on chromosome 1 was also highlighted in Sanna, et al.<sup>292</sup> Note that this study was on the Sardinian population, also an Italian isolate, and involved an interesting strategy where 256 individuals with extreme values for LDL were whole-exome sequenced. Interestingly, this variant was also highlighted in a study on an African American cohort<sup>295</sup>.

We chose a commonly used threshold for genome-wide significance of  $5 \times 10^{-8}$ . This value arose from estimates of the effective number of independent tests that are carried out during a GWAS. It has been estimated that approximately 1 million independent regions of the genome exist in European populations<sup>296</sup>. This leads to the adjusted p-value of  $5 \times 10^{-8}$ . However, in African populations, due to lower LD across the genome, the corresponding estimate is in fact in the region of 2 million independent regions<sup>296</sup>. Hence, different thresholds for significance should be used in different populations<sup>297</sup>. In isolated populations, we know that LD-blocks can be larger than in general populations suggesting, potentially, that lower thresholds should be used to avoid type-2 errors. As we have not carried out the required analysis to estimate a population specific threshold for Cilento, we stuck with  $5 \times 10^{-8}$ .

To test the non-additive component, it is essential that all three genotypes are present in the sample; hence this test was only possible on 7,291,750 variants. For the GWAS of non-additive components, the null-hypothesis would seem to hold across the genome; there are no significant results. In table 4.2.2b, we give the variants with the lowest 20 p-values from this GWAS. An exploration of the variants with the lowest p-values (beyond the first 20) gave one possible loci we

felt was worth further exploration: the two variants in Table 4.2.2b that stand out are those in the gene GPIHPB1.

<b>Table 4.2.2b</b>	<b>CHR</b>	<b>MAF</b>	<b>MAF</b>	<b>MAF</b>	<b>info</b>	<b>info</b>	<b>Data Origin</b>	<b>p-value</b>	<b>Gene</b>
<i>Variant id</i>		<i>Sample</i>	<i>1000G</i>	<i>1000G</i>	<i>370K</i>	<i>OMNI</i>		<i>(non-additive)</i>	
			<i>ALL</i>	<i>EUR</i>					
rs143444879	5	0.11	0.13	0.13	0.97	0.99	1000G	$2.0 \times 10^{-7}$	-
rs1751064	13	0.22	0.28	0.28	0.98	0.91	1000G	$3.4 \times 10^{-7}$	<b>ABCC4</b>
rs150565162	5	0.26	0.25	0.33	1.00	0.99	1000G	$4.3 \times 10^{-7}$	-
rs163360	5	0.26	0.25	0.39	1.00	0.99	1000G	$4.8 \times 10^{-7}$	-
rs149362603	5	0.12	0.13	0.06	0.94	0.97	1000G	$5.9 \times 10^{-7}$	-
rs163357	5	0.26	0.25	0.33	1.00	0.99	1000G	$6.0 \times 10^{-7}$	-
rs163361	5	0.26	0.25	0.33	1.00	0.99	1000G	$6.0 \times 10^{-7}$	-
rs59728366	21	0.40	0.32	0.41	0.98	0.94	1000G	$7.2 \times 10^{-7}$	<b>B3GALT5</b>
rs2596387	5	0.26	0.25	0.33	0.99	0.99	1000G	$8.2 \times 10^{-7}$	-
rs2596388	5	0.26	0.25	0.22	0.99	0.99	1000G	$8.2 \times 10^{-7}$	-
rs112271883	8	0.04	0.06	0.02	0.84	0.84	1000G	$9.0 \times 10^{-7}$	<b>GPIHBP1</b>
rs138876170	8	0.04	0.06	0.02	0.82	0.84	1000G	$9.8 \times 10^{-7}$	<b>GPIHBP1</b>
rs154790	5	0.26	0.25	0.34	-	1.00	370K	$9.8 \times 10^{-7}$	-
rs35725707	8	0.28	0.39	0.23	0.98	0.97	1000G	$1.0 \times 10^{-6}$	<b>MATN2</b>
rs439765	3	0.32	0.26	0.21	-	1.00	370K	$1.1 \times 10^{-6}$	-
rs2513837	8	0.26	0.29	0.18	1.00	0.99	1000G	$1.2 \times 10^{-6}$	<b>MATN2</b>
rs73324	5	0.26	0.25	0.33	1.00	1.00	1000G	$1.4 \times 10^{-6}$	-
rs696897	5	0.26	0.25	0.33	1.00	1.00	1000G	$1.4 \times 10^{-6}$	-
rs258909	5	0.26	0.25	0.33	1.00	1.00	1000G	$1.4 \times 10^{-6}$	-
rs460339	5	0.26	0.25	0.33	1.00	1.00	1000G	$1.4 \times 10^{-6}$	-

*Table 4.4.2b – Lowest 20 p-values from non-additive GWAS of LDL in Cilento. See Table 4.2.2a for a full explanation of the columns. Rows in the table are marked as grey if there is a known or conceivable relationship between the gene and the trait LDL; orange otherwise.*

The gene GPIHBP1, near the bottom end of chromosome 8 could conceivably be relevant for the distribution of LDL. Variants in this gene have been associated with various lipid measurements and hypertriglyceridemic phenotypes<sup>298-303</sup>. Indeed, mutations in this gene have already been found, under recessive models, associated with Hypertriglyceridemia<sup>304</sup>. Knowing the connections between this gene and lipid related traits, including through recessive models, is suggestive that this signal from the non-additive GWAS could have some foundation. The two variants rs112271883 and rs138876170 were imputed from the 1000G; there were however a small group (< 10) of nearby variants coming from the WES local reference panel suggesting that the imputation of this region may have benefitted from the WES SSP (See Section 3.3.5). In Figure 4.4.2d, the distribution of LDL against the different genotypes of these two variants is given, including those individuals whose genotypes were considered as missing during the GWAS.

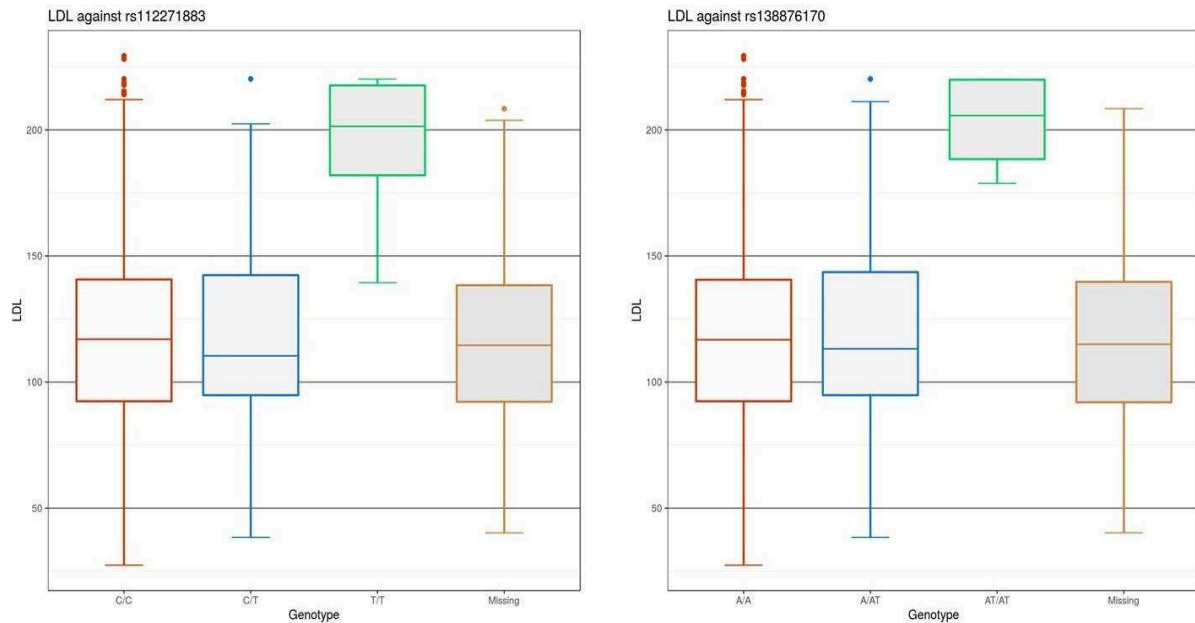
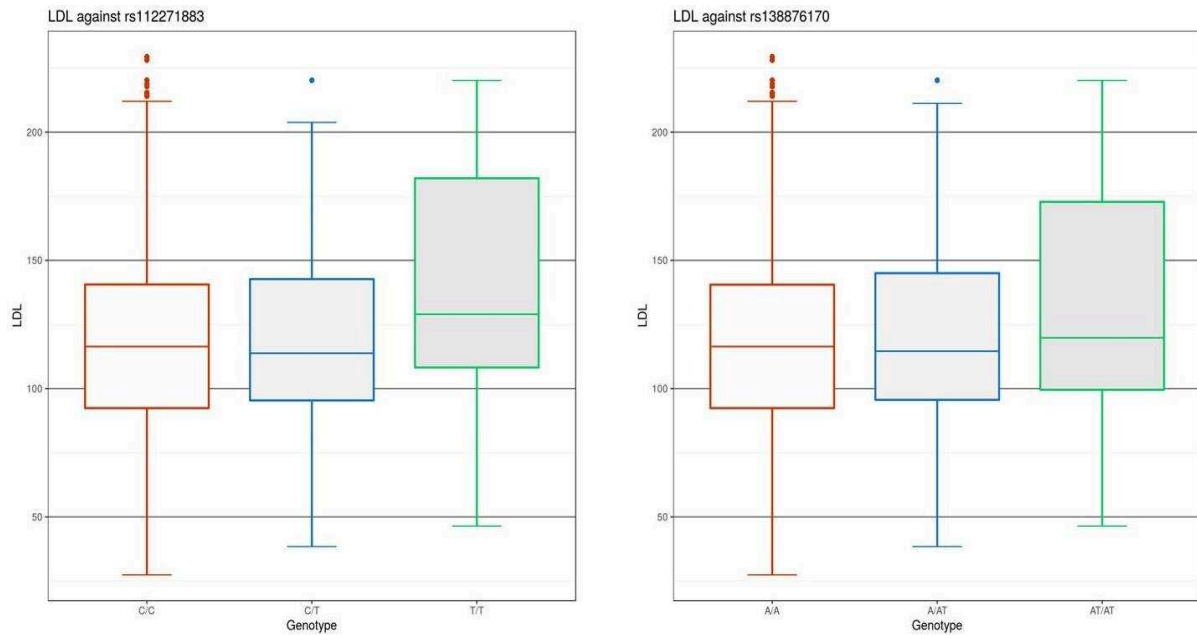


Figure 4.4.2d - The distributions of LDL depending on the genotypes at two positions of interest in GPIHPB1 that stood out from the results on the non-additive GWAS. Left panel shows LDL against rs112271883, for individuals with genotypes C/C, C/T, T/T, or 'Missing'. The right panel show the same for rs138876170 where individuals had possible genotype A/A, A/AT, AT/AT, or 'Missing'. There are 100 and 103 individuals with missing data for these positions respectively.

The evidence from these two positions may not be overwhelmingly convincing, importantly there are only a few instances of the rare genotype. For rs112271883, the major allele is C and the minor allele is T and only six individuals have the genotype T/T. Variant rs138876170 is an insertion, with the major allele being a single A nucleotide and the minor allele involving an inserted T nucleotide. Only four individuals had the rare genotype AT/AT. These four individuals all also had the genotype T/T on rs112271883 and two of these are also siblings. The 'info' scores for the two variants are not particularly high (but they are not common variants so the info is hard to interpret). Crucially, there are many genotypes that are missing due to the decision to keep only confidently imputed variants when hard-calling. To gain a better resolution of this genomic region, a stretch (50% of chromosome 8) of this hard-called imputed data was returned to SHAPEIT2 for re-phasing. This would fill in the missing genotypes for these two variants. The logic being that while IMPUTE2 could not confidently impute some genotypes in this region, SHAPEIT2 could perhaps give better results as will make inference from within the whole Cilento sample. We have already established in Chapter 3 that this phasing software is highly accurate in isolated populations; but that imputation (in particular when



using a WES SSP) was not always guaranteed to enjoy quite the same levels of success. The imputation concordance rate of SHAPEIT2 for sporadic missing genotypes was estimated as having 98% accuracy on our simulated data. Having completed this re-phasing, new boxplots were made which now include genotypes of all individuals (Figure 4.4.2e).

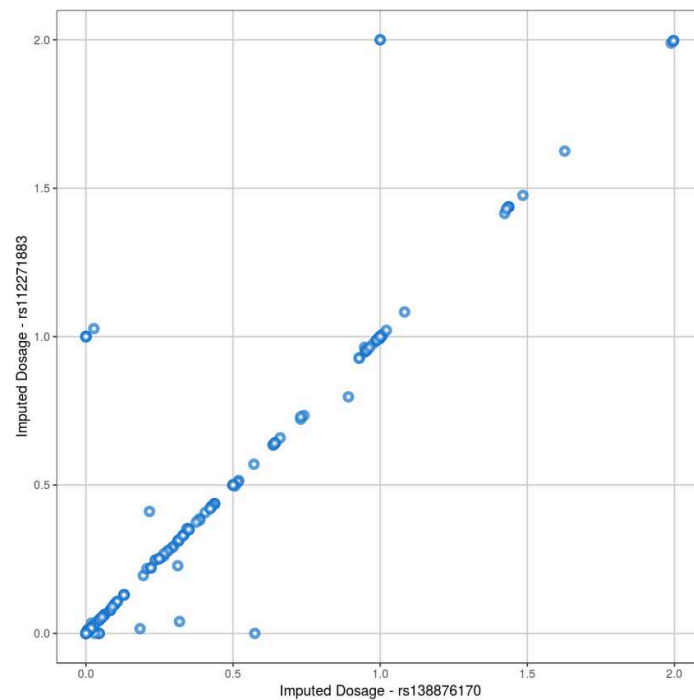


*Figure 4.4.2e - The distributions of LDL depending on the genotypes at two positions of interest in GPIHPB1. Here, missing genotypes have been ‘called’ using SHAPEIT2.*

In Figure 4.4.2e, there are now 16 individuals with genotype T/T for variant rs112271883 and there are 14 individuals with genotype AT/AT for variant rs138876170. From the boxplots, it became clear that the signal from these variants had been diluted. What is more, we could see that these additional individuals with rare genotypes brought into the analysis by SHAPEIT2 were estimated as carrying (locally) the exact same pairs of haplotypes as the initial group of individuals which originally stood out. It is therefore almost certainly not the case that the small group of individuals that stand out in Figure 4.4.2d are carrying two pairs of a very specific haplotype.

In Figure 4.4.2f, the imputed dosages of rs112271883 and rs138876170 are plotted against each other, illustrating why many individuals were excluded from the first GWAS. We can see the group of individuals with a dosage roughly equal to  $\sim 1.5$  that were excluded from the initial analysis but were brought into the second analysis by SHAPEIT2 with imputed dosages of 2. This illustrates how if we

had been able to implement our non-additive GWAS on dosage genotypes directly, we probably would never have noticed the gene *GPIHPB1* and perhaps not have been tempted by erroneous conclusions. In Section 4.5, we will give the definition of a non-additive component for dosage data that would be appropriate for such analyses in the future.



*Figure 4.4.2f – Plot of imputed dosages of rs112271883 and rs138876170 in Cilento. The correlation between these variants is high and many dosages show a high level of uncertainty.*

At this point our conclusion is that the results on *GPIHPB1* from the non-additive GWAS probably stood out by chance alone and do not represent a link between the gene and the trait in Cilento. This is not to say that a deeper investigation of this region could not produce a working hypothesis but for the moment our plan will be to turn to alternative analysis approaches for finding sources of the non-additive component for LDL in Cilento. One approach to explore non-additive effects that we would like to test for association between homozygosity and LDL level in Cilento using the methods described in Abney, et al.<sup>305</sup>

## 4.5 Prospective: Estimating K and D with Uncertain Genotypes

Here, we describe estimators of matrices  $K$  and  $D$  from imputed data or from low depth sequencing data. This section will initially follow a very similar line as the derivation of additive and non-additive components given in Section 4.1. We will first give the orthogonal decomposition of additive and non-additive effects without assuming Hardy-Weinberg proportions in the population, before using this result to give similar decompositions in the presence of genotype uncertainty.

### 1) *Departure from Hardy-Weinberg*

Given a genetic value taking values of  $u_{j0}$ ,  $u_{j1}$ , and  $u_{j2}$  as before, we will now define its expectation using genotype probabilities that do not necessarily follow Hardy-Weinberg proportions:

$$E_{\varepsilon}[g_j] = \bar{p}_{jAA}u_{j0} + \bar{p}_{jAa}u_{j1} + \bar{p}_{jaa}u_{j2}$$

Where  $\bar{p}_{jG}$  refers simply to the empirical observed probability of genotype  $G$  (at the  $j^{\text{th}}$  variant) based simply on the observed frequency of genotype  $G$  in the population. The subscript  $j$  will be dropped off  $\bar{p}_{jG}$  to ease readability. The subscript  $\varepsilon$  is added throughout to distinguish this section from the calculations in Section 4.1 as this differs from previous calculations which used the allelic frequencies  $p$  and  $q$ . A discussion of this particular approach can be found in Vitezica, et al.<sup>306</sup>

The inner product, corresponding to equation 4.1a, can now be defined as follows:

$$\langle g_j^1, g_j^2 \rangle_{\varepsilon} = E_{\varepsilon}[g_j^1 g_j^2] = \bar{p}_{AA}u_{j0}^1 u_{j0}^2 + \bar{p}_{Aa}u_{j1}^1 u_{j1}^2 + \bar{p}_{aa}u_{j2}^1 u_{j2}^2$$

Again we can then assume the following orthogonal decomposition:  $g_j = \mu_j X_1 + a_j X_j^{a,\varepsilon} + d_j X_j^{d,\varepsilon}$ .

Choosing  $X_j^{a,\varepsilon}$  to equal  $\alpha(X_{012} - \beta X_1)$  and specifying  $\langle X_j^{a,\varepsilon}, X_j^{a,\varepsilon} \rangle_{\varepsilon} = 1$  and  $\langle X_j^{a,\varepsilon}, X_1 \rangle_{\varepsilon} = 0$  gives:

$$X_j^{a,\varepsilon} = v_{\varepsilon}^{-\frac{1}{2}} [X_{012} - (\bar{p}_{Aa} + 2\bar{p}_{aa})X_1], \text{ where}$$

$$v_{\varepsilon} = \bar{p}_{Aa} + (4\bar{p}_{AA}\bar{p}_{aa} - \bar{p}_{Aa}^2).$$

On inspection, we see this corresponds closely to the classical additive component, as the term  $\bar{p}_{Aa} + 2\bar{p}_{aa}$  is equivalent to the expected minor allele count.  $v_{\varepsilon}$  represents the variance of the minor allele count; composed of the variance under Hardy-Weinberg proportions ( $\bar{p}_{Aa}$ ) and the change in the variance coming from any departure from Hardy-Weinberg proportions ( $4\bar{p}_{AA}\bar{p}_{aa} - \bar{p}_{Aa}^2$ ).

We again find  $X_j^{d,\varepsilon}$  by specifying  $\langle X_j^{d,\varepsilon}, X_j^{d,\varepsilon} \rangle_\varepsilon = 1$ ,  $\langle X_j^{d,\varepsilon}, X_1 \rangle_\varepsilon = 0$ , and  $\langle X_j^{d,\varepsilon}, X_j^{a,\varepsilon} \rangle_\varepsilon = 0$ , and we find the following:

$$X_j^{d,\varepsilon} = \left( \gamma_\varepsilon \sqrt{\frac{\bar{p}_{aa}}{\bar{p}_{AA}}}, -2\gamma_\varepsilon \sqrt{\frac{\bar{p}_{aa}\bar{p}_{AA}}{\bar{p}_{Aa}^2}}, \gamma_\varepsilon \sqrt{\frac{\bar{p}_{AA}}{\bar{p}_{aa}}} \right), \text{ where } \gamma_\varepsilon = \left( \bar{p}_{aa} + 4\frac{\bar{p}_{aa}\bar{p}_{AA}}{\bar{p}_{Aa}} + \bar{p}_{AA} \right)^{-\frac{1}{2}}.$$

## 2) Incorporating Genotype Uncertainty

We will now consider the case where there exists enough uncertainty about genotypes that we wish to retain this uncertainty in subsequent calculations. This might well be the case for either low depth sequencing data or for imputed data. The derivation will follow the same steps as above, and here, the subscript  $L$  is used. In this situation, for individual  $i$ 's data at the  $j$ th variant we have three probabilities for each potential genotype -  $L_{ij}^{AA}$ ,  $L_{ij}^{Aa}$ , and  $L_{ij}^{aa}$ .

If we consider again the genetic value  $g_j$ , this can now no longer depend on the now unobservable genotype. Instead, it can be defined by these three probabilities: under this setup we have the following continuous form for the genetic value:

$$g_j = L_j^{AA}u_{j0} + L_j^{Aa}u_{j1} + L_j^{aa}u_{j2}.$$

Across all individuals in a sample, the variation in the values of  $L_{ij}^{AA}$ ,  $L_{ij}^{Aa}$ , and  $L_{ij}^{aa}$  will however reflect the distribution of the unobserved genotypes. Hence, we can think of these three probabilities as random variables, with distributions depending on the unobserved genotypes  $G_j$ . Taking the expectation of the genetic value gives:

$$E_L[g_j] = E[L_j^{AA}]u_{j0} + E[L_j^{Aa}]u_{j1} + E[L_j^{aa}]u_{j2}.$$

In a similar way that allele frequencies would usually be empirically estimated from the ensemble of all individuals in the sample, we must do the same for each of the expected values of the genotype probabilities by estimating:

$$E[L_j^{AA}] = \frac{1}{N} \sum_i L_{ij}^{AA}, \quad E[L_j^{Aa}] = \frac{1}{N} \sum_i L_{ij}^{Aa}, \quad E[L_j^{aa}] = \frac{1}{N} \sum_i L_{ij}^{aa}$$

We denote these expectations as  $\bar{L}_{AA}$ ,  $\bar{L}_{Aa}$ , and  $\bar{L}_{aa}$ .

Therefore, we must now find orthogonal components using the following inner product:

$$\langle g_j^1, g_j^2 \rangle_L = E_L[g_j^1 g_j^2] = \bar{L}_{AA} u_{j0}^1 u_{j0}^2 + \bar{L}_{Aa} u_{j1}^1 u_{j1}^2 + \bar{L}_{aa} u_{j2}^1 u_{j2}^2$$

Now we have the exact same scenario as with the empirical example first given in this section, hence we have the following orthogonal components:

$$X_j^{a,L} = v_L^{-\frac{1}{2}} [X_{012} - (\bar{L}_{AA} + 2\bar{L}_{aa})X_1], \text{ where} \quad \text{Eq. 4.5a}$$

$$v_L = \bar{L}_{AA} + (4\bar{L}_{aa}\bar{L}_{AA} - \bar{L}_{AA}^2), \text{ and}$$

$$X_j^{d,L} = \left( \gamma_L \sqrt{\frac{\bar{L}_{aa}}{\bar{L}_{AA}}}, -2\gamma_L \sqrt{\frac{\bar{L}_{aa}\bar{L}_{AA}}{\bar{L}_{AA}^2}}, \gamma_L \sqrt{\frac{\bar{L}_{AA}}{\bar{L}_{aa}}} \right) \text{ where } \gamma_L = \left( \bar{L}_{aa} + 4\frac{\bar{L}_{aa}\bar{L}_{AA}}{\bar{L}_{AA}} + \bar{L}_{AA} \right)^{-\frac{1}{2}}.$$

### 3) Imputation Dosages

In a very similar way, we can derive a similar coding in the presence of dosage genotypes, which we will mark with the subscript  $I$  (for imputation). This is relevant to the circumstance where three posterior probabilities for each genotype from imputation are not available as they have been summarised in a single dosage. We now define our genetic value as:

$$g_{ij} = \begin{cases} (1 - w_{ij})u_{j0} + w_{ij}u_{j1}, & \text{if } w_{ij} < 1 \\ (2 - w_{ij})u_{j1} + (w_{ij} - 1)u_{j2}, & \text{if } w_{ij} \geq 1 \end{cases}$$

Where  $w_{ij}$  is the dosage of variant  $j$  for individual  $i$ . Here, there is a required assumption that a dosage between 0 and 1 suggests that the probability of the  $aa$  genotype is zero. Similarly, a dosage between 1 and 2 is assumed to mean that the probability of the  $AA$  genotype is zero. Our inner product becomes:

$$\langle g_j^1, g_j^2 \rangle_I = E_I[g_j^1 g_j^2] = \bar{W}_{AA} u_{j0}^1 u_{j0}^2 + \bar{W}_{Aa} u_{j1}^1 u_{j1}^2 + \bar{W}_{aa} u_{j2}^1 u_{j2}^2, \text{ where}$$

$$\bar{W}_{AA} = \frac{1}{N} \sum_{w_{ij} < 1} (1 - w_{ij}), \bar{W}_{Aa} = \frac{1}{N} \sum_{w_{ij} < 1} w_{ij} + \frac{1}{N} \sum_{w_{ij} \geq 1} (2 - w_{ij}), \bar{W}_{aa} = \frac{1}{N} \sum_{w_{ij} \geq 1} (w_{ij} - 1).$$

In an entirely equivalent manner to the derivations of  $X_j^{a,L}$  and  $X_j^{d,L}$ , we arrive at the following additive and non-additive orthogonal components:

$$X_j^{a,I} = v_I^{-\frac{1}{2}} [X_{012} - (\bar{W}_{AA} + 2\bar{W}_{aa})X_1] \quad \text{Eq. 4.5b}$$

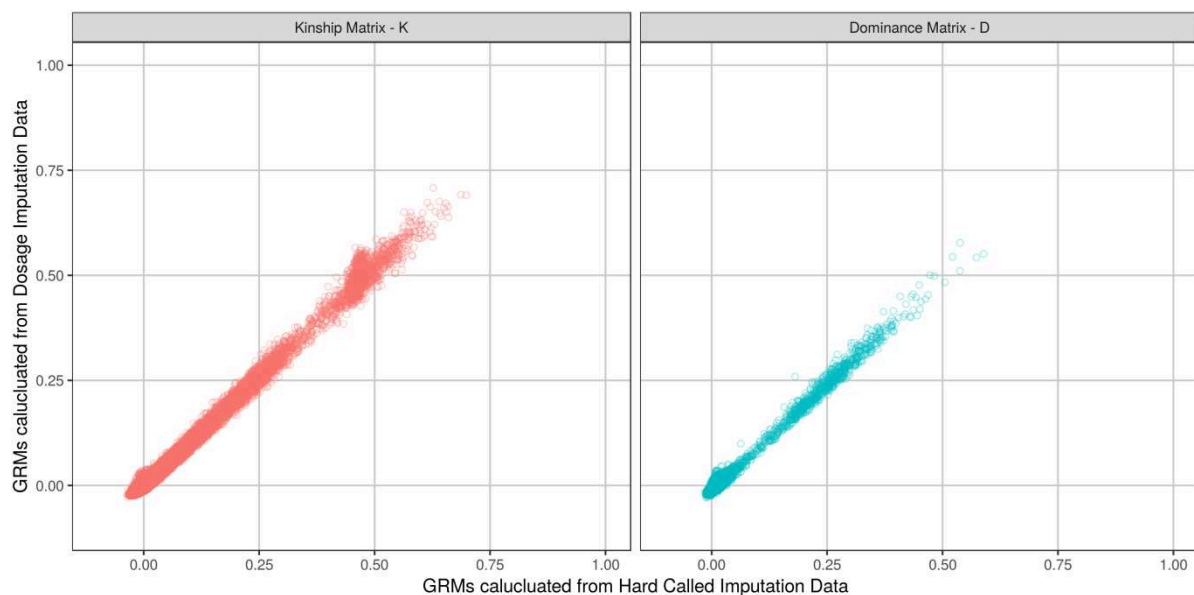
Here:  $v_I = \bar{W}_{AA} + (4\bar{W}_{aa}\bar{W}_{AA} - \bar{W}_{AA}^2)$ .

$$X_j^{d,I} = \left( \gamma_I \sqrt{\frac{\bar{W}_{aa}}{\bar{W}_{AA}}}, -2\gamma_I \sqrt{\frac{\bar{W}_{aa}\bar{W}_{AA}}{\bar{W}_{AA}^2}}, \gamma_I \sqrt{\frac{\bar{W}_{AA}}{\bar{W}_{aa}}} \right) \text{ where } \gamma_I = \left( \bar{W}_{aa} + 4\frac{\bar{W}_{aa}\bar{W}_{AA}}{\bar{W}_{AA}} + \bar{W}_{AA} \right)^{-\frac{1}{2}}.$$

Hence, we can also find orthogonal components for imputed dosage type data.

These new non-additive components  $X_j^{d,L}$  and  $X_j^{d,I}$  can be used for performing a GWAS on non-additive effects directly from posterior probabilities or from dosage data, respectively. Indeed, this is what we hope to implement for non-additive GWAS as described in Section 4.4.2.

We tested these ideas when estimating matrices  $K$  and  $D$  on dosage data that we created by performing imputation on to the simulated Array data of Cilento. We found that in this case, the estimation of the matrices  $K$  and  $D$  were almost identical to those estimated from ‘hard called’ imputed genotypes (Figure 4.5a).



*Figure 4.5a – Comparison of off diagonal elements of matrices  $K$  and  $D$  from GRMs calculated either using dosages explicitly with the components  $X_{ij}^{a,I}$  and  $X_{ij}^{d,I}$  described above, or when using hard-called genotypes and hence classical components  $X_{ij}^a$  and  $X_{ij}^d$ . Data from Chromosome 1 on the HapGen+Pedigree simulation.*

We see from Figure 4.5a that retaining the uncertainty from imputation, in this case, does not lead to large differences in the GRMs. This is because, as we have previously demonstrated, we do not need an overwhelming detailed amount of data in order to estimate a GRM in an isolate such as Cilento; in fact, we showed in our published work that simply using the Array genotypes will probably suffice.

Estimating GRMs from imputed data will be more relevant when the quality of the original data was low. A possible scenario could be in the presence of low depth sequencing data. In such data, each position will on average only be read by the genotyping machinery a very limited number of times; this average read depth could be as low as 0.1 or 0.5. For such data, the calling of the genotypes is problematic as the set of observed reads for each genotype will often be very small and there will be high levels of missing data. In general, for WGS data, three genotype likelihoods are calculated based on the set of observed reads, using for example the methods described in GATK<sup>307</sup>. These likelihoods will then indicate one genotype that has most supporting evidence from the reads and it will be clear as to which genotype should be called. If there is insufficient certainty, then genotypes may be called as missing. For example, the following sets of reads might be observed for four individuals at a given position:

$$(1) - \{A, A, A, A, A, A, A, A, A, A, A, A, A, A, A, A\},$$

$$(2) - \{A, a, A, A, A, a, a, a, A, a, A, A, a, A, a, a, A, a, a\},$$

$$(3) - \{a, a, a, a, a, a, a, a, A, a, a, a, a, a, a, a, a, a, a, a, a, a\},$$

$$(4) - \{A, A, A\}.$$

In the first example, the set of reads clearly suggests the genotype  $AA$ . The second set of reads suggests a heterozygous genotype  $Aa$ . The third set of reads suggests the genotype  $aa$ , even though there is one instance of a read of the major allele  $A$  (it is more likely that this one read is an error than the site is truly heterozygous and all but one reads came from one single haplotype). The fourth set of reads is perhaps too small to draw a meaningful conclusion.

When there are few reads, it is clear that it becomes impossible to designate a genotype with any certainty. However, genotype likelihoods,  $L^{AA}$ ,  $L^{Aa}$ , and  $L^{aa}$  can still carry useful information. The method SEEKIN<sup>308</sup> was designed to estimate kinship from low depth sequencing data, via an intermediate imputation step. Recent versions of the software BEAGLE as well as recent software

GeneImp<sup>309</sup> are compatible with genotype likelihood data and are used both to impute new sites as well as to improve the calling of low-coverage genotypes via the information stored in large panels of reference haplotypes. SEEKIN estimates kinship from such imputed data. The SEEKIN estimator for kinship for a single imputed SNP is similar to the point wise estimates that are combined to calculate a GRM. The SEEKIN estimator is given below:

$$2\bar{\varphi}_{ik} = \frac{(w_i - 2\bar{q})(w_k - 2\bar{q})}{2\bar{q}(1 - \bar{q})(r^2)^2}$$

Here,  $w_i$  and  $w_k$  are the imputed dosages of the two individuals  $i$  and  $k$ ;  $\bar{q}$  is the estimate of the MAF of the SNP; and  $r^2$  is the correlation between true genotypes and imputed genotypes. This will not be known so an estimator in the form of an imputation quality score,  $\hat{r}^2$ , will be used.

From equation 4.5b, using the derivation for dosages that we have presented, our point-wise estimator would be the following:

$$2\bar{\varphi}_{ik} = \frac{(w_i - \bar{W}_{Aa} + 2\bar{W}_{aa})(w_k - \bar{W}_{Aa} + 2\bar{W}_{aa})}{v_I}$$

Where  $\bar{w}$  is the mean dosage across all individuals and is an estimator of  $2q$ .  $v_I$  is the empirical variance of the dosages of the SNP across all individuals. In addition to SEEKIN, there exist recent methods for estimating kinship and IBD-sharing directly from genotype likelihoods<sup>310-313</sup>. The estimators that we derived allow the estimation of the non-additive component, but it will be instructive to compare our additive estimators to existing methods in the future.

We were also interested in the idea of working directly with genotype likelihoods. Working with imputed data or genotype likelihoods would appear to be equivalent as with both types of data we have three genotype probabilities rather than a single genotype. However, there is one key difference: with imputed data, there is prior information coming from the reference panel. Thus, if there is little information from an HMM (or other model), then the three posterior genotype probabilities will likely reflect the allele frequency in the reference panel and should roughly equal



$p^2$ ,  $2pq$ , and  $q^2$ . Genotype likelihoods, however, are calculated without such prior information, and whilst calling pipelines will leverage information across multiple samples to describe the quality of the mapping and calling of each variant, the individual level genotype likelihoods are based only on individual read data. Hence, genotype likelihoods of  $(1/3, 1/3, 1/3)$  and imputed posterior probabilities of  $(1/3, 1/3, 1/3)$  describe quite different things. If the variant has a low MAF, posterior imputed probabilities of  $(1/3, 1/3, 1/3)$  could, hypothetically, be quite informative; suggesting that this individual may well carry one or even two rare alleles. Genotype likelihoods of  $(1/3, 1/3, 1/3)$  however represent a complete lack of information and could even be treated as a missing value.

We have begun testing out method for directly estimating relatedness matrices  $K$  and  $D$  from low-coverage data. We took our simulated data using the Cilento genealogies for chromosome 10 and for each individual level genotype, we used a simple additional layer of simulation to create three genotype likelihoods. This process was also used in our first study (see the description of Error Models in the Supplementary Methods of Annex A). It is based on the simulation method described in Kim, et al.<sup>314</sup> Explicitly, for every position on our simulated chromosome 10, we assigned a value of mean read depth derived from the sequence of observed mean read depths from the actual WGS data of chromosome 10 from 19 individuals in Cilento. This sequence of average read depths was used in order so that we could replicate any potential patterns of low or high depth in different regions of the chromosome. Then for every position, we drew random counts of major and minor reads based on Poisson distributions with parameters depending on the true simulated genotype generated from gene-dropping. Our simulation allows error reads to occur with a rate of 1%. The simulation process is described in Box 4.5. The mean depth in the WGS panel of Cilento on chromosome 10 is roughly 54 reads. To generate low depth data, we simply scaled down the observed sequence of average depths from Cilento accordingly.

**Box 4.5 - Simulating Genotype Likelihoods**

Given a simulated genotype  $g \in \{0,1,2\}$ :

- Draw a value for the read depth from the observed mean depths in Cilento, call this  $\delta$ .
- Scale this depth down in order to simulate a low-depth scenario, giving a new depth of  $\delta'$ .
- Draw counts of major alleles and minor alleles from two Poisson distributions:  $r_A \sim \text{Poisson}(\rho_A \delta')$  and  $r_a \sim \text{Poisson}(\rho_a \delta')$  where:

$$\rho_A = \begin{cases} 0.99, & \text{if } g = 0 \\ 0.50, & \text{if } g = 1 \\ 0.01, & \text{if } g = 2 \end{cases} \quad \text{and} \quad \rho_a = \begin{cases} 0.01, & \text{if } g = 0 \\ 0.50, & \text{if } g = 1 \\ 0.99, & \text{if } g = 2 \end{cases}$$

- Calculate genotype likelihoods  $L^{AA}$ ,  $L^{Aa}$ , and  $L^{aa}$  as follows:

First compute:

$$\lambda^{AA} = 0.01^{r_a} \times 0.99^{r_A}$$

$$\lambda^{Aa} = 0.5^{r_a + r_A}$$

$$\lambda^{aa} = 0.99^{r_a} \times 0.01^{r_A}$$

And then:

$$L^{AA} = \lambda^{AA} / (\lambda^{AA} + \lambda^{Aa} + \lambda^{aa})$$

$$L^{Aa} = \lambda^{Aa} / (\lambda^{AA} + \lambda^{Aa} + \lambda^{aa})$$

$$L^{aa} = \lambda^{aa} / (\lambda^{AA} + \lambda^{Aa} + \lambda^{aa})$$

*Box 4.5 – Simulation of genotype likelihoods on top of previously simulated genotypes.*

Once we had generated sets of reads for every simulated position, genotype likelihoods were estimated from binomial distribution function. This admittedly skips several steps from the GATK calling algorithm and we only have considered a very simple error simulation and model. In Figure 4.5b, we compare the off diagonal elements of matrices  $K$  and  $D$  that have either been calculated as GRMs from the simulated genotypes, or from the simulated genotype likelihoods. We varied the mean overall read depth of the chromosome. There was an immediately obvious bias in the estimates from genotype likelihoods. This bias was greatest when the global mean depth was lowest.

It is possible to demonstrate this bias by numerically calculating the expected value of the correlation between additive and non-additive components by forming a summation over possible Poisson draws in our simulation. To elaborate, for a parent-offspring pair; then the expected covariance of their classical additive components (as defined in Section 4.1) of a single variant is equal to  $\frac{1}{2}$ . However, when computing the equivalent expectation for a parent-offspring pair using our simulated genotype likelihoods and the revised additive components (equation 4.5a) our

evaluation will systematically fall below  $\frac{1}{2}$ . The bias is depending on read depth, MAF, and error model. This explains the patterns observed in Figure 4.5b. Whilst we have shown that orthogonal additive and non-additive variance components can be estimated directly from genotype likelihoods, these estimators are not equivalent to classical GRMs as they are not unbiased moment estimates of IBD-sharing matrices  $K$  and  $D$ . To pursue this line of investigation may require a more sophisticated simulation of low depth data and better appreciation of the bioinformatics tools used to produce genetic data, an aspect that we are so far yet to explore in detail.

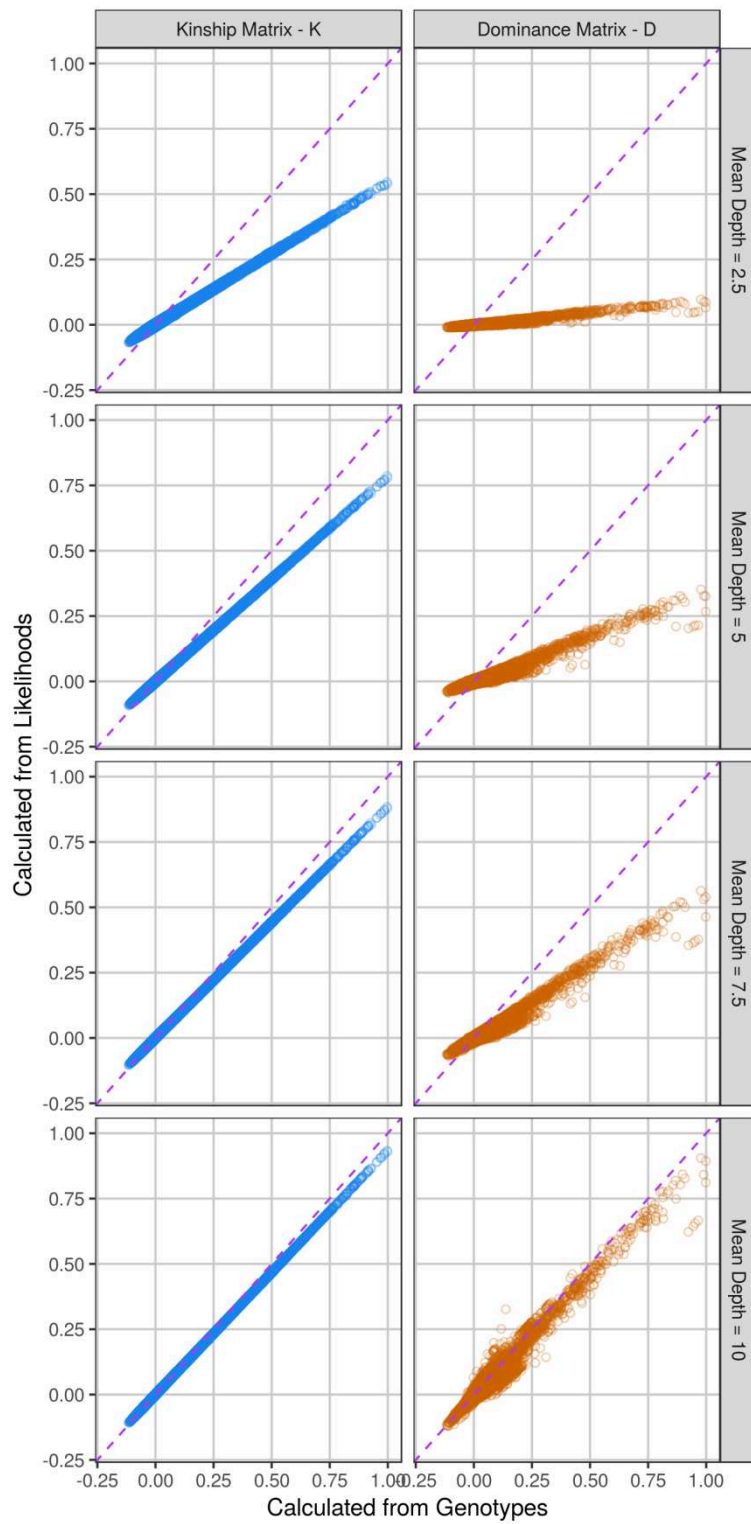


Figure 4.5b - Comparison of off diagonal elements of matrices  $K$  and  $D$  from GRMs calculated either using simulated genotypes or simulated genotype likelihoods. We varied the mean read depth on the chromosome, using values of 2.5, 5, 7.5, and 10.

## 4.6 Conclusions on Heritability

Our study of dominance components of heritability in isolated and outbred populations was very instructive for our own analyses of Cilento as well as answering questions about previously published results in the literature. By simulating data, with a similar structure to Cilento, we could gain insights into the results found on the true data of the three villages. For example, we were able to conclude that for BMI, the large estimate for  $h_a^2$  was probably driven by shared environmental effects. For LDL on the other hand, the large estimate for  $h_a^2$  may truly point to a non-zero component. We were able to show why disparate estimates for  $h_a^2$  have been presented in different populations suggesting that this has been a result of some combination of particular trait architectures, confounding with environmental factors, and insufficient sample sizes.

To fully pick apart the global architecture of a trait will require more individuals than we have in Cilento, in particular it would have been of great interest to apply the methods described in Young, et al. <sup>315</sup> for jointly exploring genetic effects and shared environmental effects. We have also proposed that studying multiple isolates might be a powerful approach, though there would be additional challenges coming from differences between isolates.

We also found success when using GRMs for estimating variance components in an isolated population. However, it is not completely clear to what extent these results were depended on our simulation set-up. Our simulated traits never deviated from the assumed polygenic model, our simulated effect sizes were drawn from normal distributions, and we therefore engendered a relationship between the size of genetic effects and the frequency of the variants; as described in section 4.1. A future direction for this particular analysis will be to continue our exploration of potential trait architectures by simulating traits that adhere less closely to the polygenic model and then see if GRMs continue to hold an advantage over IBD-based estimates.

This study has opened new avenues of investigation; and led us to complete a first GWAS using our new imputation dataset for the trait LDL. This demonstrated the potential of the new imputation dataset as we identified known genes under an additive model in this initial GWAS. As of yet, we were not able to pinpoint something specific that could help explain the estimates for  $h_a^2$  for LDL in Cilento. We have demonstrated the importance of retaining genotype uncertainty for this particular test in an anecdotal manner by looking at two positions on chromosome 8 that appeared to be possible interesting candidates for further investigation. We plan to continue to explore non-additive effects for the trait LDL in Cilento. Finally, we have also begun to explore methods for estimating additive and non-additive components from imputed data or even from low-depth sequencing data.

## Chapter 5: Discussion

The objective of this thesis has been to understand how the study of isolated populations can continue to shed light of the architecture of complex traits. Having presented numerous successful studies of isolated populations, the challenge is to predict the role of such populations in future projects. Though GWAS have been enormously successful, revolutionizing our understanding of the human genome, for many traits the research community's attention will soon turn towards more elaborate models, beyond looking at marginal effects of every genetic variant. Deeper biological understanding of the genome, driven at first by the multitude of GWAS results, will necessarily be incorporated into the future queries and models put forth by genetic epidemiologists. The innate properties of population isolates that have so far proved useful to gene-mapping in many cases should continue to benefit such future studies. The increased availability of whole-genome sequencing data will give much added richness to genetic data in isolates that will allow more detailed hypotheses to be tested. Isolated populations will continue to present good hunting grounds for finding rare and unusual types of genetic variation.

Sequencing studies in isolated populations have been quick to demonstrate their power in finding both novel and population specific levels of genetic variation. Gathering large samples of high quality sequencing data in isolated populations seems guaranteed to provide further interesting findings by exploring the specific characteristics of the genetics of each isolate with increasing resolution. To this end, though sequencing costs remain relatively high, genotype imputation methods will remain important tools. Hence, we undertook an in depth simulation study to investigate the best practices for performing such imputations in isolated populations.

From our simulations, we were able to both identify a pragmatic pipeline for the existing data in Cilento, as well as providing recommendations for the scientific community. This investigation was also very informative on the performance of IBD- and LD-based methods for phasing in isolated

populations. The algorithmic approach of SHAPEIT2 was particularly capable of modelling the mosaicism between individuals in an isolated population and we demonstrated that highly accurate phasing results should be anticipated in an isolate such as Cilento. Indeed, all phasing software tested proved to give very accurate phasing results. We were also able to reinforce previous findings regarding the benefits of using study specific reference panels for imputation by showing the strengths of a very small panel of local reference haplotypes. By comparing two different simulations of the bottleneck effect in the population, we also showed that imputation using only external public reference haplotypes can be unsatisfactory for imputation in a population with highly prevalent genetic drift. Whether or not our exact recommendations will last the test of time is perhaps questionable due to the continued advancements of available software; yet we feel that this detailed investigation into the interplay of isolate characteristics with phasing and imputation algorithms will continue to prove instructive.

This simulation study was also crucial in understanding how best to perform imputation using the WES panel in Cilento. The imputation study demonstrated how the use of a WGS study specific panel would provide highly accurate imputation. However, when using a study specific WES panel (as in Cilento), we showed that the improvement for imputation would largely be restricted to exonic regions. We were furthermore able to demonstrate an improvement in imputation for the real data of Cilento – a result that has been included in a recent research paper describing, for the first time, the sequencing data in Cilento<sup>174</sup>. Having completed imputation across the 22 autosomal chromosomes, and being confident of high quality imputation, our collaborators in Naples are in a good position to embark on analysis on the deep phenotyping data in Cilento. By performing a trial GWAS on LDL levels, partially in order to appraise the imputation, there is no evidence of any unforeseen problematic consequences from our imputation protocol. Ongoing projects that will involve this imputation dataset include looking at a genome-wide interaction study between VEGF serum levels and body-mass index and a GWAS on Bilirubin. Finally, by imputing up to the 1000G set



of positions, a commonly used and highly dense imputation panel, this will aid Cilento to continue to take part in meta-analyses with other cohorts.

The prevalence of IBD-sharing in isolated populations, including between distantly related individuals, facilitates exploration of genetic trait architecture beyond looking simply at marginal additive effects. After observing some interesting estimates for heritability components for commonly studied traits across different populations, as well as some contention in the literature, we decided to evaluate the estimation of non-additive variance components between isolated and outbred populations. Through simulation we were able to explain some of these inconsistencies in the literature; suggesting that the non-additive components of certain traits may well be overestimated in isolated populations due to shared environmental factors; but also possibly underestimated in outbred populations as they can be driven by rare variation. We also suggested that careful studies of isolated populations (potentially including data from multiple isolates) can detect non-additive components for complex traits; though the actual values of the estimates of said components will not be reliable.

Taking these ideas into the Cilento population, we presented an indication of the presence of a non-zero dominance component for the trait LDL in Cilento. This is a finding that we will continue to investigate; after initial findings from a non-additive GWAS for LDL suggested a possible link with the GPIHBP1 gene. A deeper investigation showed that this link to was not likely to be of any significance, though it could be an interesting region for continued study. One approach could be to directly sequence the gene including additional individuals from Cilento. This finding does in any case demonstrate the type of paradigm that may be required for future work on small population isolates in order to continue to compete with huge cohort studies: arising from both the imputation that was carried out in the Cilento, and by considering a specific hypothesis about trait architecture in the isolate. The non-additive GWAS also showed the great importance of retaining the uncertainty of imputed dosages. In order to fully test for non-additivity, we developed new methods for testing

for this component when using either genotype dosages or posterior imputation genotype probabilities. This led us to extend some of the methods used in our heritability study to data from low-depth sequencing data. The use of very low depth data in population isolates is a conceivable cost effective strategy (if again combined with imputation) for gathering WGS data. The methods we propose might also be suitable for off-target reads from WES data which could be explored in Cilento. WES sequencing is ‘targeted’ meaning that a list of known exons are directly sought by genotyping probes. This is opposed to whole-genome sequencing where genomic fragments are sequenced impartially. During targeted exome sequencing, it is however possible for the wrong fragments to be captured and sequenced during this process; and these reads can still be aligned to non-exonic regions<sup>316</sup>. These off-target reads are nevertheless infrequent, and so these regions end up with similar characteristics to low depth sequencing data<sup>309</sup>. We observed that the distribution of positions in WES data was problematic for imputation methods due to the gaps between exons. It may be the case that incorporating off-target reads can help various statistical methods when dealing with WES data. Hence, using methods that have been extended to incorporate the uncertainty of low depth data may be necessary and will be something that we will continue to explore.

In our investigation of heritability, we discussed the differences between heritability estimates from different study designs. Studies of related individuals are required to estimate the ‘true’ heritability rather than SNP heritability but it is well known that these estimates are subject to confounding due to shared environmental factors. Therefore, estimates from isolated populations are problematic and indeed we illustrated this in our second large simulation study. The arguments against studying phenotypic variance decomposition in isolated populations could however be spun in favour of such studies. Recently proposed methods by Young, et al.<sup>315</sup> aim to allow estimations free of said biases, based on the idea that by knowing the expected relatedness between pairs of individuals can help to pick apart the effects of shared environment and the effects of genome-sharing. The greater the range of IBD-sharing proportions between pairs of individuals in an isolate, the less likely it is for a

naturally occurring confounder to mirror the distribution<sup>110</sup>. This is an advantage of studying isolates with high levels of relatedness such as the Hutterites or very large isolates such as Iceland where the vast number of pairings gives a very high resolution of IBD-sharing. As discussed before, isolated populations are characterised by reduced environmental heterogeneity as well as genetic heterogeneity. In studies of large diverse cohorts of unrelated individuals, there will be multiple possible unobserved structures in the cohort. Isolated populations are conversely appropriate for studies that involve the gathering of environmental exposure data as well as genetic data; both allowing for testing hypotheses such as in gene–environment interaction models or when looking at variance decomposition and heritability.

A key limitation of studies of isolated population is indeed often the limited sample size available. We have discussed how a necessary prospective for studying (small) isolated populations will be to continue efforts to combine forces with other isolated populations and to participate in meta-analyses. Discoveries made in isolated populations will always have to be replicated and validated in other populations; and the goal of studying isolates has always been to have findings with implications to wider populations.

In the study of rare variants, it is conceivable that within a certain gene, different populations could harbour different rare variation with augmented frequency due to isolation. Association tests at the gene level could be made via rare variant burden meta-analysis as proposed in Feng, et al.<sup>317</sup>, Jiang and McPeck<sup>318</sup>, and Liu, et al.<sup>319</sup> who developed specific methods to this end. If we further add the possibility that these different populations could be isolated, then all of the benefits of studying isolates in terms of increases statistical power for low-frequency variants could translate through to such meta-analyses. Indeed, such an approach was debuted in Southam, et al.<sup>126</sup> with success; and it would be intriguing to see if this direction is expanded on larger and more diverse sets of isolates such as the group of isolates studied in Xue, et al.<sup>83</sup>, particularly as it has been suggested that there is increased power in such studies when datasets span multiple ethnic groups<sup>320</sup>.

A theme in both of our main studies in this thesis has been to compare IBD-based and non-IBD-based methods. In our first study we compared long range phasing and SHAPEIT2 and in our second study we compared different estimators of relatedness matrices. Both times, we were able to observe IBD-based methods performing very well. EAGLE2 achieved very low switch error rates and IBDLD was able to estimate IBD-sharing proportions in Cilento with high precision (Supplementary Material in Annex B). Methods based on IBD sharing are intuitively preferable when working with an isolated population; we know that IBD-sharing is prevalent and relevant in isolates and it seems logical to evoke methods that incorporate this structure. However, we were rather forced to accept in both our studies that methods that did not model for IBD performed best in the isolated population; SHAPEIT2 relied on haplotype-sharing (IBS) and GRM estimators of utilise single-point correlations between individuals which are in-fact moment estimators of single-locus IBD-sharing proportions.

The relatively recent review paper of Speed and Balding <sup>272</sup> concluded that explicit ideas of relatedness between individuals may no longer be strictly necessary in order to discuss the genetics of complex traits. Our findings appear to support this standpoint, but there are certain caveats. For differentiating between phasing methods, we in fact tested different phasing software, so there is also a question of implementation as well as algorithm choice. For example, we cannot rule out that SHAPEIT2 is in some way built better than EAGLE2, leading to small differences in performance. As EAGLE2 combines an IBD-based method with an LD-based method similar to SHAPEIT2, it would be reasonable to expect EAGLE2 to give at least equivalent results to SHAPEIT2, if not better; this was not the case. Furthermore, results from our heritability analysis showed an advantage for using GRMs to estimate relatedness matrices. This was however based on simulated phenotypes following ideal polygenic models. It could be that the advantage could turn in the favour of IBD-based methods for traits with different architectures.

At the conclusion of this thesis, it is clear that the subject of complex trait architecture remains open and that many potential research directions can still benefit greatly from studies of isolated populations. The characteristics of such populations are likely to be as favourable for testing new hypotheses regarding complex trait architecture, as they have proved favourable in the past. The output of this thesis will prove valuable to subsequent analyses of the Cilento dataset and our two published simulation studies regarding the study of population isolates will be of use to the scientific community.

## References

1. Liu, H., Prugnolle, F., Manica, A. & Balloux, F. A geographically explicit genetic model of worldwide human-settlement history. *Am J Hum Genet* **79**, 230-7 (2006).
2. Prugnolle, F., Manica, A. & Balloux, F. Geography predicts neutral genetic diversity of human populations. *Curr Biol* **15**, R159-60 (2005).
3. Ellwood, T. *The Book of the Settlement of Iceland. Translated from the Original Icelandic of Ari the Learned [or, Rather, from the Editions of 1774 and 1843 of the Anonymous Work Known as "Landnámabók"]*, by Rev. T. Ellwood, (T. Wilson, 1898).
4. Cann, H.M. *et al.* A human genome diversity cell line panel. *Science* **296**, 261-2 (2002).
5. Rosenberg, N.A. *et al.* Genetic structure of human populations. *Science* **298**, 2381-5 (2002).
6. Helgason, A., Sigurethardottir, S., Gulcher, J.R., Ward, R. & Stefansson, K. mtDNA and the origin of the Icelanders: deciphering signals of recent population history. *Am J Hum Genet* **66**, 999-1016 (2000).
7. Helgason, A. *et al.* Estimating Scandinavian and Gaelic ancestry in the male settlers of Iceland. *Am J Hum Genet* **67**, 697-717 (2000).
8. Gulcher, J., Helgason, A. & Stefansson, K. Genetic homogeneity of Icelanders. *Nat Genet* **26**, 395 (2000).
9. Ebenesersdóttir, S.S. *et al.* Ancient genomes from Iceland reveal the making of a human population. *Science* **360**, 1028 (2018).
10. Myerowitz, R. & Costigan, F.C. The major defect in Ashkenazi Jews with Tay-Sachs disease is an insertion in the gene for the alpha-chain of beta-hexosaminidase. *J Biol Chem* **263**, 18587-9 (1988).
11. Myerowitz, R. Splice junction mutation in some Ashkenazi Jews with Tay-Sachs disease: evidence against a single defect within this ethnic group. *Proc Natl Acad Sci U S A* **85**, 3955-9 (1988).
12. Behar, D.M. *et al.* MtDNA evidence for a genetic bottleneck in the early history of the Ashkenazi Jewish population. *Eur J Hum Genet* **12**, 355-64 (2004).
13. Bray, S.M. *et al.* Signatures of founder effects, admixture, and selection in the Ashkenazi Jewish population. *Proc Natl Acad Sci U S A* **107**, 16222-7 (2010).
14. Ostrer, H. A genetic profile of contemporary Jewish populations. *Nat Rev Genet* **2**, 891-8 (2001).
15. Hostetler, J.A. History and relevance of the Hutterite population for genetic studies. *Am J Med Genet* **22**, 453-62 (1985).
16. Boycott, K.M. *et al.* Clinical genetics and the Hutterite population: a review of Mendelian disorders. *Am J Med Genet A* **146A**, 1088-98 (2008).
17. Payne, M., Rupar, C.A., Siu, G.M. & Siu, V.M. Amish, mennonite, and hutterite genetic disorder database. *Paediatr Child Health* **16**, e23-4 (2011).
18. Kruglyak, L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat Genet* **22**, 139-44 (1999).
19. Ober, C., Abney, M. & McPeck, M.S. The genetic dissection of complex traits in a founder population. *Am J Hum Genet* **69**, 1068-79 (2001).
20. Bourgain, C. & Génin, E. Complex trait mapping in isolated populations: Are specific statistical methods required? *Eur J Hum Genet* **13**, 698-706 (2005).
21. Peltonen, L., Palotie, A. & Lange, K. Use of population isolates for mapping complex traits. *Nat Rev Genet* **1**, 182-90 (2000).
22. Arcos-Burgos, M. & Muenke, M. Genetics of population isolates. *Clin Genet* **61**, 233-47 (2002).
23. Kruglyak, L. Genetic isolates: separate but equal? *Proc Natl Acad Sci U S A* **96**, 1170-2 (1999).
24. Lander, E.S. & Schork, N.J. Genetic dissection of complex traits. *Science* **265**, 2037-48 (1994).
25. Kristiansson, K., Naukkarinen, J. & Peltonen, L. Isolated populations and complex disease gene identification. *Genome Biol* **9**, 109 (2008).

26. A comprehensive genetic linkage map of the human genome. NIH/CEPH Collaborative Mapping Group. *Science* **258**, 67-86 (1992).
27. Dib, C. *et al.* A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* **380**, 152-4 (1996).
28. Deloukas, P. *et al.* A physical map of 30,000 human genes. *Science* **282**, 744-6 (1998).
29. International Human Genome Sequencing, C. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860 (2001).
30. The International Human Genome Mapping, C. *et al.* A physical map of the human genome. *Nature* **409**, 934 (2001).
31. International Human Genome Sequencing, C. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931 (2004).
32. Tharp, W.G. & Sarkar, I.N. Origins of amyloid-beta. *BMC Genomics* **14**, 290 (2013).
33. Brothers, H.M., Gosztyla, M.L. & Robinson, S.R. The Physiological Roles of Amyloid-beta Peptide Hint at New Ways to Treat Alzheimer's Disease. *Front Aging Neurosci* **10**, 118 (2018).
34. Goate, A. *et al.* Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's disease. *Nature* **349**, 704-6 (1991).
35. Eckman, C.B. *et al.* A new pathogenic mutation in the APP gene (I716V) increases the relative proportion of A beta 42(43). *Hum Mol Genet* **6**, 2087-9 (1997).
36. Jonsson, T. *et al.* A mutation in APP protects against Alzheimer's disease and age-related cognitive decline. *Nature* **488**, 96-9 (2012).
37. Zhang, F. & Lupski, J.R. Non-coding genetic variants in human disease. *Hum Mol Genet* **24**, R102-10 (2015).
38. Edwards, S.L., Beesley, J., French, J.D. & Dunning, A.M. Beyond GWASs: illuminating the dark road from association to function. *Am J Hum Genet* **93**, 779-97 (2013).
39. Portin, P. & Wilkins, A. The Evolving Definition of the Term "Gene". *Genetics* **205**, 1353-1364 (2017).
40. Mendel, G. Versuche über Pflanzen-Hybriden. *Verh. naturf. Ver. Brünn* **4**, 3-47 (1866).
41. Lock, R.H. *Recent Progress in the Study of Variation, Heredity, and Evolution*, (Murray, London, 1906).
42. Morgan, T.H. Chromosomes and heredity. *Am. Nat.* **44**, 449-496 (1910).
43. Watson, J.D. & Crick, F.H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**, 737-8 (1953).
44. Franklin, R.E. & Gosling, R.G. Molecular configuration in sodium thymonucleate. *Nature* **171**, 740-1 (1953).
45. Wilkins, M.H., Seeds, W.E., Stokes, A.R. & Wilson, H.R. Helical structure of crystalline deoxypentose nucleic acid. *Nature* **172**, 759-62 (1953).
46. Bobadilla, J.L., Macek, M., Jr., Fine, J.P. & Farrell, P.M. Cystic fibrosis: a worldwide analysis of CFTR mutations--correlation with incidence data and application to screening. *Hum Mutat* **19**, 575-606 (2002).
47. Kerem, B. *et al.* Identification of the cystic fibrosis gene: genetic analysis. *Science* **245**, 1073-80 (1989).
48. Zielenski, J. *et al.* Identification of the M1101K mutation in the cystic fibrosis transmembrane conductance regulator (CFTR) gene and complete detection of cystic fibrosis mutations in the Hutterite population. *Am J Hum Genet* **52**, 609-15 (1993).
49. Morton, N.E. Logarithm of odds (lods) for linkage in complex inheritance. *Proc Natl Acad Sci U S A* **93**, 3471-6 (1996).
50. Elston, R.C. & Stewart, J. A general model for the genetic analysis of pedigree data. *Hum Hered* **21**, 523-42 (1971).
51. Lander, E.S. & Green, P. Construction of multilocus genetic linkage maps in humans. *Proc Natl Acad Sci U S A* **84**, 2363-7 (1987).

52. Kruglyak, L. & Lander, E.S. Faster multipoint linkage analysis using Fourier transforms. *J Comput Biol* **5**, 1-7 (1998).
53. Scott, D.A. *et al.* Nonsyndromic autosomal recessive deafness is linked to the DFNB1 locus in a large inbred Bedouin family from Israel. *Am J Hum Genet* **57**, 965-8 (1995).
54. Scott, D.A. *et al.* Identification of mutations in the connexin 26 gene that cause autosomal recessive nonsyndromic hearing loss. *Hum Mutat* **11**, 387-94 (1998).
55. Fisher, R.A. XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Transactions of the Royal Society of Edinburgh* **52**, 399-433 (1919).
56. Tattini, L., D'Aurizio, R. & Magi, A. Detection of Genomic Structural Variants from Next-Generation Sequencing Data. *Front Bioeng Biotechnol* **3**, 92 (2015).
57. Mills, M.C. & Rahal, C. A scientometric review of genome-wide association studies. *Commun Biol* **2**, 9 (2019).
58. Jacquard, A. *The genetic structure of populations*, (Springer-Verlag, New York - Heidelberg - Berlin, 1974).
59. Falconer, D.S. *Introduction to Quantitative Genetics*, (Oliver and Boyd, Edinburgh and London, 1960).
60. Serre, J.-L. *Génétique des populations*, (Nathan, Paris, 1997).
61. Haldane, J.B.S. The combination of linkage values, and the calculation of distance between the loci of linked factors. *J Genet* **8**, 299-309 (1919).
62. Perdry, H., Dandine-Roulland, C., Bandyopadhyay, D. & Kettner, L. Gaston: Genetic Data Handling (QC, GRM, LD, PCA) & Linear Mixed Models. CRAN <https://CRAN.R-project.org/package=gaston>. (2018).
63. Mackiewicz, D., de Oliveira, P.M., Moss de Oliveira, S. & Cebrat, S. Distribution of recombination hotspots in the human genome--a comparison of computer simulations with real data. *PLoS One* **8**, e65272 (2013).
64. Myers, S. *et al.* The distribution and causes of meiotic recombination in the human genome. *Biochem Soc Trans* **34**, 526-30 (2006).
65. Karigl, G. A recursive algorithm for the calculation of identity coefficients. *Ann Hum Genet* **45**, 299-305 (1981).
66. Albrechtsen, A. *et al.* Relatedness mapping and tracts of relatedness for genome-wide data in the presence of linkage disequilibrium. *Genet Epidemiol* **33**, 266-74 (2009).
67. Hatzikotoulas, K., Gilly, A. & Zeggini, E. Using population isolates in genetic association studies. *Briefings in Functional Genomics* **13**, 371-377 (2014).
68. Sheffield, V.C., Stone, E.M. & Carmi, R. Use of isolated inbred human populations for identification of disease genes. *Trends Genet* **14**, 391-6 (1998).
69. Wright, S. The Evolution of Dominance. *The American Naturalist* **63**, 556-561 (1929).
70. Baier, L.J. & Hanson, R.L. Genetic studies of the etiology of type 2 diabetes in Pima Indians: hunting for pieces to a complicated puzzle. *Diabetes* **53**, 1181-6 (2004).
71. Dabelea, D. *et al.* Increasing prevalence of Type II diabetes in American Indian children. *Diabetologia* **41**, 904-10 (1998).
72. Knowler, W.C., Bennett, P.H., Hamman, R.F. & Miller, M. Diabetes incidence and prevalence in Pima Indians: a 19-fold greater incidence than in Rochester, Minnesota. *Am J Epidemiol* **108**, 497-505 (1978).
73. Colonna, V. *et al.* Small effective population size and genetic homogeneity in the Val Borbera isolate. *Eur J Hum Genet* **21**, 89-94 (2013).
74. Devlin, B., Roeder, K., Otto, C., Tiobech, S. & Byerley, W. Genome-wide distribution of linkage disequilibrium in the population of Palau and its implications for gene flow in Remote Oceania. *Hum Genet* **108**, 521-8 (2001).
75. Service, S.K., Ophoff, R.A. & Freimer, N.B. The genome-wide distribution of background linkage disequilibrium in a population isolate. *Hum Mol Genet* **10**, 545-51 (2001).



76. Kong, A. *et al.* Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat Genet* **40**, 1068-75 (2008).
77. Uricchio, L.H., Chong, J.X., Ross, K.D., Ober, C. & Nicolae, D.L. Accurate imputation of rare and common variants in a founder population from a small number of sequenced individuals. *Genet Epidemiol* **36**, 312-319 (2012).
78. Glodzik, D. *et al.* Inference of identity by descent in population isolates and optimal sequencing studies. *Eur J Hum Genet* **21**, 1140-1145 (2013).
79. Gusev, A. *et al.* Whole population, genome-wide mapping of hidden relatedness. *Genome Res* **19**, 318-26 (2009).
80. Heutink, P. & Oostra, B.A. Gene finding in genetically isolated populations. *Hum Mol Genet* **11**, 2507-15 (2002).
81. Colonna, V. *et al.* Campora: A Young Genetic Isolate in South Italy. *Human heredity* **64**, 123-135 (2007).
82. Charlesworth, B. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nat Rev Genet* **10**, 195-205 (2009).
83. Xue, Y. *et al.* Enrichment of low-frequency functional variants revealed by whole-genome sequencing of multiple isolated European populations. *Nat Commun* **8**, 15927 (2017).
84. Genin, E. & Clerget-Darpoux, F. Association studies in consanguineous populations. *Am J Hum Genet* **58**, 861-6 (1996).
85. Wright, A.F., Carothers, A.D. & Pirastu, M. Population choice in mapping genes for complex diseases. *Nat Genet* **23**, 397-404 (1999).
86. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* **48**, 1279-1283 (2016).
87. Panoutsopoulou, K. *et al.* Genetic characterization of Greek population isolates reveals strong genetic drift at missense and trait-associated variants. *Nat Commun* **5**, 5345 (2014).
88. Sidore, C. *et al.* Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers. *Nat Genet* **47**, 1272-1281 (2015).
89. Vartiainen, E. *et al.* Thirty-five-year trends in cardiovascular risk factors in Finland. *Int J Epidemiol* **39**, 504-18 (2010).
90. McQuillan, R. *et al.* Runs of homozygosity in European populations. *Am J Hum Genet* **83**, 359-72 (2008).
91. Jonsson, T. *et al.* Variant of TREM2 associated with the risk of Alzheimer's disease. *N Engl J Med* **368**, 107-16 (2013).
92. Jakobsdottir, J. *et al.* Rare Functional Variant in TM2D3 is Associated with Late-Onset Alzheimer's Disease. *PLoS Genet* **12**, e1006327 (2016).
93. Steinberg, S. *et al.* Loss-of-function variants in ABCA7 confer risk of Alzheimer's disease. *Nat Genet* **47**, 445-7 (2015).
94. Arnar, D.O. & Palsson, R. Genetics of common complex diseases: a view from Iceland. *Eur J Intern Med* **41**, 3-9 (2017).
95. Hovatta, I. *et al.* Schizophrenia in the genetic isolate of Finland. *Am J Med Genet* **74**, 353-60 (1997).
96. Hovatta, I. *et al.* A genomewide screen for schizophrenia genes in an isolated Finnish subpopulation, suggesting multiple susceptibility loci. *Am J Hum Genet* **65**, 1114-24 (1999).
97. Muir, W.J. *et al.* Direct microdissection and microcloning of a translocation breakpoint region, t(1;11)(q42.2;q21), associated with schizophrenia. *Cytogenet Cell Genet* **70**, 35-40 (1995).
98. St Clair, D. *et al.* Association within a family of a balanced autosomal translocation with major mental illness. *Lancet* **336**, 13-6 (1990).
99. Hennah, W., Thomson, P., Peltonen, L. & Porteous, D. Genes and schizophrenia: beyond schizophrenia: the role of DISC1 in major mental illness. *Schizophr Bull* **32**, 409-16 (2006).

100. Ekelund, J. *et al.* Replication of 1q42 linkage in Finnish schizophrenia pedigrees. *Mol Psychiatry* **9**, 1037-41 (2004).
101. Ekelund, J. *et al.* Chromosome 1 loci in Finnish schizophrenia families. *Hum Mol Genet* **10**, 1611-7 (2001).
102. Palo, O.M. *et al.* Association of distinct allelic haplotypes of DISC1 with psychotic and bipolar spectrum disorders and with underlying cognitive impairments. *Hum Mol Genet* **16**, 2517-28 (2007).
103. Hennah, W. *et al.* Families with the risk allele of DISC1 reveal a link between schizophrenia and another component of the same molecular pathway, NDE1. *Hum Mol Genet* **16**, 453-62 (2007).
104. Hennah, W. *et al.* Haplotype transmission analysis provides evidence of association for DISC1 to schizophrenia and suggests sex-dependent effects. *Hum Mol Genet* **12**, 3151-9 (2003).
105. Brandon, N.J. *et al.* Understanding the role of DISC1 in psychiatric disease and during normal development. *J Neurosci* **29**, 12768-75 (2009).
106. Turunen, J.A. *et al.* The role of DTNBP1, NRG1, and AKT1 in the genetics of schizophrenia in Finland. *Schizophr Res* **91**, 27-36 (2007).
107. Mange, A.P. Growth and Inbreeding of a Human Isolate. *Hum Biol* **36**, 104-33 (1964).
108. Ober, C. *et al.* HLA and mate choice in humans. *Am J Hum Genet* **61**, 497-504 (1997).
109. Abney, M., McPeck, M.S. & Ober, C. Broad and narrow heritabilities of quantitative traits in a founder population. *Am J Hum Genet* **68**, 1302-7 (2001).
110. Abney, M., McPeck, M.S. & Ober, C. Estimation of variance components of quantitative traits in inbred populations. *Am J Hum Genet* **66**, 629-50 (2000).
111. Livne, O.E. *et al.* PRIMAL: Fast and Accurate Pedigree-based Imputation from Sequence Data in a Founder Population. *PLoS Computational Biology* **11**, e1004139 (2015).
112. Ober, C. & Yao, T.C. The genetics of asthma and allergic disease: a 21st century perspective. *Immunol Rev* **242**, 10-30 (2011).
113. Ober, C. *et al.* Genome-wide search for asthma susceptibility loci in a founder population. The Collaborative Study on the Genetics of Asthma. *Hum Mol Genet* **7**, 1393-8 (1998).
114. Ober, C., Tsalenko, A., Parry, R. & Cox, N.J. A second-generation genomewide screen for asthma-susceptibility alleles in a founder population. *Am J Hum Genet* **67**, 1154-62 (2000).
115. Bellenguez, C., Ober, C. & Bourgain, C. A multiple splitting approach to linkage analysis in large pedigrees identifies a linkage to asthma on chromosome 12. *Genet Epidemiol* **33**, 207-16 (2009).
116. Ober, C. *et al.* Effect of variation in CHI3L1 on serum YKL-40 level, risk of asthma, and lung function. *N Engl J Med* **358**, 1682-91 (2008).
117. Campbell, C.D. *et al.* Whole-genome sequencing of individuals from a founder population identifies candidate genes for asthma. *PLoS One* **9**, e104396 (2014).
118. Nicolae, D. *et al.* Fine mapping and positional candidate studies identify HLA-G as an asthma susceptibility gene on chromosome 6p21. *Am J Hum Genet* **76**, 349-57 (2005).
119. Ober, C. Asthma Genetics in the Post-GWAS Era. *Ann Am Thorac Soc* **13 Suppl 1**, S85-90 (2016).
120. Zuk, O., Hechter, E., Sunyaev, S.R. & Lander, E.S. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc Natl Acad Sci U S A* **109**, 1193-8 (2012).
121. Manolio, T.A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747-53 (2009).
122. Nolte, I.M. *et al.* Missing heritability: is the gap closing? An analysis of 32 complex traits in the Lifelines Cohort Study. *Eur J Hum Genet* **25**, 877-885 (2017).
123. Maher, B. Personal genomes: The case of the missing heritability. *Nature* **456**, 18-21 (2008).
124. Genin, E. & Clerget-Darpoux, F. The missing heritability paradigm: a dramatic resurgence of the GIGO syndrome in genetics. *Hum Hered* **79**, 1-4 (2015).

125. Gilly, A. *et al.* Cohort-wide deep whole genome sequencing and the allelic architecture of complex traits. *Nat Commun* **9**, 4674 (2018).
126. Southam, L. *et al.* Whole genome sequencing and imputation in isolated populations identify genetic associations with medically-relevant complex traits. *Nat Commun* **8**, 15606 (2017).
127. Jeronicic, A. *et al.* Whole-exome sequencing in an isolated population from the Dalmatian island of Vis. *Eur J Hum Genet* **24**, 1479-87 (2016).
128. Chheda, H. *et al.* Whole-genome view of the consequences of a population bottleneck using 2926 genome sequences from Finland and United Kingdom. *Eur J Hum Genet* **25**, 477-484 (2017).
129. Lescai, F. *et al.* Whole-exome sequencing of individuals from an isolated population implicates rare risk variants in bipolar disorder. *Transl Psychiatry* **7**, e1034 (2017).
130. Lencz, T. *et al.* High-depth whole genome sequencing of an Ashkenazi Jewish reference panel: enhancing sensitivity, accuracy, and imputation. *Hum Genet* **137**, 343-355 (2018).
131. Fakhro, K.A. *et al.* The Qatar genome: a population-specific tool for precision medicine in the Middle East. *Hum Genome Var* **3**, 16016 (2016).
132. Low-Kam, C. *et al.* Whole-genome sequencing in French Canadians from Quebec. *Hum Genet* **135**, 1213-1221 (2016).
133. Gudbjartsson, D.F. *et al.* Large-scale whole-genome sequencing of the Icelandic population. *Nat Genet* **47**, 435-44 (2015).
134. Lim, E.T. *et al.* Distribution and medical impact of loss-of-function variants in the Finnish founder population. *PLoS Genet* **10**, e1004494 (2014).
135. Benton, M.C. *et al.* A Phenomic Scan of the Norfolk Island Genetic Isolate Identifies a Major Pleiotropic Effect Locus Associated with Metabolic and Renal Disorder Markers. *PLoS Genet* **11**, e1005593 (2015).
136. Benton, M.C. *et al.* Mapping eQTLs in the Norfolk Island genetic isolate identifies candidate genes for CVD risk traits. *Am J Hum Genet* **93**, 1087-99 (2013).
137. Tachmazidou, I. *et al.* A rare functional cardioprotective APOC3 variant has risen in frequency in distinct population isolates. *Nat Commun* **4**, 2872 (2013).
138. Farmaki, A.E. *et al.* The mountainous Cretan dietary patterns and their relationship with cardiovascular risk factors: the Hellenic Isolated Cohorts MANOLIS study. *Public Health Nutr* **20**, 1063-1074 (2017).
139. Boyle, E.A., Li, Y.I. & Pritchard, J.K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169**, 1177-1186 (2017).
140. Liu, X., Li, Y.I. & Pritchard, J.K. Trans effects on gene expression can drive omnigenic inheritance. *bioRxiv*, 425108 (2018).
141. Shi, H., Kichaev, G. & Pasaniuc, B. Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data. *Am J Hum Genet* **99**, 139-53 (2016).
142. Loh, P.R. *et al.* Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nat Genet* **47**, 1385-92 (2015).
143. Visscher, P.M. & Yang, J. A plethora of pleiotropy across complex traits. *Nat Genet* **48**, 707-8 (2016).
144. Wray, N.R., Wijmenga, C., Sullivan, P.F., Yang, J. & Visscher, P.M. Common Disease Is More Complex Than Implied by the Core Gene Omnigenic Model. *Cell* **173**, 1573-1580 (2018).
145. Finucane, H.K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* **47**, 1228-35 (2015).
146. Lee, J.J. *et al.* Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat Genet* **50**, 1112-1121 (2018).
147. Evangelou, E. *et al.* Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits. *Nat Genet* **50**, 1412-1425 (2018).
148. Nielsen, J.B. *et al.* Biobank-driven genomic discovery yields new insight into atrial fibrillation biology. *Nat Genet* **50**, 1234-1239 (2018).

149. Pasaniuc, B. & Price, A.L. Dissecting the genetics of complex traits using summary association statistics. *Nat Rev Genet* **18**, 117-127 (2017).
150. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203-209 (2018).
151. Lencz, T. *et al.* Genome-wide association study implicates NDST3 in schizophrenia and bipolar disorder. *Nat Commun* **4**, 2739 (2013).
152. Pollin, T.I. *et al.* A null mutation in human APOC3 confers a favorable plasma lipid profile and apparent cardioprotection. *Science* **322**, 1702-5 (2008).
153. Pattaro, C. *et al.* A meta-analysis of genome-wide data from five European isolates reveals an association of COL22A1, SYT1, and GABRR2 with serum creatinine level. *BMC Med Genet* **11**, 41 (2010).
154. Ceballos, F.C., Joshi, P.K., Clark, D.W., Ramsay, M. & Wilson, J.F. Runs of homozygosity: windows into population history and trait architecture. *Nat Rev Genet* **19**, 220-234 (2018).
155. Joshi, P.K. *et al.* Directional dominance on stature and cognition in diverse human populations. *Nature* **523**, 459-462 (2015).
156. Lettre, G., Lange, C. & Hirschhorn, J.N. Genetic model testing and statistical power in population-based association studies of quantitative traits. *Genet Epidemiol* **31**, 358-62 (2007).
157. Visscher, P.M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet* **101**, 5-22 (2017).
158. Wenzel, S.E. Asthma phenotypes: the evolution from clinical to molecular approaches. *Nat Med* **18**, 716-25 (2012).
159. Mallal, S. *et al.* HLA-B\*5701 screening for hypersensitivity to abacavir. *N Engl J Med* **358**, 568-79 (2008).
160. Telli, M.L., Hunt, S.A., Carlson, R.W. & Guardino, A.E. Trastuzumab-related cardiotoxicity: calling into question the concept of reversibility. *J Clin Oncol* **25**, 3525-33 (2007).
161. Robson, M. & Offit, K. Clinical practice. Management of an inherited predisposition to breast cancer. *N Engl J Med* **357**, 154-62 (2007).
162. Offit, K. Personalized medicine: new genomics, old lessons. *Hum Genet* **130**, 3-14 (2011).
163. Masellis, M. *et al.* Dopamine D2 receptor gene variants and response to rasagiline in early Parkinson's disease: a pharmacogenetic study. *Brain* **139**, 2050-62 (2016).
164. Titova, N. & Chaudhuri, K.R. Personalized medicine in Parkinson's disease: Time to be precise. *Mov Disord* **32**, 1147-1154 (2017).
165. Torkamani, A., Wineinger, N.E. & Topol, E.J. The personal and clinical utility of polygenic risk scores. *Nat Rev Genet* **19**, 581-590 (2018).
166. Colonna, V. *et al.* Comparing population structure as inferred from genealogical versus genetic information. *European Journal of Human Genetics* **17**, 1635-1641 (2009).
167. Del Mercato, P. & Infante, A. *Cilento, uomini e vicende*, (Reggiani Editore, Salerno, 1980).
168. Ciullo, M. *et al.* New susceptibility locus for hypertension on chromosome 8q by efficient pedigree-breaking in an Italian isolate. *Hum Mol Genet* **15**, 1735-43 (2006).
169. Ciullo, M. *et al.* Identification and replication of a novel obesity locus on chromosome 1q24 in isolated populations of Cilento. *Diabetes* **57**, 783-90 (2008).
170. Sorice, R. *et al.* Association of a variant in the CHRNA5-A3-B4 gene cluster region to heavy smoking in the Italian population. *Eur J Hum Genet* **19**, 593-6 (2011).
171. Ruggiero, D. *et al.* Genetics of VEGF serum variation in human isolated populations of cilento: importance of VEGF polymorphisms. *PLoS One* **6**, e16982 (2011).
172. Sorice, R. *et al.* Genetic and environmental factors influencing the Placental Growth Factor (PGF) variation in two populations. *PLoS One* **7**, e42537 (2012).
173. Ruggiero, D. *et al.* Genetic variants modulating CRIPTO serum levels identified by genome-wide association study in Cilento isolates. *PLoS Genet* **11**, e1004976 (2015).

174. Nutile, T. *et al.* Whole-Exome Sequencing in the Isolated Populations of Cilento from South Italy. *Scientific Reports: Accepted February 2019* (2019).
175. The 1000 Genomes Project Consortium. A map of human genome variation from population scale sequencing. *Nature* **467**, 1061-1073 (2010).
176. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).
177. Choi, S.H. *et al.* Six Novel Loci Associated with Circulating VEGF Levels Identified by a Meta-analysis of Genome-Wide Association Studies. *PLoS Genet* **12**, e1005874 (2016).
178. Falchi, M. *et al.* A genomewide search using an original pairwise sampling approach for large genealogies identifies a new locus for total and low-density lipoprotein cholesterol in two genetically differentiated isolates of Sardinia. *Am J Hum Genet* **75**, 1015-31 (2004).
179. Helgason, A., Palsson, S., Gudbjartsson, D.F., Kristjansson, T. & Stefansson, K. An association between the kinship and fertility of human couples. *Science* **319**, 813-6 (2008).
180. Wijmsman, E.M., Rothstein, J.H. & Thompson, E.A. Multipoint Linkage Analysis with Many Multiallelic or Dense Diallelic Markers: Markov Chain–Monte Carlo Provides Practical Approaches for Genome Scans on General Pedigrees. *Am J Hum Genet* **79**, 846-858 (2006).
181. The UK10K Consortium. The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82-90 (2015).
182. Gazal, S. *et al.* Inbreeding Coefficient Estimation with Dense SNP Data: Comparison of Strategies and Application to HapMap III. *Human heredity* **77**, 49-62 (2014).
183. Su, Z., Marchini, J. & Donnelly, P. HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics* **27**, 2304-2305 (2011).
184. Reich, D.E. *et al.* Linkage disequilibrium in the human genome. *Nature* **411**, 199-204 (2001).
185. Daly, M.J., Rioux, J.D., Schaffner, S.F., Hudson, T.J. & Lander, E.S. High-resolution haplotype structure in the human genome. *Nat Genet* **29**, 229-32 (2001).
186. Burdick, J.T., Chen, W.M., Abecasis, G.R. & Cheung, V.G. In silico method for inferring genotypes in pedigrees. *Nat Genet* **38**, 1002-4 (2006).
187. Chen, W.M. & Abecasis, G.R. Family-based association tests for genomewide association scans. *Am J Hum Genet* **81**, 913-26 (2007).
188. Visscher, P.M. & Duffy, D.L. The value of relatives with phenotypes but missing genotypes in association studies for quantitative traits. *Genet Epidemiol* **30**, 30-6 (2006).
189. Abecasis, G.R., Cherny, S.S., Cookson, W.O. & Cardon, L.R. Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* **30**, 97-101 (2002).
190. Abecasis, G.R. & Wigginton, J.E. Handling marker-marker linkage disequilibrium: pedigree analysis with clustered markers. *Am J Hum Genet* **77**, 754-67 (2005).
191. Lange, K., Sinsheimer, J.S. & Sobel, E. Association testing with Mendel. *Genet Epidemiol* **29**, 36-50 (2005).
192. Lange, K., Weeks, D. & Boehnke, M. Programs for Pedigree Analysis: MENDEL, FISHER, and dGENE. *Genet Epidemiol* **5**, 471-2 (1988).
193. Scheet, P. & Stephens, M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet* **78**, 629-44 (2006).
194. Li, Y., Willer, C.J., Ding, J., Scheet, P. & Abecasis, G.R. MaCH: Using Sequence and Genotype Data to Estimate Haplotypes and Unobserved Genotypes. *Genet Epidemiol* **34**, 816-834 (2010).
195. Browning, S.R. & Browning, B.L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet* **81**, 1084-97 (2007).
196. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* **39**, 906-913 (2007).

197. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat Genet* **48**, 1284-1287 (2016).
198. Browning, B.L., Zhou, Y. & Browning, S.R. A One-Penny Imputed Genome from Next-Generation Reference Panels. *Am J Hum Genet* **103**, 338-348 (2018).
199. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G.R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet* **44**, 955-959 (2012).
200. Delaneau, O., Zagury, J.-F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat Meth* **10**, 5-6 (2013).
201. O'Connell, J. *et al.* Haplotype estimation for biobank-scale data sets. *Nat Genet* **48**, 817-820 (2016).
202. Li, N. & Stephens, M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**, 2213-33 (2003).
203. Loh, P.-R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nat Genet* **48**, 1443-1448 (2016).
204. Loh, P.-R., Palamara, P.F. & Price, A.L. Fast and accurate long-range phasing in a UK Biobank cohort. *Nat Genet* **48**, 811-816 (2016).
205. Durbin, R. Efficient haplotype matching and storage using the positional Burrows–Wheeler transform (PBWT). *Bioinformatics* **30**, 1266-1272 (2014).
206. Browning, Brian L. & Browning, Sharon R. Genotype Imputation with Millions of Reference Samples. *Am J Hum Genet* **98**, 116-126 (2016).
207. Browning, S.R. Multilocus association mapping using variable-length Markov chains. *Am J Hum Genet* **78**, 903-13 (2006).
208. Palin, K., Campbell, H., Wright, A.F., Wilson, J.F. & Durbin, R. Identity-by-Descent-Based Phasing and Imputation in Founder Populations Using Graphical Models. *Genet Epidemiol* **35**, 853-860 (2011).
209. Genovese, G., Leibon, G., Pollak, M.R. & Rockmore, D.N. Improved IBD detection using incomplete haplotype information. *BMC Genet* **11**, 58 (2010).
210. Hickey, J.M. *et al.* A combined long-range phasing and long haplotype imputation method to impute phase for SNP genotypes. *Genetics, Selection, Evolution : GSE* **43**, 12-12 (2011).
211. O'Connell, J. *et al.* A General Approach for Haplotype Phasing across the Full Spectrum of Relatedness. *PLoS Genetics* **10**, e1004234 (2014).
212. Browning, S.R. Missing data imputation and haplotype phase inference for genome-wide association studies. *Hum Genet* **124**, 439-50 (2008).
213. Baum, L.E. An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. *Inequalities* **3**, 1-8 (1972).
214. Baum, L.E., Petrie, T., Soules, G. & Weiss, N. A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. *Ann. Math. Statist.* **41**, 164-171 (1970).
215. Viterbi, A. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory* **13**, 260-269 (1967).
216. Forney, G.D. The viterbi algorithm. *Proceedings of the IEEE* **61**, 268-278 (1973).
217. Rabiner, L. & Juang, B. An introduction to hidden Markov models. *IEEE ASSP Magazine* **3**, 4-16 (1986).
218. Petrushin, V. Hidden Markov Models: Fundamentals and Applications Part 2: Discrete and Continuous Hidden Markov Models.  
<https://www.eecis.udel.edu/~lliao/cis841s06/hmmtutorialpart2.pdf>.
219. Petrushin, V. Hidden Markov Models: Fundamentals and Applications Part 1: Markov Chains and Mixture Models. <https://www.eecis.udel.edu/~lliao/cis841s06/hmmtutorialpart1.pdf>.
220. Delaneau, O., Coulonges, C. & Zagury, J.F. Shape-IT: new rapid and accurate algorithm for haplotype inference. *BMC Bioinformatics* **9**, 540 (2008).

221. Delaneau, O., Marchini, J. & Zagury, J.F. A linear complexity phasing method for thousands of genomes. *Nat Methods* **9**, 179-81 (2011).
222. Delaneau, O., Howie, B., Cox, Anthony J., Zagury, J.-F. & Marchini, J. Haplotype Estimation Using Sequencing Reads. *Am J Hum Genet* **93**, 687-696 (2013).
223. Thompson, E.A. The IBD process along four chromosomes. *Theor Popul Biol* **73**, 369-73 (2008).
224. Mitt, M. *et al.* Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel. *Eur J Hum Genet* (2017).
225. Howie, B., Marchini, J. & Stephens, M. Genotype imputation with thousands of genomes. *G3 (Bethesda)* **1**, 457-70 (2011).
226. Han, L. & Abney, M. Identity by descent estimation with dense genome-wide genotype data. *Genet Epidemiol* **35**, 557-67 (2011).
227. Han, L. & Abney, M. Using identity by descent estimation with dense genotype data to detect positive selection. *Eur J Hum Genet* **21**, 205-11 (2013).
228. Leutenegger, A.L. *et al.* Estimation of the inbreeding coefficient through use of genomic data. *Am J Hum Genet* **73**, 516-23 (2003).
229. Browning, B.L. & Browning, S.R. A fast, powerful method for detecting identity by descent. *Am J Hum Genet* **88**, 173-82 (2011).
230. Browning, B.L. & Browning, S.R. Detecting identity by descent and estimating genotype error rates in sequence data. *Am J Hum Genet* **93**, 840-51 (2013).
231. Moltke, I., Albrechtsen, A., Hansen, T.V., Nielsen, F.C. & Nielsen, R. A method for detecting IBD regions simultaneously in multiple individuals--with applications to disease genetics. *Genome Res* **21**, 1168-80 (2011).
232. Hochreiter, S. HapFABIA: identification of very short segments of identity by descent characterized by rare variants in large sequencing data. *Nucleic Acids Res* **41**, e202 (2013).
233. Naseri, A., Liu, X., Zhang, S. & Zhi, D. Ultra-fast Identity by Descent Detection in Biobank-Scale Cohorts using Positional Burrows-Wheeler Transform. *bioRxiv*, 103325 (2017).
234. Li, Y., Willer, C.J., Ding, J., Scheet, P. & Abecasis, G.R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol* **34**, 816-34 (2010).
235. Howie, B., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5**, e1000529 (2009).
236. Lunter, G. Haplotype matching in large cohorts using the Li and Stephens model. *Bioinformatics* (2018).
237. Deelen, P. *et al.* Improved imputation quality of low-frequency and rare variants in European samples using the 'Genome of The Netherlands'. *Eur J Hum Genet* **22**, 1321-6 (2014).
238. Pistis, G. *et al.* Rare variant genotype imputation with thousands of study-specific whole-genome sequences: implications for cost-effective study designs. *Eur J Hum Genet* **23**, 975-983 (2015).
239. Surakka, I. *et al.* Founder population-specific HapMap panel increases power in GWA studies through improved imputation accuracy and CNV tagging. *Genome Res* **20**, 1344-51 (2010).
240. Roshyara, N.R. & Scholz, M. Impact of genetic similarity on imputation accuracy. *BMC Genetics* **16**, 90 (2015).
241. Joshi, P.K. *et al.* Local Exome Sequences Facilitate Imputation of Less Common Variants and Increase Power of Genome Wide Association Studies. *PLOS ONE* **8**, e68604 (2013).
242. Hou, L. *et al.* A population-specific reference panel empowers genetic studies of Anabaptist populations. *Sci Rep* **7**, 6079 (2017).
243. Zhou, W. *et al.* Improving power of association tests using multiple sets of imputed genotypes from distributed reference panels. *Genet Epidemiol* **41**, 744-755 (2017).
244. Ye, S. *et al.* Comparison of genotype imputation strategies using a combined reference panel for chicken population. *Animal*, 1-8 (2018).

245. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat Rev Genet* **11**, 499-511 (2010).
246. Roshyara, N.R., Kirsten, H., Horn, K., Ahnert, P. & Scholz, M. Impact of pre-imputation SNP-filtering on genotype imputation results. *BMC Genet* **15**, 88 (2014).
247. Roshyara, N.R., Horn, K., Kirsten, H., Ahnert, P. & Scholz, M. Comparing performance of modern genotype imputation methods in different ethnicities. *Scientific Reports* **6**, 34386 (2016).
248. Liu, E.Y. *et al.* Genotype Imputation of Metachip SNPs Using a Study-Specific Reference Panel of ~4,000 Haplotypes in African Americans From the Women's Health Initiative. *Genet Epidemiol* **36**, 107-117 (2012).
249. Broeckx, B.J.G. *et al.* An exome sequencing based approach for genome-wide association studies in the dog. *Sci Rep* **7**, 15680 (2017).
250. Rowan, T.N. *et al.* A Multi-Breed Reference Panel and Additional Rare Variation Maximizes Imputation Accuracy in Cattle. *bioRxiv*, 517144 (2019).
251. Brondum, R.F., Ma, P., Lund, M.S. & Su, G. Short communication: genotype imputation within and across Nordic cattle breeds. *J Dairy Sci* **95**, 6795-800 (2012).
252. Patterson, M. *et al.* WhatsHap: Weighted Haplotype Assembly for Future-Generation Sequencing Reads. *J Comput Biol* **22**, 498-509 (2015).
253. Garg, S., Martin, M. & Marschall, T. Read-based phasing of related individuals. *Bioinformatics* **32**, i234-i242 (2016).
254. Bansal, V. & Bafna, V. HapCUT: an efficient and accurate algorithm for the haplotype assembly problem. *Bioinformatics* **24**, i153-9 (2008).
255. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-303 (2010).
256. Zheng, G.X. *et al.* Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat Biotechnol* **34**, 303-11 (2016).
257. Choi, Y., Chan, A.P., Kirkness, E., Telenti, A. & Schork, N.J. Comparison of phasing strategies for whole human genomes. *PLoS Genet* **14**, e1007308 (2018).
258. Cheung, C.Y., Thompson, E.A. & Wijsman, E.M. GIGI: an approach to effective imputation of dense genotypes on large pedigrees. *Am J Hum Genet* **92**, 504-16 (2013).
259. Ullah, E. *et al.* Comparison and assessment of family- and population-based genotype imputation methods in large pedigrees. *Genome Res* **29**, 125-134 (2019).
260. Saad, M. & Wijsman, E.M. Combining family- and population-based imputation data for association analysis of rare and common variants in large pedigrees. *Genet Epidemiol* **38**, 579-90 (2014).
261. Abney, M. & El Sherbiny, A. Kinpute: Using identity by descent to improve genotype imputation. *bioRxiv*, 399147 (2018).
262. Sargolzaei, M., Chesnais, J.P. & Schenkel, F.S. A new approach for efficient genotype imputation using information from relatives. *BMC Genomics* **15**, 478 (2014).
263. Wang, Y., Lin, G., Li, C. & Stothard, P. Genotype Imputation Methods and Their Effects on Genomic Predictions in Cattle. *Springer Science Reviews* **4**, 79-98 (2016).
264. Ma, P., Brondum, R.F., Zhang, Q., Lund, M.S. & Su, G. Comparison of different methods for imputing genome-wide marker genotypes in Swedish and Finnish Red Cattle. *J Dairy Sci* **96**, 4666-77 (2013).
265. Ni, G. *et al.* Comparison among three variant callers and assessment of the accuracy of imputation from SNP array data to whole-genome sequence level in chicken. *BMC Genomics* **16**, 824 (2015).
266. Ventura, R.V. *et al.* Assessing accuracy of imputation using different SNP panel densities in a multi-breed sheep population. *Genet Sel Evol* **48**, 71 (2016).
267. Friedrich, J. *et al.* Accuracy of genotype imputation in Labrador Retrievers. *Anim Genet* **49**, 303-311 (2018).



268. Hickey, J.M., Kinghorn, B.P., Tier, B., van der Werf, J.H. & Cleveland, M.A. A phasing and imputation method for pedigreed populations that results in a single-stage genomic evaluation. *Genet Sel Evol* **44**, 9 (2012).
269. Antolin, R., Nettelblad, C., Gorjanc, G., Money, D. & Hickey, J.M. A hybrid method for the imputation of genomic data in livestock populations. *Genet Sel Evol* **49**, 30 (2017).
270. Nettelblad, C. Breakdown of methods for phasing and imputation in the presence of double genotype sharing. *PLoS One* **8**, e60354 (2013).
271. Tenesa, A. & Haley, C.S. The heritability of human disease: estimation, uses and abuses. *Nat Rev Genet* **14**, 139-49 (2013).
272. Speed, D. & Balding, D.J. Relatedness in the post-genomic era: is it still useful? *Nat Rev Genet* **16**, 33-44 (2015).
273. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* **42**, 565-9 (2010).
274. Yang, J., Zeng, J., Goddard, M.E., Wray, N.R. & Visscher, P.M. Concepts, estimation and interpretation of SNP-based heritability. *Nat Genet* **49**, 1304-1310 (2017).
275. Dandine-Roulland, C. & Perdry, H. The Use of the Linear Mixed Model in Human Genetics. *Hum Hered* **80**, 196-206 (2015).
276. Gilmour, A.R., Thompson, R. & Cullis, B.R. Average Information REML: An Efficient Algorithm for Variance Parameter Estimation in Linear Mixed Models. *Biometrics* **51**, 1440-1450 (1995).
277. Pilia, G. *et al.* Heritability of cardiovascular and personality traits in 6,148 Sardinians. *PLoS Genet* **2**, e132 (2006).
278. Traglia, M. *et al.* Heritability and demographic analyses in the large isolated population of Val Borbera suggest advantages in mapping complex traits genes. *PLoS One* **4**, e7554 (2009).
279. Zaitlen, N. *et al.* Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. *PLoS Genet* **9**, e1003520 (2013).
280. van Dongen, J., Willemsen, G., Chen, W.M., de Geus, E.J. & Boomsma, D.I. Heritability of metabolic syndrome traits in a large population-based sample. *J Lipid Res* **54**, 2914-23 (2013).
281. Chen, X. *et al.* Dominant Genetic Variation and Missing Heritability for Human Complex Traits: Insights from Twin versus Genome-wide Common SNP Models. *Am J Hum Genet* **97**, 708-14 (2015).
282. Zhu, Z. *et al.* Dominance genetic variation contributes little to the missing heritability for human complex traits. *Am J Hum Genet* **96**, 377-85 (2015).
283. Vinkhuyzen, A.A., Wray, N.R., Yang, J., Goddard, M.E. & Visscher, P.M. Estimation and partition of heritability in human populations using whole-genome analysis methods. *Annu Rev Genet* **47**, 75-95 (2013).
284. Vitart, V. *et al.* Heritabilities of ocular biometrical traits in two croatian isolates with extended pedigrees. *Invest Ophthalmol Vis Sci* **51**, 737-43 (2010).
285. Gazal, S. *et al.* Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. *Nat Genet* **49**, 1421-1427 (2017).
286. Speed, D., Cai, N., Johnson, M.R., Nejentsev, S. & Balding, D.J. Reevaluation of SNP heritability in complex human traits. *Nat Genet* **49**, 986-992 (2017).
287. Kathiresan, S. *et al.* Common variants at 30 loci contribute to polygenic dyslipidemia. *Nat Genet* **41**, 56-65 (2009).
288. Sandhu, M.S. *et al.* LDL-cholesterol concentrations: a genome-wide association study. *Lancet* **371**, 483-91 (2008).
289. Burnett, J.R. & Hooper, A.J. Common and rare gene variants affecting plasma LDL cholesterol. *Clin Biochem Rev* **29**, 11-26 (2008).
290. Willer, C.J. *et al.* Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nat Genet* **40**, 161-9 (2008).

291. Teslovich, T.M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707-13 (2010).
292. Sanna, S. *et al.* Fine mapping of five loci associated with low-density lipoprotein cholesterol detects variants that double the explained heritability. *PLoS Genet* **7**, e1002198 (2011).
293. Rincint, R. *et al.* Recovering power in association mapping panels with variable levels of linkage disequilibrium. *Genetics* **197**, 375-87 (2014).
294. Laporte, F., Charcosset, A. & Mary-Huard, T. Estimation of the relatedness coefficients from biallelic markers, application in plant mating designs. *Biometrics* **73**, 885-894 (2017).
295. Franceschini, N. *et al.* Prospective associations of coronary heart disease loci in African Americans using the MetaboChip: the PAGE study. *PLoS One* **9**, e113203 (2014).
296. Pe'er, I., Yelensky, R., Altshuler, D. & Daly, M.J. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet Epidemiol* **32**, 381-5 (2008).
297. Sobota, R.S. *et al.* Addressing population-specific multiple testing burdens in genetic association studies. *Ann Hum Genet* **79**, 136-47 (2015).
298. Johansen, C.T. & Hegele, R.A. Genetic bases of hypertriglyceridemic phenotypes. *Curr Opin Lipidol* **22**, 247-53 (2011).
299. Sonnenburg, W.K. *et al.* GPIHBP1 stabilizes lipoprotein lipase and prevents its inhibition by angiopoietin-like 3 and angiopoietin-like 4. *J Lipid Res* **50**, 2421-9 (2009).
300. Coca-Prieto, I. *et al.* Childhood-onset chylomicronaemia with reduced plasma lipoprotein lipase activity and mass: identification of a novel GPIHBP1 mutation. *J Intern Med* **270**, 224-8 (2011).
301. Franssen, R. *et al.* Chylomicronemia with low postheparin lipoprotein lipase levels in the setting of GPIHBP1 defects. *Circ Cardiovasc Genet* **3**, 169-78 (2010).
302. Beigneux, A.P. *et al.* Chylomicronemia with a mutant GPIHBP1 (Q115P) that cannot bind lipoprotein lipase. *Arterioscler Thromb Vasc Biol* **29**, 956-62 (2009).
303. Paquette, M., Hegele, R.A., Pare, G. & Baass, A. A novel mutation in GPIHBP1 causes familial chylomicronemia syndrome. *J Clin Lipidol* **12**, 506-510 (2018).
304. Gonzaga-Jauregui, C. *et al.* Whole-exome sequencing reveals GPIHBP1 mutations in infantile colitis with severe hypertriglyceridemia. *J Pediatr Gastroenterol Nutr* **59**, 17-21 (2014).
305. Abney, M., Ober, C. & McPeck, M.S. Quantitative-trait homozygosity and association mapping and empirical genomewide significance in large, complex pedigrees: fasting serum-insulin level in the Hutterites. *Am J Hum Genet* **70**, 920-34 (2002).
306. Vitezica, Z.G., Legarra, A., Toro, M.A. & Varona, L. Orthogonal Estimates of Variances for Additive, Dominance, and Epistatic Effects in Populations. *Genetics* **206**, 1297-1307 (2017).
307. DePristo, M.A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**, 491-498 (2011).
308. Dou, J. *et al.* Estimation of kinship coefficient in structured and admixed populations using sparse sequencing data. *PLoS Genet* **13**, e1007021 (2017).
309. Spiliopoulou, A., Colombo, M., Orchard, P., Agakov, F. & McKeigue, P. GenImp: Fast Imputation to Large Reference Panels Using Genotype Likelihoods from Ultralow Coverage Sequencing. *Genetics* **206**, 91-104 (2017).
310. Vieira, F.G., Albrechtsen, A. & Nielsen, R. Estimating IBD tracts from low coverage NGS data. *Bioinformatics* **32**, 2096-2102 (2016).
311. Vieira, F.G., Fumagalli, M., Albrechtsen, A. & Nielsen, R. Estimating inbreeding coefficients from NGS data: Impact on genotype calling and allele frequency estimation. *Genome Res* **23**, 1852-61 (2013).
312. Lipatov, M., Sanjeev, K., Patro, R. & Veeramah, K. Maximum Likelihood Estimation of Biological Relatedness from Low Coverage Sequencing Data. *bioRxiv*, 023374 (2015).
313. Waples, R.K., Albrechtsen, A. & Moltke, I. Allele frequency-free inference of close familial relationships from genotypes or low depth sequencing data. *Mol Ecol* (2018).

314. Kim, S.Y. *et al.* Estimation of allele frequency and association mapping using next-generation sequencing data. *BMC Bioinformatics* **12**, 231-231 (2011).
315. Young, A.I. *et al.* Relatedness disequilibrium regression estimates heritability without environmental bias. *Nat Genet* **50**, 1304-1310 (2018).
316. Samuels, D.C. *et al.* Finding the lost treasures in exome sequencing data. *Trends Genet* **29**, 593-9 (2013).
317. Feng, S., Liu, D., Zhan, X., Wing, M.K. & Abecasis, G.R. RAREMETAL: fast and powerful meta-analysis for rare variants. *Bioinformatics* **30**, 2828-9 (2014).
318. Jiang, D. & McPeck, M.S. Robust rare variant association testing for quantitative traits in samples with related individuals. *Genet Epidemiol* **38**, 10-20 (2014).
319. Liu, D.J. *et al.* Meta-analysis of gene-level tests for rare variant association. *Nat Genet* **46**, 200-4 (2014).
320. Mensah-Ablorh, A. *et al.* Meta-Analysis of Rare Variant Association Tests in Multiethnic Populations. *Genet Epidemiol* **40**, 57-65 (2016).

## Glossary of Terms

Throughout the thesis, the following notations and terms are used very often:

$i, k = 1, \dots, N$  : Index of individuals in a sample or population

$j = 1, \dots, M$  : Index of genetic variants in a chromosome or any particular set of variants

$r = 1, \dots, R$  : Elements of a set of reference haplotypes, or similar

$\Delta_l$ ,  $l = 1, \dots, 9$  : Jacquard's 9 condensed identity coefficients

$\varphi_{ik}$  : Kinship of individuals  $i$  and  $k$

$f_i$  : Inbreeding coefficient of individual  $i$

$G_{ij}$  : Genotype of individual  $i$  at position  $j$

$Y_i$  : The phenotype of individual  $i$

$X_{ij}^*$  : Coded genotypes of individual  $i$  at position  $j$ , superscript  $*$  will indicate various different coding, in particular ' $a$ ' and ' $d$ ' refer to additive and non-additive/dominant coding

$E[*]$  : Expectation of  $*$

$var[*]$  : Variance of  $*$

$d_H$  : Genetic distance

$A$  &  $a$  : The major and minor alleles for a genetic variant

$p$  or  $p_A$  &  $q$  or  $p_a$  : Frequencies of the major and minor alleles

$AA, Aa$ , &  $aa$  : Three possible genotypes for a genetic variant

$p_{AA}, p_{Aa}$ , &  $p_{aa}$  : Frequencies of the three possible genotypes for a genetic variant

$u_0, u_1$ , &  $u_2$  : Genetic effects of the three possible genotypes for a genetic variant

$g_{ij}$  : Genetic value of individual  $i$  at position  $j$

$a_j$  &  $d_j$  : Additive and non-additive genetic effects

$V_Y, V_G$ , &  $V_E$  : Phenotypic, Genetic, and Environmental Variance Components

$\sigma_E^2$  : Environmental variance parameter

$\tau_a$  and  $\tau_D$  : Genetic additive and genetic non-additive/dominant variance parameters

$\sigma_S^2$  : Household/sibling variance parameter

$H^2$  : Broad-sense Heritability

$h_a^2$  &  $h_d^2$  ( $h_A^2$  &  $h_D^2$  in Annex B) : Additive and non-additive components of Heritability ( $H^2$ )

$h_S^2$  : Household/sibling variance components

$L_{ij}^{AA}$ ,  $L_{ij}^{Aa}$ , &  $L_{ij}^{aa}$  : Likelihoods, scaled to be probabilities (sum to 1), of the three possible genotypes of individual  $i$  at position  $j$

$w_{ij}$ : Imputed dosage of individual  $i$  at position  $j$

GWAS : Genome-Wide Association Study

Array : Genotyping Array data

370K : Illumina 370 K array

OMNI : Illumina HumanOmniExpress array

WES : Whole-Exome Sequencing

WGS : Whole-Genome Sequencing

MAF : Minor Allele Frequency

bp, Kb, and Mb : 1 base pair, 1,000 base pairs, 1,000,000 base pairs

cM : centi-Morgans (unit of Genetic Distance)

LD : Linkage Disequilibrium

IBD : Identity-By-Descent

IBS : Identity-By-State

IBD=2, IBD=1, and IBD=0 : Number of alleles shared Identity-By-Descent

LRP : Long Range Phasing

SSP : Study Specific Panel

HMM : Hidden Markov Model

HRC : The Haplotype Reference Consortium panel

1000G : The 1000 Genomes panel

UK10K : The UK10K panel

SER : Switch Error Rate

Imputation Accuracy : Correlation between dosages and true genotypes

'info' : IMPUTE2 imputation quality score

'RSQ' : MINIMAC3 imputation quality score

LMM : Linear Mixed Model

MLE : Maximum Likelihood Estimate

GRM : Genetic Relationship Matrix

K : The kinship matrix or additive GRM

D : The dominance matrix or non-additive GRM

BMI : Body-Mass Index

LDL : Low-Density Lipoprotein

'Pedigree Simulation' : The simulation using only the pedigree of Cilento

'HapGen+Pedigree Simulation' : The simulation using both HapGen and the pedigree of Cilento







'rare' : When the frequency of the variant is described as rare, this will normally pertain to some hypothetical general population, unless specified otherwise.

## Annexes

**Annex A:** Herzig, AF, Nutile, T, Babron, M-C, Ciullo, M, Bellenguez, C, & Leutenegger, A-L. (2018). Strategies for phasing and imputation in a population isolate. *Genet Epidemiol* 42,201-213.

**Annex B:** Herzig, AF, Nutile, T, Ruggiero, D, Ciullo, M, Perdry, H, & Leutenegger, A-L. (2018). Detecting the dominance component of heritability in isolated and outbred human populations. *Scientific Reports* 8, 18048.

# Strategies for phasing and imputation in a population isolate

Anthony Francis Herzig<sup>1,2</sup>  | Teresa Nutile<sup>3</sup>  | Marie-Claude Babron<sup>1,2</sup>  |  
Marina Ciullo<sup>3,4</sup>  | Céline Bellenguez<sup>5,6,7\*</sup>  | Anne-Louise Leutenegger<sup>1,2\*</sup> 

<sup>1</sup>Université Paris-Diderot, Sorbonne Paris Cité, U946, Paris, France

<sup>2</sup>Inserm, U946, Genetic Variation and Human Diseases, Paris, France

<sup>3</sup>Institute of Genetics and Biophysics A. Buzzati-Traverso—CNR, Naples, Italy

<sup>4</sup>IRCCS Neuromed, Pozzilli, Isernia, Italy

<sup>5</sup>Inserm, U1167, RID-AGE—Risk Factors and Molecular Determinants of Aging-Related Diseases, Lille, France

<sup>6</sup>Institut Pasteur de Lille, Lille, France

<sup>7</sup>Université de Lille, U1167—Excellence Laboratory LabEx DISTALZ, Lille, France

## Correspondence

Anthony Francis Herzig, INSERM U946, 27 Rue Juliette Dodu, 75010 Paris, France.  
Email: anthony.herzig@inserm.fr

## Funding information

European Sequencing and Genotyping Infrastructure, Grant/Award Number: 262055; Seventh Framework Programme; A.F.H.

\*These authors contributed equally to this study.

## ABSTRACT

In the search for genetic associations with complex traits, population isolates offer the advantage of reduced genetic and environmental heterogeneity. In addition, cost-efficient next-generation association approaches have been proposed in these populations where only a subsample of representative individuals is sequenced and then genotypes are imputed into the rest of the population. Gene mapping in such populations thus requires high-quality genetic imputation and preliminary phasing. To identify an effective study design, we compare by simulation a range of phasing and imputation software and strategies. We simulated 1,115,604 variants on chromosome 10 for 477 members of the large complex pedigree of Campora, a village within the established isolate of Cilento in southern Italy. We assessed the phasing performance of identical by descent based software ALPHAPHASE and SLRP, LD-based software SHAPEIT2, SHAPEIT3, and BEAGLE, and new software EAGLE that combines both methodologies. For imputation we compared IMPUTE2, IMPUTE4, MINIMAC3, BEAGLE, and new software PBWT. Genotyping errors and missing genotypes were simulated to observe their effects on the performance of each software. Highly accurate phased data were achieved by all software with SHAPEIT2, SHAPEIT3, and EAGLE2 providing the most accurate results. MINIMAC3, IMPUTE4, and IMPUTE2 all performed strongly as imputation software and our study highlights the considerable gain in imputation accuracy provided by a genome sequenced reference panel specific to the population isolate.

## KEYWORDS

founder effect, genotyping errors, identity by descent, linkage disequilibrium, study specific panel

## 1 | INTRODUCTION

For many complex traits, attention has turned to the search for associations with low-frequency or rare variants. This follows the success of genome-wide association studies (GWAS) in identifying associations with many common variants but without yet gaining a satisfactorily complete description of the genetic heritability for various complex traits. The large sample sizes required to achieve sufficient power to detect associations with rare variants (particularly if effect size is modest),

combined with the sequencing cost, limit the opportunities for finding such associations.

Population isolates have inherent characteristics beneficial to the study of complex traits, namely reduced environmental and genetic heterogeneity (Bourgain & Génin, 2005; Hatzikotoulas, Gilly, & Zeggini, 2014). Because of the bottleneck at the founding of the population followed by generations of genetic drift, some mutations that would be described as “rare” in general populations can occur with greater frequency in the population isolate. Fewer individuals are



hence required to achieve sufficient power for analyses. Also, unique patterns of linkage disequilibrium (LD) are expected within such populations and long haplotypes will be identical by descent (IBD) among members of the population even when not closely related.

To take advantage of the prevalence of shared IBD regions, a subset of the study population can be whole-genome sequenced (WGS) and then made available as a Study Specific Panel (SSP) for genetic imputation on to the remainder of the genotyped sample (Asimit & Zeggini, 2012; Holm et al., 2011; Zeggini, 2011). Alternatively, public reference panels could be employed for imputation: for example, the 1000 Genomes Project (1000G) (The 1000 Genomes Project Consortium, 2015) or the Haplotype Reference Consortium (HRC) (McCarthy et al., 2016). All study designs require efficient phasing and imputation, and a range of software has been developed to this end.

Methods for phasing can be classified as either LD based (Browning & Browning, 2016; Delaneau, Zagury, & Marchini, 2013; O'Connell et al., 2016) or IBD based (Glodzik et al., 2013; Hickey et al., 2011; Livne et al., 2015; Palin, Campbell, Wright, Wilson, & Durbin, 2011). O'Connell et al. (2014) found that despite the prevalence of IBD regions in an isolate, LD-based methods outperformed the IBD-based method proposed by Palin et al. (2011) when tested in several population isolates. Recently a new method was proposed to combine both LD-based and IBD-based approaches and was shown to achieve increased phasing accuracy over LD-based methods in a large outbred population (Loh et al., 2016; Loh, Palamara, & Price, 2016). However, this new approach is yet to be evaluated in a population isolate.

Several studies investigating imputation strategies have shown that using an imputation panel specific to the population under study increases imputation accuracy compared to using larger multiethnic public reference panels. This has been observed in population isolates (Joshi et al., 2013; Pistis et al., 2015; Surakka et al., 2010) and in outbred populations (Deelen et al., 2014; Mitt et al., 2017; Roshyara & Scholz, 2015). However, no study has compared imputation software and imputation strategies together in a population isolate since the recent releases of updated software versions (Browning & Browning, 2016; Bycroft et al., 2017; Das et al., 2016), new methods (Durbin, 2014), and larger and denser reference panels (McCarthy et al., 2016; The 1000 Genomes Project Consortium, 2015).

In population isolates, genealogical data may be available. There exist many methods for phasing and imputation using in part or solely pedigree data (Abecasis, Cherny, Cookson, & Cardon, 2002; Chen & Schaid, 2014; Cheung, Thompson, & Wijisman, 2013; Hickey et al., 2011; Livne et al., 2015). The size and complexity of the pedigrees typical to isolates precludes the application of some methods that use only pedigree data. However, methods that combine IBD inference from

both genetic and pedigree information should be well adapted for population isolates (Hickey et al., 2011; Livne et al., 2015).

Here we provide an updated evaluation of state-of-the-art phasing and imputation methods in the context of a population isolate. We test the latest versions of existing software as well as recently released software on simulated data with the structure of the population isolate of Campora in southern Italy. The effects of errors and missingness on the performance of each software were also assessed. The design of our study also gives the opportunity to observe in detail the effects of isolate characteristics on phasing and imputation software in order to provide recommendations for future studies of population isolates.

## 2 | METHODS

### 2.1 | Campora

Pedigree and genetic data for Campora have previously been gathered as part of the Vallo di Diano Project. The pedigree contains 2,894 members, including 495 founders and spans the 16th century to the present day (Colonna et al., 2007). The pedigree of Campora was reconstructed from parish records (supplementary Fig. S1). Although the pedigree captures many loops and connections that result in a high level of relatedness, it falls short of reaching back to the founding event of Campora. Previous analysis of sex chromosomes and mitochondrial DNA in Campora concluded that around 96.7% of the genetic variability was explained by 17 female and 20 male lineages. Hence, while the recorded pedigree contains 495 founders, the true founding event in Campora likely involved closer to 37 founders (Colonna et al., 2007).

Of the present day individuals, 477 have high-quality genotypes, all of whom have been genotyped on an Illumina 370K SNP-chip array (ARRAY). A subset of 93 individuals has whole exome sequencing (WES) data and another subset of 18 individuals has whole-genome sequencing (WGS) data. The WES subset was selected to serve as an SSP using the method described in Uricchio, Chong, Ross, Ober, and Nicolae (2012) but with genetic kinship in the place of genealogical kinship. This way we selected a subset with a high level of relatedness to the remaining unselected individuals while avoiding high levels of relatedness among the selected individuals. This resulted in a selection of 93 individuals spread across the bottom four generations of the Campora pedigree with a higher proportion coming from the bottom two generations. The set of 93 individuals does not contain multiple members of any single nuclear family.

### 2.2 | Simulation

Genetic data were simulated with similar characteristics to those observed in the real genetic data from Campora

(supplementary Fig. S2). Gene dropping of chromosome 10 (chr10) was performed on the entire pedigree using the MOR-GAN package Genedrop (Wijsman, Rothstein, & Thompson, 2006). For time efficiency, Genedrop was only provided with a coarse genetic map, we then sampled precise locations of recombination events on the far denser genetic map used in our study as in Gazal et al. (2014).

We considered two approaches to generate the founder haplotypes, both enlisting the haplotypes of the UK10K panel (UK10K) (The UK10K Consortium, 2015) (see URLs). The UK10K contains member of the TwinsUK cohort; for the purposes of the simulation one member from each pair of monozygotic or dizygotic twins was removed leading to a pool of 7,500 haplotypes. In a first simulation strategy we sampled the 990 pedigree founder haplotypes without replacement from the pool of UK10K haplotypes. In a second simulation strategy we first sampled 80 haplotypes from UK10K to approximate the founding event of roughly 37 founders in Campora and then used HapGen2 (Su, Marchini, & Donnelly, 2011) to simulate recombination events and mutations to create a pool of mosaic haplotype from which the 990 founder haplotypes of the pedigree were sampled without replacement. From hence we refer to these two simulation strategies as “Pedigree” and “HapGen+Pedigree”, respectively. Further details on HapGen2 parameters are given in Supplementary Materials. Each strategy was independently replicated 100 times with independent draws for the 990 and 80 haplotypes, respectively. In each replicate we simulated variants at ARRAY positions for all 477 individuals and WGS positions for the 93 SSP individuals. We observed that the HapGen+Pedigree simulation produced simulated data with a mean pairwise genetic kinship (estimated on ARRAY genotypes) closer to the mean observed in Campora (supplementary Fig. S3) suggesting the HapGen+Pedigree simulation better mimicked the data of Campora.

### 2.3 | Error models

Errors and sporadic missingness were simulated in the data. Both were introduced independently in the two simulated platforms (ARRAY and WGS). Missing genotypes observed in the ARRAY data in Campora were set to missing in the simulated data. Errors on the ARRAY data were simulated with a simple undirected error model where one allele from a genotype can change to the other available allele (major or minor) at that position with an error rate of 0.001.

For the WGS data, we simulated multiple reads for each genotype (including erroneous reads), from which genotype likelihoods and genotype quality scores were estimated using a similar methodology to previous studies involving next-generation sequencing data simulation (Kim et al., 2011; Vieira, Albrechtsen, & Nielsen, 2016). Genotypes

that emerged with a quality score less than 20 were set to missing, otherwise the genotype of greatest likelihood was kept. Our error model was tuned to produce missingness rates close to the observed missingness rate in Campora (between 0.01 and 0.02) and error rates similar to those expected on the sequencing platform used in Campora (between 0.003 and 0.004). Full details of our WGS data simulation and the error model are given in Supplementary Materials and specific nucleotide error rates in supplementary Table S1.

To assess the effect of genotyping errors and missingness on the performance of each phasing and imputation algorithm, we completed the same phasing and imputation steps using simulated data with both genotype errors and missingness (Imperfect data) but also without any such imperfections (Perfect data).

### 2.4 | Quality control

No quality control was performed on individuals. For Imperfect data, all genotypes in the nuclear family were set to missing each time a Mendelian error was introduced by our error models. In all files, variants were removed for low minor allele frequency (MAF), significant deviation from Hardy-Weinberg equilibrium and for high missingness in the case of imperfect data (Supplementary Materials).

### 2.5 | Phasing

Phasing algorithms can be separated into two main methodological classes:

LD-based methods that rely on hidden Markov models (HMM) are employed by phasing algorithms SHAPEIT2 (Delaneau, Zagury et al. 2013) and BEAGLE (Browning & Browning, 2016). Phase is estimated with respect to LD patterns and haplotype similarity and is built for each individual as a mosaic of current haplotype estimations of all other sample individuals as well as external reference haplotypes if they are made available to the algorithm. For SHAPEIT2 we considered the use of the “duohmm” option (O’Connell et al., 2014) that harnesses parent-offspring or duo information for phasing. We also tested SHAPEIT3 (O’Connell et al., 2016), a new version of SHAPEIT2 designed for large sample sizes.

In IBD-based methods, long stretches of IBD can be directly sought between pairs of individuals in order to phase directly each individual in turn in an approach named Long Range Phasing (Kong et al., 2008). We tested two software that employ Long Range Phasing: SLRP (Palin et al., 2011) and ALPHAPHASE (Hickey et al., 2011). ALPHAPHASE was developed for livestock populations and is able to use pedigree information in addition to genotypes. SLRP, which

was specifically designed for population isolates, uses only the genotypes.

Two releases of a new software that combines LD-based and IBD-based methods were also tested: EAGLE version 1 (EAGLE1) (Loh et al., 2016) and version 2 (EAGLE2) (Loh et al., 2016). EAGLE1 was aimed at general populations and was developed to phase data with very large sample sizes. It employs Long Range Phasing followed by an HMM in a second step. EAGLE2 focuses on harnessing an external reference panel. It no longer uses Long Range Phasing and instead is based on the positional Burrows-Wheeler transform (Durbin, 2014) and an HMM. Yet if EAGLE2 is used without a reference panel it adds the Long Range Phasing algorithm of EAGLE1 as an initial step.

BEAGLE, SHAPEIT2, SHAPEIT3, and EAGLE2 can make inference from an external reference panel when phasing. We tested all software without an external panel and SHAPEIT2 and EAGLE2 with the 1000G panel.

Switch error rate (SER) is the standard measure to assess the accuracy of an estimation of genetic phase. A switch error is observable between two consecutive heterozygous sites and occurs if phase at the second heterozygous site is incorrect with respect to that of the first. The SER is the fraction of pairs of heterozygous sites where a switch error has occurred out of the total number of possible pairs. A description of SER calculation in the presence of known genotype errors is given in the Supplementary Materials. We calculated SERs on the entirety of chr10: globally over all individuals and variants, for each individual, and for each variant. We compared the SER per variant to MAF calculated naively on the simulated ARRAY genotypes and the mean SER of each individual to the individual's mean genetic kinship with all other sample members. Kinship was estimated from the simulated ARRAY genotypes using the R package "Gaston" (see URLs).

## 2.6 | Imputation

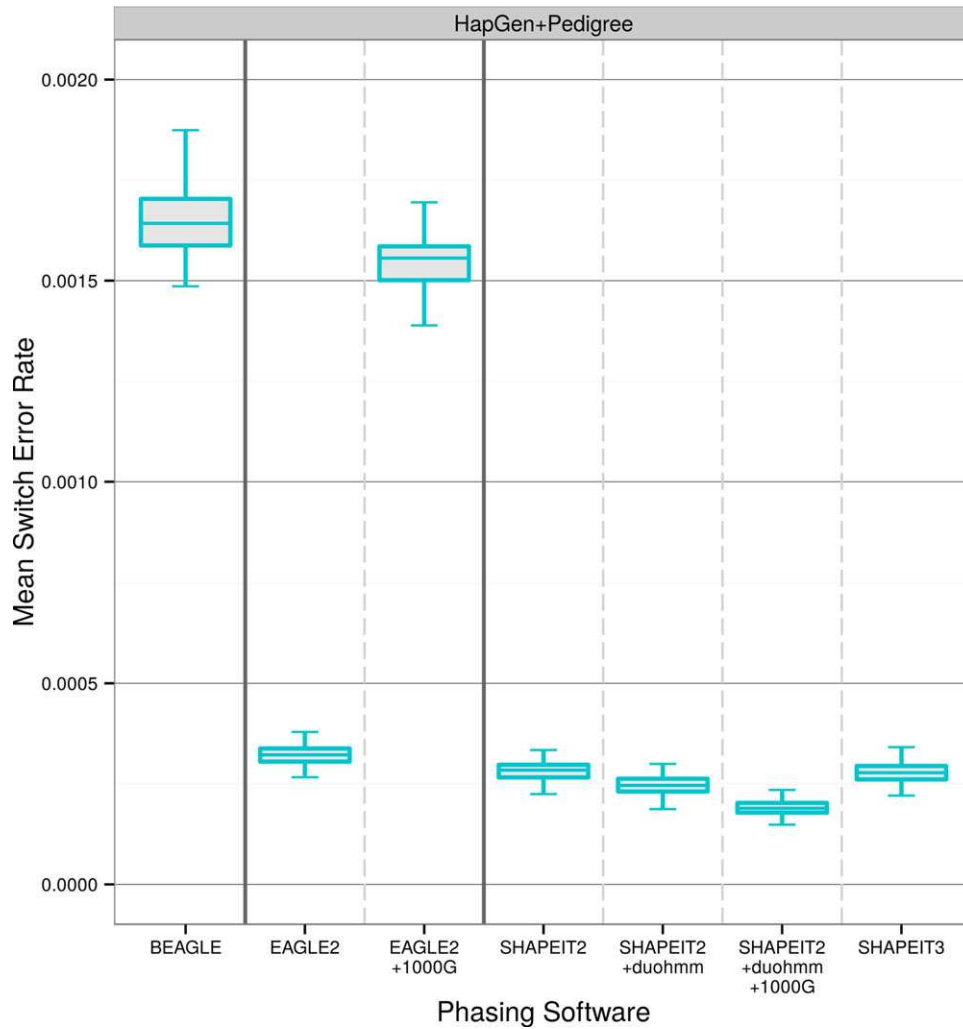
LD-based imputation methods IMPUTE2 (Howie, Donnelly, & Marchini, 2009), IMPUTE4 (Bycroft et al., 2017), BEAGLE v4.1 (Browning & Browning, 2016), and MINIMAC3 (Das et al., 2016) were compared when using the 1000G as a reference panel. We included all 2,504 individuals from all populations of the 1000G for imputation as this has been shown to be the best approach (Howie, Marchini, & Stephens, 2011). We also used the HRC panel but only for MINIMAC3 due to the computational burden associated with this panel. The HRC panel used was the version made available for download through the European Genome-phenome Archive, which contains 27,165 individuals, including all samples from the 1000G. As our simulations were based on the UK10K, we removed all UK10K haplotypes, leading to 23,450 individuals. We also tested the PBWT software

(Durbin, 2014) on 20 of our replicates through use of the Wellcome Trust's Sanger Imputation Service and again using the 1000G as a reference panel. We did not test PBWT with the HRC panel as we could not remove the UK10K haplotypes from the panel when using this imputation service. To restrict to 20 replicates per simulation strategy was a pragmatic decision based on the time required to upload data to the server.

The benefits of imputation using an SSP (either alone or combined with a public reference panel) were investigated. In each simulation replicate, we first created an SSP: WGS and ARRAY data for the 93 SSP individuals were combined (setting discordant genotypes created by our error models to missing in the case of Imperfect data) and then phased. Imputation was performed with IMPUTE2 with a combination of this SSP and the 1000G panel, using the software option that allows the combination of two reference panels through cross-imputation. We also tested MINIMAC3 with a combination of the SSP and the HRC panel. As MINIMAC3 does not offer an option for cross-imputation, the two panels were first restricted to the set of variants in common between them and then merged. We denote a phasing or imputation strategy by the name of the software added to the panels employed, for example: EAGLE2+1000G, IMPUTE2+1000G, or MINIMAC3+HRC+SSP.

Imputation accuracy of software was assessed in each replicate by the squared Pearson's correlation between imputed genotype dosages and original simulated genotypes for each biallelic SNP polymorphic in the simulated data and present in the output of every imputation software. Imputation was restricted to the telomeric region of the short arm of chr10 (20 Mb in length). As imputation scenarios involving the SSP of 93 individuals were tested, imputation accuracy was measured for all scenarios on the complementing set of 384 non-SSP individuals. Mean imputation accuracy was calculated over distinct partitions of the observed range of MAF by averaging across all variants in each MAF bin considered. MAF was estimated naively on all 7,500 UK10K haplotypes. All imputation software were run on prephased data arising from the best phased data found when comparing phasing software. For general populations, it is possible that prephasing could lead to a loss of imputation accuracy (Rosshyara, Horn, Kirsten, Ahnert, & Scholz, 2016) but this is unlikely to be significant in population isolates where highly accurate phased data are achievable (Howie, Fuchsberger, Stephens, Marchini, & Abecasis, 2012).

All imputation software provided imputation quality scores per variant; the calculation of such scores varies between imputation software but the scores have been shown to be highly correlated to each other (Marchini & Howie, 2010). We investigated the consequences of post imputation quality control based on imputation quality scores in a separate analysis.



**FIGURE 1** Global switch error rates for BEAGLE, EAGLE2, SHAPEIT2, and SHAPEIT3 for the HapGen+Pedigree simulation strategy

## 2.7 | Speed

Because we only concentrate on a single chromosome with a moderate number of individuals, computation time was not an issue for our simulation. However, many of the algorithms considered were designed with speed and low memory usage in mind. Indeed, EAGLE1, EAGLE2, BEAGLE, MINIMAC3, PBWT, IMPUTE4, and SHAPEIT3 are all geared toward performance when analyzing very large numbers of individuals or when leveraging very large external reference panels. We measured real and computational time elapsed during a single replicate of the HapGen+Pedigree simulation. All phasing and imputation executions were completed on a  $2 \times 6$  core,  $2 \times 12$  thread 2.66 GHz Intel Xeon Processor X5650 with 96 Gb of random access memory.

The options used for phasing and imputation software are discussed in the Supplementary Materials and the software versions used are detailed in the URLs.

## 3 | RESULTS

### 3.1 | LD-based phasing

For analyses of phasing performance, we present results from only the HapGen+Pedigree simulation unless otherwise indicated as the patterns of results were very similar between the two simulation strategies. Imperfect ARRAY data initially spanned 13,599 variants on chr10 and following quality control an average of 13,262 variants remained on the HapGen+Pedigree simulation strategy. Totalling over the 477 individuals and across the entirety of chr10, phasing algorithms were required to phase an average of 2,150,627 heterozygous sites in each simulation replicate. All LD-based phasing algorithms considered were able to phase the ARRAY data to a high degree of accuracy with global SERs below 0.002 (Fig. 1). EAGLE2 delivered improved SER compared to EAGLE1 (supplementary Fig. S4) and so we only present detailed

results for EAGLE2. SHAPEIT2 provided the most accurately phased data and the additions of the “duohmm” option and the 1000G as an external reference panel further improved its performance. SHAPEIT3 performed similarly to SHAPEIT2 and for subsequent analysis we will only present results for SHAPEIT2+duohmm+1000G. SHAPEIT2+duohmm+1000G achieved a mean SER of  $1.9 \times 10^{-4}$  while EAGLE2 achieved  $3.2 \times 10^{-4}$ . The mean global SERs for all phasing strategies considered are given in supplementary Table S2.

### 3.2 | IBD-based phasing

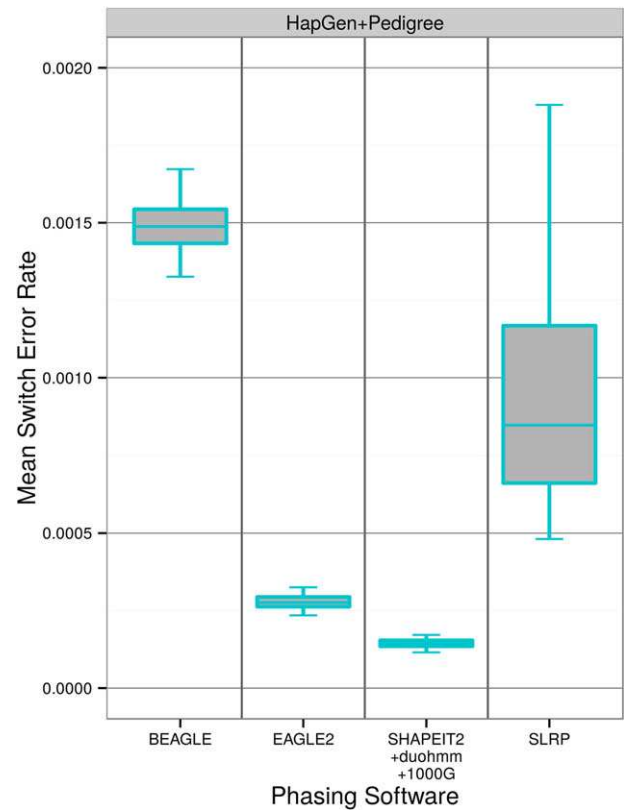
We note that EAGLE2 outperformed EAGLE2+1000G; conversely to what was observed for SHAPEIT2 (Fig. 1). This result can be interpreted as evidence of the utility of the EAGLE2 Long Range Phasing routine for population isolates as this routine is irrevocably omitted from the algorithm when using an external reference panel.

ALPHAPHASE and SLRP both provided added complications because they only phase sites that were found IBD between individuals. SLRP outperformed ALPHAPHASE in terms of SER even though ALPHAPHASE had access to the pedigree information (supplementary Fig. S5). ALPHAPHASE however phased more heterozygous sites than SLRP that may explain some of the difference in SER between the two. We chose to compare only SLRP to other software (Fig. 2) as SLRP was clearly stronger than ALPHAPHASE. Owing to the sites left unphased by SLRP, a separate calculation of SER restricted to the set of sites phased by SLRP in each replicate was carried out. SLRP produced higher SERs than SHAPEIT2+duohmm+1000G and EAGLE2 and reducing the analysis to these sites resulted in lower SERs for all other phasing software (when compared to Fig. 1). On these sites, SHAPEIT2+duohmm+1000G achieved a mean SER of  $1.4 \times 10^{-4}$  while EAGLE2 achieved  $2.7 \times 10^{-4}$  and so a considerable proportion of the switch errors observed in Figure 1 occurred on the small percentage (1.6% on average) of heterozygous sites left unphased by SLRP. This suggests that the sites left unphased by SLRP, which are by definition in areas where SLRP was unable to identify IBD between individuals, are precisely those sites that other software frequently phased incorrectly.

### 3.3 | Factors that impact phasing performance

To further explore the performance of phasing software, we performed a series of subanalyses to identify patterns in the distributions of switch errors on chr10. Variants with low MAF had demonstrably higher SERs, whether using LD-based software or EAGLE2 (supplementary Fig. S6).

The levels of IBD in the simulated populations clearly affected phasing performance as all software had improved

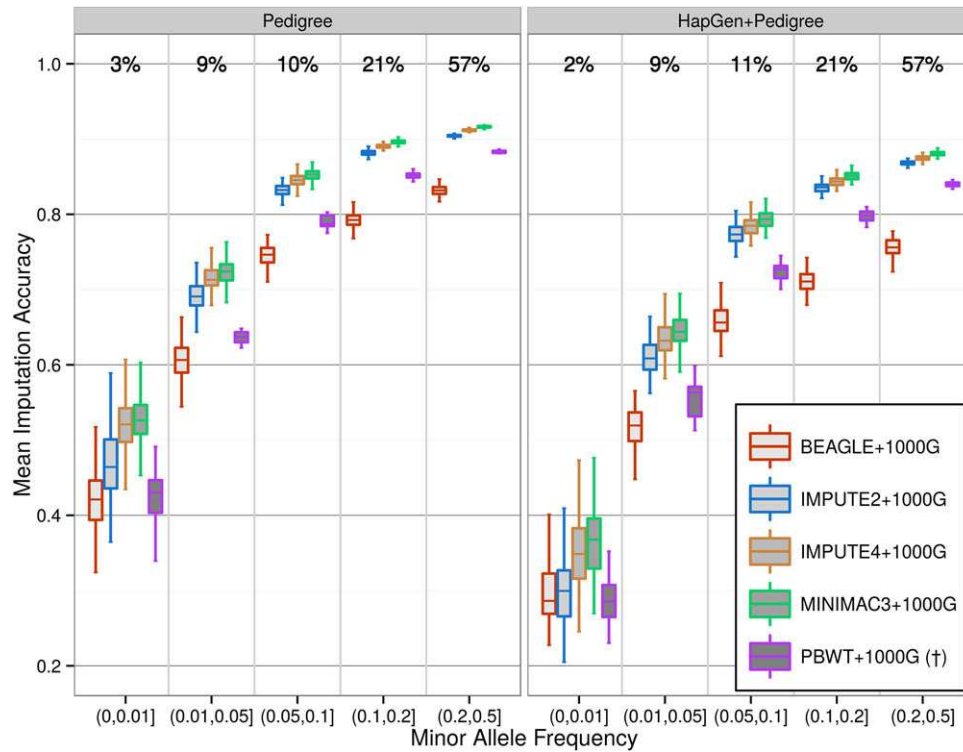


**FIGURE 2** Global switch error rates for BEAGLE, SLRP, EAGLE2, and SHAPEIT2+duohmm+1000G for the HapGen+Pedigree simulation strategy on the set of variants successfully phased by SLRP in each replicate

phasing accuracy in the presence of the elevated IBD in the HapGen+Pedigree simulation as compared to the Pedigree simulation strategy (supplementary Fig. S7). Similarly, SLRP and ALPHAPHASE both phased many more sites on the HapGen+Pedigree simulation (supplementary Fig. S8a, b). At the individual level, all phasing algorithms had lower performance for the individuals with the lowest mean pairwise genetic kinship to the rest of the sample (supplementary Fig. S9a–c).

Phasing software returned slightly higher SERs when phasing data with errors and missingness (supplementary Fig. S10) and ALPHAPHASE and SLRP phased significantly less sites when errors and missingness were present (supplementary Fig. S8a, b). The effect of imperfections within the data was noticed particularly on the Long Range Phasing algorithms (ALPHAPHASE, SLRP, and EAGLE2).

We specifically investigated the IBD status at switch errors sites in the Pedigree simulation strategy for EAGLE2 and SHAPEIT2+duohmm+1000G (Supplementary Materials and supplementary Fig. S11) as in only this simulation strategy, true IBD sharing was accessible from Genedrop. For both phasing approaches, there were a lower number of true IBD haplotypes at switch errors sites (six IBD haplotypes on



**FIGURE 3** Software imputation accuracy with the 1000G as an external reference panel and for the Pedigree and HapGen+Pedigree simulation strategies. The percentages of variants in each MAF bin are displayed atop the figure. Total number of variants for each strategy: 40,989 (Pedigree) and 40,407 (HapGen+Pedigree). † PBWT was only run on 20 replicates of each simulation strategy

average) compared to correctly phased sites (17 IBD haplotypes on average). These true IBD haplotypes are the haplotypes that the software can use as phase informative. Hence the performance of the LD-based method SHAPEIT2 was implicitly linked to the prevalence of IBD.

### 3.4 | Accuracy of imputation software

Results pertain to imputation of phased Imperfect ARRAY data from both simulation strategies unless otherwise stated. Following the results from the phasing software evaluation, we phased ARRAY and WGS data with SHAPEIT2+duohmm+1000G. This phasing strategy was also found to be the most accurate for WGS data (supplementary Fig. S12).

In each replicate, mean imputation accuracy was calculated across all polymorphic SNPs found within the output of every software. On average this entailed a selection of 40,989 SNPs for the Pedigree simulation and 40,407 SNPs for the HapGen+Pedigree simulation. This difference is ascribed to the presence of more monomorphic variants in the HapGen+Pedigree simulation.

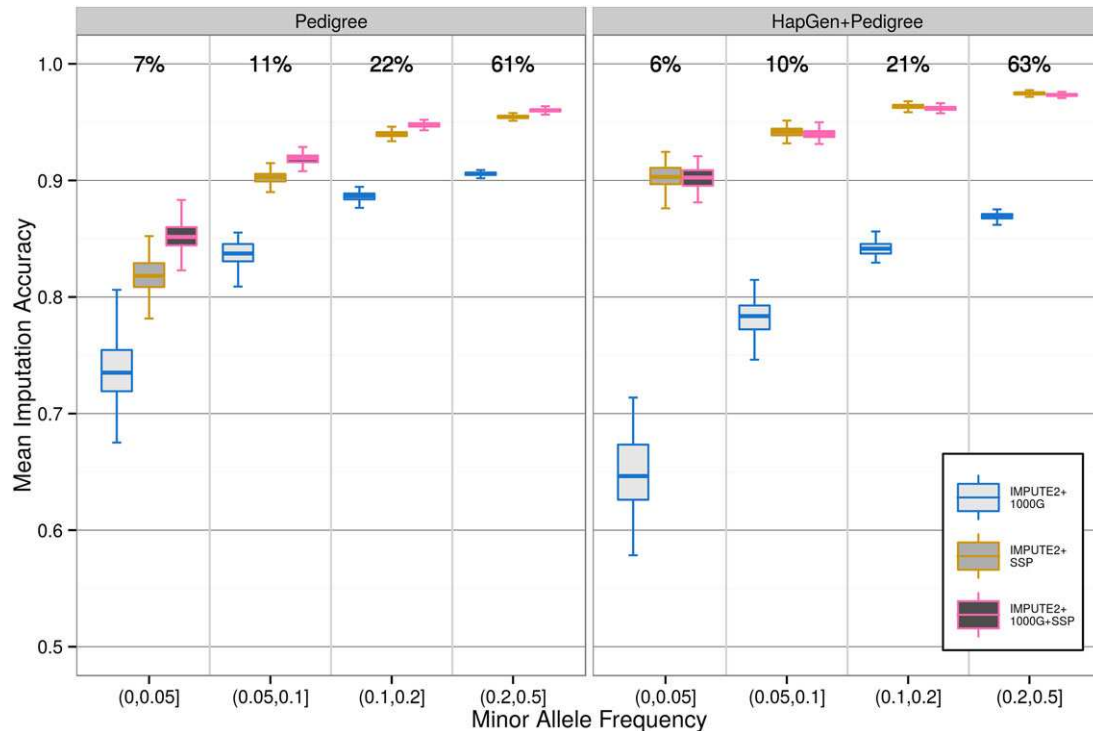
When using 1000G as the reference panel, MINIMAC3 provided the best imputation accuracy in both simulation strategies followed closely by IMPUTE4 and then IMPUTE2 (Fig. 3). Variants with low MAF were universally harder

to impute. BEAGLE and PBWT consistently delivered lower imputation accuracy than IMPUTE2, IMPUTE4, and MINIMAC3. Although IMPUTE4 marginally outperformed IMPUTE2, it currently does not offer the option to combine reference panels necessary for subsequent analyses in which we hence compare IMPUTE2 and MINIMAC3.

Genotype errors and missingness on the ARRAY data had minimal impact on imputation accuracy but such imperfections simulated on the WGS SSP had slightly more effect (supplementary Figs. S13 and S14).

### 3.5 | Impact of reference panel choice

By comparing the two simulation strategies, we were able to identify the consequences of reference panel choice in a population isolate. When the 1000G was chosen as the external reference panel, imputation accuracy was significantly lower in the HapGen+Pedigree simulation strategy than in the Pedigree one (Fig. 3). This difference in imputation accuracy may be due to differences in MAF between the simulated data and the 1000G reference panel (Supplementary Materials and supplementary Fig. S15). MAFs on the HapGen+Pedigree simulation had drifted further away from the 1000G reference panel and the variants with the highest differences in MAF to the 1000G reference panel were imputed with lower



**FIGURE 4** Imputation accuracy of IMPUTE2 when using various reference panels for the Pedigree and HapGen+Pedigree simulation strategies. The set of variants used for comparison is a reduction of the set used in Figure 3 because using only the SSP as a reference panel limits the set of possible variants to compare imputed dosages and true genotypes. This depleted the number of variants in the  $[0,0.01)$  MAF category, which was therefore merged with that of  $[0.01,0.05)$  MAF. Total number of variants for each strategy: 35,058 (Pedigree) and 34,065 (HapGen+Pedigree)

accuracy than random selections of similar variants (supplementary Fig. S16a).

Imputation with the SSP was an improvement upon imputation with the 1000G for both IMPUTE2 and MINIMAC3 (Figs. 4 and 5). When using the SSP, the simulation strategy with the highest imputation accuracy was the HapGen+Pedigree simulation, contrary to when using only the 1000G (Fig. 3). This can be ascribed the higher levels of IBD between the 93 SSP members and the 384 other individuals in this simulation strategy. Indeed, the most accurately imputed individuals were consistently those with higher values of mean pairwise kinship to the set of SSP individuals (supplementary Fig. S17).

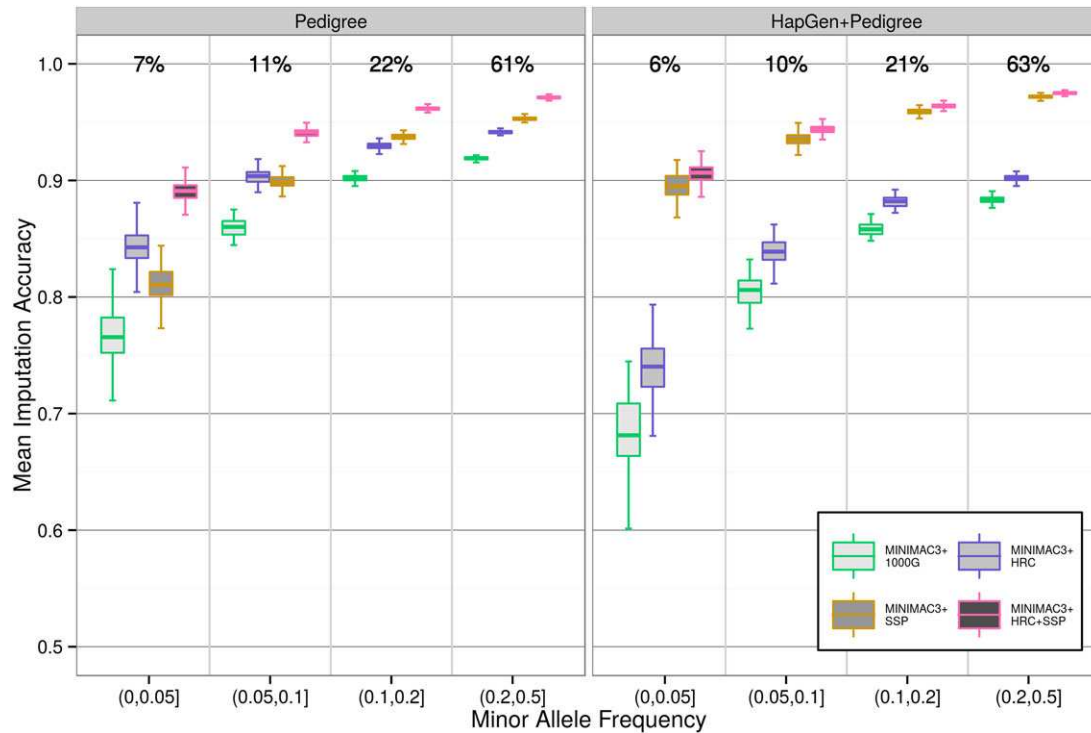
For MINIMAC3, imputation accuracy was clearly improved by using the HRC over the 1000G (Fig. 5). Imputation that included the SSP again produced more accurate results than imputation with only public reference panels on the HapGen+Pedigree simulation strategy. Rare variants were however imputed more accurately by MINIMAC3+HRC than by MINIMAC3+SSP on the Pedigree simulation. The results of Figs. 3–5 are summarized in supplementary Table S3.

The founding event in an isolate will result in higher MAFs for certain variants as compared to general populations. Variants with a high difference in MAF compared to the 1000G were imputed as well as the random selections of

comparable variants under IMPUTE2+SSP, but with lower accuracy under IMPUTE2+1000G (supplementary Fig. S16a). When changing reference panel from the 1000G to the SSP, we observed that imputation accuracy increased the most for variants with a MAF higher in the sample than in the 1000G (Supplementary Materials and supplementary Fig. S16b). Another consequence of using solely the 1000G as a reference panel was the fact that some variants that were monomorphic in the sample were imputed with dosages compatible with being heterozygous for many individuals, that is, polymorphic in the sample (supplementary Fig. S16c, d).

### 3.6 | Imputation quality scores

Finally, we analyzed the effect of applying various thresholds of the “info” score for IMPUTE2 and the “RSQ” score for MINIMAC3. Each successive threshold improved imputation accuracy for both IMPUTE2 and MINIMAC3 with the latter still providing higher accuracy in each MAF bin (Supplementary Materials and supplementary Fig. S18a, b). The “RSQ” measure gave a better indication of imputation accuracy than “info” and we also found that higher thresholds than the standard ones were arguably preferable for both rare and common variants in both simulation strategies (Supplementary Materials and supplementary Table S4).



**FIGURE 5** Imputation accuracy of MINIMAC3 with various reference panels on the same set of variants as used in Figure 4

### 3.7 | Speed

For phasing, BEAGLE, EAGLE1, and EAGLE2 were the fastest because they allow for multiple threading. SHAPEIT2 required more computation time than other algorithms. For imputation, the quickest software were BEAGLE and IMPUTE4. MINIMAC3+1000G was quicker than IMPUTE2+1000G. We observed the additional complexity encountered by IMPUTE2 when performing cross-imputation. The full list of times is given in supplementary Table S5.

## 4 | DISCUSSION

Using simulated genetic data, we have rigorously tested the performance of a range of phasing and imputation software in a population isolate. EAGLE2 (without a reference panel) and SHAPEIT2 were the strongest performing phasing software with SHAPEIT2+duohmm+1000G giving the most accurately phased data. MINIMAC3, IMPUTE4, and IMPUTE2 all performed well and we observed a slight advantage for MINIMAC3. MINIMAC3 imputation was more accurate with the HRC as an external reference panel rather than the 1000G. The use of an SSP proved to be a very successful strategy, when used alone, but even more so when combined with a large external reference panel. MINIMAC3+HRC+SSP proved the most effective imputation strategy. Genotype errors and missingness were shown

to have only a small effect on the performance of all phasing and imputation software considered.

If we compare our phasing results to published results for outbred populations, it is clear that all methods performed with greater accuracy (SERs at least one order of magnitude smaller) on our simulated data. Indeed, for outbred populations, very large sample sizes have been required to achieve the high level of phasing accuracy observed in our population isolate study. For examples, see Bycroft et al. (2017), Loh, Danecek, et al. (2016), O'Connell et al. (2016), and Mitt et al. (2017).

IBD-based phasing methods did not prove as effective as the LD-based software SHAPEIT2, which appeared itself to directly profit from IBD in the sample. O'Connell et al. (2014) also observed SHAPEIT2 benefiting from IBD. Indeed, the performance of IBD-based and LD-based software followed a similar pattern: all were less accurate when less IBD was present and all had difficulty when phasing the likely non-IBD regions of the genome and when phasing individuals with a low average kinship to the rest of the sample. IBD-based methods were the most affected by imperfections in the data.

EAGLE was expected to perform strongly on population isolate data as it should combine the appeal of Long Range Phasing and the strengths of LD-based methods such as SHAPEIT2. Though the combination of IBD-based and LD-based approaches in EAGLE1 and EAGLE2 is a clear improvement over previous Long Range Phasing software, it does not provide more accurate phasing than the LD-based



approach implemented in SHAPEIT2. This is in accord with the results of Mitt et al. (2017) in a cohort of intermediate size but not with those of Loh, Danecek, et al. (2016) in much larger cohorts. EAGLE2 was developed with the aim of handling large sample sizes but as gene-mapping studies in population isolates will remain by nature small scale, SHAPEIT2 remains the optimum choice for phasing.

Published results for SHAPEIT3 in outbred populations suggest that it may return less accurate phased data compared to SHAPEIT2 (O'Connell et al., 2016). Of the two, SHAPEIT2 is recommended for sample sizes less than 20,000, which would encompass the realm of population isolates. In our study, SHAPEIT2 and SHAPEIT3 performed very similarly.

Our comparisons on imputation strategies agree with recent literature (Deelen et al., 2014; Mitt et al., 2017; Pistis et al., 2015) in terms of the improvement in accuracy brought by a reference panel specific to the population under study. Mitt et al. (2017) concluded that for certain outbred populations, such a panel can outperform an order of magnitude larger and more diverse reference panel (the HRC). We show that for a population isolate, an SSP can be far smaller and still outperform the HRC. As discussed in Asimit and Zeggini (2012), the appropriate size of the SSP will depend on the diversity of the isolate.

The HapGen+Pedigree simulation strategy gave the best representation of a true isolate with a strong founder effect producing large disparities to general populations represented in public databases. Of the two simulation strategies, imputation accuracy was significantly lower on this simulation when using only a public reference panel. This suggests that for a population isolate with a very small set of founders and high relatedness between individuals, using public reference panels alone is not a completely appropriate strategy for imputation. A better solution is to sequence a subset of the isolate to serve as an SSP. Even with a very large external reference panel, such as the HRC (here 23,450 individuals), imputation accuracy could not match the level reached by an SSP of 93 individuals. Using an SSP was particularly effective when imputing variants with MAFs higher in the sample than in an external reference panel. As such variants are precisely those that motivate the study of population isolates, this strengthens the argument for using an SSP in a population isolate.

We observed that the best results came from combining an external reference panel and our SSP together for imputation. IMPUTE2 facilitates cross-imputation of two reference panels with variants at nonidentical sets of positions. This is an attractive strategy for isolates as all positions from both panels can be imputed including variants specific to the isolate.

The accuracy of imputation can be directly linked to the statistical power of subsequent association tests (Browning & Browning, 2009; Huang, Wang, & Rosenberg, 2009; Li, Willer, Ding, Scheet, & Abecasis, 2010; Surakka et al., 2010).

Indeed, if  $N$  is the number of individuals in a study and a variant is imputed with an imputation accuracy of  $r^2 = \alpha$ , then the statistical power of an association test using the imputed dosages is equivalent to that of a test performed on observed genotypes for  $\alpha N$  samples. This is the intended interpretation of imputation quality scores that are estimates of the true  $r^2$  statistics (Marchini & Howie, 2010). To give an example, we have observed differences in imputation accuracy of around 0.2 for rare variants ( $MAF \leq 0.05$ ) and 0.1 for common variants ( $MAF > 0.05$ ) between MNIMAC3+1000G and MINIMAC3+HRC+SSP on the HapGen+Pedigree simulation (supplementary Table S3). Imputation accuracy was measured on a sample of size  $N = 384$  (non-SSP individuals), hence, these observed differences in imputation accuracy would correspond to losses of power equivalent to removing around 77 or 38 of these individuals from subsequent analyses, respectively. Studies in isolates typically involve unavoidably modest sample sizes. Hence, there is great importance in attaining the highest imputation accuracy possible in such studies in order to preserve power.

One possible option for SHAPEIT2 that we did not consider is the PIR option that harnesses phase informative reads (Delaneau, Howie, Cox, Zagury, & Marchini 2013). To include this in our simulation would have required the creation of the original read data; this was judged to be too great a computational burden for our study. To be clearer option was tested in Mitt et al. (2017) and did not significantly improve the global performance of SHAPEIT2. Another version of SHAPEIT2, SHAPEITR (Sharp, Kretzschmar, Delaneau, & Marchini, 2016), sets out to improve phasing by concentrating on rare variants. However, as it is so far only available through the Oxford Statistics Phasing Server (see URLs), it is not suitable for an in-house simulation.

One software that we have not tested is PRIMAL, which uses Long Range Phasing and is designed for phasing and imputation in population isolates (Livne et al., 2015). PRIMAL specifically requires pedigree information for phasing and an SSP for imputation. We were unable to successfully setup and run PRIMAL on our simulated datasets and we have been advised by the authors to wait for a new version that is soon to be released.

In this study, we have strived to create realistic isolate data to thoroughly test a range of phasing and imputation software and strategies. Our study design allowed us to observe how phasing and imputation algorithms are impacted by certain characteristics of isolate data, namely IBD between sample members and characteristics arising from isolation such as divergent MAFs compared to reference populations. We found that the best strategy for phasing in a population isolate was to use SHAPEIT2 with the "duohmm" option and with an external reference panel. For imputation, if no SSP is sequenced in the isolate, it is desirable to use the largest public reference panel available. This would lead to the use of

MINIMAC3 or IMPUTE4 as these software can handle very large reference panels. If an SSP is available in the isolate, it should be used and the option in IMPUTE2 that combines reference panels through cross-imputation makes it an attractive choice of imputation software. In this case the largest available public reference panel compatible with IMPUTE2 should be used with the SSP. At the time of publication, IMPUTE4 and MINIMAC3 do not offer the option of combining two reference panels, but, if such options do become available, then a strategy that both combines the HRC and an SSP by cross-imputation would likely be both fast and highly accurate in a population isolate.

## ACKNOWLEDGMENTS

We address special thanks to the people of Campora for their participation in the study. We kindly thank the European Genome-phenome Archive at the European Bioinformatics Institute for making available the UK10K imputation panel (EGAD00001000776) and HRC imputation panel (EGAD00001002729) for the use in our simulation study. We also thank the two anonymous reviewers for their comments that greatly improved the manuscript. ESGI—The research leading to these results has received funding from the Seventh Framework Programme [FP7/2007-2013] under grant agreement no. 262055. A.F.H. was funded by an international Ph.D. fellowship from Sorbonne Paris Cité (convention HERZII5RDXMTSPC1LIETUE).

## CONFLICT OF INTEREST


None declared.

## URLS

1. ALPHAPHASE (v1.2), <https://www.alphagenes.roslin.ed.ac.uk/alphasuite-softwares/alphaphase/>
2. BEAGLE (v4.1), <https://faculty.washington.edu/browning/beagle/beagle.html>
3. EAGLE2 (v2.3.2) and EAGLE1 (v1.0), <https://data.broadinstitute.org/alkesgroup/Eagle/>
4. SHAPEIT2 (v2.837), [https://mathgen.stats.ox.ac.uk/genetics\\_software/shapeit/shapeit.html](https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html)
5. SHAPEIT3 (v1.0), <https://jmarchini.org/shapeit3/>
6. SLRP (v1.0), <https://github.com/kpalin/SLRP>
7. IMPUTE2 (v2.3.2), [https://mathgen.stats.ox.ac.uk/impute/impute\\_v2.html](https://mathgen.stats.ox.ac.uk/impute/impute_v2.html)
8. IMPUTE4 (v1.0), <https://jmarchini.org/impute-4/>
9. MINIMAC3 (v2.0.1), <https://genome.sph.umich.edu/wiki/Minimac3>
10. 1000 Genomes data set (Phase 3), <http://www.internationalgenome.org/>


11. Haplotype Reference Consortium, [www.haplotype-reference-consortium.org/](http://www.haplotype-reference-consortium.org/)
12. UK10K Project, <https://www.uk10k.org/>
13. Sanger Imputation Service, <https://imputation.sanger.ac.uk/>
14. Michigan Imputation Server, <https://imputationserver.sph.umich.edu/>
15. Oxford Statistics Phasing Server, <https://phasingserver.stats.ox.ac.uk/>
16. R-package “Gaston,” <https://cran.r-project.org/web/packages/gaston/index.html>
17. European Genome-phenome Archive, <https://www.ebi.ac.uk/ega/home>

## ORCID

Anthony Francis Herzig 

<http://orcid.org/0000-0001-9392-9924>


Teresa Nutile  <http://orcid.org/0000-0001-7062-8352>

Marie-Claude Babron 

<http://orcid.org/0000-0002-4100-0299>

Marina Ciullo  <http://orcid.org/0000-0002-9621-9984>

Céline Bellenguez  <http://orcid.org/0000-0002-1240-7874>

Anne-Louise Leutenegger 

<http://orcid.org/0000-0002-1302-4357>

## REFERENCES

- Abecasis, G. R., Cherny, S. S., Cookson, W. O., & Cardon, L. R. (2002). Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics*, 30(1), 97–101. <https://doi.org/10.1038/ng786>
- Asimit, J. L., & Zeggini, E. (2012). Imputation of rare variants in next generation association studies. *Human Heredity*, 74(0), 196–204. <https://doi.org/10.1159/000345602>
- Bourgain, C., & Génin, E. (2005). Complex trait mapping in isolated populations: Are specific statistical methods required? *European Journal of Human Genetics*, 13(6), 698–706. <https://doi.org/10.1038/sj.ejhg.5201400>
- Browning, B. L., & Browning, S. R. (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *American Journal of Human Genetics*, 84(2), 210–223. <https://doi.org/10.1016/j.ajhg.2009.01.005>
- Browning, Brian L., & Browning, Sharon R. (2016). Genotype imputation with millions of reference samples. *American Journal of Human Genetics*, 98(1), 116–126. <https://doi.org/10.1016/j.ajhg.2015.11.020>
- Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., ... Marchini, J. (2017). Genome-wide genetic data on ~500,000 UK Biobank participants. bioRxiv, 166298, 1–28. <https://doi.org/10.1101/166298>
- Chen, W., & Schaid, D. J. (2014). PedBLIMP: Extending linear predictors to impute genotypes in pedigrees. *Genetic Epidemiology*, 38(6), 531–541. <https://doi.org/10.1002/gepi.21838>

- Cheung, C. Y., Thompson, E. A., & Wijsman, E. M. (2013). GIGI: An approach to effective imputation of dense genotypes on large pedigrees. *American Journal of Human Genetics*, *92*(4), 504–516. <https://doi.org/10.1016/j.ajhg.2013.02.011>
- Colonna, V., Nutile, T., Astore, M., Guardiola, O., Antoniol, G., Ciullo, M., & Persico, M. G. (2007). Campora: A young genetic isolate in South Italy. *Human Heredity*, *64*(2), 123–135. <https://doi.org/10.1159/000101964>
- Das, S., Forer, L., Schonherr, S., Sidore, C., Locke, A. E., Kwong, A., ... Fuchsberger, C. (2016). Next-generation genotype imputation service and methods. *Nature Genetics*, *48*(10), 1284–1287. <https://doi.org/10.1038/ng.3656>
- Deelen, P., Menelaou, A., van Leeuwen, E. M., Kanterakis, A., van Dijk, F., Medina-Gomez, C., ... Swertz, M. A. (2014). Improved imputation quality of low-frequency and rare variants in European samples using the “Genome of The Netherlands.” *European Journal of Human Genetics*, *22*(11), 1321–1326. <https://doi.org/10.1038/ejhg.2014.19>
- Delaneau, O., Howie, B., Cox, A. J., Zagury, J.-F., & Marchini, J. (2013). Haplotype estimation using sequencing reads. *American Journal of Human Genetics*, *93*(4), 687–696. <https://doi.org/10.1016/j.ajhg.2013.09.002>
- Delaneau, O., Zagury, J.-F., & Marchini, J. (2013). Improved whole-chromosome phasing for disease and population genetic studies. *Nature Methods*, *10*(1), 5–6. <https://doi.org/10.1038/nmeth.2307>
- Durbin, R. (2014). Efficient haplotype matching and storage using the positional Burrows–Wheeler transform (PBWT). *Bioinformatics*, *30*(9), 1266–1272. <https://doi.org/10.1093/bioinformatics/btu014>
- Gazal, S., Sahbatou, M., Perdry, H., Letort, S., Génin, E., & Leutenegger, A. L. (2014). Inbreeding coefficient estimation with dense SNP data: Comparison of strategies and application to HapMap III. *Human Heredity*, *77*(1–4), 49–62. <https://doi.org/10.1159/000358224>
- Glodzik, D., Navarro, P., Vitart, V., Hayward, C., McQuillan, R., Wild, S. H., ... McKeigue, P. (2013). Inference of identity by descent in population isolates and optimal sequencing studies. *European Journal of Human Genetics*, *21*(10), 1140–1145. <https://doi.org/10.1038/ejhg.2012.307>
- Hatzikotoulas, K., Gilly, A., & Zeggini, E. (2014). Using population isolates in genetic association studies. *Briefings in Functional Genomics*, *13*(5), 371–377. <https://doi.org/10.1093/bfpg/elu022>
- Hickey, J. M., Kinghorn, B. P., Tier, B., Wilson, J. F., Dunstan, N., & van der Werf, J. H. J. (2011). A combined long-range phasing and long haplotype imputation method to impute phase for SNP genotypes. *Genetics, Selection, Evolution: GSE*, *43*(1), 12, 1–13. <https://doi.org/10.1186/1297-9686-43-12>
- Holm, H., Gudbjartsson, D. F., Sulem, P., Masson, G., Helgadóttir, H. T., Zanon, C., ... Stefansson, K. (2011). A rare variant in MYH6 is associated with high risk of sick sinus syndrome. *Nature Genetics*, *43*(4), 316–320. <https://doi.org/10.1038/ng.781>
- Howie, B., Donnelly, P., & Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics*, *5*(6), e1000529, 1–15. <https://doi.org/10.1371/journal.pgen.1000529>
- Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., & Abecasis, G. R. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature Genetics*, *44*(8), 955–959. <https://doi.org/10.1038/ng.2354>
- Howie, B., Marchini, J., & Stephens, M. (2011). Genotype imputation with thousands of genomes. *G3 (Bethesda)*, *1*(6), 457–470. <https://doi.org/10.1534/g3.111.001198>
- Huang, L., Wang, C., & Rosenberg, N. A. (2009). The relationship between imputation error and statistical power in genetic association studies in diverse populations. *American Journal of Human Genetics*, *85*(5), 692–698. <https://doi.org/10.1016/j.ajhg.2009.09.017>
- Joshi, P. K., Prendergast, J., Fraser, R. M., Huffman, J. E., Vitart, V., Hayward, C., ... Navarro, P. (2013). Local exome sequences facilitate imputation of less common variants and increase power of genome wide association studies. *PLoS One*, *8*(7), e68604, 1–6. <https://doi.org/10.1371/journal.pone.0068604>
- Kim, S. Y., Lohmueller, K. E., Albrechtsen, A., Li, Y., Korneliusen, T., Tian, G., ... Nielsen, R. (2011). Estimation of allele frequency and association mapping using next-generation sequencing data. *BMC Bioinformatics*, *12*, 231, 1–16. <https://doi.org/10.1186/1471-2105-12-231>
- Kong, A., Masson, G., Frigge, M. L., Gylfason, A., Zusmanovich, P., Thorleifsson, G., ... Stefansson, K. (2008). Detection of sharing by descent, long-range phasing and haplotype imputation. *Nature Genetics*, *40*(9), 1068–1075. <https://doi.org/10.1038/ng.216>
- Li, Y., Willer, C. J., Ding, J., Scheet, P., & Abecasis, G. R. (2010). MaCH: Using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology*, *34*(8), 816–834. <https://doi.org/10.1002/gepi.20533>
- Livne, O. E., Han, L., Alkorta-Aranburu, G., Wentworth-Sheilds, W., Abney, M., Ober, C., & Nicolae, D. L. (2015). PRIMAL: Fast and accurate Pedigree-based imputation from sequence data in a founder population. *PLoS Computational Biology*, *11*(3), e1004139, 1–14. <https://doi.org/10.1371/journal.pcbi.1004139>
- Loh, P.-R., Danecek, P., Palamara, P. F., Fuchsberger, C., A Reshef, Y., K Finucane, H., ... L Price, A. (2016). Reference-based phasing using the Haplotype Reference Consortium panel. *Nature Genetics*, *48*(11), 1443–1448. <https://doi.org/10.1038/ng.3679>
- Loh, P.-R., Palamara, P. F., & Price, A. L. (2016). Fast and accurate long-range phasing in a UK Biobank cohort. *Nature Genetics*, *48*(7), 811–816. <https://doi.org/10.1038/ng.3571>
- Marchini, J., & Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, *11*(7), 499–511. <https://doi.org/10.1038/nrg2796>
- McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., & Teumer, A. (2016). ... The Haplotype Reference ConsortiumA reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics*, *48*(10), 1279–1283. <https://doi.org/10.1038/ng.3643>
- Mitt, M., Kals, M., Parn, K., Gabriel, S. B., Lander, E. S., Palotie, A., ... Palta, P. (2017). Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel. *European Journal of Human Genetics*, *25*(7), 869–876. <https://doi.org/10.1038/ejhg.2017.51>
- O’Connell, J., Gurdasani, D., Delaneau, O., Pirastu, N., Ulivi, S., Cocca, M., ... Marchini, J. (2014). A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS Genetics*, *10*(4), e1004234, 1–21. <https://doi.org/10.1371/journal.pgen.1004234>

- O'Connell, J., Sharp, K., Shrine, N., Wain, L., Hall, I., Tobin, M., ... Marchini, J. (2016). Haplotype estimation for biobank-scale data sets. *Nature Genetics*, *48*(7), 817–820. <https://doi.org/10.1038/ng.3583>
- Palin, K., Campbell, H., Wright, A. F., Wilson, J. F., & Durbin, R. (2011). Identity-by-descent-based phasing and imputation in founder populations using graphical models. *Genetic Epidemiology*, *35*(8), 853–860. <https://doi.org/10.1002/gepi.20635>
- Pistis, G., Porcu, E., Vrieze, S. I., Sidore, C., Steri, M., Danjou, F., ... Sanna, S. (2015). Rare variant genotype imputation with thousands of study-specific whole-genome sequences: Implications for cost-effective study designs. *European Journal of Human Genetics*, *23*(7), 975–983. <https://doi.org/10.1038/ejhg.2014.216>
- Roshyara, N. R., Horn, K., Kirsten, H., Ahnert, P., & Scholz, M. (2016). Comparing performance of modern genotype imputation methods in different ethnicities. *Scientific Reports*, *6*, 34386, 1–12. <https://doi.org/10.1038/srep34386>
- Roshyara, N. R., & Scholz, M. (2015). Impact of genetic similarity on imputation accuracy. *BMC Genetics*, *16*, 90, 1–16. <https://doi.org/10.1186/s12863-015-0248-2>
- Sharp, K., Kretschmar, W., Delaneau, O., & Marchini, J. (2016). Phasing for medical sequencing using rare variants and large haplotype reference panels. *Bioinformatics*, *32*(13), 1974–1980. <https://doi.org/10.1093/bioinformatics/btw065>
- Su, Z., Marchini, J., & Donnelly, P. (2011). HAPGEN2: Simulation of multiple disease SNPs. *Bioinformatics*, *27*(16), 2304–2305. <https://doi.org/10.1093/bioinformatics/btr341>
- Surakka, I., Kristiansson, K., Anttila, V., Inouye, M., Barnes, C., Moutsianas, L., ... Ripatti, S. (2010). Founder population-specific HapMap panel increases power in GWA studies through improved imputation accuracy and CNV tagging. *Genome Research*, *20*(10), 1344–1351. <https://doi.org/10.1101/gr.106534.110>
- The 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. <https://doi.org/10.1038/nature15393>
- The UK10K Consortium. (2015). The UK10K project identifies rare variants in health and disease. *Nature*, *526*(7571), 82–90. <https://doi.org/10.1038/nature14962>
- Uricchio, L. H., Chong, J. X., Ross, K. D., Ober, C., & Nicolae, D. L. (2012). Accurate imputation of rare and common variants in a founder population from a small number of sequenced individuals. *Genetic Epidemiology*, *36*(4), 312–319. <https://doi.org/10.1002/gepi.21623>
- Vieira, F. G., Albrechtsen, A., & Nielsen, R. (2016). Estimating IBD tracts from low coverage NGS data. *Bioinformatics*, *32*(14), 2096–2102. <https://doi.org/10.1093/bioinformatics/btw212>
- Wijsman, E. M., Rothstein, J. H., & Thompson, E. A. (2006). Multi-point linkage analysis with many multiallelic or dense diallelic markers: Markov Chain–Monte Carlo provides practical approaches for genome scans on general Pedigrees. *American Journal of Human Genetics*, *79*(5), 846–858. <https://doi.org/10.1086/508472>
- Zeggini, E. (2011). Next-generation association studies for complex traits. *Nature Genetics*, *43*(4), 287–288. <https://doi.org/10.1038/ng0411-287>

## SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

**How to cite this article:** Herzig AF, Nutile T, Babron MC, Ciullo M, Bellenguez C, Leutenegger AL. Strategies for phasing and imputation in a population isolate. *Genet Epidemiol.* 2018;1–13. <https://doi.org/10.1002/gepi.22109>

## Strategies for Phasing and Imputation in a Population Isolate

### Supplementary Material

Anthony Francis Herzig (1,2)

Teresa Nutile (3)

Marie-Claude Babron (1,2)

Marina Ciullo (3,4)

Céline Bellenguez (5,6,7,8)

Anne-Louise Leutenegger (1,2,8)

(1) Université Paris-Diderot, Sorbonne Paris Cité, U946, F-75010 Paris, France

(2) Inserm, U946, Genetic variation and Human diseases, F-75010 Paris, France

(3) Institute of Genetics and Biophysics A. Buzzati-Traverso - CNR, Naples, Italy

(4) IRCCS Neuromed, Pozzilli, Isernia, Italy

(5) Inserm, U1167, RID-AGE - Risk factors and molecular determinants of aging-related diseases, F-59000 Lille, France

(6) Institut Pasteur de Lille, F-59000 Lille, France

(7) Université de Lille, U1167 - Excellence Laboratory LabEx DISTALZ, F-59000 Lille, France

(8) These authors contributed equally to this study

### Correspondence:

Anthony Francis Herzig

INSERM UMR 946

27 Rue Juliette Dodu

75010, PARIS, FRANCE.

anthony.herzig@inserm.fr

+33172639313

## Supplementary Materials

### Data Simulation

Founder haplotypes were created with HapGen using a theoretical population size of 3,000 and the default mutation rate. This choice led to simulated data with similar kinship to the observed genotypes in Campora (Supplementary Figure 3). The percentage of sites phased by SLRP serves as a good proxy for the proportion of IBD that can be found within the sample. When phasing the true data from Campora, 99% of heterozygous sites were phased, a similar percentage to those observed on the HapGen+Pedigree simulation, Supplementary Figure 8a.

### Error Models.

Here we describe our error model for the WGS data of the 93 SSP individuals. For each simulated genotype, a set of bases was sampled from the two possible alleles of the genotype in order to represent the bases across multiple reads containing the position. Error bases are simulated within this set and can take any value out of A,C,T and G. The depth of the set was randomly selected from the depths observed in the Campora WGS data at the corresponding position. We then used the approach of Kim et al. (2011) to make approximate calculations of genotype likelihoods from which we calculated genotype qualities based on the models implemented by the next-generation sequence calling software GATK (DePristo et al., 2011).

As in to Kim et al. (2011), our error models were not symmetric. Lower genotype quality was observed in Campora on AT and GT SNPs. Hence we simulated higher error rates for error types  $A \rightarrow T$ ,  $T \rightarrow A$ ,  $G \rightarrow T$ , and  $T \rightarrow G$  as shown in Supplementary Table 1. For all genotypes, the error rate from the true base to the base corresponding to the other possible allele at the position (according to the simulated genotype) was augmented by 1/120. For example, when simulating error bases for a true base of A at a position with alleles A and G present in the simulated data, the error rate  $A \rightarrow G$  was  $1/120 + 1/120$ . However, if the alleles present in the data were A and T, the error rate  $A \rightarrow G$  would remain at 1/120 and the error rate  $A \rightarrow T$  would be augmented by 1/120. Such error models were chosen in order to create similar distributions of genotype quality as had been observed in Campora and overall genotyping error rates high enough to be of interest when analysing their effect on phasing and imputation. Our error models do not attempt to provide a faithful representation of the calling of genotypes from raw sequence reads but simply to create errors and missingness simply, randomly, and in similar patterns to those observed in WGS data in Campora.

### Quality Control

ARRAY variants were removed for high missingness ( $> 5\%$ ), low MAF ( $< 1\%$ ), and significant deviation from Hardy-Weinberg equilibrium ( $p < 10^{-5}$ ). WGS variants were removed for high missingness ( $> 10\%$ ), low Minor Allele Count ( $< 2$ ), and significant deviation from Hardy-Weinberg equilibrium ( $p < 10^{-5}$ ).

### **Phasing**

When using EAGLE2+1000G we set the parameter 'pbwtiter' to 3 which significantly improved the phasing and ensures that phasing inference was made not just from the 1000G but from estimated haplotypes of other individuals in the sample. When phasing with BEAGLE we found that results were generally robust to changes in the 'window' and 'overlap' parameters. For EAGLE2 and BEAGLE we allowed multiple threading (four threads) after observing that restricting to one thread did not significantly change results. However, for SHAPEIT2 it is recommended to not use multiple threading so we used a single thread for each phasing run. Using BEAGLE with the 1000G as an external reference panel proved problematic as many variants were removed from the analysis by the algorithm due to high differences in MAFs between the sample and the reference panel. As population isolate data has been simulated, it is to be expected that MAFs differ from those observed in 1000G. We thus did not present BEAGLE results for this option. We did not test SHAPEIT3 with the 1000G as a reference panel due to the similarity between SHAPEIT2 and SHAPEIT3. Otherwise software were used with default settings.

Note that it is not possible to calculate SER at the exact site of a genotype error or missing genotype as there is no true phase from which to make a comparison with. Hence all calculations are made irrespective of error sites in each replication. An error can still cause a SER in a direct way but this would be measured at the preceding and following heterozygous sites on the chromosome (Supplementary Figure 10).

### **IBD-Sharing**

To create Supplementary Figure 11 for SHAPEIT2+duohmm+1000G and EAGLE2 we randomly selected 200 heterozygous sites with switch errors and 200 heterozygous sites which were phased correctly. For each of the 400 sites we then counted the number of haplotypes in the sample IBD to the individual at the site. This analysis was performed on ARRAY data with no genotyping errors or missingness simulated and on the Pedigree simulation where the exact IBD sharing information was accessible. This is because on the Pedigree simulation, we know exactly which founder haplotype has been copied at every site for each individual. On the HapGen+Pedigree simulation, we could not keep track of this information.

### **Imputation**

Following user manual recommendations for IMPUTE2, the region was split into four regions each of width 5Mb and using a buffer region of 0.25Mb. Identical settings were used for IMPUTE4. Outputs from the four runs were then concatenated after imputation. We experimented with the 'k-haps' parameter in IMPUTE2 and were unable to observe significant changes in accuracy and so the default parameter was used. MINIMAC3 was run with default parameters as was BEAGLE as we found that results were not sensitive to changes in the 'window', 'overlap' and 'Ne' (effective population size) parameters. A detailed investigation on the effects of model parameters on IMPUTE2, MINIMAC3 and BEAGLE is to be found in Browning and Browning (2016). As population isolates are the subject of this investigation it might be suggested that a lower value of 'Ne' would theoretically be suitable. In the context of imputation, the 'Ne' parameter controls the expected rate of recombination. Whilst our simulated individuals were constructed as mosaics of founder haplotypes with relatively few recombinations, a high recombination rate is still required in order to model each individual as an imperfect mosaic of external reference haplotypes. For IMPUTE2 we took advantage of the 'merge-ref-panel' option to perform cross imputation between the 1000G and our WGS SSP.

#### **Difference in MAF between sample and reference panel**

We compared the absolute difference in MAF between the simulated data in each replicate and either the 1000G Europeans populations or the complete 1000G. We averaged these differences over all variants used to estimate imputation accuracy and compared them to the baseline mean difference in MAF between the UK10K (our source of founding haplotypes) and the 1000G (Supplementary Figure 15). Compared to the Pedigree simulation, the HapGen+Pedigree simulation strategy produced simulated data with greater disparity in MAF compared to the 1000G. It was possible to observe a pattern between this disparity in MAF and lower imputation accuracy (Supplementary Figure 16a). To illustrate the importance of this difference in MAF we selected variants with a high MAF in our simulated data set ( $>0.3$ ) and a large difference in MAF compared to the 1000G (top 10% of all MAF differences). We also excluded variants with imputation quality score ('info') below 0.7 coming from imputation using IMPUTE2+1000G. In the Pedigree simulation an average of 2,340 variants fulfilled these criteria and the average 90th percentile of absolute MAF differences was 0.17. In the HapGen+Pedigree simulation there were an average of 2,166 variants and the average 90th percentile of absolute MAF differences was 0.20. We compared the mean imputation accuracy from imputation using IMPUTE2+1000G and IMPUTE2+SSP over this selection of variants to the mean imputation accuracy over a random selection of variants with similar MAFs (Supplementary Figure 16a). In both simulation strategies



variants with a large difference in MAF to the 1000G were harder to impute under IMPUTE2+1000G but were imputed with similar accuracy under IMPUTE2+SSP.

To investigate further, we also selected variants with either a MAF significantly higher in the 1000G reference panel than in the sample or vice-versa. We then calculated the percentage increase in imputation accuracy by changing reference panel from the 1000G reference panel to the SSP (Supplementary Figure 16b). To put these increases into context, we again selected random selections of variants with a similar MAF to the chosen variants but without the large differences in MAF between the sample and the 1000G. Variants with significantly higher MAF in the sample (compared to the 1000G reference panel) experienced the most benefit from the change of reference panel for imputation.

A final analysis was made on variants which were monomorphic in the sample. Such variants may represent the greatest difference in MAF between the sample and an external reference panel. We have compared imputation accuracy on the telomeric region of the short arm of chr10 (20Mb in length). In this region 102,100 variants (found in the UK10K) were simulated and 22% and 31% of these variants became monomorphic in the Pedigree and HapGen+Pedigree simulation strategies respectively due to the founder effects that we simulated. From each replicate of each strategy, we selected 100 monomorphic variants at random. From this selection, an average of 18% and 19% of the variants passed a 0.4 threshold on the IMPUTE2 imputation quality score 'info'. For each variant that passed the threshold, we called imputed genotypes from imputed dosages by assigning the genotype to the highest genotype likelihood if and only if one genotype likelihood exceeded 0.8. In Supplementary Figure 16c we present the number of individuals with an incorrect called genotype for the assembly of all variants considered across replicates. A few variants present extreme results, these were noted to be variants with extremely different MAF between the UK10K and the 1000G. For example, the highest point on the left panel of Supplementary Figure 16c has a MAF of 0.47 in the 1000G and 0.0026 in the UK10K. Supplementary Figure 16d shows a zoom-in on Supplementary Figure 16c.

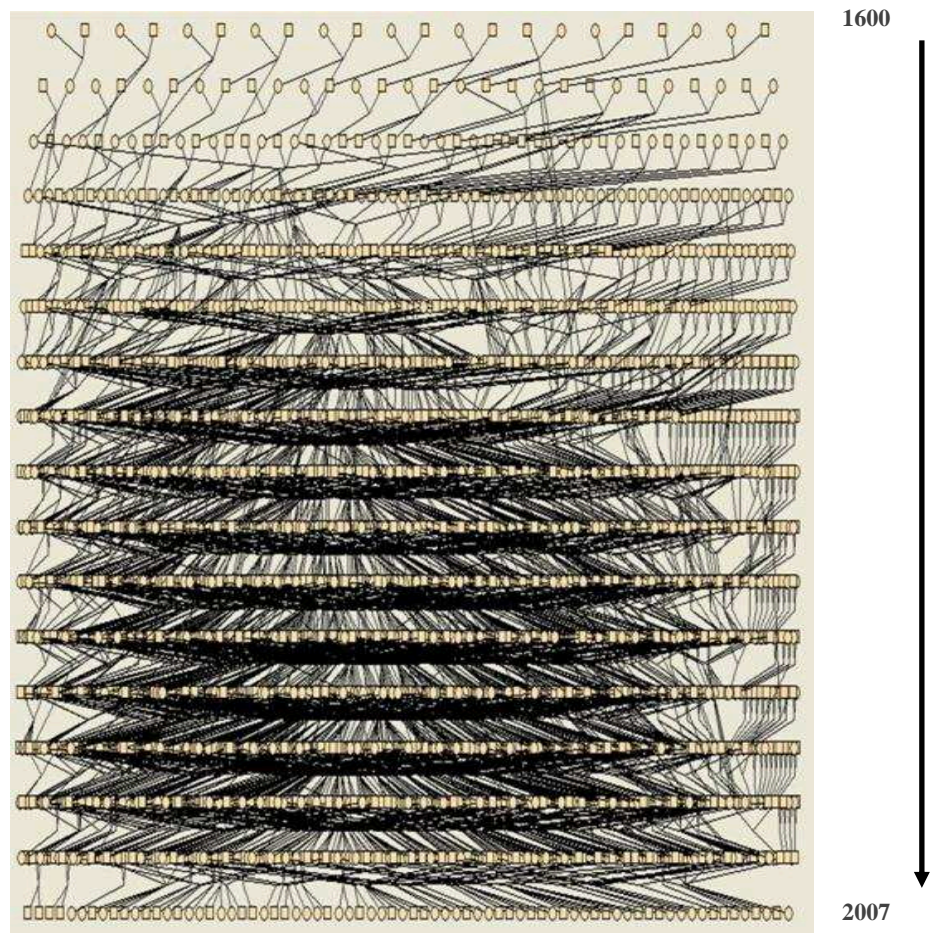
Genetic kinship coefficients between all 477 simulated individuals and the 1000G Europeans were computed using WGS positions after LD-pruning. The mean pairwise kinship on the Pedigree simulation was  $1.2 \times 10^{-4}$  and  $2.5 \times 10^{-6}$  on the HapGen+Pedigree simulation. Again this demonstrated greater dissimilarity between the HapGen+Pedigree simulated data and 1000G than between the Pedigree simulated data and 1000G.

#### **Imputation Quality scores: 'info' and 'RSQ'**

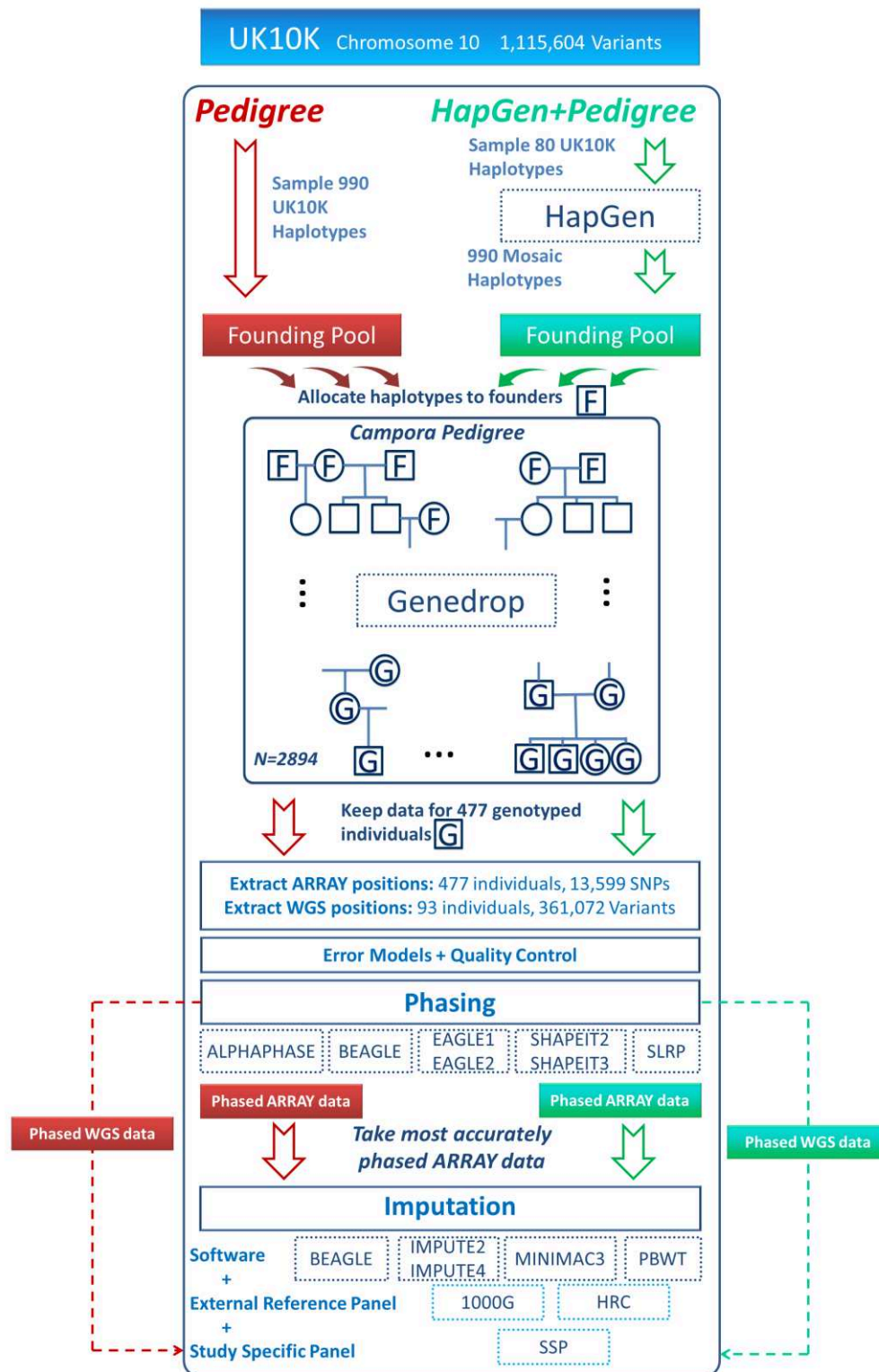
First, we applied the standard thresholds for common variants of 0.4 for 'info' and 0.3 for 'RSQ' (Li, Willer, Ding, Scheet, & Abecasis, 2010; Pistis et al., 2015) (Supplementary Figure 18a).

We then specified different thresholds for ‘info’ and ‘RSQ’ and calculated the resulting mean imputation accuracy in the remaining variants (Supplementary Figure 18b) under MINIMAC3+1000G and IMPUTE2+1000G on the HapGen+Pedigree simulation strategy. We observed that increasing the thresholds continued to give gains in mean imputation accuracy at a price of removing large numbers of variants. Particularly for low MAF variants, greater increases in imputation accuracy were observed by placing thresholds on the ‘RSQ’ measure than ‘info’. Furthermore, the mean imputation accuracy of remaining variants became almost equal across all MAF bins when using the ‘RSQ’ measure while greater differences remain between MAF bins when using the ‘info’ score.

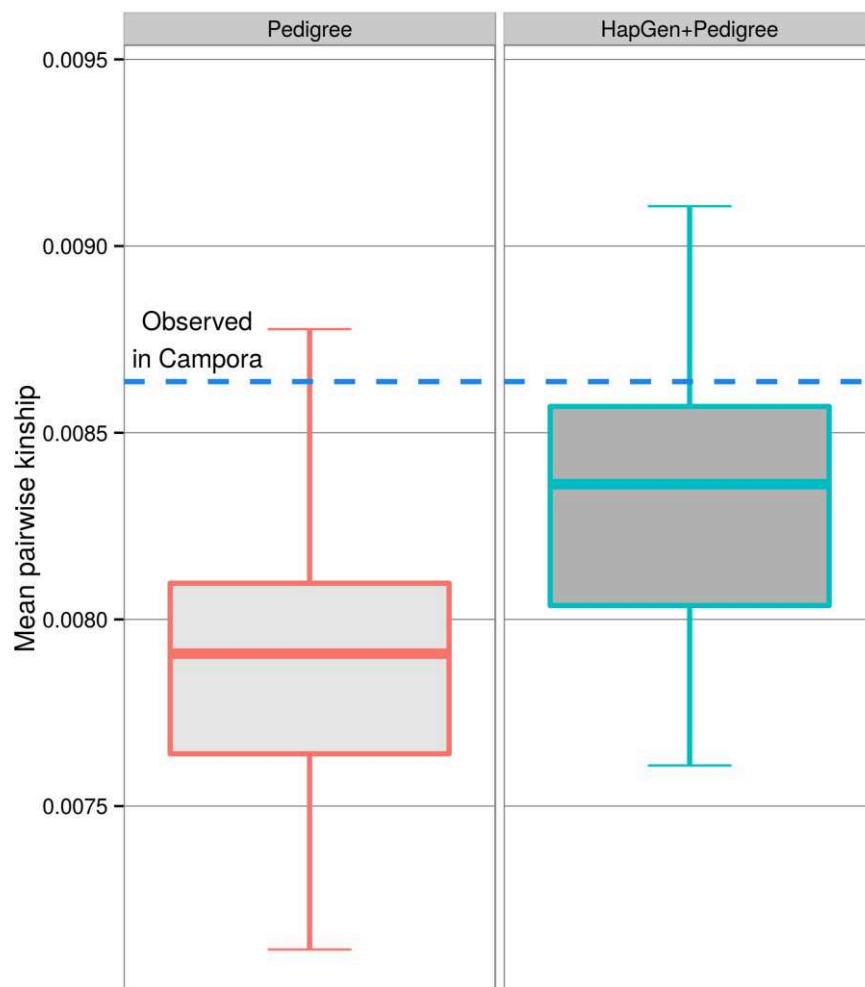
By defining sets of well and poorly imputed variants (imputation accuracy above 0.5 or below 0.2) we observed that the standard thresholds of 0.4 for ‘info’ and 0.3 for ‘RSQ’ fail to remove many poorly imputed variants (Supplementary Table 4). Furthermore for rare variants, the quality scores have less ability to separate well imputed variants from poorly imputed variants as reported by others (Liu et al., 2012; Pistis et al., 2015). To ensure that the majority of poorly imputed low MAF variants will be removed, higher thresholds than the standard ones are required. For common variants we observed that similarly high thresholds could be used and only a small number of well imputed variants would be lost and more poorly imputed variants would be removed. The choice of threshold represents a compromise between attempting to remove all badly imputed variants while hoping to not discard too many well imputed variants that could be highly valuable to subsequent analyses. Poorly imputed variants could give false positive results. However, if the motivation for imputation was envisaged single point analyses, the damage would be minimal as the researcher could still access the ‘info’ or ‘RSQ’ scores in order to see whether significantly associated variants had a very high imputation quality score or one just above the threshold. If multipoint analyses (gene-based or haplotype based) were envisaged, then poorly imputed variants have the potential to cause false negative results which would be harder to rectify; suggesting that in this scenario higher thresholds should be taken.



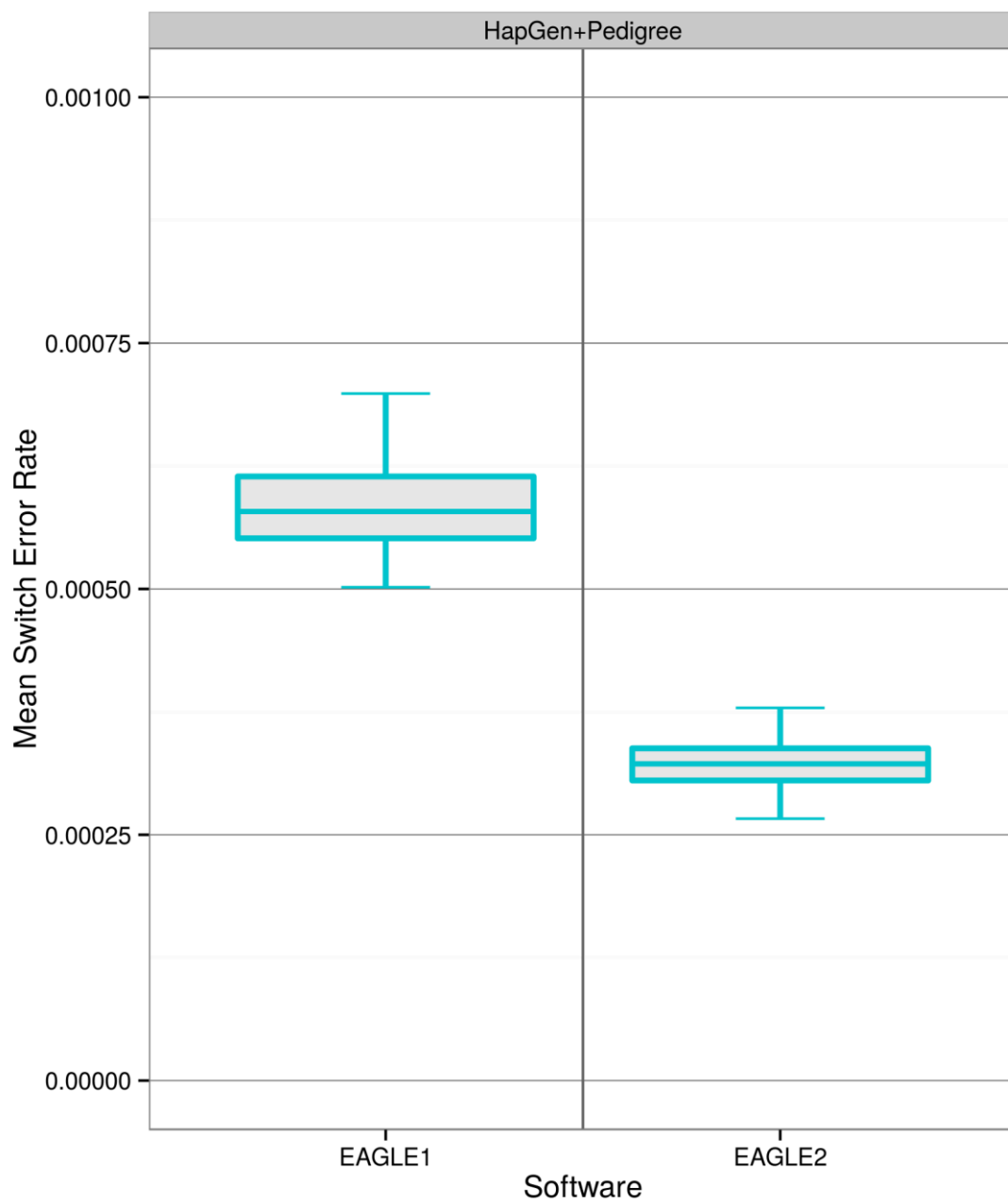
**Supplementary Figure 1.** The pedigree of Campora as recorded from parish records.



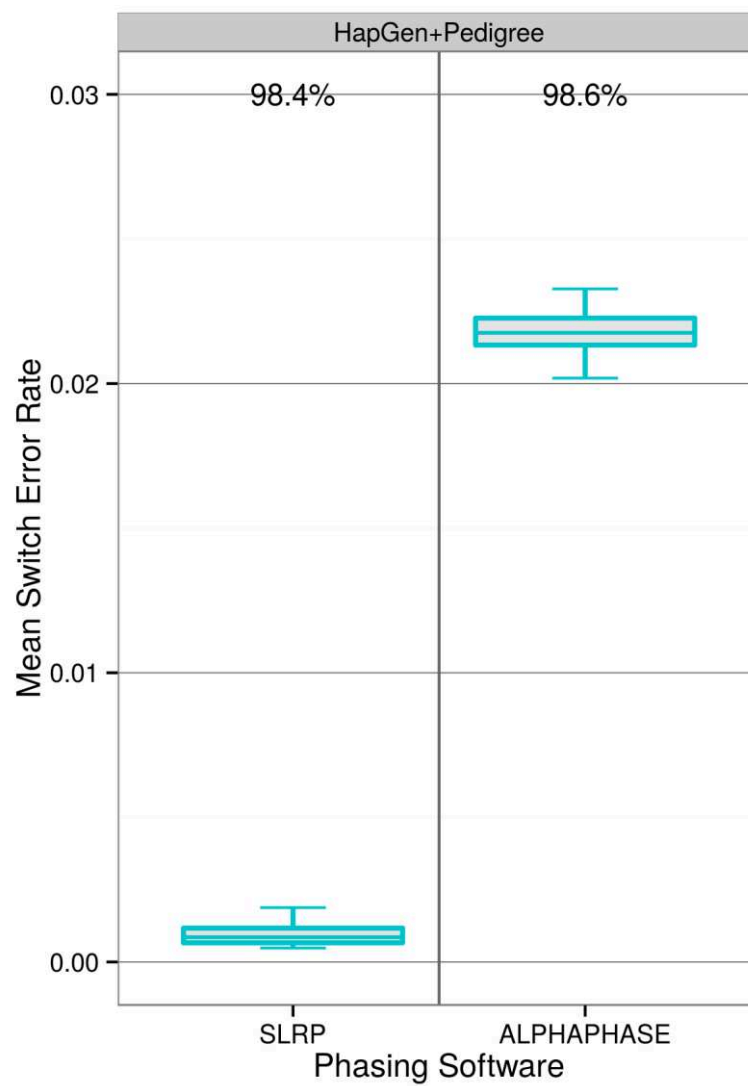
**Supplementary Figure 2.** Schematic of the two simulation strategies.



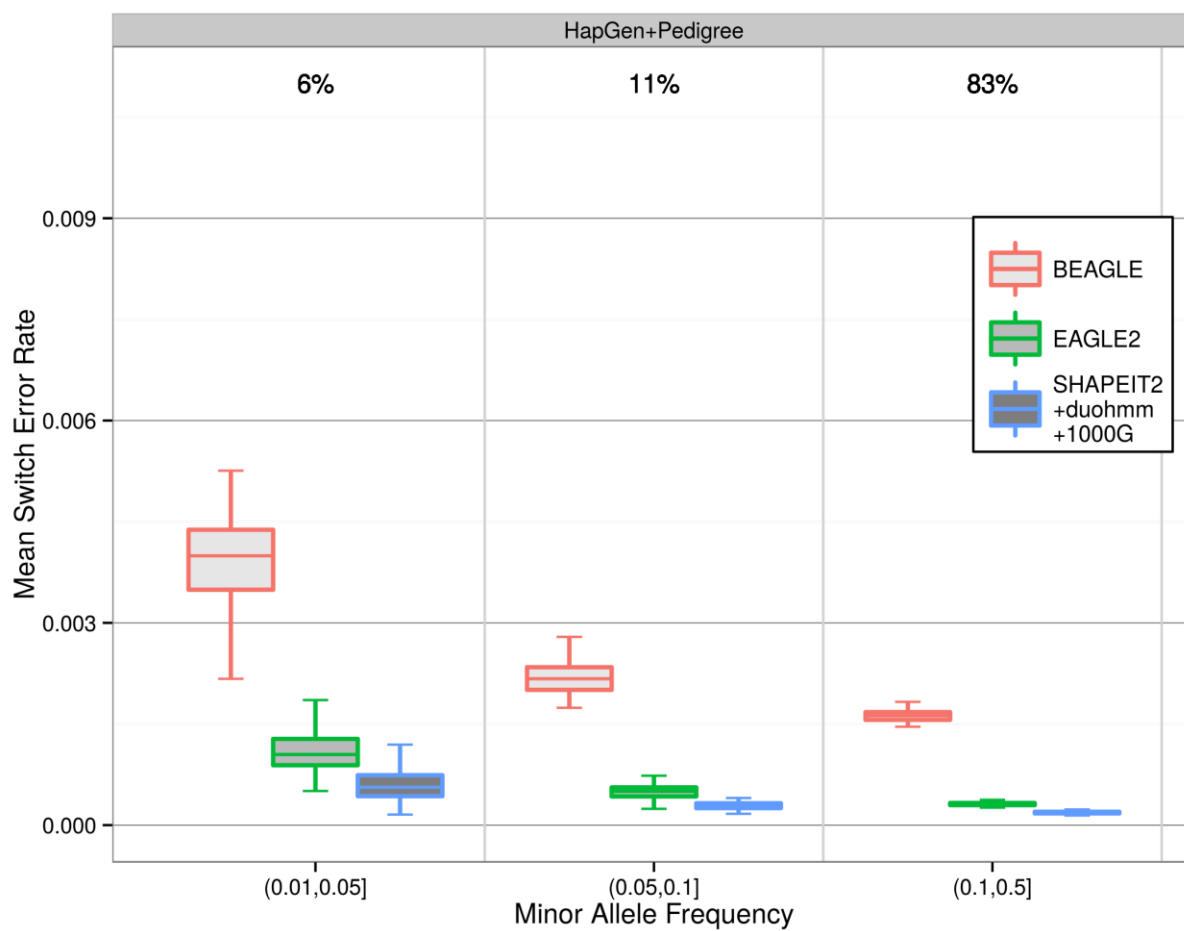
**Supplementary Figure 3.** Comparison of mean pairwise genetic kinship coefficients estimated on simulated ARRAY data for 477 individuals for both simulation strategies. The HapGen+Pedigree simulation created data with closer mean pairwise genetic kinship to the mean pairwise genetic kinship calculated on the observed genotypes in Campora for the same individuals (dashed line). As every pedigree founder haplotype is first generated from 80 UK10K haplotypes in the HapGen+Pedigree simulation, the pedigree founders are no longer independent and share regions of IBD. Proportions of IBD are consequentially elevated throughout the sample and surpass those predicted solely by the pedigree information.



**Supplementary Figure 4.** Mean Switch Error Rates for EAGLE1 and EAGLE2 on the HapGen+Pedigree simulation strategy.

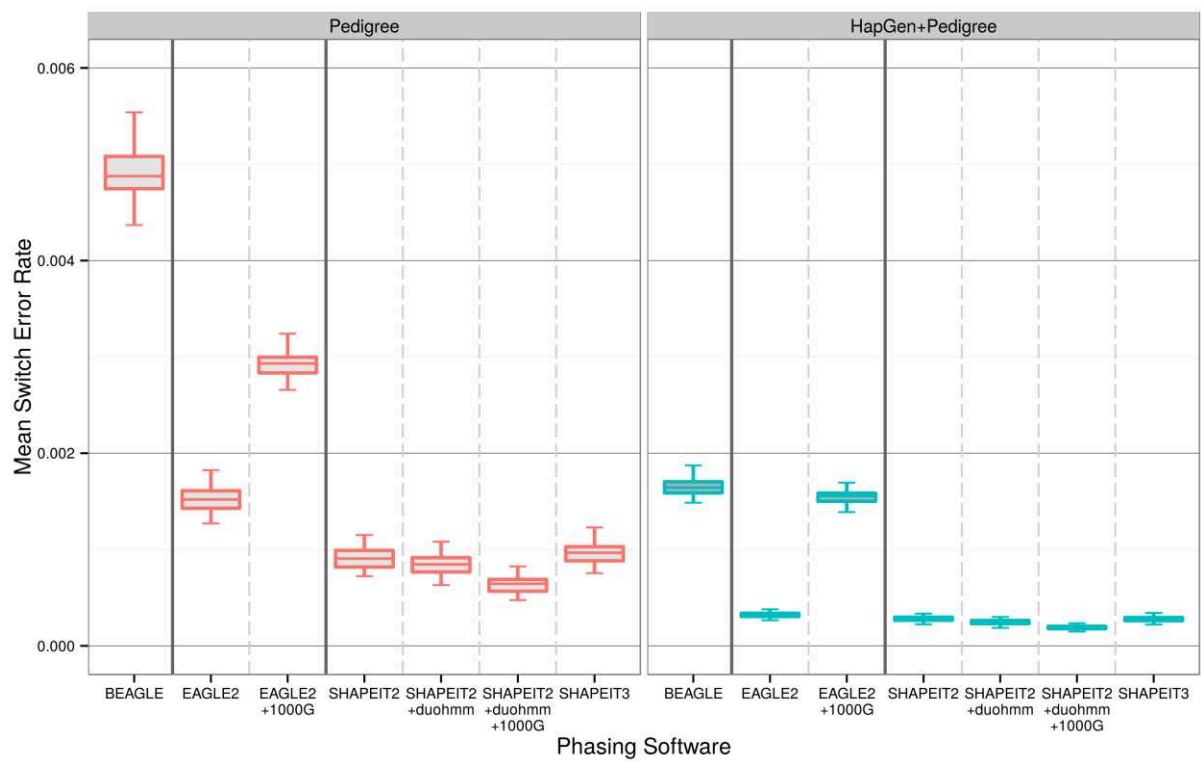


**Supplementary Figure 5.** Comparison of Long Range Phasing Software SLRP and ALPHAPHASE on the HapGen+Pedigree simulation strategy. The percentages of heterozygous sites phased are displayed atop the figure.

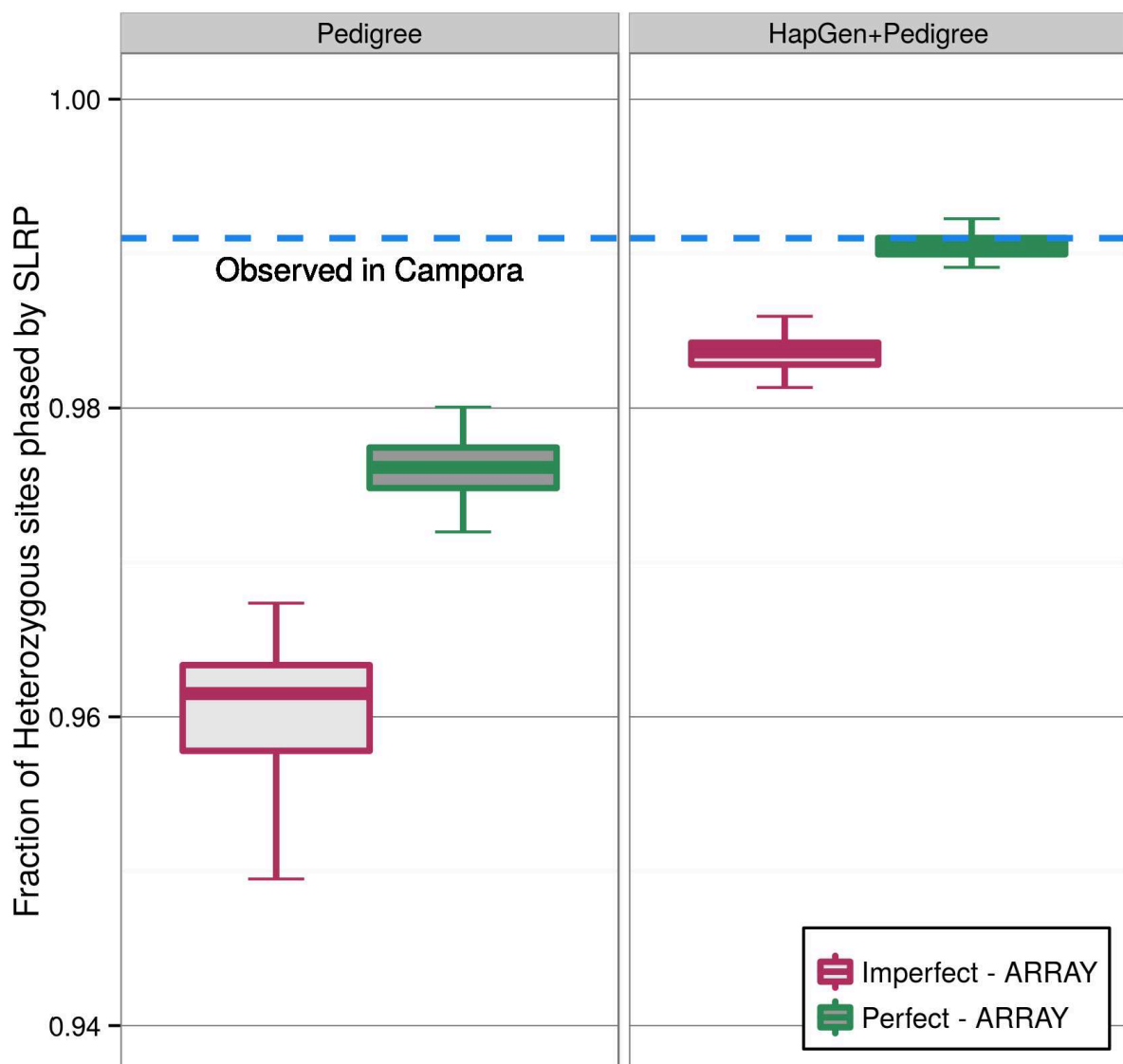


**Supplementary Figure 6.** Comparison of SERs according to MAF for BEAGLE, EAGLE2 and SHAPEIT2+duohmm+1000G on the HapGen+Pedigree simulation strategy. In each MAF bin, the mean SER over all variants is displayed. The percentages of variants in each MAF bin are displayed atop the figure.

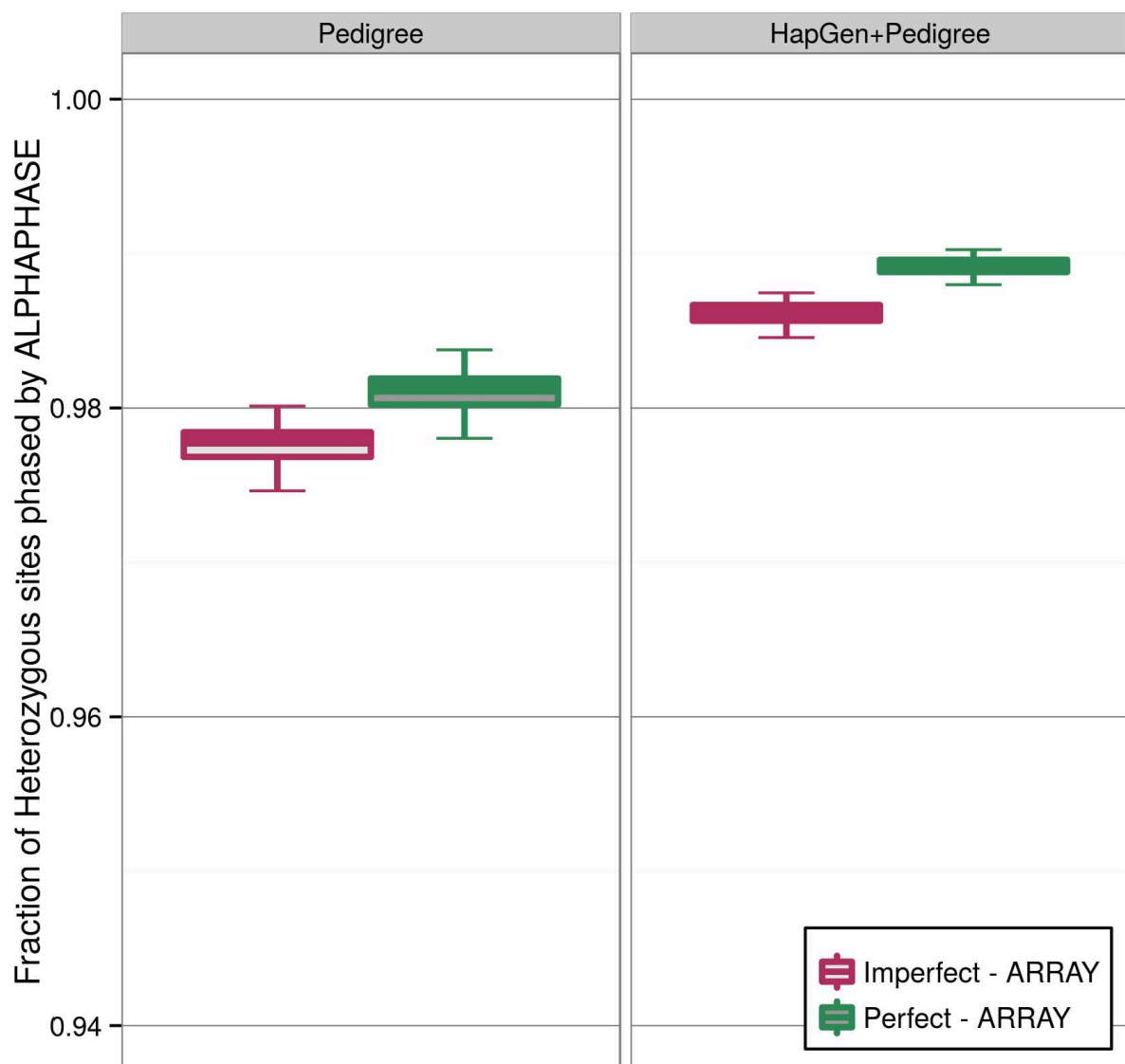




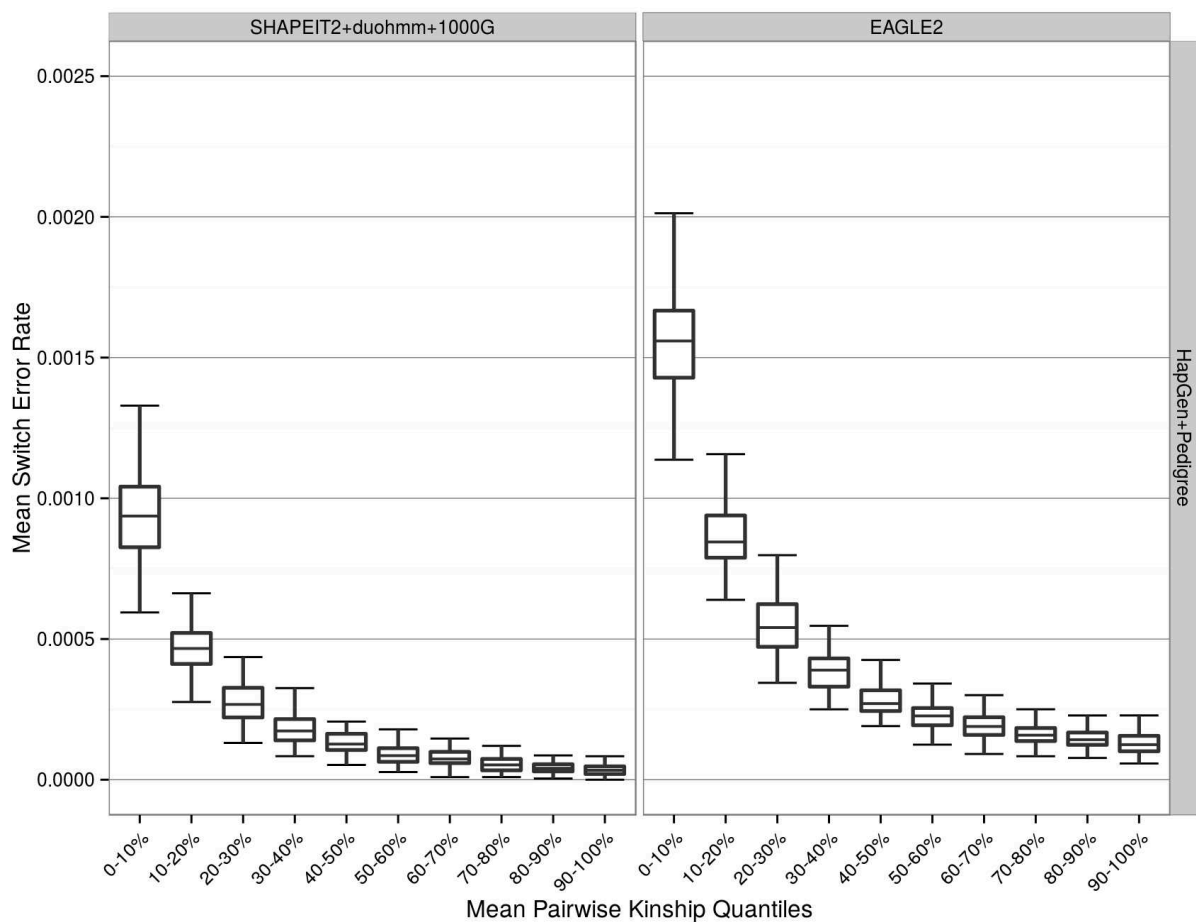
**Supplementary Figure 7.** Global SERs from both simulation strategies for all LD-based software.



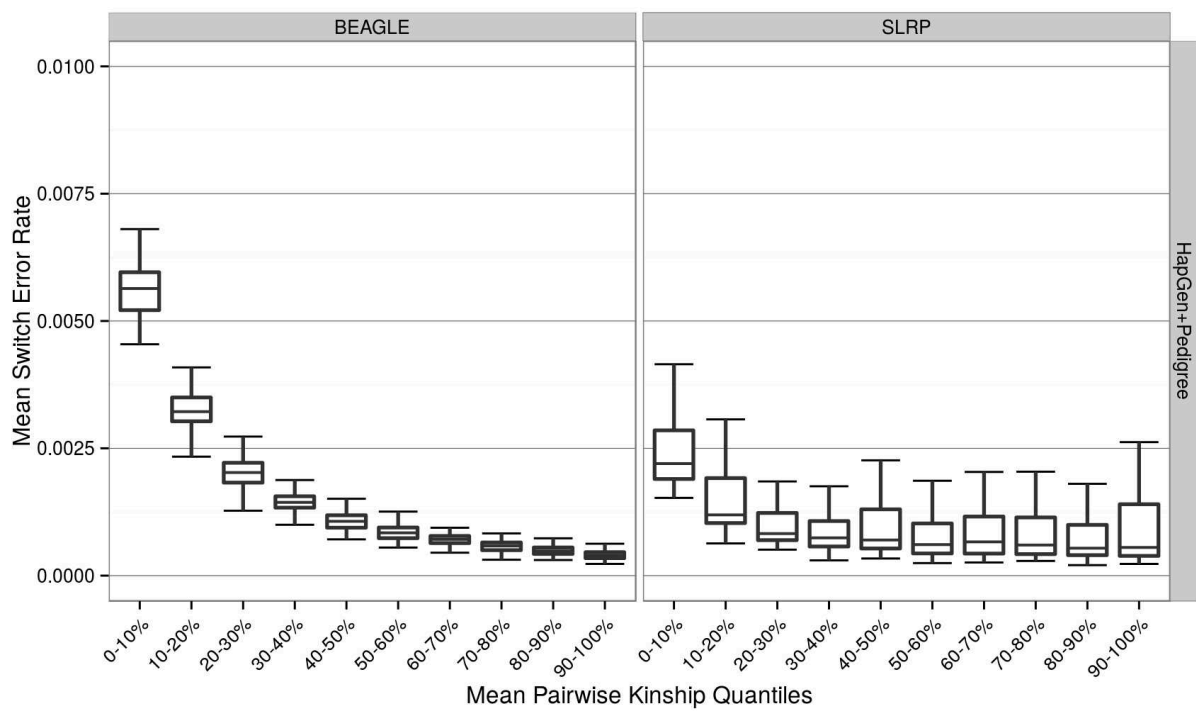
**Supplementary Figure 8a.** Comparison of the fraction of heterozygous sites phased by SLRP for both simulation strategies. SLRP phased a higher proportion of sites when applied to the HapGen+Pedigree simulation, similar to the proportion of sites as when applied to the observed ARRAY genotypes in Campora by SLRP (blue line). Genotype errors and missingness led to a reduction in the number of sites that SLRP was able to phase in both simulation strategies.



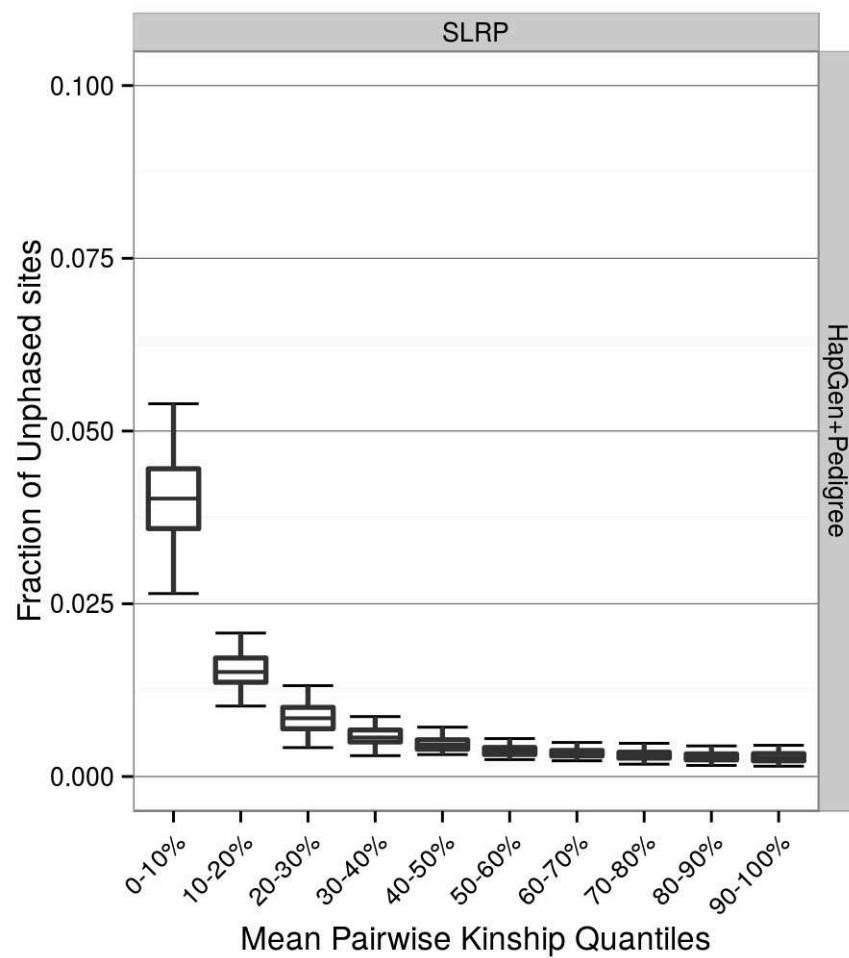
**Supplementary Figure 8b.** Comparison of the fraction of heterozygous sites phased by ALPHAPHASE for both simulation strategies. ALPHAPHASE phased a higher proportion of sites when applied to the HapGen+Pedigree simulation. Genotype errors and missingness led to a reduction in the number of sites that ALPHAPHASE was able to phase in both simulation strategies.



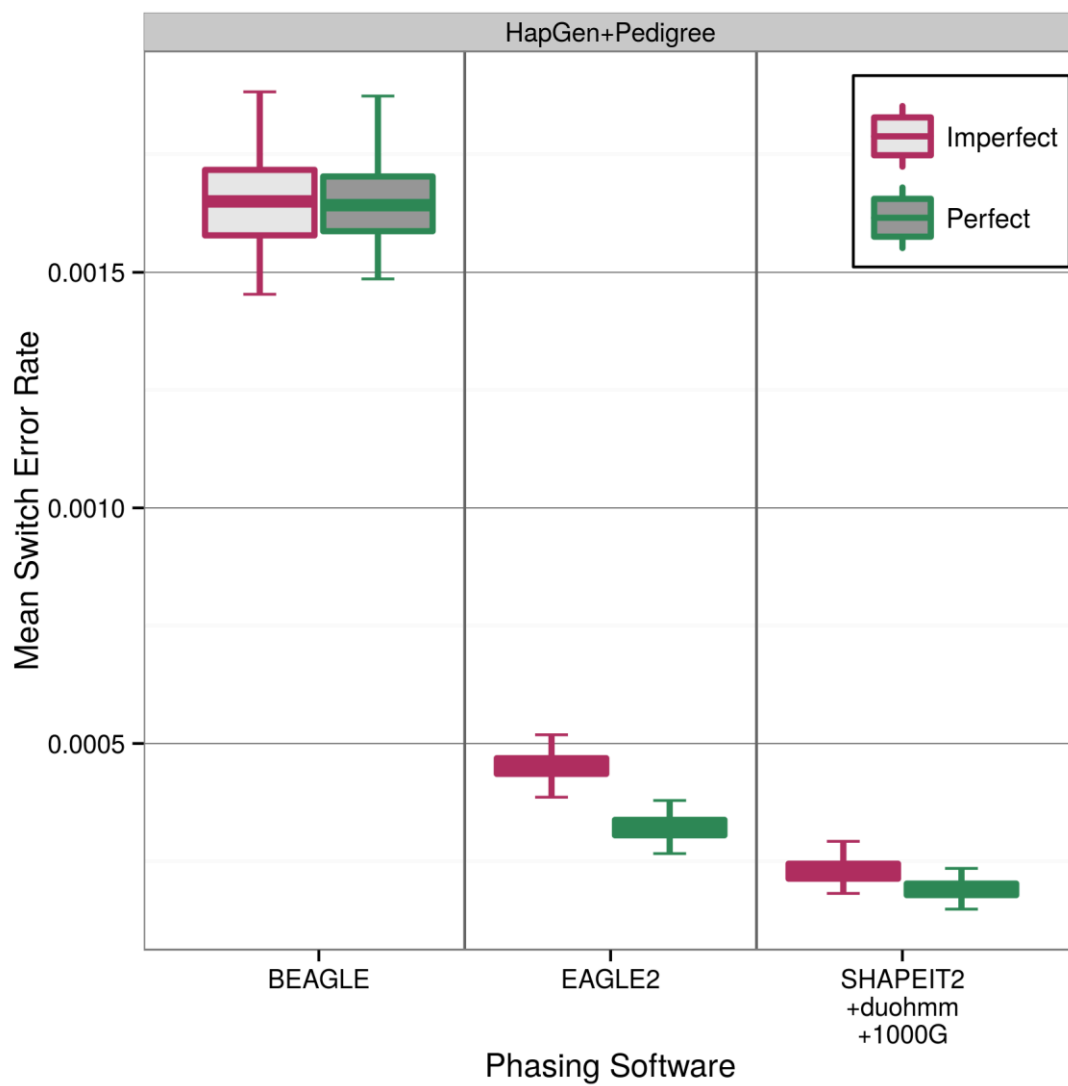
**Supplementary Figure 9a.** Relationship between mean pairwise genetic kinship and individual SER for SHAPEIT2+duohmm+1000G and EAGLE2. In each replicate, ARRAY genotypes were used to calculate the mean pairwise genetic kinship coefficient of each individual to all others. We considered 10 equally sized bins of mean pairwise genetic kinship based on the quantiles of the distribution of mean pairwise genetic kinship. In each group we then calculated the mean SER for all individuals in the group.



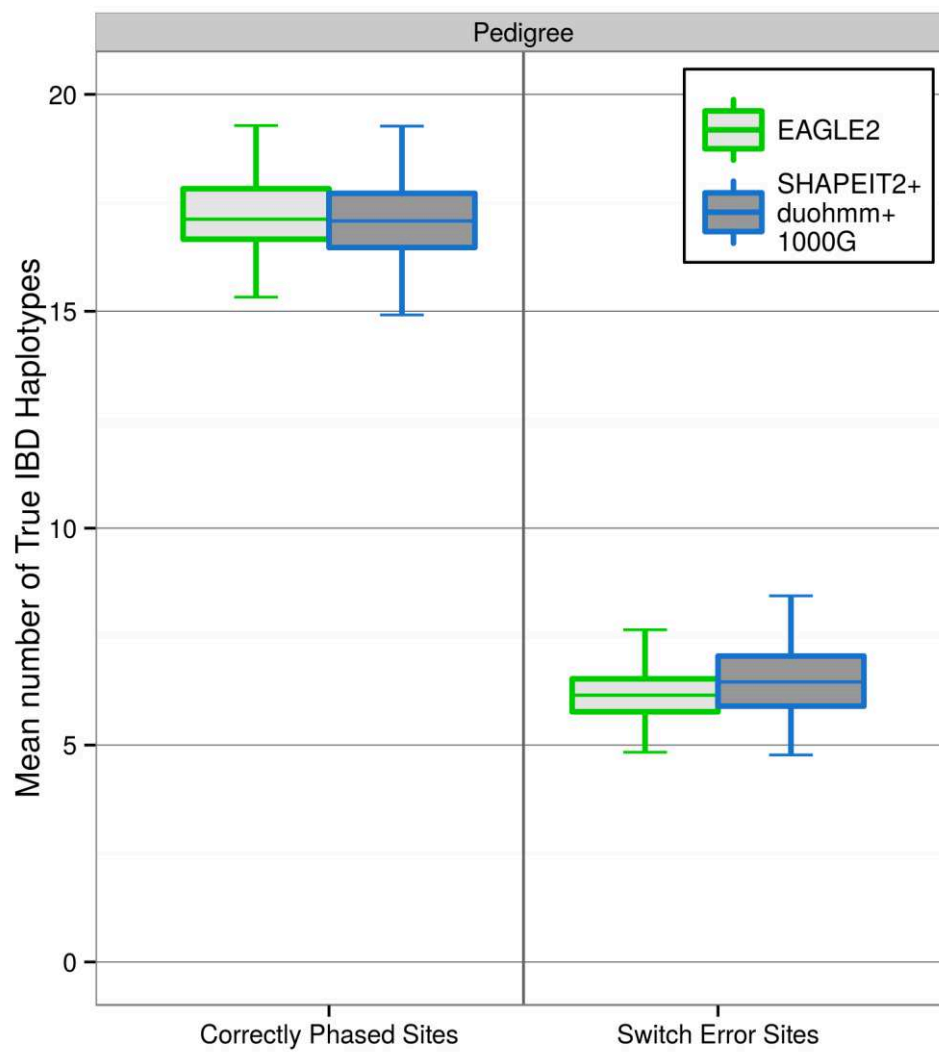
**Supplementary Figure 9b.** As Supplementary Figure 9a, but for BEAGLE and SLRP. Note the different scale on the y-axis compared to Supplementary Figure 9a.



**Supplementary Figure 9c.** Fraction of all heterozygous sites left unphased by SLRP according to mean pairwise genetic kinship.

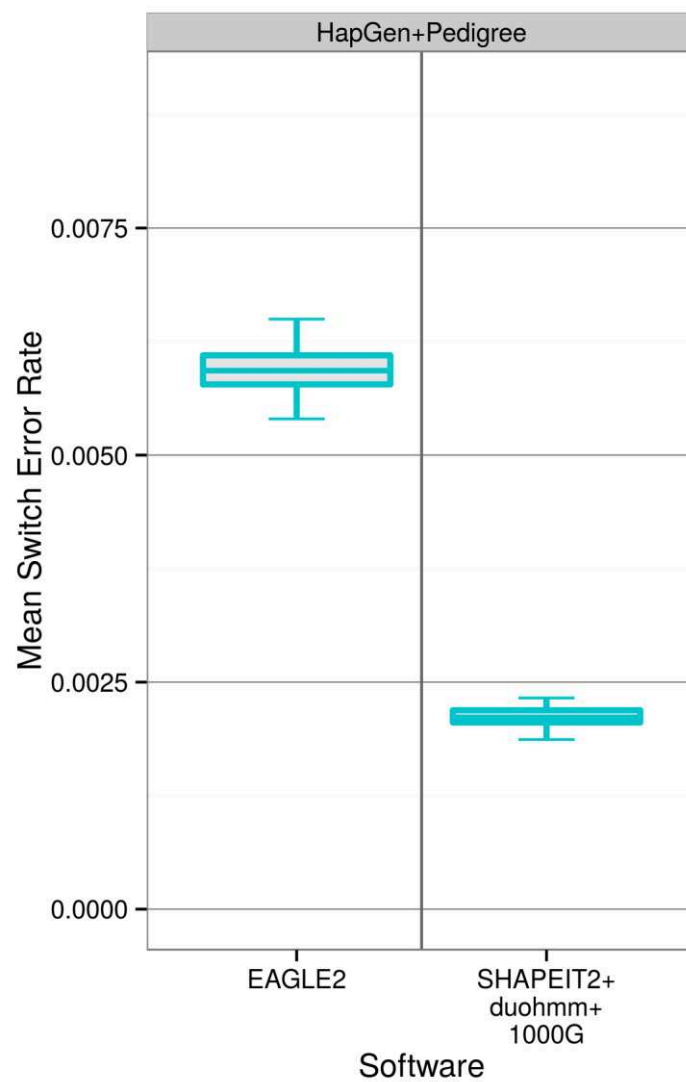


**Supplementary Figure 10.** Effect of genotype errors and missingness (Imperfect) on mean Switch Error Rate according to software and on the HapGen+Pedigree simulation strategy.

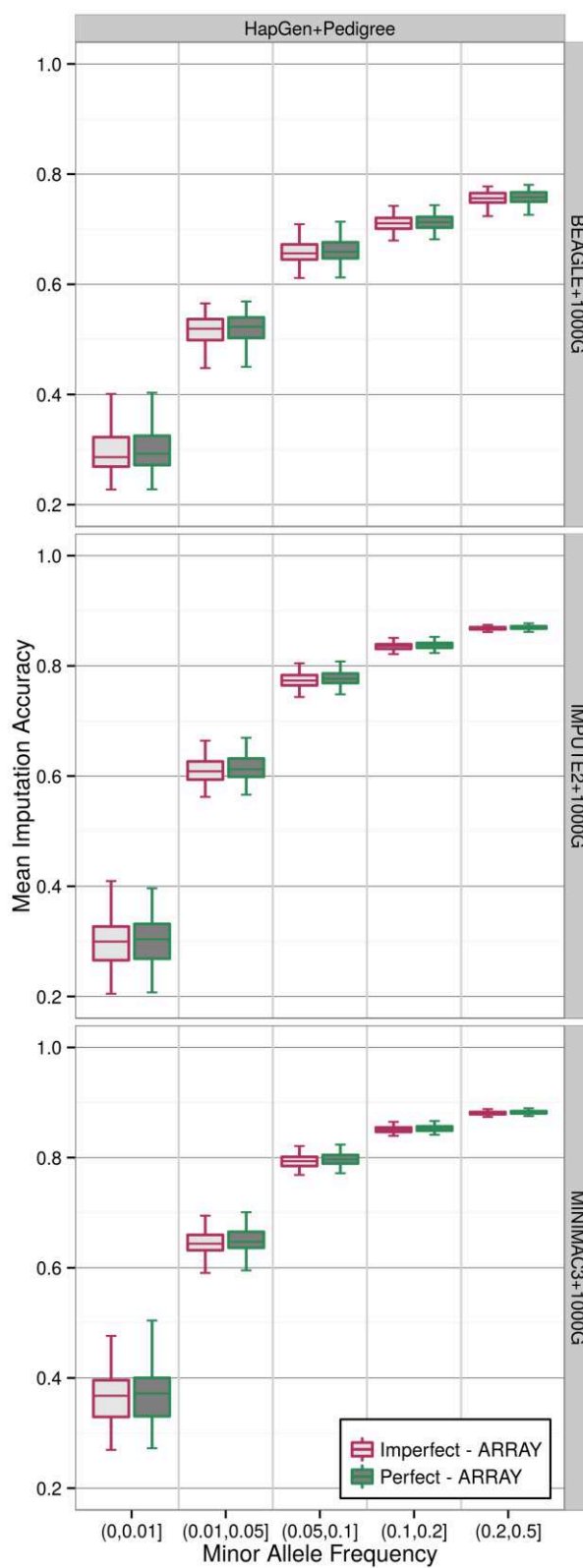


**Supplementary Figure 11.** Comparison of mean number of true IBD haplotypes at either correctly phased sites or switch error sites for SHAPEIT2+duohmm+1000G or EAGLE2. This analysis was only possible on the Pedigree simulation where the exact locations of simulated IBD-sharing were known.

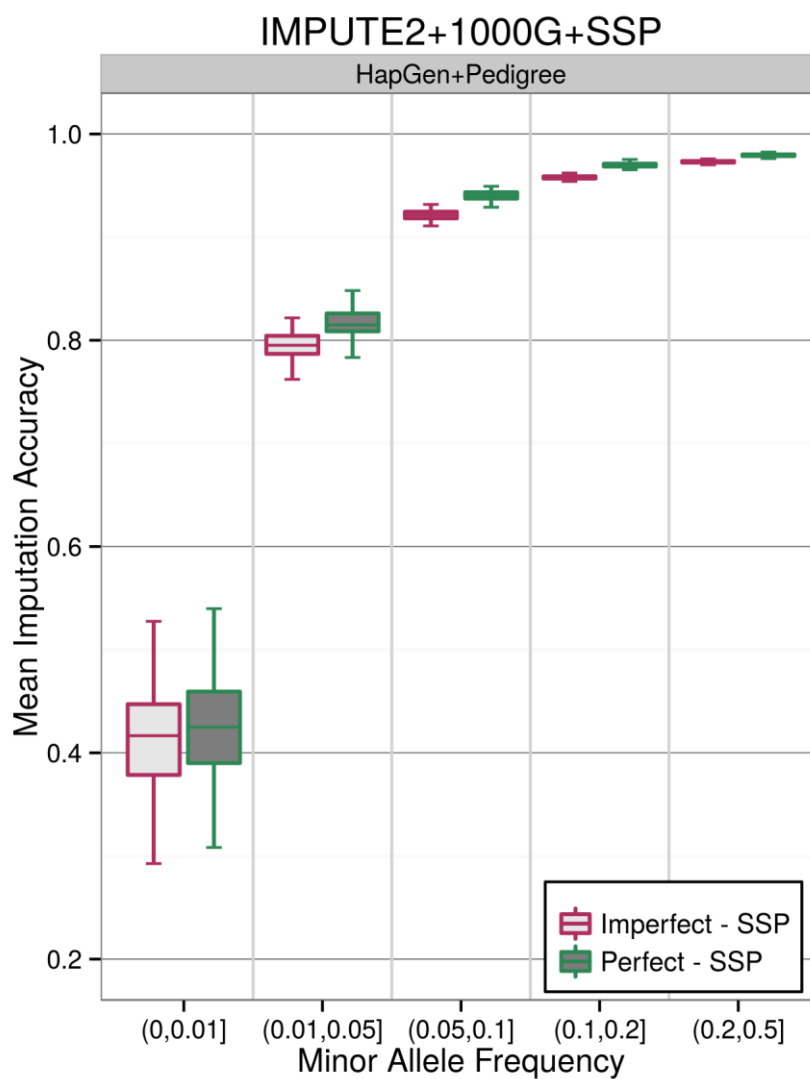




**Supplementary Figure 12.** Mean SER for the phasing of the 93 WGS SSP individuals with SHAPEIT2+duohmm+1000G and EAGLE2. There are two likely causes for the higher SERs as compared to the phasing of ARRAY data: firstly, a smaller number of individuals are involved, and secondly the WGS data contained a higher proportion of variants with MAF below 0.05.

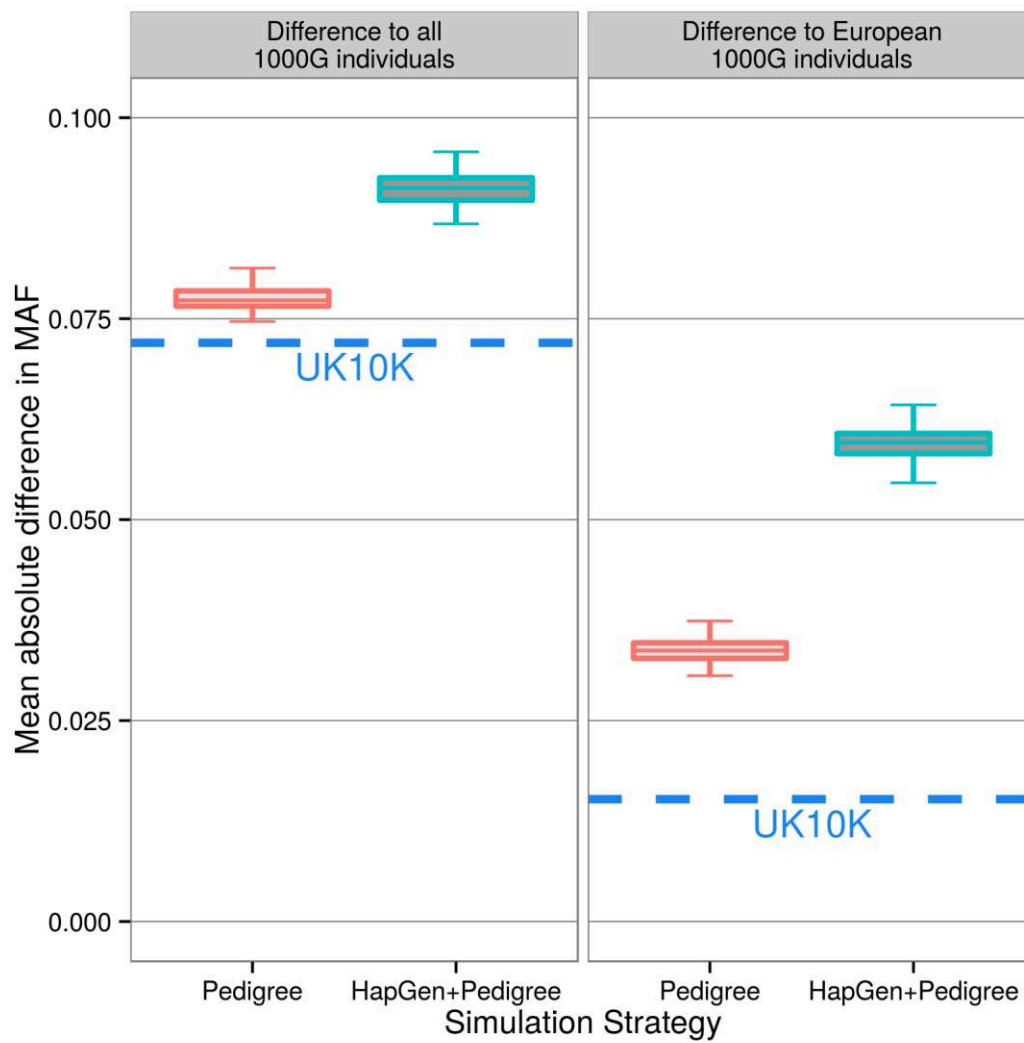


**Supplementary Figure 13.** Effect of genotype errors and missingness on the ARRAY data on imputation accuracy.

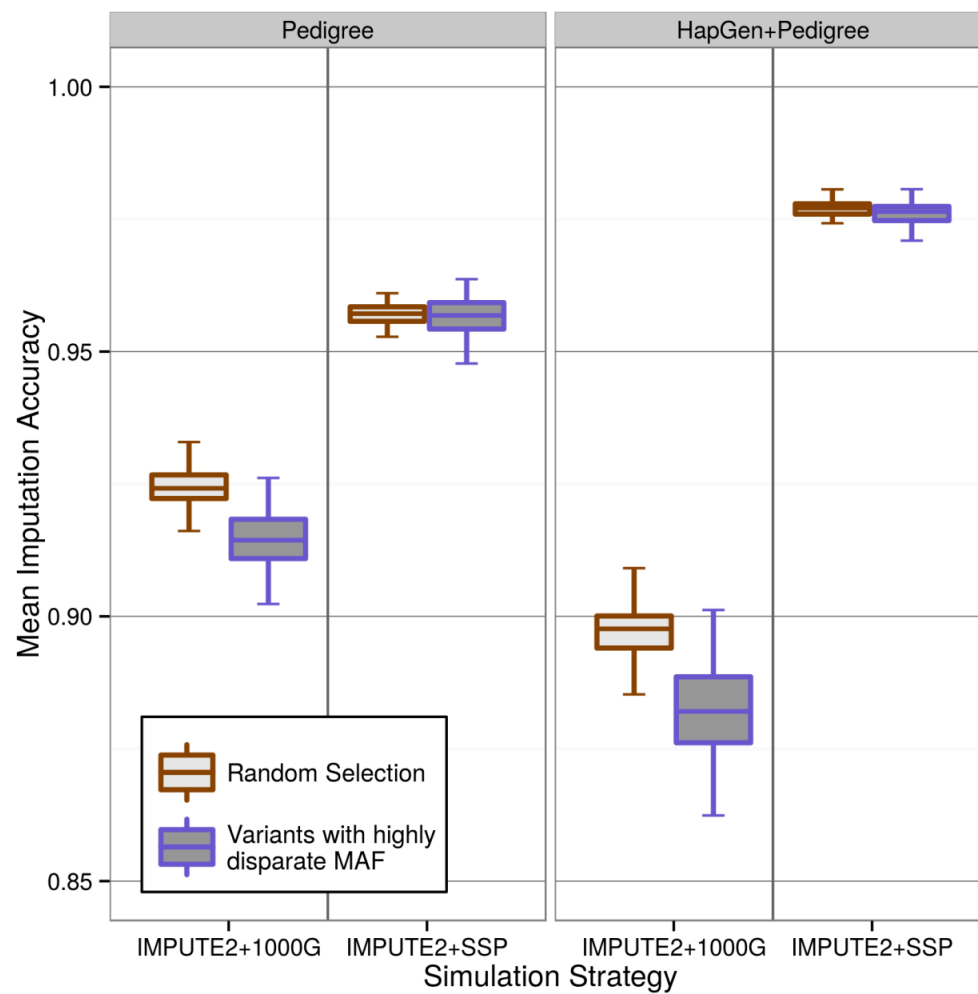


**Supplementary Figure 14.** Effect of genotype errors and missingness on the SSP on imputation accuracy.

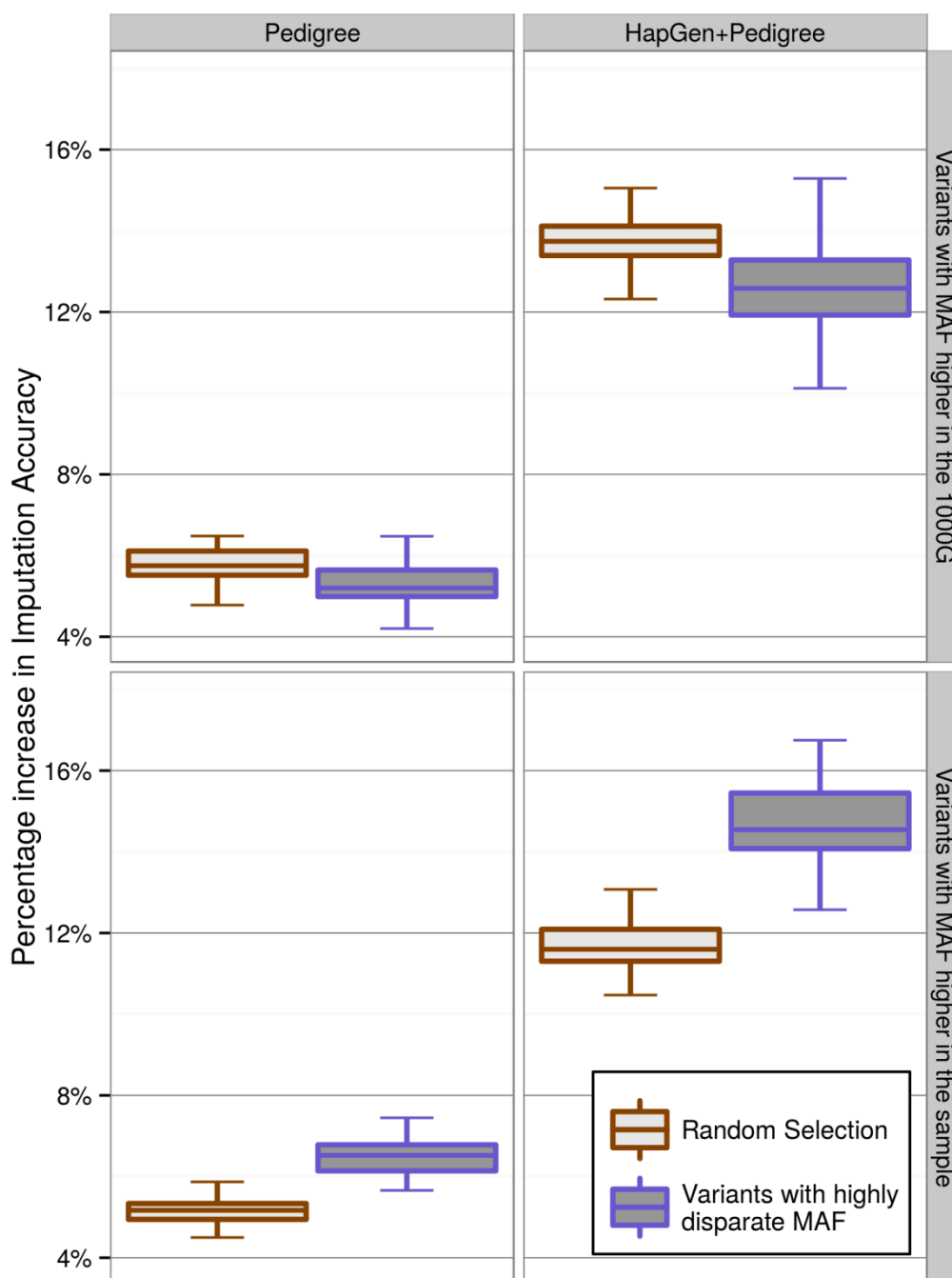
Imputation accuracy calculated from imputation strategy IMPUTE2+1000G+SSP.



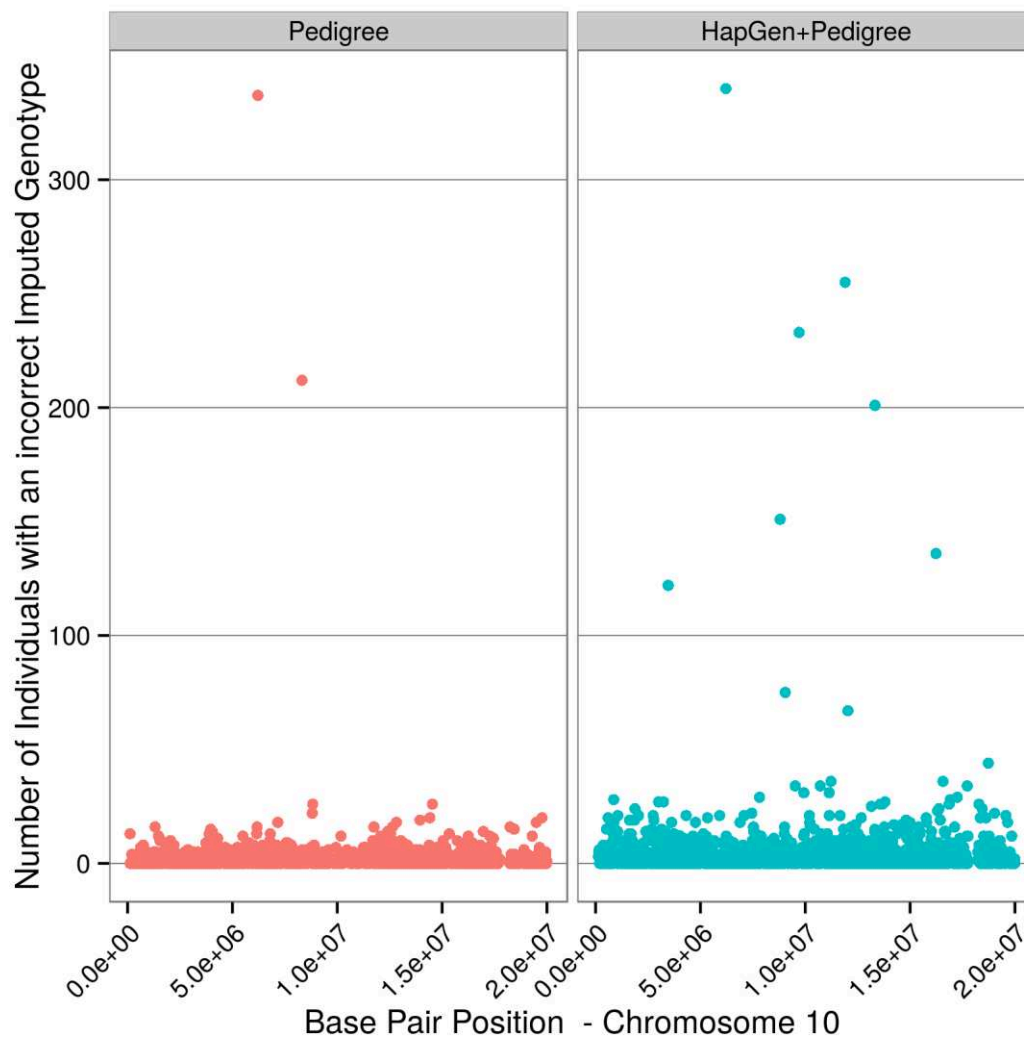
**Supplementary Figure 15.** Comparison of absolute difference in MAF between simulated data and the 1000G panel for both simulation strategies. Dashed lines represent the mean difference in MAF between the UK10K (founding pool used for the simulation) and the 1000G.



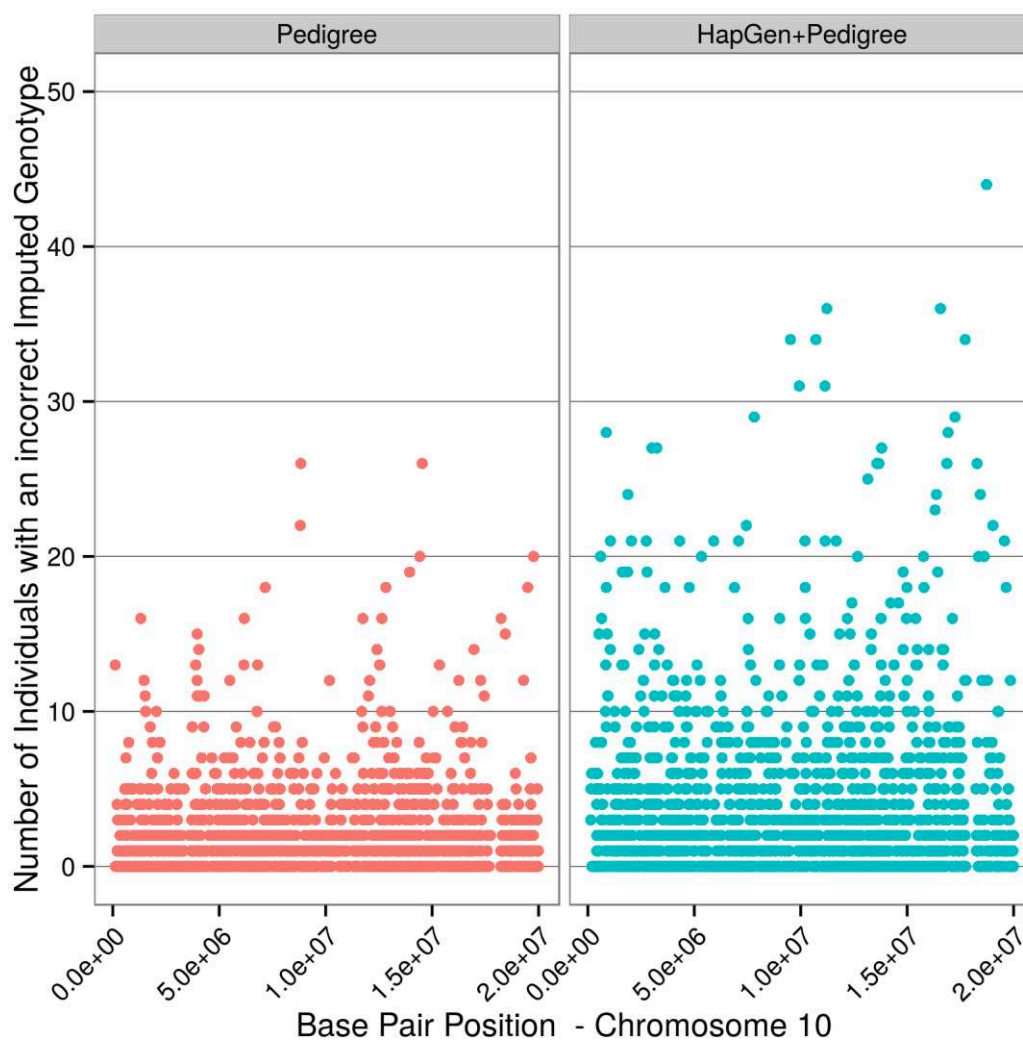
**Supplementary Figure 16a.** Comparison of imputation accuracy for sets of variants with particularly high differences in MAF compared to the 1000G panel against random selections of similar variants without such elevated disparities. Imputation accuracy was calculated from the imputation strategies IMPUTE2+1000G and IMPUTE2+SSP.



**Supplementary Figure 16b.** Increase in imputation accuracy by changing from IMPUTE2+1000G to IMPUTE2+SSP for sets of variants with either MAF greater in the 1000G reference panel compared to the sample or vice-versa. Once again, for each set of chosen variants for comparison, we selected a random selection of control variants with similar MAF in the sample to the chosen set.

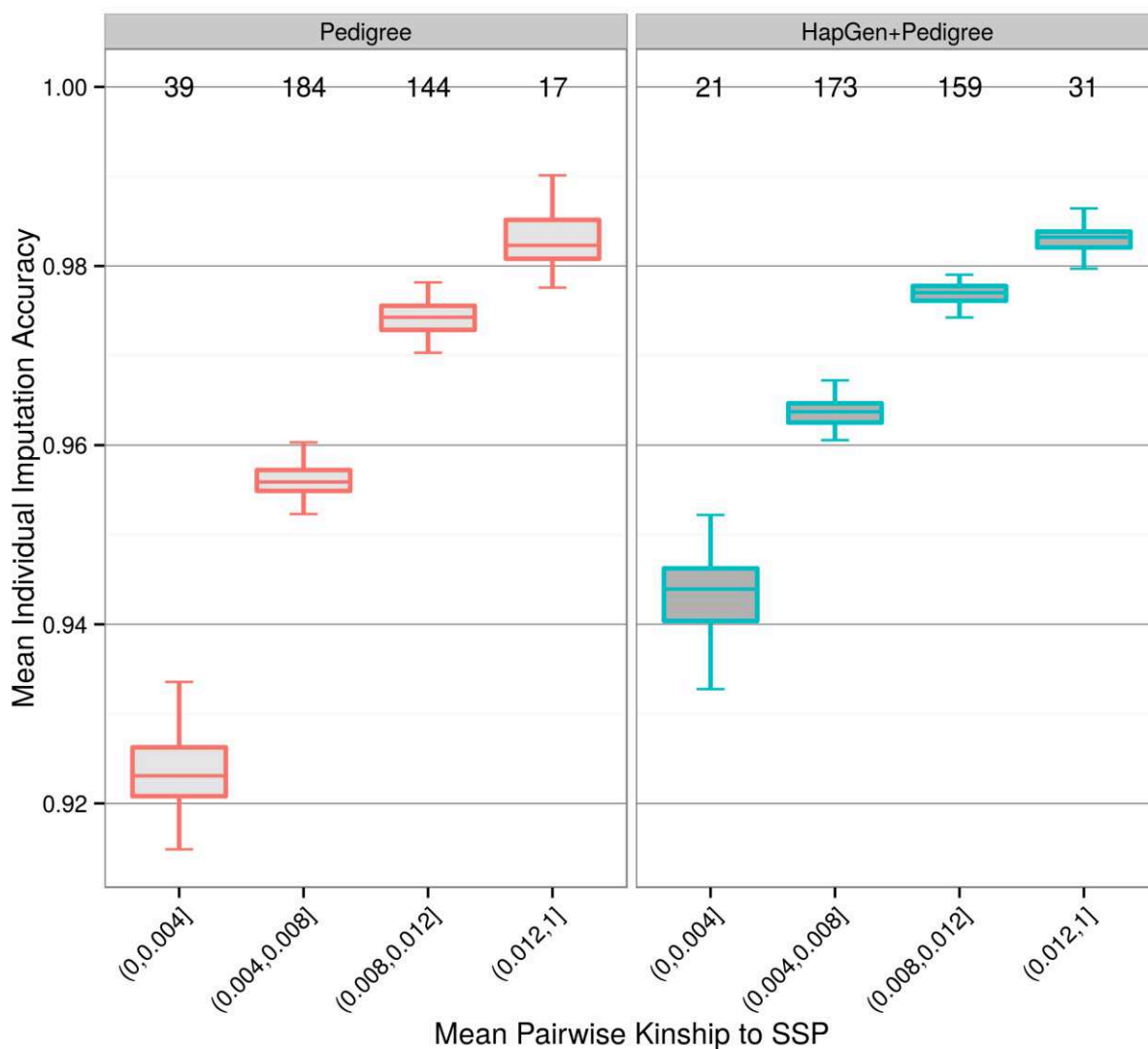


**Supplementary Figure 16c.** Imputation of monomorphic variants in the sample under IMPUTE2+1000G. The number of individuals with an incorrectly imputed genotype (after taking a hard call) against base pair position on chromosome 10.

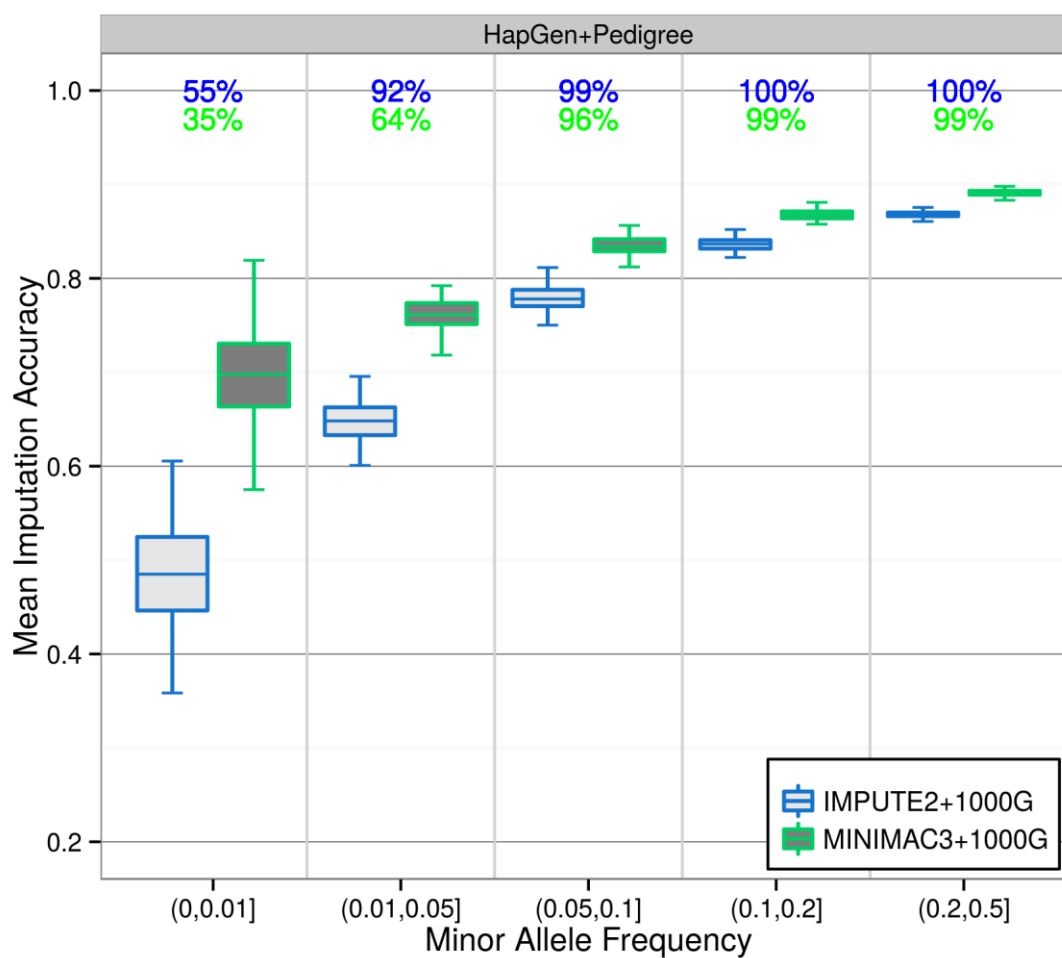


**Supplementary Figure 16d.** Zoom-in onto Supplementary Figure 16c showing the distribution of points with y-axis values less than 50.

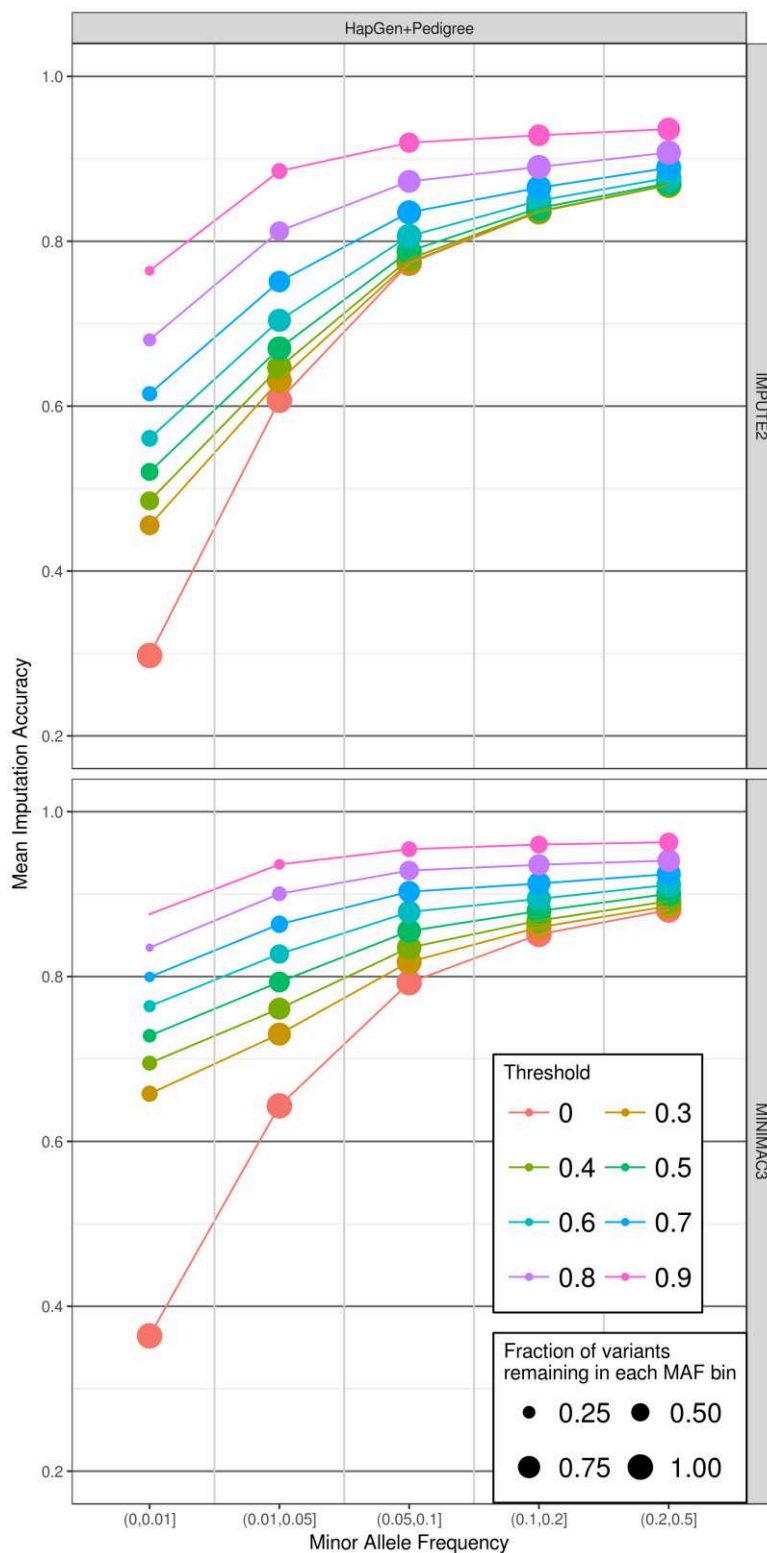




**Supplementary Figure 17.** Comparison of individual imputation accuracy against the mean pairwise genetic kinship between each non-SSP member and all 93 SSP members. Mean numbers of individuals contributing to each bin of individual mean pairwise genetic kinship are displayed atop the figure. The minimum observed imputation accuracy for a single individual was just above 0.85 and just above 0.88 for the Pedigree and HapGen+Pedigree simulation strategies respectively.



**Supplementary Figure 18a.** Imputation accuracy across all MAFs following post imputation quality control based on a 0.4 threshold on 'info' scores for IMPUTE2 and a 0.3 threshold on 'RSQ' scores for MINIMAC3. These are the often recommended thresholds for 'info' and 'RSQ'. The percentages of variants that remain in each MAF bin after thresholding are displayed atop the figure (blue for IMPUTE2 and green for MINIMAC3).



**Supplementary Figure 18b.** Imputation accuracy across all MAFs following post imputation quality control based on either ‘info’ scores for IMPUTE2 or ‘RSQ’ scores for MINIMAC3. Imputation accuracy and imputation quality scores are derived from IMPUTE2+1000G or MINIMAC3+1000G imputation.

True Base	Error Base rates				
	G	T	C	A	Total
G	-	1/60	1/120	1/120	1/30
T	1/60	-	1/120	1/60	1/24
C	1/120	1/120	-	1/120	1/40
A	1/120	1/60	1/120	-	1/30

**Supplementary Table 1.** Error rates between specific bases for the simulation of WGS data.

Phasing Software + Options	Mean Switch Error Rate	
	Pedigree	HapGen+Pedigree
ALPHAPHASE †	0.0235	0.0218
BEAGLE	0.00490	0.00165
EAGLE1	0.00267	0.000589
EAGLE2	0.00152	0.000321
EAGLE2+1000G	0.00293	0.00155
SHAPEIT2	0.000910	0.000283
SHAPEIT2+duohmm	0.000845	0.000247
SHAPEIT2+duohmm+1000G	0.000638	0.000191
SHAPEIT3	0.000957	0.000279
SLRP †	0.00117	0.000950

† Not all variants were phased.

**Supplementary Table 2.** Mean global SER across simulation replicates for all phasing strategies considered.

Imputation Software + Reference Panel	Mean Imputation Accuracy										
	MAF	Pedigree					HapGen+Pedigree				
		(0,0.01]	(0.01,0.05]	(0.05,0.10]	(0.10,0.20]	(0.20,0.50]	(0,0.01]	(0.01,0.05]	(0.05,0.10]	(0.10,0.20]	(0.20,0.50]
BEAGLE+1000G †	0.423	0.605	0.744	0.792	0.832	0.296	0.518	0.658	0.710	0.757	
IMPUTE2+1000G †	0.472	0.670	0.833	0.882	0.904	0.299	0.608	0.774	0.836	0.868	
IMPUTE4+1000G †	0.524	0.714	0.845	0.890	0.912	0.351	0.633	0.786	0.845	0.877	
MINIMAC3+1000G †	0.530	0.722	0.852	0.900	0.916	0.366	0.644	0.793	0.851	0.881	
PBWT+1000G (20 replicates) †	0.426	0.640	0.791	0.851	0.883	0.290	0.555	0.724	0.798	0.840	
MAF	(0,0.05]	(0.05,0.10]	(0.10,0.20]	(0.20,0.50]	(0,0.05]	(0.05,0.10]	(0.10,0.20]	(0.20,0.50]			
IMPUTE2+1000G ‡	0.749	0.863	0.883	0.907	0.670	0.811	0.834	0.871			
IMPUTE2+SSP ‡	0.845	0.921	0.938	0.954	0.916	0.951	0.963	0.974			
IMPUTE2+1000G+SSP ‡	0.872	0.933	0.946	0.960	0.914	0.950	0.961	0.973			
MINIMAC3+1000G ‡	0.779	0.882	0.900	0.920	0.703	0.831	0.855	0.884			
MINIMAC3+HRC ‡	0.844	0.918	0.930	0.942	0.752	0.860	0.879	0.903			
MINIMAC3+SSP ‡	0.840	0.917	0.935	0.953	0.909	0.946	0.958	0.971			
MINIMAC3+HRC+SSP ‡	0.905	0.951	0.961	0.971	0.918	0.953	0.964	0.974			

† Corresponds to a comparison on 40,989 and 40,407 variants on the Pedigree and HapGen+Pedigree simulation strategies respectively.

‡ Corresponds to a comparison on 35,058 and 34,605 variants present in the SSP on the Pedigree and HapGen+Pedigree simulation strategies respectively.

**Supplementary Table 3.** Mean Imputation accuracy across simulation replicates split by MAF, these results correspond to Figures 3, 4, and 5 in the main text.

MAF	(0,0.01]		(0.01,0.05]		(0.05,0.10]		(0.10,0.20]		(0.20,0.50]	
	Good	Bad	Good	Bad	Good	Bad	Good	Bad	Good	Bad
<b>N</b>	313	595	2232	670	3568	272	7621	210	21682	359
	<b>Variants remaining (%) after threshold was applied</b>									
<b>info</b>										
0.3	96	37	100	80	100	97	100	99	100	100
0.4	94	30	100	69	100	90	100	96	100	99
0.5	91	23	100	54	100	73	100	79	100	82
0.6	89	17	99	37	100	48	100	47	100	43
0.7	80	11	98	20	98	23	99	20	100	13
0.8	68	6	93	8	92	9	95	7	97	3
0.9	46	2	77	2	74	2	80	3	85	1
<b>RSQ</b>										
0.3	85	13	96	37	100	55	100	60	100	65
0.4	79	9	93	23	99	34	100	35	100	33
0.5	71	6	88	14	97	18	99	18	100	10
0.6	62	4	80	7	92	9	96	8	98	3
0.7	51	2	67	4	84	4	90	4	94	1
0.8	37	1	51	2	70	2	77	2	84	1
0.9	20	0	29	0	47	1	53	0	60	0

**Supplementary Table 4.** Mean number of variants (N) in each MAF bin that were well imputed (Good) or poorly imputed (Bad) as defined by whether Imputation accuracy exceeded 0.5 or fell below 0.2 respectively. The body of the table displays the mean percentage of variants remaining after ‘info’ or ‘RSQ’ thresholds have been applied. ‘Info’ and ‘RSQ’ scores pertain to IMPUTE2+1000G and MINIMAC3+1000G imputation respectively.

<b>Phasing</b>	<b>Real Time</b>	<b>Computational Time</b>
BEAGLE	0:05:53	0:36:58
SLRP	3:39:20	3:34:38
ALPHAPHASE	0:04:05	0:03:59
EAGLE1	0:04:33	0:16:47
EAGLE2	0:04:11	0:15:27
EAGLE2+1000G	0:15:46	0:44:32
SHAPEIT2	1:00:46	1:00:40
SHAPEIT2+duohmm	1:01:58	1:01:53
SHAPEIT2+duohmm+1000G	1:14:23	1:09:08
SHAPEIT3	0:46:46	0:46:45
<b>Imputation</b>		
BEAGLE+1000G	0:17:54	3:35:49
IMPUTE2+1000G	2:03:24	1:57:39
IMPUTE4+1000G	0:13:55	0:12:56
IMPUTE2+SSP	0:02:49	0:02:42
IMPUTE2+1000G+SSP	5:00:29	4:53:01
MINIMAC3+1000G †	1:07:13	1:05:05
MINIMAC3+SSP †	0:03:30	0:03:20
MINIMAC3+HRC †	7:39:29	7:35:56

† Part of the duration taken by MINIMAC3 was attributed to reformatting the reference panel into a specialised MINIMAC3 format.

**Hours:Minutes:Seconds**


**Supplementary Table 5.** Time requirements for phasing ARRAY data on the whole of chromosome 10 for and imputing 20Mb of chromosome 10.

## References

- Browning, Brian L., & Browning, Sharon R. (2016). *Genotype Imputation with Millions of Reference Samples*. *Am J Hum Genet*, 98(1), 116-126. doi: 10.1016/j.ajhg.2015.11.020
- DePristo, M. A., Banks, E., Poplin, R. E., Garimella, K. V., Maguire, J. R., Hartl, C., . . . Daly, M. J. (2011). *A framework for variation discovery and genotyping using next-generation DNA sequencing data*. *Nat Genet*, 43(5), 491-498. doi: 10.1038/ng.806
- Kim, S. Y., Lohmueller, K. E., Albrechtsen, A., Li, Y., Korneliussen, T., Tian, G., . . . Nielsen, R. (2011). *Estimation of allele frequency and association mapping using next-generation sequencing data*. *BMC Bioinformatics*, 12, 231-231. doi: 10.1186/1471-2105-12-231
- Li, Y., Willer, C. J., Ding, J., Scheet, P., & Abecasis, G. R. (2010). *MaCH: Using Sequence and Genotype Data to Estimate Haplotypes and Unobserved Genotypes*. *Genet Epidemiol*, 34(8), 816-834. doi: 10.1002/gepi.20533
- Liu, E. Y., Buyske, S., Aragaki, A. K., Peters, U., Boerwinkle, E., Carlson, C., . . . Li, Y. (2012). *Genotype Imputation of MetaboChip SNPs Using a Study-Specific Reference Panel of ~4,000 Haplotypes in African Americans From the Women's Health Initiative*. *Genet Epidemiol*, 36(2), 107-117. doi: 10.1002/gepi.21603
- Pistis, G., Porcu, E., Vrieze, S. I., Sidore, C., Steri, M., Danjou, F., . . . Sanna, S. (2015). *Rare variant genotype imputation with thousands of study-specific whole-genome sequences: implications for cost-effective study designs*. *Eur J Hum Genet*, 23(7), 975-983. doi: 10.1038/ejhg.2014.216



# SCIENTIFIC REPORTS



OPEN

## Detecting the dominance component of heritability in isolated and outbred human populations

Anthony F. Herzig<sup>1,2</sup>, Teresa Nutile<sup>3</sup>, Daniela Ruggiero<sup>3,4</sup>, Marina Ciullo<sup>3,4</sup>, Hervé Perdry<sup>5</sup> & Anne-Louise Leutenegger<sup>1,2</sup>

Inconsistencies between published estimates of dominance heritability between studies of human genetic isolates and human outbred populations incite investigation into whether such differences result from particular trait architectures or specific population structures. We analyse simulated datasets, characteristic of genetic isolates and of unrelated individuals, before analysing the isolate of Cilento for various commonly studied traits. We show the strengths of using genetic relationship matrices for variance decomposition over identity-by-descent based methods in a population isolate and that heritability estimates in isolates will avoid the downward biases that may occur in studies of samples of unrelated individuals; irrespective of the simulated distribution of causal variants. Yet, we also show that precise estimates of dominance in isolates are demonstrably problematic in the presence of shared environmental effects and such effects should be accounted for. Nevertheless, we demonstrate how studying isolates can help determine the existence or non-existence of dominance for complex traits, and we find strong indications of non-zero dominance for low-density lipoprotein level in Cilento. Finally, we recommend future study designs to analyse trait variance decomposition from ensemble data across multiple population isolates.

For a plethora of human traits, there is an observable resemblance between close relatives. This suggests the presence of genetic constituents in the architectures of such traits and leads to an obvious question: for a pair of individuals, can one describe a relationship between their degree of relatedness (genomic sharing) and the degree of similarity of their trait values? Fisher unravelled this question by proposing a decomposition of the variance of a trait, with components attributed to each individual's genome and to the amassment of environmental exposures in each individual's history. This genetic component of the variability is known as the heritability of the trait which Fisher connected to the correlation of trait values between relatives. Heritability has been estimated extensively for a multitude of traits and through diverse models and study designs. Importantly, the recent availability of dense genetic data in large cohorts has enabled the estimation of heritability from samples of unrelated individuals whereas previous estimations had been driven by studies of close relatives such as twins or nuclear families. A review of heritability estimation in related individuals can be found in Tenesa & Haley<sup>1</sup> and a recent discussion of heritability estimation in unrelated individuals can be found in Yang *et al.*<sup>2</sup>

An important distinction is to be made between broad-sense heritability ( $H^2$ ) and the more commonly communicated narrow-sense heritability ( $h^2$ ). This stems from the innovative modelling of complex traits by Fisher who demonstrated the interest of splitting the genetic variance of a trait into additive, dominant (interaction of alleles within a genotype of a single locus), and epistatic (interaction between genotypes of multiple loci) components<sup>3</sup>. For details on more elaborate models, we refer the reader to Abney *et al.*<sup>4</sup> and Young & Durbin<sup>5</sup>. Briefly put,  $h^2$  describes the additive contributions of each allele received from one's parents while  $H^2$  encompasses the effect of one's whole genome and is the sum of  $h^2$  and the contributions of non-additive effects. For the purposes of this study, we term this non-additive fraction of variance as 'dominant' as we do not here consider epistasis or higher order variance terms; we will denote this component as  $h_D^2$  (equal to  $H^2 - h^2$ ). In terms of phenotypic

<sup>1</sup>Inserm, U946, Genetic variation and Human diseases, Paris, France. <sup>2</sup>Université Paris-Diderot, Sorbonne Paris Cité, U946, Paris, France. <sup>3</sup>Institute of Genetics and Biophysics A. Buzzati-Traverso - CNR, Naples, Italy. <sup>4</sup>IRCCS Neuromed, Pozzilli, Isernia, Italy. <sup>5</sup>Université Paris-Saclay, University. Paris-Sud, Inserm, CESP, Villejuif, France. Hervé Perdry and Louise Leutenegger jointly supervised this work. Correspondence and requests for materials should be addressed to A.F.H. (email: [anthony.herzig@inserm.fr](mailto:anthony.herzig@inserm.fr)) or M.C. (email: [marina.ciullo@igb.cnr.it](mailto:marina.ciullo@igb.cnr.it))

similarities between family members, the parent/offspring correlation is equal to  $\frac{1}{2}h^2$  while the sibling correlation is equal to  $\frac{1}{4}h_D^2 + \frac{1}{2}h^2$ . To give clarity, we define  $h_A^2 = h^2$ .

We will consider the estimation of heritability through maximum-likelihood estimation of variance parameters of linear mixed models (LMMs). For a setting of  $N$  individuals and  $Y$  a vector of observed phenotypes, we will consider the following model with fixed effects  $X$  and a variance-covariance structure split into genetic additive, genetic dominant, and environmental components:

$$Y \sim MVN(\beta_0^T X, \tau_A K + \tau_D D + \sigma_E^2 I_N) \quad (1)$$

We then are able to estimate the heritabilities as follows:

$$H^2 = \frac{\tau_A + \tau_D}{\tau_A + \tau_D + \sigma_E^2}, \quad h_A^2 = \frac{\tau_A}{\tau_A + \tau_D + \sigma_E^2}, \quad h_D^2 = \frac{\tau_D}{\tau_A + \tau_D + \sigma_E^2} \quad (2)$$

There are various possible choices of the  $N \times N$  matrices  $K$  and  $D$ . Historically,  $K$  and  $D$  are defined in terms of identity-by-descent (IBD) probabilities<sup>4,6,7</sup>.  $K$  is equal to  $2\varphi$ , where  $\varphi_{i,j}$  is the kinship coefficient of individuals  $i$  and  $j$ , defined as the probability of two alleles, randomly sampled from each of individuals  $i$  and  $j$ , at the same locus will be IBD.  $D_{i,j}$  is the probability that individuals  $i$  and  $j$  share exactly two pairs of alleles IBD at a given locus. Both  $\varphi_{i,j}$  and  $D_{i,j}$  are themselves expressions of Jacquard's nine coefficients of identity:  $\varphi_{i,j} = \Delta_1 + \frac{1}{2}(\Delta_3 + \Delta_5 + \Delta_7) + \frac{1}{4}\Delta_8$ , and  $D_{i,j} = \Delta_1 + \Delta_7$ <sup>6</sup>. In studies of family data or isolated populations, these coefficients have been classically estimated from pedigree information but with the advent of dense genomic information, they can now be estimated reliably from genotype data by either estimating genome-wide IBD sharing probabilities or detecting and counting IBD segments<sup>8–10</sup>. Such methods have also been developed for studies of unrelated individuals<sup>11</sup>, though the predominant approach in such studies is to use moment estimators of  $K$  and  $D$  by taking correlations between each pair of individuals' (orthogonal) additive and dominant genetic components, respectively<sup>12,13</sup>. These latter estimators are known as genetic relationship matrices (GRMs) and can be used in any study design.

This leads to two distinct interpretations of the matrices  $K$  and  $D$  which both come with potential drawbacks. If IBD probabilities are used to estimate  $K$  and  $D$ , they represent the level of relatedness between pairs of individuals based on the presence of recent common ancestors but if  $K$  and  $D$  are estimated as GRMs, then they represent simply the correlation between pairs of individuals' genotypes. For the former interpretation, coefficients of identity can only be approximated either by their expected values based on the pedigree structure linking individuals or by estimating the proportions of IBD-sharing between individuals based on their genotypes. However, exhaustive pedigree information is never available and indeed the concept of IBD is similarly problematic due to the ambiguity of how many generations to consider when looking back for evidence of shared genetic ancestors. After many generations, mutations and recombinations cause the IBD segments to become increasingly short and not completely identical and thus difficult to distinguish from background genetic variation<sup>14–16</sup>. For the latter interpretation involving GRMs, there is the immediate problem that such correlations are computed from a large set of variants which are not specific to the trait being studied in the hope that these variants will be representative of the unknown set of causal variants via linkage disequilibrium (LD) (correlations between variants)<sup>17</sup>. Consequentially, if heritability is estimated with GRMs, it corresponds to only a proportion of the phenotypic variation coming from the subset of causal variants that are in LD with the genotyped variants<sup>18</sup>. This can lead to downwardly biased estimate of heritability as causal variants may often be held at low frequencies by selection<sup>19,20</sup> and so will be in weak LD with common genotyped variants. Furthermore, if there exist relatively few causal variants, the large numbers of non-causal variants used to estimate the genetic correlations might mask the desired correlation of causal variants between individuals<sup>21</sup>. Genomic-based IBD methods applied to unrelated individuals has been suggested as an approach to improve upon genetic correlation methods as detected stretches of IBD can cover some un-typed genetic variation<sup>11</sup>.

The main motivation for employing GRMs is that this allows for the estimation of heritability from unrelated individuals, thus leveraging data from large cohorts and avoiding shared environment biases<sup>13,22</sup>. However, there has been a trend towards using genomic-based estimates even when pedigree data is available due to the increased precision of relatedness estimation from genetic data, both in human studies<sup>16,23–27</sup> and in animal/plant studies<sup>28–31</sup>.

For complex human traits, it has been suggested that one can assume that any contributions from non-additive genetic components ( $h_D^2$ ) are relatively small compared to the additive genetic components<sup>32</sup> and thus often only estimates of  $h_A^2$  are presented. In a recent study, Zhu *et al.*<sup>12</sup> illustrated this characterization of diminutive dominant genetic variance for 79 traits in two large samples of unrelated individuals. This result was then re-enforced in Nolte *et al.*<sup>33</sup>. Yet, many others have presented incongruent results on this subject. Chen *et al.*<sup>34</sup> compared the same approach as Zhu *et al.*<sup>12</sup> with a twin-based analysis and concluded that whilst the genetic variances of 19 traits were predominantly additive, dominant genetic components were nonetheless more prominently apparent than when described elsewhere. Aside from these studies, dominance heritability estimation using GRMs has rarely been carried out, and the authors who are more interested in dominance tend to rely on family data<sup>35,36</sup>. Of particular note is the observation that significant non-additive genetic components for many traits have been found in some studies on population isolates: Abney *et al.*<sup>37</sup>, Pilia *et al.*<sup>38</sup>, and Traglia *et al.*<sup>39</sup> (Table 1).

An isolate is characterized as a population arising from a small group of founders and experiencing subsequent demographic growth in isolation. Such populations will include pairs of distantly related individuals who nonetheless share long haplotypes IBD, and may even share both haplotypes IBD in some regions. The presence of both pairs of closely related individuals and pairs of cryptically related individuals suggests that isolates could be ideally suited to heritability analyses. Furthermore, isolates are of interest for assessing the existence of

Phenotype	Abney, McPeck, & Ober <sup>37</sup> , N = 806, Isolate (1)		Pilia <i>et al.</i> <sup>38</sup> , N = 6,148, Isolate (1) (2)		Traglia <i>et al.</i> <sup>39</sup> , N = 1,803, Isolate (1) (2)		Zaitlen <i>et al.</i> <sup>41</sup> , N ≈ 15,000, Extended Genealogies (3)		van Dongen <i>et al.</i> <sup>35</sup> , N ≈ 7,500, Twin Study (4)		Chen <i>et al.</i> <sup>34</sup> , N = 7,740, Twin Study (5)		Chen <i>et al.</i> <sup>34</sup> , N = 5,779, Outbred (5) (6)		Zhu <i>et al.</i> <sup>42</sup> , N = 8,682, Outbred (6)		Nolte <i>et al.</i> <sup>33</sup> , N = 13,436, Outbred (6)	
	$h_A^2$	$h_D^2$	$h_A^2$	$h_D^2$	$h_A^2$	$h_D^2$	$h_A^2$	$h_D^2$	$h_A^2$	$h_D^2$	$h_A^2$	$h_D^2$	$h_A^2$	$h_D^2$	$h_A^2$	$h_D^2$	$h_A^2$	$h_D^2$
Height	—	—	0.77	0.23 *	0.78	0.22 *	—	—	0.81	0.09	0.77	0.09*	0.62	0.00	0.48	0.02	0.49	0.00
BMI	0.54	0.00	0.36	0.32 *	0.33	0.17	0.16	0.09	0.41	0.37	0.28	0.41*	0.21	0.02	0.23	0.15*	0.25	0.02
TGLY	0.37	0.00	0.30	0.42 *	0.39	0.35 *	—	—	0.33	0.25	0.42	0.14	0.31	0.28*	—	—	0.19	0.01
HDL	0.63	0.00	0.47	0.11	0.62	0.00	0.42	0.14*	0.40	0.27	0.66	0.00	0.24	0.01	0.25	0.07	0.19	0.00
Total Chol	—	—	0.38	0.29 *	0.23	0.77 *	—	—	0.51	0.16	0.28	0.19*	0.15	0.00	0.21	0.01	0.23	0.00
LDL	0.36	0.60 *	0.37	0.27 *	0.33	0.66 *	0.20	0.26*	0.51	0.18	0.23	0.24*	0.16	0.00	0.26	0.02	0.27	0.00

**Table 1.** Published results for additive and dominant genetic variability from various study designs. \*Estimates of  $h_D^2$  presented as statistically significant at the 5% level. ‘—’ Trait not studied for dominance in the article. (1) Estimates based on estimating  $K$  and  $D$  from expected proportions of identity-by-descent (IBD) sharing coming from pedigree information. (2) The depth of pedigree information in these studies did not allow the differentiation between a dominance model (including non-additive genetic variation) and a household model (including an effect of shared environment between siblings). (3) The authors of this study analysed a large sample from the Icelandic population for whom extensive pedigree data was available, Matrices  $K$  and  $D$  were estimated by locating and counting stretches of IBD between pairs of individuals. (4) This study analyses a large cohort of monozygotic and dizygotic adult twins. Standard errors are only presented for broad-sense heritability, though it is likely that the estimates for  $h_D^2$  for all traits other than height were significantly different to zero. (5) The authors of this study performed separate analysis, firstly a twin based study using structural equation methods with adjustments for reported levels of time spent in a shared environment between twins, and secondly a study of a large sample of unrelated which included one individual out of most twin pairs in the first analysis. (6) Estimates based on calculating correlations between additively and non-additively coded genotypes to compute matrices  $K$  and  $D$ . Abbreviations: BMI: Body-mass index; TGLY: Triglycerides; HDL: High-density lipoproteins; Total Chol: Total cholesterol; LDL: Low-density lipoproteins; N: Sample size.

genetic components as one can assume that less heterogeneity in environmental exposures will be present in the population.

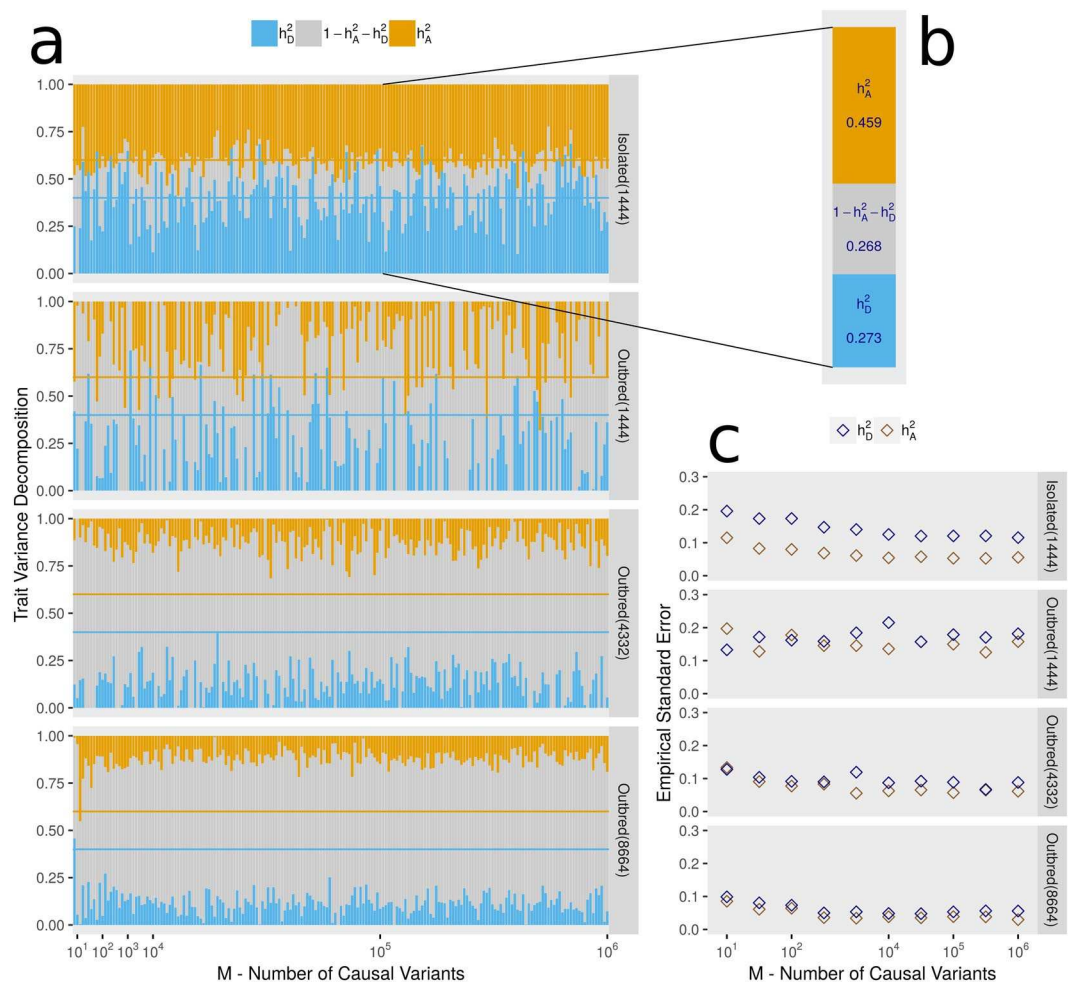
Studying dominance in samples of human twins or siblings can be problematic due to confounding between the sharing of genotypes and shared environmental factors<sup>34</sup>. In a large population isolate, such confounding had been deemed as unlikely to arise due to the extensive range of possible degrees of relatedness between individuals<sup>37,40</sup>. However, the presence of numerous sibling pairs in the sample could easily lead to confounding with the proportions of sharing two alleles IBD ( $IBD = 2$ ) and indeed such confounding between estimates for dominance and shared environmental factors between relatives has recently been observed by Zaitlen *et al.*<sup>41</sup> who performed a study on extended genealogies from the Icelandic populations, itself a moderate isolate.

Genetic dominance has often been considered in the study of various animal species (mammals, poultry, and fish are most commonly studied). Here, by design, confounding with shared environmental factors can often be avoided and extensive and highly accurate pedigree data can be recorded. For many traits, dominance heritability is often found to be significantly different from zero and the inclusion of dominance has been shown to give improved performance of prediction models in animal studies<sup>42–47</sup>. Negative results regarding the improvement of prediction given by including genetic dominance have also been presented (eg. Heidaraitabar *et al.*<sup>48</sup>) and indeed debate continues in regards to the practical value of non-additive variation; for recent reviews we refer the reader to Varona *et al.*<sup>49</sup> and Wolak & Keller<sup>50</sup>. The increased interest in non-additive variation in this domain suggests that there may be value in not discounting such variation in human studies.

We propose to compare heritability estimations in a range of simulated study designs in order to contrast studies in population isolates and in samples of unrelated individuals. In this way we hope to determine whether the differences between studies in isolates and in unrelated samples stem from particular trait architectures, specific population characteristics, or non-equivalence between interpretations of heritability in differing study settings. We will also assess different methods for estimating the matrices  $K$  and  $D$  in an isolate as well as the effect of shared environmental factors between siblings on the estimation of  $h_D^2$  in an isolate. We then proceed to analyse anew the six complex traits displayed in Table 1 in the genetic isolate of Cilento in Southern Italy where we will validate conclusions from our simulation study and search for evidence of significant non-additive genetic components.

## Results

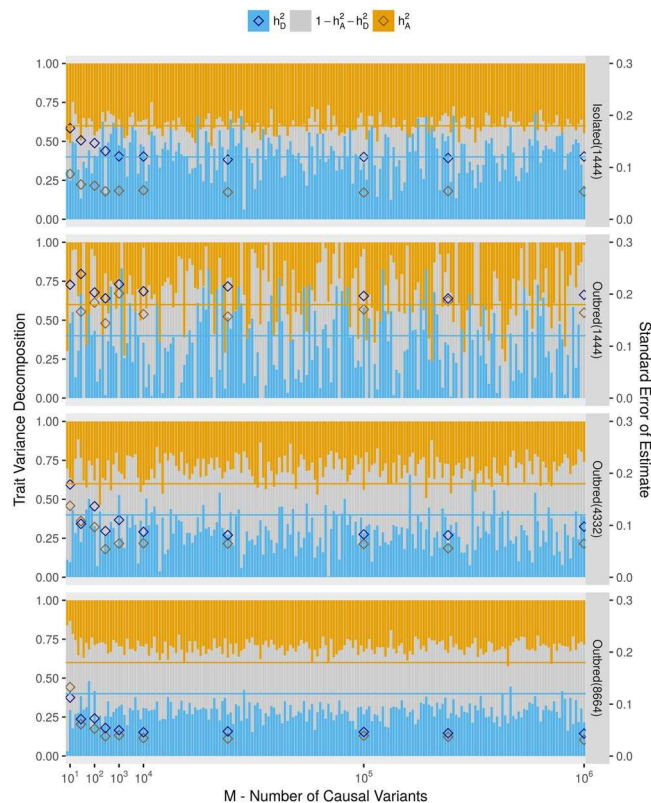
**Effect of population structure.** We assessed the ability of an LMM to detect the additive and dominant genetic variance components in four simulated populations, including firstly one population labelled “Isolated(1444)” which mimics the population structure of the genetic isolate of Cilento from Southern Italy (this cohort is described fully in the Methods section), along with three simulated outbred populations, “Outbred(1444)”, “Outbred(4332)”, and “Outbred(8644)” where the numbers in parentheses indicate the sample sizes. All populations are formed from mosaic haplotypes arising from the UK10K imputation panel<sup>51</sup>. We



**Figure 1.** Estimating heritability components in simulated populations with different structures. **(a)** Maximum Likelihood Estimates (MLEs) of  $h_A^2$  (gold) and  $h_D^2$  (blue) are presented for each simulated phenotype by vertical descending gold and ascending blue bars respectively. The middle grey bars represent the remaining environmental variation ( $1 - h_A^2 - h_D^2$ ). Each phenotype was simulated using different numbers of causal variants ( $M$ ) for each variance component which corresponds to the x-axis. Causal variants are mostly rare, as they are selected completely at random (Causal Variant Scenario A). All MLEs are displayed for the 4 populations either Isolated( $N$ ) or Outbred( $N$ ), where the value of  $N$  denotes the sample size. Horizontal gold and blue lines indicating the values used for simulation ( $h_A^2 = 0.4$ ,  $h_D^2 = 0.4$ ). Matrices  $K$  and  $D$  were calculated using roughly 5.8 million frequent UK10K positions. A missing bar for  $h_A^2$  or  $h_D^2$  indicates the maximum likelihood estimate of the parameter was zero. **(b)** An example of one set of MLEs from section A is given for the population Isolated(1444) and a value of  $M$  of  $10^5$ . **(c)** Gold and blue diamonds represent the empirical standard errors of the MLEs for a selection of values of  $M$ . Simulation repeated 500 times.

simulated phenotypes with the following characteristics:  $h_A^2 = h_D^2 = 0.4$ ,  $M$  causal additive variants, and  $M$  causal dominant variants. Causal variants are selected at random and effect sizes are drawn from normal distributions. Full details of the simulation of genotypes, phenotypes, and population structure are given in the Methods section. We chose 200 values of  $M$  between 1 and 1,000,000, and for some values of  $M$  we repeated the simulation 500 times in order to empirically estimate the standard errors of the estimates of  $h_A^2$  and  $h_D^2$ . We have considered either selecting causal variants completely at random (Causal Variant Scenario A) or from only the set of variants with  $MAF > 0.01$  (Causal Variant Scenario B). Results for Scenarios A and B are presented in Figs 1 and 2, respectively. Here, we have calculated  $K$  and  $D$  for each population as GRMs from a dense set roughly 5.8 million of frequent UK10K variants ( $MAF > 0.05$ ). We also performed the simulation with  $K$  and  $D$  calculated on roughly 170,000 single nucleotide polymorphisms (SNPs) which are those also available in the real data of Cilento (Supplementary Figs 1 and 2).

Fitting the LMM for Isolated(1444) resulted in accurate estimates of  $h_A^2$ , estimations of  $h_D^2$  were also unbiased but were clearly more problematic as seen by the low precision of the estimates. The results from Isolated(1444) were neither affected by the MAF range of the causal variants or the density of the genetic data used to estimate  $K$  and  $D$ . However the, precision of the estimates was low. The estimates in all of the simulated outbred populations were evidently downwardly biased when causal variants were selected completely at random and therefore



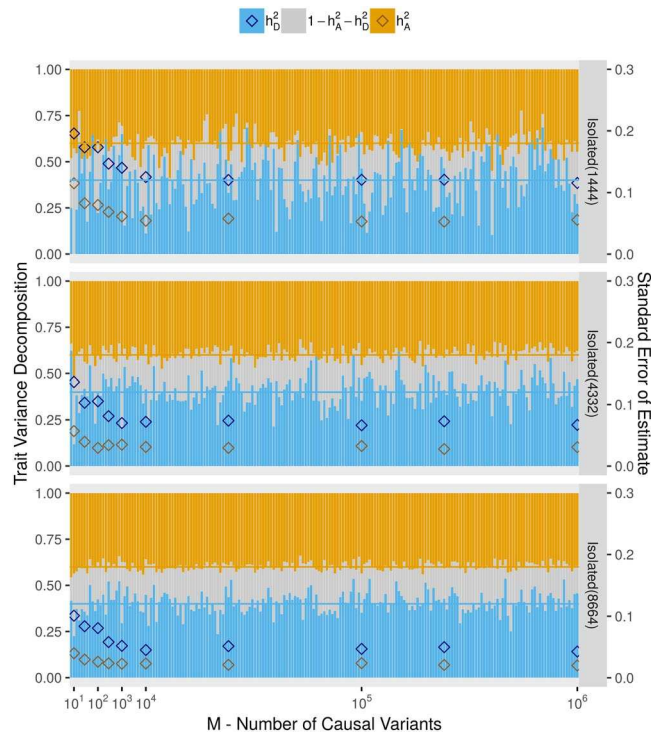
**Figure 2.** Heritability estimates when causal variants are non-rare. Here, phenotypes are simulated by choosing causal variants that are all non-rare, as they are selected to have  $MAF > 0.01$  (Causal Variant Scenario B). Legends and the configuration of this plot are identical to those of Fig. 1A. Here, and for subsequent figures, we overlay the empirical standard error estimates, whose values correspond to the second y-axis on the right of the figure.

included many rare variants as in the UK10K panel (from which all simulated data is based on), over 50% of the variants have a  $MAF$  below 0.01. As the size of the outbred population increases, the precision of the estimates increases but downward biases remain, even when all causal variants are non-rare. The number of causal variants for each variance component ( $M$ ) did not affect the results other than we observed that a small number of causal variants led to lower precision in the results obtained when simulations were repeated. This is shown by the diamonds representing empirical standard errors measured for certain values of  $M$  shown in Figs 1 and 2 and in Supplementary Figs 1 and 2.

We observed increased precision in the estimation of heritability components as we increased the size of the simulated outbred population (Figs 1 and 2). To explore the effect of sample size when studying isolates, we simulated populations with isolate characteristics of sizes 4,332 and 8,664 labelled as Isolated(4332) and Isolated(8664), respectively. A description of the simulation is given in the Methods section. For these populations, we simulated phenotypes under Causal Variant Scenarios A (displayed in Fig. 3) and B (displayed in Supplementary Fig. 3). The precisions of the estimates of  $h_A^2$  and  $h_D^2$  from these larger samples was increased compared to the population Isolated(1444) and estimates remained unbiased for both heritability components. Indeed, the population Isolated(8664) gave the most accurate heritability estimates of all populations thus far considered.

Subsequent analyses will focus on the population Isolated(1444). This will be of particular interest as for this population results are directly comparable with analyses of the real data of Cilento.

**Effect of the choice of relatedness matrices.** To compare methods for calculating  $K$  and  $D$  in a population isolate, we performed similar simulations of phenotypes and tested the estimation of  $h_A^2$  and  $h_D^2$  from our LMM from each of the following strategies:  $K$  and  $D$  calculated from the pedigree of Cilento,  $K$  and  $D$  calculated from exact IBD-sharing recorded during the data simulation (true IBD),  $K$  and  $D$  calculated as GRMs, and finally  $K$  and  $D$  calculated using either the IBDLD<sup>9</sup> or GIBDLD<sup>52</sup> software (see Methods section). Comparisons of off-diagonal elements of these matrices are given in Supplementary Fig. 4a–d. There was clear additional variation in the true proportions of IBD-sharing as compared to the expected values calculated by the pedigree (Supplementary Fig. 4a) and this was captured by the GRMs (Supplementary Fig. 4b). The matrix  $K$  as estimated by a GRM was very similar to the true IBD-sharing probabilities but there were some differences for the matrix  $D$  (Supplementary Figure 4c). The software IBDLD and GIBDLD were able to accurately estimate the true IBD-sharing in the simulated isolate (Supplementary Fig. 4d).

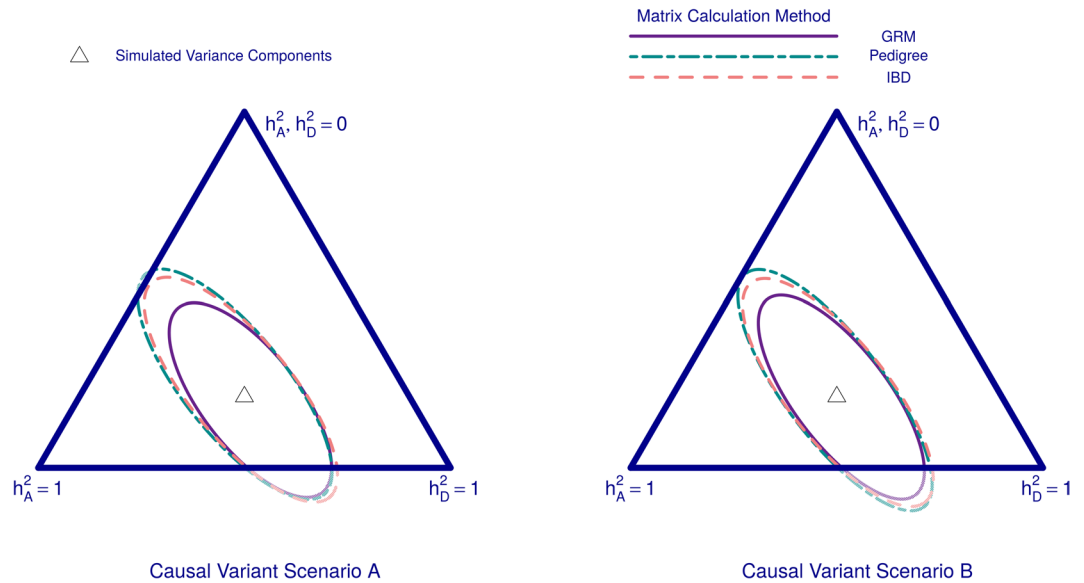


**Figure 3.** Effect of sample size on heritability estimates in an isolate. Estimates of  $h_A^2$  and  $h_D^2$  are compared for populations with isolate characteristics of size 1,444, 4,332, and 8,664. Phenotypes are simulated under Causal Variant Scenario A and under the setting  $h_A^2 = 0.4$ ,  $h_D^2 = 0.4$ . Legends and the configuration of this plot are identical to those of Fig. 2.

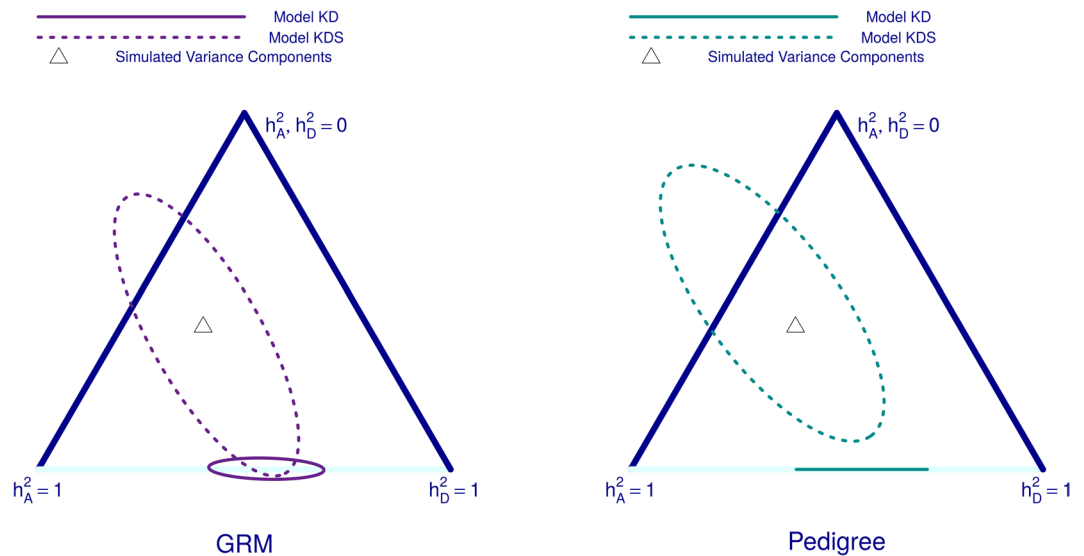
The maximum likelihood estimates (MLEs) of  $h_A^2$  and  $h_D^2$  from each simulated phenotype can be positioned on a simplex to represent the range of possible values of the two parameters  $h_A^2$  and  $h_D^2$ . We present results from 500 simulated phenotypes with  $M = 100,000$  where we display minimal ellipses that contain 95% of all MLEs obtained from each strategy (Fig. 4).

First we compare GRM estimators using roughly 5.8 million frequent ( $MAF > 0.05$ ) UK10K positions with estimates of  $K$  and  $D$  using either pedigree information or true IBD-sharing information (Fig. 4). The method-of-moment GRM estimates appear most accurate, while true IBD-sharing based matrices performed very similarly to expected IBD-sharing matrices derived from the pedigree. This trend in results occurred irrespective of the MAFs of causal variants or the number of causal variants (Fig. 4 and Supplementary Fig. 5). The advantage observed for the GRM method is mostly evident in the estimate of  $h_D^2$  as the ellipses were similarly sized in their minor axes (which describes variation in  $h_A^2$ ) but more differentiable when examining their major axes (which describes variation in  $h_D^2$ ). Indeed, it was on the dominance matrix  $D$  that we observed noticeable differences between off-diagonal elements when comparing GRMs to IBD-based methods (Supplementary Fig. 4c,d). Genomic IBD-based estimates from IBDLD or GIBDL were also used to calculate  $K$  and  $D$ . These Hidden Markov Model (HMM) based methods are not suitable for millions of variants and so were applied to the set of roughly 170,000 SNPs present in all three Cilento villages. These methods were compared to the use of GRMs based on the same set of variants and to using pedigree information or true IBD-sharing information (Supplementary Fig. 6a,b). Such HMM methods could have improved upon the strategy using true IBD proportions as such methods could potentially uncover additional hidden IBD in our simulated population arising from IBD-sharing within the UK10K. We found that IBDLD and GIBDL led to similar estimates of  $h_A^2$  and  $h_D^2$  to using either pedigree information or true IBD-sharing; and again no method was observed to outperform the use of GRMs.

**Effect of the presence of a shared environment.** To investigate how shared environmental factors can affect the estimation of  $h_D^2$  in a population isolate, we simulated additional phenotypes for the population Isolated(1444) under causal variant Scenario A, with  $M = 100,000$ , and with  $h_A^2 = 0.4$ ,  $h_D^2 = 0.4 - h_S^2$ , for the following values of  $h_S^2$ : 0.00, 0.02, 0.05, 0.10, 0.20, and 0.40. For each of these phenotypes, we added positive covariance between the environmental components of siblings. This covariance between siblings creates a confounding between non-additive genetic effects and shared environment effects. Full details of this phenotype simulation and the confounding created are found in the Methods section. We present the estimations of  $h_A^2$  and  $h_D^2$  from analyses with (model KDS) or without (model KD) the inclusion of a variance component (S) indicating pairs of siblings in the sample for  $h_S^2 = 0.20$  (Fig. 5). Throughout, model names indicate the set of variance-covariance matrices included in the LMM. Results for further values of  $h_S^2$  are displayed in Supplementary Fig. 7a–f. Here, we used either GRMs or pedigree based estimates for  $K$  and  $D$  as these were predominantly the methods used in



**Figure 4.** Effect of relatedness matrix estimation method in an isolate. Here, we compare methods of estimating matrices  $K$  and  $D$  for the simulated population isolate ‘Isolated(1444)’.  $K$  and  $D$  are estimated using either genetic relationship matrices (GRM), Pedigree information, or true IBD-sharing (IBD). Results are displayed on a simplex governed by the two parameters  $h_A^2$  and  $h_D^2$ , which both could range between 0 and 1. The heritability scenario used to simulate all phenotypes ( $h_A^2 = h_D^2 = 0.4$ ) is marked by the triangular point in the centre of each simplex. Minimal ellipses containing 95% of the maximum likelihood estimates (MLEs) from 500 simulated phenotypes under either Causal Variant Scenario A or B (see Figs 1 and 2) are presented. Here, phenotypes are simulated from a large set of causal variants ( $M = 100,000$ ).



**Figure 5.** Effect of shared environmental factors on heritability component estimates in an isolate. Comparison of estimates of  $h_A^2$  and  $h_D^2$  under models with and without a shared environment component (model KDS and model KD, respectively). As in Fig. 4, minimal ellipses containing 95% of the maximum likelihood estimates (MLEs) from 500 simulated phenotypes but now under the setting  $h_A^2 = 0.4$ ,  $h_D^2 = 0.2$ ,  $h_S^2 = 0.2$ . Matrices  $K$  and  $D$  are calculated either using genotype relationship matrices (GRMs) or pedigree information. In the case of model KD when using pedigree information (right), all MLEs were found to be directly on the bottom edge of the simplex, and so the minimal ellipsoid degenerated into a line segment. Here, phenotypes are simulated from a large set of causal variants ( $M = 100,000$ ).

forementioned studies that calculated dominant genetic components for widely studied traits (Table 1). Our simulations indicate that once a significant correlation between siblings is introduced, our unadjusted estimates for the broad-sense heritability became close to or equal to 1 (MLEs falling on the bottom axis of the simplex for model KD). Again, in these analyses using GRMs appears to outperform the use of pedigree based estimates.

Phenotype	GRM Model: K		GRM Model: KD		GRM Model: KS		GRM Model: KDS			Pedigree Model: K		Pedigree Model: KD		Pedigree Model: KS		Pedigree Model: KDS		
	$h_A^2$	$h_D^2$	$h_A^2$	$h_D^2$	$h_A^2$	$h_S^2$	$h_A^2$	$h_D^2$	$h_S^2$	$h_A^2$	$h_D^2$	$h_A^2$	$h_D^2$	$h_A^2$	$h_S^2$	$h_A^2$	$h_D^2$	$h_S^2$
Height	0.76	0.74	0.13	0.74	0.04	0.04	0.74	0.12	0.01	0.75	0.74	0.15	0.74	0.04	0.74	0.15	0.00	
BMI	0.40	0.35	0.58	0.31	0.23	0.31	0.00	0.23	0.44	0.35	0.65	0.35	0.21	0.35	0.00	0.21		
TGLY	0.27	0.24	0.26	0.21	0.11	0.21	0.00	0.11	0.28	0.23	0.45	0.23	0.11	0.23	0.41	0.01		
HDL	0.49	0.49	0.00	0.44	0.02	0.44	0.00	0.02	0.48	0.49	0.00	0.48	0.01	0.48	0.00	0.01		
Total Chol	0.29	0.23	0.55	0.23	0.18	0.22	0.27	0.12	0.29	0.21	0.72	0.22	0.18	0.21	0.47	0.06		
LDL	0.32	0.25	0.52	0.24	0.17	0.23	0.29	0.10	0.33	0.24	0.66	0.24	0.16	0.24	0.45	0.06		

**Table 2.** Maximum likelihood estimates for the contribution of each variance components considered in a Linear Mixed Model (LMM). Model names refer to the set of variance components included. *K* denotes the additive genetic component, *D* the non-additive or dominant genetic component, and *S* the component accounting for shared environmental effects between siblings. The previously reported results from Table 1 can be compared to our results under the model KD. Matrices *K* and *D* are calculated either as genetic relationship matrices (GRMs) or from pedigree information.

Adjusting for such correlation between siblings in the LMM did substantially correct for this bias but it is clear that in a population such as Cilento, there is little hope in effectively discriminating between dominant genetic variability and shared environmental factors between siblings if both occur simultaneously.

An obvious approach to avoid such ambiguity would be to remove one individual from every pair of siblings but in Cilento this would greatly reduce the sample size. Therefore, we removed one individual from each pair of siblings from the population Isolated(8664), creating a sibling free population which we label as “Isolated(5136)\_nosibs”. Full details of the simulation of this population are found in the Methods section. From this population, we observed improved estimates of both  $h_A^2$  and  $h_D^2$  as compared to the Outbred(8664) under Causal Variant Scenario A; with the two populations performing similarly under Causal Variant Scenario B (Supplementary Fig. 8a,b). When compared to the results from Isolated(1444), the absence of pairs of individuals with high  $IBD = 2$  probabilities led to a slight underestimation of  $h_D^2$ , but the increased sample size led to lower standard errors across replications of phenotype simulation. If no dominant genetic component was simulated, the Isolated(1444) population was most likely to give large (more erroneous) estimates for  $h_D^2$  compared to Isolated(5136)\_nosibs and Outbred(8664) (Supplementary Fig. 8c,d).

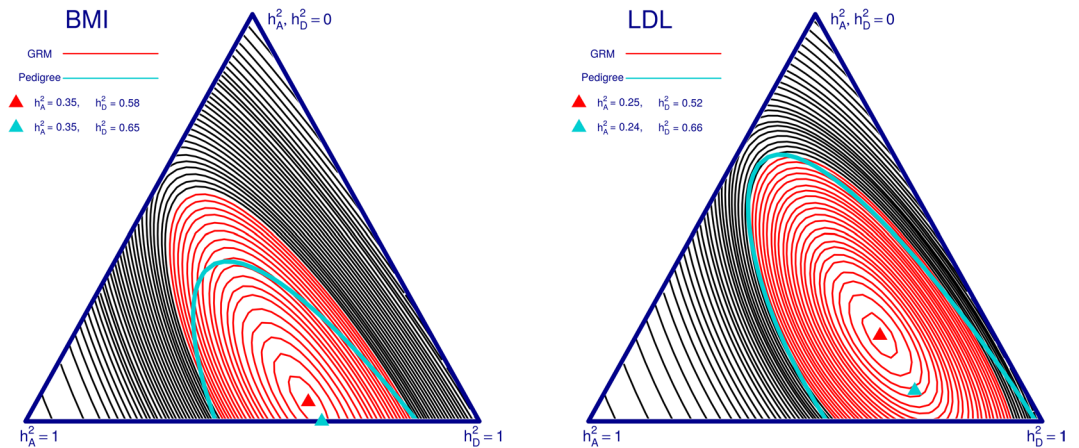
**Analysis of the Cilento Isolates.** We first calculated the matrices *K* and *D* using different approaches and then compared the resulting values. We calculated *K* and *D* using either the pedigree information, or as GRMs using genotype data before or after imputation. Results were in accordance with those from the simulated population isolate Isolated(1444) (Supplementary Fig. 9). However, we observed greater differences between the off-diagonal elements calculated with the pedigree and those in the GRMs when analysing the real Cilento data as compared to Isolated(1444). This is likely to stem from the explicit use of the pedigree information within the simulation. The inclusion of imputed variants led to similar estimates for the matrices *K* and *D* (Supplementary Fig. 10).

Following quality control and imputation (full details are given in the Supplementary Materials); we fitted LMMs to the data in Cilento having estimated matrices *K* and *D* as GRMs (using all variants with  $MAF > 0.05$  and imputation quality score  $> 0.7$ ). Several traits displayed significant dominant genetic components and our results (Table 2) are not distant to those found in the literature of previous studies in population isolates (Table 1). LMMs were fitted with different combinations of the matrices *K*, *D*, and *S* (the sibling indicator matrix). Full details are given in the Methods section; as above in the simulation study, the model names indicate the variance components included in the LMM. The orthogonality between the additive and non-additive genetic components is apparent as estimates for  $h_A^2$  are similar across models with or without the inclusion of the non-additive genetic variance component. For each phenotype considered, we estimated the entire likelihood surface as well as the MLEs for the parameters  $h_A^2$  and  $h_D^2$  under the model KD. Likelihood surfaces governed by  $h_A^2$  and  $h_D^2$  for BMI and LDL are displayed in Fig. 6 and corresponding results for other traits are found in Supplementary Fig. 11a–d. We observed similar profiles in the likelihood contours as were observed in the distributions of MLEs from repeated phenotype simulation in the simulation study. We are able to have a reasonable level of confidence in the estimates of the additive genetic component, but the dominant genetic component is problematic as our confidence regions are very wide. The MLEs found when using pedigree information to estimate matrices *K* and *D* had equivalent estimates for the additive genetic components to the MLEs found when using GRMs, however the dominant genetic components were always estimated as equal or greater when using pedigree information.

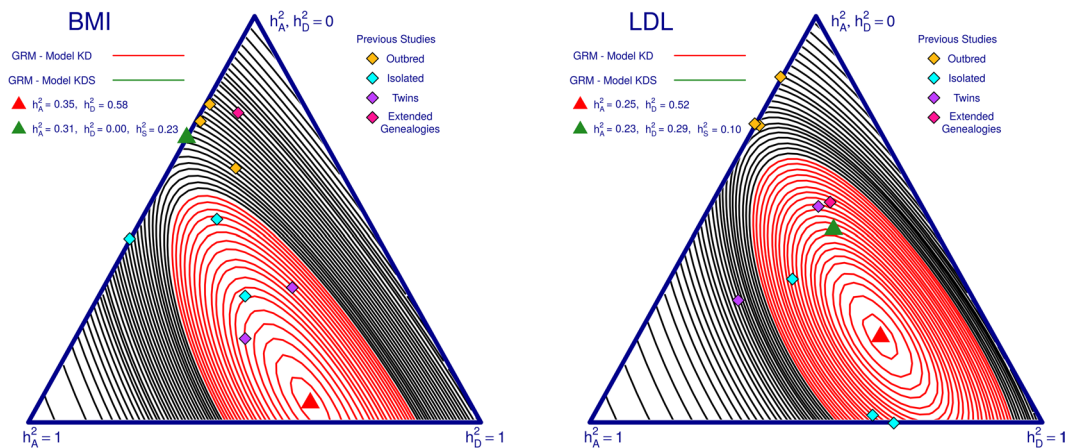
The traits of BMI, LDL, and Total Chol were all estimated as having dominant genetic components higher than their respective additive genetic components in the KD model. By examining the 95% confidence regions, there is some indication that the dominant genetic components are unlikely to be equal to zero. This is due to the observation that the red zones either do not intersect or only briefly intersect the upper left boundary ( $h_D^2 = 0$ ) of their respective simplexes (Fig. 6 and Supplementary Fig. 11d).

Adding the shared environmental component between siblings drastically changed the estimates of  $h_D^2$  for many traits as seen by comparing models KD and KDS in Table 2; for our two example traits (BMI and LDL) we present again the likelihood profiles from the original analysis and then new MLE and 95% confidence interval for  $h_A^2$  and  $h_D^2$  from the KDS model as well as the previous estimates for  $h_A^2$  and  $h_D^2$  found in the literature (Fig. 7). Equivalent plots for our other studied traits are given in Supplementary Fig. 12a–d.





**Figure 6.** Heritability analysis for BMI and LDL in Cilento. Black contours represent the likelihood profile from the model KD (see Fig. 5), with matrices  $K$  and  $D$  calculated as genetic relationship matrices (GRMs). The red zone represents the 95% confidence interval for the red maximum likelihood estimate (MLE) (red triangular peak). The corresponding MLE and 95% confidence boundary for the analysis using pedigree information to estimate  $K$  and  $D$  are added to the plot in blue.



**Figure 7.** Effect of shared environmental factors on heritability analysis for BMI and LDL in Cilento. Here we compare models KD and KDS (see Fig. 5) for the two traits in Cilento. Black contours represent the likelihood profile for the model KD, with the red zone indicating the 95% confidence interval for the red maximum likelihood estimate (MLE) (red triangular peak). The corresponding MLE for the KDS model is added in green. We also add in the previously observed estimates from the literature (Table 1).

For BMI, the unadjusted heritability estimate was distant from previously reported results, but once we allow for a shared environmental component between siblings, we find similar estimates for  $h_D^2$  to previous studies. For LDL, the unadjusted heritability estimates lay close to previous results from isolated populations, with the adjusted results moving towards previous results in studies of outbred populations but remaining quite large at 0.29.

## Discussion

Across all analyses, whether on simulated or real Cilento data, we observed that estimates of  $h_D^2$  had less precision than estimates of  $h_A^2$ .

Isolated populations exhibit favourable characteristics for uncovering the contribution of  $h_D^2$  due to the increased proportions of  $IBD = 2$  between individuals. Our simulation elaborates on this by showing that in the absence of shared environmental effects, estimating  $h_D^2$  (and indeed  $h_A^2$ ) from an LMM in a population isolate will yield unbiased results for polygenic phenotypes with wide a range of characteristics. However, we saw that shared environmental factors pose a non-trivial obstacle to analysing dominant genetic variance of a trait in an isolated population. In the presence of even small shared environmental effects between siblings in the simulated isolate, we observed that estimates of  $h_D^2$  are heavily biased. Improved estimates may be attainable by including a sibship matrix in the variance decomposition analysis but accurately partitioning between dominance effects and shared environmental effects through linear mixed modelling in a population such as Cilento may not be possible.

We compared different methods to estimate the covariance matrices  $K$  and  $D$ . In the simulated isolate, the precision of the estimates of  $h_D^2$  was either larger or equivalent when using GRMs as compared to IBD-based methods. This had previously also been noted by Browning & Browning<sup>53</sup> when estimating  $h_A^2$ . Furthermore, it would appear that only a relatively small number of SNPs are required to compute such GRMs in an isolate as using far denser sets of variants (either in our simulation or through imputation in the Cilento dataset) did not noticeably affect the fitting of the LMM. The advantage observed for GRMs could be because they can capture similarities between all types of pairs of individuals in the isolate; including similarities not described by the recorded pedigree structure or originating before the founding event of the population. Therefore this approach combines the classical interpretation of heritability regarding closely related individuals with the more recent approaches involving samples of unrelated individuals.

Foreseeably, the simulated outbred populations led to underestimation of both  $h_A^2$  and  $h_D^2$  in most of the settings of phenotype simulation. This may go some way to explain the differences between estimates of  $h_D^2$  that we observed in the literature for many complex traits. Our results suggest that observing very different estimations for non-additive genetic components between isolates and outbred populations could indicate the presence of many causal variants that occur at low frequencies across populations and that have non-zero dominant genetic effects. However, such an observation could also indicate the presence of bias due to the shared environmental factors in the studies of isolates. We note that estimation from outbred populations can also suffer from biases arising from shared environmental factors due to hidden structures existing within the population; a scenario that we have not considered in our simulation study. Population stratification within a cohort is a known example of a structure that can lead to bias in heritability studies of unrelated individuals<sup>54,55</sup>.

The heritability analyses that we have carried out in Cilento did indeed suggest the presence of non-additive genetic variance for some of the traits considered. However, the phenotypes studied in Cilento behaved in similar ways to the simulated phenotypes with added non-genetic correlation between siblings. The simulation study suggested that even a very small shared environmental effect between siblings could result in the disparate heritability estimates we observed in Cilento between fitting LMMs with and without a variance component for covariance between siblings. When the simulated shared environmental component was large, broad-sense heritability estimates approached 1; this is a result we observed in both previous studies of isolates for many traits<sup>38,39</sup> (see Table 1) and in Cilento for the trait BMI. Combining this observation with the wide observed ranges of estimates for  $h_D^2$  in the literature strongly suggests that previous results in isolates have thus far been inflated by shared environmental effects and that  $h_D^2$  statistics have been overestimated. For a trait such as LDL, we still observed high estimates for  $h_D^2$  even when accounting for a shared environment effect in the model, a result which our simulation suggests would be unlikely if indeed  $h_D^2 = 0$  for this trait.

It has been argued that the classical separation of the two additive and non-additive genetic components may lead to higher estimates for the additive genetic variance over the non-additive genetic variance<sup>56</sup>. However, proposed alternative definitions are far less interpretable and lead to variance decompositions with less applicable value. Higher order non-additive genetic variance components could be contributing to our estimates of dominance in Cilento<sup>5</sup>. Indeed, we recognise that ignoring the presence of epistatic effects has been shown to lead to overestimations of  $H^2$  by Zuk *et al.*<sup>57</sup> who also proposed a non-parametric method for estimating heritability in a population isolate. Such approaches require large samples of pairs of individuals with identical expected relatedness coefficients. Similar approaches include those based on Haseman-Elston regression<sup>58</sup> and studies focusing on populations of siblings or nuclear families. However, for the data of Cilento such methods proved not to be applicable due to the variety of relationships between pairs, such that looking at each pair type separately resulted in sample sizes too small to provide realistic estimations. There exist a wide range of sophisticated approaches for calculating narrow-sense heritability in sample of unrelated individuals<sup>59–61</sup>. Zaitlen *et al.*<sup>41</sup> proposed to dissect narrow-sense heritability in samples containing close relatives by splitting variance between GRMs and thresholded GRMs, and isolated populations could prove a valuable resource for future studies using such approaches. However, as we include non-additive genetic components and wish to compare our results to studies using pedigree based methods, we have not explored such concepts here.

In this study, we have demonstrated various phenomena which can either result in under-estimation of  $h_D^2$  in studies of outbred populations or over-estimation in studies including closely related individuals. At this juncture, the existence of significant non-zero dominant genetic variation for many traits remains uncertain, but this could be elucidated through the continued gathering of estimates from diverse populations. Whilst different populations harbour differing levels of environmental variation, and hence one cannot expect agreement on heritability estimations, studies of isolated populations could lead to more reliable conclusions as to the existence or non-existence of genetic dominance for complex traits. If significant estimates for  $h_D^2$  are found when accounting for a shared environment effect between siblings, this is indicative of a true non-zero dominance component.

One possible future direction would be to increase the sample size in a study of an isolate. However, as this will not usually be feasible for a single isolate, one strategy that could be particularly interesting would be to combine data from several isolates with similar ancestral origins. Such an approach could give high precisions of the estimates of both  $h_A^2$  and  $h_D^2$  due to the large sample size. Importantly, this strategy could also provide a large enough sample to complete analyses without sibling pairs, and to facilitate appropriate sensitivity analyses regarding the presence of siblings.

## Methods

**The Cilento Isolate.** The Cilento isolate comprise three villages from the South of Italy; Campora, Cardile, and Gioi. Pedigree, phenotypic, and genetic data have previously been gathered as part of the Cilento Study. A pedigree structures which connects all three village has been reconstructed from parish records. The three villages have been shown to represent characteristics of population isolates intermediate between the large isolate

population of Iceland<sup>62</sup> and the highly isolated Hutterite population<sup>63,64</sup>. Aggregating over the three villages, we have a pedigree of 7,585 members including 1,444 genotyped members. The high quality of the reconstructed genealogy in Cilento makes it an appropriate tool for simulating a realistic example of data from an isolated population. Individuals from Campora and Cardile have been genotyped on an Illumina 370 K array, whilst individuals from Gioi have been genotyped on an Illumina HumanOmniExpress array. Deep phenotyping has been performed in Cilento for a range of anthropometric, cardiometabolic, and haematological traits. For the purposes of this study, we have concentrated on phenotypes that have been often analyzed in the literature of both other population isolates and in samples of unrelated individuals (Supplementary Table 1).

**Simulation of genotypes and phenotypes.** To create simulated datasets, we created mosaic haplotypes using the same stochastic recombination model as in the generation of control individuals by the software HapGen2<sup>65</sup>. We took the UK10K imputation panel as reference haplotypes having first removed one individual from every pair of twins present in the panel. To simulate unrelated individuals we sampled 22 pairs of mosaic chromosomes, where each section of their mosaics is copied from a randomly sampled haplotype from the pool of UK10K haplotypes. In this manner, we created a sample of 8,664 ( $6 \times 1,444$ ) unrelated individuals. To create isolate type data, for each chromosome, we randomly selected 200 UK10K haplotypes, from which 2,940 mosaic haplotypes were simulated in order to simulate the 1,470 founders of the combined pedigree of Cilento. This set of founder haplotypes were supplied to the software Genedrop (part of the MORGAN<sup>66</sup> package) along with the pedigree of Cilento in order to simulate phased genetic data for the 1,444 genotyped members of Cilento. Our gene-dropping approach was identical to the methods used in Herzig *et al.*<sup>67</sup> We have made comparisons on four potential populations: the 1,444 individuals from Genedrop with isolate type data, labelled “Isolated(1444)”, and three possible sets of the 8,664 simulated unrelated individuals: “Outbred(1444)”, “Outbred(4332)”, “Outbred(8664)”, that represent outbred populations of the same size as Cilento, three times the size, and six times the size, respectively. We chose this range of sample sizes based on an analysis of the variance of eigenvalues<sup>68</sup> of GRMs estimated on the populations Isolated(1444) and Outbred(1444) (Supplementary Materials and Supplementary Table 2). The choice of 200 haplotypes for the generation of founder haplotypes for Cilento stems from the previous work which estimated that 96.7% of the genetic diversity in Campora is accounted for by 17 female and 20 male lineages<sup>63</sup>. This would suggest that 74 ( $37 \times 2$ ) autosomal haplotypes would be appropriate for the generation of simulated data for Campora and we decided to scale this up to 200 for the generation of simulated data for the three villages. We checked that this created simulated data with a similar structure as the observed data in Cilento (Supplementary Table 2 and Supplementary Fig. 13).

Our method for simulating isolate-type data requires a pedigree for gene-dropping. To create larger datasets with isolate characteristics, we used the Cilento pedigree multiple times. In detail, we simulated six populations of size 1,444, each using the Cilento pedigree but with different random draws of founding haplotypes. We then combined the first three and all six of these populations to create the populations Isolated(4332) and Isolated(8664), respectively. In addition, we randomly discarded one individual from each sibling pair of the population Isolate(8664) to create a population with no sibling pairs of size 5,136, labelled as “Isolated(5136)\_nosibs”.

Phenotypes were simulated repeatedly for each population as the sum of normally distributed errors (Equation 3).

$$Y = \beta_A^T G_A + \beta_D^T G_D + \varepsilon \quad (3)$$

$G_A$  and  $G_D$  are the additive genetic components of the genotypes of the randomly selected  $M$  causal additive variants and the non-additive genetic components of the randomly selected  $M$  causal dominant variants, respectively. Effect sizes  $\beta_A$  and  $\beta_D$  were drawn from normal distributions. Variants may exhibit both additive and dominant effects and a maximum of  $2M$  variants could have non-zero effect sizes. We varied the heritability by scaling the effect-sizes accordingly in the knowledge that  $\tau_A = \sum \beta_A^2$  and  $\tau_D = \sum \beta_D^2$ . We have simulated a range of possible phenotype characteristics by varying the number of causal variants and the MAFs of causal variants.

To estimate the variance parameters, and hence heritability, we fitted the model of Equation 1 in the R-package ‘Gaston’<sup>69</sup> and estimated parameters  $\tau_A$ ,  $\tau_D$ , and  $\sigma_\varepsilon^2$  using Average Information Restricted Maximum Likelihood Estimation (AIREML)<sup>70</sup>. Matrices  $K$  and  $D$  were estimated using the method-of-moment techniques described in Zhu *et al.*<sup>12</sup>, and we either used all variants present on the UK10K, or the variants present in the real data from all three Cilento villages. The exact set of variants used for these calculations were those with MAF  $> 0.05$  and those passing a quality control threshold on the Hardy-Weinberg p-values ( $> 10^{-5}$ ).

In the case of Isolated(1444), we also estimated  $K$  and  $D$  from the pedigree structure of Cilento using software IdCoefs<sup>4</sup> that calculates  $\Delta_1, \dots, \Delta_9$  through the recursive algorithm described by Karig<sup>71</sup>. Furthermore, we were able to record the origin of every mosaic segment simulated during the HapGen based and gene-dropping stages. This allowed us to calculate true proportions of IBD-sharing between every pair of individuals in the Isolated(1444) population. We also tested the software IBDLD<sup>9</sup> and GIBDLD<sup>52</sup> which directly estimate  $\Delta_1, \dots, \Delta_9$ . For IBDLD, we used the LD-RR mode, default parameters, and we supplied the software with the expected values of  $\Delta_1, \dots, \Delta_9$  between all pairs from the pedigree (calculated by IdCoefs). Conversely, GIBDLD used only the genotypes; we also ran this software with default parameters. For both IBDLD and GIBDLD, we used only the SNPs present in both genotyping arrays in Cilento as the software were not designed for sequence data.

Here we introduce the sibship matrix, denoted as  $S$ , which has values of 1 on the diagonal and at every off-diagonal element corresponding to pairs of siblings in the sample; all other entries are zero. To simulate phenotypes for the population Isolated(1444) with additional correlation between pairs of siblings, approximating an effect of shared environmental exposure, we simulated phenotypes under the same model as Equation 3 except that the environmental components were no longer drawn independently from normal distributions, but from a

multi-variate normal distribution with zero mean and a covariance structure of  $(\sigma_E^2 + \sigma_S^2)I_N + \sigma_S^2S$ ; a matrix with  $\sigma_E^2 + \sigma_S^2$  on the diagonal and  $\sigma_S^2$  on every off-diagonal entry corresponding to a pair of siblings in the sample. We chose values of  $\sigma_S^2$  in order to create phenotypes with  $h_S^2$ : 0.00, 0.02, 0.05, 0.10, 0.20, and 0.40 where  $h_S^2 = \sigma_S^2 / (\tau_A + \tau_D + \sigma_S^2 + \sigma_E^2)$ .

**Analysis of Cilento Data.** After quality control on both phenotypes and genetic data (details in the Supplementary Materials), we used the same approach as with the simulated data to estimate the heritabilities of the seven traits considered in this study. The only difference being that for the analyses of Cilento data, we added the following covariates to the LMM: age, sex, age  $\times$  sex, and indicators of village membership (Campora, Cardile, or Gioi). For one trait (Triglycerides) we transformed the phenotype to a logarithmic scale, whereas other traits were left untransformed after excluding very small numbers of outliers. LDL and Total Chol were both pre-adjusted for use of lipid-lowering medication. Matrices  $K$  and  $D$  were again estimated on the basis of pedigree or genetic information. To calculate GRMs from genetic data, we were restricted to using the set of variants on the intersections of the two arrays used for genotyping of Cilento data. As this set was relatively sparse, we also performed genetic imputation with the following pipeline: phasing by SHAPEIT2<sup>72</sup> with the “duohmm” option<sup>73</sup> and informed by the Haplotype Reference Consortium<sup>74</sup> (HRC) reference panel followed by imputation by IMPUTE4<sup>75</sup> with the HRC as the reference panel.  $K$  and  $D$  were then computed on hard called imputed genotypes<sup>76,77</sup> after removing variants with imputation quality scores below 0.7.

In a recent study of the Icelandic population, Young *et al.*<sup>78</sup> presented an IBD-based method for nuclear families in the Icelandic population aimed at eliminating environmental bias by looking at deviations in observed kinship from expected values. In Cilento data, the sample size precluded this approach as there are insufficient numbers of pairs of individuals with the required expected level of IBD-sharing and with both sets of parent’s genotypes. However, we are able to add a shared environment effect by adding into our model a variance component indicating pairs of individuals who share the same mother. A similar approach was shown to lead to unbiased results in many simulation settings in Young *et al.*<sup>78</sup> As pairs of siblings have by far the highest probability of sharing two alleles IBD as each locus (one chance in four), correlations caused by shared environmental exposures between siblings are very likely to confound the estimation of  $h_D^2$ . If there is significant confounding, this should be indicated by a large difference in results when including such a matrix indicating siblings in the LMM. We fitted four LMMs for every trait which we denote as model K, model KD, model KS, and model KDS to indicate the set of variance-covariance matrices included in the model.

## Data Availability

The UK10K panel of haplotypes is available from the European Genome-phenome Archive and the simulation scripts are available from Anthony Francis Herzig (anthony.herzig@inserm.fr) on reasonable request. The Cilento datasets analysed during the current study are available from Marina Ciullo (marina.ciullo@igb.cnr.it) on reasonable request and on a collaborative basis.

## References

- Tenesa, A. & Haley, C. S. The heritability of human disease: estimation, uses and abuses. *Nat Rev Genet* **14**, 139–49 (2013).
- Yang, J., Zeng, J., Goddard, M. E., Wray, N. R. & Visscher, P. M. Concepts, estimation and interpretation of SNP-based heritability. *Nat Genet* **49**, 1304–1310 (2017).
- Fisher, R. A. XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Transactions of the Royal Society of Edinburgh* **52**, 399–433 (1919).
- Abney, M., McPeck, M. S. & Ober, C. Estimation of variance components of quantitative traits in inbred populations. *Am J Hum Genet* **66**, 629–50 (2000).
- Young, A. I. & Durbin, R. Estimation of epistatic variance components and heritability in founder populations and crosses. *Genetics* **198**, 1405–16 (2014).
- Jacquard, A. *The genetic structure of populations*, (Springer-Verlag, New York - Heidelberg - Berlin, 1974).
- Falconer, D.S. *Introduction to Quantitative Genetics*, (Oliver and Boyd, Edinburgh and London, 1960).
- Gusev, A. *et al.* Whole population, genome-wide mapping of hidden relatedness. *Genome Res* **19**, 318–26 (2009).
- Han, L. & Abney, M. Identity by descent estimation with dense genome-wide genotype data. *Genet Epidemiol* **35**, 557–67 (2011).
- Browning, B. L. & Browning, S. R. A fast, powerful method for detecting identity by descent. *Am J Hum Genet* **88**, 173–82 (2011).
- Evans, L. M. *et al.* Narrow-sense heritability estimation of complex traits using identity-by-descent information. *Heredity (Edinb)* **121**, 616–630 (2018).
- Zhu, Z. *et al.* Dominance genetic variation contributes little to the missing heritability for human complex traits. *Am J Hum Genet* **96**, 377–85 (2015).
- Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* **42**, 565–9 (2010).
- Speed, D. & Balding, D. J. Relatedness in the post-genomic era: is it still useful? *Nat Rev Genet* **16**, 33–44 (2015).
- Powell, J. E., Visscher, P. M. & Goddard, M. E. Reconciling the analysis of IBD and IBS in complex trait studies. *Nat Rev Genet* **11**, 800–5 (2010).
- Thompson, E. A. Identity by descent: variation in meiosis, across genomes, and in populations. *Genetics* **194**, 301–26 (2013).
- Lee, S. H., Wray, N. R., Goddard, M. E. & Visscher, P. M. Estimating missing heritability for disease from genome-wide association studies. *Am J Hum Genet* **88**, 294–305 (2011).
- Visscher, P. M., Yang, J. & Goddard, M. E. A commentary on ‘common SNPs explain a large proportion of the heritability for human height’ by Yang *et al.* (2010). *Twin Res Hum Genet* **13**, 517–24 (2010).
- Pritchard, J. K. Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet* **69**, 124–37 (2001).
- Pritchard, J. K. & Cox, N. J. The allelic architecture of human disease genes: common disease-common variant...or not? *Hum Mol Genet* **11**, 2417–23 (2002).
- Golan, D. & Rosset, S. Accurate estimation of heritability in genome wide studies using random effects models. *Bioinformatics* **27**, i317–23 (2011).
- Vinkhuyzen, A. A., Wray, N. R., Yang, J., Goddard, M. E. & Visscher, P. M. Estimation and partition of heritability in human populations using whole-genome analysis methods. *Annu Rev Genet* **47**, 75–95 (2013).

23. Golan, D., Lander, E. S. & Rosset, S. Measuring missing heritability: inferring the contribution of common variants. *Proc Natl Acad Sci USA* **111**, E5272–81 (2014).
24. Lee, S. H., Goddard, M. E., Visscher, P. M. & van der Werf, J. H. Using the realized relationship matrix to disentangle confounding factors for the estimation of genetic variance components of complex traits. *Genet Sel Evol* **42**, 22 (2010).
25. Hill, W. G. & Weir, B. S. Variation in actual relationship as a consequence of Mendelian sampling and linkage. *Genet Res (Camb)* **93**, 47–64 (2011).
26. Leutenegger, A. L. *et al.* Estimation of the inbreeding coefficient through use of genomic data. *Am J Hum Genet* **73**, 516–23 (2003).
27. Wang, K., Gaitsch, H., Poon, H., Cox, N. J. & Rzhetsky, A. Classification of common human diseases derived from shared genetic and environmental determinants. *Nat Genet* **49**, 1319–1325 (2017).
28. Stanton-Geddes, J., Yoder, J. B., Briskine, R., Young, N. D. & Tiffin, P. Estimating heritability using genomic data. *Methods in Ecology and Evolution* **4**, 1151–1158 (2013).
29. Berenos, C., Ellis, P. A., Pilkington, J. G. & Pemberton, J. M. Estimating quantitative genetic parameters in wild populations: a comparison of pedigree and genomic approaches. *Mol Ecol* **23**, 3434–51 (2014).
30. Gay, L., Siol, M. & Ronfort, J. Pedigree-free estimates of heritability in the wild: promising prospects for selfing populations. *PLoS One* **8**, e66983 (2013).
31. Wang, H., Misztal, I. & Legarra, A. Differences between genomic-based and pedigree-based relationships in a chicken population, as a function of quality control and pedigree links among individuals. *J Anim Breed Genet* **131**, 445–51 (2014).
32. Hill, W. G., Goddard, M. E. & Visscher, P. M. Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genet* **4**, e1000008 (2008).
33. Nolte, I. M. *et al.* Missing heritability: is the gap closing? An analysis of 32 complex traits in the Lifelines Cohort Study. *Eur J Hum Genet* **25**, 877–885 (2017).
34. Chen, X. *et al.* Dominant Genetic Variation and Missing Heritability for Human Complex Traits: Insights from Twin versus Genome-wide Common SNP Models. *Am J Hum Genet* **97**, 708–14 (2015).
35. van Dongen, J., Willemsen, G., Chen, W. M., de Geus, E. J. & Boomsma, D. I. Heritability of metabolic syndrome traits in a large population-based sample. *J Lipid Res* **54**, 2914–23 (2013).
36. Boomsma, D. I. *et al.* An Extended Twin-Pedigree Study of Neuroticism in the Netherlands Twin Register. *Behav Genet* **48**, 1–11 (2018).
37. Abney, M., McPeck, M. S. & Ober, C. Broad and narrow heritabilities of quantitative traits in a founder population. *Am J Hum Genet* **68**, 1302–7 (2001).
38. Pilia, G. *et al.* Heritability of cardiovascular and personality traits in 6,148 Sardinians. *PLoS Genet* **2**, e132 (2006).
39. Traglia, M. *et al.* Heritability and demographic analyses in the large isolated population of Val Borbera suggest advantages in mapping complex traits genes. *PLoS One* **4**, e7554 (2009).
40. Vitart, V. *et al.* Heritabilities of ocular biometrical traits in two croatian isolates with extended pedigrees. *Invest Ophthalmol Vis Sci* **51**, 737–43 (2010).
41. Zaitlen, N. *et al.* Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. *PLoS Genet* **9**, e1003520 (2013).
42. Sun, C., VanRaden, P. M., Cole, J. B. & O’Connell, J. R. Improvement of prediction ability for genomic selection of dairy cattle by including dominance effects. *PLoS One* **9**, e103934 (2014).
43. Moghaddar, N. & van der Werf, J. H. J. Genomic estimation of additive and dominance effects and impact of accounting for dominance on accuracy of genomic evaluation in sheep populations. *J Anim Breed Genet* **134**, 453–462 (2017).
44. Nagy, I. *et al.* The contribution of dominance and inbreeding depression in estimating variance components for litter size in Pannon White rabbits. *J Anim Breed Genet* **130**, 303–11 (2013).
45. Serenius, T., Stalder, K. J. & Puonti, M. Impact of dominance effects on sow longevity. *J Anim Breed Genet* **123**, 355–61 (2006).
46. Joshi, R., Woolliams, J. A., Meuwissen, T. & Gjoen, H. M. Maternal, dominance and additive genetic effects in Nile tilapia; influence on growth, fillet yield and body size traits. *Heredity (Edinb)* **120**, 452–462 (2018).
47. Ebrahimi, K., Dashab, G. R., Faraji-Arough, H. & Rokouei, M. Estimation of additive and non-additive genetic variances of body weight in crossbreed populations of the Japanese quail. *Poult Sci* (2018).
48. Heidaritabar, M. *et al.* Impact of fitting dominance and additive effects on accuracy of genomic prediction of breeding values in layers. *J Anim Breed Genet* **133**, 334–46 (2016).
49. Varona, L., Legarra, A., Toro, M. A. & Vitezica, Z. G. Non-additive Effects in Genomic Selection. *Front Genet* **9**, 78 (2018).
50. Wolak, M. & Keller, L. *Dominance genetic variance and inbreeding in natural populations*, p. 104–127 (Oxford University Press, Oxford, 2014).
51. The UK10K Consortium. The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82–90 (2015).
52. Han, L. & Abney, M. Using identity by descent estimation with dense genotype data to detect positive selection. *Eur J Hum Genet* **21**, 205–11 (2013).
53. Browning, S. R. & Browning, B. L. Identity-by-descent-based heritability analysis in the Northern Finland Birth Cohort. *Hum Genet* **132**, 129–38 (2013).
54. Dandine-Roulland, C. *et al.* Accuracy of heritability estimations in presence of hidden population stratification. *Sci Rep* **6**, 26471 (2016).
55. Browning, S. R. & Browning, B. L. Population structure can inflate SNP-based heritability estimates. *Am J Hum Genet* **89**, 191–3; author reply 193–5 (2011).
56. Huang, W. & Mackay, T. F. The Genetic Architecture of Quantitative Traits Cannot Be Inferred from Variance Component Analysis. *PLoS Genet* **12**, e1006421 (2016).
57. Zuk, O., Hechter, E., Sunyaev, S. R. & Lander, E. S. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc Natl Acad Sci USA* **109**, 1193–8 (2012).
58. Haseman, J. K. & Elston, R. C. The investigation of linkage between a quantitative trait and a marker locus. *Behav Genet* **2**, 3–19 (1972).
59. Evans, L. M. *et al.* Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits. *Nat Genet* **50**, 737–745 (2018).
60. Speed, D., Cai, N., Johnson, M. R., Nejentsev, S. & Balding, D. J. Reevaluation of SNP heritability in complex human traits. *Nat Genet* **49**, 986–992 (2017).
61. Gazal, S. *et al.* Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. *Nat Genet* **49**, 1421–1427 (2017).
62. Kong, A. *et al.* Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat Genet* **40**, 1068–75 (2008).
63. Colonna, V. *et al.* Campora: A Young Genetic Isolate in South Italy. *Human heredity* **64**, 123–135 (2007).
64. Colonna, V. *et al.* Comparing population structure as inferred from genealogical versus genetic information. *European Journal of Human Genetics* **17**, 1635–1641 (2009).
65. Su, Z., Marchini, J. & Donnelly, P. HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics* **27**, 2304–2305 (2011).
66. Wijsman, E. M., Rothstein, J. H. & Thompson, E. A. Multipoint Linkage Analysis with Many Multiallelic or Dense Diallelic Markers: Markov Chain–Monte Carlo Provides Practical Approaches for Genome Scans on General Pedigrees. *Am J Hum Genet* **79**, 846–858 (2006).

67. Herzig, A. F. *et al.* Strategies for phasing and imputation in a population isolate. *Genet Epidemiol* **42**, 201–213 (2018).
68. Raffa, J. D. & Thompson, E. A. Power and Effective Study Size in Heritability Studies. *Stat Biosci* **8**, 264–283 (2016).
69. Perdry, H., Dandine-Roulland, C., Banddyopadhyay, D. & Kettner, L. Gaston: Genetic Data Handling (QC, GRM, LD, PCA) & Linear Mixed Models. CRAN, <https://CRAN.R-project.org/package=gaston> (2018).
70. Gilmour, A. R., Thompson, R. & Cullis, B. R. Average Information REML: An Efficient Algorithm for Variance Parameter Estimation in Linear Mixed Models. *Biometrics* **51**, 1440–1450 (1995).
71. Karigl, G. A recursive algorithm for the calculation of identity coefficients. *Ann Hum Genet* **45**, 299–305 (1981).
72. Delaneau, O., Zagury, J.-F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat Meth* **10**, 5–6 (2013).
73. O'Connell, J. *et al.* A General Approach for Haplotype Phasing across the Full Spectrum of Relatedness. *PLoS Genetics* **10**, e1004234 (2014).
74. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* **48**, 1279–1283 (2016).
75. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
76. Yang, J. *et al.* Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat Genet* **47**, 1114–20 (2015).
77. Mancuso, N. *et al.* The contribution of rare variation to prostate cancer heritability. *Nat Genet* **48**, 30–5 (2016).
78. Young, A. I. *et al.* Relatedness disequilibrium regression estimates heritability without environmental bias. *Nat Genet* **50**, 1304–1310 (2018).

## Acknowledgements

We address special thanks to the people of Campora, Cardile, and Gioi for their participation in the study. We kindly thank the European Genome-phenome Archive at the European Bioinformatics Institute for making available to us the UK10K imputation panel (EGAD00001000776) and HRC imputation panel (EGAD00001002729). A.F.H. was funded by an international Ph.D. fellowship from Sorbonne Paris Cité (convention HERZI15RDXMTSPC1LIETUE) and by the Fondation Recherche Médicale (convention FRM FDT201805005384).

## Author Contributions

A.F.H. carried out the simulation study and analyses of Cilento and wrote the initial manuscript. A.F.H., H.P. and A.-L.L. formulated the simulation study design and main analyses strategies. T.N., D.R. and M.C. prepared the genotypic and phenotypic data of Cilento and M.C. also advised on the subsequent analyses approaches. All authors contributed to the final redaction of the paper.

## Additional Information

**Supplementary information** accompanies this paper at <https://doi.org/10.1038/s41598-018-36050-7>.

**Competing Interests:** The authors declare no competing interests.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2018

**Detecting the dominance component of heritability in isolated and outbred human populations**

Anthony F. Herzig<sup>1,2,†</sup>, Teresa Nutile<sup>3</sup>, Daniela Ruggiero<sup>3,4</sup>, Marina Ciullo<sup>3,4,†</sup>, Hervé Perdry<sup>5,‡</sup>, Anne-Louise Leutenegger<sup>1,2,‡</sup>

1. *Inserm, U946, Genetic variation and Human diseases, F-75010 Paris, France*

2. *Université Paris-Diderot, Sorbonne Paris Cité, U946, F-75010 Paris, France*

3. *Institute of Genetics and Biophysics A. Buzzati-Traverso - CNR, Naples, Italy*

4. *IRCCS Neuromed, Pozzilli, Isernia, Italy*

5. *Université Paris-Saclay, Univ. Paris-Sud, Inserm, CESP, Villejuif, France*

† *Corresponding authors*

‡ *Co-senior authors*

## Supplementary Materials

### Quality Control

Using the known pedigree structure of the Cilento isolates, we scanned for Mendelian errors within the genotype data using Plink <sup>76</sup> and set all genotypes to missing within nuclear families wherever such errors were found. We then removed eight individuals due to very high levels of missingness (over 5%) and restricted to the set of shared SNPs between the two genotyping arrays used in Cilento. Finally we removed variants with minor allele frequencies less than 0.01, Hardy-Weinberg p-values less than  $10^{-5}$  and with missingness greater than 5%. This left 173,911 SNPs from which to calculate GRMs for subsequent heritability analysis. Finally, it became apparent that three pairs of monozygotic twins were present within the sample. We decided to remove one member at random from each pair from subsequent heritability analysis. These twin pairs can be observed in Supplementary Figures 9 and 10.

In Supplementary Table 1 we detail the seven phenotypes studied here, for each phenotype we removed values lying more than three standard deviations away from the observed mean (after transformations (if any) had been applied).

### Imputation

Phasing and imputation were completed separately on the two sets of individuals coming from different genotyping arrays in Cilento. We reconstructed genetic phase in Cilento from SHAPEIT2 <sup>69</sup> with the 'duohmm' option <sup>70</sup>. SHAPEIT2 was employed with 15 burn-in iterations, 15 pruning iterations, 35 main iterations and we used a reduced version of the HRC panel <sup>71</sup> to inform phasing. Following this, we performed haplotype imputation using IMPUTE4 <sup>72</sup> and the same version of the HRC panel. This reduced HRC panel used here included 27,165 individuals and was made available to us from the European phenome-genome archive. IMPUTE4 was applied using default parameters in windows of 5Mb with 250Kb buffer regions. Imputation quality scores ('info') were calculated with the software QCTOOL.

Following imputation, we removed all variants with an 'info' score less than 0.7 in either of the two genotyping arrays and called most likely genotypes whenever an individual genotype had a posterior probability greater than 0.9. Otherwise, imputed genotypes were set to missing. This led to a final count of 3,757,339 confidently imputed variants.

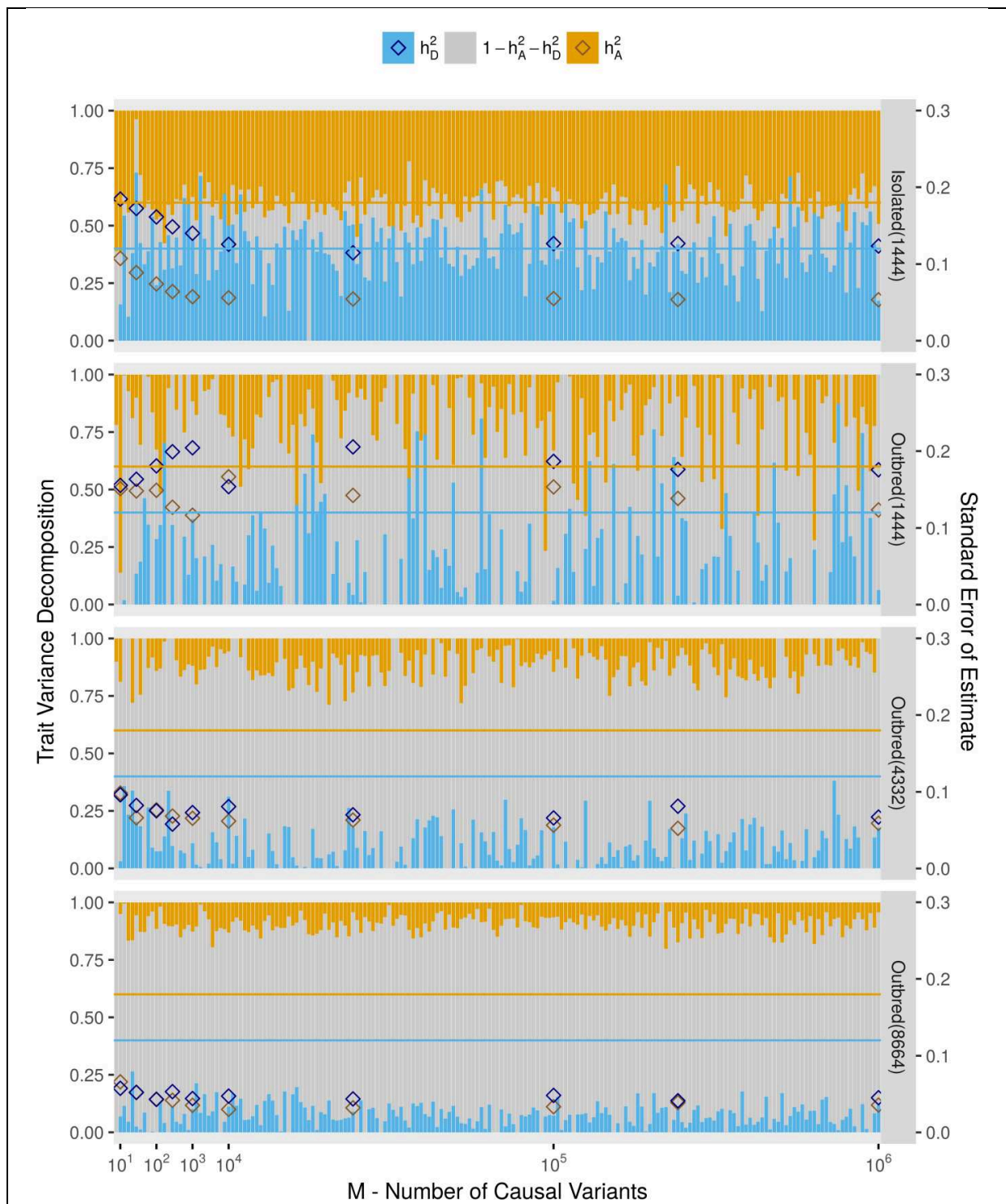
### Simulation

Our HapGen <sup>64</sup> like haplotype mosaic simulated was tuned to produce mosaic pieces of average size of 1-2cM, both in the case of creating simulated populations of outbred individuals, and when creating founding haplotypes for gene-dropping onto the Cilento pedigree. Here we set the effective population size to 3000 and we chose not to simulate mutations, genotyping errors, or missingness in our datasets. The mosaicism was tuned in order to achieve similar kinship and structure in the simulated isolated population as observed in the observed data in Cilento. Indeed, in Supplemental Table 2 we give the variances of the eigenvalues of all GRM matrices calculated both on the Cilento dataset studies here as well as the various simulated datasets and in Supplementary Figure 13 we compare a simple principle component analysis in Cilento and in the simulated population isolate



based on the pedigree of Cilento - Isolated(1444). Through these analyses, we observe that our simulation of an isolate appears to have successfully created similar structure to the Cilento data.

It can be shown that the precision of the estimate of a variance parameter in a linear mixed model is proportional to the product of the sample size and the variance of eigenvalues of the associated variance-covariance matrix<sup>66</sup>. Hence, as we observed that the variances of eigenvalues of the matrix  $K$  in the Outbred(1444) were roughly 36 times smaller than the corresponding values in Isolated(1444), we reasoned that an outbred population of  $6 \times 1444$  (8664) individuals would approximately give us equivalent precision to the population Isolated(1444). Thus we simulated an Outbred population of size 8664, as well as an intermediate population of size  $3 \times 1444$  (4332) in order to observe any trends in the results relating to sample size.

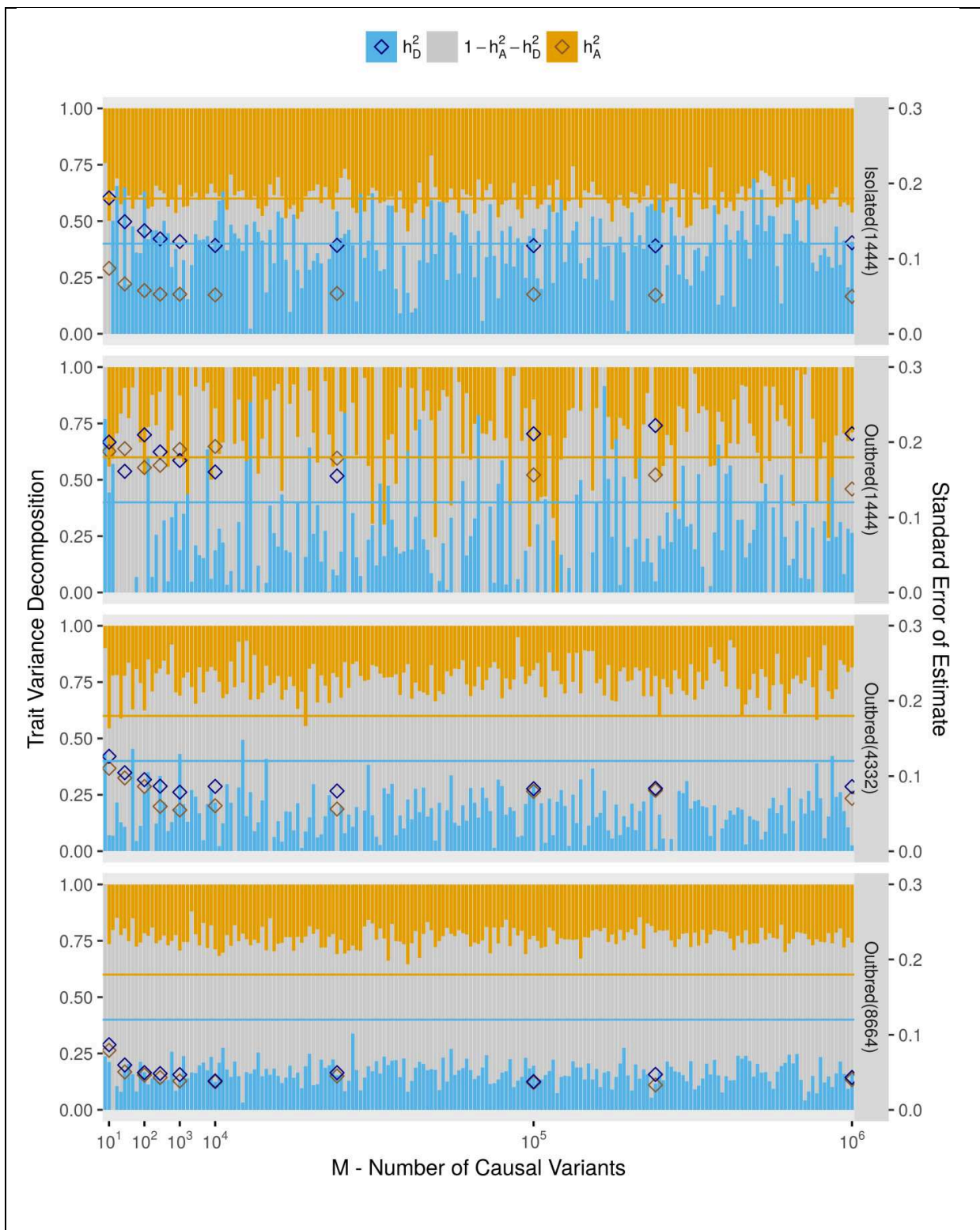


Supplementary Figure 1.

Estimates of  $h_A^2$  and  $h_D^2$  are represented for each simulated phenotype. Each phenotype was simulated using different numbers of causal variants ( $M$ ) for each variance component. Results from four simulated populations are given, either Isolated( $N$ ) or Outbred( $N$ ), where the value of  $N$  denotes the sample size.

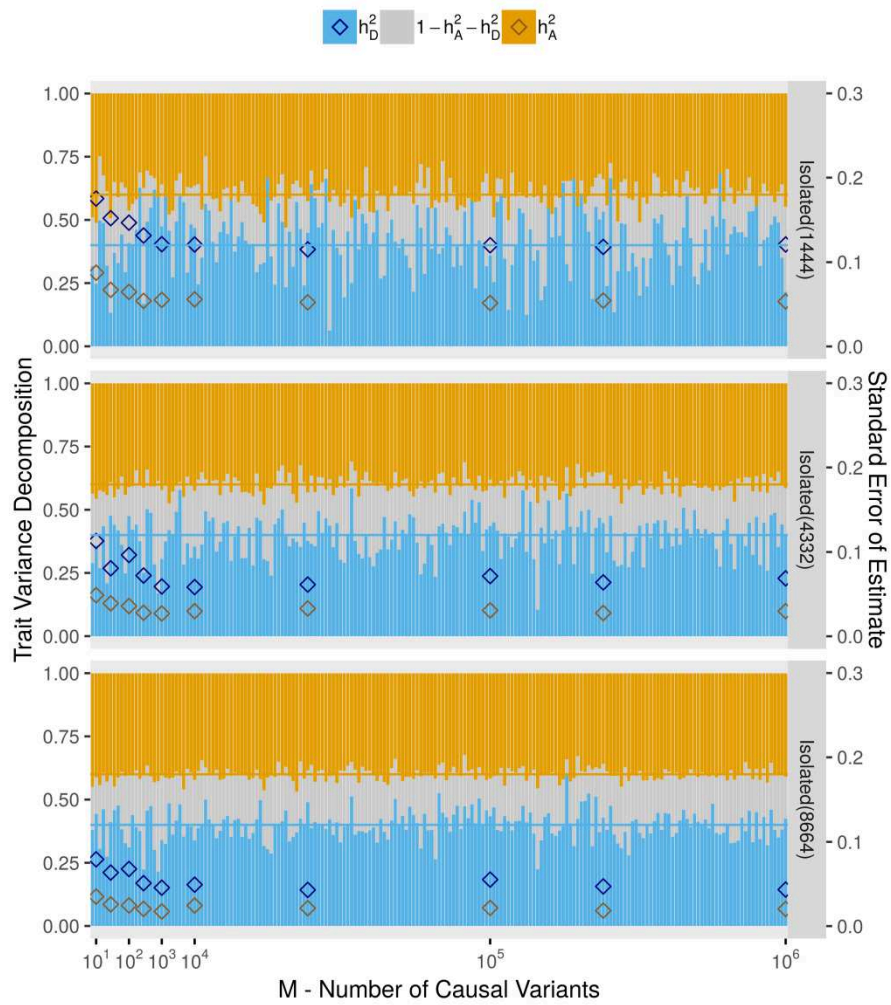
Matrices  $K$  and  $D$  are calculated using roughly 170,000 variants present in all villages in Cilento, causal variants are selected completely at random (Causal Variant Scenario A).

Diamonds represent empirical standard errors measured for certain values of  $M$  which are measured on the right vertical axes. A missing bar for  $h_A^2$  or  $h_D^2$  indicates the maximum likelihood estimate of the parameter was zero.



Supplementary Figure 2.

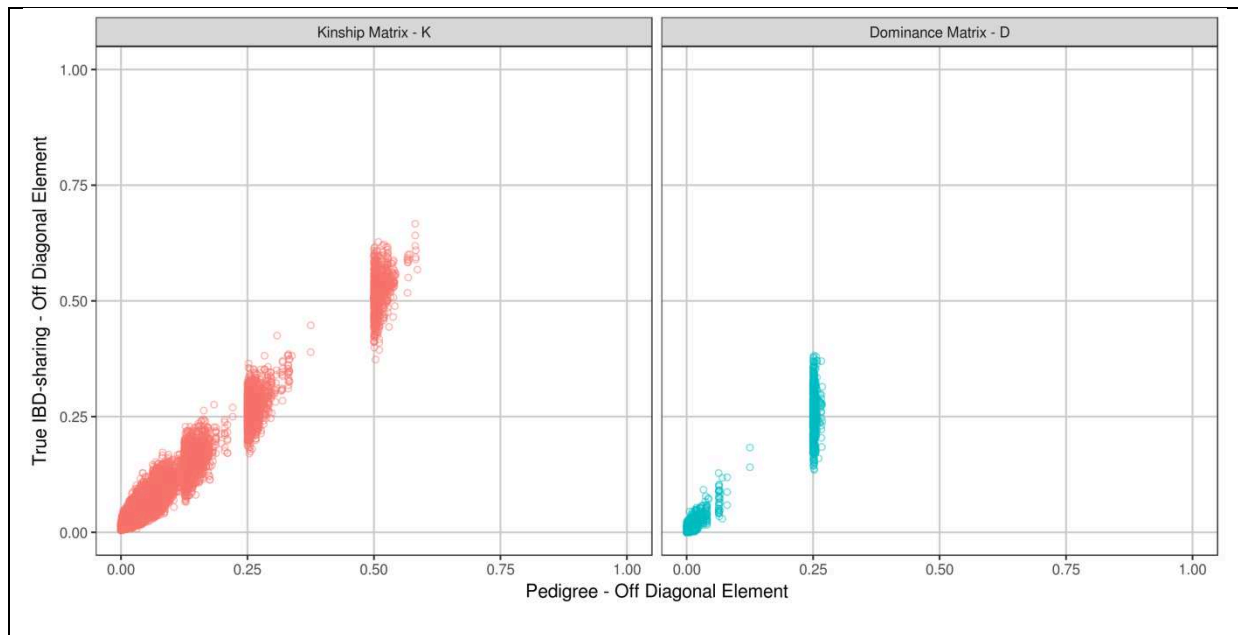
Identical to Supplementary Figure 1 apart from here K and D are calculated using roughly 170,000 variants present in all villages in Cilento, causal variants are selected to have MAF > 0.01 (Causal Variant Scenario B).



Supplementary Figure 3.

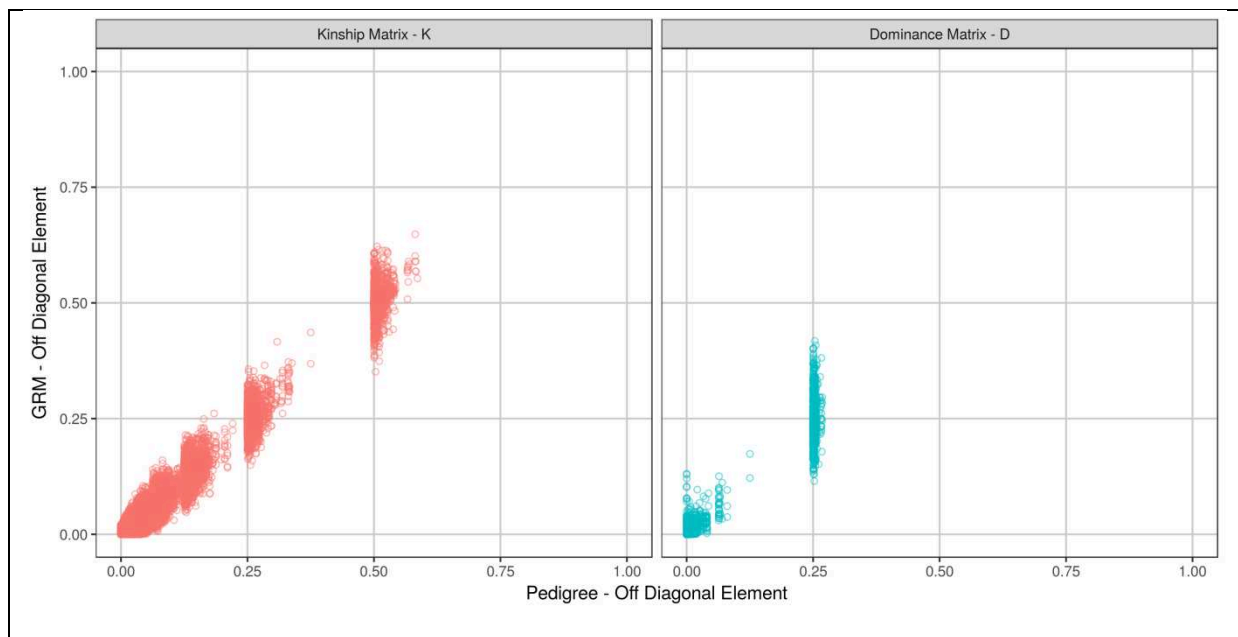
Comparison of heritability estimates from a single simulated isolated population with populations constructed by combining isolated populations.

Identical to Figure 3 in the main text but here causal variants are selected to have MAF > 0.01 (Causal Variant Scenario B).



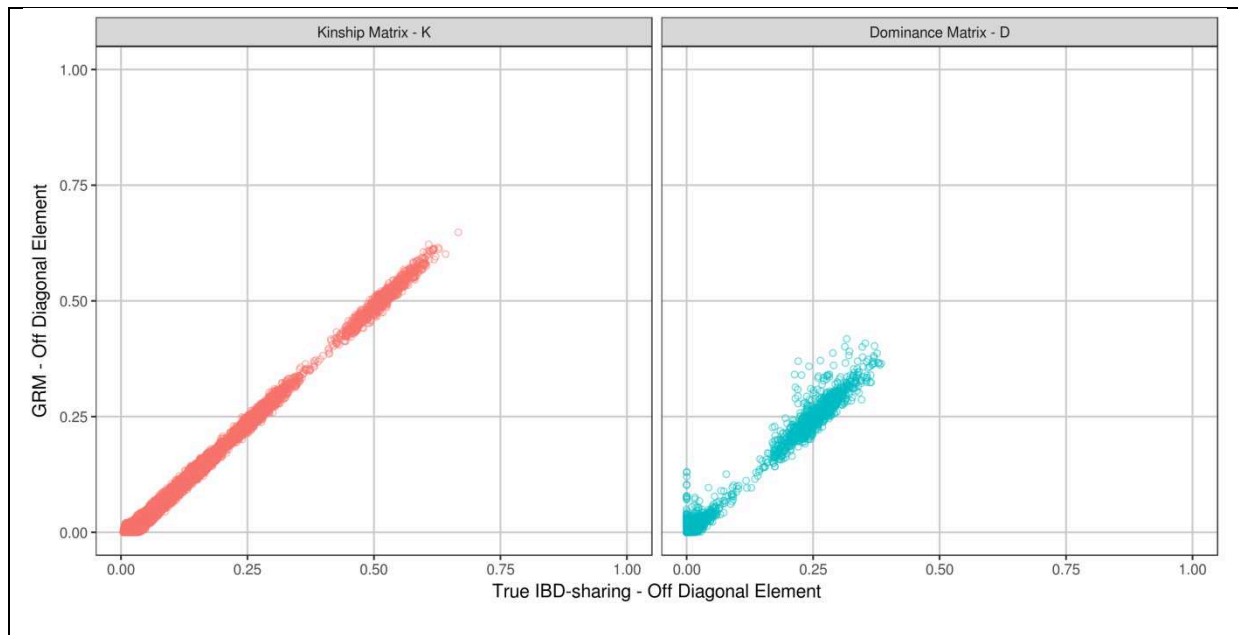
Supplementary Figure 4a.

Comparison of off-diagonal elements of the matrices K and D calculated on the simulated isolated population 'Isolated(1444)'. K and D are calculated either from the true proportions of IBD-sharing that occurred during the simulation of the data or from the pedigree information of Cilento.



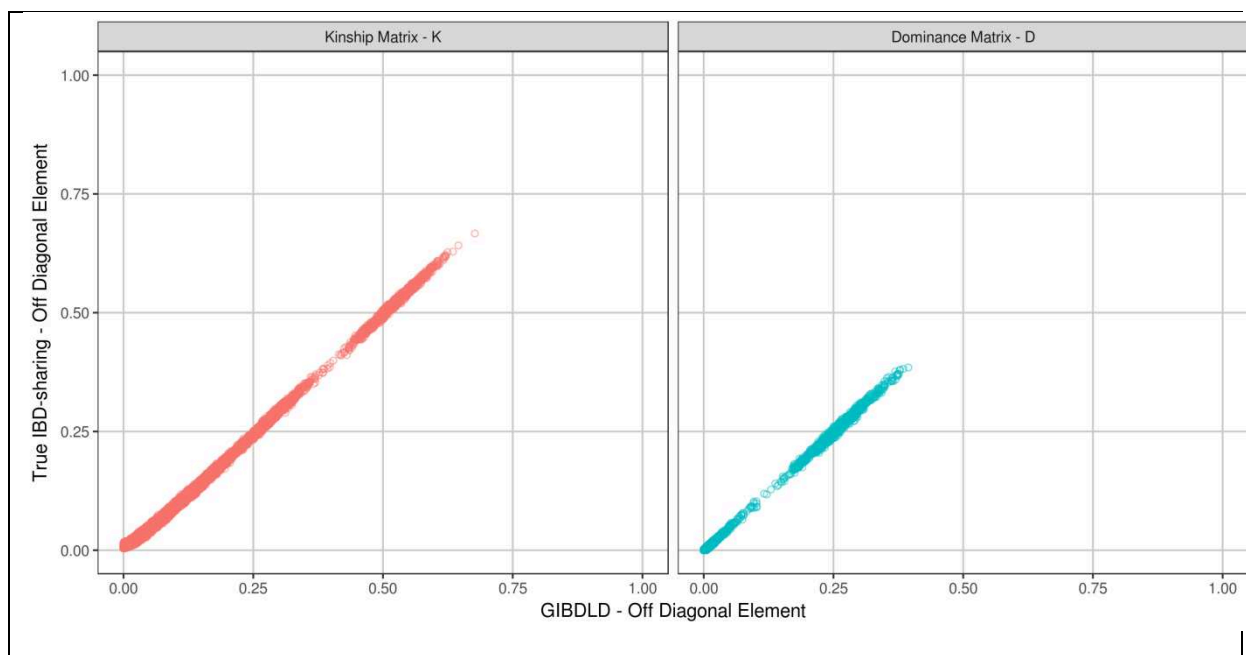
Supplementary Figure 4b.

Comparison of off-diagonal elements of the matrices K and D calculated on the simulated isolated population 'Isolated(1444)'. K and D are calculated either as genetic relationship matrices (GRMs) or from the pedigree information of Cilento.



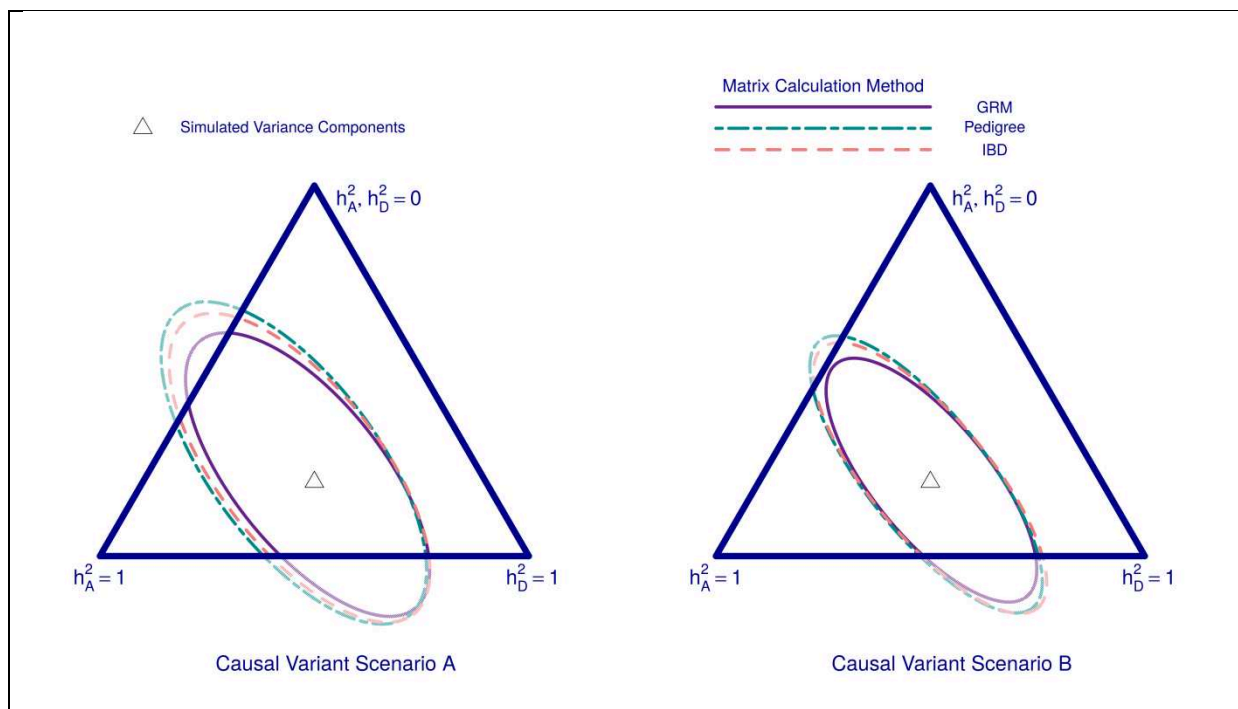
Supplementary Figure 4c.

Comparison of off-diagonal elements of the matrices K and D calculated on the simulated isolated population 'Isolated(1444)'. K and D are estimated either as genetic relationship matrices (GRMs) or from the true proportions of IBD-sharing that occurred during the simulation of the data.



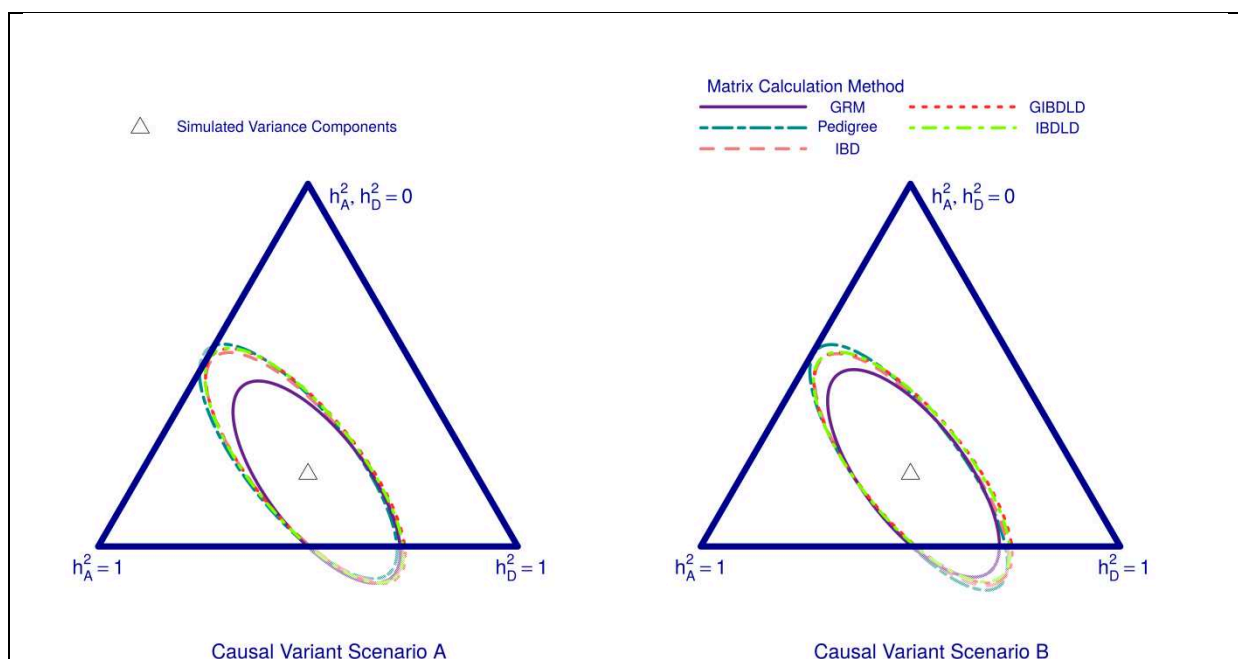
Supplementary Figure 4d.

Comparison of off-diagonal elements of the matrices K and D calculated on the simulated isolated population 'Isolated(1444)'. K and D are estimated using either the software GIBDLD or the true proportions of IBD-sharing that occurred during the simulation of the data. Off-diagonal elements estimated by IBDLD were equally similar.



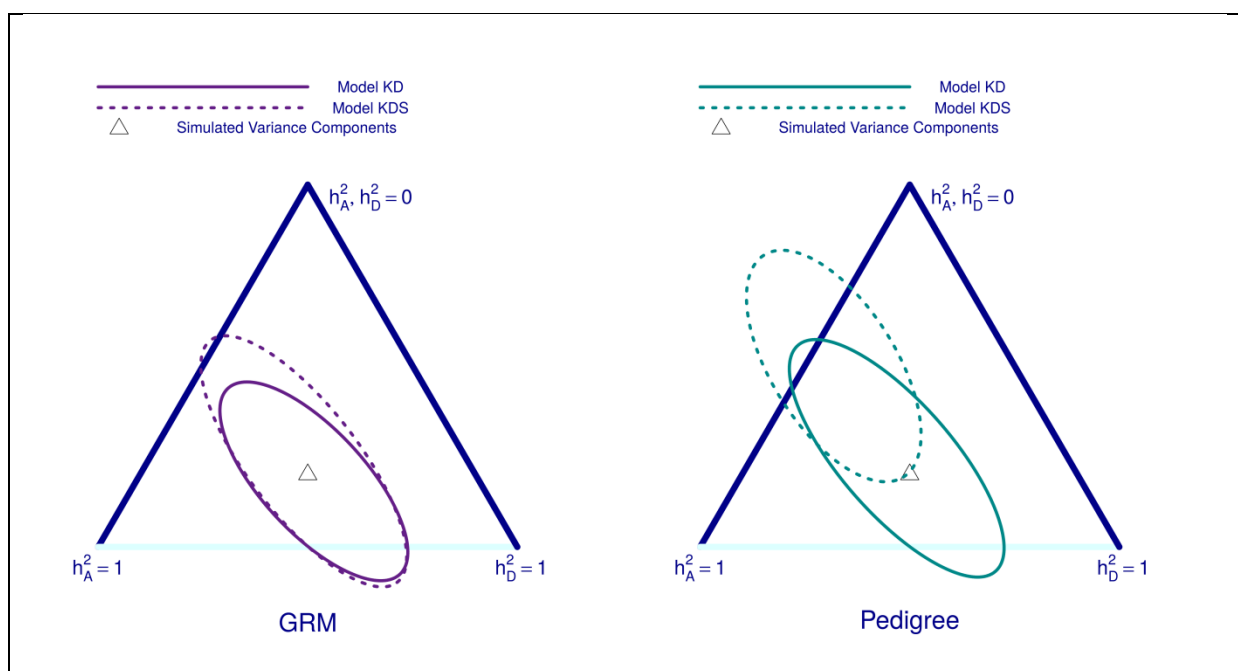
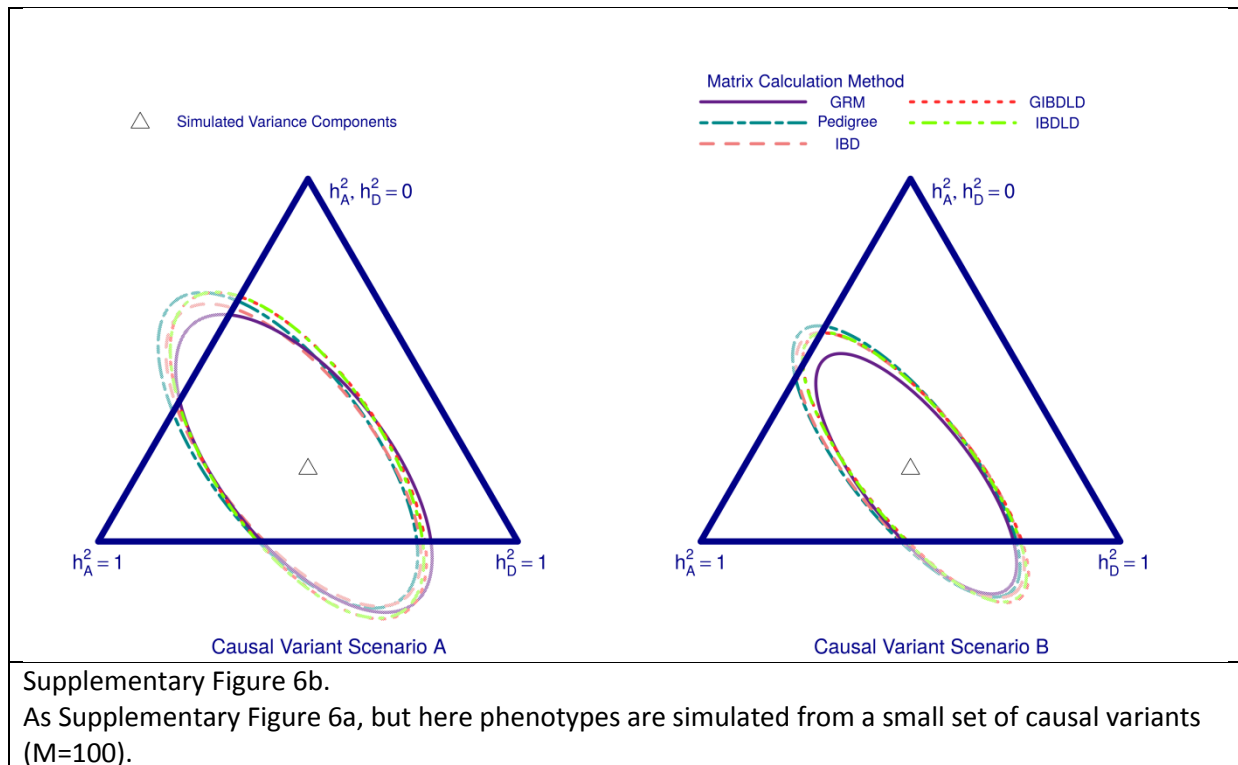
Supplementary Figure 5.

Comparison of strategy for computing matrices K and D for the Isolated(1444) population. Identical to Figure 4 in the main text apart from here, phenotypes are simulated from a small set of causal variants ( $M=100$ ).

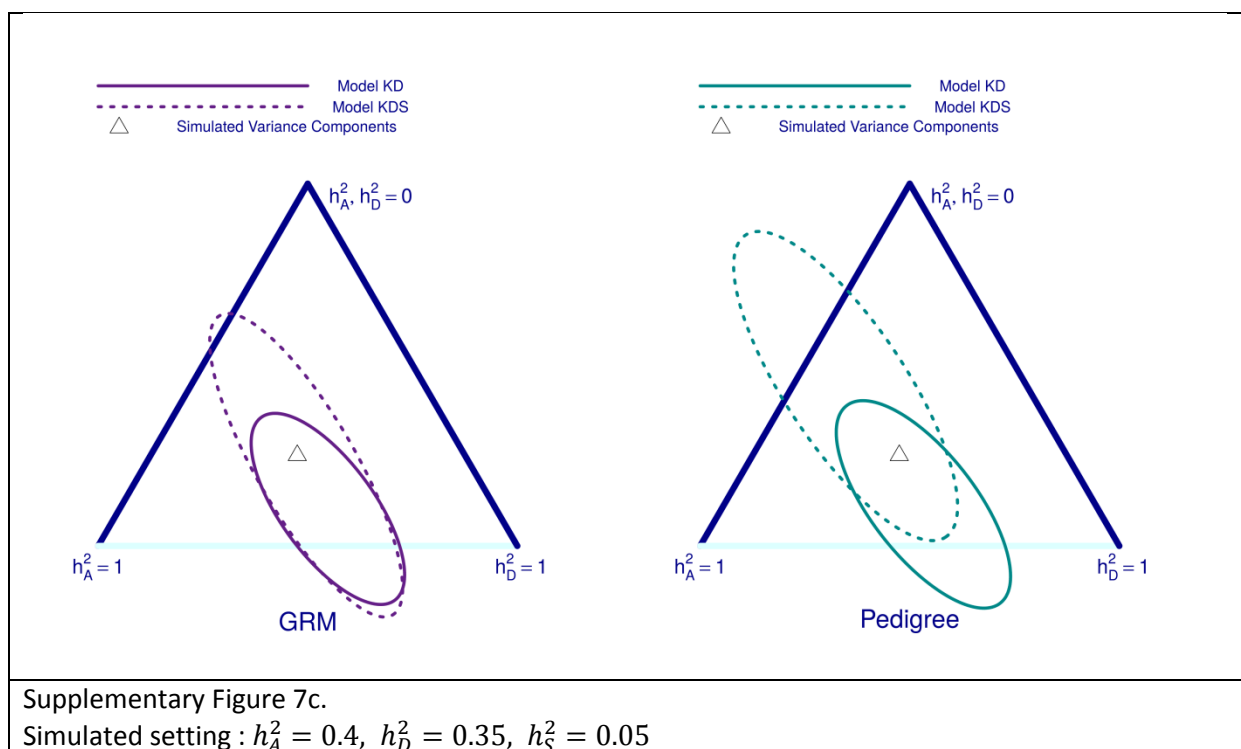
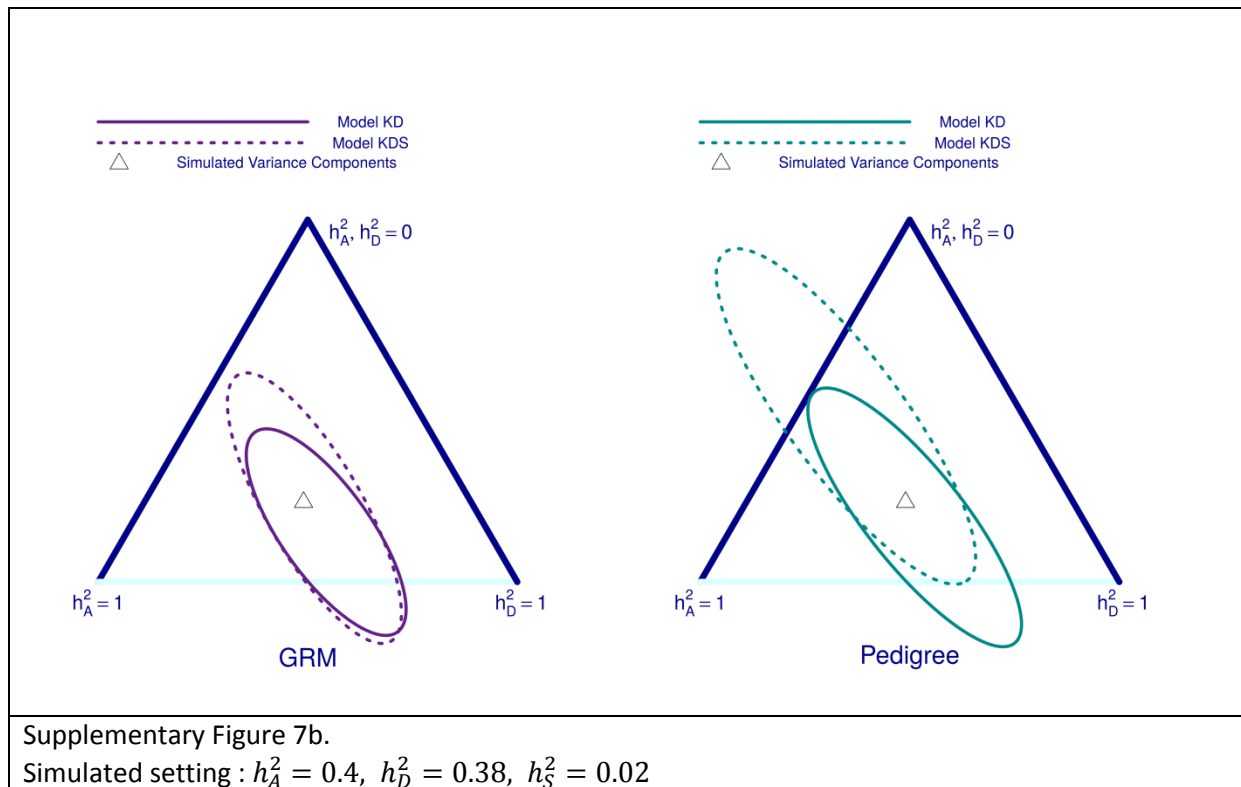


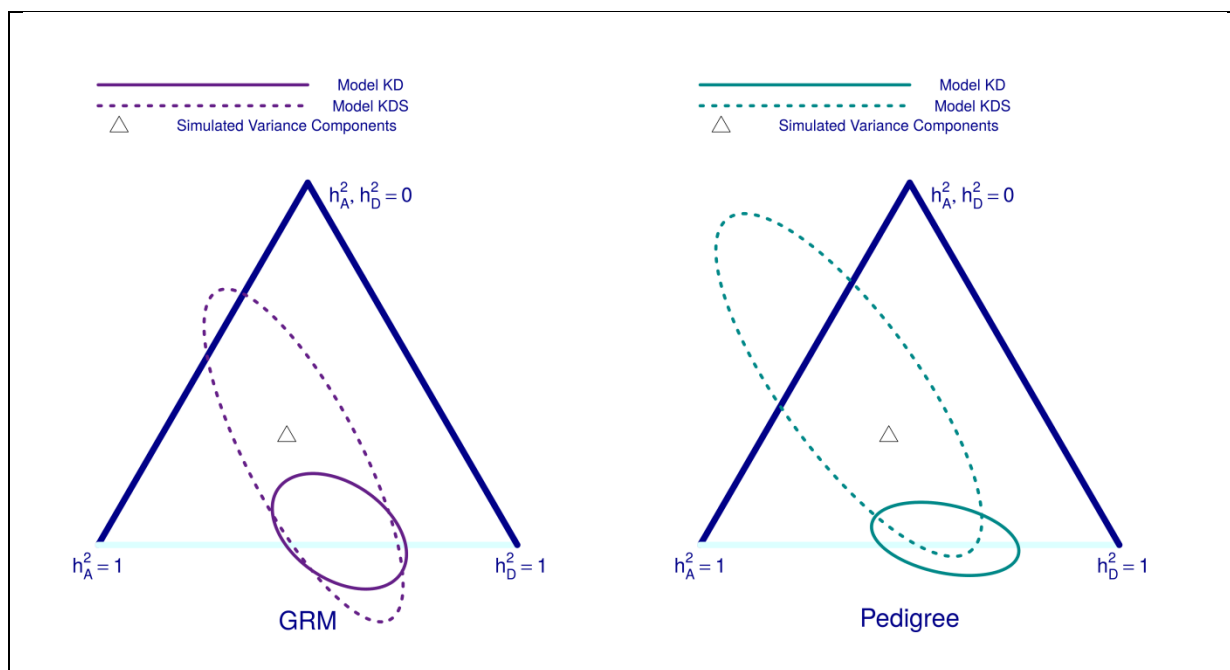
Supplementary Figure 6a.

Comparison of strategy for computing matrices K and D for the Isolated(1444) population. The set of roughly 170,000 SNPs present in the Cilento isolates are used here to calculate K and D for the Isolated(1444) population, either as GRMs or using the IBDLD or GIBDL software. Estimates using either pedigree information or recorded IBD-sharing from the simulating are also given. Here, phenotypes are simulated from a large set of causal variants ( $M=100,000$ ).



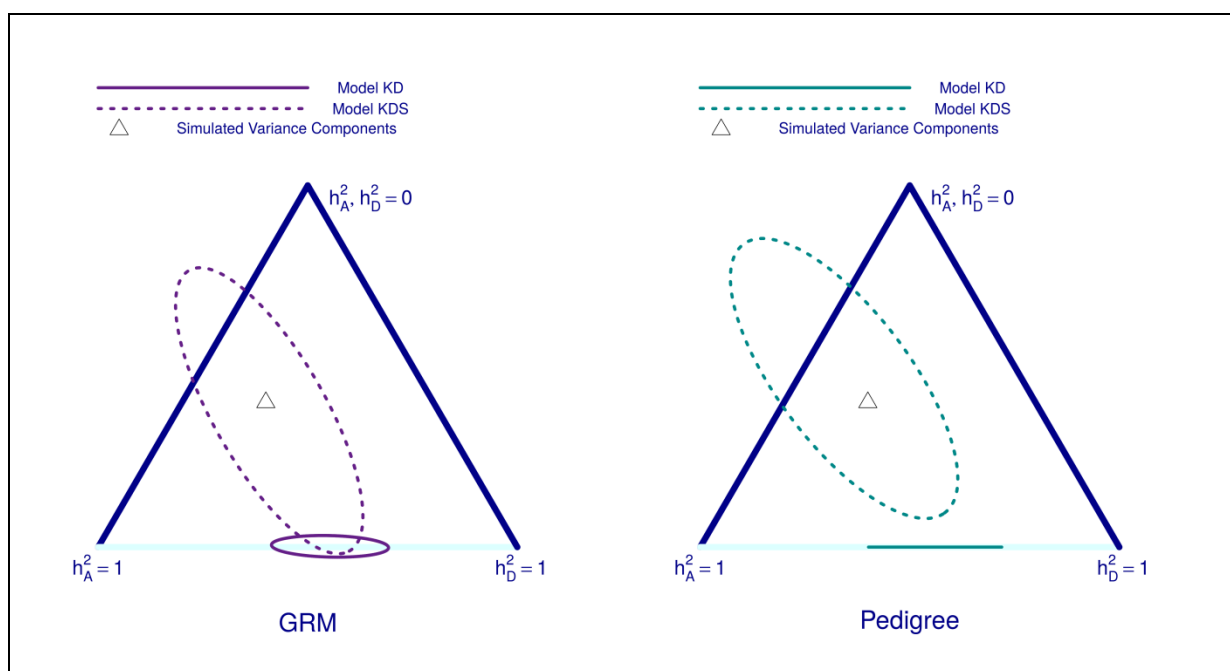






Supplementary Figure 7d.

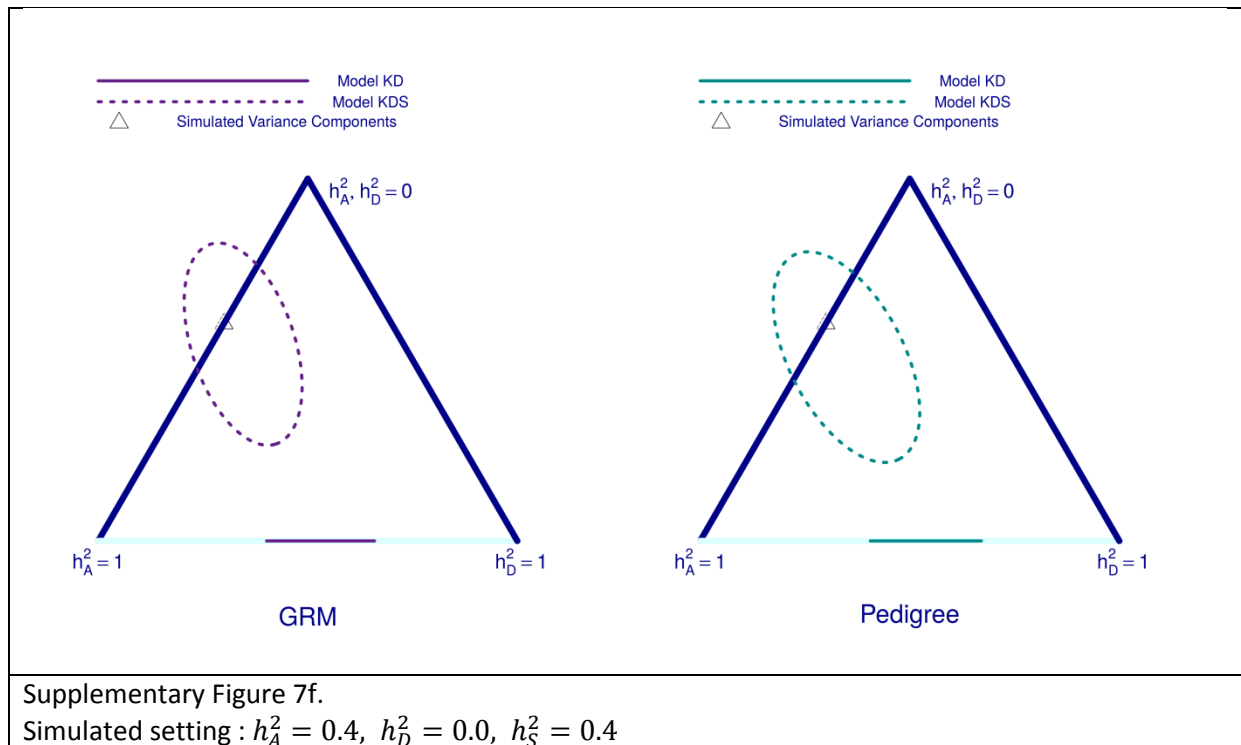
Simulated setting :  $h_A^2 = 0.4$ ,  $h_D^2 = 0.3$ ,  $h_S^2 = 0.1$

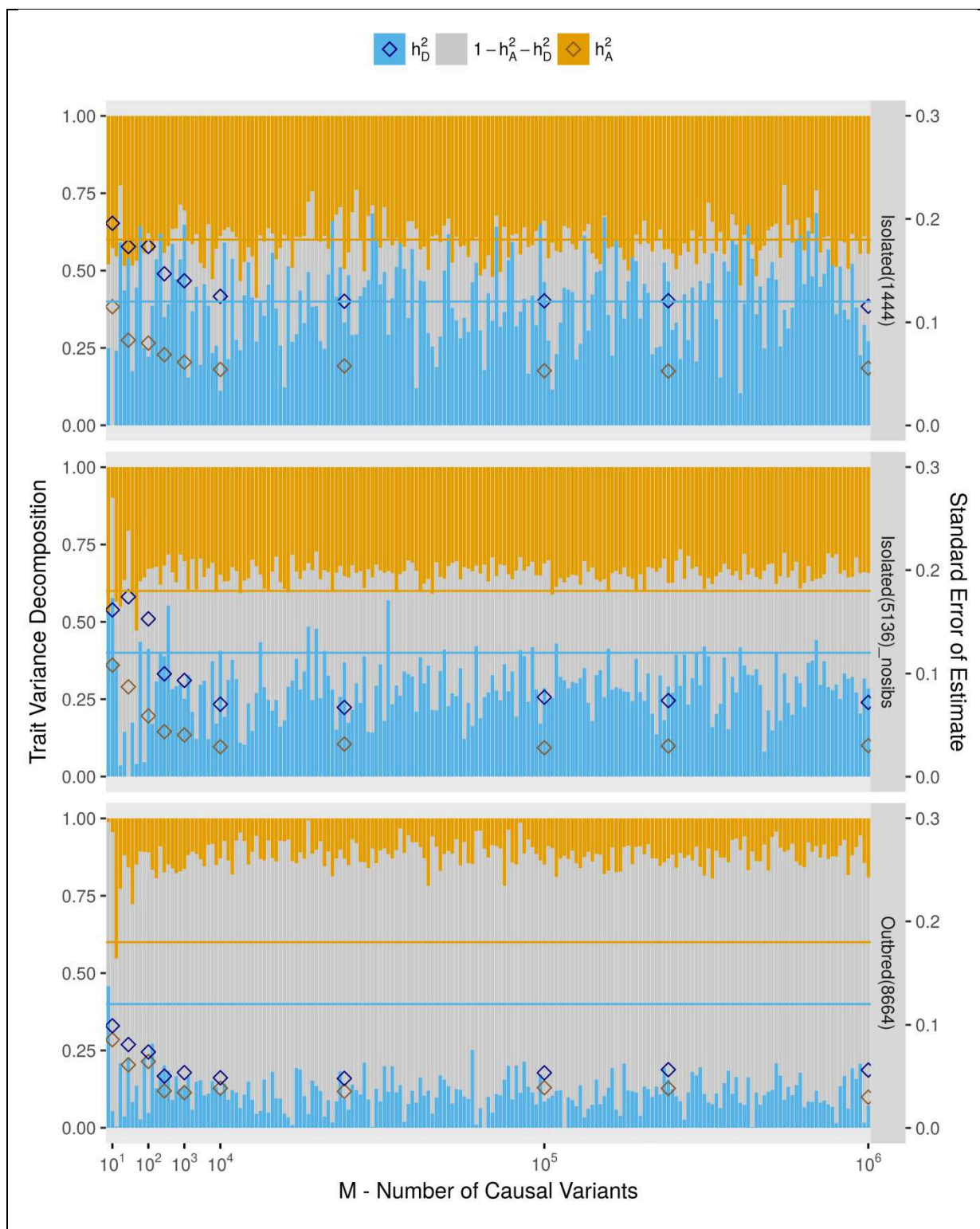


Supplementary Figure 7e.

Simulated setting :  $h_A^2 = 0.4$ ,  $h_D^2 = 0.2$ ,  $h_S^2 = 0.2$

Completely the same as Figure 5 in the main text but is given here also for continuity with other Supplementary Figures.



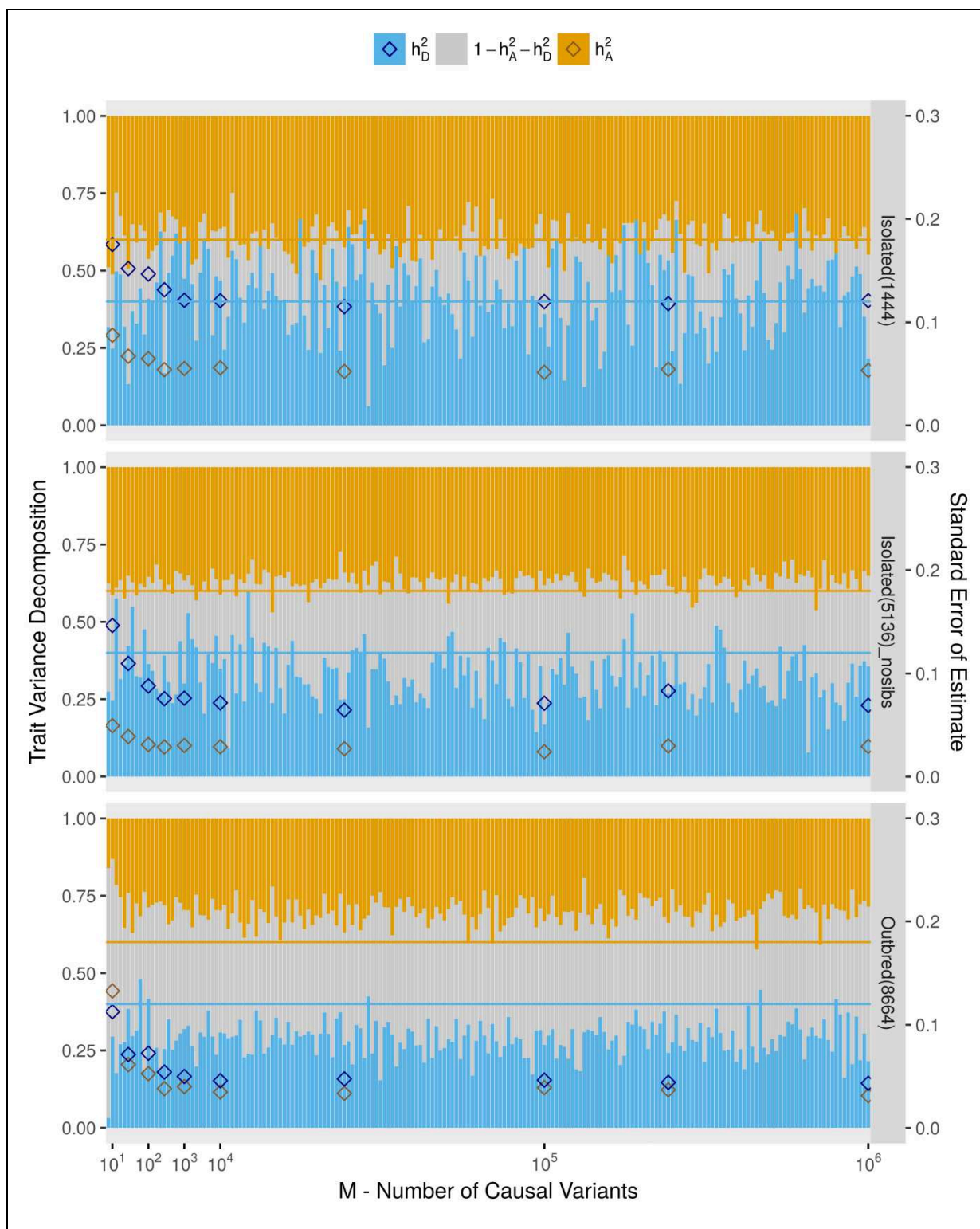


Supplementary Figure 8a.

Comparison of heritability analysis for three simulated populations.

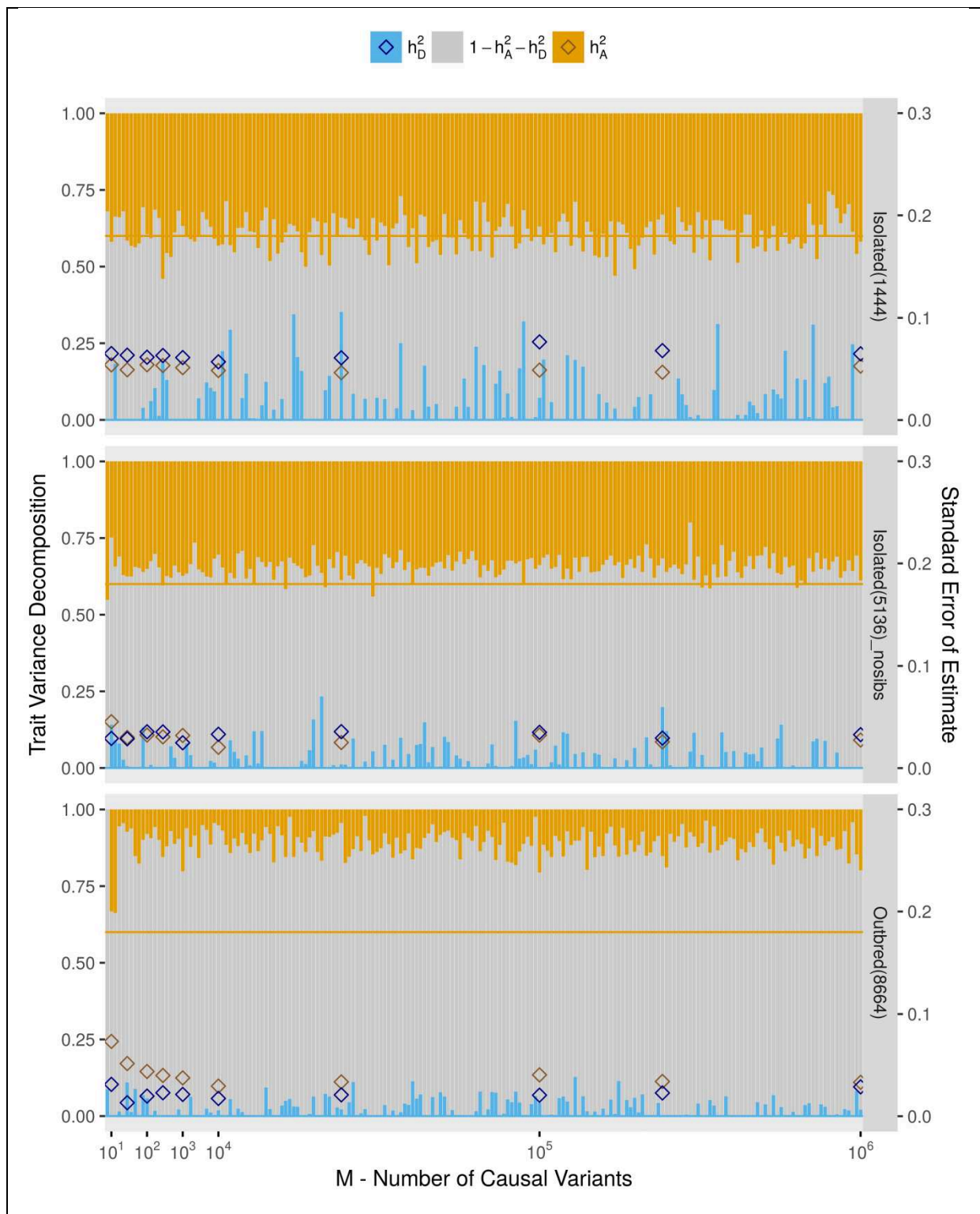
Phenotypes are simulated under the setting :  $h_A^2 = 0.4$ ,  $h_D^2 = 0.4$

Similar to earlier Figures but here we include the larger simulated isolated population, a composite of isolates with no sibling pairs. Causal variants are selected completely at random (Causal Variant Scenario A).



Supplementary Figure 8b.

Identical to Supplementary Figure 8a apart from here, causal variants are selected to have MAF > 0.01 (Causal Variant Scenario B).

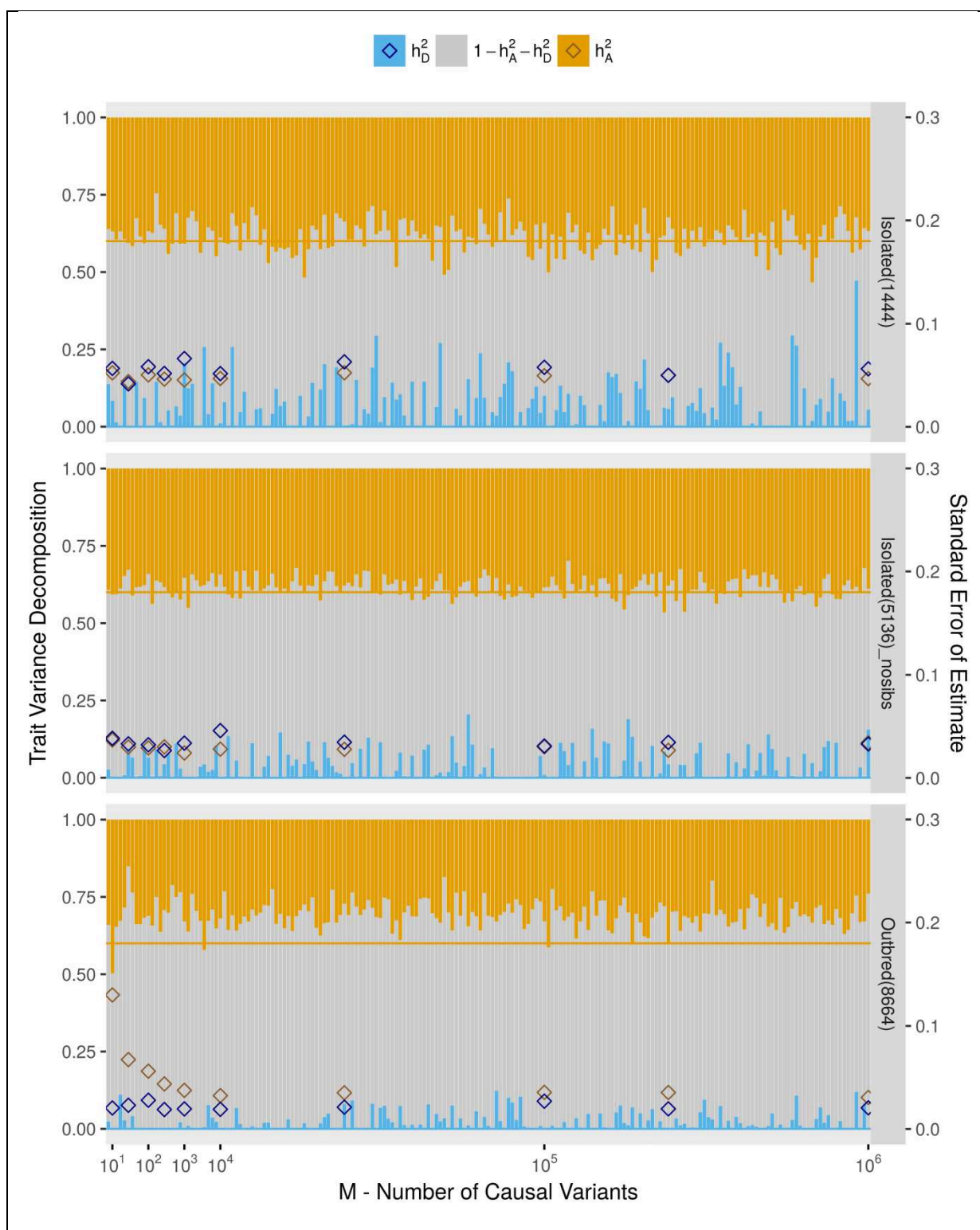


Supplementary Figure 8c.

Comparison of heritability analysis for three simulated populations.

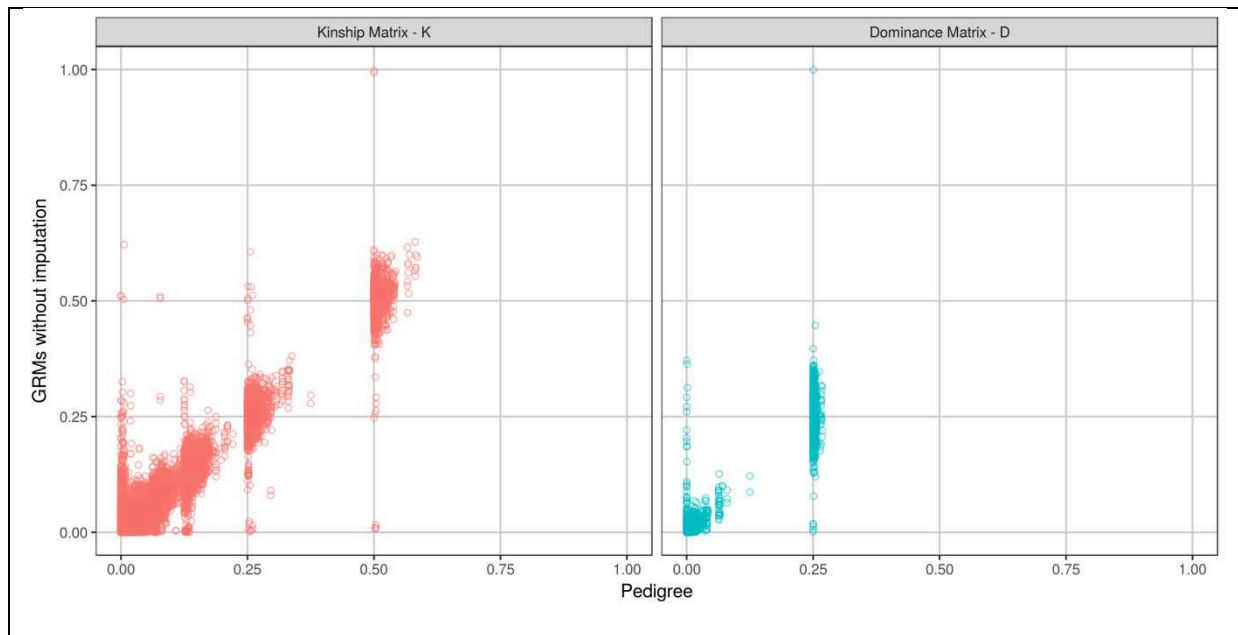
Phenotypes are simulated under the setting :  $h_A^2 = 0.4$ ,  $h_D^2 = 0.0$

Similar to earlier Figures but here we include the larger simulated isolated population, a composite of isolates with no sibling pairs. Causal variants are selected completely at random (Causal Variant Scenario A).



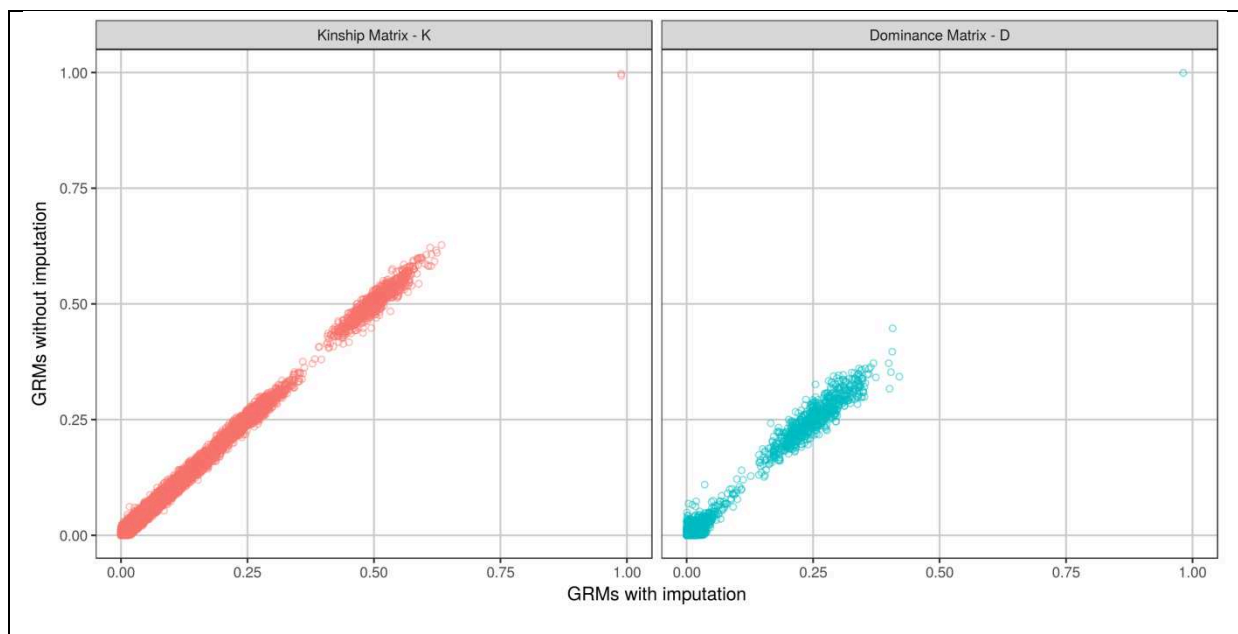
Supplementary Figure 8d.

Identical to Supplementary Figure 8c apart from here, causal variants are selected to have MAF > 0.01 (Causal Variant Scenario B).



Supplementary Figure 9.

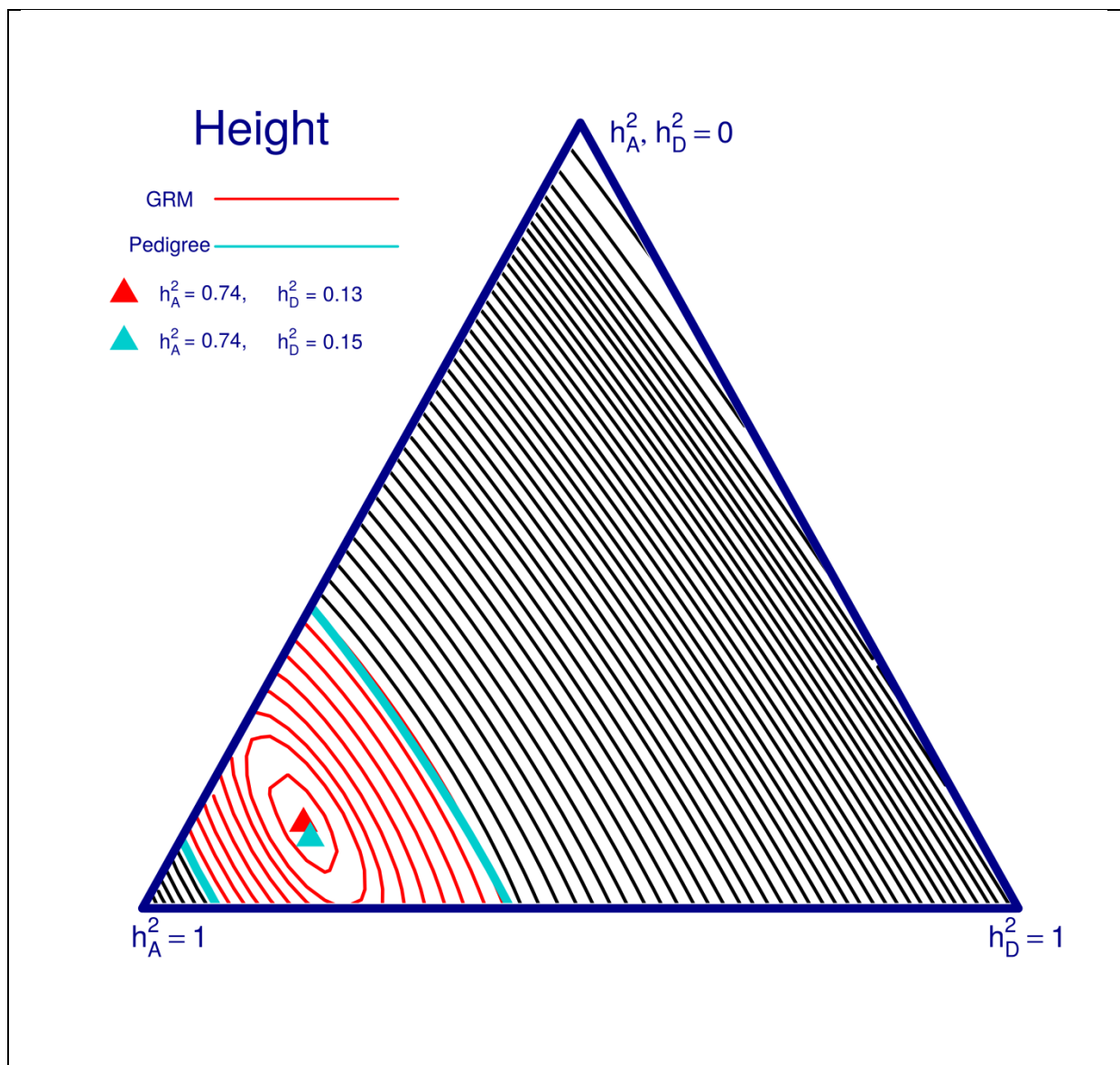
Comparison of off diagonal elements of matrices  $K$  and  $D$  calculated for the Cilento dataset using either pedigree information or genetic relationship matrices (GRMs) from the observed genotypes in Cilento.



Supplementary Figure 10.

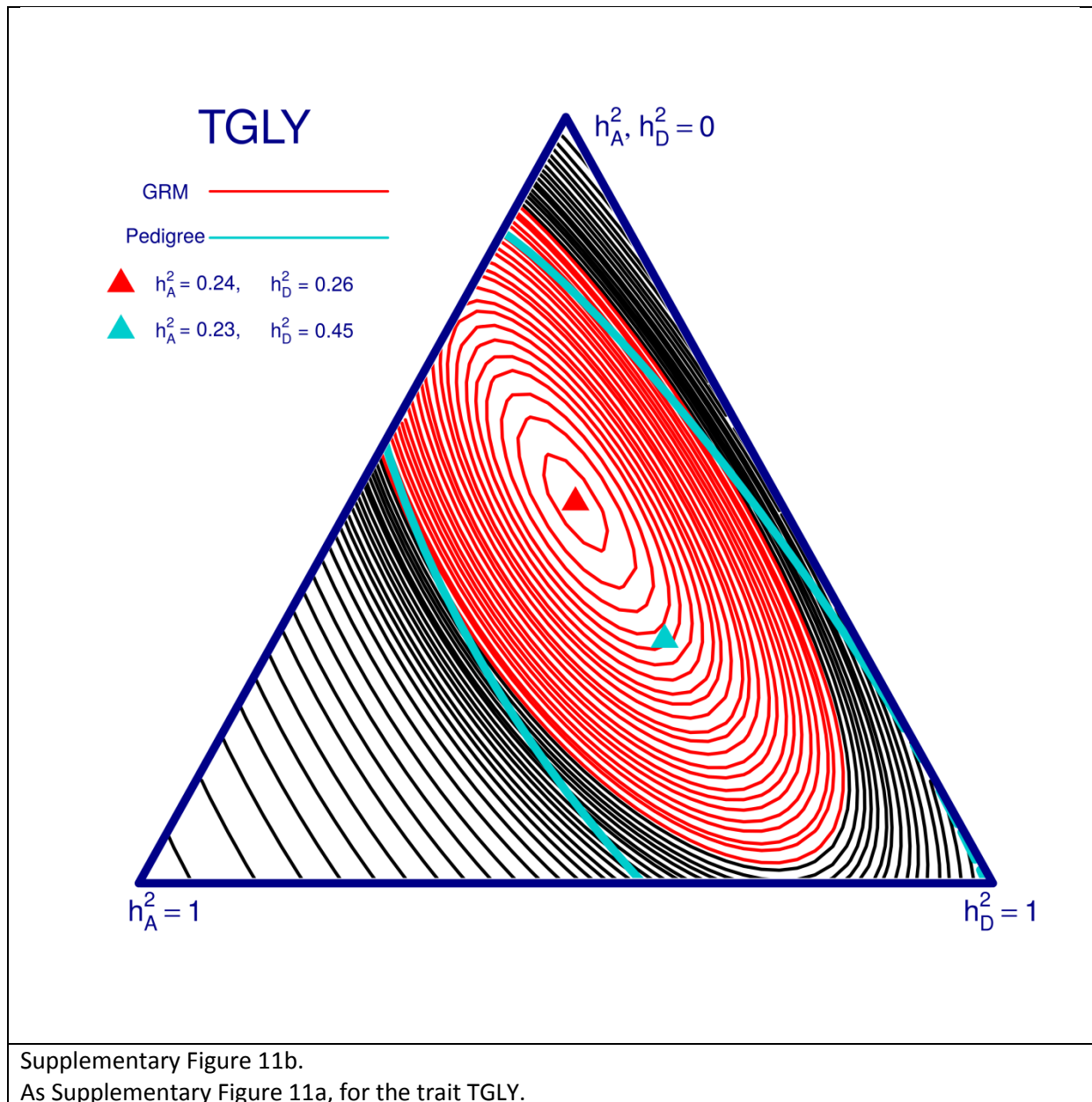
Comparison of off diagonal elements of matrices  $K$  and  $D$  calculated for the Cilento dataset using genetic relationship matrices (GRMs) before and after the inclusion of imputed variants in the Cilento dataset.

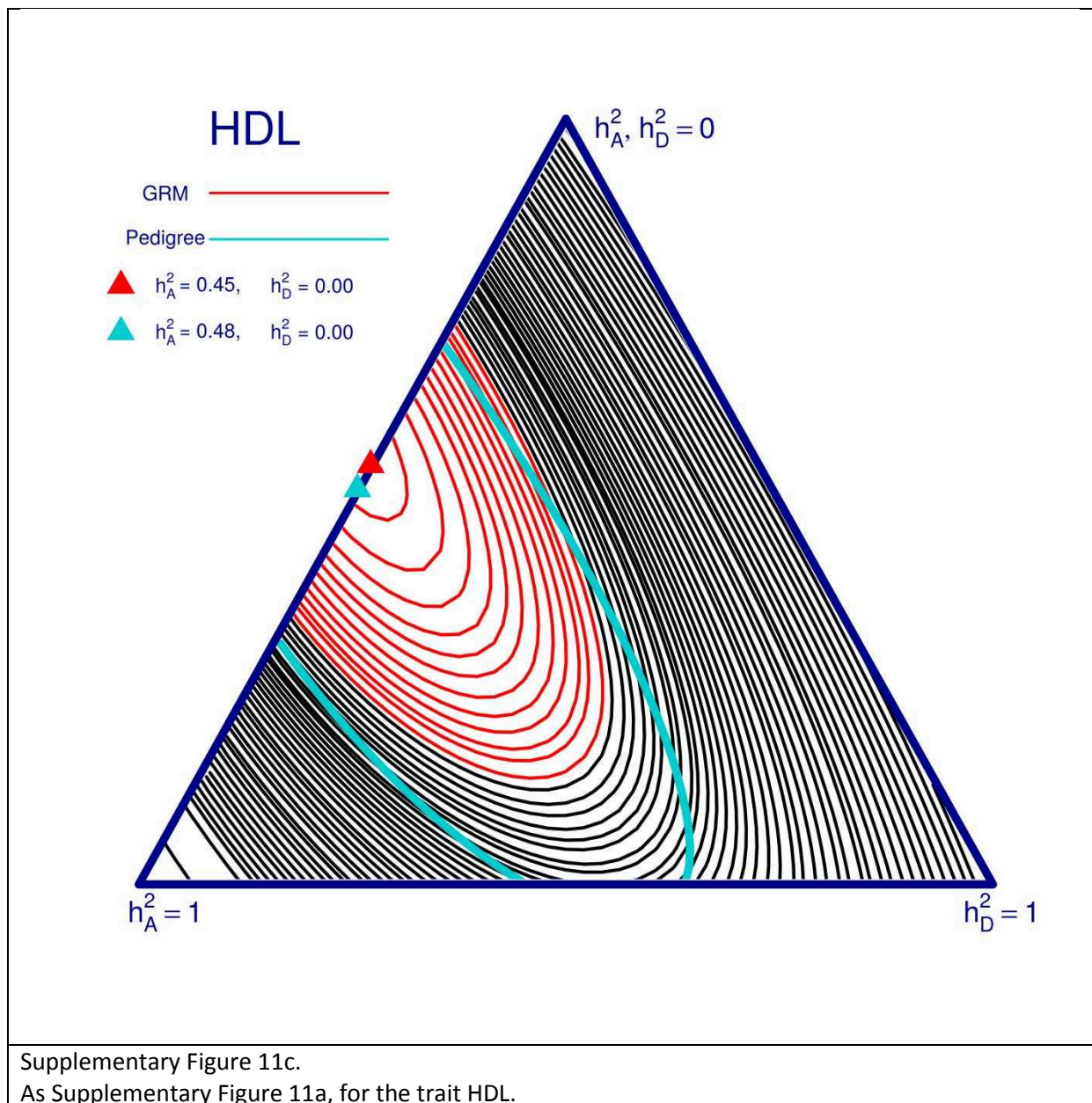


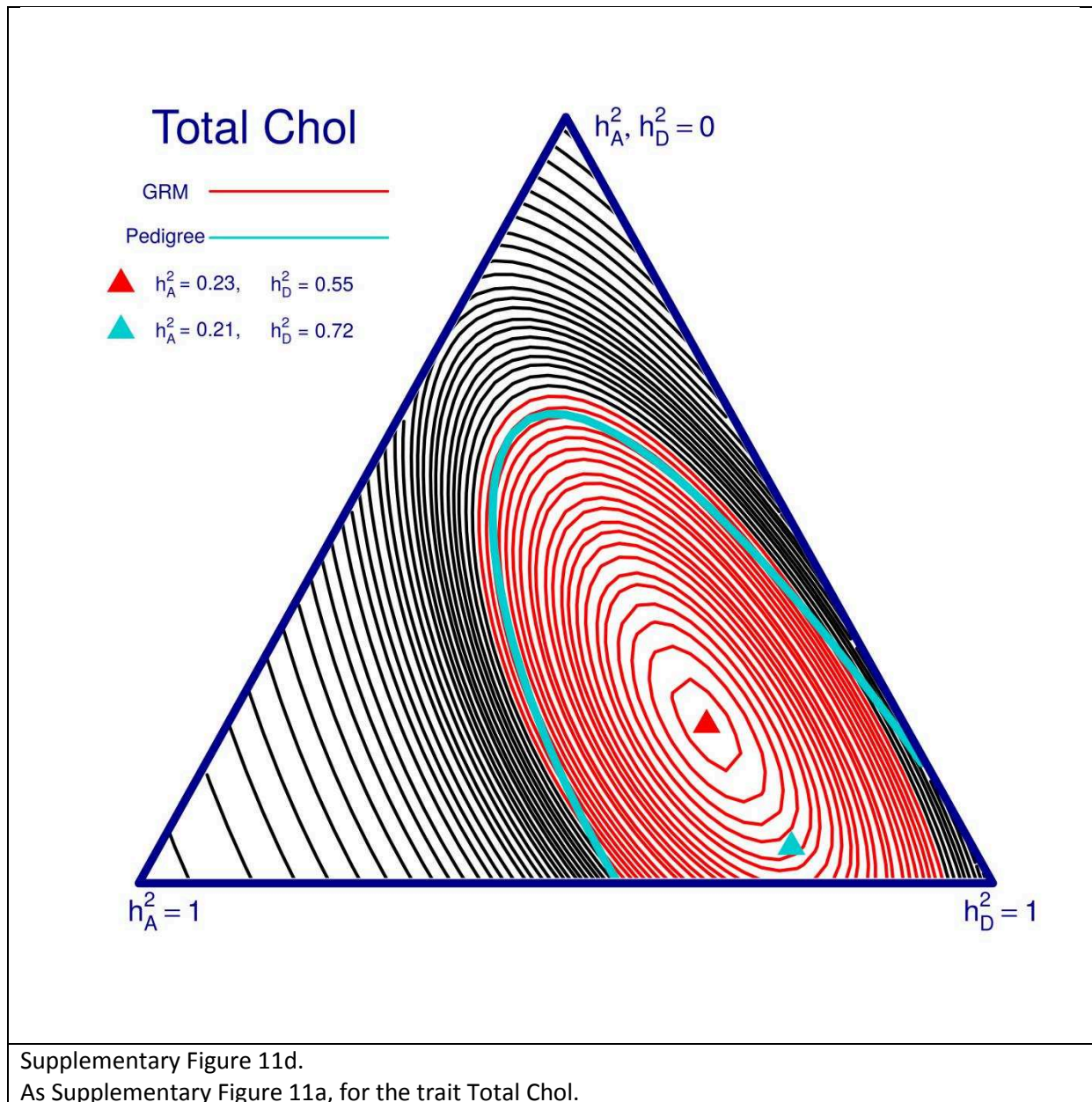


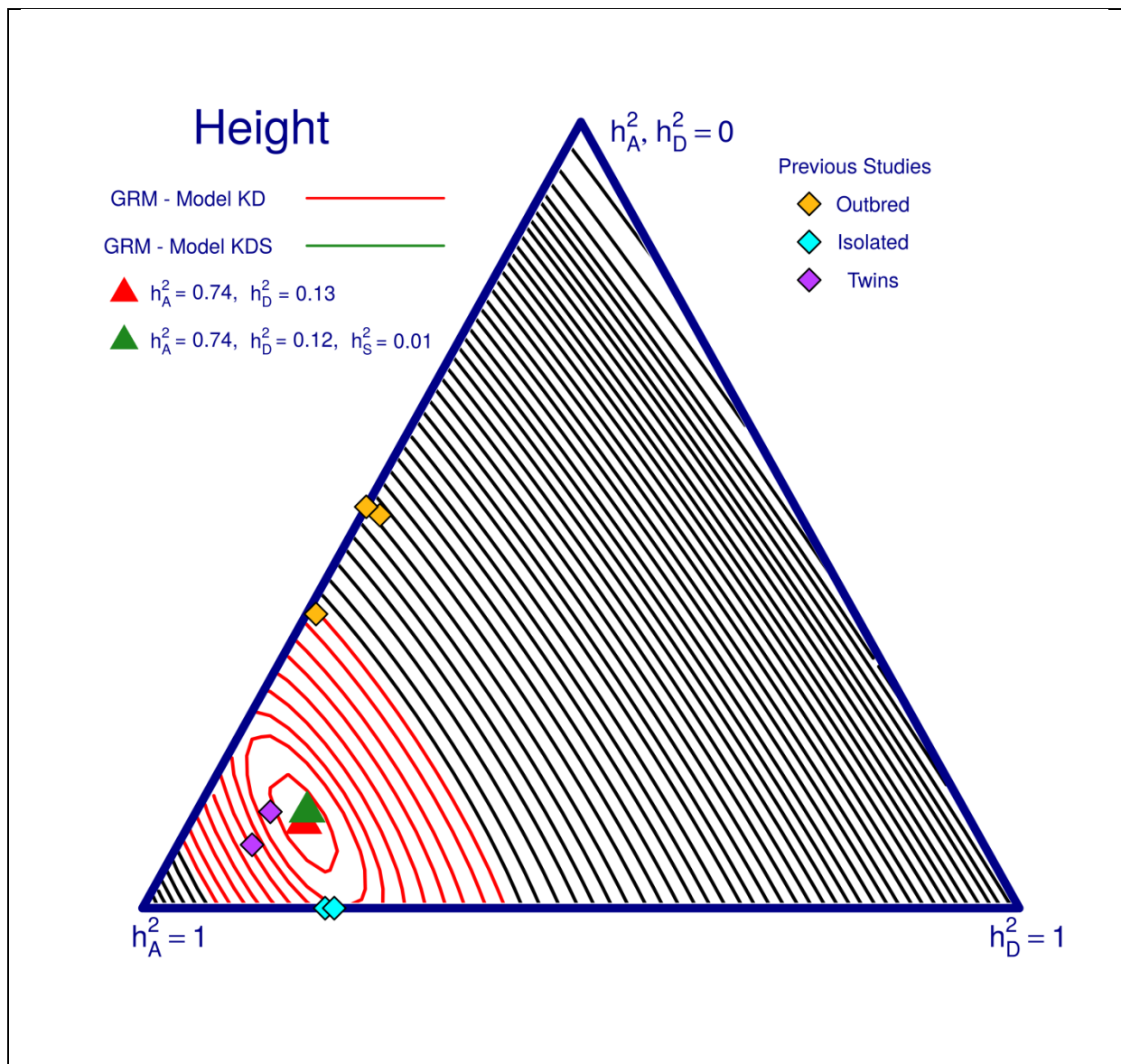
Supplementary Figure 11a.

Heritability analysis for Height in Cilento. Black contours represent the likelihood profile from the model KD (see Figure 5 in main text), with matrices  $K$  and  $D$  calculated as genetic relationship matrices (GRMs). The red zone represents the 95% confidence interval for the red maximum likelihood estimate (MLE) (red triangular peak). The corresponding MLE and 95% confidence boundary for the analysis using pedigree information to estimate  $K$  and  $D$  are added to the plot in blue.





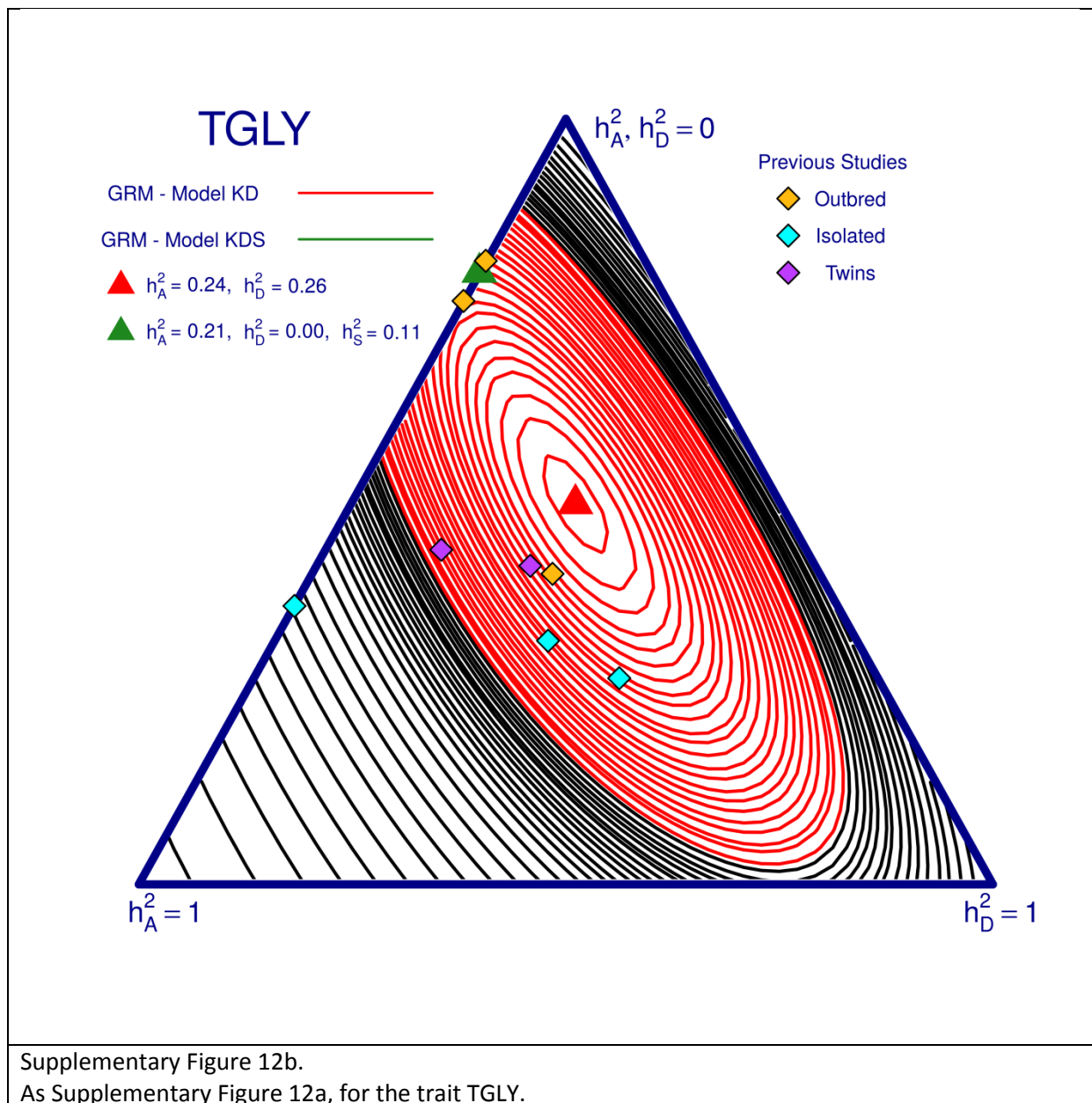


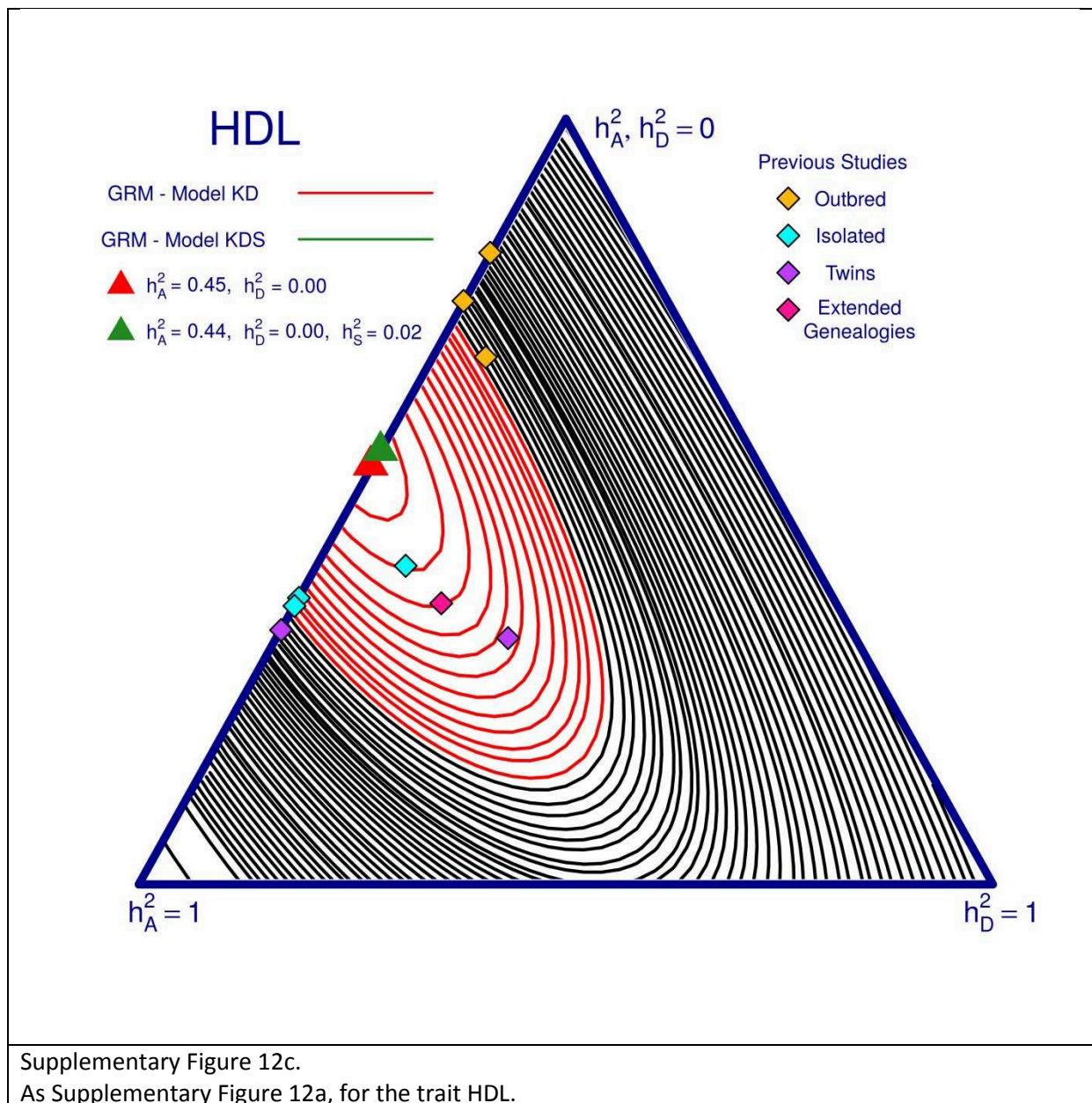


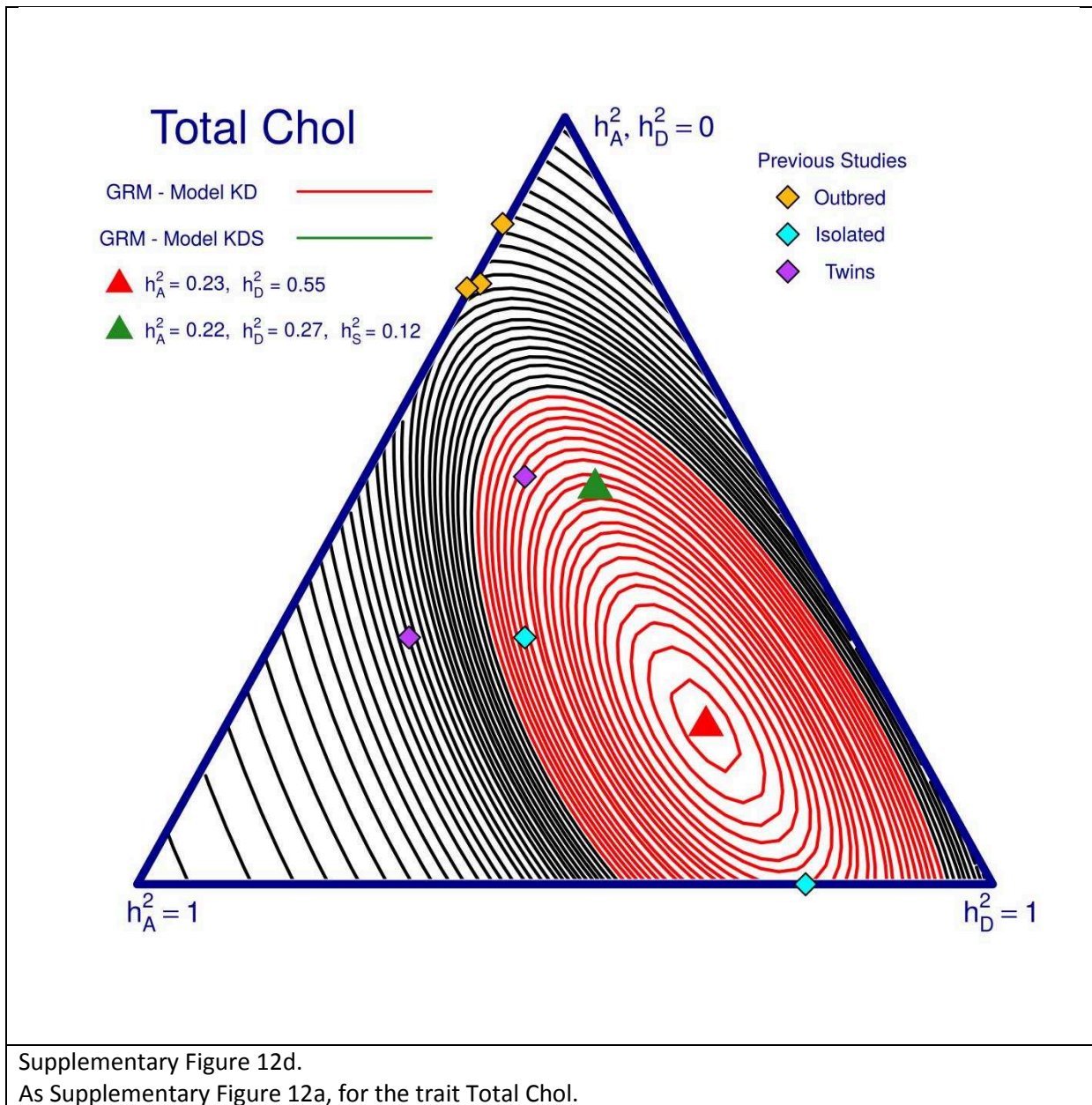
Supplementary Figure 12a.

Comparison of models KD and KDS (see Figure 7 in main text) in Cilento for the trait Height. Black contours represent the likelihood profile for the model KD, with the red zone indicating the 95% confidence interval for the red maximum likelihood estimate (MLE) (red triangular peak).

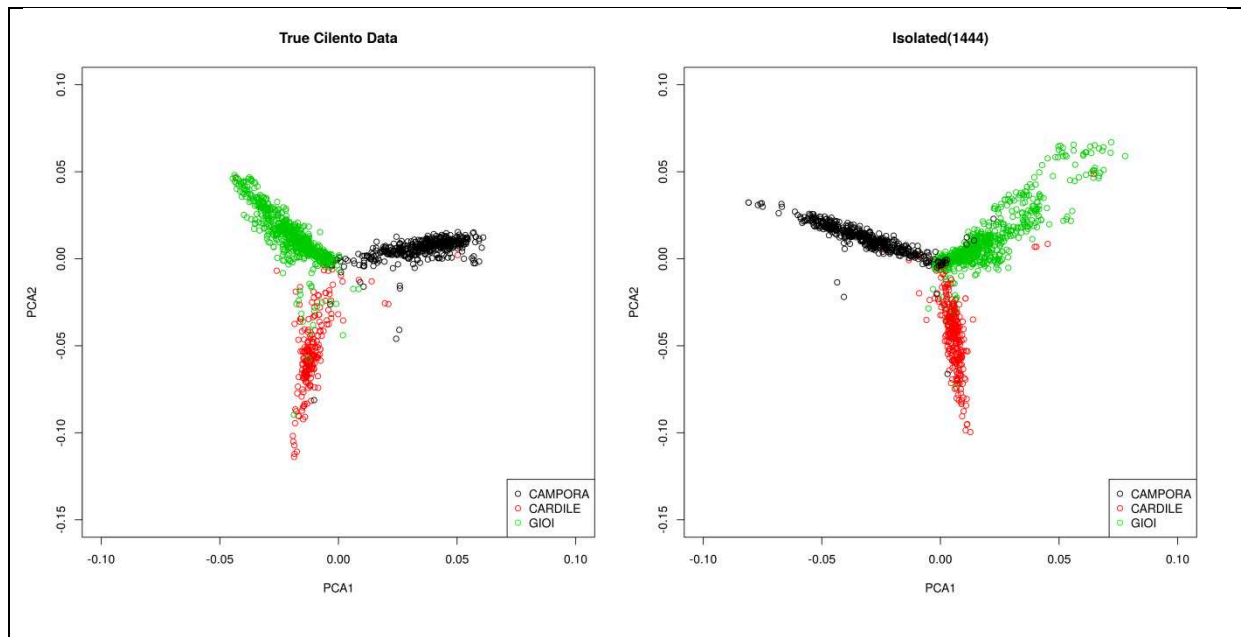
The corresponding MLE for the KDS model is added in green. We also add in the previously observed estimates from the literature (Table 1 in the main text).











Supplementary Figure 13.

Principle components analysis, performed both in the real data of Cilento and the simulated population Isolated(1444) which aimed to mimic Cilento.

<b>Supplementary Table 1</b>	Mean (sd)	Non-missing values following Quality Control	Number of outliers removed	Transformation
Phenotype				
Height	162.26(9.45)	1193	3	-
BMI	26.12(4.31)	1184	12	-
TGLY	4.77(0.52)	1326	16	Logarithmic
HDL	60.50(15.66)	1328	14	-
Total Chol	207.19(41.67)	1331	12	Adjusted for medications
LDL	118.16(35.71)	1299	12	Adjusted for medications

Abbreviations: BMI: Body-mass index; TGLY: Triglycerides; HDL: High-density lipoproteins; Total Chol: Total cholesterol; LDL: Low-density lipoproteins; sd: standard deviation.

Summary statistics for each of the seven traits studied after the removal of outliers and application of transformations. For TGLY we performed a logarithmic transformation, and the traits Total Chol and LDL had been adjusted for a small number of individuals who were recorded as having prescriptions for lipid lowering medications.

<b>Supplementary Table 2</b>	Cilento	Isolated(1444)	Outbred(1444)	Outbred(4332)	Outbred(8664)
$var(\lambda_a)$	1.192	1.127	0.030	0.089	0.176
$var(\lambda_d)$	0.109	0.096	0.015	0.044	0.088

Comparisons of estimated variances of the eigenvalues of the matrix K ( $\lambda_a$ ) and of the matrix D ( $\lambda_d$ ) across different simulated populations as well as the observed data in Cilento.

1 **References**

- 2 64. Su, Z., Marchini, J. & Donnelly, P. HAPGEN2: simulation of multiple disease SNPs.  
3 *Bioinformatics* **27**, 2304-2305 (2011).
- 4 66. Raffa, J.D. & Thompson, E.A. Power and Effective Study Size in Heritability Studies. *Stat*  
5 *Biosci* **8**, 264-283 (2016).
- 6 69. Delaneau, O., Zagury, J.-F. & Marchini, J. Improved whole-chromosome phasing for disease  
7 and population genetic studies. *Nat Meth* **10**, 5-6 (2013).
- 8 70. O'Connell, J. *et al.* A General Approach for Haplotype Phasing across the Full Spectrum of  
9 Relatedness. *PLoS Genetics* **10**, e1004234 (2014).
- 10 71. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat*  
11 *Genet* **48**, 1279-1283 (2016).
- 12 72. Bycroft, C. *et al.* Genome-wide genetic data on ~500,000 UK Biobank participants. *bioRxiv*  
13 (2017).
- 14 76. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based  
15 linkage analyses. *Am J Hum Genet* **81**, 559-75 (2007).
- 16
- 17