The Impact of Low-Coverage Whole-Genome Sequencing on Advancing Genetic Epidemiology Studies in Global Populations

Zhipeng Zhang

University of Oxford, St Cross College

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2025

Statement of authorship

All laboratory works on whole-genome sequencing and genotyping involved in this thesis was contributed by co-workers. Experimental work on two pilot cohorts (Chapter 2) and whole-genome sequencing of the Gambian Malaria Cases and Controls (GAMCC) samples (Chapter 3, 4, and 5) was conducted by Amy Trebes (formerly of Oxford Genomics Centre [OGC]) and James Docker (Technology Platforms Group [TPG]) and overseen by David Buck [OGC] and Paolo Piazza [TPG] at the Centre for Human Genetics, University of Oxford. Sample selection, quantification and preparation for shipment on the GAMCC cohort were undertaken by Annie Forster, who also provided advice on the GAMCC dataset, and Gavin Band. Whole-genome genotyping and Human Leukocyte Antigen typing of the same cohort were conducted by Eurofins Genetics and HistoGenetics, respectively.

The Vietnamese samples described in Chapter 6 were obtained from two studies organised by Graham Cooke and Barnaby Flower, extracted and collected by Chau Le Ngoc in Vietnam, and subsequently coordinated by George Airey at the University of Oxford. Whole-genome sequencing was performed by James Docker [TPG], Gabrielle Rockett [TPG], and Yanxia Wu [TPG], overseen by Paolo Piazza [TPG]. All computational analyses presented in this thesis represent my own work, except for the prototype of the low-coverage structural variant (lcSV) method in Chapter 5, which was

first developed by Gavin Band and subsequently adopted by myself. The massive computation tasks involved in this thesis were supported by the Biomedical Research Computing Cluster and team stuff. Lastly, I am deeply grateful for valuable advice from my supervisors Gavin Band and Azim Ansari that shaped this whole thesis.

Abbreviations: GAMCC: Gambian Malaria Cases and Controls; OGC: Oxford Genomics Centre; TPG: Technology Platforms Group; lcSV: low-coverage structural variant.

Zhipeng Zhang

8th of October, 2025

Acknowledgements

It has been a great while since the very beginning of my journey in Oxford, so distant that I barely remember any feeling and occurrence whilst jotting down this paragraph on my last flight back to Heathrow Terminal 2, last in the sense of living, studying, and being a part of the city and the university. Words fail me.

This three-year of pursuing a DPhil in a medical degree has doubtlessly witnessed the greatest transition of me the person and my life. It was initiated by the last question I was asked in the interview with Azim and Gavin, whom later became my supervisors and surely (I guess) forgot that little seed they placed into the deepest me. I have graduated from a mathematical finance degree and (attempted to) illustrate my professionalism in the basic subjects of mathematics and statistics with a naïve presentation of a deep learning model in pricing a stock, and it was the first time that I started to (and was inspired to) deeply think about why I am doing something. Not only is such a logical thinking crucial to my research throughout the years, but also makes me a better myself. Besides, the help I gained from them was enormous: be it the very rudimentary knowledge of statistical genomics my incredible project in low-coverage whole-genome sequencing that I have enjoyed so much (albeit not the originally proposed), or the countless insightful discussion about life and future. I present the deepest appreciation for guiding me through this intriguing DPhil journey, for bearing my mistakes and laziness, and for relentlessly sharing the wisdom of work and life.

The second bit is dedicated to my parents, without whose support I was never able to achieve any of my current self. I was not brilliant enough (and similarly now) to receive a scholarship from either the university or other entities, so my parents are the ones who bear all my expenses. I was taught since extremely young (maybe before the first time of my breathe) that the family is always more than happy to afford any cost associated with studying, which I have never taken for granted. Both of them are first-generation in our families attending college, and it was not somewhen recently that I felt their biggest wish on me (second to healthiness) was to accomplish a PhD, which later became one of the greatest motivations of doing so (albeit they never pressured so). I am thus deeply grateful to my parents, not only to their endless support, be it material or spiritual, but also this unconscious lesson I learned that education enlarges our horizon on life and endows us more possibilities, without which I would never be able to live a better life than currently. Likewise, I love you all.

Yet another source of enormous mental support was from my partner Peisha Kou, who stepped into my life when I was experiencing the most mudded period before my Transfer of Status. It was so abyssal that I barely remembered a smile for several months due to the great academic pressure and the similarly hazing weather in UK (in the second edition of this Acknowledgements bit, I want to add the last two months before submitting this thesis, which was equally plaguing). I was suffocating. Peisha spared

long time keeping me company by doing nothing special but just being around (though not physically but at the distal end of this planet). It was not until then that I felt love: love not in terms of doing amazing things, but greater love in merely being together and supporting relentlessly. We have always joked about tearing her certificate of graduation to half which should have been endowed to me as I helped in terms of writing or alike (though not totally wrong), and it is nothing different from this side. Thanks for always being around and bolstering without any complaints, both within the darkest period of mine and now, and we journey further.

Additionally, I am also grateful to the following many people: the Beijing Big Bens (my invaluable friends since high school with this name coming from some ancient memes but kept so as the same from acknowledgements of my Master's thesis, comprising Yiran Fu, Hanhao Hu, Chang Li, Huaiyu Peng, and Zizhang Cheng), the *earth dog bros* (a direct translation from Chinese for being non-chic named by Xijia [Cici] You, one of the dogs, comprising Cici You, Jinke Chang, Xianghe Zhu, Jiahao Ye, and Tianhuai Xu), my college advisor of four years Jo Ashbourn and all the lovely people of St Cross College (many retired), as well as my other invaluable friends yet to mention. Again, in the second edition of this Acknowledgements, I am thinking of dedicating more words to Jinke Chang and Xiaowei Liang, who are among the most people affecting me most in terms of this degree. I collaborated with Xiaowei many years ago on a paper (back then I was still doing organic chemistry, an absolutely

amazing field) and was thinking of doing a PhD. Xiaowei influenced me a lot on my thoughts about academia, or more specifically, life in academia. Then, I met Jinke Chang (thanks again Cici for that very beginning formal Dinner at Green Templeton College), from whom I learned considerable wisdom and knowledge (that I should or shouldn't) about work and life. We also spent much of the time complaining life in research, ate new dancing dragon, and drank pints and pints and pints of beers (and get drunk). I am to carry these enlightening perspectives (which surely already changed my life) to the future, and enormously grateful to both of you.

To colleagues. I am grateful to colleagues for performing experimental work involved in this thesis and their helpful discussions, constructive feedback, and encouragement during the course of this study. Specifically, I deeply appreciate James Docker, Gabrielle Rockett, Yanxia Wu, Paolo Piazza from the Technology Platforms Group (Centre for Human Genetics, University of Oxford), as well as Amy Trebes and David Buck from the previous Oxford Genomic Centre for all lab works and sequencing of the samples. I sincerely thank Robert Davies for the amazing QUILT software and all his help regarding the method, Yang Luo, Alex Mentzer, Annie Forster, and Ruth Nanjala for all valuable discussions regarding the HLA and low-coverage imputation in this region, Chau Le Ngoc for all invaluable conversations about the HCV samples, and Annie Forster (again) for sample selection, quantification and preparation for shipment of the GAMCC samples. Lastly, I would like to express my special thanks to

the Biomedical Research Computing Cluster and the team for supporting the massive computation involved in my project (and of course my supervisors for paying for this):

The computational aspect of this research employs the Biomedical Research Computing cluster, with the following quote: the research was supported by the Wellcome Trust Core Award Grant Number 203141/Z/16/Z with additional support from the NIHR Oxford BRC. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

Abstract

Genetic epidemiology connects genomics with population health, offering insights into disease susceptibility, progression, and treatment outcomes. Understanding how genetic variation shapes these traits reveals biological mechanisms and informs clinical interventions. Yet, most existing knowledge stems from European, American, and select Asian populations, leaving major gaps in underrepresented regions where infectious diseases are most common. Expanding genomic studies to these populations is essential for scientific equity and a fuller understanding of human biology. Low-coverage whole-genome sequencing (lcWGS) offers a cost-effective, unbiased alternative to genotyping arrays and deep sequencing. LcWGS with imputation enables accurate variant discovery without ascertainment bias, enhancing sensitivity to population-specific and complex variants and making it especially valuable for large-scale global studies.

This thesis explores the potential of lcWGS in genetic epidemiology through applications in the context of global populations with two cohorts from Africa and Asia. Using samples from The Gambia, I benchmarked lcWGS performance in detecting genome-wide variants, HLA alleles, and structural variants, demonstrating its strong potential for comprehensive variant discovery. In addition, I developed strategies to further improve imputation accuracy by refining computational workflows and enhancing population representation. Then, I applied lcWGS in a genome-wide association study of Hepatitis C Virus infection in a Vietnamese cohort, identifying a

regulatory variant of *OSBPL2* that likely influences viral replication. Together, these findings demonstrate lcWGS as a scalable, accurate, and inclusive approach for global genetic studies.

Extended Abstract

Genetic epidemiology studies represent a critical intersection of genomics and population health, providing essential insights for both clinical practice and fundamental research. The study of genetic variation across populations provides valuable insights into a wide range of traits, including pathogenic infections, where host genetic factors can influence susceptibility, disease progression, and treatment outcomes. Understanding these genetic determinants can reveal crucial mechanisms underlying disease pathophysiology, potentially guiding the development of novel therapeutic approaches and preventive strategies. To achieve the goal, accurate genomic data serves as the foundation for robust association studies that link genetic variants to disease phenotypes, thereby enabling the identification of risk factors. This precision is especially important in regions with high infectious disease burdens, where genetic association studies could reveal biology that is currently underexplored. At present, the benefits of such studies are concentrated in European, Asian, and American populations, while data from other populations remain scarce. This gap poses challenges, limiting both the equity and the breadth of biological insights that can be gained.

Central to the discourse on human genomic data generation are concerns regarding its comprehensiveness and cost-effectiveness. Traditional DNA microarrays provide reliable genotype calls at known genetic variants but focuses on predetermined markers. Whole-genome sequencing methods offer more comprehensive genetic information but

at considerably higher costs that often prohibit large-scale implementation, which is essential for genome-wide association studies where power is greatly affected by sample size. Low-coverage whole-genome sequencing (lcWGS) is now emerging as an innovative approach, combining affordability with impressive genome-wide accuracy rendered by recent developments of imputation algorithms like QUILT and GLIMPSE. Unlike microarrays, which are potentially prone to ascertainment bias due to their design based on specific populations (predominantly European due to historic reasons) or predetermined disease markers, lcWGS represents an unbiased approach that captures the complete spectrum of genetic variation, at least in principle. This characteristic makes lcWGS particularly valuable for studying global populations where standard arrays may miss population-specific variants crucial to understanding disease mechanisms. Additionally, lcWGS has been shown to accurately capture complex variation in clinically relevant regions such as the Human Leukocyte Antigen (HLA) region, which is strongly associated with immune-mediated diseases and pathogenic infections. Beyond this, lcWGS holds promise for capturing large and common structural variations, although methods for this task have not yet been established and remain under development.

This thesis aims to demonstrate the substantial potential of lcWGS applied to two global populations. In the first part of this thesis, I compared the lcWGS approach to genotyping data from traditional DNA microarrays on a single set of samples from The

Gambia. I demonstrated the capability of lcWGS with imputation in capturing genome-wide variants with 1000 Genomes Project reference panel by comparing with genotyping data. To further explore the larger TOPMed reference panel which is only accessible from the imputation server, I employed a two-stage imputation approach by retaining reliably imputed variants and resubmitting for further refinement, resulting in superior accuracy. Second, I investigated the performance of lcWGS in complex genomic variants and reinforce its utility in incorporating these into genome-wide association studies (GWAS). Specifically, I performed HLA imputation and compared the results with truth by sequence-based typing. Despite the compromised initial imputation, I optimised the method by refining the sequence-to-allele alignment and revise an HLA phasing strategy, which was later employed to construct a Gambian-enriched reference panel. I also developed and refined a population-based method in calling large structural variants, demonstrated its practical utility in inferring some previously established deletion haplotypes, and highlighted some challenges inherent by lcWGS. These results exemplify in practice that lcWGS is a viable approach for GWAS studies in African. In the last part of this thesis, I applied lcWGS to a GWAS of Hepatitis C Virus phenotypes in a Vietnamese cohort to investigate genetic determinants of disease outcomes and identified a plausible regulatory variant in *OSBPL2* that may impair HCV replication by altering lipid homeostasis. Collectively, these results provide a comprehensive characterisation of the lcWGS workflow, demonstrating its robustness and practical applicability in contemporary genetic epidemiology studies.

List of Abbreviations

ABO	ABO blood group
ATP2B4	Plasma Membrane Calcium-Transporting ATPase 4
BIC	Bayesian Information Criteria
bp	base pair
CCS	Circular Consensus Sequencing
CD	Cluster of Differentiation
CNV	Copy Number Variation
dbSNP	Single Nucleotide Polymorphism Database
ddNTP	dideoxynucleotide
dNTP	deoxynucleotide
DKK2	Dickkopf WNT Signalling Pathway Inhibitor 2
DNA	deoxyribonucleic acid
DNAP	Illumina DNA Prep
FDR	False Discovery Rate
FCGR	Fc Gamma Receptor
GAMCC	Gambian Malaria Cases and Controls
GTEx	Genotype-Tissue Expression
gnomAD	Genome Aggregation Database
GRCh37/38	Genome Reference Consortium Human Build 37/38
GWAS	Genome-Wide Association Study
GYP	Glycophorin
H3Africa	Human Heredity and Health in Africa
HBA	Haemoglobin Subunit Alpha
HBB	Haemoglobin Subunit Beta
HCV	Hepatitis C virus
HiFi	High Fidelity
HIV	Human Immunodeficiency Virus
HLA	Human Leukocyte Antigen
indel	insertion-deletion
IFNL	Interferon Lambda
IPD-IMGT	Immuno Polymorphism Database and international ImMunoGeneTics information system
JAK	Janus Kinase
kb	kilo base
lcSV	low-coverage Structural Variant
lcWGS	low-coverage Whole-Genome Sequencing
LD	Linkage Disequilibrium
MalariaGEN	Malaria Genomic Epidemiology Network
Mb	Mega base

NEB	New England Biolabs Next® Ultra™ DNA Library Prep Kit for Illumina
NETO1	Neuropilin and Tolloid-like 1
ng	nanogram
NGS	Next-Generation Sequencing
NIPT	Non-Invasive Prenatal Testing
Omni5M	Illumina HumanOmni5M-4v1 array
ONT	Oxford Nanopore Technologies
OSBPL2	Oxysterol Binding Protein Like 2
PacBio	Pacific Biosciences
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
QC	Quality Control
RNA	ribonucleic acid
SNP	Single Nucleotide Polymorphism
sQTL	splice Quantitative Trait Locus
STAT	Signal Transducer and Activator of Transcription
SV	Structural Variant
TOPMed	Trans-Omics for Precision Medicine
UIIFS	New England Biolabs Next® Ultra™ II FS DNA Library Prep Kit for Illumina
VEP	Variant Effect Predictor
vs.	versus
µl	microlitre

Table of contents

List of Figures

List of Tables

Chapter 1 Introduction

1.1 Thesis objectives and composition

The aim of this thesis is to investigate and demonstrate the potential of low-coverage whole-genome sequencing (lcWGS) as a scalable and unbiased approach for generating high-resolution genomic data in the context of genetic epidemiology. Specifically, I evaluate lcWGS as a viable alternative to traditional genotyping platforms from the perspectives of its unbiasedness, accuracy, and capability of capturing a whole scope of genetic variants, particularly in studies involving global populations. Then, I apply lcWGS to a GWAS of Hepatitis C Virus (HCV) infected Vietnamese cohort to investigate the impact of genomic variants on infectious disease outcomes, further demonstrating the utility of lcWGS in conducting genetic epidemiology studies.

The first part of the thesis focuses on optimising lcWGS data generation and evaluating its accuracy. Chapter 2 explores different library preparation conditions to assess their influence on sequencing quality. With a high-performing protocol, genomic data for a Gambian cohort is generated, along with those obtained from DNA microarrays to benchmark the accuracy of lcWGS. Chapters 3 and 4 delve into the performance of lcWGS in detecting genome-wide variants and resolving the highly polymorphic Human Leukocyte Antigen (HLA) region that is known to be biologically essential yet

challenging due to its high polymorphic nature, respectively. Chapter 5 extends the analysis by illustrating the capacity of lcWGS in capturing large structural variants (SVs), an important class of genomic variation with known implications for disease outcomes. Overall, these chapters highlight the substantial technical potential that lcWGS offers for providing comprehensive and accurate genomic information.

Building upon these methodological insights, Chapter 6 applies lcWGS in a genetic association study involving a Vietnamese population infected with HCV. This application aims to identify host genetic factors that contribute to differential immune responses and clinical outcomes, demonstrating the real-world utility of lcWGS in uncovering biologically meaningful associations in diverse populations. By integrating technological validation with applied epidemiological research, this thesis aims to demonstrate the value of lcWGS as a cost-effective and accurate approach for generating comprehensive genomic data across diverse populations, thereby enabling deeper insights into the genetic underpinnings of infectious diseases.

1.2 DNA genotyping and sequencing methods

Fundamental to genetic research is the ability to generate reliable and comprehensive genomic data. Broadly, this can be achieved through two primary approaches: sequencing-based methods, which directly read nucleotide sequences and rely on computational algorithms to assemble and identify genetic variants, as well as

genotyping-based methods, which detect predefined variants using array-based designs. In this section, I outline the evolution of sequencing and genotyping technologies for generating human genomic data, from early low-throughput approaches to current high-throughput platforms. I then focus on the emergence of lcWGS, highlighting the motivations behind its development and its growing role as a cost-effective solution for large-scale genetic studies. Since the context of this thesis only involves human DNA samples, the following sessions are phrased as such and do not review technologies relevant to other types of genomic data.

1.2.1 Whole-genome sequencing techniques

The earliest sequencing method, known as Sanger sequencing or the chain-termination method, was developed by Frederick Sanger in 1977 and remained the gold standard for accuracy in variant detection [1-5]. Together with other key technological innovations in sequencing and computation tools, Sanger sequencing was largely employed in the Human Genome Project [6]. The process begins with polymerase chain reaction (PCR) amplification of the DNA region of interest using sequence-specific primers that flank the target and DNA polymerase. The amplified DNA is then subjected to a modified enzymatic synthesis of DNA strands in the presence of both normal non-fluorescent deoxynucleotides (dNTPs) and a small proportion of fluorescent dideoxynucleotides (ddNTPs), which lack a 3'-hydroxyl group required for chain elongation. When a ddNTP is incorporated into a growing DNA strand, it causes termination at that specific base. The resulting DNA fragments of varying lengths are

then separated by capillary electrophoresis, in which fragments migrate through a thin polymer-filled capillary. As they pass a detector window, the terminal base of each fragment, identified by its fluorescently labelled ddNTP, is recorded [7]. The order of fragment lengths reveals the original DNA sequence. While Sanger sequencing offers high per-base accuracy and thus still thrives in mutation detection, it is inherently low-throughput and cost-prohibitive for large-scale applications, making it impractical for population-level genetic studies.

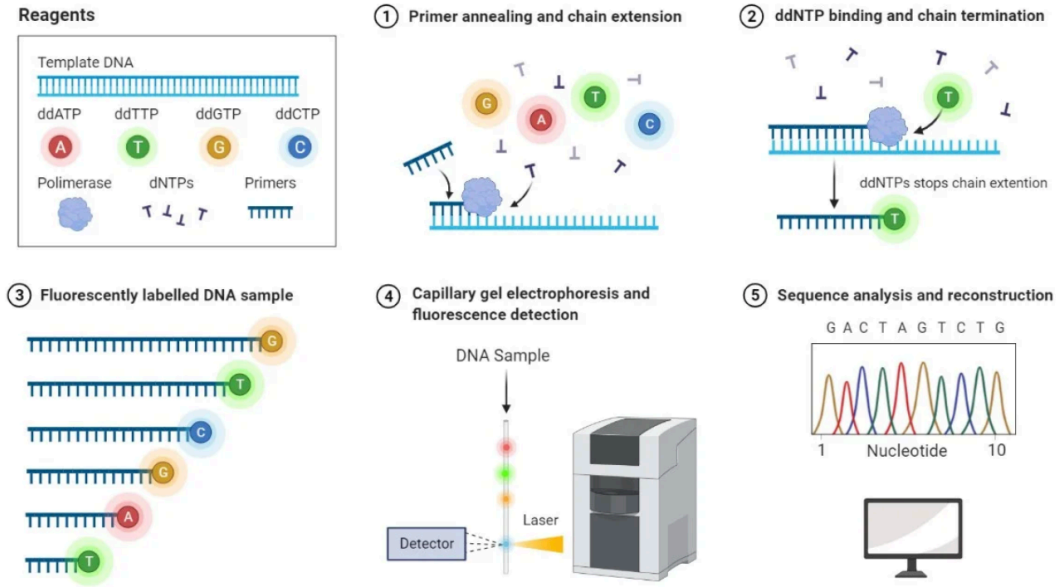


Figure 1.2.1. Sanger Sequencing. This figure was created by Ona, S. (2025), BioRender [8].

The advent of next-generation sequencing (NGS) technologies in the mid-2000s revolutionised the field of genomics by enabling massively parallel sequencing of millions of DNA fragments simultaneously. The first decade of 21st century has

witnessed a wide flourish of various technologies that surpass the older Sanger sequencing by a factor of 100 to 1000 in daily throughput while substantially reducing per-base cost [9-17], thereby subsequently termed as NGS [18]. This transformation was driven largely by key breakthroughs pioneered by Solexa, including the engineering of a polymerase capable of efficiently incorporating the fluorescently modified nucleotides [19, 20] and the adoption of the molecular clustering approach [21, 22] to reduce noise and error in the sequencing process [23]. Following its acquisition of Solexa in early 2007, Illumina further developed and scaled the technology to enable large-scale, high-throughput sequencing: for example, the Illumina NovaSeq platform which was widely used to sequence samples involved in this thesis [24]. In this method, genomic DNA is fragmented, ligated to adaptors which contains oligonucleotides that bind to the flowcell and individual identifiers, and optionally subject to PCR especially if the input amount of DNA materials are limited. These fragments are then transferred to a flowcell where they hybridise with oligos that are covalently attached to the solid surface and thus immobilised. To enhance the signal in the following sequencing step, each molecule is amplified into clusters through an isothermal bridge amplification while the undesired strands are washed away. Now that each cluster contains the same molecule, fluorescently labelled nucleotides are incorporated one base at a time by DNA polymerase. After each incorporation, a laser excites the fluorophores and a high-resolution camera captures the signal, identifying the added base before the terminator is cleaved to allow the next cycle. This cycle

continues across all clusters in parallel, generating millions of short reads in a single run. Since each DNA fragment has two ends, this process can be optionally repeated to reversely read the same molecule to maximise data output (to form read 2 in this method of paired-end sequencing). While NGS dramatically reduces per-base cost and saves considerable time thereby making whole-genome sequencing feasible at a population scale, it generally produces shorter reads (typically up to 600 bases [25]) with compromised accuracy compared to Sanger sequencing. These shorter and less accurate reads pose challenges for reliably assembling regions with complex genomic architecture, such as segmental duplications, homopolymer stretches, and large SVs.

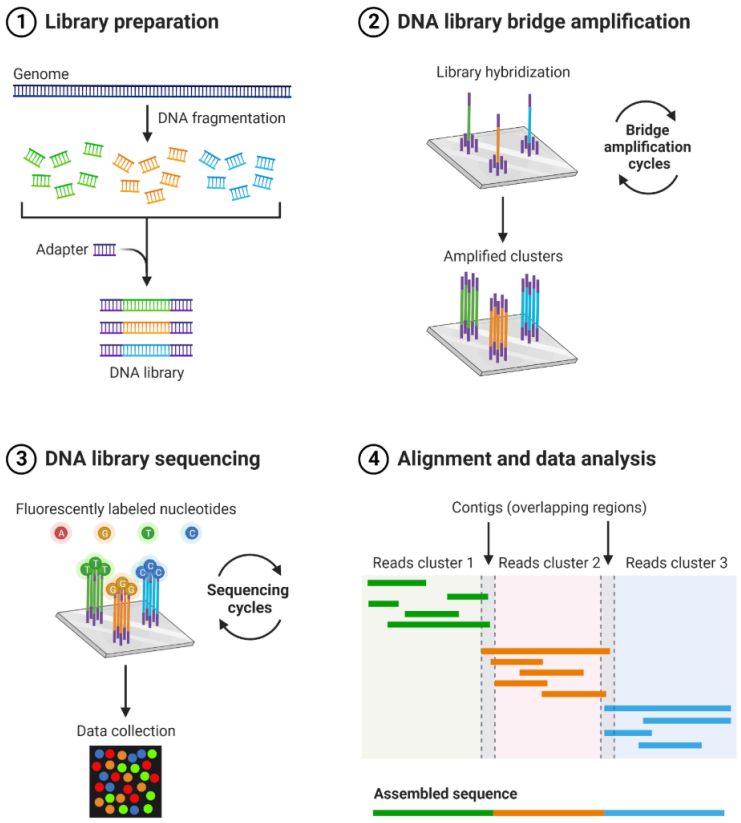


Figure 1.2.2. Illumina next-generation sequencing. This figure was created by Ona, S. (2025), BioRender [26].

To overcome these limitations, Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) have developed long-read sequencing technologies, capable of producing reads that range from kilobases to even megabases in length [27-32]. The PacBio High Fidelity (HiFi) chemistry involves ligating adapters to double-stranded DNA to form a circular molecule, which is then repeatedly sequenced by DNA polymerase incorporating fluorescently labelled nucleotides, allowing high-accuracy consensus reads from multiple passes of the same template. This strategy is known as circular consensus sequencing [32]. In contrast, ONT sequences native DNA molecules as they pass through a biological nanopore embedded in a membrane. As the double-stranded DNA molecule is unwounded to single-stranded DNA which subsequently translocates through the pore, changes in electric current are measured and translated into nucleotide sequences via computational algorithms. Long-read technologies offer clear advantages in resolving complex regions and the possibility of preserving epigenetics if PCR-free workflows are employed [33], making them a valuable complement to short-read sequencing in genomics research. Nevertheless, long-read technologies remain higher cost than short-read sequencing, limiting cost-effectiveness for large-scale cohorts.

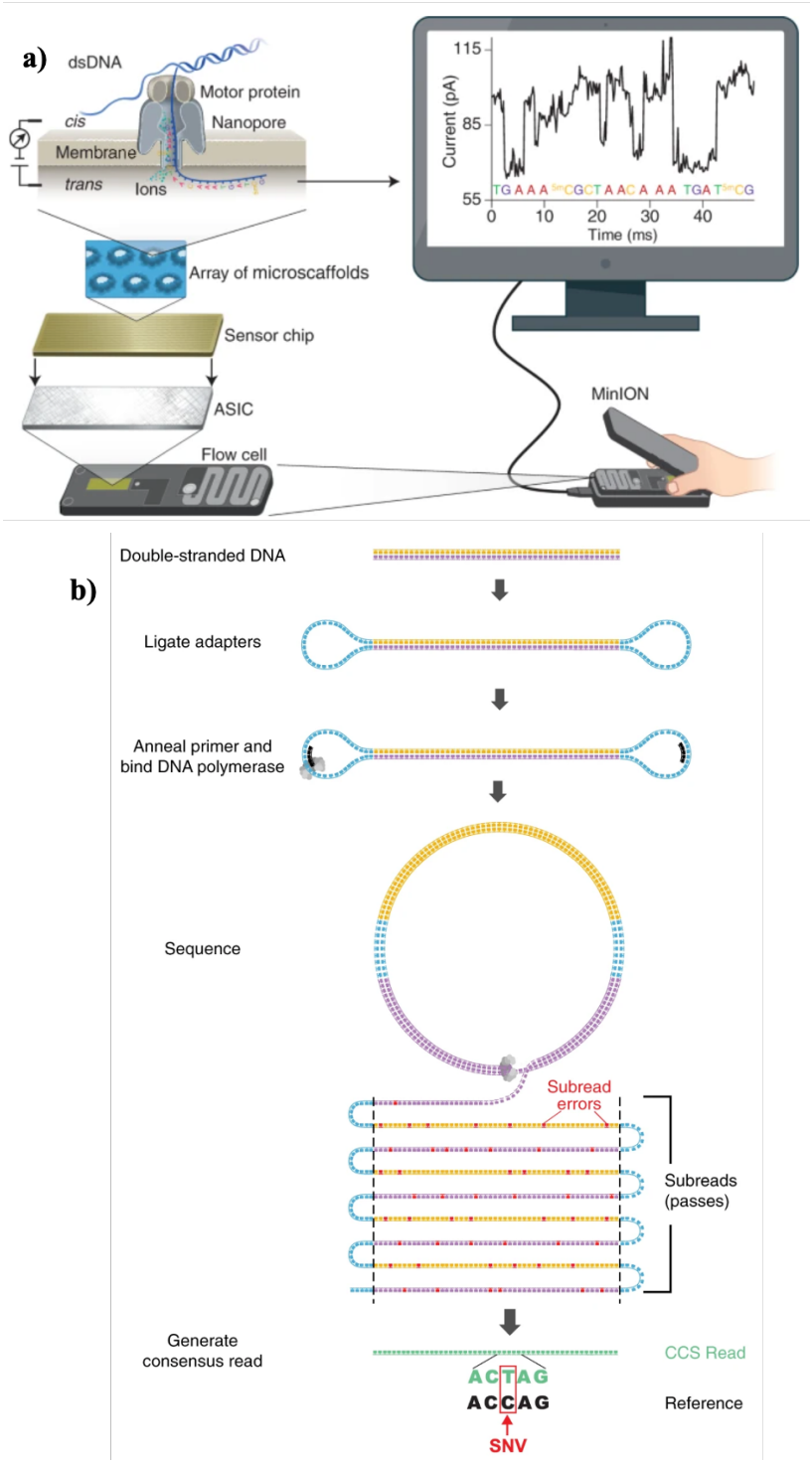


Figure 1.2.3. Long-read sequencing. a) ONT long-read sequencing (Figure 1 in [31]). **b)** PacBio long-read circular consensus sequencing (Figure 1 in [32]).

1.2.2 Genotyping approaches

Microarrays, also known as chips, were among the earliest high-throughput technologies developed to capture genetic variation. First demonstrated in 1995 with the introduction of dual-labelled quenched probes that enabled real-time fluorescence detection [34], microarrays have since been widely applied for genome-wide association studies (GWAS), phasing, fine mapping, and linkage disequilibrium (LD) studies at population scale [35-42]. Unlike sequencing-based approaches that read nucleotide sequences directly, microarrays are designed to detect known genetic variants through a process of hybridisation between fragmented genomic DNA and pre-synthesised oligonucleotide probes immobilised on a solid surface. Each probe targets a specific single nucleotide polymorphism (SNP) or insertion-deletion (indel), and the hybridisation signal (typically detected via fluorescence) indicates the presence or absence of a particular allele in a sample [43]. The resulting intensities are resolved to genotype calls via computational algorithms, distinguishing between homozygous and heterozygous states at each targeted variant site, thus enabling accurate downstream analyses across millions of loci [37].

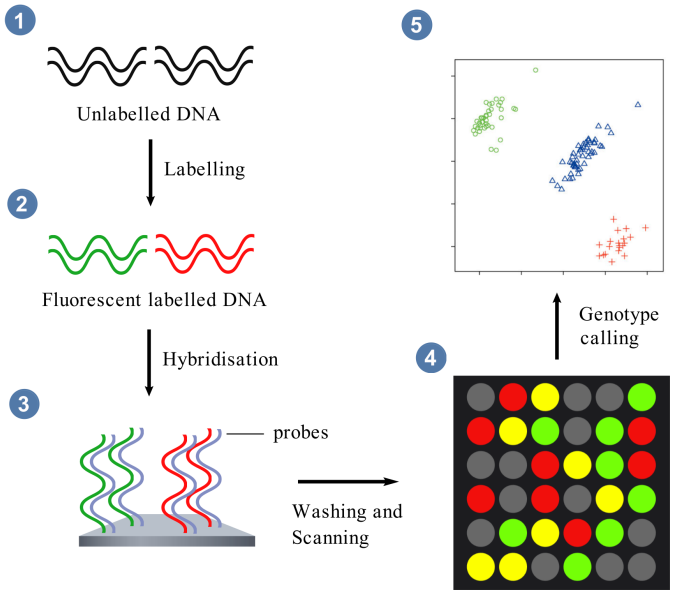


Figure 1.2.4. An example of DNA microarray workflow. This figure was created with BioGDP.com [44]. The general workflow was adapted from Figure 1 in [43]. The microarray illustration (circled number 4) was created with BioRender.com (BioRender, 2025). [45]. The cluster plot (circled number 5) was adapted from Figure 1 in [46].

Microarrays have evolved significantly in sophistication and versatility over the past three decades [35, 47]. Early designs were primarily informed by variants identified in individuals of European ancestry, particularly through the International HapMap Project, resulting in limited representation of global genetic diversity [48]. With the advent of larger and more diverse reference databases such as the 1000 Genomes Project [49, 50] and the Genome Aggregation Database (gnomAD) [51], as well as an expanding understanding of genetic variants linked to disease phenotypes, modern arrays now capture a broader spectrum of variation, including clinically relevant SNPs, rare variants, and population-specific markers [52]. Furthermore, arrays are available in a range of densities, from cost-efficient, low-density panels for targeted screening to

high-density platforms optimised for comprehensive variant coverage, supporting broader applications across diverse and historically underrepresented populations [35].

Nevertheless, microarrays still bear several limitations by nature. First, microarrays suffer from ascertainment bias, as the included markers are predetermined by prior knowledge that is often derived from specific populations or disease contexts [53]. Second, even high-density arrays provide only a sparse sampling of the genome, capturing a subset of common variants and relying heavily on imputation to infer missing genotypes. While imputation algorithms can partially address this issue by inferring untyped variants using reference panels, the resulting accuracy may be compromised, especially in genetically diverse populations [54-58]. Third, microarrays have limited power of detecting large SVs and copy number variants in the capability of defining breakpoints, resolving single copy gains and heterozygous deletions, and accessing repeat-rich regions due to limited probe set and sensitivity [59]. These inherently complex variants often require sequence-based approaches for reliable identification yet are closely linked to disease phenotypes [60, 61]. Thus, while microarrays remain a powerful and cost-effective tool in genetic epidemiology studies, genotyping methods are not without challenges in certain settings.

1.2.3 The emergence of low-coverage whole-genome sequencing

Ongoing reductions in DNA sequencing costs have positioned lcWGS as a compelling alternative to traditional genotyping arrays for genetic epidemiology studies, offering a scalable and cost-effective strategy for capturing genome-wide variation in large cohorts [62, 63]. Unlike deep whole-genome sequencing, in which each base is sequenced multiple times to ensure reliable genotype calls, lcWGS reduces sequencing depth to as low as 0.1×, that is, on average, every ten bases is covered once, by pooling large number of samples and sequencing them together on a single flowcell. This design reduces per-sample sequencing costs while retaining the ability to capture common variants, which are typically the primary focus of GWAS, as rare variants remain difficult to detect without substantially larger cohorts. However, these costs are variable and can be influenced by factors such as bulk purchases, institutional discounts, and other considerations. In the study described in Chapter 3, we achieved approximately £10 for sequencing and £29 for library preparation per sample at an average coverage of 1.21×. Although lcWGS produces far fewer reads per genome, recent advances in genotype imputation algorithms, such as GLIMPSE [64, 65] and QUILT [66, 67], have improved the ability to infer genotypes with high accuracy from sparse data and been proven in recent studies across Hispanics [68], Europeans [69-72], Africans [70, 73, 74], Near Oceania populations [75], global populations [76, 77], as well as non-human genomes [78-86].

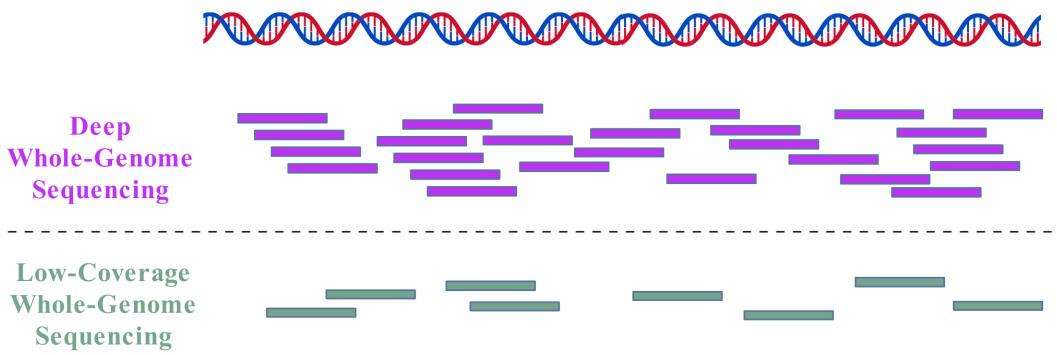


Figure 1.2.5. Deep whole-genome sequencing and lcWGS. This figure was created with BioGDP.com [44].

Motivated by these insights, we have been applying lcWGS to real-world cohorts from developing countries that exhibit both complex genetic backgrounds and high burdens of pathogenic infection. As part of this work, I aim to develop robust pipelines to process and leverage lcWGS data in real datasets and to benchmark its performance against whole-genome microarray genotyping. Specifically, I enhance the current imputation workflow to maximise data acquisition, assess its ability to accurately capture variants in complex genomic regions such as the HLA locus without incurring additional costs, and demonstrate its capacity to call large and common SVs that are clinically crucial. These efforts culminate in our own GWAS analyses presented in this thesis and thenceforth, highlighting the transformative role of lcWGS in advancing genetic epidemiology studies particularly from global populations.

1.3 Genotype calling and imputation

Accurate genotype information forms the foundation of genetic association studies. In

deep whole-genome sequencing, genotypes can be reliably obtained by aligning sequencing reads to a reference genome, identifying variant sites, and directly calling genotypes for each individual, as high per-base coverage provides sufficient data to observe most variants with high confidence [87]. However, in more cost-effective approaches such as lcWGS or DNA microarrays, much of the genome is only sparsely captured by the respective technologies, resulting in either missing or only vaguely characterised representation. In these cases, imputation becomes essential to infer genotypes at unobserved or uncertain positions by leveraging population-level LD patterns and large reference panels [54, 55, 58, 88].

The foundational model in genotype imputation, known as the Li and Stephens model [58], employs a probabilistic framework for modelling LD through a hidden Markov model. This model forms the theoretical backbone for most contemporary phasing and imputation algorithms, although alternative methodologies are present and similar [64-67, 89-92]. The Li and Stephens model assumes that each individual haplotype can be represented as a mosaic of segments copied from a reference panel of known haplotypes, reflecting the historical recombination events and LD structure in the population. At any given position, the model either copies (possibly imperfectly) the current haplotype or switches to a different one, with transition probabilities determined by recombination rate between adjacent markers. This copying process is formalised through a Markov chain, where each hidden state corresponds to a reference haplotype, and emissions

represent the observed alleles. Such an implementation not only overcomes the limitations of previous approaches that usually assumes block-like LD structure, overlooks recombination events, or only measures adjacent sites independently, but also maintains the computational tractability for long chromosome regions [58].

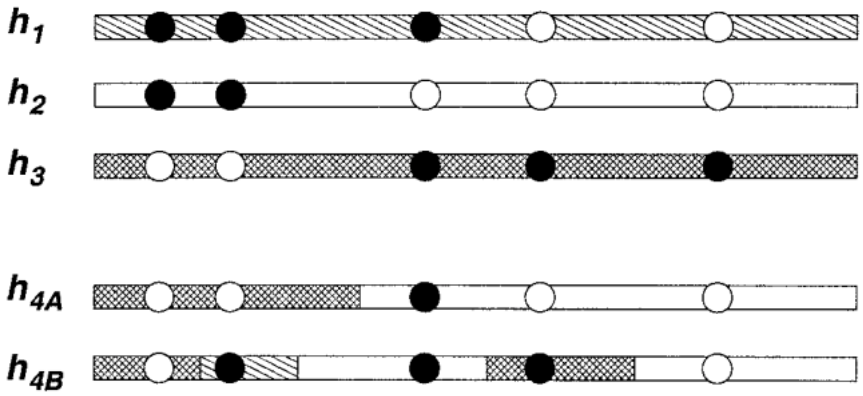


Figure 1.3.1. Li and Stephens model. This figure was adapted from Figure 2 in [58]. Suppose h_1, h_2, and h_3 are reference haplotypes (each with different shading) and the circles are SNPs. The graph below (h_{4A} and h_{4B}) depicts two possible values of an inferred haplotype, where they are mosaic copies of the segments from the reference panel. An imperfect copy exemplified in the third locus where h_{4A} and h_{4B} has a black allele yet copied from h_2 possessing a white allele, which might originate from mutations. Li and Stephens model is by essence a hidden Markov model, with the hidden state being each shading and the emission as observed SNPs, giving rise to the probability of observing each possible haplotype.

Building upon the Li and Stephens framework, recent imputation methods such as QUILT [66, 67] and GLIMPSE [64, 65] have adapted for lcWGS data and achieved higher accuracy compared to previous methods [93-96]. These advances have enabled lcWGS to emerge as a viable alternative to microarrays [97, 98]. On one hand, the dependence between nearby SNPs in sequencing reads contrasts with the independence of short genotyping array probes, presenting unique challenges for variant interpretation.

This complexity is further amplified in highly polymorphic regions, diverse populations, and with long-read platforms like ONT and PacBio. Additionally, a read comes from either the paternal or the maternal haplotype. If the parental origin of each read can be inferred, which is not possible with microarrays, haploid imputation becomes feasible, substantially reducing computational complexity. This simplification not only accelerates computation but also enables scalable and efficient imputation across thousands of samples. These factors highlight the need for specialised imputation algorithms tailored to the characteristics of lcWGS data. Accordingly, GLIMPSE takes raw diploid genotype likelihoods, employs Gibbs sampling to refine them by iterating over selecting haplotypes from reference panel and estimating the target haplotypes with the Li and Stephens model, and finally emits genotype posteriors. On the other hand, QUILT processes the sequencing reads directly. It also relies on Gibbs sampling between updating partitioned read labels into maternal and paternal sets based on the current subset of reference panel and full haploid imputation to update this subset based on the current read labels. After a pre-specified number of iterations, it emits the finally imputed genotype dosages. By enabling fast, accurate, and scalable imputation, these methods make it feasible to harness the full potential of expansive and diverse reference panels.

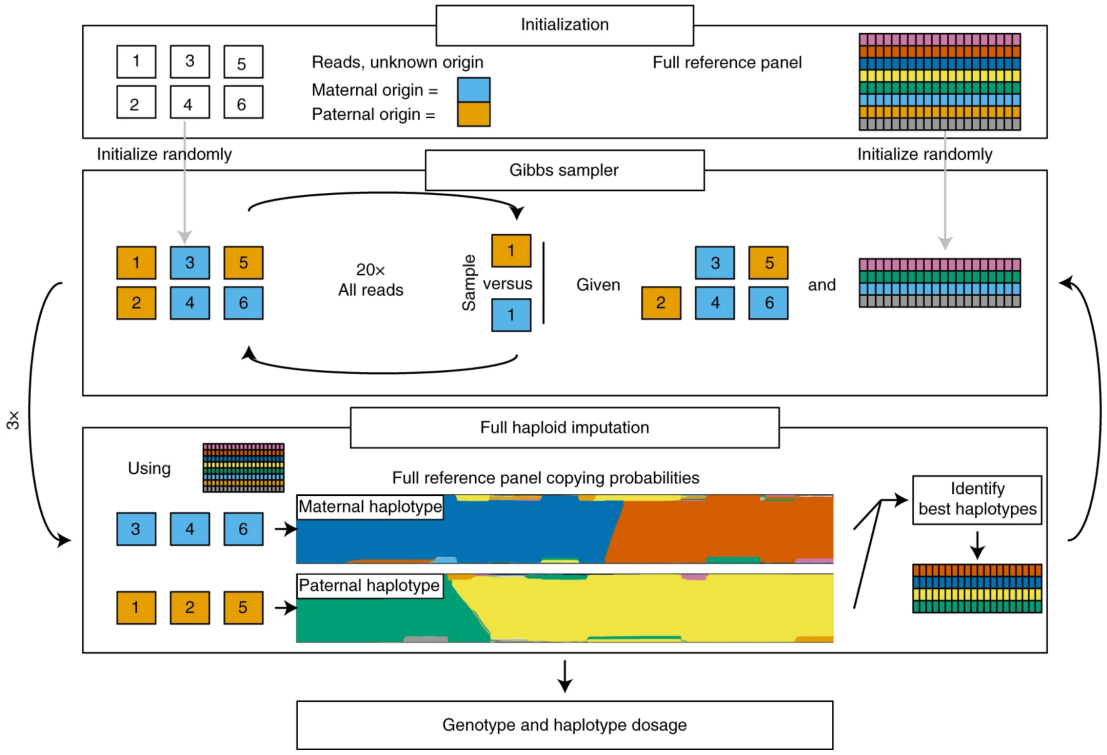


Figure 1.3.2. The QUILT method on lcWGS imputation. This figure was adapted from Figure 1 in [66]. In brief, the QUILT method starts with sequencing reads and a reference panel. It proceeds between Gibbs sampling to determine the parental originality of reads using a subset of reference panel as well as perform haploid imputation with hidden Markov model using the reads to refine the current subset of reference panel. After a pre-specified number of iterations, QUILT outputs genotype dosage as well as the final phasing.

Imputation accuracy is influenced by many factors. Variants with lower allele frequency, complicated genetic architecture (for example, large SVs), or those that reside in regions with complex LD structure are harder to impute [88]. In addition to choosing appropriate imputation software as discussed in the previous paragraph, the selection of reference panels plays a crucial role in imputation accuracy. Reference panels that contain a larger number of haplotypes and are genetically similar to the target population are generally more favourable, as they offer better representation of allele frequencies and haplotype diversity. Over the past two decades, an increasing number

of reference panels have been released along with an expansion in the number of samples included to support genotype imputation. The HapMap Project was the first publicly available reference panel and laid the groundwork for cataloguing common genetic variation [48, 99, 100]. To gain a deeper understanding of human genome variation for further elucidating genotype-phenotype associations, the 1000 Genomes Project was setup by incorporating a broader set of populations [49, 50, 101, 102], where the final phase contains 3,202 individuals. The Haplotype Reference Consortium thrived to build a larger haplotype reference panel (with 64,976 haplotypes) and to provide a centralised resource for genotype imputation by combining datasets from 20 different studies, although it is still overwhelmed by individuals with European ancestry [103]. More recently, the Trans-Omics for Precision Medicine (TOPMed) programme aimed to uncover genetic factors underlying cardiopulmonary and blood disorders to advance diagnosis and treatment, where its initial phase released deep whole-genome sequencing data and phenotyping for 53,831 ethnically and ancestrally diverse samples [104]. In parallel with these global initiatives, several population-specific reference panels emerged to improve imputation performance in non-European populations, including GenomeAsia100K [105], the Consortium on Asthma among African-ancestry Populations in the Americas [106], the African Genome Resources panel [107], BioBank Japan [108, 109], the Genome of Netherlands Project [110], The UK10K Cohorts [111], a Sardinia cohort [112], the Korean Variant Archive [113, 114], the Gambian Genome Variant Project [115], and so on. The continued development and

expansion of both global and local reference panels remain critical for achieving high-resolution and population-relevant genotype imputation.

Apart from genome-wide imputation, there remains a critical need to develop specialised methods to infer other genetically more complex variants for lcWGS, as standard imputation approaches often struggle to accurately capture highly polymorphic regions like the HLA locus and large SVs due to complex LD patterns, population specificity, and sequencing and alignment artifacts. Motivated by these, I discuss methodological efforts to perform HLA imputation and SV calling in Chapters 4 and 5 to address this research gap, respectively.

1.4 The HLA region

Apart from genome-wide variants that capture a broad reservoir of genetic diversity underlying different disease aetiologies, particular attention should be given to genomic regions of known biological importance. These may be loci previously implicated in specific diseases or regions more broadly associated with diverse disease outcomes, even when they present technical challenges. Such regions often exhibit complex genetic architectures, extremely high polymorphism, and special LD patterns that can be population specific, making them difficult to genotype directly.

The HLA region, which is part of the Major Histocompatibility Complex, is a key

genomic locus that plays a central role in immune responses, disease susceptibility, and organ transplantation compatibility. It is located on the short arm of chromosome 6 (6p21) and spans a physical ~3.6 Mega base (Mb) region, though it can effectively encompass a much larger genomic area through LD and extended haplotypes. The region contains a dense cluster of genes, many of which encode proteins essential for immune system function, including those involved in antigen presentation, antibody signalling, and apoptosis. The HLA region is broadly divided into three major classical gene classes: I, II, and III. Class I genes (HLA-A, HLA-B, and HLA-C) encode molecules that present endogenous peptides to $CD8^+$ T cells, whereas class II genes (HLA-DP, HLA-DM, HLA-DO, HLA-DQ, and HLA-DR) encode molecules that present exogenous peptides to $CD4^+$ T cells, facilitating self-antigen recognition and immune responses against extracellular pathogens. Apart from the antigen-presenting HLA class I and class II genes, there are also many other genes present in the region, although they generally exhibit much lower allelic variability. Class III genes encode components of the complement system, which orchestrates human immune response by leveraging both innate and adaptive immune system [116].

To date, a plethora of HLA class I and class II alleles have been shown to be associated with a wide range of diseases [117]. For example, HLA-DRB1*15:01 is associated with an increased risk of multiple sclerosis [118] and DQB1*02 and DQB1*03:02 with risk of coeliac disease [119]. Class I alleles, HLA-A*24:02 and HLA-B*39:06, as well as

HLA-DQB1*03:02 are associated with type 1 diabetes, as well as extended haplotypes comprised of alleles across multiple class I and II loci [120-122]. Evidence from studies in a Gambian population suggests that HLA-DRB1*13:02 confers a protective effect against Hepatitis B Virus (HBV) infection [123, 124], exemplifying the relationship between HLA alleles and viral infection. HLA class I and class II alleles encode polymorphic heterodimeric molecules (allotypes) that bind peptides derived from self-antigens and present them to T cells both during thymic development and throughout immune surveillance in the periphery, to establish and maintain immune tolerance. They also recognise non-self or foreign antigens triggering T and B cell responses to infectious organisms [125-128]. Examining genetic associations between specific HLA alleles or haplotypes with diseases or pathogenic infection is thus essential, underscoring the need for accurate and cost-effective HLA typing and phenotypic data.

In this section, I begin by introducing the complex HLA nomenclature system used to describe polymorphic HLA alleles. I then provide a systematic review of key HLA imputation algorithms, with an emphasis on those developed for genotyping and whole-genome sequencing data. This thesis investigates HLA inference in Chapter 4.

1.4.1 HLA alleles and nomenclature

More than 40,000 HLA alleles have been recorded in the IPD-IMGT/HLA database (abbreviation of the Immuno Polymorphism Database and the international ImMunoGeneTics information system) [129]. Accompanying the expanding number of

HLA alleles, we have also witnessed a more than 50-year progress on naming conventions since the very beginning 1968 nomenclatures [130, 131]. Taking allele HLA-A*02:09:01:03 as an example, the current nomenclature specifies first the region where this allele resides (HLA-A) followed by a separator (*). Afterwards, colons (known as field separators) are used to delineate the boundaries of each field, which carry different level of details of the allele:

- The first field (02) indicates serological antigen carried by the allotype.
- The second field (09) specifies different nucleotide substitutions that alter the encoded amino acid sequence.
- The third field (01) distinguishes alleles differ only by synonymous nucleotide substitutions within the coding sequence.
- The fourth field (03) corresponds to different genomic sequences in introns or the 5' and 3' untranslated regions that flank the exons and introns.

An optional letter can be added to the end of an HLA allele to indicate its expression status [132].

Optional suffix	Expression status
N	**N**ull indicating the allele is not expressed.
L	**L**ow cell surface expression compared to normal levels.
S	**S**ecreted as the protein is soluble but not present on cell surface.
C	Present in the **C**ytoplasm and not on cell surface.
A	**A**berrant as whether the protein is expressed is doubtful.
Q	**Q**uestionable given the allele has been shown to affect normal expression levels in other alleles, yet not confirmed.

Table 1.4.1. Optional suffixes in HLA nomenclature.

1.4.2 HLA typing and imputation

Albeit the highly polymorphic nature and complex LD patterns in the HLA region, accurately typing HLA alleles is possible through the earliest serological methods to sequencing methods [133-140]. Since serological methods fail to capture HLA polymorphism which is most prominent in the exonic regions, HLA sequence-based typing has emerged as the primary method for accurate HLA typing [135, 136]. Sequence-based typing typically involves capturing the HLA region, amplifying the exons of the HLA genes, and finally high-throughput sequencing to generate read data, which are then used for downstream HLA calling [133]. The advances of NGS [11] and long-read sequencing (PacBio [29, 30] and ONT [27, 28, 141]) have further enlarged the scope of HLA typing by generating genomic sequences and assembling to reference sequences rather than merely focusing on specific exons, which in turns rendering even higher (3-field and 4-field) resolution typing, higher accuracy, as well as the cost-effective nature of such technologies [142-149].

Nevertheless, the extra cost and complexity of such approaches have driven the development of alternative solutions. As a result, various HLA imputation algorithms have been developed, aiming to predict HLA genotypes as a side product of readily available data, such as RNA sequencing, whole-genome sequencing, whole-exome sequencing, or DNA microarrays [150-153]. The accuracy and reliability of these imputation algorithms, however, are highly dependent on the availability of large and

ethnically diverse HLA reference panels, which should contain SNP haplotypes tagged with high-resolution HLA allelic information. While many studies utilise population-specific reference panels that are often limited in size and accessibility, the first global HLA reference panel was assembled incorporating 2,693 individuals across 26 diverse populations from the 1000 Genomes Project [154-156]. More recently, an expanded multi-ethnic HLA reference panel encompassing 21,546 individuals spanning global populations [157] was constructed and became available through the Michigan Imputation Server [158]. Expanding the size and diversity of these reference datasets not only extends the potential of HLA inference across populations but also enables the detection of rare alleles, thereby broadening the applicability of HLA imputation and facilitating a deeper understanding for disease aetiology.

Early imputation algorithms focused on modelling the complicated LD pattern [159] in the HLA region with genotype data, including LDMhc [160] and HLA*IMP [161-163]. These methods leverage the relatedness between the querying SNP haplotypes of unknown HLA type to those observed in reference panel, with HLA*IMP polishes the SNP selection scheme of LDMhc and enhances numerical performance. HIBAG [164] employs attribute bagging to classify HLA types from randomly selected SNPs and bootstrapping to aggregate the classifying results. The algorithm was also implemented by a web server HLA-IMPUTER [165], which expanded the original Han Chinese reference panel to multiple populations. SNP2HLA [166] implements HLA imputation

in the BEAGLE [92] framework by embedding known HLA alleles and amino acids from the reference panel into the SNP haplotypes and subsequently performing imputation and phasing. This approach has also recently been refined by CookHLA [167] in two main aspects. In contrast to SNP2HLA, which assigns HLA markers primarily to the central region of each gene, CookHLA imputes each exon separately and then generates a consensus call to better capture polymorphism in the HLA region. Additionally, CookHLA employs a dynamic genetic map that is adaptively learned from the data, allowing for improved modelling of local recombination patterns. Benchmarking results demonstrated that CookHLA achieved superior imputation accuracy when compared to existing tools such as HIBAG and SNP2HLA, highlighting the effectiveness of this new approach.

Advances in next-generation sequencing have also spurred the development of different bioinformatic tools for HLA inference. Graph-based methods have emerged as a prominent early innovation. Tools such as MHC*PRG [168] and HLA*PRG [169] represent polymorphic HLA alleles using population reference graphs, enabling the efficient encoding of shared sequence regions resulting from common evolutionary origins, while simultaneously preserving nucleotide-level polymorphisms. Then, the methods implement an alignment tool to the graph with both non-paired long reads and paired-end short reads. At 30× coverage, the method was almost error-free, though at a sacrifice of computational performance, and was subsequently improved by HLA-LA

[170]. In contrast, Kourami [171] assembles reads to HLA haplotypes rather than infers the best matching allele in the database, enabling the discovery of novel HLA alleles. GRIMM [172] also employs a graph design in HLA imputation but specialises in the application of donor-recipient matching. Other software includes HLA-VBSeq [173, 174] that opts for a Bayesian approach to align reads to alleles, as well as DEEP*HLA [175] and HLARIMNT [176] that utilise convolutional neural network and transformer-based architectures, respectively, exemplifying the potential of deep learning approaches to accurately infer rare HLA alleles that is otherwise challenging for conventional methods.

Additionally, software designed for RNA-seq data and whole-exome sequencing data have also flourished in different applications. Methods employing RNA-seq data such as seq2HLA [177], HLAforest [178], and PHLAT [179] follow the alignment-calling scheme. OptiType [180] and xHLA [181] formulate the problem via integer linear programming, whereas the latter uses protein-level alignment of the typing exons. HLAminer [182, 183], ATHLATES [184], HLAscan [185], and HLA-HD [186] account for the intron-exon structure of HLA genes, whereas Polysolver [187] has been specifically optimised for tumour samples, addressing the unique challenge posed by alignment to canonical HLA references.

Despite these advancements, few HLA imputation algorithms have been developed

specifically for lcWGS. In Chapter 4, I examine the only HLA imputation method implemented in the QUILT software (QUILT-HLA) in this setting, which serves as the foundation for subsequent methodological improvements presented in this thesis.

1.5 Infectious diseases and Genome-Wide Association Study

Infectious diseases remain a leading cause of morbidity and mortality worldwide, particularly in low- and middle-income countries where access to healthcare and preventive measures may be limited [188]. While environmental and socio-economic factors play major roles in infection risk and disease progression, especially the greater mobility of populations as a result of globalisation, it is difficult to overstate the significant contribution of host genetic variation in shaping individual susceptibility to infectious agents that results in varying immune responses, pathogen clearance, and disease severity among individuals [189]. Table 1.5.1 summarises key host genetic variants shown to modulate the risk of severe malaria and HCV infection, pathogens that are relevant to this thesis:

Variant	Type	Protective Allele	Alternative Allele	Gene	Pathogen
rs334	SNP	A	T	HBB	Malaria
rs4951377	SNP	A	G	ATP2B4	Malaria
rs8176719	Indel	T	TC	ABO	Malaria
$-\alpha/--$	SV	Deletion	Normal	HBA1/HBA2	Malaria
Dantu	SV	Dantu	Normal	GYP	Malaria
rs368234815	Indel	TT	ΔG	IFNL4	HCV

Table 1.5.1. Key genetic variants associated with severe malaria and HCV infection. All SNPs and indels were reported according to the gnomAD convention, using the forward (reference) strand. *HBB*: Haemoglobin subunit beta; *ATP2B4*: Plasma membrane calcium-transporting ATPase 4; *ABO*: ABO blood group; -α/--: α-globin deletion haplotypes associated with α-thalassaemia; SV: Structural variant; *HBA*: Haemoglobin subunit alpha; Dantu: The Dantu blood group variant; *GYP*: Glycophorin; *IFNL4*: Interferon lambda 4.

The most established genetic factor influencing malaria susceptibility is the *HBB* SNP rs334, which results in a missense mutation producing the sickle-cell trait; this altered haemoglobin impairs parasite growth within erythrocytes and reduces rosette formation, thereby protecting against severe malaria disease [115, 190-192]. Other variants include rs8176719 in the ABO gene (O-blood group), which also lowers erythrocyte rosetting [193]; rs4951377 in *ATP2B4*, which alters red blood cell calcium transport [115]; the Dantu blood group antigen, which increases membrane tension and limits parasite invasion [194]; and α-thalassaemia deletion haplotypes, which generate microcytotic erythrocytes that restrict parasite growth and enhance clearance of infected cells [195]. In the context of HCV infection, the rs368234815 polymorphism in the interferon lambda 3 and 4 (IFNL3/IFNL4) region has been shown to influence both spontaneous and treatment-induced viral clearance [196-201]. The dinucleotide TT allele causes a frameshift and prevents the expression of *IFNL4*, which induces the production of the

more effective *IFNL3* molecules, promotes the Janus activated kinase-signal transducer and activator of transcription (JAK-STAT) pathway, and results in the expression of IFN-stimulated genes to resist pathogens [202-205]. Understanding these genetic determinants provides critical insights into host-pathogen interactions and elucidates the molecular mechanisms of disease aetiology.

Early efforts to study genetic contribution to disease outcomes relied on linkage and admixture mapping and candidate gene studies [36]. Linkage mapping relies on familial data to study the inheritance and segregation of traits and genetic variants [206], whereas admixture mapping identifies disease loci by correlating ancestry with traits in recently admixed populations [207]. While useful in certain contexts, both approaches have low resolution for complex traits and non-Mendelian diseases. The establishment of the Single Nucleotide Polymorphism Database (dbSNP) for cataloguing genetic variants [208, 209], the characterisation of LD patterns through the HapMap Project [48, 99, 100], and the advent of high-throughput genotyping technologies accompanied by reduced cost for generating genomic data have revolutionised the field of genetic epidemiology studies and enabled genome-wide association studies (GWAS) in the early 2000s [210]. GWAS is a statistical approach that scans the entire genome to identify genetic variants associated with quantitative or qualitative traits and has been initially applied to age-related macular degeneration, where a common intronic variant in the complement factor H gene was identified as the risk factor [211]. Since then,

GWAS methodologies and software have been extensively developed and applied across numerous genetic studies [36, 212-214]. Due to its agnostic design that does not rely on prior biological assumptions and allows for the discovery of novel loci, especially those outside classical immune-related pathways, and its capability of capturing the polygenic contributions to phenotypes, GWAS becomes particularly valuable at inferring genetic variants that influence complex diseases and traits resulting from multiple genetic and environmental factors, although functional replicates are usually required to confirm the causality [215-218].

GWAS has long been conducted with microarrays plus imputation to capture genome-wide variation, as the prohibitive cost of performing deep whole-genome sequencing limits its scalability. As introduced in Section 1.2.3, lcWGS with imputation as an alternative of microarrays has been established and illustrated to exhibit at least comparable accuracy in capturing genome-wide variants [76], and thereby successfully applied to real GWAS across a range of species with various traits: pigs [219-224], goats [225], dogs [226, 227], rabbits [228], chickens [229-231], mice [232], fish [233-236], molluscs [237], drosophila [238], spinach [239], peanuts [240], and human [72, 241-260]. With the continued decline in sequencing and library preparation costs, together with advances in constructing large and dense reference panels, refining imputation algorithms, and assembling bioinformatics pipelines, GWAS conducted with lcWGS continues to play an essential role in elucidating the genetic architecture

of complex traits and infectious diseases across diverse populations and species.

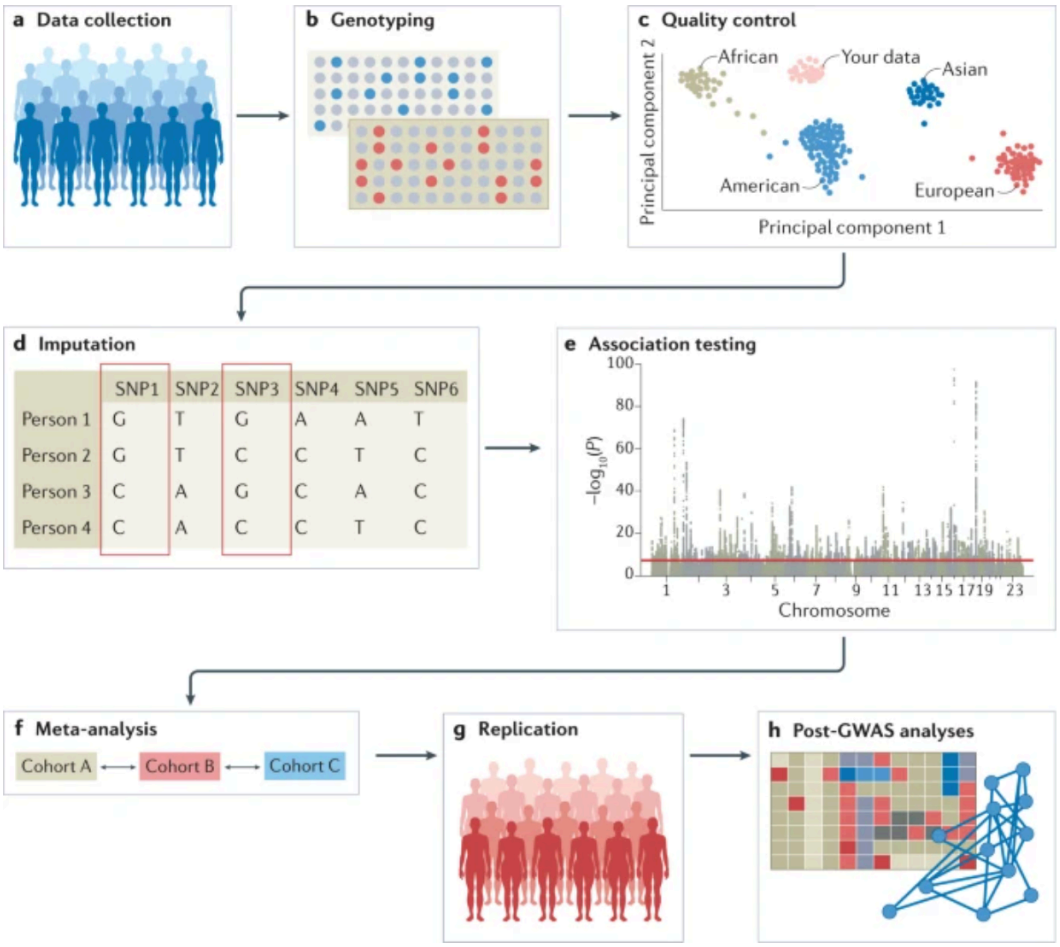


Figure 1.5.1. A general workflow of GWAS. This figure was adapted from Figure 1 in [214].

1.6 Conclusion

This chapter reviews the concepts, technologies, and methodologies that are essential for the establishment of this thesis about evaluating the feasibility of lcWGS as a scalable and unbiased alternative to traditional genotyping platforms in genetic epidemiology studies in the context of infectious diseases across global populations. The rest of this thesis presents our investigations in lcWGS and imputation from both

the technical and biological perspective. Chapter 2 assesses the impact of different library preparation conditions on the resulting lcWGS data. Chapters 3, 4, and 5 illustrate the great potential of lcWGS from the perspective of capturing genome-wide variants, complex HLA alleles, and large SVs by optimising the imputation workflow or developing specific methodology that is suitable in this setting. These findings are then applied to our own GWAS dataset in Chapter 6, which uses lcWGS to investigate genetic factors influencing HCV infection in a Vietnamese cohort. Together, these chapters show the versatility and promising future of lcWGS as a cost-effective tool for genomic discovery, with the potential to advance understanding of the genetic underpinnings of infectious disease outcomes.

Chapter 2 Evaluation of Low-Coverage Whole-Genome Sequencing Library Preparation Conditions and Sequencing Platforms

LcWGS is an emerging approach for capturing genomic variation, potentially offering notable advantages in both accuracy and cost-effectiveness [63]. However, many previous studies relied on *in silico* down-sampling of high-coverage data, which did not capture the real-world challenges posed by sequencing chemistry (for example, in [78]). Unlike microarray technologies, lcWGS lacks standardised protocols, prompting us to generate empirical sequencing data from real populations affected by infections to better evaluate its performance.

In this chapter, I explored various practical and theoretical aspects of lcWGS in terms of library preparation and sequencing platforms. Section 2.1 focuses on 66 NA12878 samples, where I assessed the impact of varying input DNA mass and library preparation kits on data quality. In Section 2.2, I analysed a separate pilot study designed to stress test the limits of input DNA mass, identifying conditions that lead to sample failure. Although these studies were not perfectly controlled with multiple replicates for each condition and confounded by experimental issues, the findings provided critical guidance for generating lcWGS data by allowing us to select a feasible library preparation method as well as to investigate fragment lengths as a key

determinant of read error rates. Additionally, I simulated lcWGS data across different sequencing platforms that were implemented at Oxford to explore other possibilities in Section 2.3, which provides practical insights to researchers willing to conduct their studies in-house.

2.1 Assessments of different library preparation conditions using 66 NA12878 samples

2.1.1 Library preparation conditions

Sequencing library preparation methods can have a large impact on the quality of the data produced. The aim of this pilot study was to determine the optimal experimental conditions for lcWGS by assessing the data quality of 66 NA12878 samples, an extensively sequenced and studied cell line with a known genomic assembly [99, 261-264], generated in different library preparation conditions as presented in Table 2.1.1. Notably, this study was not a perfectly controlled experiment with multiple replicates for each combination of conditions, but was essentially exploratory. The prepared samples were loaded on a NovaSeq 6000 platform for 151 bp paired-end sequencing.

Library preparation kits	● UIIFS: NEBNext® Ultra™ II FS DNA Library Prep Kit for Illumina® [265, 266]. ● DNAP: Illumina DNA Prep [267]. ● NEB-modified: a modified version of NEBNext® Ultra™ DNA Library Prep Kit for Illumina®, aiming for longer inserts (~400 base pairs, bps) compared to the original (~200 bps) [268].
Input mass (nanogram, ng)	The amount of original DNA material used in the library preparation: 10, 50, 100, 500.

Table 2.1.1. Library preparation conditions. Aiming for longer fragments length in the NEB-modified kit is essential, as short fragments result in undesired overlapping bases from read 1/2 and thus lead to bias by an uneven coverage of the genome.

2.1.2 Quality metrics for assessing low-coverage whole-genome sequencing library preparation performance

Sequencing data quality can be affected by several factors, including uneven sample balance in pooled libraries, duplication introduced during amplification, coverage bias across the genome, and base-calling errors associated with sequencing reads and platforms. I used five quality metrics to evaluate these factors in our datasets (see Methods for details):

- Total yield: defined as the total number of reads generated for each sample in a pool. This measure is used to assess whether samples are effectively balanced in the pooling process.
- Duplication rate: defined as the proportion of reads identified as duplicates. In particular, this reflects possible fragment duplication during DNA amplification.
- Sequencing skew: the proportion to which genomic regions remain not covered by reads, considering only regions accessible by short-read sequencing [49, 269]. It quantifies amplification or coverage bias that lead to missing information.
- Effective coverage: defined as average depth of genome coverage after duplicate reads are removed. It provides a measure of the non-redundant sequencing information.

- (Read 1/ Read 2) k-mer error rate: defined as the proportion of erroneous k-mers in each read direction. It captures systematic base-calling error dependent on read length, where read 2 is typically less accurate than read 1 [270], potentially impacting imputation performance.

2.1.3 Fragment length is the primary determinant of read 2 error rates

It has previously been observed that longer fragment sizes result in higher read 2 error rates [270], which may subsequently affect imputation performance. Since different library preparation kits target fragments of varying lengths, I investigated if similar behaviour was observed in our dataset (see Methods for details).

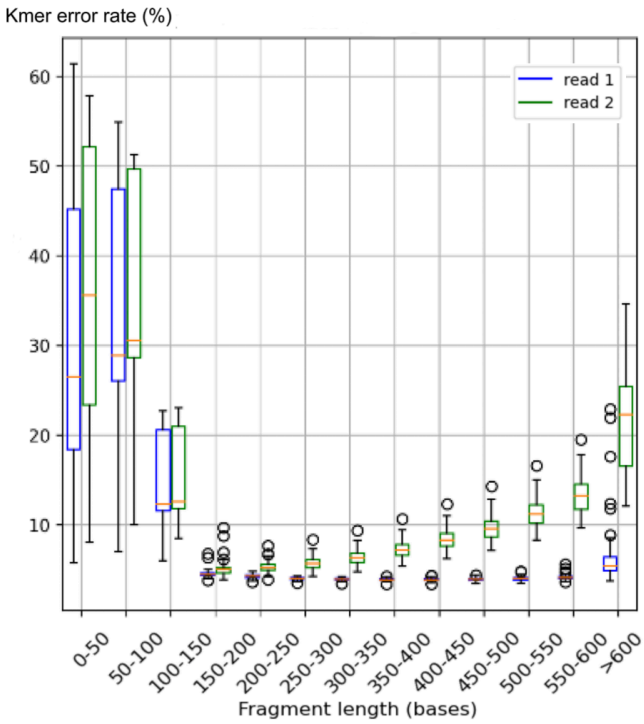


Figure 2.1.1. Read error rate distribution stratified by fragment lengths. All fragments of all samples are separated into 13 bins according to their lengths, from less than 50 bases to greater than 600 bps. Boxes on the left/right (blue/green) for each pair display read 1/read 2 error rates.

In line with the previous findings, read 2 error rates were generally higher compared to read 1 due to the error introduced in the bridge amplification step at the paired-end turnaround stage before read 2 was sequenced [271]. The increasing error rate associated with longer fragments reflects fluorescent signal crosstalk between adjacent reads due to the larger physical space occupied during imaging [272]. Additionally, the significant higher error rate for fragments shorter than 150 bases (which is shorter than 151 bp paired-end sequencing we performed) reflect adaptor contamination [273]. Across this dataset, different library preparation kits produced varying average fragment lengths: longer fragments were associated with higher sequencing errors, whereas shorter fragments resulted in increased read overlap. The error rate in read 1 displayed less variability and did not exhibit a systematic correlation with fragment length. As a result, subsequent analyses primarily focused on the error rate observed in read 1.

2.1.4 Selection of a high-performing library preparation combination

I used read 1 k-mer error rate, sequencing skew, and duplication rate to evaluate sample performance.

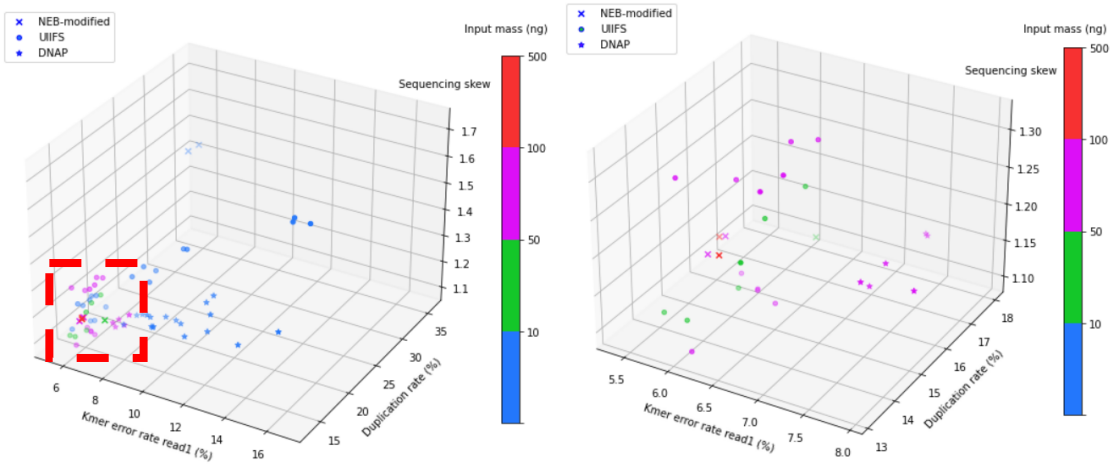


Figure 2.1.2. Sequencing performance of 66 NA12878 samples. Read 1 k-mer error rate (x-axis), duplication rate (y-axis), and sequencing skew (z-axis) are plotted for each sample, coloured by input mass and marked by library preparation kit. The left panel includes all samples, whereas the right is a zoom-in plot after removing all 10 ng samples.

Favourable samples should have low per-base error rate, less duplicative reads, and even representation of the entire genome, which corresponds to the encircled bottom left corner of the left panel. Samples prepared with 10 ng input mass (coloured blue) showed high variability and inconsistent sequencing results due to the potentially compromised library complexity introduced by these limited DNA molecules, which could hardly be remedied by PCR [274]. By excluding these samples in the right panel, we did not observe any kit or input mass that consistently produced superior data quality. No statistical tests were performed due to the exploratory nature of this pilot study. Based on these observations, we selected the combination that generated the best-performing sample, UIIFS kit with 100 ng input mass, to process the real samples in Chapters 3, 4, and 5.

2.2 Impact of input materials on low-coverage whole-genome sequencing data quality with 91 Vietnamese samples from a pilot SEARCH cohort

2.2.1 Library preparation conditions

The objective of this study was to evaluate lcWGS in a non-European population from the SEARCH project, which originally aimed to assess the effectiveness of HCV treatment in a Vietnamese cohort [275]. In this chapter, we obtained samples from this cohort to investigate the impact of different input DNA masses on lcWGS library preparation outcomes. This dataset contained 91 samples and was targeted at ~1× coverage. All the samples were prepared with the NEB-modified library preparation kit, which was the only cost-effective option available in the Oxford Genomic Centre when these experiments were conducted (chronologically, these samples were prepared earlier than those in Section 2.1). Though the input mass for all samples were originally targeted 300 ng, some resulted in low amount of input DNA (< 50 ng) due to lab failure. These samples became valuable in stress-testing the lower limit of feasible input mass as the input of lcWGS, so we performed 14 PCR cycles rather than 10 PCR cycles as in the normal counterparts.

2.2.2 Impact of input DNA quantity on sample preparation outcome

To examine the impact of low input mass on sample performance, I first investigated

the effective coverage.

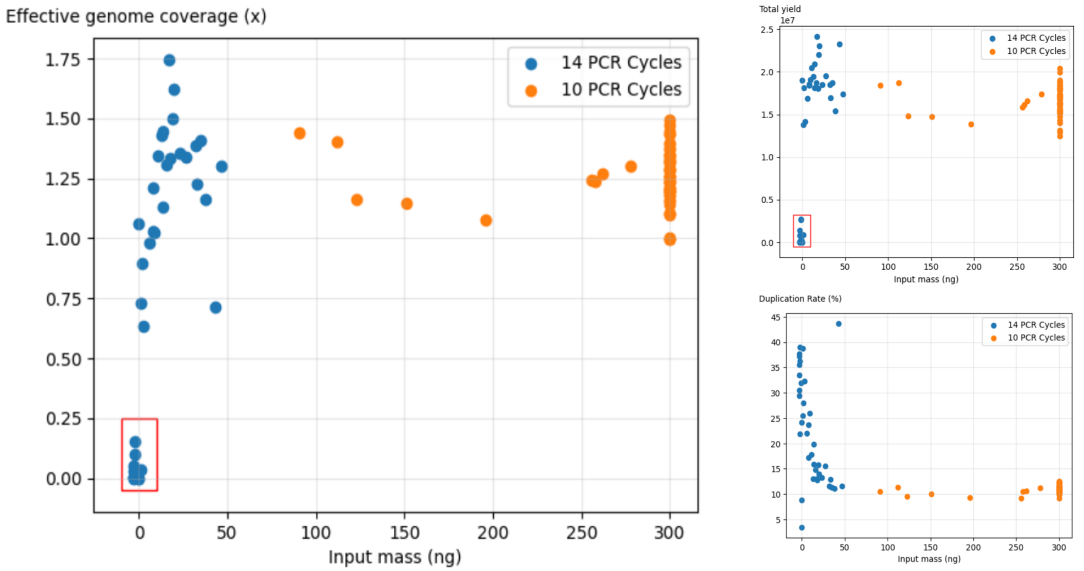


Figure 2.2.1. Sample performance of 91 Vietnamese samples. The left, top right, and bottom right panel show effective genome coverage, total yield, and duplication rate on y-axis, respectively. The blue points represent samples subjected to 14 PCR cycles (which had input mass < 50 ng), while the orange points represent samples that underwent 10 PCR cycles.

The bottom left cluster in the left panel includes 13 samples that failed to generate meaningful coverage for downstream analysis, indicating that a minimal input mass of 10 ng is necessary. After removing these outliers, I conducted a t-test to determine whether the effective genome coverage differed between the blue and orange points. The resulting t-statistic was -0.79 and p value of 0.432. Therefore, no statistically significant difference in effective coverage was observed between samples with high and low input mass. Samples with less than 50 ng of input DNA underwent four additional PCR cycles, which increased yield but also led to a higher duplication rate, as shown in the top-right and bottom-right panels, respectively. These results suggest

that additional PCR cycles should be avoided for samples with sufficient input DNA. For samples with less than 50 ng of input DNA, performing one or two extra PCR cycles may be preferable. In conclusion, extremely low amounts of input DNA led to sample failure, indicating that at least 10 ng is required for reasonable data. Apart from these low-input samples, lcWGS was successfully performed on all other samples.

2.3 Impact of different sequencing platforms on low-coverage whole-genome sequencing imputation: a simulation study

In the last section of this chapter, I investigated the impact of different sequencing platforms on generating lcWGS data, particularly from the perspective of different read lengths and error rates. For the Gambian cohort and Vietnamese cohort introduced in Chapters 3 and 6, the Illumina NovaSeq 6000 machine and the Element Biosciences AVITI model were employed for sequencing, respectively [70, 276]. Apart from the short-read platforms, there were also interests in applying low-coverage PacBio HiFi long-read sequencing, given its longer read length that aids in phasing especially in regions with repetitive units or hard to align [30]. Another practical reason for investigating these sequencing platforms was their local availability, as they are housed in the Centre for Human Genetics at the University of Oxford.

To assess whether AVITI or PacBio offers additional advantages for lcWGS, I simulated

data from both the paternal and maternal haplotypes (0.6× each) for the HG02886 sample from the 1000 Genomes Project (Mandinka), whose full genome was previously assembled [277]. The simulation experiments of short-read sequencing tested four scenarios, summarised in the Table 2.3.1.

	Read length	Read 1 error rate (%)	Read 2 error rate (%)	Insert size mean	Insert size standard deviation
NovaSeq-optimal	151	0	0	500	13
NovaSeq-real	151	0.24-0.71	0.34-1.05	329	50
AVITI-optimal	300	0	0	1000	200
AVITI-real	300	0.01	0.01	830	200

Table 2.3.1. Parameters involved in short-read simulation.

The optimal simulations were error-free and contained no overlapping bases, achieving maximum genome coverage. The NovaSeq-real simulation replicated the 151 bp samples derived from the 210 GAMCC lcWGS cohort presented in Chapter 3. In contrast, the AVITI-real 300 bp read-length simulation used parameters estimated from advertised specifications, as the AVITI machine was not yet set up when this simulation was performed. Full details of the simulation procedure are provided in the Methods section. Imputation results were then compared with the truth defined in the 1000 Genomes Project.

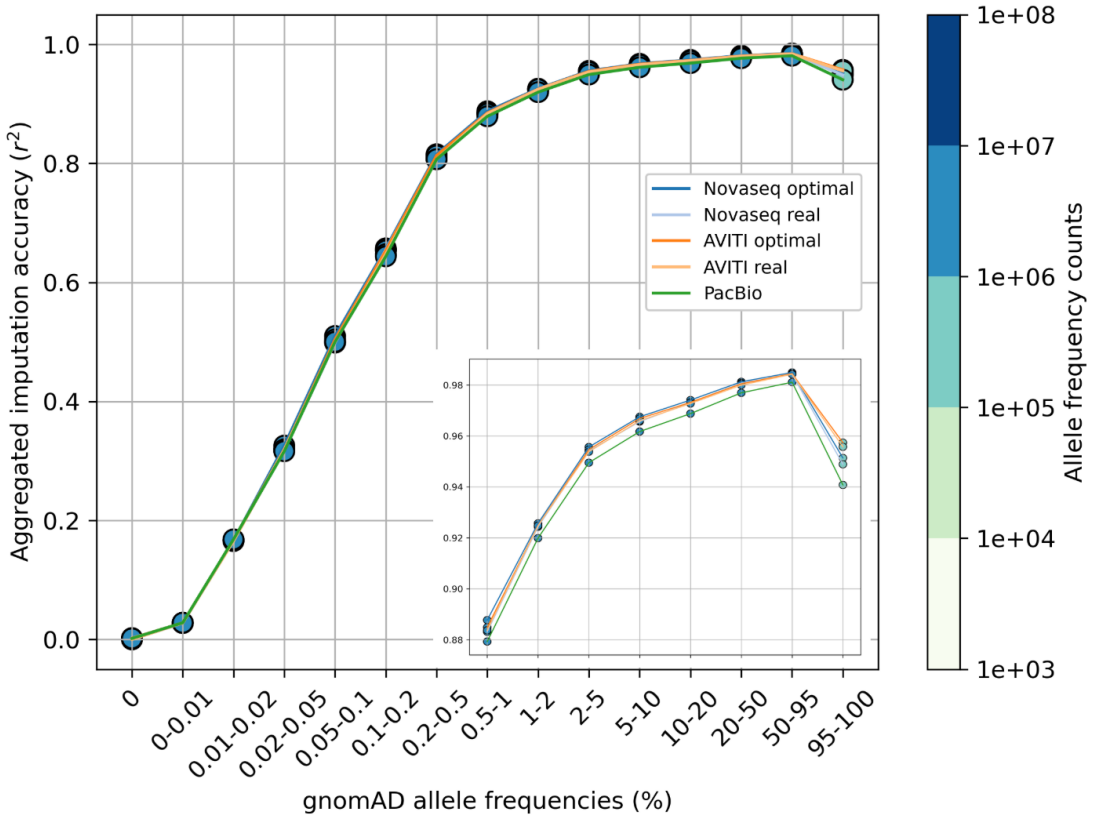


Figure 2.3.1. Imputation accuracy of simulated lcWGS data across different platforms. Mean accuracy (y-axis, computed as r^2 between imputed and directly-typed SNPs and averaged across individuals) for variants stratified by different gnomAD African allele frequencies (x-axis). Details of the imputation metrics are presented in Chapter 3. Each marker is coloured according to the number of variants involved in the calculation, and lines are labelled by different simulations. The PacBio simulation in this plot has a mean read-length of 500 bases. The bottom right subplot provides a zoom-in view of common variants, with the x-axis aligned to the main panel.

Figure 2.3.1 shows nearly identical results across different runs, although accuracy was slightly lower for the PacBio simulation (r^2 ranged from 0.954 to 0.956 for 2%-5% variants in short-read sequencing, compared with 0.950 for PacBio sequencing). This suggests that lcWGS imputation did not benefit much from the reduced base-calling error rate of the AVITI platform, and that the NovaSeq machine provided satisfactory accuracy for this task. To further explore the impact of read length on low-coverage long-read sequencing, I increased the mean read length in the simulation to 20 kb while

keeping all other parameters unchanged.

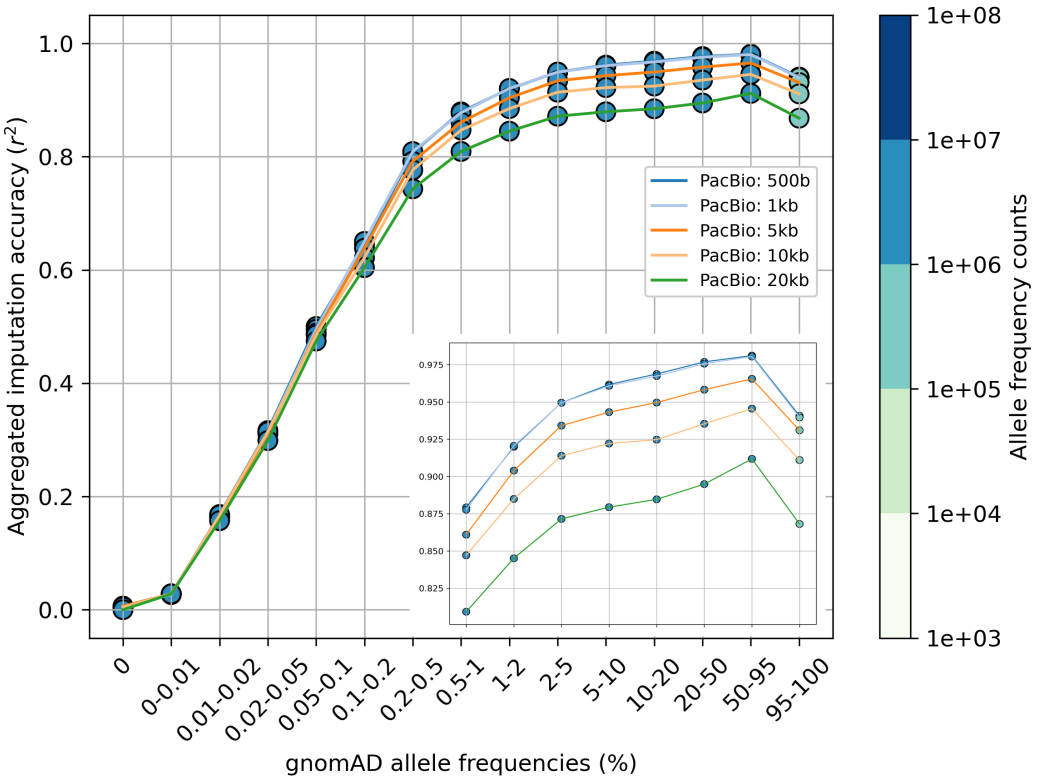


Figure 2.3.2. Imputation accuracy of simulated low-coverage PacBio data at different read lengths.

Except for 500 and 1,000 bases read-length, accuracy decreases substantially as read length increases (r^2 = 0.950, 0.949, 0.934, 0.914, and 0.872 for 2-5% variants, respectively). One reason is that, to keep computation tractable, QUILT's hidden Markov model assumes that each read informs the haplotype state only at its central SNP, effectively collapsing all other information into a single point. This approximation works well for short reads, which span only a few SNPs. For long reads, however, the assumption becomes less accurate because the read may traverse regions with weaker LD or switch between haplotypes (Supplementary information Section 1.1.2 in [66]).

On the other hand, long reads yield fewer reads at the same sequencing depth, reducing independent genomic anchors. They also increase the distance before observing the alternate haplotype, limiting local LD information. In the extreme, a single ultra-long read represents only one haplotype, leaving the other unobserved and dependent solely on the reference panel. These factors make the QUILT approximation better suited to short-read data and may explain the lower accuracy of low-coverage long-read lcWGS. Combined with higher cost, we did not pursue further PacBio sequencing.

2.4 Methods

lcWGS data processing for NA12878 samples. This pilot comprised 66 NA12878 samples, an extensively sequenced and studied cell line of known genomic assembly from European ancestry [99, 261-264]. Libraries were prepared under conditions mentioned in Section 2.1.1 and sequenced on a NovaSeq 6000 platform for 151 bp paired-end sequencing. The computational workflows were integrated into the main Snakemake pipeline described in Chapter 3 [278], which is publicly available on my GitHub page (https://github.com/Suuuuuuuus/LCWGS_pipeline).

Sequences were aligned to the human reference genome GRCh38 (accession GCA_000001405.15, released on Dec 2013) in fasta format using bwa mem v0.7.18 with six threads and default parameters [279]. Duplicative reads were removed using

picard v2.27.4 MarkDuplicates to minimise sequencing bias, yielding a mean coverage of 4.24× (range 1.25×-7.42×) across all samples, restricted to autosomes [280].

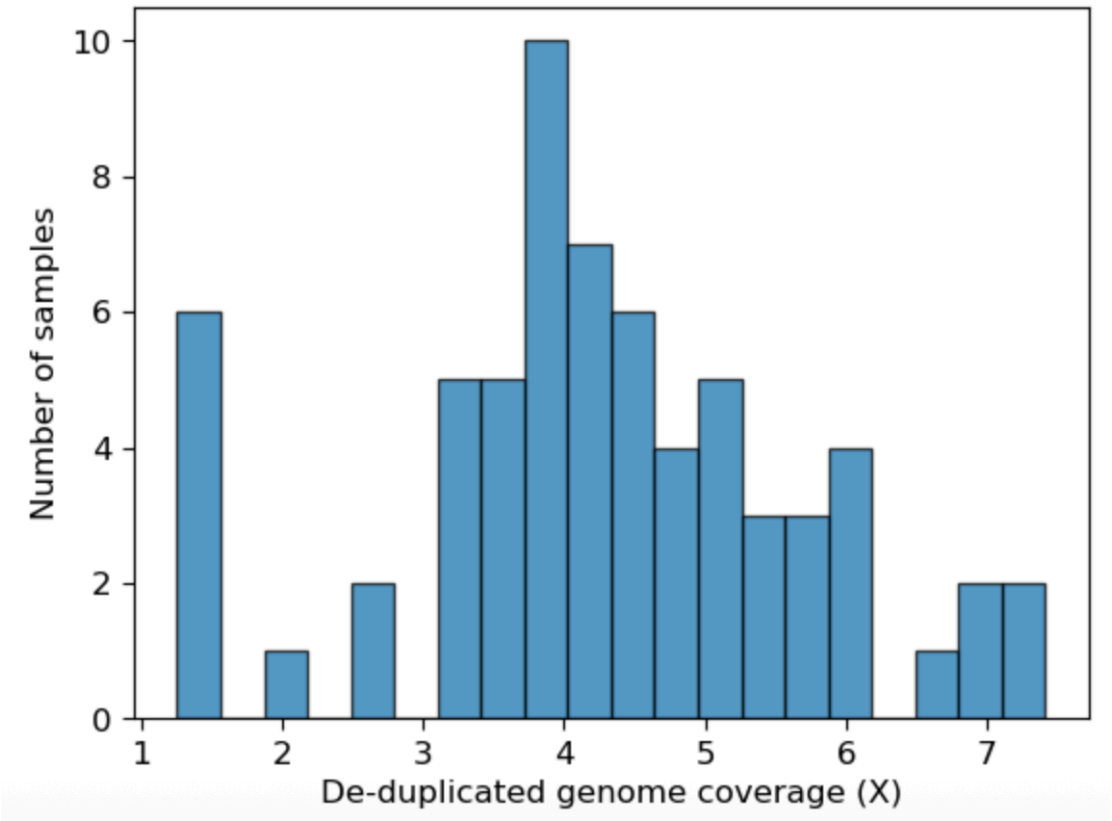


Figure 2.4.1. Genome coverage of 66 NA12878 samples. Genome coverage ranges from 1.25× to 7.42× (mean 4.24×).

Since genome coverage was uneven for this batch of samples, I down-sampled all samples to ~1× coverage with seqtk v1.4 [281] to 10,596,026 paired-end reads per sample, the number equivalent to 3.2 billion bases in the complete human genome with 151 bp paired-end sequencing.

Quality metrics and imputation accuracy calculations. The mathematical formulation of the five quality metrics are presented in this section:

- Total yield: metric of the total number of reads produced by the sequencer. I counted the number of lines in the fastq file with wc -l and divided the number by four.
- Duplication rate: a measure of duplicative reads, reported using samtools v1.16.1 markdup and flagstat [282-284].
- Sequencing skew: a measurement of sequencing bias that resulted in uneven genome coverage to quantify distortion introduced by DNA amplification, which can result in uneven read distribution across genomic regions and potentially obscure important genetic information. In the absence of sequencing bias, genome coverage is expected to follow a Poisson distribution, with the shape parameter corresponding to the targeted coverage. To quantify deviations from this expectation, the proportion of the genome not covered was calculated, defined as the fraction of bases with zero coverage across all autosomal regions. This analysis was restricted to regions accessible by short-read sequencing, as defined by the 1000 Genomes Project, to avoid alignment artefacts (Section 9.2 of the supplementary, ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/accessible_genome_masks/20141020.pilot_mask.whole_genome.bed) [49]. The observed proportion was then normalised by the expected value under the Poisson assumption, yielding a metric referred to as sequencing skew. A skew

value of 1 indicates perfectly uniform genome coverage, while values greater than 1 indicate that some regions receive higher coverage than others.

- Effective coverage: de-duplicated genome coverage. Effective genome coverage was calculated as the total coverage obtained from samtools coverage, multiplied by one minus the duplication rate, restricted to accessible regions.
- (Read 1/ Read 2) k-mer error rate: a measurement of sequencing error. This metric was calculated as the proportion of erroneous k-mers, defined as k-mers containing at least one nucleotide mismatch compared with the true k-mers derived from the reference genome, among all k-mers in all fragments. A k-mer length of 31 was used, which is the longest supported by jellyfish v2.3.1 (the term *k-mer* here was used interchangeably to refer either to a genuine k-mer of any length or specifically to a 31-mer) [285]. I calculated the k-mer error rates for both read 1 and read 2 with the classify-kmer utility from package Iorek (commit 2d46e6d61037b8185f1f91519318476a67d0cf71) [286]. As this calculation required knowledge of the true k-mers, it was performed only for the NA12878 samples [261].

lcWGS data processing for SEARCH samples. These samples underwent the same bioinformatics pipeline as the NA12878 samples.

Simulation of short-read and long-read data for HG02886. I used dwgsim v0.1.14

[287] with parameters described in Section 2.3 to simulate both parental haplotypes to 0.6× and set -y equals 0 to avoid contamination reads. I then merged the simulation for both parental haplotypes and aligned them to GRCh38 reference with bwa mem [279]. On the other hand, I simulated PacBio long-read circular consensus sequencing (CCS) data [32] for the same sample. Simulated reads were generated using pbsim v1.0.3 [288] with --data-type set to CCS, --model_qc as provided in the GitHub repository, --length-mean set to 500, 1,000, 5,000, 10,000, or 20,000, and --length-max set to 50,000. The simulated reads were then aligned using minimap2 v2.28-r1209, setting -x as map-hifi [58]. The results were imputed with QUILT using 1000 Genomes Project reference panel excluding related individuals (HG02884, HG02885 and HG02886).

2.5 Conclusions

This chapter explored two pilot cohorts that underwent lcWGS library preparation and sequencing, demonstrating the feasibility of this technology in real-world settings. Section 2.1 examined the impact of different library preparation conditions on final data quality, assessed by genome-wide representation (duplication rate and sequencing skew) and per-base accuracy (k-mer error rate). While samples with 10 ng input DNA showed inconsistent results across runs, no library preparation kit or input amount consistently produced superior data. Based on these findings, the best-performing condition was selected for processing real samples in Chapters 3, 4, and 5. We also observed challenges associated with Illumina sequencing chemistry, where read 2 exhibited

lower per-base quality and longer fragments are associated with declining accuracy. Section 2.2 evaluated the lower limit of input DNA for feasible lcWGS libraries, confirming that inputs below 10 ng lead to substantial sample failure and variable genome capture. Finally, Section 2.3 reported *in silico* experiments comparing three sequencing platforms for lcWGS. Low-coverage long-read sequencing showed inferior accuracy and is therefore discouraged. This analysis offers practical insights relevant to the genetics community. Further discussion of the limitations and implications of this chapter is provided in Section 7.1.

Chapter 3 Genome-Wide Imputation Performance of the Low-Coverage Whole-Genome Sequencing Method

Chapters 3, 4, and 5 transition the focus from data generation to variant calling and imputation. Collectively, these chapters aim to demonstrate the considerable potential of lcWGS to capture genomic variants, particularly in genetically diverse populations, and to benchmark its performance against microarrays. Chapter 3 is devoted to assessing genome-wide variants derived from lcWGS. Section 3.1 introduces the Gambian cohort from West Africa, which also forms the basis for analyses in Chapter 4 and Chapter 5. Section 3.2 presents a comprehensive comparison of imputation performance between lcWGS and microarrays. Section 3.3 examines the use of lcWGS for population structure inference. Moving beyond genome-wide analyses, subsequent sections evaluate specific regions and variants of particular interest both to this cohort and to the broader aim of applying lcWGS in GWAS. Section 3.4 characterises imputation accuracy for subsets of variants of particular biological interest and highlights regions of reduced accuracy across the genome, while Section 3.5 focuses on blood group variants and replicates a well-established marker relevant to this population.

3.1 The Gambian Malaria Cases and Controls cohort

3.1.1 Data collection and generation

To investigate imputation performance on lcWGS data in a Gambian population, we focused on a set of 210 DNA samples originally collected from children enrolled in an epidemiological study of clinical malaria [289]. Samples were ascertained either as severe or mild malaria cases on admission to the Royal Victoria Teaching Hospital, Fajara, The Gambia in 1988-1990 and donated a venous blood sample. A control set was also obtained from non-malaria-infected children in the same hospital. For the purposes of this study, we selected 70 samples from each of the severe, mild, and control sets for genotyping and sequencing. All samples were normalised to standard concentration and were aliquoted for three purposes. Whole-genome genotyping was carried out by Eurofins using the Affymetrix Axiom™ Precision Medicine Research Array. Although a more direct comparison could have involved the Human Heredity and Health in Africa (H3Africa) Consortium Array (> 2 million markers), access to this array remains limited and its content may not optimally represent genetic variation within Gambian populations [290-292]. Accordingly, we proceeded with the Affymetrix platform due to its practicality and comparable cost to lcWGS. Only 190 samples were genotyped in this way, of which four failed post-genotyping quality control (QC) and were excluded from further analyses. LcWGS for all 210 samples was carried out using the Illumina NovaSeq 6000 platform, targeting ~1× host genome

coverage with equal pooling of samples, by Oxford Genomic Centre at the University of Oxford, using 100 ng of input material per sample. Eight samples also underwent deep whole-genome sequencing on the same platform where ~30× was targeted as gold standard genomes for further comparisons. Finally, we obtained HLA genotyping for all samples through HistoGenetics (Ossining, New York, United States of America).

Source of data	Number of samples	Technologies or Services
LcWGS	210	Illumina NovaSeq 6000
DNA microarray	186	Affymetrix Axiom™ Precision Medicine Research Array
Deep whole-genome sequencing	8	Illumina NovaSeq 6000
HLA typing	210	HistoGenetics

Table 3.1.1. **Data generated for the GAMCC cohort.**

3.1.2 Sample performance

The quality metrics introduced in Chapter 2 were used to assess the performance of lcWGS. Figure 3.1.1a characterises lcWGS data quality with sequencing skew, duplication rate, and effective genome coverage. Figure 3.1.1b visualises genome coverage separately as a histogram. After removing all duplicative reads, satisfactory average genome coverage was achieved at average 1.21×, ranging from 0.71× to 1.61× calculated at accessible regions to short-read sequencing. Sequencing skew (reflecting unevenness of genome coverage as introduced in the previous chapter) for all samples ranged from 1.15 to 1.48 (mean 1.30), although most lay in the window of 1.20 to 1.35. Mean duplication rate was 8.5%, spanning 7.03% to 10.33%. Notably, samples with higher sequencing depth tend to exhibit a higher skew metric, reflecting the

amplification-induced bias (GC-bias) rather than experimental failure [49, 293-295].

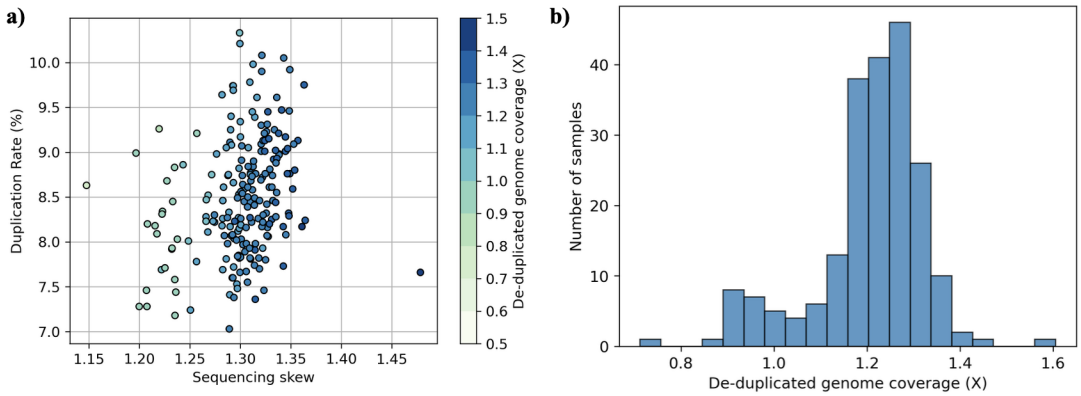


Figure 3.1.1. Sample summary statistics for 210 GAMCC lcWGS samples. a) Duplication rate (y-axis, proportion of duplicative reads calculated by samtools flagstat) against sequencing skew (x-axis, proportion to which genomic regions accessible to short-read sequencing are not covered by any read divided by the expected value assuming even coverage), coloured by de-duplicated genome coverage. **b)** Histogram of effective genome coverage (range 0.71×-1.61×, mean 1.21×).

An additional advantage of sequencing data is the ability to infer biological sex directly from coverage patterns on the sex chromosomes, which is particularly relevant in the context of GWAS. As shown in Figure 3.1.2, genomic sex can be inferred from chromosome coverage: the left cluster corresponds to males, who exhibit comparable coverage on both the X and Y chromosomes, whereas the right cluster corresponds to females, who show minimal coverage on the Y chromosome. Five samples were identified in which the self-reported sex did not align with the genomically inferred sex. The origin of these discrepancies remained unclear in this dataset. An attempt to cross-check historical HLA-B53 typing yielded no additional insights. Since the primary objective of this cohort was to evaluate lcWGS imputation performance rather than to

conduct association analyses, no further efforts were made to resolve these inconsistencies.

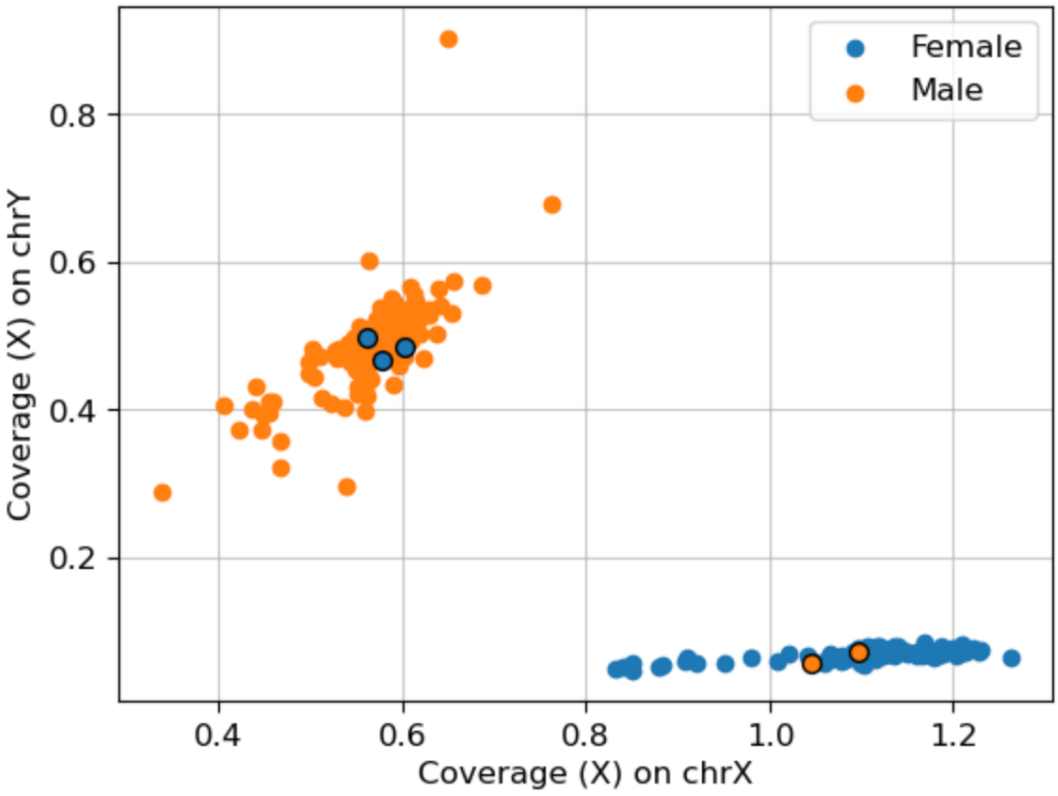


Figure 3.1.2. Genome coverage on sex chromosomes. Markers are coloured by self-reported sex, with five samples showing discrepancies compared with the sex inferred from genome coverage.

3.1.3 A bioinformatics pipeline for low-coverage whole-genome sequencing imputation

One of the primary objectives of this study was to develop a robust bioinformatics pipeline for processing lcWGS data, which was implemented using Snakemake [278]. In this section, an overview of the main steps involved in data processing and imputation is provided, with full methodological details presented in the Methods chapter. Although the pipeline is primarily designed for lcWGS data, it also includes

scripts for processing microarray data, deep whole-genome sequencing data, and HLA typing data. All scripts are available on my GitHub repository (https://github.com/Suuuuuuuus/LCWGS_pipeline). In addition, several components of the pipeline rely on a custom Python package (https://github.com/Suuuuuuuus/lcwgsus) that I developed in parallel with the pipeline.

For lcWGS, raw reads were aligned to the GRCh38 reference genome following preprocessing steps, including adapter trimming and duplicate removal. Genome-wide imputation was subsequently performed with QUILT [66], using both the 1000 Genomes Project 30× reference panel [50] and an enhanced 1000 Genomes Project panel incorporating additional African haplotypes from the Malaria Genomic Epidemiology Network (MalariaGEN) [115]. To ensure compatibility, the MalariaGEN-enhanced 1000 Genomes Project reference panel was lifted over from GRCh37 to GRCh38, filtered to retain only bi-allelic SNPs, and converted into required formats. For computational efficiency, autosomes were divided into 5 Mb windows to parallelise the imputation stage. The resulting imputed lcWGS genomes were further filtered and uploaded to the TOPMed server, a design that enhanced imputation accuracy by leveraging the large, global TOPMed reference panel.

For microarray data, raw calls from the Affymetrix Axiom™ Precision Medicine Research Array were subjected to quality control filters, including the removal of non-

autosomal variants, sites with high missingness or questionable allele frequencies, and variants deviating from the Hardy-Weinberg equilibrium threshold. The resulting high-confidence dataset was subsequently uploaded to imputation servers using both the 1000 Genomes Project and TOPMed reference panels.

For high-coverage data, eight replicate genomes underwent the same preprocessing and alignment steps described above. The genomes were divided into 5 Mb windows to parallelise computation and processed using the GATK workflow, including base quality recalibration, variant discovery with HaplotypeCaller, and variant quality score recalibration.

Together, these pipelines produced seamless and reproducible datasets across sequencing depths, enabling subsequent analyses of imputation performance, population structure inference, and association testing. This pipeline has also been made publicly available in the hope that it will support future research and facilitate the broader application of lcWGS.

3.2 Assessment of genome-wide low-coverage whole-genome sequencing imputation accuracy

The datasets assembled above enabled a series of comparisons aimed at evaluating the performance of the lcWGS approach. In the following sections, strategies to improve

lcWGS imputation accuracy were investigated, and performance was assessed against high-confidence reference data derived from microarray and deep whole-genome sequencing, which served as a practical benchmark. A summary of these comparisons is provided in Table 3.2.1.

Section 3.2.1 presents an initial comparison between lcWGS, subjected to QUILT imputation using the 1000 Genomes Project reference panel, and typed variants on the microarray. Section 3.2.2 explores improvements achieved by refining imputation with the MalariaGEN-enriched 1000 Genomes Project reference panel [115]. Finally, Section 3.2.3 demonstrates the optimal lcWGS workflow by leveraging the large and diverse TOPMed reference panel through the TOPMed imputation server. These results were benchmarked against microarray-based imputation, with all analyses evaluated relative to high-confidence deep whole-genome sequencing data as a reference standard.

<table>
<tr><th>Section</th><th>Source of data</th><th>Reference panel</th><th>Truth data</th></tr>
<tr><td>3.2.1</td><td>lcWGS</td><td>1000 Genomes Project</td><td>Microarray</td></tr>
<tr><td>3.2.2</td><td>lcWGS</td><td>1000 Genomes Project + MalariaGEN</td><td>Microarray</td></tr>
<tr><td rowspan="5">3.2.3</td><td rowspan="3">lcWGS</td><td>1000 Genomes Project</td><td rowspan="5">Deep whole-genome sequencing</td></tr>
<tr><td>1) 1000 Genomes Project
2) TOPMed*</td></tr>
<tr><td>1) 1000 Genomes Project + MalariaGEN
2) TOPMed*</td></tr>
<tr><td rowspan="2">Microarray</td><td>1000 Genomes Project *</td></tr>
<tr><td>TOPMed*</td></tr>
</table>

Table 3.2.1. Comparisons of lcWGS genome-wide imputation accuracy. *: via imputation server, otherwise by QUILT.

3.2.1 Evaluation of low-coverage whole-genome sequencing imputation at microarray typed variants

To investigate the imputation performance of lcWGS, I initially compared imputed genotypes to direct genotyping in the 186 samples with available microarray data. Specifically, I focused on n = 708,100 SNPs that passed microarray QC (Methods) and treated the array genotypes at these variants as the reference. Imputation accuracy was then quantified by calculating r^2 and genotype concordance (homozygous reference, heterozygous, and homozygous alternative) for each sample. Because imputation quality typically depends on the representation of variant haplotypes in the reference panel, and rare variants are generally harder to impute, results were stratified by allele frequency estimated from African data in the Genome Aggregation Database (gnomAD, extracted by gsutil with Google Cloud Public Datasets from https://gnomad.broadinstitute.org/downloads) [51, 296]. Full details of the above calculation are provided in Methods. At this stage, the variants considered included only bi-allelic SNPs, which was the only type of variant that QUILT could impute [66].

I characterised lcWGS imputation accuracy in Figure 3.2.1 using a 0.5% allele frequency threshold from gnomAD, due to the limited power to impute rare variants with the 1000 Genomes Project reference panel. The colour bar on the right of the graph correlates to the number of variants in each allele frequency bin. Imputation accuracy increases as variants become more common, with $r^2 = 0.970$ for variants with allele

frequency > 1% (n = 541,973) and r^2 = 0.974 for variants with allele frequency > 5% (n = 350,595). When variants were filtered using the INFO score, which reflects imputation confidence, accuracy further improved for variants with allele frequency > 1% (r^2 = 0.982, n = 445,746), consistent with observations from other studies and demonstrating satisfactory performance [66, 70, 73]. I also categorised variants by sample genotypes and calculated non-reference concordance (genotype concordance excluding homozygous reference genotypes) at the same threshold, yielding 0.99 for heterozygous and 0.991 for homozygous alternative genotypes. On the other hand, imputing rare homozygous alternative genotypes remained challenging at 1× genome coverage using the 1000 Genomes Project reference panel, with a concordance of 0.698 for the 930 homozygous alternative genotypes observed at allele frequency 0.5%-1%. Although lcWGS provides reliable genotyping for common variants, sufficient for most GWAS applications, these results highlight that further improvements are possible to unlock the full potential of lcWGS data.

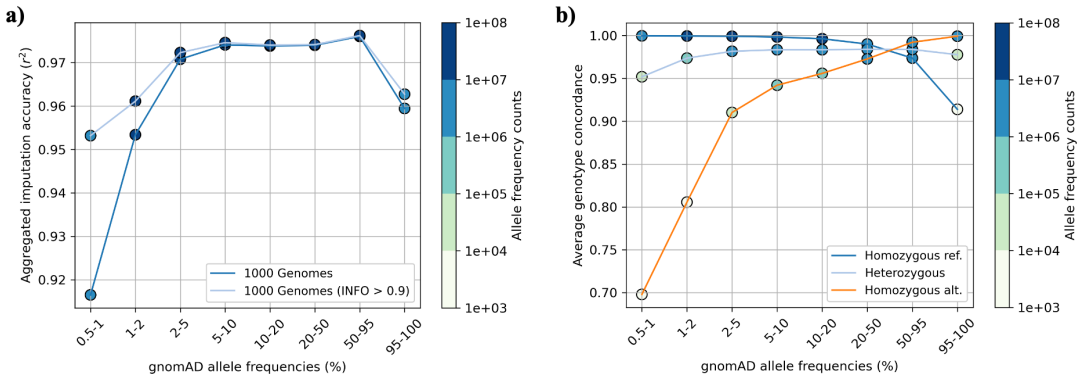


Figure 3.2.1. Imputation accuracy of 186 GAMCC samples with 1000 Genomes Project reference panel at microarray typed variants. a) Mean accuracy (y-axis, computed as r^2 between imputed and

directly-typed SNPs and averaged across individuals) for variants stratified with different gnomAD African allele frequencies (x-axis), using either all variants (dark blue) or variants with high imputation confidence (INFO > 0.9, light blue). **b)** Mean genotype concordance (y-axis) between directly-typed and the imputed SNPs, separated by allele frequencies (x-axis) and the microarray genotype (colours).

To assess whether imputation accuracy was consistent across the four ethnic groups in this cohort, I stratified the QUILT 1000 Genomes Project imputation results (restricted to variants with INFO > 0.9) by ethnic group (Figure 3.2.2). Overall, imputation accuracy was comparable across groups, with the exception of the Mandinka population (dark orange), which showed slightly reduced performance. This reduction is likely attributable to the lower mean genome coverage in Mandinka individuals (1.18× compared with 1.20×, 1.22×, and 1.23× for Jola, Wollof, and Fulani, respectively). Another contributing factor may be reduced haplotype similarity to those represented in the 1000 Genomes Project reference panel. Most Gambian haplotypes in the reference panel originate from the Mandinka population, and the greater genetic divergence of the Fulani (Figure 3.3.1) may partly explain why their comparatively higher coverage did not translate into superior imputation accuracy. However, differences in both coverage and imputation accuracy across groups were small and may reflect stochastic variation in coverage distribution or the imputation process. These results were considered satisfactory, and no further conclusions were drawn.

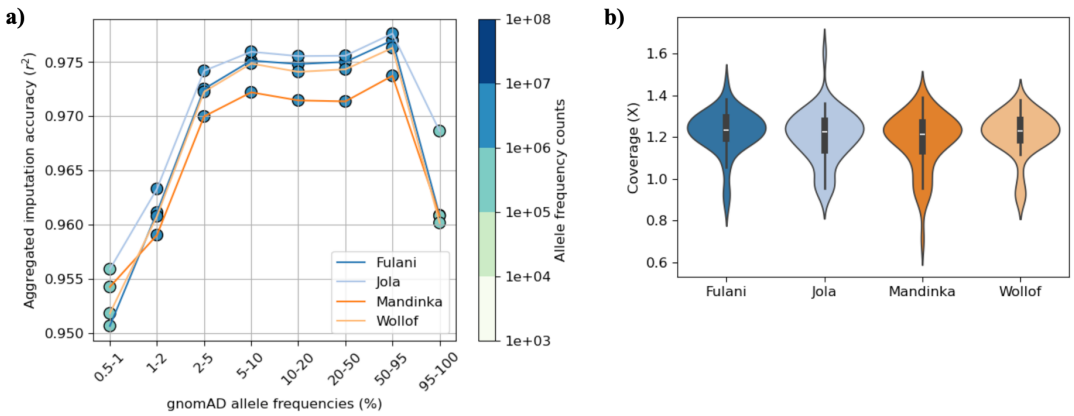


Figure 3.2.2. Imputation accuracy of 186 GAMCC samples with 1000 Genomes Project reference panel separated by ethnic group. a) Mean accuracy (y-axis) for variants stratified by different gnomAD African allele frequencies (x-axis), separated by ethnic group (colours). **b)** Genome coverage for each sample, separated by ethnic group.

3.2.2 Impact of additional African genomes on low-coverage whole-genome sequencing imputation performance

The performance of imputation is largely influenced by the abundance and diversity of haplotypes in the reference panel, and incorporating population-specific haplotypes enhances variant inference in diverse populations [73, 297-299]. Thus, I explored an expanded 1000 Genomes Project reference panel, including 765 additional samples from The Gambia, Burkina Faso, Cameroon, and Tanzania from the MalariaGEN study [115], and the combined reference panel harbours 3,967 genomes. Figure 3.2.3 shows that incorporating these African haplotypes consistently improves imputation accuracy compared to the previous run, with gains generally diminishing as variant frequency increases (r^2 = 0.916, 0.940 at 0.5%-1%; 0.974, 0.982 at 20%-50%, respectively). Moreover, genotype concordance across all types and allele frequencies improved with the addition of these genomes to the reference panel, with homozygous alternative genotypes showing the largest gain (concordance = 0.698, 0.740 at 0.5%-1%; 0.973,

0.981 at 20%-50%, respectively). Nevertheless, the power to impute rare variants remained limited at this reference panel size, and a substantial number of variants were still difficult to impute accurately.

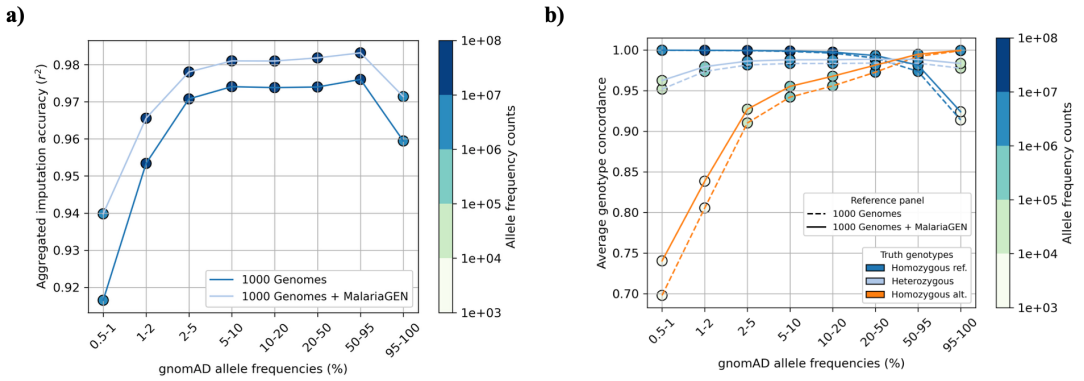


Figure 3.2.3. Imputation accuracy of 186 GAMCC samples with the MalariaGEN-enriched 1000 Genomes Project reference panel, compared at microarray typed variants. a) Mean accuracy (y-axis) for variants stratified by different gnomAD African allele frequencies (x-axis), coloured by different reference panels. **b)** Mean genotype concordance (y-axis) by different imputation reference panel (linestyles) between directly-typed and the imputed SNPs, separated by allele frequencies (x-axis) and the microarray genotype (colours).

3.2.3 A two-stage imputation workflow leveraging the Trans-Omics for Precision Medicine reference panel

The introduction of the TOPMed reference panel in 2020 represented a major advancement in genotype imputation, integrating the most comprehensive collection of variants and the largest sample sizes from diverse global populations, thereby enabling more accurate interrogation of rare variants [104, 158]. However, access to this powerful resource is limited: TOPMed imputation can only be performed via the TOPMed imputation server and cannot be run locally with QUILT due to data restrictions. This poses a critical challenge for lcWGS, as the accuracy of rare variant

imputation is particularly sensitive to the size and diversity of the reference panel, and smaller panels such as the 1000 Genomes Project are insufficient. To overcome this limitation, a two-stage imputation strategy was devised, effectively leveraging QUILT for lcWGS imputation while incorporating the large TOPMed reference panel. In the first stage, lcWGS data were imputed against a smaller reference panel (for example, 1000 Genomes Project reference panel) and filtered to retain only high-confidence, common variants. These variants were then restricted to loci represented on the densest commercial array, effectively creating an array-like input compatible with the TOPMed imputation server. In the second stage, the server imputed against the full TOPMed reference panel, thereby capturing rarer variants and indels that could not be accessed with QUILT.

Specifically, variants from the first-stage QUILT imputation were filtered to retain only common SNPs (allele frequency > 1%) with high confidence (INFO > 0.9). Loci corresponding to those targeted by the Illumina HumanOmni5M-4v1 array (Omni5M) were also retained to meet the density requirements of the TOPMed server (see Methods for details). Finally, the prepared dataset was submitted for second-stage imputation using the TOPMed r3 release, which comprises 133,597 reference samples. Using eight high-coverage samples (29.79×-35.22×, mean 32.05×) as a gold standard, genome-wide imputation accuracy was compared across different workflows and microarray data. Figure 3.2.4 shows that lcWGS consistently outperformed microarrays with the 1000

Genomes Project reference panel, particularly for low-frequency variants ($r^2 = 0.902$ vs. 0.782 at 1%-2%). Leveraging the larger TOPMed reference panel substantially improved imputation accuracy, especially for rare variants, due to the greater abundance and genetic diversity of haplotypes. The two-stage imputation strategy for lcWGS (light and dark orange lines) further demonstrated exceptional performance, achieving r^2 values of 0.985 and 0.972 at 10%-20% allele frequency, compared with 0.963 for direct TOPMed imputation on microarray data. This comparison also underscored the critical importance of incorporating additional haplotypes into reference panels, as illustrated in Figure 3.2.3. While direct TOPMed imputation on microarray data achieved slightly higher accuracy for rare variants (gnomAD allele frequency < 0.1%), this likely reflected inherent uncertainty in rare variant calls during the first-stage imputation, which relied on a relatively smaller reference panel. However, this limitation is not a major concern in practical applications, such as GWAS, where rare variants are often excluded due to limited statistical power.

Additionally, another advantage of TOPMed-based imputation is its ability to recover indels. Figure 3.2.4d shows that the two-stage lcWGS imputation (QUILT on the MalariaGEN-enhanced 1000 Genomes Project reference panel followed by TOPMed imputation) retrieved a substantial number of indels with allele frequency > 0.5% (2,480,734 out of 45,674,756 total variants, 5.43%). Benchmarking against the eight high-coverage genomes revealed that SNPs were imputed with higher accuracy than

indels, reflecting the greater challenges of imputing indels due to their complex genomic architecture and higher susceptibility to calling errors [300]. Again, lcWGS imputation consistently outperformed microarray-based imputation using TOPMed. This highlights a methodological need to incorporate indels into QUILT imputation, as they contribute substantially to genetic diversity and disease associations. However, this aspect was not addressed in the current thesis and remained only accessible via TOPMed imputation. Overall, the two-stage imputation strategy demonstrates the potential of lcWGS as a robust alternative to DNA microarrays for capturing both SNPs and indels across the human genome.

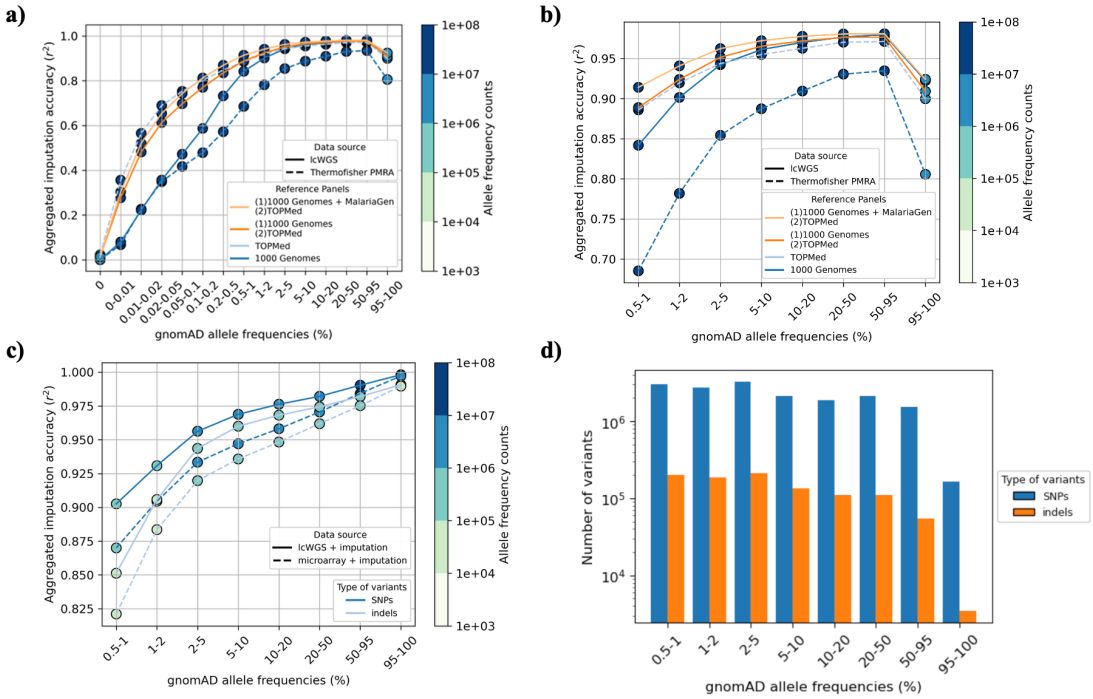


Figure 3.2.4. Two-stage genome-wide imputation with TOPMed reference panel. a) Mean accuracy (y-axis) for variants stratified by different gnomAD African allele frequencies (x-axis), coloured by different reference panels and imputation workflow. Microarray and lcWGS data are represented by dashed and solid lines, respectively. **b)** Same as a), restricted to common variants. **c)** Mean accuracy (y-axis) by SNPs (dark blue) and indels (light blue), calculated from the best imputed microarray (dashed) and lcWGS data (solid). **d)** Number of variants (y-axis, plotted on log scale) captured by imputation for variants stratified with different gnomAD African allele frequencies (x-axis), coloured by variant type.

3.3 Inference on population structure in the Gambian Malaria Cases and Controls cohort

In the previous sections, I demonstrated that lcWGS with imputation achieved high genome-wide accuracy, particularly for common variants. It is of interest whether a similar population structure can be revealed by these variants compared to genotyping data, particularly in this population in which several genetically distinguishable ancestral groups (Fulani, Jola, Wollof, and Mandinka) living in close proximity. To explore this, I performed principal component analyses (PCA) on three datasets: lcWGS data imputed with QUILT using the 1000 Genomes Project reference panel, genotyping data, and a combined dataset comprising 50% of individuals from each. Analyses were restricted to 742,235 autosomal variants typed on the microarray and 184 individuals (two excluded due to high kinship coefficients) that were common across all datasets (see Methods for details). The mix-up dataset is of particular interest for studies combining data generated from different sources, especially in meta-analysis researches. The results in Figure 3.3.1 reflect genetic differences between individuals who reported as Fulani, Wollof, Mandinka and Jola ethnic groups, similar to previous data [115], and this separation was indifferent on data generation: Figure 3.3.1a-c reveal almost the same structure, suggesting the feasibility of mixing data from different sources via this fake study. Figure 3.3.1e shows a pairwise correlation analyses by the first five PCs derived from lcWGS and microarray, again indicating strong concordance

for the first three PCs (correlation > 0.98), while subsequent PCs reveal substructure and are dominated by single SNPs. The reduced correlation observed for minor PCs across technologies is likely due to the combination of small variance explained by these PCs, imputation imperfection, or genotyping errors. For example, SNP rs72668543 contributed substantially to all PCs in the lcWGS dataset but had negligible loadings across PCs calculated from microarrays. Imputation accuracy at this site was 0.46, with one individual carrying a heterozygous genotype incorrectly imputed as homozygous reference. This error was likely due to the rarity of the variant (gnomAD allele frequency 1.7%) and could subsequently distort principal components. Overall, although lcWGS shows reduced imputation accuracy at rare variants and may overlook subtle genetic relationships, it reliably captures the major axes of population structure and remains suitable for standard GWAS analyses.

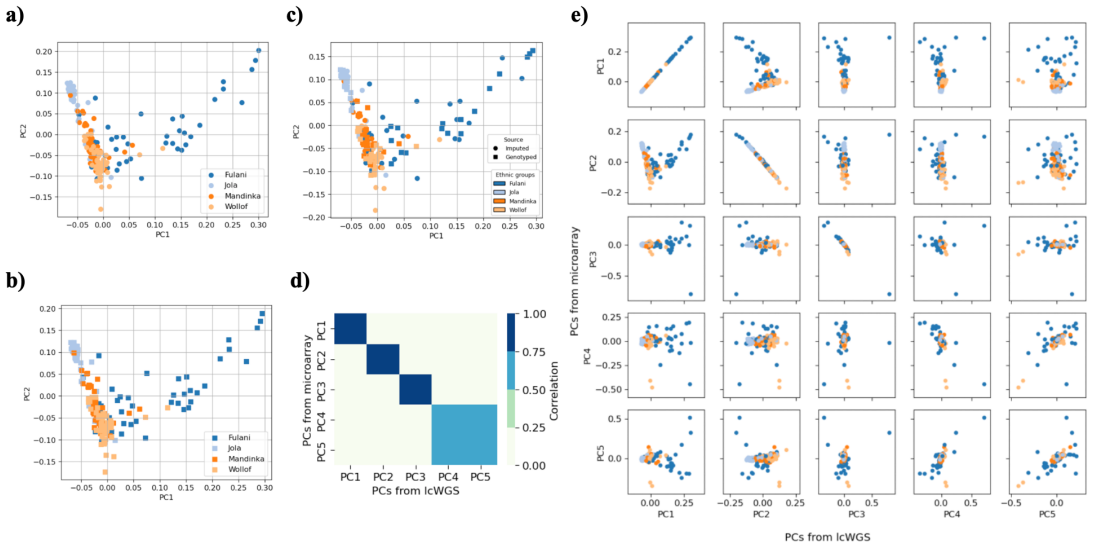


Figure 3.3.1. PCA on lcWGS and microarray data. a) The first two PCs calculated from QUILT imputation with 1000 Genomes Project reference panel. **b)** The first two PCs calculated from genotyping data. **c)** The first two PCs calculated by mixing half of the samples by imputation and half by genotyping.

All analyses are performed on the same set of variants and individuals. **d)** Absolute values of correlations on the first 5 PCs by lcWGS (x-axis) and microarray (y-axis). **e)** Pairwise scatter plot of first 5 PCs from lcWGS (x-axis) and microarray (y-axis).

3.4 Characterisation of low-coverage whole-genome sequencing imputation accuracy at specific regions

3.4.1 Genome-wide association study catalogue variants

As a key approach in genetic epidemiology studies, GWAS relies on accurate genotyping to generate reliable and meaningful associations. Apart from genome-wide imputation results presented in Section 3.2, I focused on specific regions and variants of particular interest. I downloaded GWAS catalogue variants from https://hgdownload.soe.ucsc.edu/goldenPath/hg38/database/gwasCatalog.txt.gz, a curated collection of SNP-trait associations with $p < 1.0 \times 10^{-5}$ [301, 302], to assess the ability of lcWGS to capture biologically relevant variants. Additionally, a set of 47 genes encoding blood group antigens was obtained from the HUGO Gene Nomenclature Committee [303], many of which are established risk factors for parasite infection via erythrocytes [304]. All variants within these genes were extracted for this analysis. Due to the limited sample size and reduced imputation accuracy at rare variants, SNPs with gnomAD allele frequency below 2% were excluded from this analysis. The performance of the remaining variants was then compared to genome-wide results using lcWGS data subjected to two-stage imputation (QUILT imputation using MalariaGEN-enhanced 1000 Genomes Project reference panel followed by

TOPMed imputation [89, 104, 158, 305]) and evaluated with the concordance metric. Figure 3.4.1 shows that imputation accuracy for GWAS-associated SNPs and genetic variants in blood group genes is comparable to genome-wide levels across all allele frequency bins, with no evidence of systematic bias. These results indicate that lcWGS-based imputation reliably captures biologically and clinically relevant variants, providing a solid foundation for downstream association analyses.

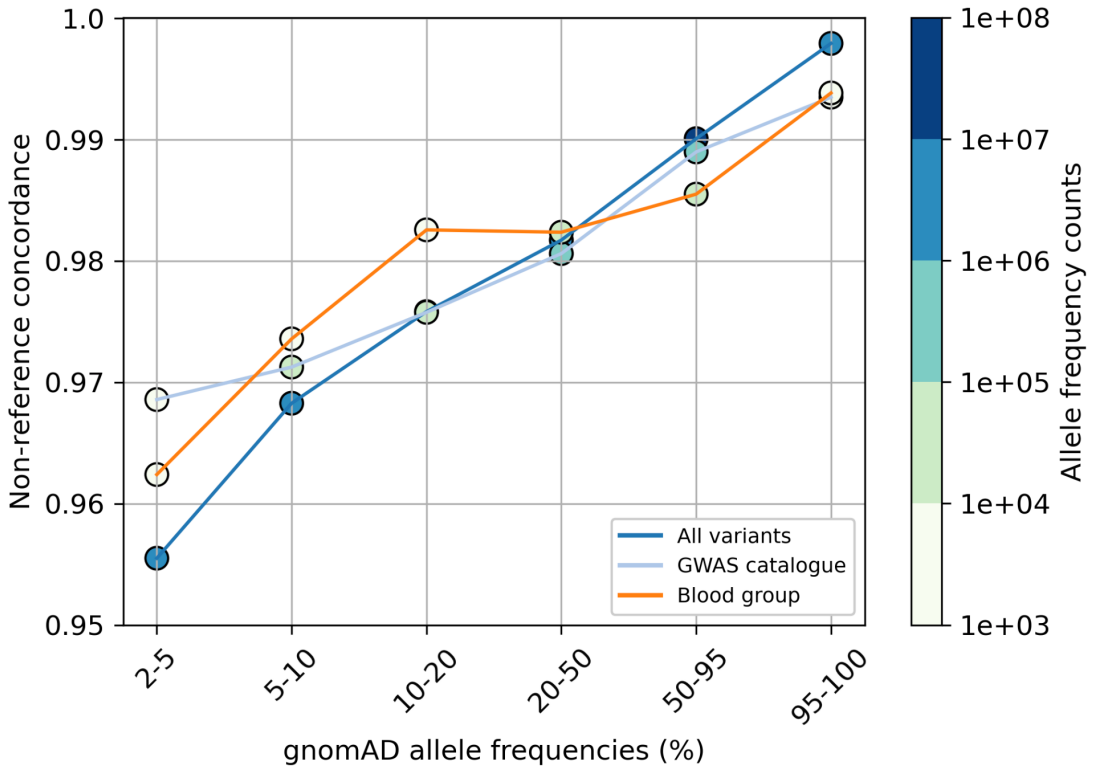


Figure 3.4.1. Imputation accuracy for different sets of variants from lcWGS two-stage imputation results. Non-reference concordance (y-axis) for variants stratified with different gnomAD African allele frequencies (x-axis), coloured by all variants (dark blue, total 11,753,267 variants), GWAS catalogue variants (light blue, total 169,022 variants), and variants in blood group genes (orange, total 7,855 variants).

3.4.2 A least accurate segmental duplication region

On the contrary, for variants that are difficult to capture accurately by sequencing due

to mapping issues or other technical limitations, microarray genotyping becomes particularly valuable, as it can reliably ascertain these sites. To characterise regions where lcWGS has poor performance, I identified 4,315 SNPs typed on the microarray where QUILT imputation on lcWGS was compromised ($r^2 < 0.5$), took a flanking region of 1 Mb centred around each SNP, and calculated average r^2 for each region. Figure 3.4.2 presents genome coverage and mapping quality, as calculated by coverotron [286], around the examined variants plotted against allele frequency. Fourteen SNPs were excluded from the figure for clarity due to elevated genome coverage (> 1.5×), of which ten exhibit compromised mapping quality (< 59), indicative of regions prone to alignment artefacts. Allele frequency is the primary factor underlying low imputation accuracy at these sites, with 86.7% of variants exhibiting a minor allele frequency below 5%, inherently limiting their imputation performance; other factors, such as GC content or genotype calling artefacts, may also contribute but were not explored further in this section.

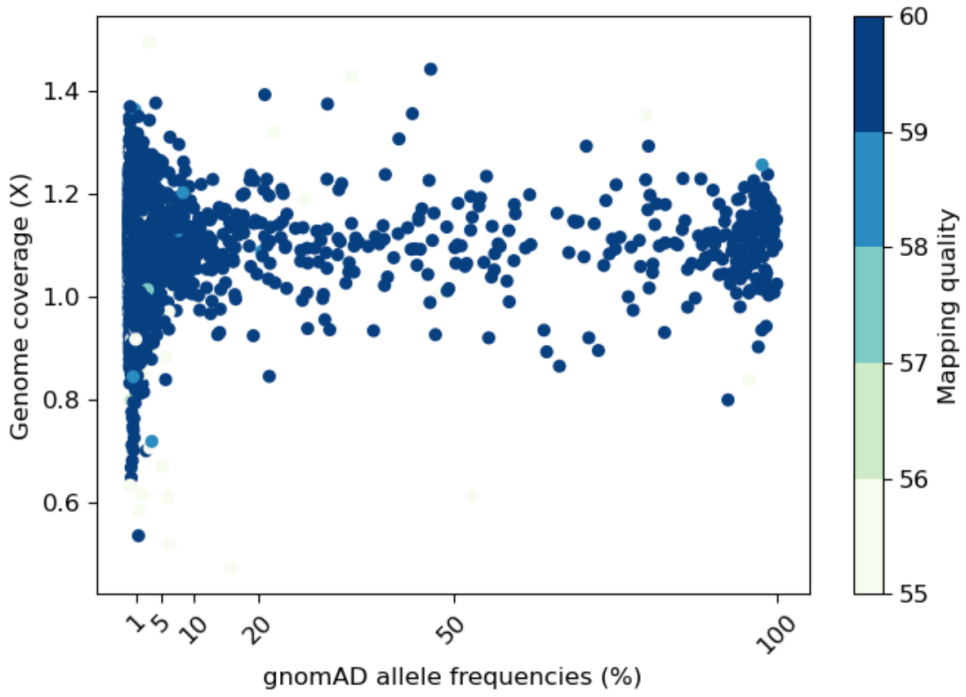


Figure 3.4.2. Quality metrics of SNPs with low imputation accuracy. Genome coverage (y-axis) and mapping quality (colour, both averaged across the nearest 1 kb genomic bin of each SNP) are plotted against gnomAD allele frequency.

Figure 3.4.3 characterises the least accurately imputed region (chr1:161,113,000-162,113,000, rounded to the nearest kb), which contains several Fc gamma receptor (FCGR) genes. FCGR genes are key components of the immune system, mediating antibody-dependent immune responses by binding the Fc region of immunoglobulin G and triggering diverse immune mechanisms [306, 307]. The FCGR locus contains approximately eight closely related paralogous genes, many clustered within the 1 Mb window on chromosome 1q23. These genes exhibit high sequence similarity and duplicative complexity, making short-read sequencing alignment and imputation particularly challenging [308]. This observation is evident from the mappability track and the average mapping quality calculated from sequencing data (see Methods for

details), both of which align with the SNPs exhibiting reduced imputation performance. Complex segmental duplication regions are challenging to be well captured by lcWGS, and higher-throughput long-read sequencing offers a more effective approach for resolving them.

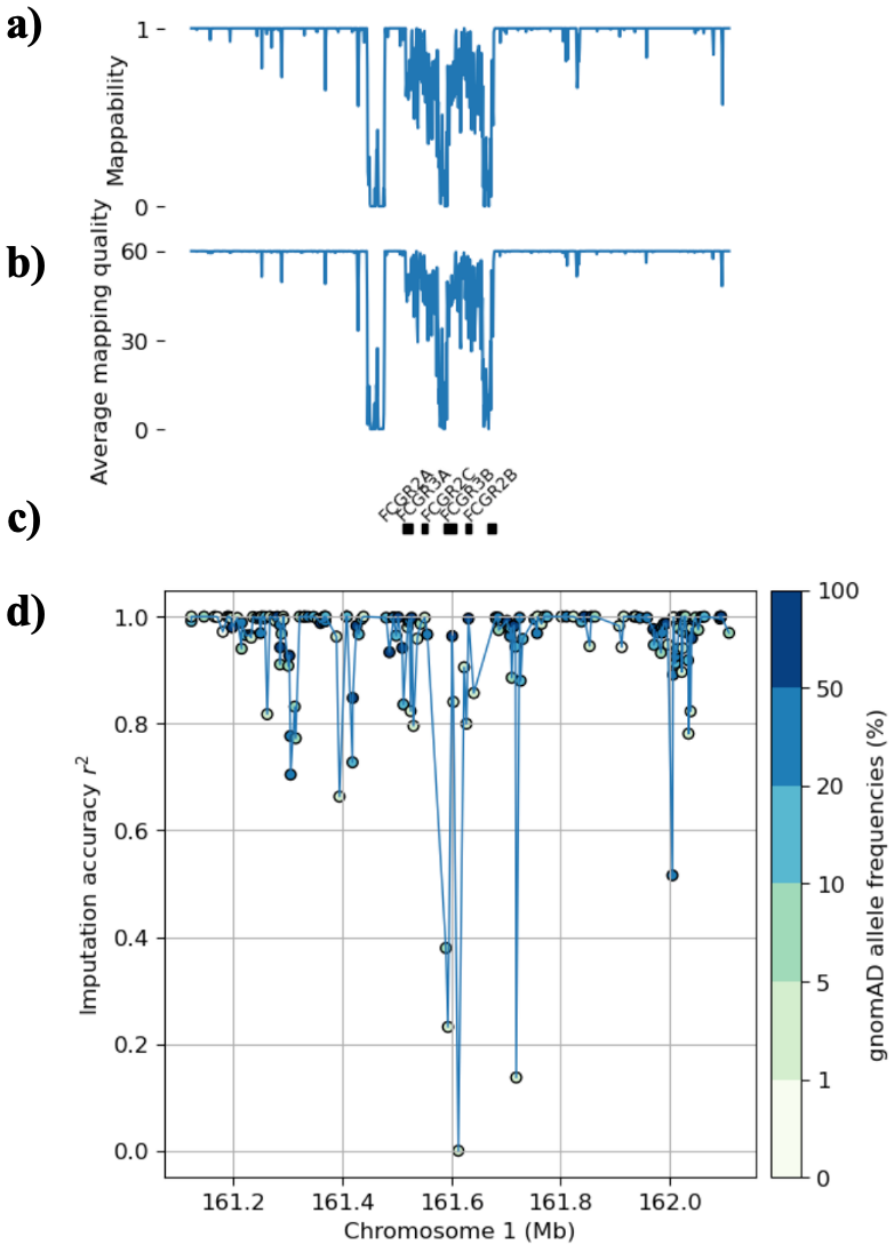


Figure 3.4.3. The least accurate 1 Mb region (chr1:161,113,000-162,113,000, rounded to the nearest kb) identified from lcWGS imputation. a) Mappability in the region, calculated in genomic windows of 1 kb. **b)** Mapping quality averaged across the individuals in the region, calculated in genomic windows of 1 kb. **c)** The FCGR genes in this region. **d)** Imputation accuracy by QUILT with 1000 Genomes Project reference panel, compared with microarray typed variants.

3.5 Blood group genetic variants implicated in malaria resistance

With the life-threatening endemic exposure to severe malaria, the African populations have been extensively studied for genetic variants associated with pathogenesis [115, 192, 309-312]. This cohort provided a unique opportunity to investigate genetic associations with malaria susceptibility, demonstrating the practical utility of lcWGS for uncovering biologically meaningful insights. Given the limited sample size and statistical power, the analysis focused on well-established blood group genes, which are known to harbour significant risk factors for malaria [304]. I characterised the reliability of association tests by assessing imputation accuracy (using deep whole-genome sequencing genotypes as the reference) for variants within 47 blood group antigen genes defined by the HUGO Gene Nomenclature Committee [303] (Figure 3.5.1). Except for three regions where no variants were shared between the imputed results and high-coverage genotypes, consistently high imputation accuracy was observed across blood group genes, with the lowest-performing region (*FUT3*) still achieving an average r^2 of 0.878. This reliable genotyping information enabled subsequent association tests. Specifically, three previously reported variants [115] associated with malaria severity were examined, and logistic regressions were conducted using case-control status in the cohort (grouping mild and severe malaria as cases, see Methods for details). Table 3.5.1 summarises the results.

Chromosome	Position	Ref.	Alt.	Nearest gene	Rsid	Alt. allele frequency	Non-reference genotype frequency (case)	Non-reference genotype frequency (control)	Odds ratio	P-value
11	5227002	T	A	HBB	rs334	0.043421	0.032520	0.174603	0.139825	0.001381
9	133257521	T	TC	ABO	rs8176719	0.303388	0.634146	0.523810	1.635259	0.126239
1	203689343	A	G	ATP2B4	rs4951377	0.681095	0.902439	0.857143	1.729267	0.255957

Table 3.5.1. Three established common variants associated with severe malaria. Association test on malaria severity conducted with imputed microarray data (with TOPMed). All SNPs and indels were reported according to the gnomAD convention in GRCh38 coordinates, using the forward (reference) strand. Alternative allele frequency was obtained from gnomAD. Ref.: Reference; Alt.: Alternative; *HBB*: Haemoglobin subunit beta; *ABO*: ABO blood group; *ATP2B4*: Plasma membrane calcium-transporting ATPase 4.

The protective effects of rs8176719 (T allele) and rs4951377 (A allele) were replicated in this dataset and were consistent with previous reports, although neither reached statistical significance [115]. In contrast, another well-established SNP, the sickle haemoglobin variant in *HBB* (rs334, chr11:5227002:T>A), reached statistical significance [115, 192, 309, 312]. It is important to note that our allele coding followed the gnomAD convention; because *HBB* is located on the reverse strand, some studies may report the protective allele as T, whereas in our data this corresponded to its reverse complement. Figure 3.5.1b and Figure 3.5.1c compare genotype composition between lcWGS-imputed and directly typed microarray data. Only three samples with heterozygous genotypes were wrongly imputed as homozygous reference, resulting in nominal significance ($p = 0.007$ for lcWGS, 0.001 for microarray, respectively). This discrepancy reflects uncertainty introduced during imputation, as also illustrated in the forest plot. The protective effect of the alternative allele (A) is suggested by the odds ratio < 1, compared to the reference (T). This proof-of-principle association analysis demonstrates the potential of lcWGS in a GWAS context.

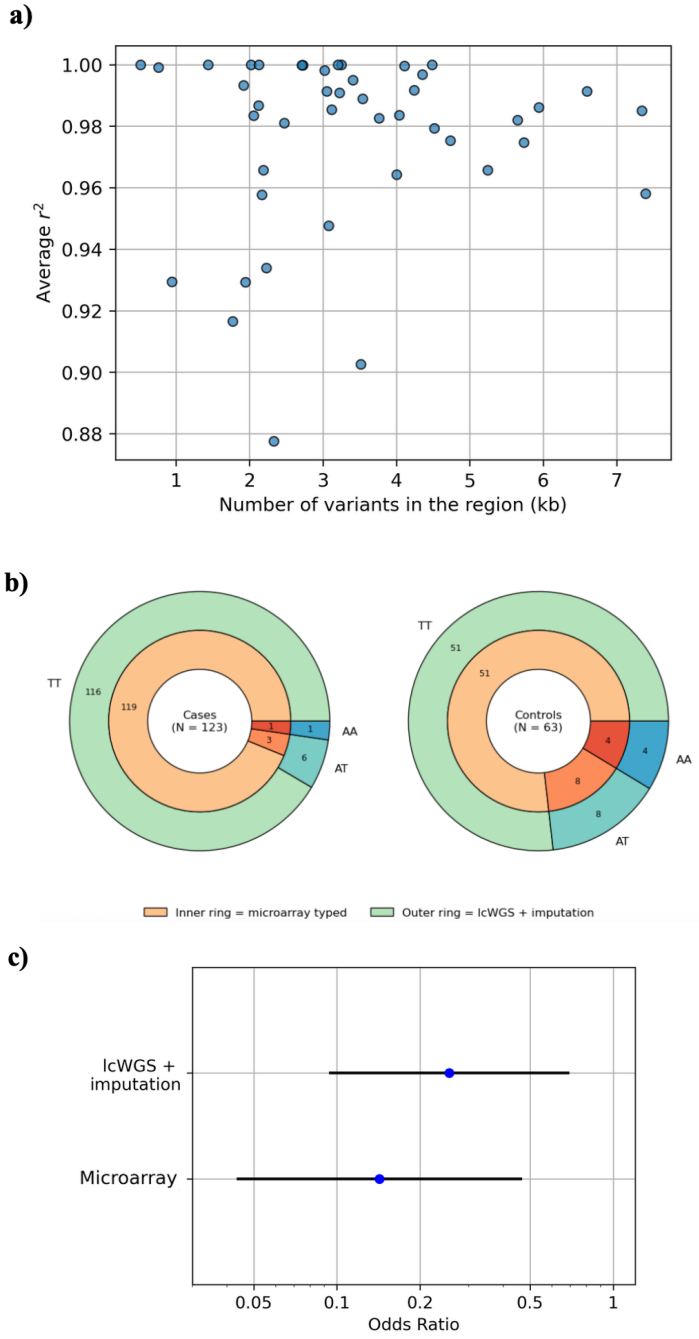


Figure 3.5.1. Association analysis of rs334. a) Average imputation accuracy for variants in blood group genes. **b)** Genotype compositions of SNP rs334 (chr11:5,227,002:T>A). **c)** Forest plot of association tests with rs334 on malaria severity.

3.6 Methods

Data collection. The GAMCC dataset was generated from 210 Gambia originally

obtained from children enrolled in an epidemiological study of clinical malaria [289] (Section 3.1.1). For lcWGS, samples were prepared with the NEBNext Ultra II FS DNA Library Prep Kit for Illumina kit, with 100 ng input DNA and 6 PCR cycles (as identified in Chapter 2). These samples were then sequenced with the Illumina NovaSeq 6000 platform to generate 151 bp paired-end reads. To compare our low-coverage samples with DNA microarrays, two plates comprising 190 samples from the total 210 were genotyped using the Affymetrix Axiom™ Precision Medicine Research Array, of which four failed post-genotyping QC and were excluded. Additionally, eight samples with both microarray and low-coverage sequencing data were selected for deep whole-genome sequencing to serve as a gold standard for comparison.

lcWGS data processing. The processing procedure for these samples largely followed the workflow described in Chapter 2, with a few modifications. First, adapters were trimmed using trimmomatic v0.3 [313], and duplicate reads were removed using fastuniq v1.1 [314]. Second, instead of aligning solely to the human genome, I also included the malaria genome (Pf3D7_v3) and icosahedral bacteriophage ΦX174 (accession NC_001422.fna, released on Jan 2023) to account for possible malaria sequences in the infected samples [3]. Incorporation of the ΦX174 genome was previously shown to improve base diversity and sample de-multiplexing in NGS experiments [315].

lcWGS QUILT imputation. Genome-wide imputation was performed on 210 lcWGS samples using QUILT (v2.0.0) [66], with both the 1000 Genomes Project reference panel (3,202 individuals, 30× sequencing by the New York Genome Centre, GRCh38) and a MalariaGEN-enriched 1000 Genomes Project reference panel, which incorporates an additional 756 African genomes from the MalariaGEN project [115]. The latter was obtained from a previous study in GRCh37 coordinates, so I used picard LiftoverVcf and the b37ToHg38 chain file (https://raw.githubusercontent.com/broadinstitute/gatk/master/scripts/funcotator/data_sources/gnomAD/b37ToHg38.over.chain) to perform liftover [49]. Both reference panels were restricted to bi-allelic SNPs to satisfy QUILT requirements and converted to haplegend format using bcftools v1.1.6 [283].

I downloaded the YRI genetic map file from the 1000 Genomes Project (ftp-trace.ncbi.nih.gov/1000genomes/ftp/technical/working/20130507_omni_recombination_rates/YRI_omni_recombination_20130507.tar), along with the liftOver tool from the UCSC Genome Browser (http://hgdownload.cse.ucsc.edu/admin/exe/linux.x86_64/liftOver) and its accompanying chain file (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/liftOver/hg19ToHg38.over.chain.gz). The map was converted to GRCh38 coordinates using the make_b38_recomb_map utility included in the QUILT package. Human autosomes were then divided into 5 Mb windows, and the reference data were prepared in QUILT format using the QUILT_prepare_reference programme, with all

parameters set to default except nGen =100. Lastly, I imputed the samples in each autosomal region with QUILT v2.0.0.

DNA microarray data processing. A total of 190 samples underwent microarray genotyping using the Affymetrix Axiom™ Precision Medicine Research Array (ThermoFisher), of which 186 passed QC, resulting in 872,149 typed variants mapped to GRCh38. I refined the variants by removing 34,212 sites on non-autosomal chromosomes, 855 sites with genotype missingness greater than 5%, and 365 sites deviating from Hardy-Weinberg equilibrium ($p < 1 \times 10^{-6}$). To filter out alleles whose observed frequencies deviated from previous data, allele frequencies were obtained from the 1000 Genomes Project reference panel and supplemented with the Gambian Genome Variation Project where data was missing, providing a benchmark for comparison [316].

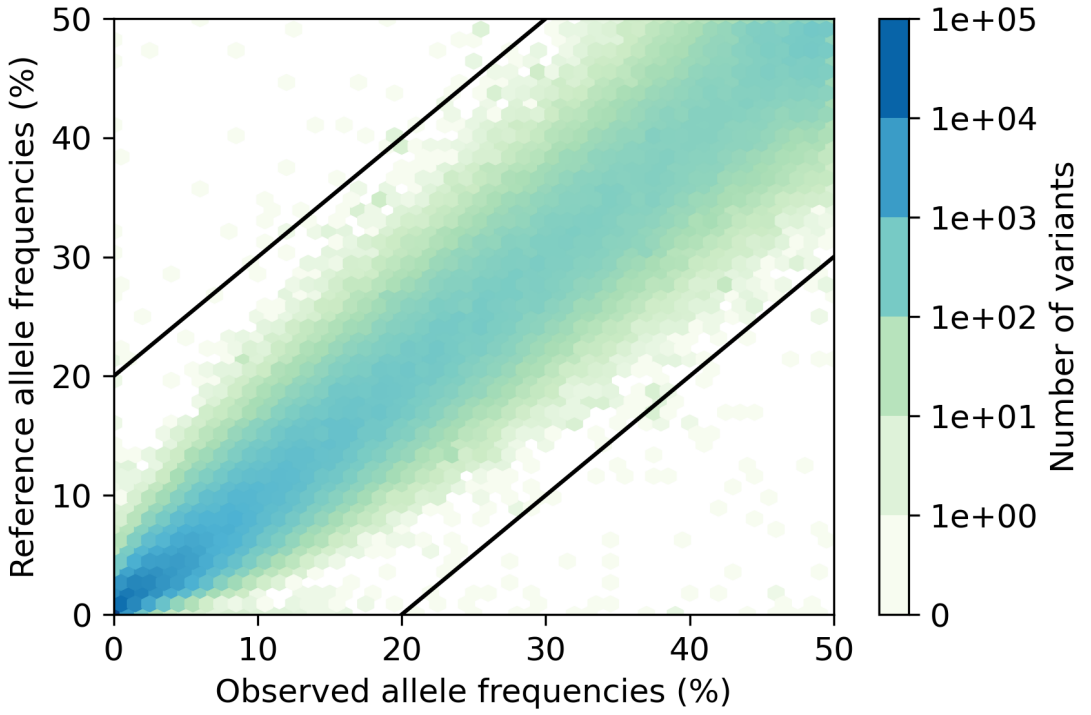


Figure 3.6.1. Allele frequency control on microarray data. The y-axis reflects allele frequency curated with the African population in the 1000 Genomes Project reference panel, supplemented by the Gambian Genome Variation. These results were then matched to the observed genotype data (x-axis). Variants with an absolute difference in allele frequency greater than 20% between the reference and observed data were excluded from subsequent analyses (bold line). The colour track of the hexagons shows varying number of variants within each bin, with darker shades indicating higher variant density. I removed 189 genotyped variants in this step.

In total, 836,528 (95.92%) sites were retained. No samples were excluded due to high missingness (> 5%), although two samples exhibited relatively high kinship coefficients, as identified with QCTOOL v2.2.2, and were therefore removed from the PCA plot in Figure 3.3.1 [317]. Finally, I uploaded the final microarray dataset to the imputation servers using both the 1000 Genomes Project Phase3 30× reference panel and the TOPMed r3 reference panel [54] for genome-wide imputation.

High-coverage data processing. Eight replicates underwent deep whole-genome sequencing (range 29.79×-35.22×, mean 32.05×) on the NovaSeq 6000 platform. The workflow differed from lcWGS after genome alignment, as I proceeded with GATK v4.3.0.0 BaseRecaliberator and ApplyBQSR methods for base quality recalibration [318]. To improve computational efficiency, autosomal regions were divided into 5 Mb chunks and processed in parallel. I used GATK HaplotypeCaller setting -L and --alleles options, using the 1000 Genomes Project vcf files separated by SNPs or indels, and --output-mode EMIT_VARIANTS_ONLY to force variant calling at these sites for downstream benchmarking. Finally, I ran GATK VariantRecalibrator and ApplyVQSR to calculate the variant quality score log-odds, providing an assessment of calling quality.

Imputation accuracy and confidence. Imputation accuracy was assessed using two metrics: non-reference concordance and imputation r^2, both ranging from 0 to 1, where higher values indicate greater accuracy. For a set of variants (typically stratified by gnomAD allele frequency), non-reference concordance is defined as the proportion of correctly imputed non-reference genotypes (heterozygous or homozygous alternative) among all non-reference genotypes, averaged across variants. Imputation r^2 is a continuous measure representing the squared correlation coefficient between the imputed diploid dosage and the true genotype, with the reference and alternative alleles coded as 0 and 1, respectively, and averaged across samples. Both metrics require the

joint presence of imputed genotypes, true genotypes, and non-missing gnomAD allele frequencies. Imputation confidence is quantified using the INFO score, a statistical metric that estimates the reliability of genotype imputation without requiring knowledge of the true genotypes [319]. For a given variant, it is computed as

INFO = 1 - mean(variance in imputed genotype / variance if only allele frequency were known)

The numerator averages across samples of the confidence of imputation by calculating variance of the genotype probabilities. The denominator is the Hardy-Weinberg equilibrium genotype variance estimated from the observed allele frequency based on the imputed dosages. INFO equals 1 if all genotypes are completely certain and drops to 0 as certainty decreases.

lcWGS two-stage imputation with TOPMed. Starting with imputation result, I kept only the genotype field (GT) in the FORMAT column and retained confidently imputed (INFO > 0.9) and common (gnomAD African allele frequency > 1%) SNPs as reliable callset. Since the imputation servers required microarray data as input by restricting the maximum number of variants in each chunk of chromosomes, I further subsetted SNPs probed by the densest commercially available DNA microarray, namely the Illumina HumanOmni5M-4v1 genotype array due to the density of SNPs after first filtering step

being off threshold. I then uploaded the newly curated vcf files to the TOPMed imputation server [104, 158].

Population structure inference. PCA was performed on three set of data: lcWGS data imputed with QUILT using the 1000 Genomes Project reference panel, genotyping data, and a combined dataset comprising 50% of individuals each from the two. These analyses were restricted to a common set of 742,235 autosomal variants, obtained from an intersection of post-QC microarray variants and post-QC QUILT-imputed variants (gnomAD allele frequency > 1% and INFO > 0.9), and a common set of 184 individuals, two excluded due to high kinship estimates that biased the decomposition (kinship coefficient > 0.1, estimated by QCTOOL). The combined dataset was constructed by split individuals from each ethnic group to each half for the two technologies. Lastly, I calculated PCs with PLINK v1.90b6.21 [320, 321].

The least accurate *FCGR* region. Average imputation accuracy for 1 Mb flanking window centred around each SNP with $r^2 < 0.5$ was calculated by comparing QUILT imputation using 1000 Genomes Project reference panel on lcWGS with microarray typed variants. The anchor SNP was chr1:161,612,689:A>G, so I focused on the region of chr1:161,113,000-162,113,000, where boundaries were rounded to the nearest kb to ease the following analyses. I calculated the mappability, obtained from a previously curated multi-read mappability track with 100-mers throughout the human genome

[322], and the mapping quality of sequencing reads, reported by coverotron [286] and averaged across individuals, both average per 1 kb genomic bin. Details of these calculations are described in the Methods section of Chapter 5. I also obtained a list of genes that overlaps with this region from the Ensembl REST API (https://rest.ensembl.org/overlap/region/human/1:161113000-162113000?feature=gene) and extracted the genomic coordinates of FCGR genes in this region [323]. These information were then assembled with the imputation accuracy plot.

Blood group genes and association tests. I queried blood group antigen genes from the HUGO Gene Nomenclature Committee website (https://www.genenames.org/), yielding in total of 47 regions [303]. I measured imputation accuracy across these regions by comparing the best lcWGS imputation results (QUILT imputation with MalariaGEN-enhanced 1000 Genomes Project reference panel followed by TOPMed) to the eight high-coverage replicates. For the sickle haemoglobin allele rs334, I compared the same imputation results with microarray data as the SNP was directly typed with the 186 common samples. Phenotype data corresponded to the malaria status of individuals in the GAMCC cohort. Original samples were categorised as controls, mild malaria, and severe malaria. I combined mild and severe malaria cases to increase statistical power, resulting in 123 cases and 63 controls. A separate multinomial logistic regression without grouping produced effect estimates in the same direction. SNPTEST v2.5.6 was utilised to perform logistic regression [319], setting -frequentist with dom

for dominant model and -method with newml, holding all else default (see https://www.chg.ox.ac.uk/~gav/snptest/ for details). The p value, effect sizes, and standard errors were reported by frequentist_dom_wald_pvalue_1, frequentist_dom_beta_1:dom/malaria=1, and frequentist_dom_se_1 in the result file, respectively. I constructed the 95% confidence interval from these metrics. This analysis was repeated for the other two variants (rs8176719, chr9:133,257,521:T>TC, and rs4951377, chr1:203,689,343:A>G), both previously reported to have large effect sizes and replicated associations with malaria severity [115].

3.7 Conclusion

This chapter evaluated genome-wide imputation accuracy and characterised specific regions of lcWGS in the GAMCC cohort. Section 3.1 introduced the study cohort and confirmed the success of experiments using the library conditions identified in Chapter 2. Section 3.2 presented a comprehensive analysis of lcWGS imputation, benchmarking results against both microarray typed variants and high-coverage data, and explored strategies to enhance imputation performance. First, I assessed the impact of incorporating additional African haplotypes into the reference panel to improve representation of endemic genetic diversity. Second, I implemented a two-stage imputation workflow, in which only high-quality, genome-wide variants were retained from initial lcWGS imputation and subsequently uploaded to the TOPMed imputation server. This method leveraged the extensive and diverse TOPMed reference panel,

which was not otherwise locally accessible, and achieved superior accuracy compared with microarray data at common variants. In Section 3.3, I demonstrated that population structure could be reliably inferred from imputed lcWGS data. When compared with results derived from microarray data, the first three PCs showed near-perfect concordance. Population inference remained robust even when combining datasets generated by different technologies. Additionally, I focused on specific regions or variants to complete the panoramic picture of lcWGS imputation. Section 3.4 benchmarked two biologically relevant subsets of variants, the GWAS catalogue variants, and variants in blood group genes, against genome-wide performance, revealing no systematic bias. I also identified the least accurately imputed region from lcWGS, a segmental duplication encompassing the FCGR genes, which is prone to mapping and alignment artefacts as well as exemplified the limitations of this approach. Section 3.5 replicated the protective effect of rs334 A allele (*HBB*), rs8176719 T allele (*ABO*), and rs4951377 A allele (*ATP2B4*) against malaria from lcWGS imputation, although statistical significance was achieved only for rs334. Collectively, this work establishes a solid foundation for the use of lcWGS with imputation as a reliable approach for genome-wide variant inference. Further discussion of the limitations and implications of this chapter is provided in Section 7.2.1.

Chapter 4 Low-Coverage Imputation in the HLA Region

In addition to evaluating genome-wide imputation performance, it is important to assess regions with complex genomic architecture and high polymorphism, especially those with established biological relevance to various diseases and phenotypes [117]. This chapter focuses on evaluating the ability of lcWGS to capture alleles in the HLA region on chromosome 6q21, which are known to influence numerous human immune traits. Section 4.1 presents the African HLA diversity observed in our data. Section 4.2 characterises the imputation performance of HLA alleles using the QUILT-HLA utility. Sections 4.3 and 4.4 describe improvements to the original method from two perspectives: assembling an Gambian-enriched reference panel and refining the imputation workflow. These efforts culminate in Section 4.5, which presents the optimal HLA imputation results and compares them with server-based imputation from microarrays.

4.1 Gambian HLA diversity

In addition to lcWGS and genotyping data, all samples were sent to HistoGenetics for sequence-based typing to obtain true HLA types. Figure 4.1.1 illustrates the unique haplotype structure and HLA allelic diversity in the Gambian cohort (allele frequency < 5% grouped as *Other*) after phasing the alleles into haplotypes (see Section 4.3 for

details). The most common class I haplotype is HLA-A*26:01-C*03:04-B*08:01 with 1.7% frequency, known to partly associate with idiopathic hypoparathyroidism in an India population [324]. Both the HLA-DRB1*13:04 (allele frequency = 18.3%) and HLA-DQB1*03:19 (33.3%), along with the HLA-DRB1*13:04-DQB1*03:19 haplotype (12.4%), are particularly common in the Gambian population, although they do not appear to be in strong LD ($r^2 = 0.122$). Notably, the HLA-DRB1*13:04 allele is particularly underrepresented in the 1000 Genomes Project reference panel, resulting in 0.84% globally, 3.09% among all African populations, and 15% in the Gambia. These findings underscore the complexity of accurately imputing HLA alleles in populations with unique genetic backgrounds.

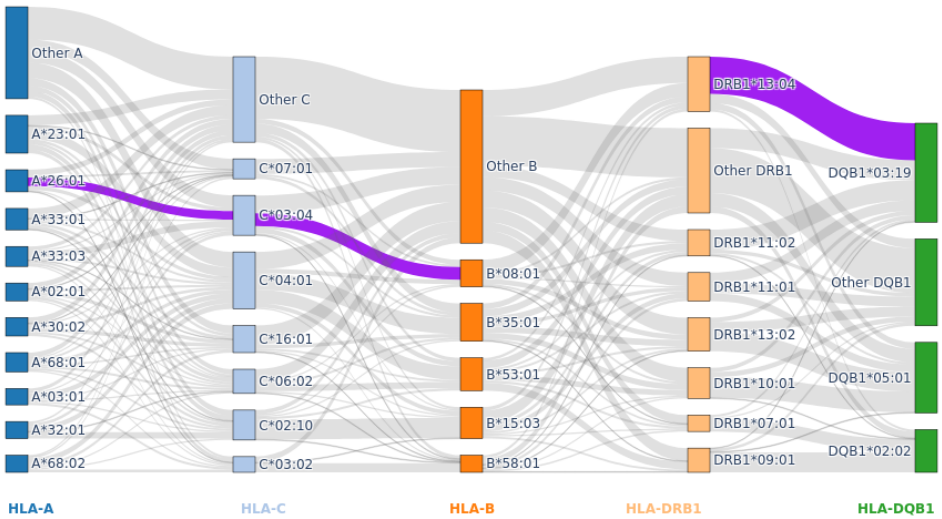


Figure 4.1.1. HLA diversity in the GAMCC cohort. The HLA alleles are obtained from sequence-based typing, whereas phasing on these alleles is conducted in Section 4.3 with a custom method. The length of the bars corresponds to the observed allele frequency observed. Grey lines connect the haplotype structure and are also proportional to the haplotype frequency, with the most dominant HLA class I/class II haplotypes coloured purple.

4.2 HLA imputation performance with QUILT-HLA

To infer the HLA alleles from lcWGS, I ran QUILT-HLA, which is part of the QUILT software, on the bam files of all GAMCC samples. For a specific gene, it leverages LD-based inference outside the gene using a reference panel where HLA alleles are phased onto SNP haplotypes and read-based imputation within the gene by aligning sequencing reads to nucleotide sequences from the IPD-IMGT/HLA database, combines the likelihood from the two parts, and emits posterior probabilities for an individual diplotype (Figure 4.2.1; Methods).

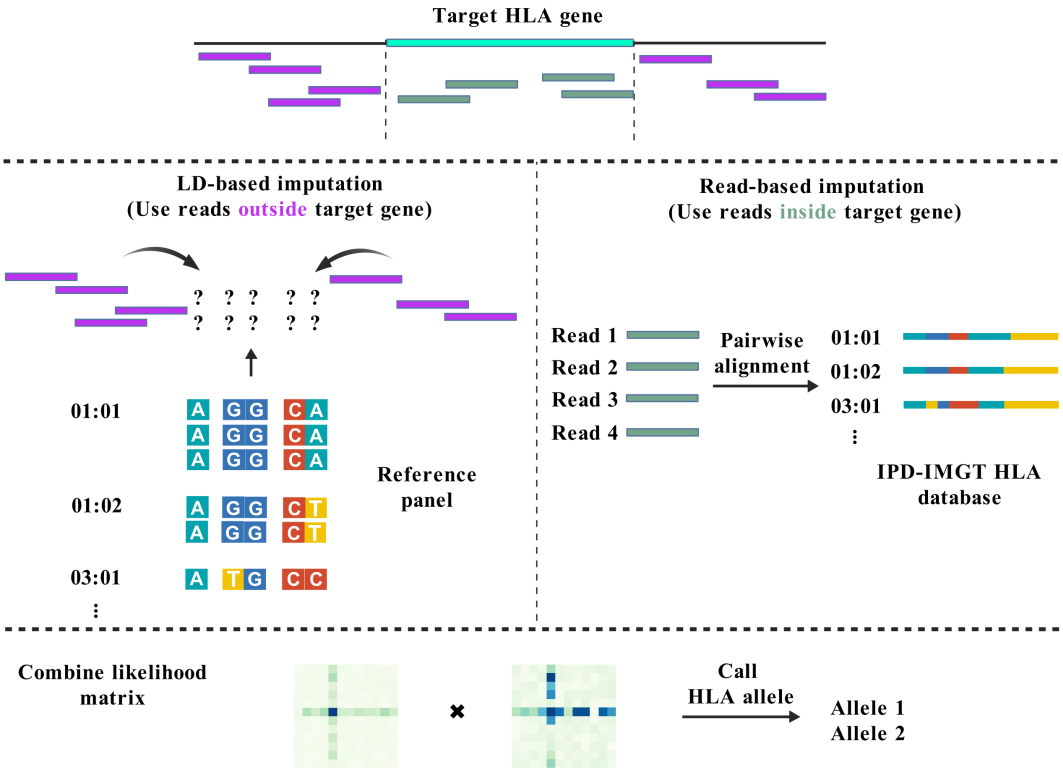


Figure 4.2.1. The QUILT-HLA workflow. This figure was created with BioGDP.com [44]. The method partitions reads aligned to the target HLA region based on their genomic positions. Reads outside the gene (purple) undergo QUILT imputation and are compared probabilistically to SNP haplotypes in the reference panel. Reads within the gene (green) are re-aligned to all HLA alleles (up to 4-field resolution) from the IPD-IMGT/HLA database due to limited HLA representation in GRCh38. Alignment scores are averaged across higher-resolution alleles to form likelihoods, and the two likelihood matrices are multiplied to produce final HLA allele calls.

I focused the analysis on the five most polymorphic HLA genes, HLA-A, -B, -C, -DQB1 and -DRB1. I ran QUILT-HLA to impute lcWGS data with the 1000 Genomes Project reference panel and calculated concordance at 2-field resolution compared to true HLA calls by sequence-based typing. Figure 4.2.2 illustrates variable imputation accuracy across HLA alleles. While imputation is accurate at the HLA-B, -C, and -DQB1 genes (mean concordance = 0.936, 0.974, and 0.979, respectively), certain common alleles at HLA-A and HLA-DRB1 show reduced accuracy (0.836 and 0.588, respectively). Notably, HLA-DRB1*13:04 (allele frequency = 18.3% in truth data) shows only 24.7% concordance. Several common HLA-A alleles, HLA-A*33:01, A*33:03, and A*68:02 (allele frequencies = 7.14%, 6.67%, and 5.71%), also exhibited low concordance (26.7%, 53.6%, and 58.3%, respectively). This is partly due to underrepresentation of these alleles in the 1000 Genomes Project reference panel (for example, HLA-DRB1*13:04). Additional contributing factors may include limitations in the QUILT implementation and the inherent structural complexity at specific HLA loci [159, 325, 326]. Given the limited imputation performance, I aimed to improve the HLA inference by constructing a Gambian-enriched 1000 Genomes Project reference panel and refining the QUILT-HLA imputation workflow in the following sections.

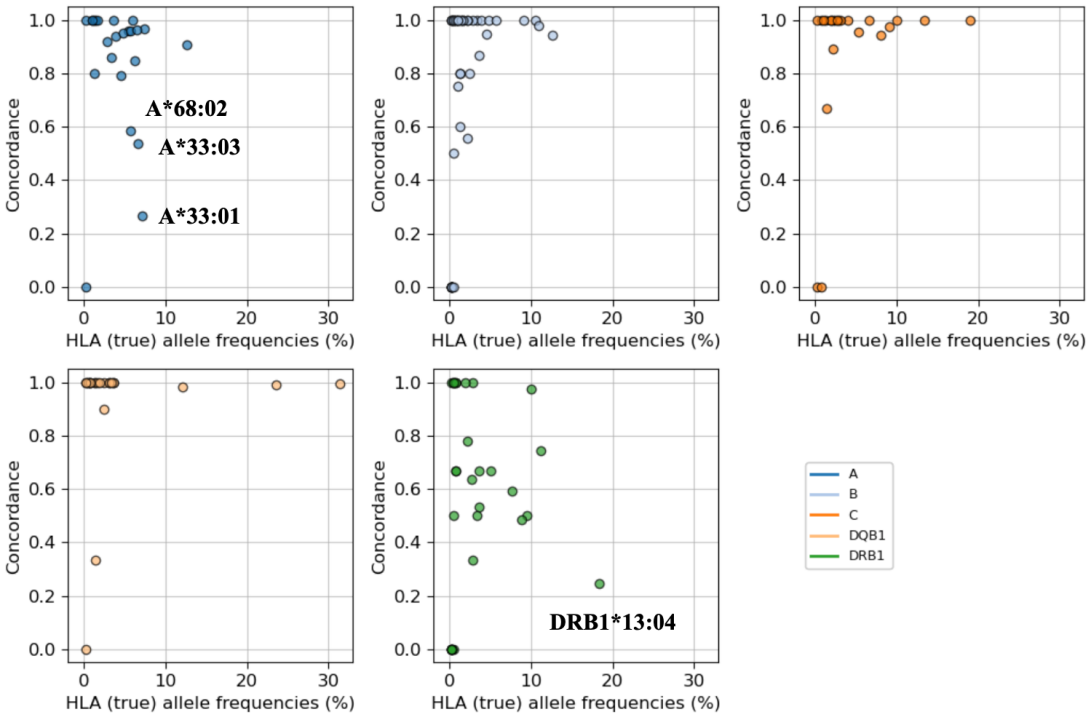


Figure 4.2.2. Variable performance of basic QUILT-HLA imputation with 1000 Genomes Project reference panel across five classical genes at 2-field resolution. The x-axis is the allele frequency observed in the truth data across GAMCC samples. The y-axis is the imputation concordance of that specific allele. Colours reflect different HLA genes. Four common alleles but have compromised imputation accuracy are marked on the figure.

4.3 An African HLA reference panel

4.3.1 An HLA phasing approach

The suboptimal imputation results, together with concerns about the reference panel and methodology, highlighted the need for further exploration of the potential of lcWGS to accurately capture HLA alleles. The original QUILT-HLA method partitions reads into two sets, applies LD-based imputation outside the target gene and read-based imputation within it, and combines the results afterwards. However, the underrepresentation of Gambian individuals necessitates the development of a Gambian-enriched 1000 Genomes Project reference panel for lcWGS, which, to our

knowledge, currently does not exist.

Combining the current dataset with the 1000 Genomes Project reference panel required phasing the true HLA calls (from HistoGenetics) onto SNP haplotypes (from QUILT imputation). To achieve this goal, I implemented a custom method inspired by the original QUILT framework and subsequently merged the phased haplotypes into the 1000 Genomes Project reference panel. Full methodological details are provided in the Methods section. The method takes the true HLA calls and phased SNP haplotypes as inputs and determines the correct phasing by evaluating two possible configurations. The intuition behind this method is simple: each pair of HLA alleles can be phased onto SNP haplotypes in only two possible ways. By matching the genomic positions of the two SNP haplotypes to the true HLA allele sequences obtained from the IPD-IMGT/HLA database, the configuration resulting in fewer nucleotide mismatches is considered the correct phasing.

To implement this, I first intersected the genomic positions of the SNP haplotypes with HLA database variant sites within each target gene. The total number of nucleotide mismatches for both phasing options was then calculated by comparing the parental haplotype sequences with the true HLA sequences documented in the database, subjected to a strict criterion to determine whether a sample was successfully phased to enhance reliability. For unphased samples, this comparison was iteratively extended by

50 bases beyond the target region on both sides until a pre-specified limit was reached or all samples were successfully phased. In this second stage of phasing, rather than comparing to sequences extracted from the database, I compared to the average sequences of already phased samples carrying the same HLA allele, since nucleotide sequences outside the target gene are not available in the database. Figure 4.3.1 and 4.3.2 further elucidate this phasing procedure with two examples.

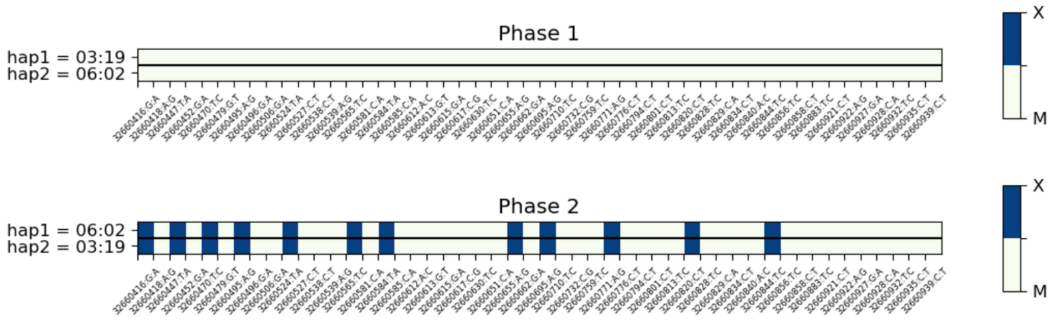


Figure 4.3.1. An example of the phasing procedure at part of HLA-DQB1. Sequence-based typing determined this individual carries HLA-DQB1*03:19 and HLA-DQB1*06:02. The x-axis represents variant sites within the gene, where mismatches are shown in blue (X) and matches in white (M). Phase 1 represents no mismatch, indicating that haplotype 1 corresponds exactly to HLA-DQB1*03:19 and haplotype 2 to HLA-DQB1*06:02, thus confirming the correct phasing. In contrast, phase 2 exhibits numerous mismatches, suggesting an incorrect configuration. Overall, the correct phasing with zero mismatches clearly outperforms the alternative configuration, which shows 24 mismatches, confirming accurate phasing at the first-stage comparison.

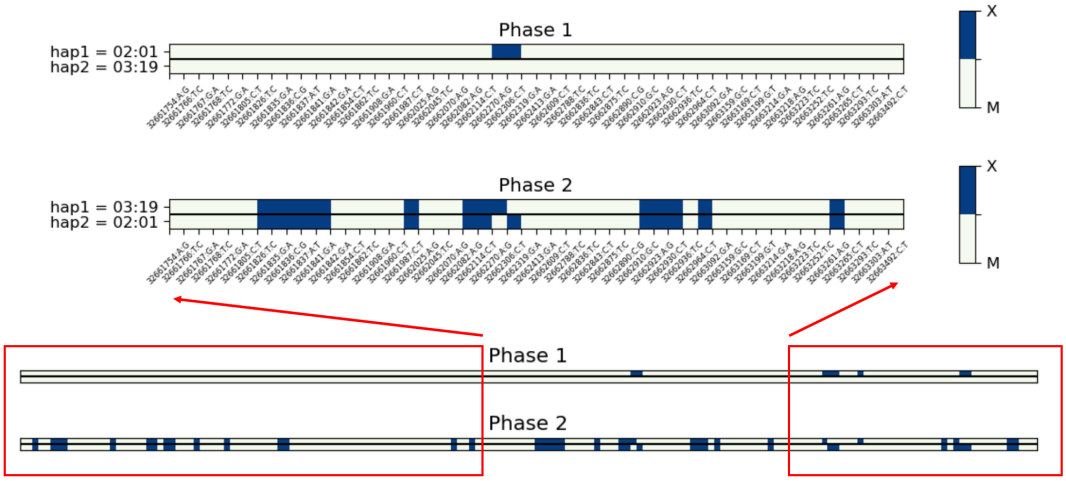
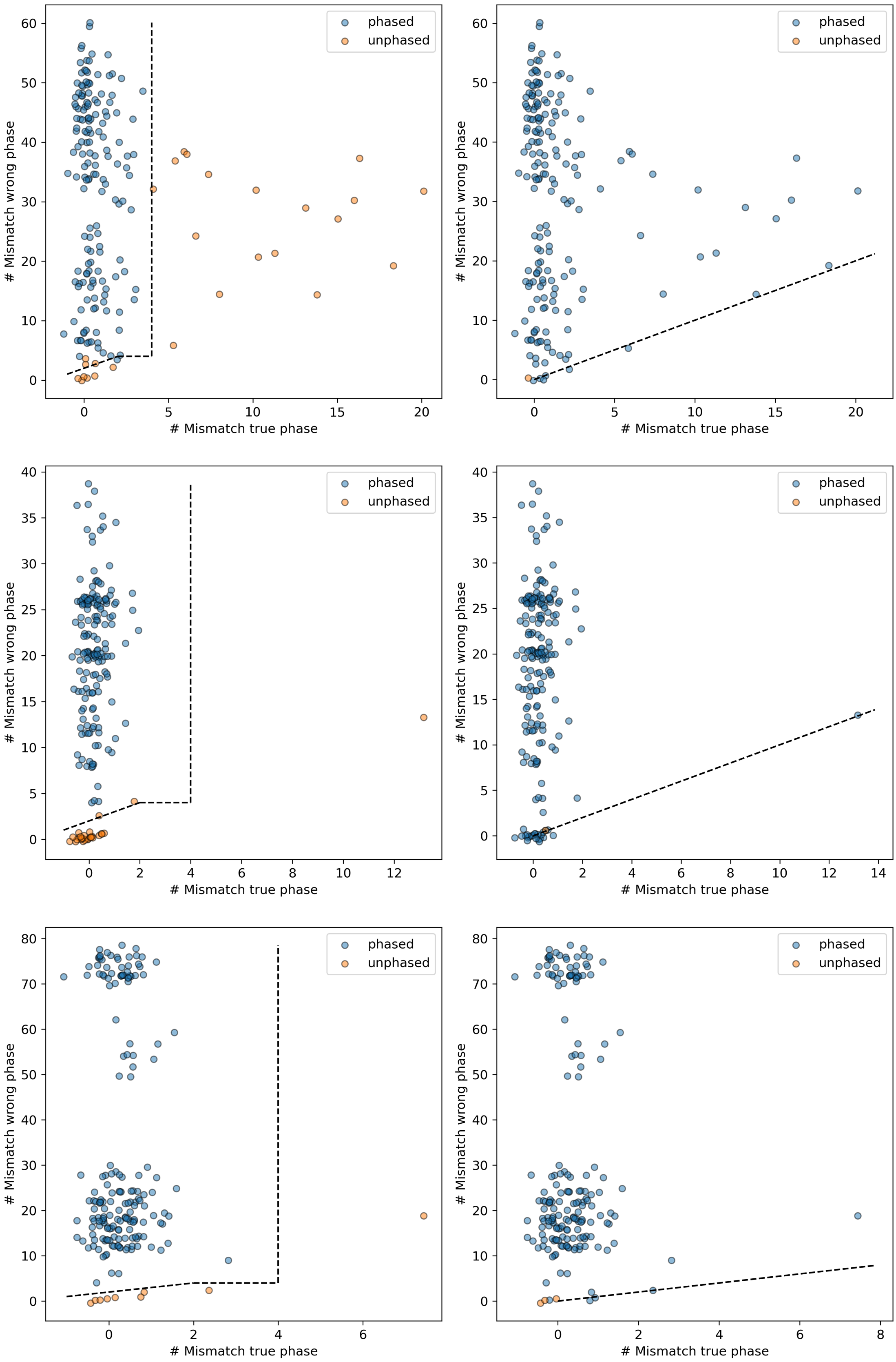


Figure 4.3.2. Another example of the phasing procedure at HLA-DQB1, extending outside the gene. This example illustrates a case where the first-stage comparison fails to phase the individual (top panel), carrying HLA-DQB1*03:19 and HLA-DQB1*02:01. Although phase 2 exhibits many more mismatches, the plausible phasing 1 also contains two mismatches. As the method applies a strict filter requiring the correct phase to have fewer than two mismatches, this sample is deemed unreliable and advanced to the second-stage comparison (bottom panel). In this stage, the analysis extends 50 bases upstream and downstream of the HLA-DQB1 gene (indicated by the circled regions), again suggesting that phase 1 is the more likely configuration. With relaxed phasing criteria at this stage, the final assignment determines haplotype 1 as HLA-DQB1*02:01 and haplotype 2 as HLA-DQB1*03:19.

Mismatches observed in the correct phasing may arise from switch errors, inaccurate genotype imputations, or imperfections within the HLA reference database. Nevertheless, the method demonstrates robustness to such errors by jointly considering a large number of variant sites. Figure 4.3.3 summarises the phasing results for all samples across all HLA loci.


phased
unphased
Mismatch wrong phase
Mismatch true phase
phased
unphased
Mismatch wrong phase
Mismatch true phase
phased
unphased
Mismatch wrong phase
Mismatch true phase
phased
unphased
Mismatch wrong phase
Mismatch true phase
phased
unphased
Mismatch wrong phase
Mismatch true phase
phased
unphased
Mismatch wrong phase
Mismatch true phase

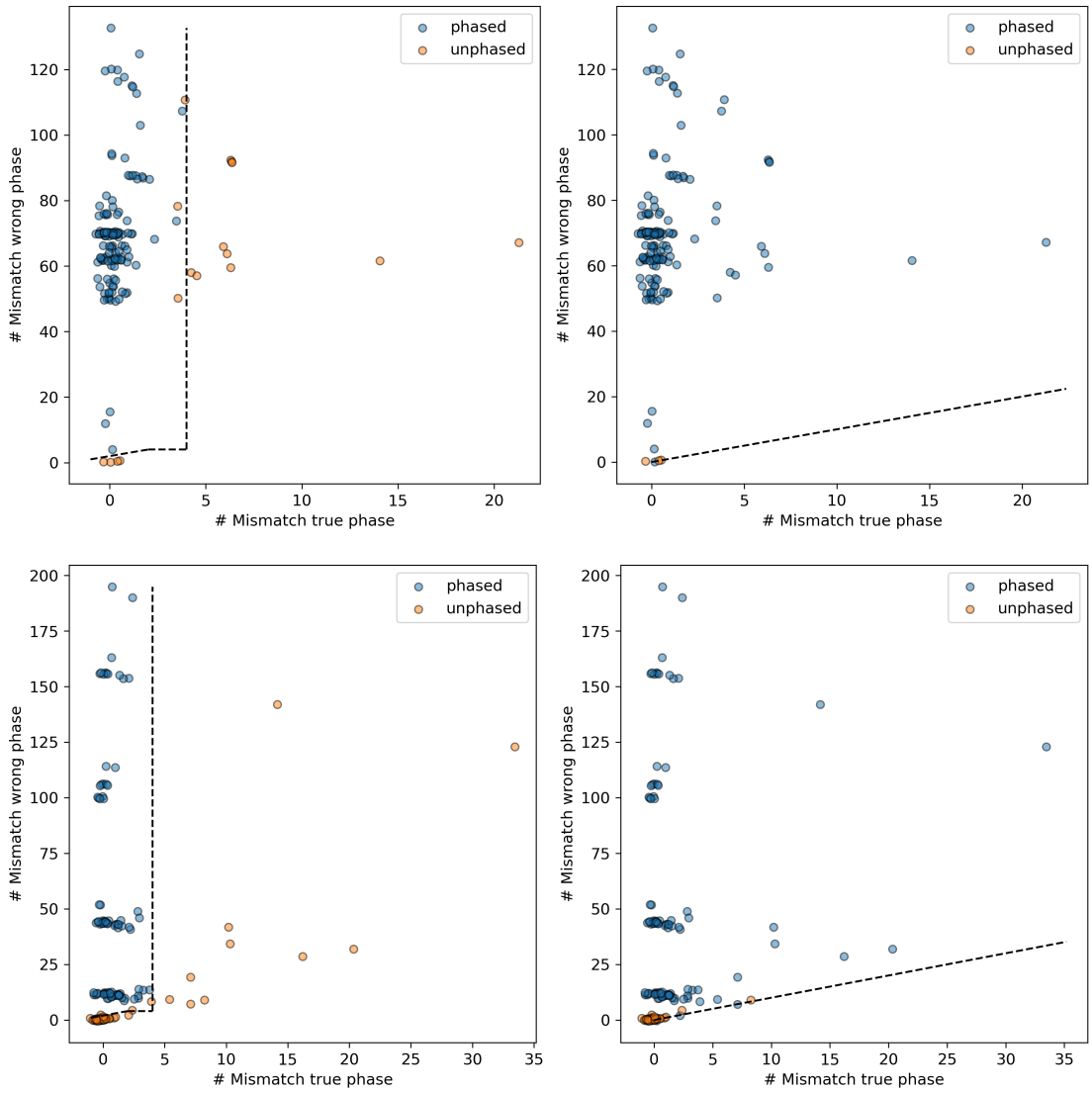


Figure 4.3.3. Phasing results at the HLA loci. From top to bottom, panels correspond to HLA-A, -B, -C, -DQB1, and -DRB1. Taking HLA-A as an example, the left panel shows the first-stage phasing results, with phased samples coloured blue and unphased samples coloured yellow. The dashed line represents the phasing criteria, with samples above/to the left considered phased. The x-axis and y-axis indicate the number of mismatches for the two possible phasings, with the plausible true phasing (smaller values) plotted on the x-axis. Markers are slightly jittered for visualisation. The right panel shows the final phasing results after the second-stage procedure. The other four genes followed the same pattern.

Individuals with homozygous HLA alleles did not require phasing, and the method successfully phased the majority of remaining individuals at all loci. Samples that remained unphased were subjected to BEAGLE phasing. Table 4.3.1 summarises the

number of individuals phased at each step. Ultimately, one individual could not be phased at the HLA-DRB1 locus due to missing truth data (from sequence-based typing), and was therefore removed. The resulting phased dataset was then combined with the 1000 Genomes Project samples to construct the Gambian HLA reference panel.

HLA loci	Number of heterozygous individuals	Number phased by custom method	Number phased by BEAGLE	Number unphased
HLA-A	187	186	1	0
HLA-B	190	189	1	0
HLA-C	187	184	3	0
HLA-DQB1	161	158	3	0
HLA-DRB1	186	163	22	1

Table 4.3.1. Number of individuals phased in each step of reference panel construction. Only individuals with heterozygous HLA alleles were included, as homozygous alleles did not require phasing.

4.3.2 HLA inference by improved reference panel

To evaluate the additional benefit conferred by the Gambian-enriched 1000 Genomes Project HLA reference panel, I ran QUILT-HLA imputation using a leave-one-out approach, and the resulting improvements in imputation confidence and concordance were compared with the previous run. As this modification primarily affected LD-based imputation, the analysis focused on the results obtained from QUILT state inference. Figure 4.3.4 demonstrates that the enhanced reference panel has modest impact on the imputation accuracy for most alleles. Although notable differences are observed for rare alleles, these largely reflect the limited number of individuals carrying those alleles. However, Figure 4.3.4b shows that the enriched reference panel improves imputation confidence, as measured by the posterior probability of HLA calls. Across 1,050

imputation jobs (210 individuals times five HLA loci), 460 (43.8%) showed increased confidence while 213 (20.1%) showed decreased confidence. The remaining cases exhibited changes smaller than 0.01 and were excluded from this histogram. Overall, the impact of the enriched reference panel on HLA inference was not substantial, likely reflecting the relatively small number of additional individuals included compared to the original 1000 Genomes Project reference panel (210 of 2,568, 8.18%).

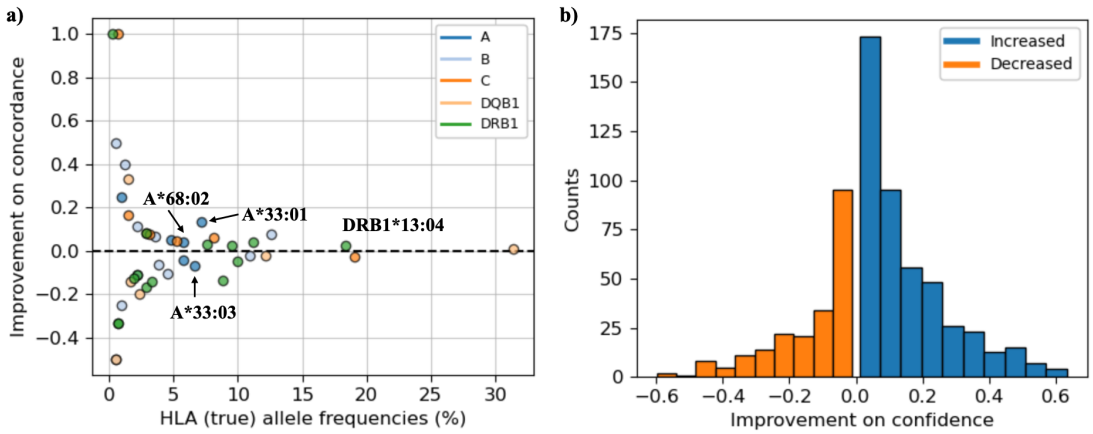


Figure 4.3.4. QUILT-HLA state inference by the Gambian-enriched 1000 Genomes Project reference panel. a) Improvements on HLA inference concordance (y-axis, negative values for decreased accuracy) against allele frequencies (x-axis), coloured by HLA loci. Only alleles with different imputation results are shown in this figure. **b)** Histogram of changes of imputation confidence for each individual and each loci (absolute difference smaller than 0.01 was excluded), coloured by increased (blue) and decreased (orange) confidence.

4.4 Refinement of the QUILT-HLA imputation workflow

4.4.1 Replacing k-mer alignment with Wavefront aligner

In the original QUILT read-based imputation framework, reads mapping to a target gene are aligned and filtered against all HLA alleles in the database using only four 10-mers per read. While this improves computational efficiency, it discards much of the

sequence information and increases the risk of missing variants that distinguish closely related alleles, further limiting an already sparse dataset and complicating imputation. To address this, I modified the workflow by re-aligning all lcWGS reads to all HLA sequences from the IPD-IMGT/HLA database using the Wavefront aligner in a pairwise manner. Although this strategy substantially increased computational demands, it maximised data retention and fully exploited the information contained in each read.

I employed a Python implementation of Wavefront aligner [327, 328] with the default scoring scheme (match = 0, mismatch = -4, gap opening = -6, and gap extension = -2) and specified min_aligned_bases_right = 10 to trim short matches at the end of alignments to pursue an approximate local alignment. Alignment scores for individual reads were calculated as the sum of base contributions, adjusted by error rates inferred from base quality scores, and floored by -4 times the read length to avoid numeric overflow. When both reads in a pair were retrieved, their alignment scores were summed to represent the underlying DNA molecule. Read pairs were subsequently discarded, to minimise spurious reads from potentially paralogous genes, if their best alignments failed to reach the minimum threshold, if neither read mapped uniquely to the target gene, or if the alignment score did not sufficiently exceed that of a secondary alignment. As alignment scores can also be interpreted as log-likelihoods [329], they were normalised by averaging across higher-resolution alleles (to 2-field) and incorporated into the main QUILT workflow (LD-based imputation). Further details are

provided in the Methods section.

Although sacrificing computational performance, this modified workflow retrieved more informative reads than the original QUILT-HLA method (Figure 4.4.1). A read was considered informative as long as one of its best aligned HLA alleles match one of the two called HLA alleles from sequence-based typing at 2-field resolution. The modified read-based imputation retained more informative reads for almost all samples at HLA-A, HLA-B, HLA-C, and HLA-DQB1 as compared to the original workflow. However, alignment at HLA-DRB1 remains challenging, potentially due to the high sequence homology with the other HLA-DRB genes and pseudogenes [325, 326, 330-332].

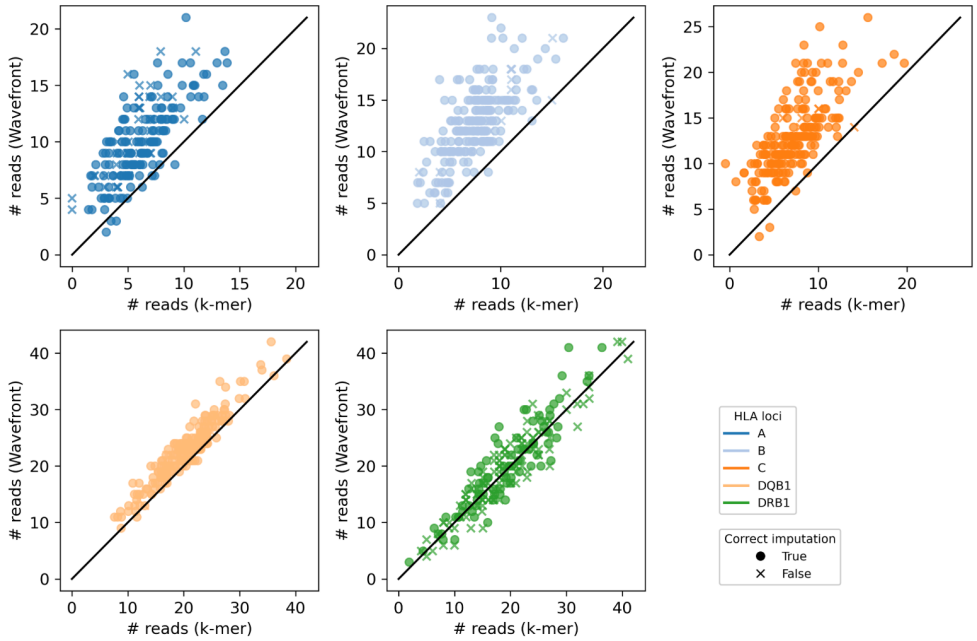


Figure 4.4.1. Modified QUILT-HLA read-based imputation with Wavefront aligner retrieves more informative reads. The figure is divided into five panels corresponding to different HLA genes. The x-axis (jittered for visualisation) shows the number of true reads retrieved by the original k-mer method, and y-axis shows reads retrieved by the new method. Each sample is marked with a circle if both HLA alleles are correctly imputed by the new approach, and with a cross otherwise.

4.4.2 HLA inference by improved read alignment

In Section 4.3.2, I assessed the marginal impact of the improved read alignment step on HLA inference relative to the default QUILT-HLA results. Because the original QUILT software did not output the likelihood matrix from its read-based inference, I modified the software to extract this information. Figure 4.4.2 illustrates substantial improvements resulting from the more robust alignment approach, in contrast to the original k-mer method, which is prone to significant data loss. Overall, read-based imputation improved for 81 out of 99 alleles (81.8%) across all HLA loci, consistent with the enhancements observed in Figure 4.4.1 for constructing a more precise likelihood matrix. Finally, I applied the improved QUILT-HLA method in combination with the Gambian-enriched 1000 Genomes Project reference panel to generate the most accurate HLA imputations for our data in the concluding section of this chapter.

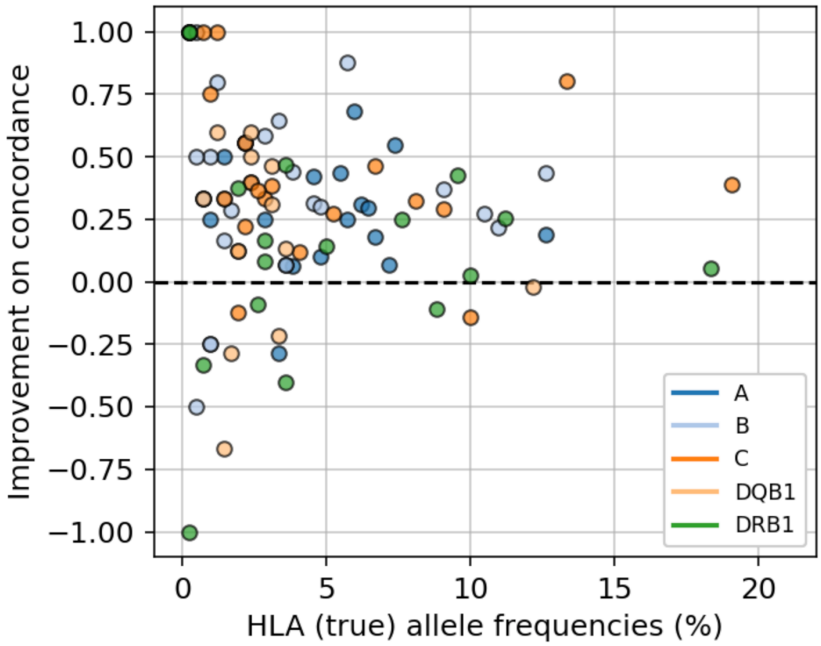


Figure 4.4.2. QUILT-HLA read-based inference by Wavefront aligner. Improvements on HLA inference concordance (y-axis, negative values for decreased accuracy) against allele frequencies (x-axis), coloured by HLA loci. Only alleles with different imputation results are shown in this figure.

4.5 The optimal HLA imputation performance

Efforts in Section 4.3 and 4.4 led to improved lcWGS HLA imputation results compared to Section 4.2, as illustrated in Figure 4.5.1. The modified QUILT-HLA approach, which replaces k-mer alignment with the Wavefront aligner and leverages the Gambian-enriched 1000 Genomes Project reference panel, achieved increases of 1.67%, 0.71%, 2.14%, 0.71%, and 6.70% for HLA-A, -B, -C, -DQB1, and -DRB1, respectively (Figure 4.5.1b). Among the 51 HLA alleles where imputation performance changed, 34 (66.7%) showed improvements, with notable gains for HLA-DRB1*13:04 (4.62%) and HLA-A*33:01 (20%). Imputing HLA-DRB1 is particularly challenging due to its complex LD patterns extending over long genomic distances as well as the presence of multiple paralogous pseudo- and non-pseudogenes, which have also been reported in many previous studies [66, 325, 326, 332-335].

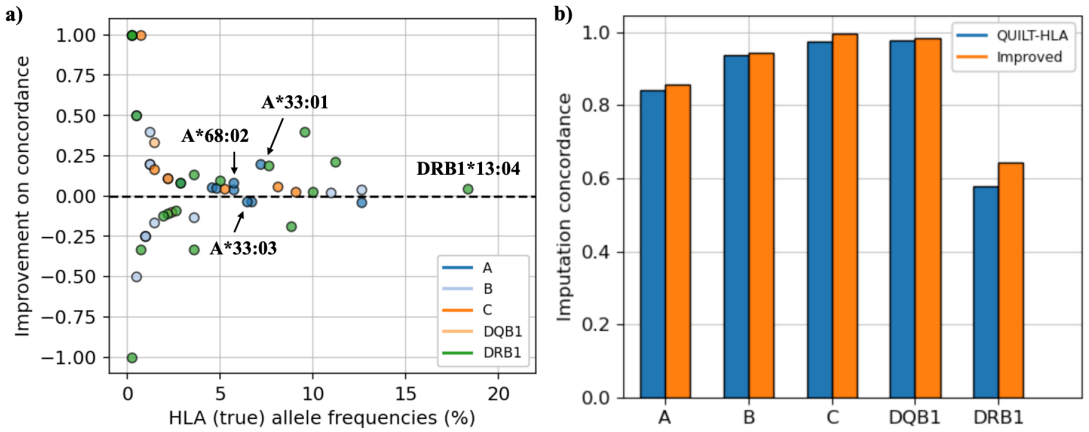


Figure 4.5.1. Improved HLA imputation accuracy with the modified QUILT-HLA workflow using the Gambian-enriched 1000 Genomes Project reference panel. a) Improvements on HLA inference concordance (y-axis, negative values for decreased accuracy) against allele frequencies (x-axis), coloured by HLA loci. Only alleles with different imputation results are shown in this figure. **b)** Imputation concordance for all individuals at all loci, coloured by the default QUILT-HLA (blue) or improved method (orange).

Although the modified QUILT-HLA method and the Gambian-enriched HLA reference panel improved imputation accuracy, results for HLA-A and HLA-DRB1 remain suboptimal for certain alleles. Notably, a multi-ethnic HLA reference panel, comprising 21,546 individuals from five ethnic groups, is available through the Michigan imputation server [157, 158] but cannot be obtained locally. Emulating the two-stage imputation approach in genome-wide imputation (Section 3.2.3), I leveraged this large HLA multi-ethnic reference panel for lcWGS data by retaining only high-confidence SNPs from the optimal genome-wide imputation (two-stage QUILT imputation with MalariaGEN-enriched 1000 Genomes Project reference panel followed by TOPMed) and uploaded to the Michigan server. This allowed benchmarking against HLA imputation performed on our microarray data using the same reference panel. Details of the imputation procedure are provided in Methods.

Figure 4.5.2 compares three imputation workflows across two data types and reference panels. For each locus, bars on the left represent imputations from the HLA multi-ethnic reference panel, either starting with microarray (dark blue) or imputed lcWGS data (light blue). As expected, microarray-based imputation consistently outperforms lcWGS, reflecting imperfect genome-wide SNP imputation from lcWGS. Bars on the right compares two different imputation workflow starting with lcWGS data, that is, leveraging the two-stage approach with the HLA multi-ethnic reference panel (light blue) or directly with the modified QUILT-HLA method using previously constructed

Gambian-enriched 1000 Genomes Project reference panel. This comparison reveals different patterns at different loci: QUILT-HLA imputation outperforms the HLA multi-ethnic reference panel at HLA-C (6.7%) and HLA-DQB1 (13.6%), is comparable at HLA-B (-0.5%), and is less well-represented at HLA-A (-14.3%) and HLA-DRB1 (-30.1%). Several points are worth highlighting in this comparison. First, unlike all previous comparisons, concordance reported in this graph was measured at G-group resolution (same antigen recognition domain) rather than 2-field resolution (same whole protein), as the HLA multi-ethnic reference panel by default reports G-group allele calls (although there is an important caveat on the way alleles are reported, see text below the figure and the Methods section). An advantage of QUILT-HLA is then by imputing HLA alleles at higher resolution, which may harbour significant genomic variations that are otherwise omitted. Second, QUILT-HLA achieves better imputation results at some locus with a considerably smaller reference panel (12.9% in number of individuals), illustrating its effective design by combining read-based imputation and LD-based imputation to fully leverage the sparse lcWGS data. The reduced imputation accuracy at HLA-A and HLA-DRB1 is likely attributable to the underrepresentation of this population in the relatively small reference panel. Implementing QUILT-HLA on the imputation server could further improve HLA imputation in the future. Overall, by combining multiple workflows, > 90% accuracy was achieved across all five HLA loci in this population, demonstrating that lcWGS can provide reliable HLA inference without additional cost.

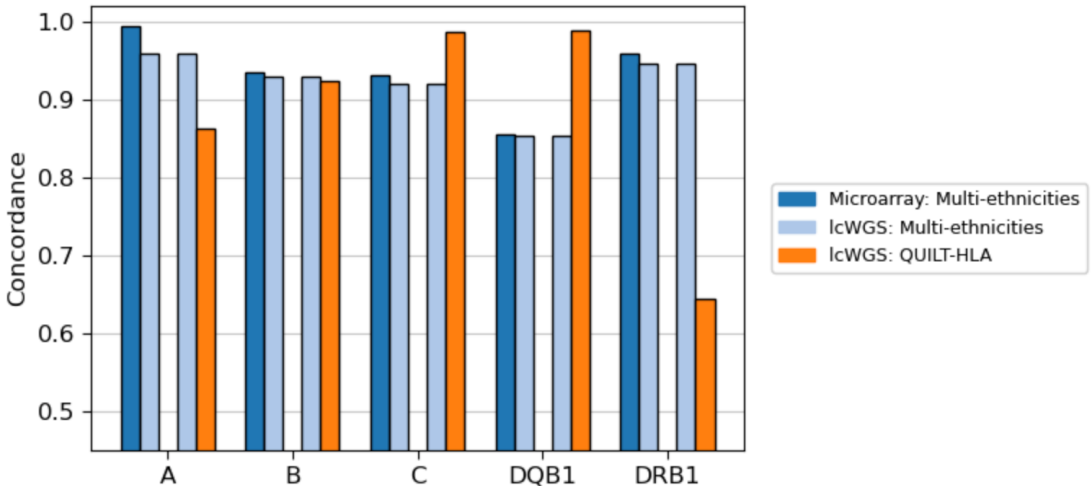


Figure 4.5.2. HLA imputation accuracy comparison. For each HLA locus on the x-axis, concordance (calculated at G-group resolution) from three different runs is shown as a bar plot. Dark blue, light blue, and orange represent HLA multi-ethnic panel imputation on microarray data, two-stage HLA multi-ethnic panel imputation on QUILT imputed lcWGS data, and the optimal QUILT-HLA imputation on lcWGS data, respectively. Each locus is separated as two parts with the light blue bars duplicated to facilitate and reinforce comparisons. Bars on the left share the same HLA inference reference panel (HLA multi-ethnic reference panel); bars on the right are all derived from lcWGS data.

4.6 Methods

HLA sequence-based typing. All lcWGS samples were sent to HistoGenetics for sequence-based typing, which selectively amplifies the HLA region and calls HLA types by sequencing. The original results reported alleles at mixed resolutions (2, 3, or 4-field), but I collapsed all calls to 2-field or G-group resolution required by different comparisons [129, 132]. Some alleles remained ambiguous at this level, and these were retained so.

HLA imputation of lcWGS data. HLA imputation on lcWGS data with QUILT-HLA occurs in two steps. In the preparation step, the software generates several auxiliary

files it needs to proceed. I obtained the HLA reference panel from the 1000 Genomes Project website (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HLA_types/20181129_HLA_types_full_1000_Genomes_Project_panel.txt) with the same vcf file and recombinant rates described above [155]. Only 2,568 samples of the 1000 Genomes Project reference panel have paired HLA typing and SNP haplotypes with which I proceeded forward. I obtained HLA allelic sequences from the IPD-IMGT/HLA database (https://github.com/ANHIG/IMGTHLA/, release v3.57.0) [129, 336, 337]. I ran QUILT_HLA_prepare_reference with default settings and imputed lcWGS data using the QUILT-HLA utility with a minor bug fix (commit bcbcb67, https://github.com/Suuuuuuuus/QUILT/tree/hla_rl_fix). For all comparisons, I extracted HLA diplotypes and the associated posterior probability for each individuals from the *combined* output that leverages information from both reads aligned to the HLA loci and LD (see https://github.com/rwdavies/QUILT/blob/master/example/QUILT_hla_reference_panel_construction.Md for details).

The original QUILT-HLA workflow. The QUILT-HLA workflow is briefly described here, with full details available in the original publication [66] and on the GitHub repository. QUILT-HLA leverages two sources of information. On one hand, it infers the target HLA gene using reads outside the gene (in terms of genomic coordinates) by LD and compares with the reference panel. On the other hand, it aligns reads inside the genes to all alleles in the IPD-IMGT/HLA database (2, 3, or 4-field) by comparing four

10-mers extracted from each read (two consecutive 10-mers from each side of a read, excluding the starting and ending sequences; for 151 bp reads, the four 10-mers are sub-sequences at position 11-20, 21-30, 122-131, 132-141). After a filtering step to remove unconfidently or ambiguously aligned reads, it calculates the likelihood by comparing the reads and the aligned allelic sequences, along with base qualities associated to the read accounting for possible base-calling errors. In the last step, it deresolutes read-based likelihood to 2-field resolution by averaging across higher resolution alleles, intersects alleles appeared in LD-based imputation, and finally multiplies the likelihood to call HLA diplotypes. An internal assumption of this workflow is that all HLA alleles are fully represented in both the database and the reference panel, as neither currently assigns likelihoods to unseen alleles. This limitation is generally not a practical concern, since the reference panel and HLA database typically provide comprehensive allele coverage, though future improvements could further address this issue.

HLA imputation assessments. HLA imputation accuracy was quantified using concordance (at 2-field resolution unless otherwise specified), defined as the proportion of correctly imputed HLA alleles for each gene. When true HLA calls from sequence-based typing were ambiguous (for example, HLA-DQB1*02:02/02:156), an imputed allele was considered concordant if it matched any of the possible alleles. This approach was unlikely to substantially inflate accuracy, since the alternative alleles were generally rare.

Improving on HLA LD-based imputation: a phasing approach to construct the Gambian HLA reference panel. The method required the true HLA allele calls from sequence-based typing and corresponding SNP haplotypes obtained via QUILT imputation with the MalariaGEN-enriched 1000 Genomes Project reference panel as input. If a sequence-based allele call was ambiguous at 2-field resolution, the ambiguity was resolved by selecting the allele with the smallest second field numeric identifier, as alternative alleles are generally rare. Variant sites are first identified from multiple sequence alignments of all alleles for a target gene, as provided by the IPD-IMGT/HLA database [129, 336, 337]. These variants were then matched with observed SNP haplotypes by chromosome coordinates. I counted the number of mismatches between SNP haplotypes and HLA alleles (from the database in the first-stage phasing and averaging across potentially ambiguous alleles in the second-stage phasing) and summed over the two possible phasing. A mismatch was called if the difference of two probabilities was greater than 0.9. I then proceeded with a two-step approach. In the first step (first comparison), I applied a strict comparison criteria by designating a phasing as correct if it had fewer than four mismatches and either outperformed the alternative by at least two mismatches or the alternative had more than four mismatches. Unphased individuals proceed to the second step, where I averaged empirical SNP haplotypes observed across confidently phased samples sharing the same alleles and compared them to unphased SNP haplotypes. The comparison criteria were relaxed by only requiring one phasing improves from the other to designate the phase of an

individual. This process iteratively extended the phasing region by 50 base pairs until either all samples were phased or a predefined maximum of 1,000 bases was reached.

For samples remaining unphased, I applied LD-based phasing using BEAGLE v5.4 emulating the SNP2HLA approach to maximise sample retention [92, 166]. Correct BEAGLE phasing was determined by comparing the closest upstream and downstream heterozygous SNPs with the QUILT result. Still unphased samples were ultimately removed from the reference panel.

Improvement on HLA read-based imputation: alignments with Wavefront aligner. The original QUILT-HLA workflow combines alignment and filtering using a k-mer approach. Specifically, four 10-mer sequences at the 11-20, 21-30, 121-130, and 131-140 bases of each read (assuming a predefined read length of 150 bases, which was later fixed with commit bcbcb67 to allow varying read lengths, https://github.com/Suuuuuuuus/QUILT/tree/hla_rl_fix). However, this strategy overlooks most base information, leading to over-filtering and suboptimal alignments. To address this issue, I replaced the k-mer strategy with Wavefront aligner, applying to each read-allele pair using default scoring parameters (match = 0, mismatch = -4, gap opening = -6, and gap extension = -2) [327, 328]. I also set min_aligned_bases_right = 10 to trim short matches at the end of alignments to pursue an approximate local alignment in this scenario. Based on alignment positions, I calculated an alignment

score by summing that of each base, adjusted by error rates derived from the associated base quality, and floor by -4 times the read length to prevent numeric overflow. The results were compiled into a matrix with dimension corresponding to the number of reads and all alleles in the database. A read pair was discarded if either of the followings was true: 1) the best alignment of each read as well as jointly was worse than a predefined alignment score (equivalent to five mismatches with uniform base quality 30); 2) neither read was uniquely aligned to the target gene; 3) the sum of the alignment score for the pair did not improve upon the secondary alignment by a score equivalent to two mismatches (-8). The final results were normalised and integrated with the main QUILT workflow (LD-based imputation). This modification was implemented in Python, with corresponding changes made in QUILT (hla_functions.R) to enable reading the updated results. The modified code is publicly available on my GitHub page (https://github.com/Suuuuuuuus/QUILT_sus/).

Imputation with the HLA multi-ethnic reference panel via Michigan imputation server. I obtained a manifest of variants in GRCh37 coordinates from the HLA multi-ethnic reference panel (22,733 variants from chr6:27,970,031-33,965,553), which was subsequently lifted over to GRCh38 (17,245 retained). The optimal lcWGS imputation result (two-stage QUILT imputation with MalariaGEN-enriched 1000 Genomes Project reference panel followed by TOPMed) and microarray data were intersected with these genomic positions, before uploading to the Michigan imputation server for HLA

imputation using the 4-digit Multi-ethnic HLA reference panel v2 and Eagle v2.4 for phasing. Notably, this imputation was performed at G-group resolution but was incorrectly coded as 2-field by selecting the first 2-field allele from each G-group (https://genepi.github.io/michigan-imputationserver/pipeline/). For example, if the true HLA 2-field allele was HLA-DQB1*03:19, the server reported the results as HLA-DQB1*03:01, as HLA-DQB1*03:19 belonged to the G-group HLA-DQB1*03:01:01G.

4.7 Conclusion

In this chapter, I explored the capability of inferring HLA alleles with lcWGS in the Gambian population. The original QUILT-HLA utility, as described in Section 4.2, exhibited sub-optimal performance due to limited allele representation in the 1000 Genomes Project reference panel and the k-mer alignment process. To address these issues, I constructed a Gambian-enriched 1000 Genomes Project reference panel to improve representation of local genomic diversity and implemented an HLA phasing method to phase HLA alleles onto SNP haplotypes, which I demonstrated to function effectively in Section 4.3. This combined reference panel is also available for future studies. In addition, the QUILT-HLA workflow was modified in Section 4.4 to use the Wavefront aligner for read-to-allele alignment, which retrieved more informative reads across nearly all HLA loci and increases imputation confidence. Together, these improvements led to higher HLA imputation accuracy, as shown in Section 4.5. Comparison with imputation using the larger HLA multi-ethnic reference panel

indicates that the modified QUILT-HLA workflow with the Gambian-enriched 1000 Genomes Project reference panel still performed better at HLA-C and HLA-DQB1, and integrating multiple imputation workflows resulted in over 90% accuracy across all five HLA loci in this population. These results underscore the promise of lcWGS for HLA inference and highlight the importance of incorporating locally relevant reference panels due to distinct allele compositions in different populations, as illustrated in Section 4.1. Further discussion of the limitations and implications of this chapter is provided in Section 7.2.2.

Chapter 5 Detection of Large Structural Variants with Low-Coverage Whole-Genome Sequencing Data

Chapters 3 and 4 evaluate the ability of lcWGS to capture genome-wide variants and highly polymorphic HLA alleles. Chapter 5 extends this work to a different class of complex genomic variation: large structural variants (SVs). Accurate detection of SVs is biologically important, as they are implicated in a wide range of disease phenotypes [338]. Although not primarily designed for SV detection, microarrays have been used in several ways to capture structural variation, including methods such as PennCNV leveraging signal intensity from SNP probes to infer copy-number variation (CNV) [339-341], CNV-targeting probes, or from LD with nearby SNPs [59]. There are also many computational tools developed to call SVs from deep whole-genome sequencing data, whereas methods tailored to lcWGS remain lacking. This chapter presents a proof-of-principle demonstration that lcWGS can capture large SVs through a custom method, which I named lcSV (low-coverage structural variant). Section 5.1 introduces the mathematical formulation of the method and conducts a proof-of-principle analysis in the glycophorin region, which is of special interest to this Gambian cohort as it associates with severe malaria. To explore the broader applicability of lcSV, Section 5.2 evaluates its performance across SVs of different lengths and allele frequencies using simulations. These results then guide my analysis of previously reported SVs from

Eichler et al. in Section 5.3 [342, 343].

5.1 A population model with shotgun stochastic search algorithm

5.1.1 A motivating example in the glycophorin region

The glycophorin genes are located on chromosome 4q28-q31, namely *GYPA*, *GYPB*, and *GYPE*, encoding sialic-acid rich sialoglycoproteins that form major components of the erythrocyte membrane and render red blood cells their hydrophilic surface. This mechanism enables erythrocytes to circulate without adhering to other cells or vessel walls [344]. Due to extensive segmental duplications and sequence homology, the glycophorin region is highly prone to SVs, many of which conferring partial protection against pathogenic infection by *Plasmodium falciparum* that is responsible for severe malaria by altering the erythrocyte surface properties and thus reducing parasite invasion [345]. Motivated by its biological importance and relevance to the GAMCC cohort, I sought to investigate the feasibility of detecting these SVs using lcWGS data.

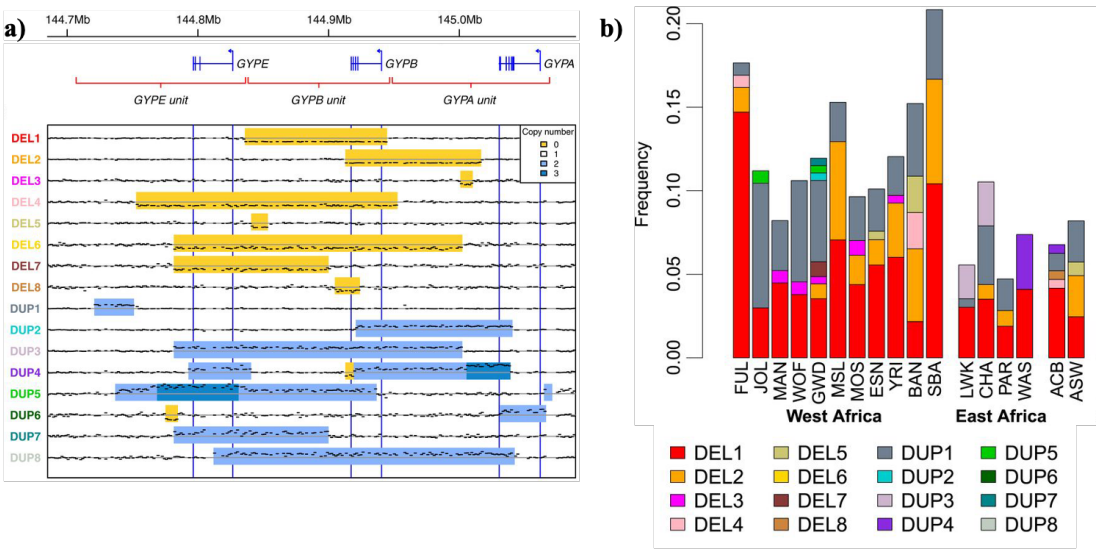


Figure 5.1.1. The glycophorin genes and SVs. These figures are adapted from Figure 1 and 2 from [345]. **a)** Genomic position in GRCh37 coordinates of the glycophorin genes (top) and the identified SVs (bottom). Deletions (yellow) and insertions (blue) are reflected by copy numbers, which is measurable by genome coverage from read pile-ups. **b)** Allele frequency of the SVs measured across different populations. The GAMCC cohort includes individuals from the first four ethnic groups (FUL: Fulani, JOL: Jola, MAN: Mandinka, and WOF: Wollof).

Figure 5.1.1 reveals the previously identified SVs in the glycophorin region. Since the number of individuals from the GAMCC cohort was limited, I only attempted to identify SVs that are reasonably large and common and decided to focus on DEL1 (~120 kb chr4:143,910,000-144,030,000) and DUP1 (~40 kb chr4:143,790,000-143,830,000). These SVs directly alter the number of DNA copies in the affected regions, producing corresponding changes in sequencing read coverage. Even at low coverage, such copy-number changes generate systematic deviations relative to neighbouring non-variant regions: duplications are reflected by increased coverage, as reads from multiple copies align to the same locus, whereas deletions result in coverage loss. To exploit this signal, I applied PCA to coverage profiles aggregated in 10 kb

genomic bins (by taking averages), which reduces random noise that would be prominent at per-base resolution in lcWGS data. PCA captures correlated patterns across samples, effectively separating technical variation from signals indicative of true SVs (see Methods for details).

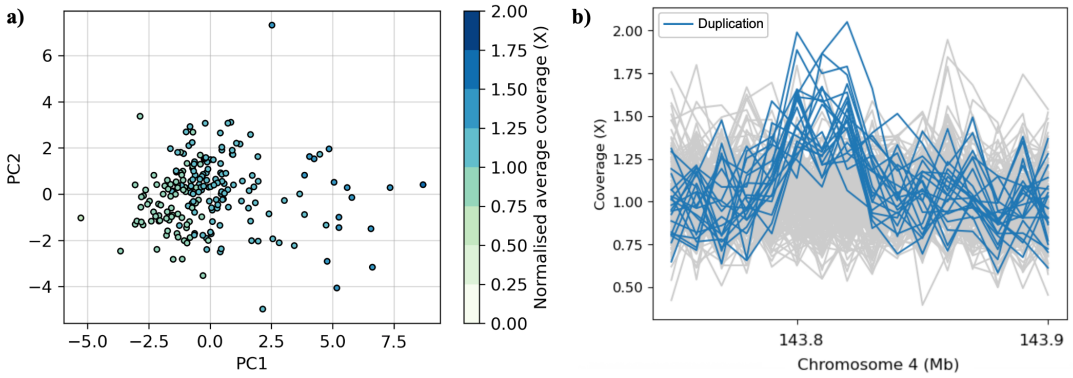


Figure 5.1.2. PCA for the DUP1 region. a) The first two PCs for each individual (marker) are showed, coloured by normalised genome coverage at this region. **b)** Genome coverage is plotted against genomic region for each individual (per line), with those PC1 > 5 coloured blue and the rest coloured grey.

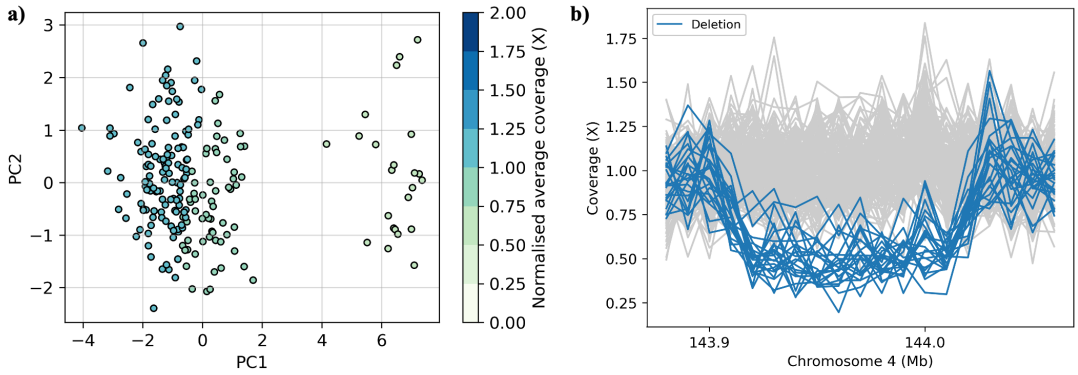


Figure 5.1.3. PCA for the DEL1 region. a) The first two PCs for each individual (marker) are showed, coloured by normalised genome coverage at this region. **b)** Genome coverage is plotted against genomic region for each individual (per line), with those PC1 > 3 coloured blue and the rest coloured grey.

As Figure 5.1.3 shows, PCA clearly separates individuals with DEL1 haplotypes from the normal genotype (individuals on the right have reduced genome coverage).

Nevertheless, no clear segregation is observed for the DUP1 region in Figure 5.1.2, although samples on the right tend to exhibit higher coverage and the line plot in Figure 5.1.2b suggests plausible duplication haplotypes for these individuals. The difference between PCA results for the DEL1 and DUP1 region is likely due to the different SV length for these haplotypes, where DEL1 is more than three times longer than DUP1 and thus a more obvious difference in genome coverage pattern between the SV region and the flanking region can be captured. Nevertheless, several concerns persist for this approach. First, the threshold on PC1 > 5 for the DUP1 region was discretionary and may not necessarily reflect the true calling of genotypes. Second, although a clear separation was observed in DEL1, manual calling on samples was required and thus less applicable to large number of regions. A more systematic way of discovering SVs with different length and allele frequency using lcWGS data as well as evaluating the calling results is thereby indispensable.

5.1.2 An alternative strategy: low-coverage structural variant

Unlike the PCA approach discussed in the previous section, which has several limitations, Leffler and colleagues employed a more systematic hidden Markov model to call SVs by examining genome coverage patterns across individuals [345]. However, this approach was not directly applicable to the present study and had several drawbacks. First, the sparsity of reads in the lcWGS setting made Gaussian-based modelling of coverage, as used for deep whole-genome sequencing, unsuitable due to overdispersion: a Poisson or negative binomial model was more appropriate in this context. Second,

their method called variants for each individual separately and reconciled SVs across samples *post hoc* by clustering nearby breakpoints, whereas a more integrated approach that allowed all individuals to explore a common pool of SVs during inference may be preferable. Additionally, the method did not model SVs as segregating haplotypes within individuals, which motivated a new approach that explicitly accounts for ploidy.

To address these considerations, I developed lcSV (low-coverage structural variant), a population model for detecting large SVs from lcWGS data. I implemented lcSV as a Python package that integrates the core algorithm with upstream preprocessing and downstream analysis utilities. lcSV requires precomputed genome coverage on sequencing data across genomic bins, which can be obtained using coverotron from the Iorek repository [286], and is publicly available on my GitHub page (https://github.com/Suuuuuuuus/lcSV). The method aims to find the best SV model for the observed genome coverage for all individuals in the studying cohort. Unlike using long-read technologies and assembling reads directly, lcSV takes the piled-up coverage as inputs and learns different patterns from that information. Intuitively, a deletion results in reduced coverage at part of a genome, since reduced or no reads are aligned to the reference region; a duplication is not observed as a contiguous copy of some genomic sequence (although it really is), but both pieces are aligned to the same genomic position in the reference, which translates into an abnormally higher coverage at the original locus compared to flanking regions. Hence, aligning reads to a reference

genome is an essential step, from which lcSV deciphers the presence or absence of an SV in a probabilistic way. Specifically, an SV model, denoted by γ, consists numerous SV haplotypes and their associated allele frequencies to represent the different genome coverage pattern likely possessed by individuals, with each haplotype being a list of copy numbers at each genomic bin. The choice of modelling genomic bins rather than per-base resolution is motivated by the inherent read sparsity of lcWGS and the non-independence of adjacent bases, which are sequenced within contiguous read segments. Importantly, enumeration of all possible SVs in a target region by combinatorics is computationally prohibitive, so lcSV employs a stochastic search framework to iteratively refine the SV model by proposing new *neighbourhood* models and evaluating these models [346], defined as every model with haplotype structure or allele frequency perturbed. Next, the evaluation step calculates the likelihood of each genome being modelled by all possible diplotypes in the current model, takes the average by taking into account the allele frequencies, and finally sums across all individuals (thus a population model) to emit a model score. Comparing these scores to the last best instruct the next move in the model space until convergence or pre-specified number of iterations. Eventually, individual genotypes are called with the best model.

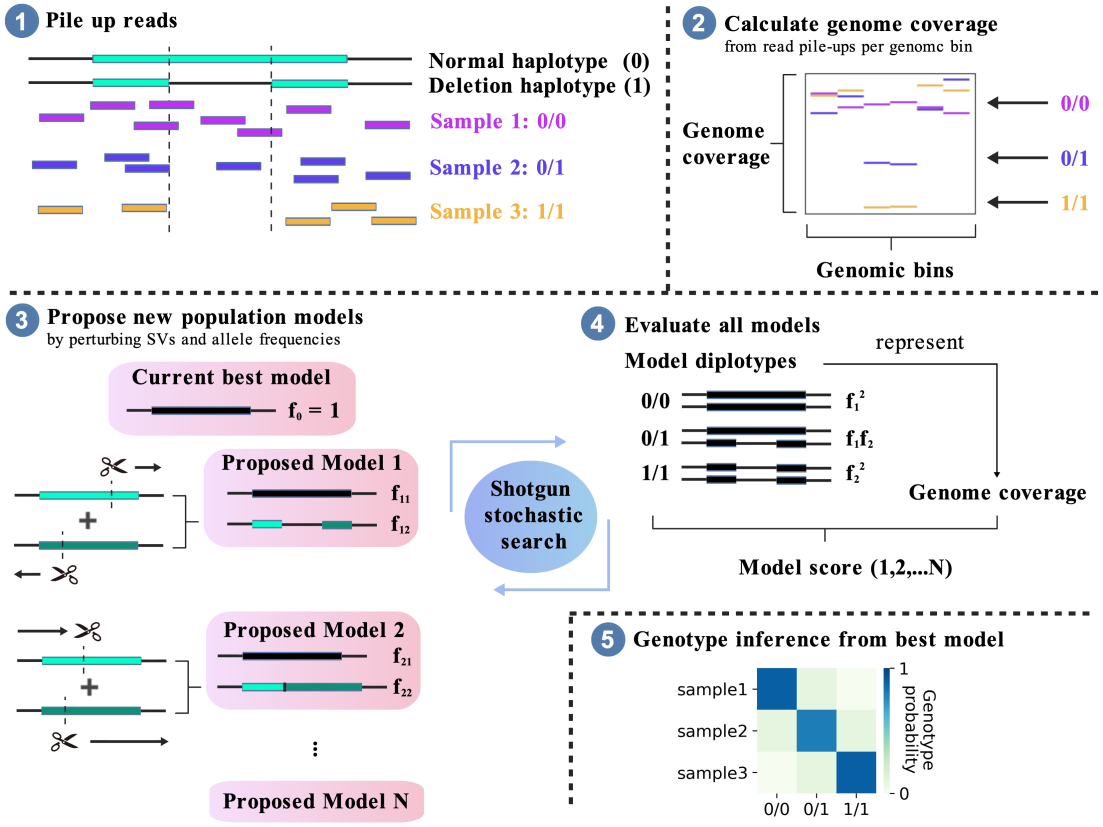


Figure 5.1.4. Overview of the lcSV method. Sequencing reads are aligned to the reference genome and converted to genome coverage. Then, the shotgun stochastic search framework is employed to iteratively propose new neighbourhood models by perturbing SV haplotypes and allele frequencies based on the current best model and evaluate these models based on observed genome coverage patterns of the individuals. After convergence or a pre-specified number of iterations, the best model is used to call sample genotypes at the target locus.

Now, I present all mathematical details of lcSV and conclude this section with an algorithm. Suppose N individuals and L genomic bins (variable with default size 1000) of a segment on the genome, again due to read sparsity in the lcWGS setting and non-independence of adjacent bases. Realised genome coverage is a matrix $\boldsymbol{C} \in \mathbb{N}_0^{L\times N}$, where each entry stores the depth measured by pile-up of all primary alignments within the chromosomal bin for that individual. To efficiently capture SVs in this region (if any), I propose models $\gamma \in \Gamma$ with different haplotype structures $h \in \mathbb{N}_0^L$, each associated with a frequency π, where each entry in h stores the copy number of that segment. Assume a maximum possible copy number 10, the full haplotype space H has 10^L different possible haplotypes, making it intractable to enumerate. Hence, I employ a shotgun stochastic search method to quickly explore plausible models through the

notion of neighbours, which I will formulate below.

Denote a single model $\gamma = (H, \pi)$. With Bayes' Formula:

$$P(\gamma \mid \boldsymbol{C}) \propto P(\boldsymbol{C} \mid \gamma) \cdot P(\gamma)$$

The likelihood term evaluates the fitness of this model on realised genome coverage across individuals. It suffices to formulate likelihood in the context of a single individual i due to independence and then sum across individuals to generate a full likelihood for a model. Suppose for an individual with realised genome coverage $\boldsymbol{c} = \boldsymbol{C}_{\cdot,i} \in \mathbb{N}_0^L$, K different haplotypes in the model and associated frequencies π, namely,

$$\begin{aligned} H &= \{h^{(1)}, h^{(2)}, \cdots, h^{(K)}\} \\ h^{(k)} &= (h_1^{(k)}, h_2^{(k)}, \cdots, h_L^{(k)}), \quad h_l^{(k)} \in \mathbb{N}_0 \\ \pi &= (\pi_1, \pi_2, \cdots, \pi_K), \quad \sum_{k=1}^{K} \pi_k = 1 \end{aligned}$$

Again, $h_l^{(k)}$ is the copy number at bin l in haplotype k and the aim is to compute likelihood

$$\mathscr{L} = P(\boldsymbol{c} \mid \gamma) = P(\boldsymbol{c} \mid H, \pi)$$

This is measured across all possible diplotypes in H, so for $a, b \in \{1, \cdots, K\}$, the diplotype is $h^{(ab)} = h^{(a)} + h^{(b)}$ and associated frequency is

$$\pi_{ab} = \pi_a \pi_b (1 + \mathbb{1}_{a \neq b})$$

Hence, the desired likelihood is marginalised over all possible pairs of segregating haplotypes:

$$\begin{aligned} P(\boldsymbol{c} \mid H, \pi) &= \sum_{a=1}^{K} \sum_{b=a}^{K} \pi_{ab} \cdot P(\boldsymbol{c} \mid h^{(ab)}) \\ &= \sum_{a=1}^{K} \sum_{b=a}^{K} \pi_{ab} \cdot \prod_{l=1}^{L} P(c_l \mid h_l^{(ab)}) \end{aligned}$$

where the last probability comes from a negative binomial (normal, if in the context of deep whole-genome sequencing) distribution parameterised by the measurements of a non-SV flanking region to account for differences in sequencing depth across individuals.

Instead of using prior $P(\gamma)$, a model selection criteria emulating that of Bayesian Information Criteria (BIC) is employed in this method. Suppose the model likelihood is $\mathscr{L}$, the original BIC is defined as

$$\text{BIC} = k \ln n - 2 \ln \mathscr{L}$$

where k is the number of parameters estimated by the model and n is the number of data points. BIC is a regularised model comparison metric that balances goodness-of-fit and model complexity (by penalising free paramters), with smaller values indicating a preferred model. In the lcSV setting, the number of parameters equals the number of haplotypes in the model $K-1$ (with the model by default contains the reference haplotype) times the number of genomic bins L. The number of data points is merely the input coverage matrix, which equals the number of individuals N times the number of genomic bins L. To account for noisy genome coverage in the lcWGS setting and to prevent overfitting on individual haplotypes, I include an additional penalty on the number and complexity of SVs, extending the BIC framework to regularise haplotype inference. Hence, the final model score S is:

$$S = ((K-1)L + X) \cdot \ln(NL) - 2 \ln \mathscr{L}$$

Term X is inversely related to the harmonic mean of run lengths and defined as:

$$X = \sum_{k=1}^{K} \sum_{r=1}^{R} \frac{1}{l_r}$$

where l_r is the length of each run r (contiguously possessing the same copy number) and R is the number of runs in a haplotype, with $L = \sum_{r=1}^{R} l_r$. For example, if haplotype $h^{(1)} = (1110000111)$, this term contributes

$$\sum_{r=1}^{3} \frac{1}{l_r} = \frac{1}{3} + \frac{1}{4} + \frac{1}{3} = \frac{11}{12}$$

to X. I do not rely solely on a prior on haplotype configurations because the inherent variability

of lcWGS data requires stronger regularisation to prevent overfitting on fragmented haplotypes. Given the number of plausible haplotypes at a non-SV hotspot locus is generally limited, this method provides appropriate model selection in this setting.

The above formulates our calculation of the likelihood for a single model γ. Then, the shotgun stochastic search framework enables an effective selection of a best-performing model:

- Initialise a default model $\gamma^{(0)}$ with a single haplotype that contains no SV and set the current best $\gamma^* = \gamma^{(0)}$.

$$h = \mathbf{1}_L \qquad \pi = 1$$

- At each iteration t, it proposes a series of models in the neighbourhood of the current best $\gamma^{(t)} \in \mathcal{N}(\gamma^*)$. Neighbourhood models are slightly tweaked versions of the current best model. Suppose $|\gamma^*| = K$, $\mathcal{N}(\gamma^*) = \{\gamma^+, \gamma^o, \gamma^-\}$ where
  - $|\gamma^+| = K + 1$ has an additional haplotype incorporated into the model. In our method, this is a recombinant generated from two random haplotypes in the current model with random breakpoints.
  - $|\gamma^o| = K$ has either one of the following:
    - Replace the a recombinant haplotype with an existing haplotype.
    - Adjust the allele frequencies of existing haplotypes in the model via Dirichlet sampling.
  - $|\gamma^-| = K - 1$ has an existing haplotype randomly removed from the model.

  For each γ, compute the model score S. Model with the smallest score S is stored as the current best, denoted by γ^*.

- Repeat this process until convergence or a pre-specified number of iterations.

With the best model γ^*, the marginalised likelihood for each individual is normalised into posterior probabilities, which becomes the final genotype probability and is used to call genotypes.

Algorithm 1 The lcSV method.

Input: Coverage matrix $\boldsymbol{C} \in \mathbb{N}_0^{L\times N}$.
Output: Best model $\gamma^* = (H, \pi)$ and individual genotype probabilities.
Initialise null-SV model $\gamma^{(0)}$ with haplotype $h = \mathbf{1}_L$, $\pi = 1$; set $\gamma^* = \gamma^{(0)}$.
for $i = 1, \cdots, I$ **do**
  Propose a set of neighbouring models $\mathcal{N}(\gamma^*)$ by perturbing haplotypes or frequencies.
  for all $\gamma \in \mathcal{N}(\gamma^*)$ **do**
    Compute model likelihood $\mathcal{L}(\gamma)$.
    Compute penalisation term X on haplotype complexity.
    Compute penalised model score $S(\gamma) = ((K-1)L + X)\ln(NL) - 2\ln\mathcal{L}(\gamma)$.
  end for
  Update $\gamma^* \leftarrow \arg\min_{\gamma\in\mathcal{N}(\gamma^*)} S(\gamma)$.
end for
Compute individual genotype probability according to γ^*.

5.1.3 Proof-of-principle simulation analysis with low-coverage structural variant

Having established the method, I first conducted proof-of-principle simulation experiments on simple cases to demonstrate both its feasibility and its strengths. In the first simulation case, I simulated a 20 kb genomic region containing two overlapping deletion haplotypes, each with a 6 kb deletion, to illustrate the ability of lcSV to model individual genome coverage as segregating haplotypes. The allele frequencies of the two deletions were set at 10% and 5%, and genotypes were randomly assigned to a total of 210 individuals. Genome coverage for each individual across genomic bins was then generated according to the assigned genotypes and subsequently re-inferred using lcSV. Additional details for lcSV on simulated dataset are presented in the Methods section. Figure 5.1.5 shows the model exploration of haplotype space: the true haplotype h_4 was refined from h_1, and h_5 from h_2 and h_3, thereby identifying the correct breakpoints while

simultaneously estimating the associated allele frequencies. After eleven iterations, the model converged to the best-fitting SV configuration, which was then used to call genotypes for each individual.

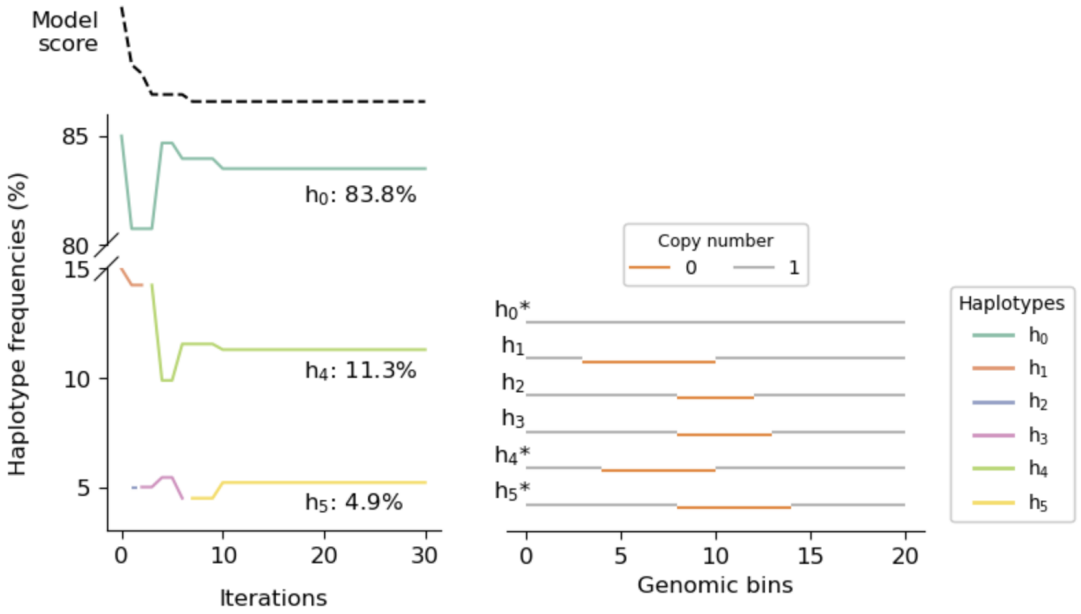
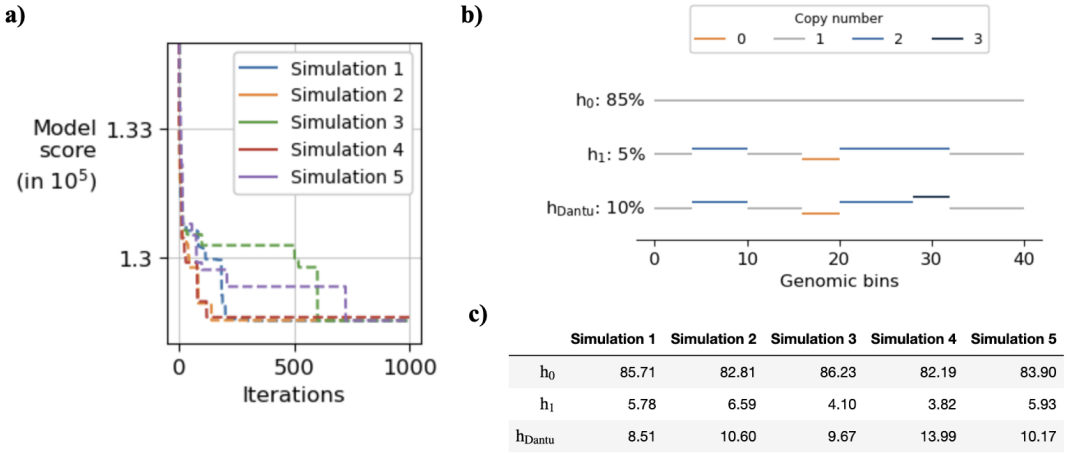


Figure 5.1.5. Simulation on a region with overlapping SV haplotypes. As iterations progressed, the model score decreased (top), and multiple candidate haplotypes together with varying allele frequency were evaluated (bottom). The reference non-SV haplotype h_0 and the target SV haplotypes h_4 and h_5 (labelled with stars) were retained in the final iteration following refinement of intermediate haplotypes. The inferred allele frequency also matches the simulation setting (11.3% and 4.9% for h_4 and h_5, respectively). The copy number structures of all haplotypes explored during the run are displayed on the right, with copy numbers 0 and 1 represented in orange and grey, respectively.

In the second experiment, I explored the performance of lcSV in detecting complex SV haplotypes, which can arise through multiple recombination events. A biologically relevant example is the DUP4 haplotype in the glycophorin region, encoding a serologically distinct blood group antigen known as Dantu [345, 347, 348]. This variant has been shown to reduce the risk of severe malaria and is common in certain East

African populations (for example, Kenya) but absent in West Africa (Figure 5.1.1). Given its biological importance and structural complexity, this experiment assessed whether lcSV is powered to detect a Dantu-like SV. To simulate this scenario, haplotypes were drawn from a hypothetical multi-allelic locus with three haplotypes: the reference (h_0), a *chimeric* haplotype mimicking Dantu (h_{Dantu}), and an intermediate variant from which Dantu was presumed to have arisen through a recombination event (h_1), with allele frequencies of 85%, 10%, and 5%, respectively (Figure 5.1.6). Because of the structural complexity of the chimeric haplotype, I ran lcSV on five simulation replicates, allowing up to 1,000 iterations per run. Across all simulations, lcSV successfully identified the correct breakpoints and copy number states aided by the presence of the intermediate haplotype. However, each run required a substantial number of iterations to converge to the optimal SV model, (for example, simulations 3 and 5 required 601 and 723 iterations, respectively) reflecting the extensive model search needed to resolve complex haplotype structures. While these results demonstrate the capability of lcSV to disentangle such configurations, it is likely that similar cases would pose greater challenges in real-world data, where loci containing chimeric haplotypes often coincide with SV hotspots and harbour a greater diversity of haplotypes accumulated through evolutionary processes, thereby necessitating larger sample sizes for precise pinpointing.



c)

	Simulation 1	Simulation 2	Simulation 3	Simulation 4	Simulation 5
h_0	85.71	82.81	86.23	82.19	83.90
h_1	5.78	6.59	4.10	3.82	5.93
h_{Dantu}	8.51	10.60	9.67	13.99	10.17

Figure 5.1.6. Simulation on a Dantu-like haplotype. a) Model score (y-axis) against iterations (x-axis), coloured by simulation replicates. **b)** The copy number structures of the reference haplotype (h_0), the intermediate haplotype (h_1), and the target chimeric haplotype (h_{Dantu}), with copy numbers coloured differently and true allele frequency displayed in text. **c)** Allele frequencies of haplotypes inferred from the optimal model in each of the five simulations.

5.1.4 Detection of DEL1 and DUP1 haplotypes in the glycophorin region with low-coverage structural variant

The previous section demonstrates that lcSV performs as expected and is somewhat capable of detecting even complex variants such as the Dantu haplotype. In this section, I applied lcSV to the glycophorin region in the real GAMCC dataset. As anticipated, only the DEL1 and DUP1 haplotypes were identified, as these are the only common variants in the Gambian population comprising this cohort. Consistent with Section 5.1.1, I used 10 kb genomic bins to mitigate noise, given that these variants are sufficiently large to be detected at this resolution. Figure 5.1.7 illustrates the iterative model refinement process, in which SV breakpoints and allele frequencies were progressively refined, ultimately converging on a model with three haplotypes: the reference, DEL1, and DUP1. Figure 5.1.7b shows the average genome coverage across

bins stratified by genotype, with DEL1 and DUP1 displaying reduced and elevated coverage, respectively, at approximately the expected coordinates. Two individuals were further called as heterozygous for both DEL1 and DUP1, and their genome coverage profiles supported these genotype assignments.

This analysis highlights both strengths and limitations of lcSV. A key advantage is its population-aware design: individual genotypes are inferred from a common pool of segregating haplotypes, avoiding the need for *post hoc* consolidation of SV calls across samples. In practice, however, since DUP1 and DEL1 do not overlap in genomic coordinates, it would be more efficient to model them separately to reduce complexity, and this run is intended to illustrate the capability of lcSV. A limitation of lcSV lies in its resolution of SV breakpoints. While the method suffices for genotype calling, it provides only an approximate delineation of breakpoints. For example, the flanking bins of DEL1 display intermediate coverage values relative to the normal and deleted states (Figure 5.1.7b), suggestive of breakpoint locations but insufficient for precise mapping. This limitation arises from the intrinsic sparsity of lcWGS data. A potential methodological refinement could involve a hierarchical approach by first operating lcSV at a broader bin level to identify candidate regions and followed by breakpoint localisation within transitional bins at higher resolution. However, this strategy was not pursued in the present work.

Together with the *in silico* simulations presented earlier, this application demonstrates the utility of lcSV for SV discovery and genotyping in lcWGS data. In the following sections, I extended its use to other genomic regions systematically.

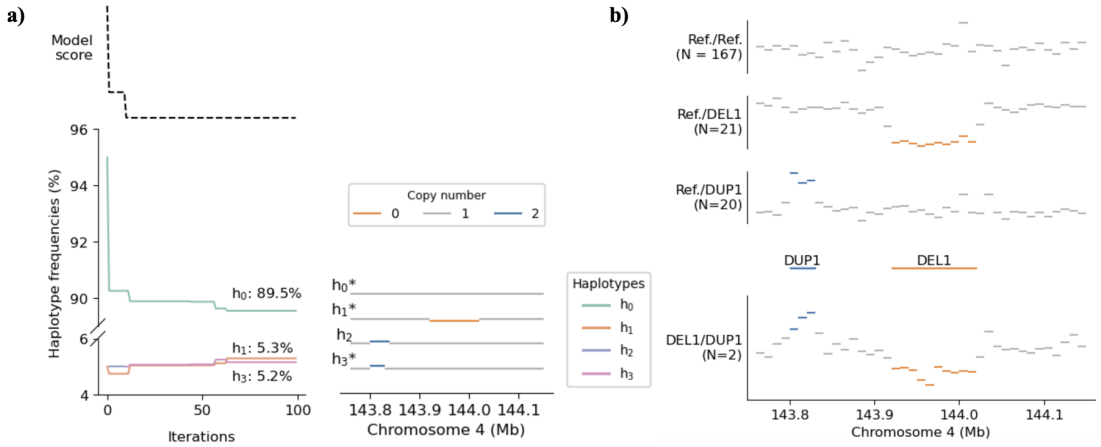


Figure 5.1.7. Calling SVs with lcSV in the glycophorin region. a) Model exploration of SV haplotypes. As iterations progress, the model score decreases (top), and multiple candidate haplotypes are evaluated (bottom). The reference non-SV haplotype h_0 and the glycophorin SV haplotypes DEL1 (h_1) and DUP1 (h_3) are retained in the final iteration following refinement of intermediate haplotypes. The copy number structures of all haplotypes explored during the run are displayed, with copy numbers 0, 1, and 2 represented in orange, grey, and blue, respectively. **b)** Genotype calls derived from the optimal SV model identified by lcSV. The average genome coverage (normalised by flanking mean for each individual) across the glycophorin region stratified by genotype class: homozygous reference, heterozygous reference/DEL1, heterozygous reference/DUP1, and heterozygous DEL1/DUP1 (top to bottom) is shown as horizontal lines per genomic bin. Genomic coordinates of DEL1 and DUP1 are indicated, and copy numbers are coloured using the same scheme as in a).

5.2 Assessment of simulated data with different parameters

It is crucial to note that the DEL1 and DUP1 haplotypes in the glycophorin regions are common and large enough to be captured by lcSV. To fully assess the potential of lcSV and apply this method to real data in the next section, I simulated SVs with different

lengths and allele frequency as well as the associated genome coverage for individuals, run lcSV on all scenarios, and assess the performance by comparing to the known truth. Notably, the resolution of lcSV is inherently constrained by the read sparsity of lcWGS, which limits its ability to detect short SVs. Consequently, the method only applies to common variants of larger size and operates again on genomic bins rather than single-base resolution. Importantly, lcSV is designed to be more effective for calling known large SVs where breakpoints may be imprecisely defined rather than for discovering entirely novel variants, as an approximate delineation of the target region is required as input. Specifically, I simulated 210 diploid individuals carrying SVs of varying lengths (2, 4, 6, 8, or 10 kb, again in 1 kb genomic bins) and allele frequencies (5%, 10%, 20%, 30%, 40%, or 50%), each with 100 replicates. I did not simulate shorter SVs due to the low sequencing depth at 1×, nor rare SVs given the population size of the real GAMCC cohort. Each locus was modelled with two haplotypes: a reference sequence with copy number one at all genomic bins and an alternate haplotype in which the central third of the region was deleted (with copy number zero). Individual sequencing coverage was drawn from a normal distribution with realistic mean and variance measured from our cohort, and read depth across loci was simulated using a negative binomial distribution to capture overdispersion of the nature of lcWGS. For each simulated individual, I generated both baseline coverage profiles and true coverage signals reflecting their SV genotype, thereby producing realistic datasets for downstream model evaluation. Full details of the above is described in the Methods section.

It is obvious from Figure 5.2.1 that SVs larger than 4 kb can be almost confidently detected by lcSV at as low as 5% allele frequency, a typical GWAS threshold for cohorts with several hundred individuals. For SVs with 2 kb length, in less than 20% of replicates at all allele frequency level was I able to correctly call the SV, limiting the utility of lcSV in calling SVs smaller than this threshold given the current population size and sequencing depth. As expected, Figure 5.2.1a shows a generally increasing trend as SVs become larger and more frequent, which is in line with the concordance plot in Figure 5.2.1b and reinforces that genotype INFO highly correlates with concordance, enabling an estimation of calling confidence without knowing the underlying truth in real world analysis.

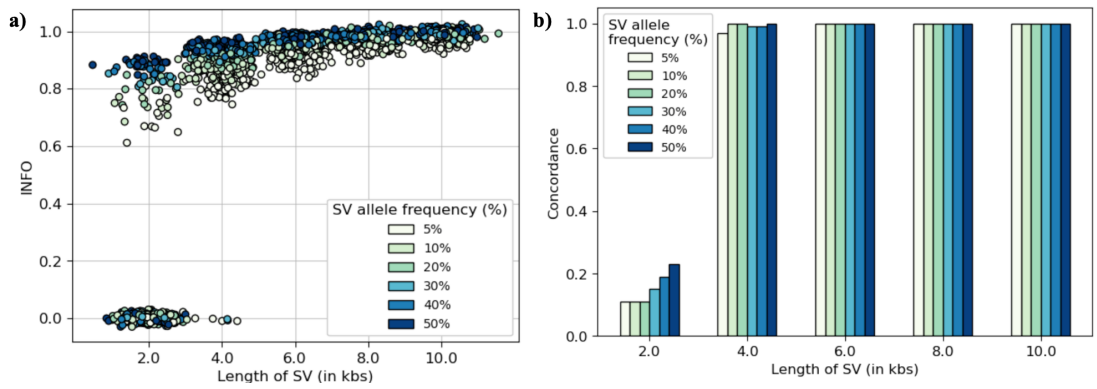


Figure 5.2.1. LcSV simulation results on 1 kb genomic bins. a) Genotype INFO (y-axis) against SV length (x-axis) and allele frequency (colours) for each run. Again, only five different cases of SV lengths are simulated as shown in the ticks, and markers are jittered on the x-axis and slightly on the y-axis to facilitate visualisation. **b)** Genotype concordance (y-axis) against SV length (x-axis) and allele frequency (colours), averaged over replicates.

5.3 Application of the low-coverage structural variant method to known structural variants

Simulation results from the previous section confirmed the feasibility of using lcSV to call SVs larger than 4 kb and with an allele frequency above 5%. Guided by this observation, I applied the method to lcWGS data from the GAMCC cohort, evaluated its overall accuracy, and examined representative examples of both successful and unsuccessful calls to provide further insights.

5.3.1 Genotyping known structural variant regions

A recent study by Ebert and co-workers presents 64 high-quality, haplotype-resolved human genome assemblies from 32 diverse individuals using long-read technologies and identifies 107,590 SVs, of which 68% were thought previously undetectable by short-read sequencing, providing a resourceful map of human structural variation [342, 343]. With this invaluable resource, we were particularly interested if some of these known SVs could be recovered in the GAMCC cohort with lcSV.

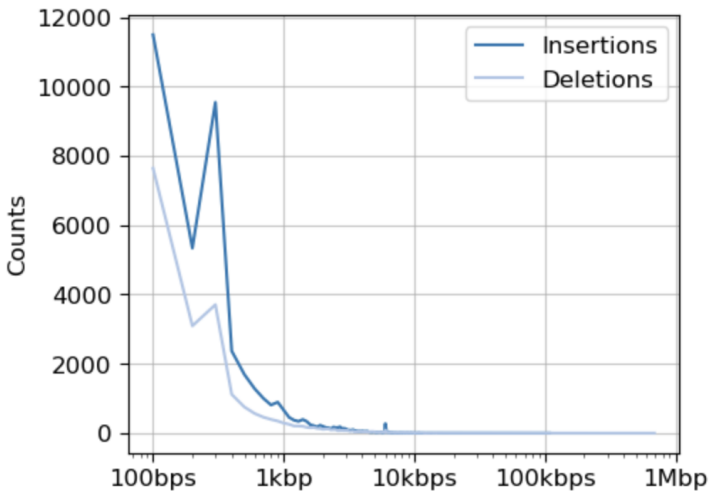


Figure 5.3.1. 107,590 previously identified SVs. These SVs were reported by Ebert et al. [343]. This plot shows the distribution of insertions (dark blue) and deletions (light blue) over SV size (x-axis).

The lcSV method relies on accurate genome coverage estimates derived from read pile-ups, which are sensitive to alignment artefacts, particularly in short-read lcWGS data such as the 151 bp paired-end reads used in the GAMCC cohort. In this configuration, coverage loss caused by deletions is readily detectable as a local reduction in read depth, whereas insertions may not be captured because additional sequence can be mapped elsewhere in the genome rather than adjacent to the insertion site depending on the various SV formation mechanisms also discussed by the original paper [343]. Since whether an SV insertion is a duplication from flanking region was not indicated in the list, we decided to focus the analysis on a curated set of 621 large (> 5 kb) and common (> 5% allele frequency) deletion SVs (see Methods for details).

5.3.2 Calling large deletion structural variants with low-coverage structural variant

As noted above, effectively mitigating alignment issues was critical for reliable SV detection. For each region, I excluded genomic bins that are inaccessible to short-read sequencing or contain reads with low mapping quality, and I avoided calling variants in regions with particularly severe alignment problems. Although no true SV type was available for each individual due to the lack of truth data, identified SV haplotypes could be compared at the model level with the SV list based on inferred genomic positions and copy numbers. Full details of the runs and assessments are provided in the Methods section. Finally, lcSV was run on 429 of 621 deletions (69.1%), each containing at least 25% mappable genomic bins based on the threshold used in the original glycophorin study, with results shown in Figure 5.3.2.

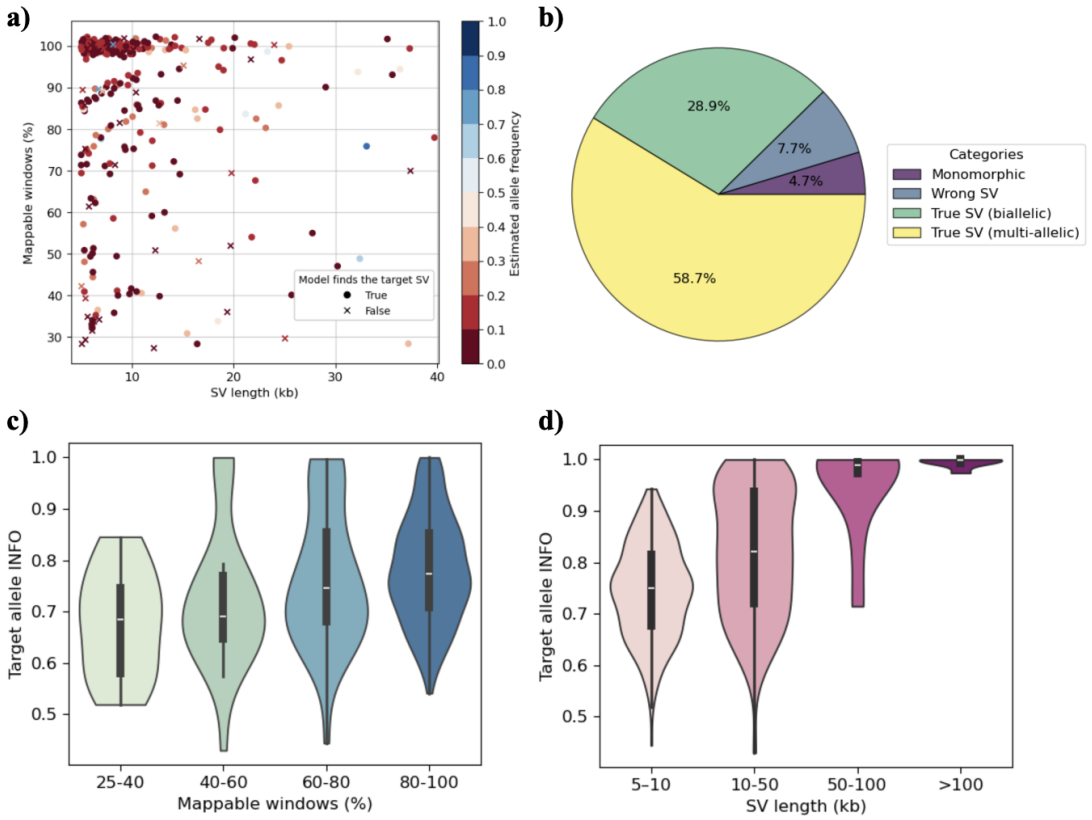


Figure 5.3.2. SV calling results on 429 deletion regions with lcSV. To reinforce, these regions are with at least 5% allele frequency, 5 kb in size, and 25% mappable windows. **a)** Scatter plot of the percentage of mappable windows (y-axis) over the length of SVs (x-axis). For visualisation purposes, the x-axis is truncated to SVs smaller than 40 kb (retaining 395 sites), with values slightly jittered on both axes to facilitate visualisation. Each marker is coloured by the model estimated allele frequency and labelled by a circle (cross) to indicate our method correctly (incorrectly) identifies the target SV haplotype. If the true haplotype is not found by lcSV, the most frequent alternative haplotype is plotted, labelled by a cross. **b)** The calling results by categories. The blue (bi-allelic, the target deletion haplotype is the only alternative SV detected in this region) and purple (multi-allelic, more than two haplotypes are present in this population at this locus) colour represent the proportion of cases where the target SVs are retrieved; the yellow colour indicates at least an alternative SV has been identified, although none of them are in line with the target one; the red colour depicts all monomorphic sites, namely, the final model contains only the reference haplotype. **c)** Distribution of target allele INFO by mappability. **d)** Distribution of target allele INFO by SV length. I removed all monogenic SV regions (376 regions) in c) and d).

Since lcSV detected SVs using genomic bins, I considered an identified deletion haplotype to match the target haplotype if the model-inferred segment with copy number zero overlapped the reported SV coordinates, allowing a tolerance of ±1 kb

around both breakpoints. The confidence INFO metric was measured as the ratio of the observed variance of the expected allele dosage across samples to the expected variance under Hardy-Weinberg equilibrium (see Methods for details). Across all 429 regions, I recovered 87.6% of the target SV haplotypes using the comparison criteria mentioned earlier [343]. However, since this set of SVs originated from a relatively small and globally diverse reference, it was expected that some SVs are absent from the GAMCC samples or previously uncharacterised haplotypes exist. Although these haplotypes were not directly accessible for this cohort due to lack of truth data, their presence could be inferred from the INFO metric (Figure 5.3.2c and Figure 5.3.2d). SVs that were larger and less affected by mappability issues were more confidently detected by lcSV. In 58.7% of regions, lcSV identified the target locus as multi-allelic, containing at least one additional alternative deletion haplotype beyond the known variant. On the other hand, Figure 5.3.3 illustrates lcSV calls across several genic regions with high-confidence signals, providing further evidence that the method performs as expected. Visualisation of genome coverage stratified by inferred genotypes clearly reveals the presence of SVs. The left column highlights bi-allelic sites, while the right column presents multi-allelic regions. For instance, the top- and bottom-right panels display overlapping SVs, where a smaller variant is embedded within a larger one, whereas the central panel demonstrates two adjacent SVs. Nonetheless, lcSV calls may subject to spurious signals or haplotypes, which could arise either from genuine smaller SVs in proximity to the target region or as false positive copy numbers due to sparsity of

lcWGS data. Moreover, in the absence of validated truth genotypes for these variants and samples, instances where the expected haplotype was not recovered may reflect true allele frequency differences across. To further contextualise these results, I presented three representative cases to provide additional insights.

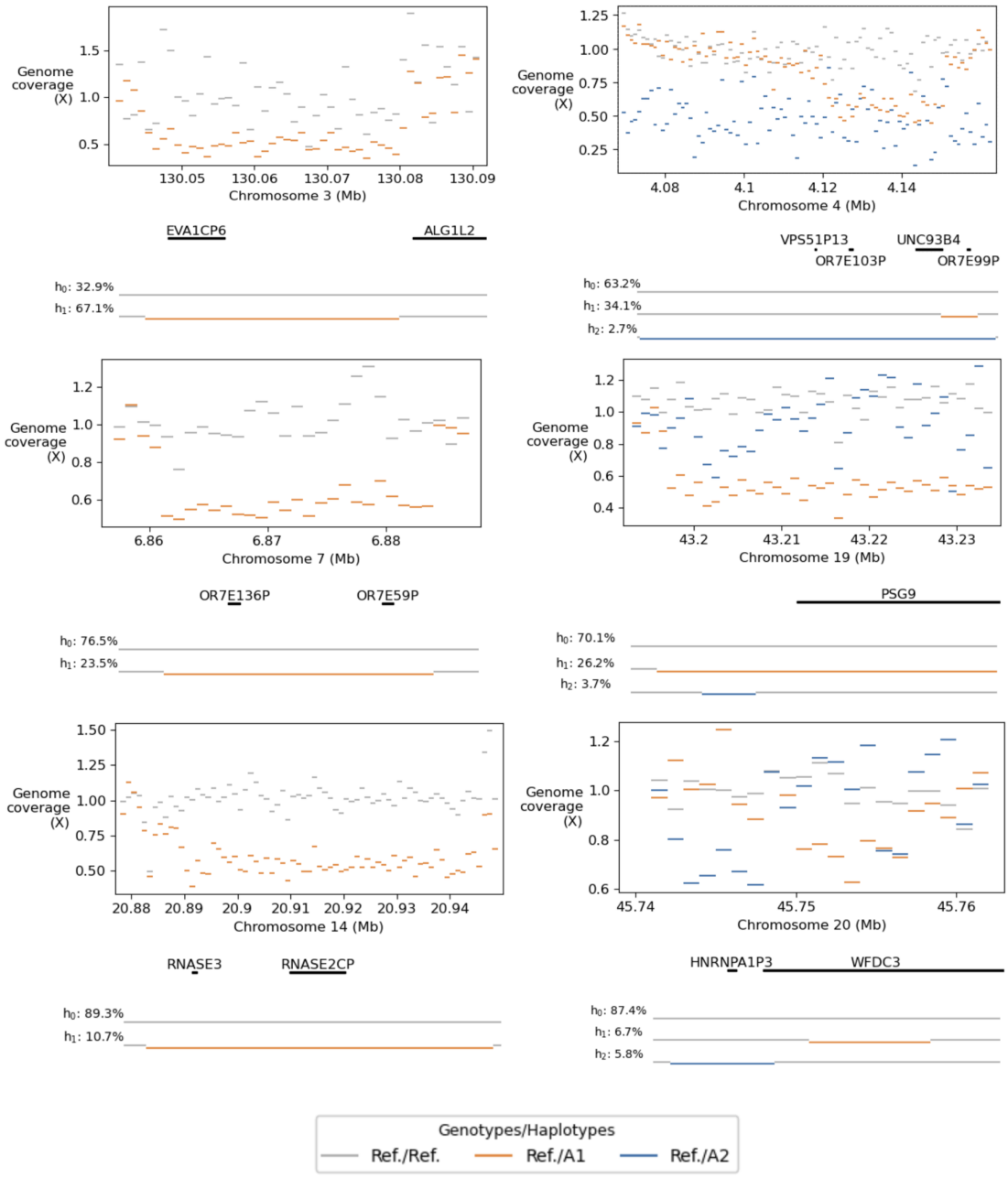


Figure 5.3.3. SV genotype calls on six loci with lcSV. The left panels show three bi-allelic sites, and the right panels show three multi-allelic sites. For each plot, normalised average genome coverage (y-axis;

scaled by the mean coverage of flanking non-SV regions) is displayed across 1 kb genomic bins (x-axis). Individuals are grouped by genotype, with those carrying the reference haplotype (Ref.), alternative allele 1 (A1), or alternative allele 2 (A2) shown in grey, orange, and dark blue, respectively. Individuals carrying both non-reference haplotypes are excluded due to limited representation. Annotated genes and haplotypes called from lcSV with the associated allele frequency within each region are indicated in the tracks below, with A1 or A2 deletions coloured in a similar pattern as genome coverage.

5.3.3 Investigation of factors affecting structural variant calling results using low-coverage structural variant across three cases

Even though lcSV successfully identified most of the target deletions, as shown in the previous sections, several factors could influence the calling process and the reliability of its results. In this section, I examined three representative cases that highlight challenges in accurate SV detection from lcWGS data.

Long intergenic non-protein coding RNA 2055 (LINC02055) harbours a large deletion locus spanning ~180 kb (GRCh38 coordinate chr8:136,668,039-136,850,135). In Section 5.1, I demonstrated that both PCA and lcSV agreed in their calling of the large DEL1 SV in the glycophorin region, which was of similar size to this variant. Nevertheless, lcSV failed to identify any non-reference haplotype in this population (i.e., the best model contains only the reference haplotype, with copy number equal to one across all genomic bins). The heatmap in Figure 5.3.3 reveals that the labelled sample Δ has reduced genome coverage in the middle of this locus (lighter colour), which is confirmed by comparison with average of the rest in Figure 5.3.3b. Genome coverage halves for sample Δ, suggesting this individual is heterozygous at this locus. The observed allele frequency of this SV in our cohort was 1/420, thereby too low to

be detected by the current method; a more relaxed penalisation formulation might reveal this SV. This observation reflects population-specific variation in allelic composition, as the original literature reports an allele frequency exceeding 5% among African individuals, and underscores that the inability to call SVs with lcSV is likely attributable to the limited resolution and data sparsity inherent to the approach and lcWGS.

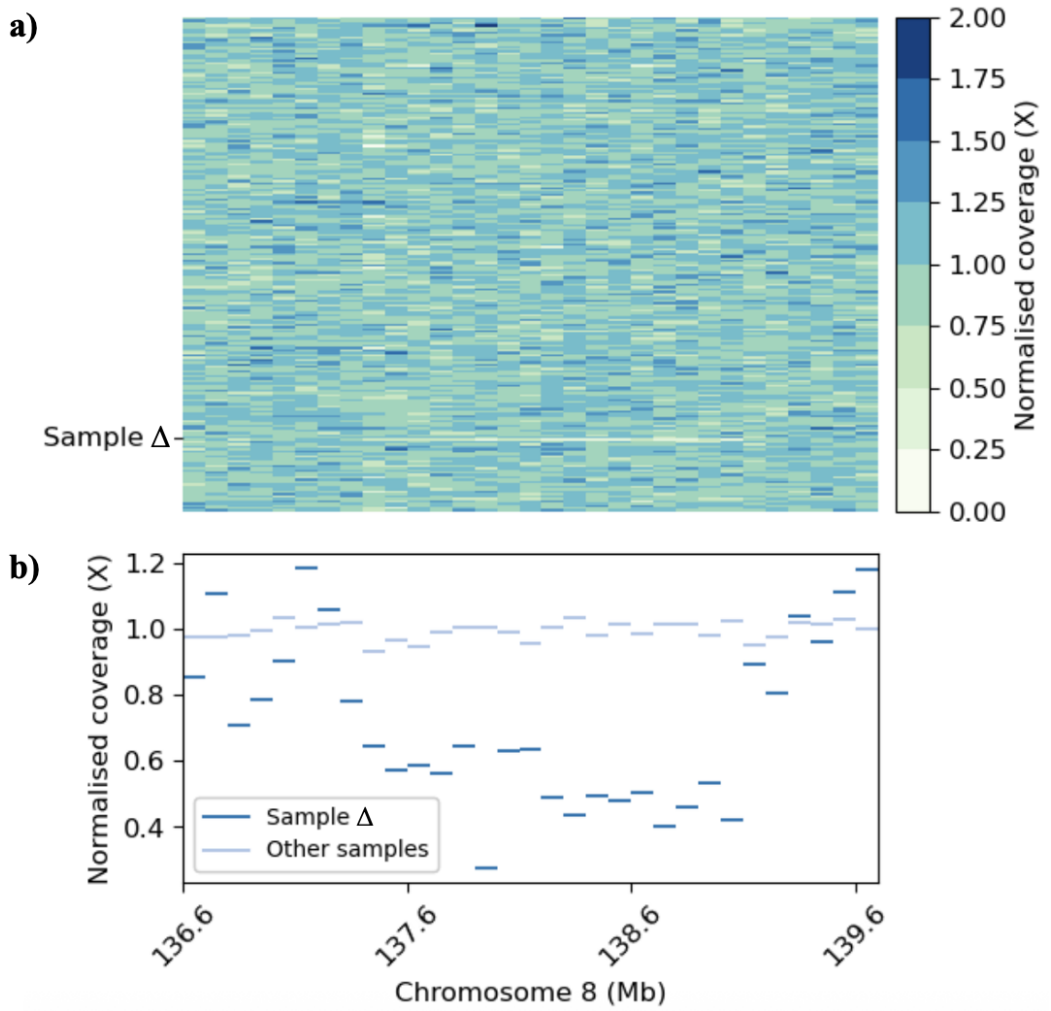


Figure 5.3.4. Genome coverage at LINC02055. a) Average genome coverage per 10 kb window (normalised by flanking region) for each individual (y-axis) across this region (chr8:136,600,000-139,610,000, x-axis): darker colour reflects higher coverage. Sample Δ has reduced coverage in this region and is marked out. **b)** Average genome coverage (y-axis) across individuals excluding sample Δ and sample Δ in this region.

In 7.7% of regions, lcSV detected at least one SV haplotype but failed to identify the target haplotype. I examined one such case using an example of > 9 kb deletion (chr4:107,231,553-107,240,789) within the Dickkopf WNT signalling pathway inhibitor 2 (DKK2) gene and found that the discrepancy arose from inaccurately inferred SV boundaries, where excessive noise in the coverage data prevented precise breakpoint detection. Instead of the reported deletion in the original paper, lcSV identified a longer deletion (h_1). This result was unlikely to reflect alignment artifacts, as the locus was accessible to short-read sequencing, nor was it attributable to nearby smaller SVs, since this was the only variant detected in the region. Rather, it represented an incorrect delineation of breakpoints. To further examine this issue, I constructed three SV models (Model1, Model2, and Model3), each comprising the reference haplotype and each of the three deletion haplotypes characterised by distinct breakpoints (h_1, h_2, and h_3). Here, Model1 was output by lcSV, whereas Model3 contained the reported SV haplotype. Holding the alternative allele frequency constant, I evaluated their respective model scores and genotype calling confidence. Model2 yielded the lowest genotype confidence (INFO, Figure 5.3.5f) and the least improvements over the default model (model score, panel Figure 5.3.5e), representing the worst over the three models. By contrast, Model3 and Model1 produced similar scores and INFO metrics. The heatmap in Figure 5.3.5c shows 39 individuals identified as carrying the deletion by at least one of Model1 or Model3. Visual inspection reveals no consistent pattern in genome coverage across the two genomic bins adjacent to the

reported boundary (107.25-107.27 Mb). To conclude, it is not possible to clearly determine whether the deletion haplotype spans 10 kb or 12 kb with this dataset, as the signal cannot be confidently distinguished from lcWGS read sparsity and high variability, although the model clearly disfavours the intermediate 11 kb configuration. These findings underscore a key limitation: while our approach and lcWGS are suitable for genotyping known variants, reliably detecting novel variants or resolving precise breakpoints would require higher sequencing coverage.

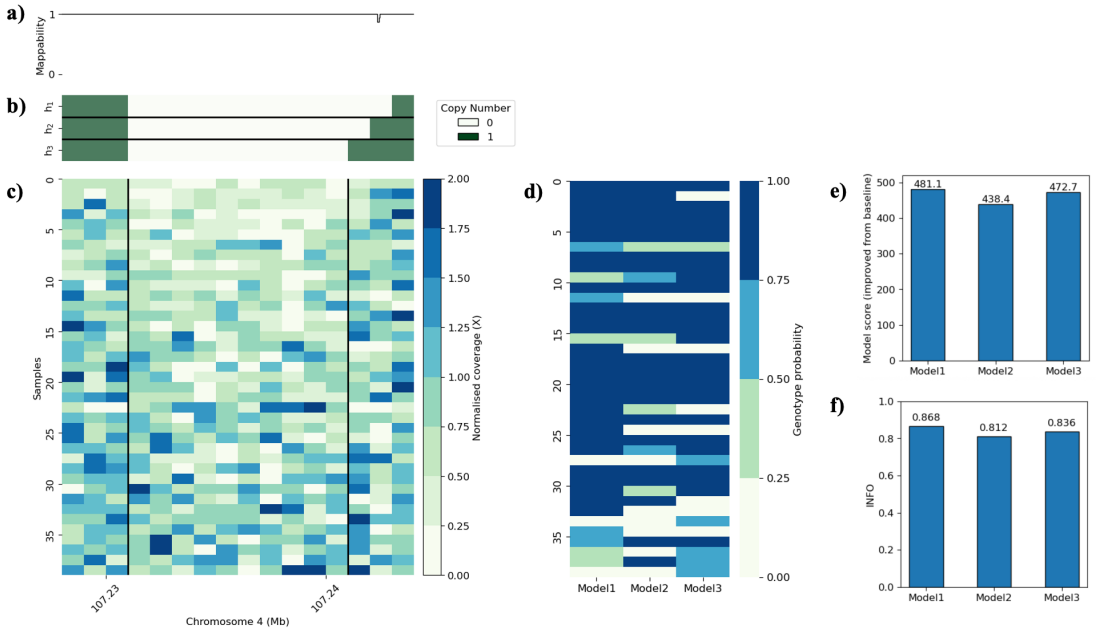


Figure 5.3.5. LcSV calling result at *DKK2*. a) Mappability in the DKK2 region. **b)** The three SV haplotypes (h_1, h_2, and h_3) inspected in this section, corresponding to three SV models (Model1, Model2 and Model3). Genomic bins with the normal copy number one are coloured green and deletion left white. Among the three SVs, h_1 is identified by lcSV, whereas h_3 matches the genomic coordinates recorded previously. The alternative allele frequency for each model is set to be that identified by lcSV. **c)** Average genome coverage per 1 kb window (normalised by flanking region) for individuals (y-axis) across this region (chr4:107,228,000-107,244,000, x-axis): darker colour reflects higher coverage. Only 39 individuals that are identified to have the deletion haplotype by at least one of the three SV models are retained, ordered by increasing average genome coverage across this region. The bolded vertical lines indicate the plausible breakpoints of this site (from h_3). **d)** Heterozygous genotype probability for each individual (row, ordered accordingly as d) by each model (column), with higher probability coloured darker. **e)** Improved model scores relative to the default model (with only the reference haplotype) by introducing the respective SV haplotypes. **f)** Genotype INFO score for the alternative alleles.

In the last case, I presented a multi-allelic locus where lcSV identified the target SV but also suggested another unknown haplotype. I obtained a list of four common SVs in this region (chr3:1,876,125-1,881,529) from gnomAD with African allele frequency greater than 1%, where lcSV identified the leading common variant (H_1, which agrees the h_1 variant recorded in the database) as well as an uncharacterised haplotype that was not previously acknowledged (H_2). Notably, telomeric and subtelomeric regions are known to be prone to SV hotspots, owing to their repetitive architecture that foster non-allelic homologous recombination and other rearrangement mechanisms [349-351], exhibiting elevated SV density compared to genomic averages, are structurally dynamic, and often pose challenges in replication and end-maintenance [352]. Figure 5.3.6d shows the average genome coverage of individuals stratified by the genotypes inferred using the optimal lcSV model. A distinct drop in genome coverage across a 4 kb region (1,869,000-1,873,000) supports the presence of a deletion on haplotype H_2, potentially representing a novel SV. Twelve individuals were identified as heterozygous reference/H_2 at this locus, which is less likely to be spurious. This observation highlights that the reported SV catalogue does not fully capture population-specific variation, and that lcSV may also suggest previously unreported haplotypes. However, in the absence of truth data, confirmation of whether this SV is genuinely novel still requires molecular validation.

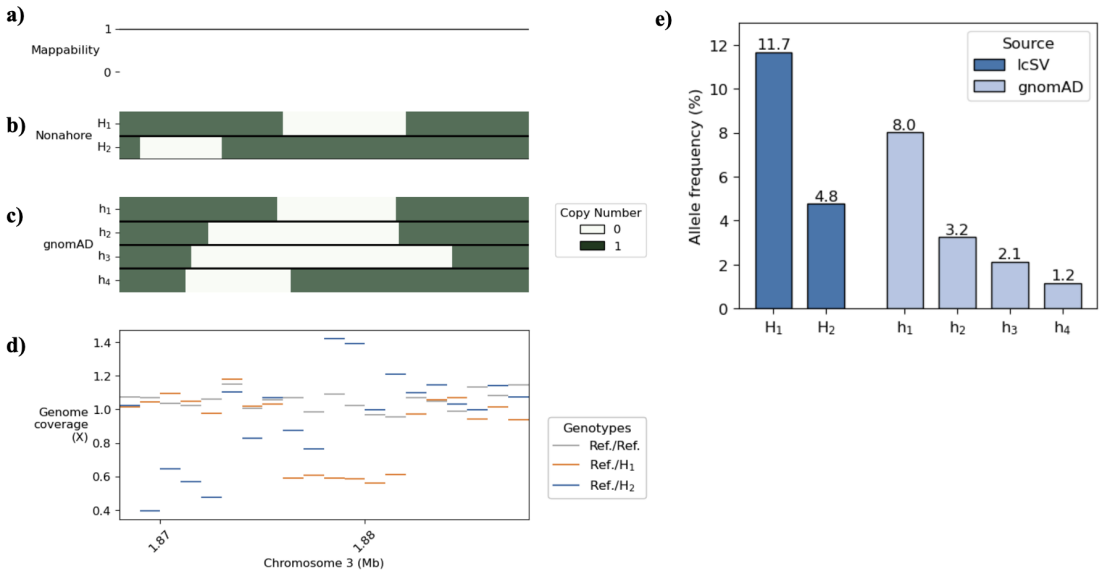


Figure 5.3.6. SV haplotypes in chr3:1,876,125-1,881,529. a) Mappability in this region. **b)** The two SV haplotypes identified by lcSV (H_1 and H_2) inspected in this section. Genomic bins with the normal copy number one are coloured green and deletion left white. **c)** The four SV haplotypes recorded in gnomAD (h_1, h_2, h_3, and h_4). **d)** The average genome coverage (normalised by flanking mean for each individual) across this region stratified by genotype class: homozygous reference, heterozygous reference/H_1, and heterozygous/H_2 (top to bottom) is shown as horizontal lines per genomic bin. **e)** Allele frequencies of the SVs, either estimated by lcSV or obtained from gnomAD.

Hence, many factors complicate SV detection in the telomeric regions and should thus be treated cautiously with lcSV. Integrating the results from previous analyses, I identified 324 non-monogenic, non-telomeric deletion regions (> 50% mappable windows, allele frequency > 5%, and size > 5 kb; Figure 5.3.2 and Section 5.2) that can be inferred using lcSV in the Gambian population at 1× sequencing depth.

5.4 Methods

Measuring genome coverage from read pile-ups. The presence or absence of SVs are reflected from different coverage patterns across genomic regions, where an

exceptionally high coverage suggests plausible duplications and low coverage suggests deletions. This section outlines the methodology used to measure individual genome coverage throughout this chapter. Starting with bam files (from Chapter 3), I retained only properly paired, primary, and non-supplementary reads with samtools -f2 -F2304 and chunked by 5 Mb genomic regions as in Chapter 3 for performance purposes. Then, I ran coverotron v0.9 (revision 19ca4a1) on each chunked region, which is a utility from Iorek (commit 19ca4a1152e10c81ff87686ee3d4594d6f2873de), with 1 kb bin size and default parameters [286]. The software output the observed coverage within each consecutive 1 kb window along with a mean mapping quality for each individual. These results can be optionally deresoluted to 10 kb bins by averaging over consecutive regions.

PCA on glycophorin DEL1 and DUP1. I took rough GRCh38 coordinates of the DEL1 (chr4:143,910,000-144,030,000) and DUP1 (chr4:143,790,000-143,850,000) haplotypes identified in previous literature [345, 353], as knowledge of positions of the exact breakpoints is not necessary in this approach. To account for different sequencing depth of each individual introduced from pooling and PCR, I normalised nominal coverage by the mean depth of flanking in both directions of the same length (500 kb) as the target region. Since both DEL1 and DUP1 are large, I worked with 10 kb genomic bins to avoid extra noise by smaller regions and standardised before PCA decomposition.

Simulation of genome coverage and true genotypes across SV regions. For all simulation analyses in this chapter, I generated genome coverage data and true genotypes for 210 individuals. The raw sequencing depth of each individual was assumed to follow a Gaussian distribution with mean 1.21× and standard deviation 0.12×, chosen to approximate the empirical characteristics of the GAMCC cohort. SV haplotypes were represented as arrays of copy numbers across genomic bins (default size 1 kb) and assigned allele frequencies. Diploid genotypes were then randomly allocated to individuals according to these frequencies. For each bin, sequencing coverage was sampled from a negative binomial distribution, parameterised by the following mean and variance

$$mean = copy\ number \times bin\ size \times sequencing\ depth$$

$$variance = copy\ number \times bin\ size \times sequencing\ depth \times per\text{-}base\ variance$$

where the last term (set to 100) captures stochastic variation in read sampling and accounts for overdispersion in the lcWGS setting. This simulation workflow was also implemented in the lcSV package and enabled different distributions (for example, Gaussian and Poisson) for application to deep whole-genome sequencing data. For each individual, two outputs were produced: a *training* coverage profile representing genome-wide background depth where mean and variance of coverage can be inferred from, representing data observed from a flanking non-SV region as well as a *coverage*

profile reflecting the specific haplotype structure and associated copy number of the simulated region. The resulting haplotypes and coverage profiles served as ground truth data for assessing the accuracy of lcSV inference.

Assessments on the size and allele frequency of SVs detectable by lcSV. Simulation was performed exactly as described above except that each scenario (varying SV length and allele frequency) was simulated with 100 replicates. In each scenario, only two possible haplotypes were present in the population: the normal haplotype with copy number one at all positions as well as a deletion haplotype with copy number zero in the central third. I examined SV length of 2, 4, 6, 8, or 10 kb (translating into 6, 12, 18, 24, 30 kb chromosome region and the equal numbers of genomic bins) and allele frequency with 5%, 10%, 20%, 30%, 40%, and 50%, where each haplotype was randomly assigned a normal or a deletion haplotype based on these parameters. I ran lcSV on the true coverage profile by learning mean and variance of the training, called the SVs on the simulated individuals, and compared to the known haplotype designations.

Evaluation of the results was based on genotype concordance and the INFO metric. Because the true haplotypes for each individual were known, concordance was set to 1 if the final SV model contained exactly the correct two haplotypes, the SV haplotype and the normal haplotype, and 0 otherwise. The INFO metric was computed from the

genotype probabilities following the approach in Section 3.7, with adaptations for multi-allelic variants. Each simulation scenario was repeated 100 times, and the measurements were averaged across replicates to reduce spurious variation.

Obtaining and filtering the list of known SVs. I downloaded two files from this repository (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/release/v2.0/integrated_callset/): a manifest of all SVs recording their basic information as well as a vcf file that stores the exact genomic sequences and some extra information [343, 354]. Based on the simulation results, I only attempted deletions that were large (SV length > 5 kb) and common enough (at least one of the estimated African allele frequency in the original paper, POP_AFR_AF or PG_INFO_AFR, is greater than 5%). This filter retrieved 621 entries for downstream analysis.

Accessibility of SVs by short-read sequencing data. To address alignment artefacts by short-read sequencing data, I adopted a strategy akin to the original glycophorin paper [345]. I obtained a previously curated multi-read mappability track with 100-mers throughout the human genome [322]. At each base, it represents the probability that a randomly selected 100-mer which overlaps with a given position is uniquely mappable to the region. For each genomic bin, I calculated the average mappability by averaging values for each base and the mapping quality by averaging values for each read aligned to the region (output by coverotron), and I considered it to be mappable if average

mappability was greater than 0.9 and mapping quality was greater than 20. I excluded SV regions where less than 25% genomic bins were mappable (192 out of 621 was removed) and excluded non-mappable bins in the inference procedure for the rest.

Running lcSV on the SV list variants. Since different SV alleles may cover overlapping genomic regions, I expanded the original breakpoints for a target SV to enclose all adjacent SVs as identified in the full SV list (before filtering on SV length and allele frequency) that intersected with the current region, until no overlapping was present. Three additional genomic bins (always of size 1 kb in this run, although customisable for other purposes) in both upstream and downstream direction were included to allow for clear visualisation of the boundaries. This strategy is crucial for our method, as lcSV is capable of finding all SVs that are large and common enough in a population due to its design in modelling SVs as segregating haplotypes, which otherwise biases the results by chunking surrounding SVs to incomplete segments. Then, mean and variance of genome coverage were then estimated from flanking windows of 1 Mb on each side, excluding non-mappable regions or until reaching a chromosome end. These coverage metrics were supplied to lcSV for SV calling. The algorithm was configured to generate 200 recombinant haplotypes per iteration, with premature termination if no improved model (lower score) was found in the preceding 100 iterations, and a maximum of 2,000 iterations otherwise.

To evaluate each lcSV run, all haplotypes in the best SV model were extracted. Since true individual-level SV genotypes were not available, accuracy was assessed at the model haplotype level. Model concordance was defined as 1 if the true haplotype was retrieved by the model, by inferring the type and position of copy numbers for each haplotype, or 0 otherwise. Two types of INFO were also output by the model: the true allele INFO if concordance was 1 and 0 if otherwise, as well as the max allele INFO that was associated with the most frequent non-reference SV haplotype. Individual genotypes were called based on the best SV model, and all metrics were recorded in the output file for downstream analysis.

5.5 Conclusion

This chapter begins with a motivating example in Section 5.1, focusing on the glycophorin region, a biologically important locus that harbours multiple SV haplotypes and confers resistance to parasite invasion of erythrocytes. To assess whether SVs could be inferred from lcWGS data, PCA was first applied to genomic coverage profiles, revealing distinct patterns among individuals. While this approach demonstrated feasible for SV detection, it was limited by reliance on manual interpretation, restricted resolution, and indirect genotype inference. Motivated by these deficiencies as well as the lack of specific method for calling SVs in the lcWGS context, I developed lcSV, a population-based model that searches for SVs by stochastically generating recombinant haplotypes, evaluating their fit to observed coverage data, and

inferring genotypes from a common pool of haplotypes. The utility of lcSV was illustrated through two *in silico* experiments. In the first, lcSV successfully resolved overlapping deletions in a haplotype-aware manner by jointly inferring diplotypes from a shared haplotype pool, rather than treating each individual separately and recombining haplotypes across individuals *post hoc*. In the second simulation, lcSV successfully reconstructed complex haplotype resembling the Dantu antigen, which likely arose from multiple recombination events, although convergence required a large number of iterations. In practical settings, pinpointing the precise SV breakpoints remained challenging and would necessitate molecular validation for confirmation. Finally, applying lcSV to the glycophorin region in the GAMCC cohort reassured the presence of the common DEL1 and DUP1 haplotypes, supported by distinct coverage patterns across individuals with different genotypes and fully consistent with PCA.

To further evaluate the potential of lcSV across diverse genomic regions, I conducted simulation studies in Section 5.2 to define the detectable thresholds of allele frequency and SV length under a population size of 210 individuals sequenced at ~1× coverage, mirroring the setting of the real GAMCC cohort. Because deletions and insertions differ in their underlying characteristics, with insertions may involve alignments to non-flanking sequences related to various SV formation mechanisms and in turns distort coverage estimates, I restricted the following analyses to only deletions. The simulations revealed that SV size has a major impact on calling results, with reliable

identification typically requiring SV larger than 4 kb, while allele frequency has a milder effect, with variants common than 5% allele frequency detected robustly.

Building on these results, Section 5.3 applied lcSV to previously reported deletion SV regions, attempting calls at 429 loci and successfully identifying plausible target haplotypes in 87.6% of cases. These results corroborated earlier observations that read alignments and SV size strongly influence SV calling outcomes, as reflected in the INFO score distributions. Visualisation of several genic regions further supported these calls, with distinct genome coverage patterns observed for individuals with different genotypes, in both bi-allelic and multi-allelic contexts. To better understand factors influencing our method and limitations inherent to lcWGS, I examined three illustrative examples: a rare allele (allele frequency 1/420) missed by lcSV; an allele called with incorrect breakpoints, likely due to data sparsity; and an SV hotspot near a chromosome end comprising four possible haplotypes, where previous uncharacterised haplotypes in this cohort may present. Collectively, while lcSV effectively identifies common, well-mapped SVs from lcWGS, variant detection is constrained by allele frequency, local genome coverage, breakpoint resolution, and genomic context. Careful region selection is therefore critical to ensure reliable population-scale SV genotyping. Hence, I compiled a set of 324 regions that are large, common, non-telomeric, and accessible by short-read sequencing in the Gambian population as reliably identifiable by lcSV.

Together with Chapter 4, this chapter demonstrated the utility of lcWGS not only for capturing genome-wide variants but also for interrogating highly polymorphic loci, such as the HLA region, and complex SVs. These findings further underscore the viability of lcWGS for variant detection in population-scale studies across global populations. Further discussion of the limitations and implications of this chapter is provided in Section 7.2.3.

Chapter 6 Genome-Wide Association Study Using Low-Coverage Whole-Genome Sequencing Data in a Vietnamese Hepatitis C Virus Cohort

The previous chapters demonstrated that lcWGS combined with imputation can provide accurate genotyping across a diverse range of genomic variants, including SNPs, indels, and larger classes of variation such as SVs. This performance can be further enhanced by incorporating population-specific haplotypes and large-scale global reference panels. Building on this foundation, this chapter presents a GWAS conducted using lcWGS to investigate the influence of host genomic factors on Hepatitis C Virus (HCV) infection in a Vietnamese cohort. Section 6.1 describes the study cohort, sample demographics, sequencing performance, and imputation results, while Section 6.2 details the main GWAS analyses on genome-wide markers and HLA alleles. Collectively, this chapter highlights the practical utility of lcWGS for genotype-phenotype association studies, demonstrating its value not only for variant discovery but also for deriving meaningful biological insights.

6.1 A Vietnamese cohort with Hepatitis C Virus infection

6.1.1 Overview of the Hepatitis C Virus

HCV is an enveloped, positive-sense, single-stranded RNA virus of the *Flaviviridae* family and a major cause of chronic liver disease worldwide [355]. More than 50

million people are estimated to live with chronic infection, which can progress to fibrosis, cirrhosis, and hepatocellular carcinoma if untreated [356, 357]. HCV is genetically diverse, with at least seven major genotype groups and numerous subtypes that differ in their global distribution, transmission patterns, and therapeutic responses [355, 358]. Despite the availability of highly effective direct-acting antivirals, the global burden of HCV remains substantial and far from eradication [359]. Viral replication primarily occurs in hepatocytes, and disease outcomes are shaped by a complex interplay between viral genetic variation, host immune responses, and environmental factors [360].

In this chapter, I focused on investigating the impact of host genetics variants on HCV infection. Previous GWAS have identified many variants that are associated with viral clearance and treatment response [361-363]. Among them, rs12979860 residing upstream of the IFNL3 gene was identified as a strong SNP associated with both treatment response and spontaneous clearance [197, 364, 365], while subsequent work revealed a dinucleotide frameshift variant rs368234815, in strong LD with rs12979860, as the likely causal variant [202] and thus creating *IFNL4* as a new gene. Nevertheless, expression of *IFNL4* is thought to impair viral clearance by inducing a less effective interferon response [366]. Other variants, including rs2596542 (near *MICA* [367]), rs1012068 (near *DEPDC5* [368]), and rs17047200 (intronic variant of *TLL1* [369]) in Japanese populations, rs2071286 (intronic variant of *NOTCH4* [370]) across global

populations, as well as several HLA alleles [371-373], were also shown to influence HCV outcomes.

6.1.2 Data collection and generation

To investigate host genetic factors on the impact of HCV disease outcome in a Vietnamese population, as well as demonstrate the utility of lcWGS in providing biological insights in GWAS, we obtained 874 samples, along with metadata and HCV viral load measurements, from two previous studies of adults with chronic HCV infection in Vietnam, including individuals with mild liver disease (the SEARCH project [275]) and with advanced fibrosis or compensated cirrhosis (the VIETNARMS project [374, 375]). Majority of patients were infected with HCV genotype 6 determined from virus sequencing, with full sample demographics summarised in Table 6.1.1. Library constructions were again performed using the UIIFS kit by aiming for 100 ng input, although dependent on the supplied sample. Unlike the GAMCC cohort in Chapter 3, we performed lcWGS on all samples in batches using the Element Biosciences AVITI platform, generating 151 bp paired-end reads with a target of ~1× genome coverage, at the University of Oxford. The AVITI platform has been shown to generate higher-quality base-calls (Q40 versus Q30 on Illumina platforms), particularly for longer libraries, as it does not require the additional PCR amplification during cluster generation which is error prone [70, 376]. To improve coverage of the highly polymorphic HLA region, we applied an additional hybridisation capture approach targeting HLA class I and II loci on 711 samples, while the remaining samples could

not be processed due to insufficient DNA material. This is described further below. Unlike the GAMCC cohort in Chapter 3, no truth data (deep whole-genome sequencing replicates, microarray-based genotyping, or HLA sequence-based typing) were generated for comparison for this cohort.

Age in years	45 (18-80)
log_{10} viral load	6.0 (1.2-8.2)
Sex	
Male	421 (73.5%)
Female	152 (26.5%)
HCV genotype	
Genotype 1	225 (39.3%)
Genotype 2	49 (8.6%)
Genotype 3	14 (2.4%)
Genotype 6	285 (49.7%)
Total patients	**573**

Table 6.1.1. Sample demographics for 573 samples in the Vietnamese cohort. These samples were included in the final GWAS analysis (Section 6.2) that passed sample QC (Sections 6.1.3, Section 6.1.4, and Methods) and had paired viral load data.

6.1.3 Genotype imputation and HLA inference

In this section, I summarised the main steps of lcWGS processing and imputation. Full details are given in the Methods section. As in Chapter 3, I ran all samples through the lcWGS pipeline that I developed with several changes. First, for a subset of samples, sequencing yield was initially insufficient due to library preparation failure or low DNA input. These samples were subsequently boosted with additional sequencing runs to achieve adequate coverage. Together with the HLA capture, all data generated for each sample was merged at fastq level before running through the lcWGS pipeline for adapter trimming, read alignment to the GRCh38 reference genome, and duplicate

removal. Similar to the strategy in Chapter 3 to enhance genotype imputation, I combined 1,163 genomes in the pilot GenomeAsia 100K reference panel (in GRCh37 coordinates, subsequently liftover to GRCh38 using picard) [105], which I was granted access, with the 1000 Genomes Project reference panel (subset to 2,504 unrelated individuals), using IMPUTE2 v2.3.2 [377] by imputing the two reference panel into each other (see Methods for details). The resulting reference panel comprised 3,667 individuals and was subsequently used for QUILT imputation. To facilitate timely processing of samples as well as for computational purposes, imputation was performed separately for each batch as samples arrived and was not affected in accuracy as QUILT treats each individual genome separately. Then, all batches were merged at this stage to minimise batch artefacts, together subjected to INFO, allele frequency, and the Omni5M variant filters as in Chapter 3. Eventually, these results were uploaded to the TOPMed server for the second stage imputation.

On the other hand, the additional HLA capture enabled alternative approaches for HLA inference. In addition to running the modified QUILT-HLA workflow (referred as QUILT-HLA throughout this chapter) with the 1000 Genomes Project reference panel and applying the two-stage multi-ethnic reference panel approach described in Chapter 4, I also attempted direct HLA typing using HLA-LA [170], a graph-based method known for high accuracy with deep whole-genome sequencing data. HLA-LA was developed from its precursor HLA-PRG [169], with improved computational efficiency

and reduced genome coverage requirements for reliable HLA typing (although still needs 15×). Many genomes in the Vietnamese cohort exceeded this coverage at various loci, providing an opportunity to evaluate the accuracy of HLA-LA in this setting and to explore the additional benefits of a blended sequencing strategy, that is, targeted capture of highly polymorphic regions combined with lcWGS. Again, all analyses were restricted to HLA-A, -B, -C, -DQB1, and -DRB1, and allele calls from all methods were converted to G-group resolution to match the output of the two-stage imputation using the HLA multi-ethnic panel.

6.1.4 Sample performance and imputation results

To characterise sequencing performance in the Vietnamese cohort, Figure 6.1.1 summarises genome coverage and other quality metrics. Genome coverage ranged from 0.01× to 4.99× (mean 1.17×), reflecting moderate pooling imbalance and variation in capture. Even after repeat sequencing of low-coverage samples, a subset remained underrepresented. In total, 7% of all samples had at least 60% of the genome uncovered. Because imputation performance is strongly influenced by genome coverage, genotype inference for these individuals was less reliable (as confirmed in Figure 6.1.2), and they were therefore excluded from downstream analyses. In contrast, the use of the AVITI platform resulted in higher-quality sequencing data than that observed in the GAMCC cohort described in Chapter 3, with substantially reduced duplication rates (mean 0.88% vs. 8.50%). Among individuals with sequencing depths between 0.5× and 1.5× (where most GAMCC samples fall) the Vietnamese cohort also showed a smaller proportion

of the genome not covered (mean 38.2% vs. 41.3%) achieved by lower average coverage (mean 1.02× vs. 1.13×).

As shown in Figure 6.1.1b, eleven individuals with mismatched genomic and self-reported sex were identified and thus also excluded from subsequent analyses. Figure 6.1.1d reflects variability in genome coverage across HLA loci (HLA-A: 16.34× [0-146.0×]; HLA-B: 13.82× [0-178.4×]; HLA-C: 15.20× [0-186.0×]; HLA-DQB1: 1.18× [0-4.8×]; HLA-DRB1: 7.00× [0-41.5×]). The suboptimal capture of HLA class II loci is evident from their reduced coverage, particularly at HLA-DQB1 where capture was largely unsuccessful. In contrast, coverage at HLA class I loci improves by approximately ten-fold compared with raw lcWGS data.

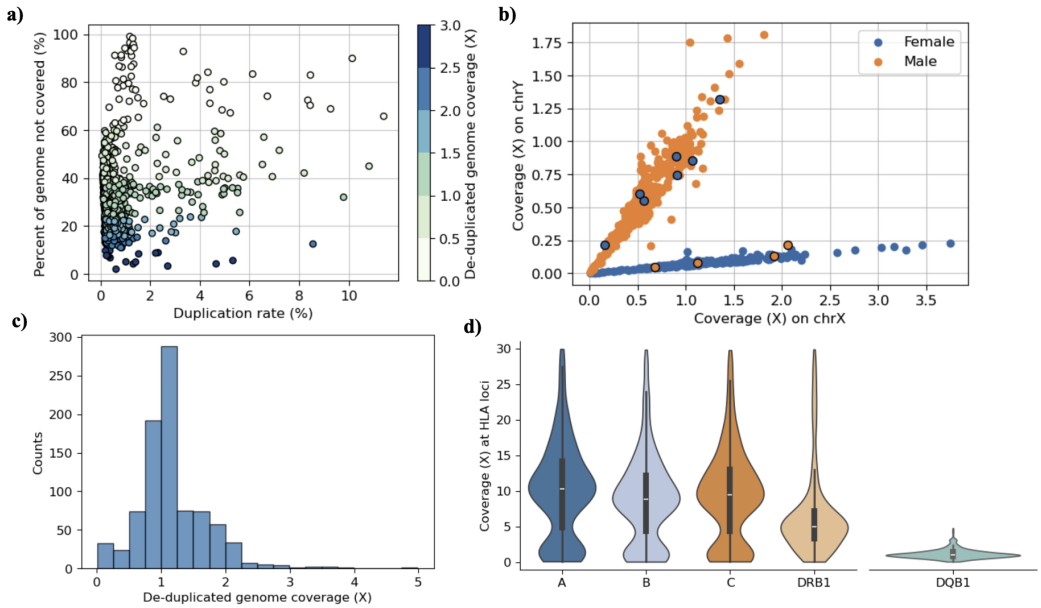


Figure 6.1.1. Sample performance for 874 GAMCC lcWGS samples. a) Proportion of genome not covered (y-axis) against duplication rate (x-axis), coloured by de-duplicated genome coverage. **b)** Genome coverage on sex chromosomes. Markers are coloured with self-reported sex. A sample was called female if X-chromosome coverage was more than four times that of the Y-chromosome, and male

otherwise. Eleven samples show discrepancies compared to sex inferred from genome coverage. **c)** Histogram of de-duplicated genome coverage (range 0.01×-4.99×, mean 1.17×). **d)** Distribution of genome coverage at HLA-loci. For visualisation purposes, HLA-A, HLA-B, HLA-C, and HLA-DRB1 are plotted together and capped at a maximum of 30× coverage, while HLA-DQB1 is displayed separately with a disproportionally narrow width (though in reality it would correspond to a much wider representation). A total of 205 samples had an average HLA coverage of < 5×, of which 163 did not undergo additional HLA capture due to insufficient DNA material, resulting in the two peaks of HLA coverage in d).

Since no truth data were generated for this cohort, lcWGS imputation performance could only be assessed through imputation confidence (Figure 6.1.2). For this evaluation, I focused on a 20 Mb region at chr6:40,000,000-60,000,000. At the per-individual level, Figure 6.1.2a shows that imputation performance, measured as the average maximum genotype probability across loci, did not correlate with HCV viral load. Figure 6.1.2b explores the relationship between duplication rate and proportion of genome uncovered with imputation confidence. While duplication rate was not a major determinant, samples with greater genomic regions lacking reads consistently showed reduced imputation confidence. This finding illustrates that imputation is facilitated by available sequence information and further supports the principle that higher and more evenly distributed genome coverage improves imputation accuracy. Finally, at variant level, Figure 6.1.2c demonstrates that imputation confidence increases with allele frequency. These results show that QUILT imputation produced data of sufficient quality. At this stage, I removed 11 individuals of discordant sex, 21 assigned to different identifiers (Methods), and 59 with less than 40% genome covered by at least one base out of 874 sequenced samples, retaining 783 for downstream analyses.

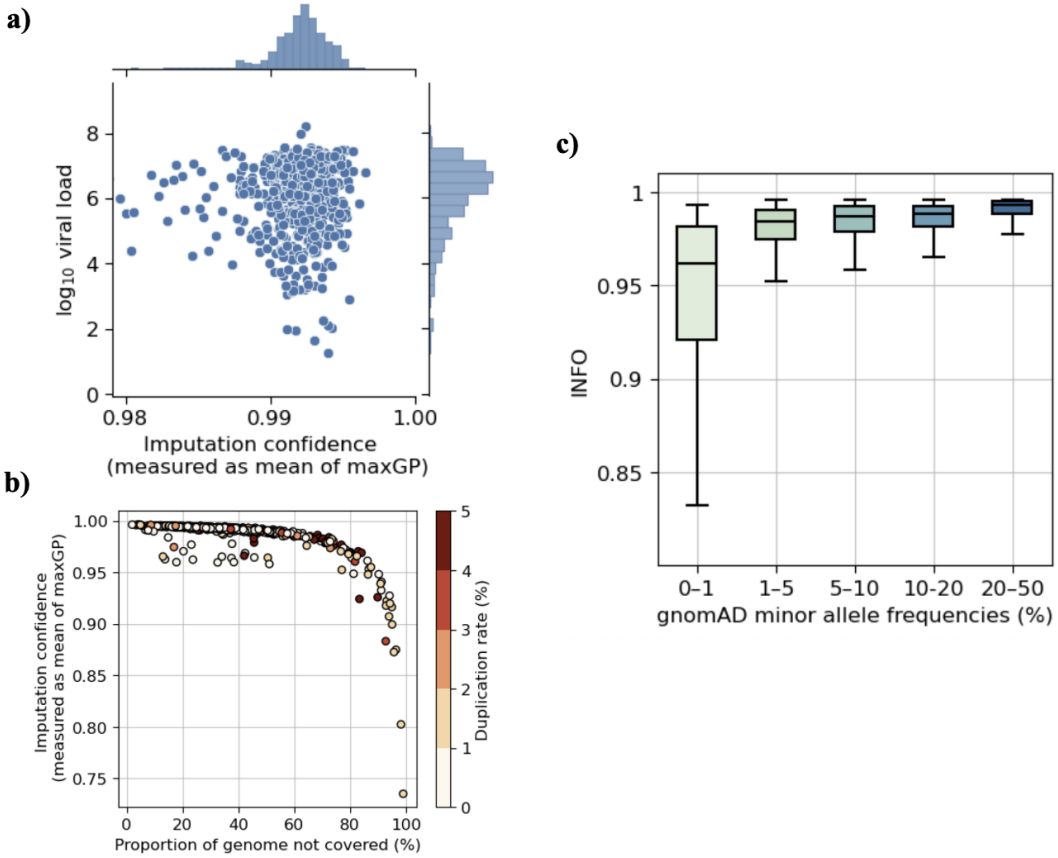


Figure 6.1.2. Genome-wide imputation performance in the Vietnamese cohort, measured on a 20 Mb region (chr6:40,000,000-60,000,000). a) Sample $\log_{10}$ viral load (y-axis) against imputation confidence (x-axis, calculated as average maximum genotype probability of an individual across loci), clamped at minimum 0.979 on x-axis for visualisation purposes. Distributions of both axes are shown on the marginal histograms. **b)** Imputation confidence (y-axis, similar as a) against proportion of genome not covered (x-axis), coloured by duplication rate. Each marker represents an individual in both a) and b). **c)** Distribution of imputation INFO (y-axis) across gnomAD minor allele frequencies (x-axis) for each variant. If an alternative allele frequency recorded in gnomAD was greater than 0.5, I reverted the reference and alternative allele to obtain its minor allele frequency.

In addition to genome-wide variants, three workflows (HLA-LA calling, QUILT-HLA imputation, and two-stage imputation using the HLA multi-ethnic reference panel) were employed for HLA inference at G-group resolution, as described in Section 6.1.2. Since no sequence-based typing was performed in this cohort, HLA imputation performance could only be assessed by comparing results across the methods. Figure 6.1.3a shows

that QUILT-HLA and the HLA multi-ethnic imputation were the most concordant, although substantial discrepancies remained. This comparison does not necessarily indicate underperformance of either method, as differences may arise from multiple factors, including the absence of truth data for precise validation, imputation uncertainty due to limited representation of Vietnamese haplotypes in the reference panels [292], and technical artefacts in allele representation. Concordance with HLA-LA was generally lower, particularly at HLA class II loci, especially for HLA-DQB1. This pattern is illustrated in Figure 6.1.3b, where HLA-LA shows high uncertainty, with many estimates clustering near zero. The mean confidence values for HLA-LA at HLA-A, -B, -C, -DQB1, and -DRB1 are 0.790, 0.770, 0.769, 0.013, and 0.411, compared with 0.968, 0.971, 0.993, 0.986, and 0.926 for QUILT-HLA, as the former depends only on reads aligning directly to HLA genes while the latter incorporates LD information from surrounding genomic regions. Confidence metrics were not calculated for the HLA multi-ethnic reference panel due to allele missingness (as bi-allelic calls were not forced at each locus), which would distort the measure. Similar patterns were observed in Figure 6.1.1 and could be attributed to limited HLA coverage even with additional capture, which constrained the full potential of HLA-LA.

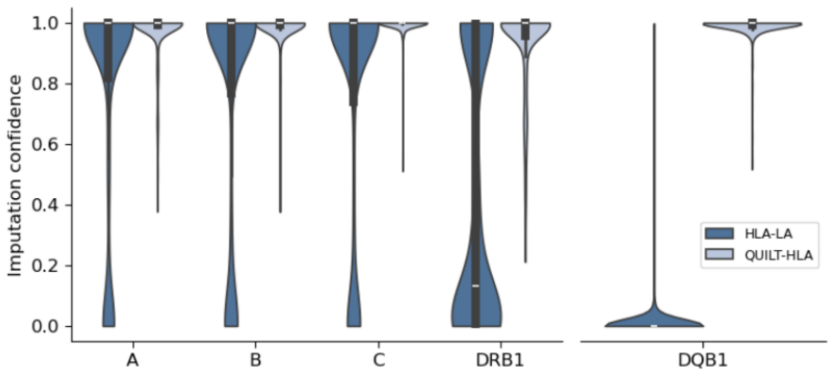
To generate consensus HLA calls for each individual, I combined the three methods using a *post hoc* strategy that considered both called alleles and confidence scores (see Methods for details). First, allele calls with confidence below 0.1 were discarded. Next,

for each locus in each individual, confidence scores were averaged across reported alleles to reflect overall support, and the two highest-confidence alleles were retained as probable calls. Finally, loci with combined confidence below 0.7 were set to missing to reduce false positives due to uncertainty. Using this approach, 7,202 allele copies (out of 7,830; 91.98%) were assigned as non-missing among 783 diploid individuals passing QC across the five HLA loci. Figure 6.1.3c shows the concordance of all methods relative to this consensus. QUILT-HLA imputation was over 90% concordant at HLA-B and HLA-C and contributed most to the consensus calls. However, I could not determine absolute accuracy, as no truth data were available.

a)

Locus	HLA-LA vs QUILT-HLA	HLA-LA vs multi-ethnic	QUILT-HLA vs multi-ethnic
A	0.757	0.653	0.760
B	0.679	0.653	0.858
C	0.670	0.645	0.894
DQB1	0.370	0.351	0.786
DRB1	0.408	0.305	0.422

b)



c)

Locus	HLA-LA vs consensus	QUILT-HLA vs consensus	multi-ethnic vs consensus
A	0.791	0.890	0.789
B	0.709	0.921	0.886
C	0.686	0.931	0.895
DQB1	0.321	0.842	0.753
DRB1	0.474	0.730	0.615

Figure 6.1.3. HLA calling of 783 individuals in the Vietnamese cohort. a) Pairwise concordance

between the three HLA workflow at each locus, averaged across individuals. **b)** Distribution of inference confidence for HLA-LA and QUILT-HLA. As in Figure 6.1.1, HLA-DQB1 is plotted on a compressed scale to facilitate visualisation. **c)** Concordance of the three HLA workflow to a consensus call (see main text and Methods) at each locus, averaged across individuals.

To summarise, I obtained 783 samples with 5,145,539 high-quality autosomal variants and 109 consensus-based HLA G-group allele calls, which I took forward to the GWAS study described in Section 6.2.

6.1.5 Inference on population structure in the Vietnamese cohort

In Section 3.3, I showed that common SNPs from lcWGS imputation can reveal population structure in the GAMCC cohorts. To assess population structure within this Vietnamese cohort, I performed PCA using an approximately independent set of markers at least 50 kb apart, omitting the HLA (chr6:25,000,000-40,000,000) and a common inversion region (chr8:7,500,000-12,000,000) [378] to avoid LD distortion, yielding 38,981 markers (see Methods for details). Twenty-one individuals related at second degree or closer, identified by KING [379], were excluded from PCA decomposition and projected back afterwards. Figure 6.1.4 shows no obvious population structure within the cohort, suggesting a relatively homogeneous population and no artifacts related to sequencing batch, viral load, or genome coverage. When analysed alongside individuals from the GenomeAsia-enriched 1000 Genomes Project reference panel, the cohort clustered with other East Asian populations, confirming that the data are suitable for GWAS analyses.

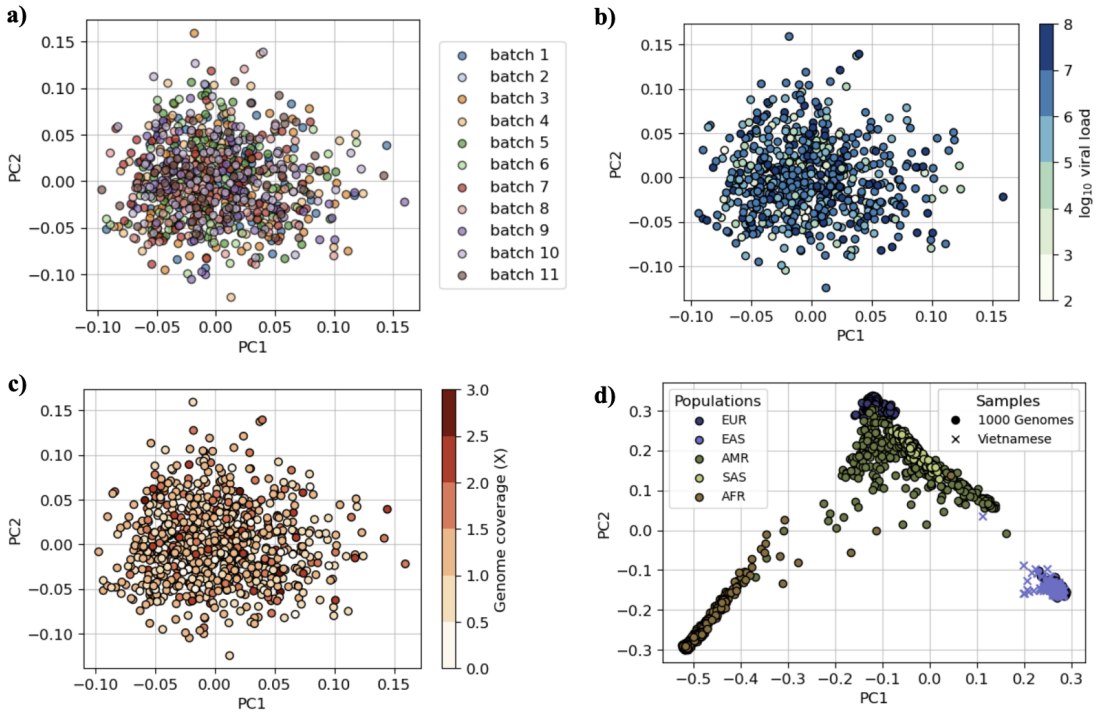


Figure 6.1.4. Population structure in the Vietnamese cohort. a-c) The first two PCs coloured by sequencing batches (a), viral load (b), and genome coverage (c), respectively. **d)** The first two PCs calculated from a joint PCA to the 783 individuals in the Vietnamese cohort together with 3,667 individuals in the GenomeAsia-enhanced 1000 Genomes Project reference panel, running on the same set of markers. Markers are coloured based on super populations and labelled with circles and crosses for individuals from the reference panel and the Vietnamese cohort, respectively.

6.2 Genome-wide association study of host genetic variants influencing Hepatitis C Virus viral load

In this cohort, I then performed a GWAS investigating the relationship between host genetic variants and HCV viral load. Of the 783 individuals passing QC, 210 had missing phenotype data and were excluded, leaving 573 individuals for further GWAS analysis. Table 6.1.1 presents sample demographics for these individuals. Both genome-wide variants and HLA alleles were examined for their contribution to the phenotype using a linear mixed model, which additionally accounts for sample relatedness [380].

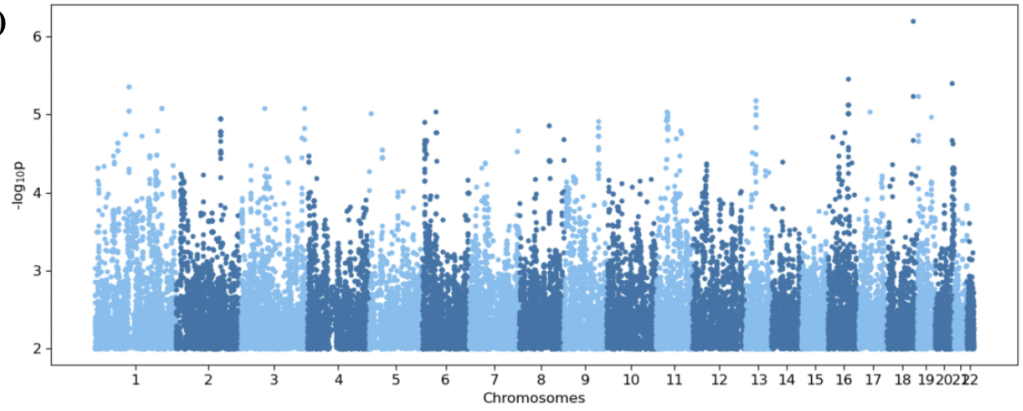
The genetic relatedness matrix (GRM) was estimated using the same 38,981 markers employed in PCA (Section 6.1) from post-QC QUILT-imputed data using the software GEMMA [381]. Genome-wide variants were obtained from two-stage imputation (QUILT imputation with the GenomeAsia-enriched 1000 Genomes Project reference panel followed by TOPMed imputation), subject to filters on Hardy-Weinberg equilibrium $p > 1\times10^{-6}$, minor allele frequency > 5%, and TOPMed imputation $R^2 > 0.5$ to the final 5,145,539 autosomal variants. From the consensus calls described in Section 6.1.4, I identified 109 HLA alleles at G-group resolution with allele frequencies > 1%. Eventually, GWAS was then conducted using the GRM and covariates (age, sex, HCV genotype, and the first five host PCs), assuming additive effects of allele dosages, on both genome-wide variants and on HLA alleles. Further details are available in the Methods section.

6.2.1 Impact of genome-wide variants on viral load

Figure 6.2.1 shows the GWAS results for host genetic variants, with the top ten variants summarised in Figure 6.2.1b and annotated with their nearest protein-coding genes within a 1 Mb flanking region (retrieved from the Ensemble REST API, as in Chapter 5 and see Methods for details). No variants reached $p < 5 \times 10^{-8}$ or Benjamini-Hochberg False Discovery Rate (FDR) 5% [382]. The most statistically significant signal was rs34292649 ($p = 6.32 \times 10^{-7}$) on chromosome 18, approximately 150 kb upstream of the gene neuropilin and tolloid-like 1 (*NETO1*). NETO proteins help control the activity and placement of certain glutamate receptors in neurons, making them important for

proper communication between brain cells and for processes like learning and memory [383, 384]. To investigate potential functional consequences, the top ten variants were uploaded to the Ensembl Variant Effect Predictor (VEP) [385] website (https://www.ensembl.org/Tools/VEP), which predicted rs34292649 as an intergenic variant. Because this variant has a low allele frequency (5.4%) and lacks supporting biological evidence, I instead focused on rs6142998 (chr20:62228407:A>G), predicted by VEP as a regulatory variant residing in an enhancer of Oxysterol Binding Protein Like 2 (*OSBPL2*). This variant is further explored in Figure 6.2.2.

a)



b)

Chr.	Position	Rsid	Ref.	Alt.	Alt. allele frequency (%)	Effect size (β)	Standard Error	P value	Nearest gene (within 1 Mb)	Distance (bases)
18	73027897	rs34292649	A	C	5.4	-0.718287	0.142541	6.32e-07	NETO1	159751
16	63322764	rs8056922	A	C	6.1	-0.624907	0.133306	3.47e-06	N/A	0
20	62228407	rs6142998	A	G	35.4	-0.317436	0.068120	3.96e-06	OSBPL2	3515
1	97504833	rs147850929	G	C	5.6	-0.633878	0.136679	4.39e-06	DPYD	0
1	97504730	rs142001424	A	G	5.6	-0.633878	0.136679	4.39e-06	DPYD	0
19	1721958	rs73513267	G	T	5.5	-0.606392	0.132483	5.81e-06	ONECUT3	31548
18	73027869	rs146499545	AT	A	5.8	-0.626031	0.136835	5.86e-06	NETO1	159723
13	70976925	rs1553591	A	T	31.6	0.308405	0.067779	6.58e-06	DACH1	461041
13	70976967	rs1553592	T	A	31.6	0.308405	0.067779	6.58e-06	DACH1	460999
16	63342313	rs7185751	A	G	6.4	-0.592595	0.131080	7.52e-06	N/A	0

Figure 6.2.1. GWAS of host genetic variants associated with $\log_{10}$ HCV viral load in the Vietnamese cohort of 573 individuals. a) P values were calculated under an additive model of association, and all variants with p > 0.01 were excluded from this graph. **b)** Ten leading variants annotated with nearest protein-coding genes. Distance was set to 0 if no protein-coding genes exist or the variant resides within the nearest gene. All SNPs and indels were reported according to the gnomAD convention in GRCh38 coordinates, using the forward (reference) strand. Chr.: Chromosome; Ref.: Reference; Alt.: Alternative.

The genomic context of rs6142998 is shown in Figure 6.2.2a. This variant is in LD with multiple markers across the OSBPL2 gene, and association analysis conditioning on rs6142998 abolishes those signals, suggesting its potential functional role in modulating *OSBPL2* expression or splicing. This is further supported by colocalisation analysis using data from the Genotype-Tissue Expression (GTEx) project [386] (https://www.gtexportal.org/home/), where the alternative G allele is associated with reduced splicing efficiency at the *OSBPL2* intron and likely reflect altered transcriptional regulation. A plausible mechanism underlying this association is that rs6142998 reduces *OSBPL2* expression by interfering splicing and thereby decrease oxysterol binding protein-related protein 2-mediated enrichment of plasma membrane cholesterol level [387]. These alterations in lipid homeostasis disrupt the intracellular lipid environment required for HCV replication [388] and thus result in reduced viral load [389, 390]. While no previous literature has reported this association, this finding suggests a potential link between rs6142998 and HCV susceptibility, which requires experimental validation in relevant cell models to confirm causality.

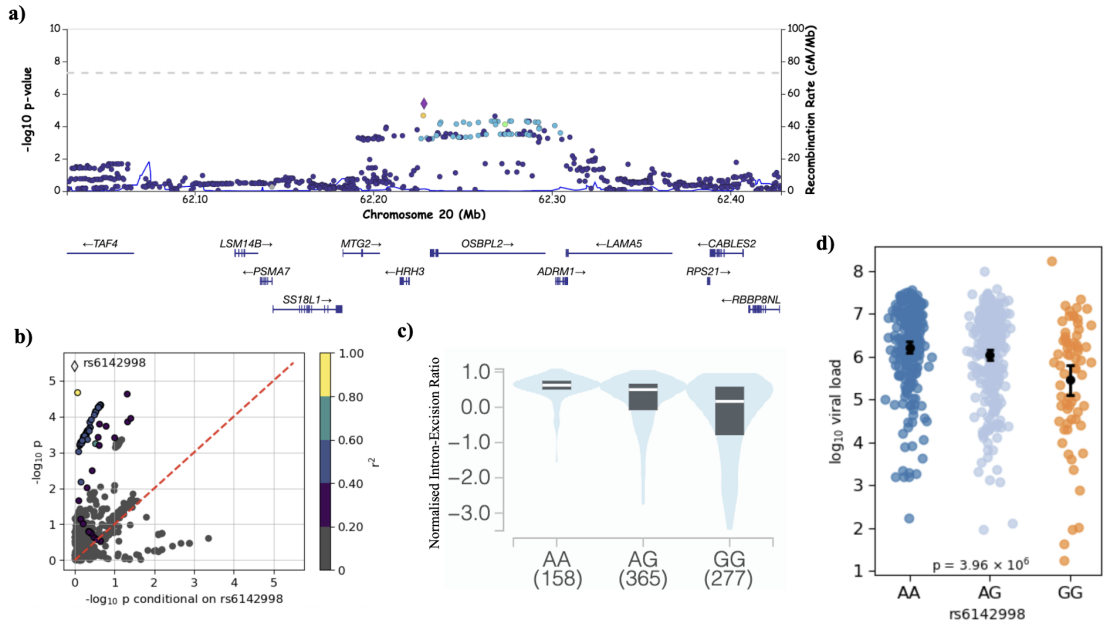


Figure 6.2.2. Evidence of association of rs6142998. a) Hitplot of the leading variant rs6142998 (chr20: 62228407:A>G), centred at a 400 kb region (chr20:62028407-62428407, x-axis) with protein-coding genes shown below. Markers are coloured by LD estimated from TOPMed imputed genotypes. **b)** Association statistics -$\log_{10}$ p from the main GWAS (y-axis) against -$\log_{10}$ p conditioned on rs6142998, assuming additive effect of allele dosage. **c)** Colocalisation with splice quantitative trait locus (sQTL) from whole blood, obtained from GTEx website. **d)** Distribution, estimated mean and 95% confidence interval (shown as black dots and bars, respectively) of $\log_{10}$ viral load stratified by rs6142998 genotypes.

On the other hand, I examined the well-known association of rs12979860 in *IFNL4*, which is in strong LD with the causal dinucleotide variant rs11322783 (the single base deletion in rs368234815). Imputation results confirmed perfect linkage between the two variants in this cohort and supported the favourable effect of the C allele. Nevertheless, the association did not reach nominal significance (p = 0.222, effect size -0.153, and standard error 0.125), corresponded to only 23% power at nominal significance level at 0.05 and essentially no power at genome-wide significance [391]. Independent genotyping of rs12979860 by PCR showed a similar effect and reduced the standard error (p = 0.160, effect size -0.153, and standard error 0.125), and comparison against

imputed genotypes for this SNP revealed concordance of 0.912. While many other cohorts have reported rs12979860 as strongly associated (at genome-wide significance) with viral load or viral clearance at a similar population size [197, 364, 365, 392], it is important to note that this cohort primarily comprises HCV genotype 6 infections. To our knowledge, no large GWAS has yet documented the effect of rs12979860 in genotype 6 [393, 394]. This suggests that the impact of rs12979860 on viral control may be dependent on viral genotype (effect sizes are different for HCV genotype 1 and genotype 6, Figure 6.2.3c), as also suggested by previous studies [395, 396]. Consequently, the effect of rs12979860 could be more context-specific, highlighting the need for dedicated studies to confirm this.

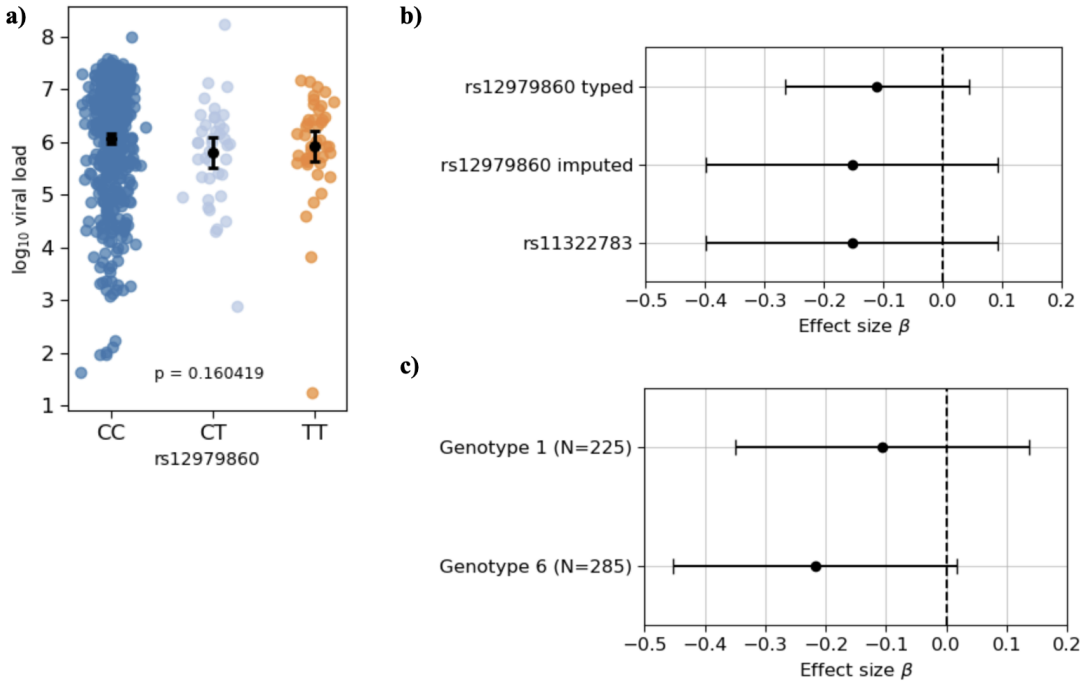


Figure 6.2.3. Evidence of association for rs12979860 and rs11322783. a) Distribution, estimated mean and 95% confidence interval (shown as black dots and bars, respectively) of $\log_{10}$ viral load stratified by rs12979860 genotypes based on independent genotyping rather than lcWGS and imputation. The indel

rs11322783 is in perfect LD with rs12979860. **b)** Forest plot of the 95% confidence intervals for effect sizes of typed and imputed rs12979860, identified as -0.111 (standard error 0.0789), and -0.153 (standard error 0.125), respectively. The dashed line indicates effect size 0. **c)** Forest plot of the 95% confidence intervals for effect sizes of typed rs12979860 stratified by HCV genotypes, identified as -0.106 (standard error 0.124), and -0.217 (standard error 0.120), respectively. The dashed line indicates effect size 0. The number of individuals infected with each HCV genotype is also indicated in the figure.

6.2.2 Impact of HLA alleles on viral load

I also investigated the impact of HLA alleles on HCV viral load, analysing 109 alleles with at least 1% frequency from the consensus HLA calls. The only allele approaching nominal significance was HLA-DQB1*06:01:01G (p = 0.0527, Figure 6.2.4). This allele has been previously associated with chronic HBV infection [397, 398] and several autoimmune diseases including systemic lupus erythematosus [399] and Sjögren's disease [400]. Its association with increased viral load in our cohort may involve competition in antigen presentation between viral and self-peptides. However, I did not pursue further analyses, both due to this allele's marginal association and the limited HLA imputation accuracy, which could compromise the reliability of association results.

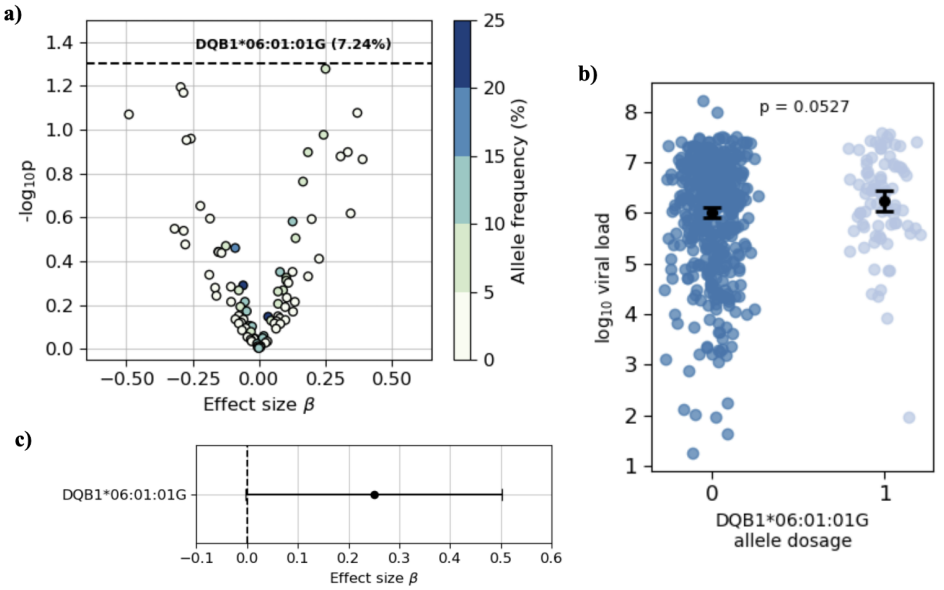


Figure 6.2.4. Evidence of association of the HLA alleles. a) Association statistics -$\log_{10}$ p (y-axis)

against effect size (x-axis), with each HLA allele coloured with by allele frequency in this cohort. The dashed line indicates nominal significance (p = 0.05). The most significant allele is HLA-DQB1*06:01:01G, yet only marginally significant (p = 0.0527). **b)** Distribution, estimated mean and 95% confidence interval (shown as black dots and bars, respectively) of $\log_{10}$ viral load stratified by HLA-DQB1*06:01:01G allele dosage. **c)** Forest plot of the 95% confidence intervals for effect size of HLA-DQB1*06:01:01G (effect size 0.250, standard error 0.129).

6.3 Methods

Data collection. Samples involved in this chapter were obtained from two previous studies of adults with chronic HCV infection in Vietnam, including individuals with mild liver disease (the SEARCH project [275]) and with advanced fibrosis or compensated cirrhosis (the VIETNARMS project [374, 375]). Sample demographics were described in the main text. A total of 874 samples underwent lcWGS with the Element Biosciences AVITI platform for 151 bp paired-end sequencing, targeting ~1× genome coverage. An additional sequencing targeting the HLA region (with in-house probes) was applied to all samples to capture HLA alleles, and the resulting files were merged at fastq level.

Data processing and genome imputation. The processing procedure largely followed that of previous cohorts (Chapters 2 and 3), with only two main differences in the bioinformatics workflow highlighted in this section. First, because this cohort consisted of Vietnamese individuals, an East Asian-enhanced reference panel was created for QUILT imputation, emulating the approach used in Chapter 3 with MalariaGEN genomes. Specifically, I was granted access to 1,163 genomes from the pilot

GenomeAsia 100K reference panel (in GRCh37 coordinates, subsequently lifted over to GRCh38) [105], which was later joined with the 1000 Genomes Project reference panel (subset to 2,504 unrelated individuals) with IMPUTE2 v2.3.2 [377]. The software imputed variants unique to each panel into the other, expanding both panels to include the union of all variants. The resulting set of haplotypes was then treated as a single, enriched reference panel for subsequent imputation. For computational efficiency, the genome was divided into 5 Mb chunks, and only bi-allelic SNPs were retained, as required by QUILT. I set -k_hap 198 56 to instruct the software the number of unrelated Vietnamese haplotypes in each reference panel and an effective population size (-Ne) 20,000 as recommended. The imputed chunks were assembled by chromosomes, and the combined reference panel was composed of 3,667 individuals. Second, since samples were sequenced in batches, I ran QUILT imputation on each batch separately and merged all samples together. This design was primarily chosen for computational efficiency and did not affect imputation results, as QUILT handled each genome independently. The imputed results were merged into a single file per chromosome containing all individuals and were then subset to variants with an INFO score greater than 0.9, an East Asian gnomAD allele frequency greater than 0.01, and the Omni5M markers (as in Chapter 3), retaining only reliably imputed SNPs while satisfying the density requirements for TOPMed imputation. The resulting files were subsequently uploaded to the TOPMed imputation server for second-stage imputation.

Sample and SNP QC. Starting with 874 individuals who underwent lcWGS, 11 samples with discordant genomic and recorded sex (Section 6.1.3) were excluded, along with 21 duplicate samples registered in both the SEARCH and VIETNARMS projects, retaining the sequencing data with higher depth in each duplicate pair. Additionally, 59 individuals with less than 60% of the genome covered, reflecting compromised imputation performance, were removed. The resulting dataset comprised 783 individuals for population structure analysis (detailed in the following section). For GWAS analyses, a further 210 individuals lacking viral load phenotype information were excluded. Genome-wide variants were filtered by removing those with Hardy-Weinberg equilibrium $p < 1\times10^{-6}$, minor allele frequency < 5%, and TOPMed imputation $R^2 > 0.5$. The final GWAS dataset included 573 individuals and 5,145,539 autosomal variants.

HLA allele inference. To infer HLA alleles for individuals, three complementary strategies were employed. As described in Chapter 4, HLA alleles were inferred using the enhanced QUILT-HLA workflow with the 1000 Genomes Project reference panel as well as a two-stage imputation approach by SNP imputation with QUILT using the GenomeAsia-enriched 1000 Genomes Project reference panel followed by leveraging the HLA multi-ethnic reference panel via the Michigan imputation server. Finally, because the HLA regions were enriched during sequencing, HLA-LA v1.0.4 was also used to call HLA alleles, a method that has demonstrated high accuracy for deep

sequencing data in the HLA region [169, 170]. For all calls, alleles were converted to G-group resolution to align the QUILT-HLA and HLA-LA outputs with the resolution of the multi-ethnic reference panel and to resolve ambiguity between 2-field allele calls. Consensus HLA genotypes for each individual were then determined by *post hoc* integration of these imputation results. Alleles were combined by accounting for the confidence associated with each method: the Q1 column from HLA-LA, posterior probabilities from QUILT-HLA, and a uniform posterior probability of 0.5 for imputation using the multi-ethnic reference panel. First, I removed all allele calls if confidence is below 0.1. Then, for each locus in each individual, confidence scores were averaged across all reported alleles to reflect cumulative support, and the top two alleles were retained as probable calls. Alleles with combined confidence below 0.7 were assigned as missing to minimise false positives arising from uncertainty. Using this approach, among 783 diploid individuals that passed QC across five HLA loci (HLA-A, B, C, DQB1, and DRB1), 7,202 allele copies (out of 7,830, 91.98%) were assigned as non-missing. In the association analysis, only 109 HLA alleles were tested as allele frequency for these variants are greater than 1%.

Population structure inference. To gauge the population structure in this cohort, I took the first-stage post-filtering imputation result (before uploaded to TOPMed), combined each chromosome, and subset variants with minor allele frequency greater than 5%. I encoded the results to BGEN format [401] using QCTOOL [317]. To gain an

approximately independent set of markers for PCA analysis, I ran inthinnerator v2.2.2 to thin genome-wide SNPs by randomly choosing one SNP and then excluding nearby SNPs within a 50 kb region (set by -min-distance). Two chromosome regions were excluded in this analysis to avoid distortion on eigen decomposition due to their abnormal LD patterns: the HLA region (chr6:25,000,000-40,000,000) and a common inversion region (chr8:7,500,000-12,000,000) [378]. These regions were excluded using QCTOOL [402] with the -excl-range option, resulting in a final set of 38,981 markers. I excluded 21 related individuals identified as at least second degree relatives by KING with --unrelated --degree 2 [379] in the PCA decomposition and projected them back afterwards. The same analysis was also applied jointly to the 783 individuals in the Vietnamese cohort and the 3,667 individuals from the GenomeAsia-enhanced 1000 Genomes Project reference panel.

GWAS of host genome-wide variants on viral load. I performed a linear mixed model GWAS of $\log_{10}$ viral load to account for both covariates and genetic relatedness within the study cohort [380], which comprised 573 Vietnamese individuals. The analysis included 5,145,539 autosomal variants from two-stage imputation with allele frequencies of at least 5% and 109 HLA alleles at G-group resolution with allele frequencies of at least 1%. GRM was estimated in GEMMA v0.98.5 [381] using the -gk option, based on the same markers employed in the PCA. Finally, I performed GWAS with the GRM and covariates (age, sex, HCV genotype, and the first five host

PCs), assuming additive effects of allele dosages:

$$log_{10}\ viral\ load \sim SNP + GRM + age + sex + HCV\ genotype + 5\ host\ PCs$$

Post-GWAS analyses. Gene annotations were retrieved from the Ensembl REST API using Python (as in Chapter 5), and leading SNPs were annotated with either the gene in which they are located or, if intergenic, the nearest protein coding gene within 1 Mb (or none if no such gene exists). Hitplots of top SNPs were plotted by a jupyter notebook implementation of LocusZoom [403] (named jupyter-locuszoom, which can be accessed from the GitHub page at https://github.com/krassowski/jupyter-locuszoom) [404].

6.4 Conclusions

This chapter presented a GWAS of host genomic variants, including genome-wide markers and HLA alleles, on HCV viral load. Section 6.1 described the study cohort, data processing, and imputation pipeline, emphasising the rigorous quality control required for lcWGS data. Sample performance varied considerably, and individuals with discordant sex, duplicates, insufficient genome coverage, or missing phenotypes are excluded. Genome-wide imputation performed as expected, whereas HLA regions remained challenging despite the blended sequencing strategy with additional capture on HLA loci. Three HLA imputation workflows were applied to infer HLA alleles, but

performance varied across loci and methods, necessitating a consensus approach to produce a joint call. In Section 6.2, the first GWAS using lcWGS with imputation for HCV viral load was conducted. No genome-wide significant associations were identified in this cohort, either under the conventional threshold $p < 5 \times 10^{-8}$ or Benjamini-Hochberg FDR 5%, likely reflecting imputation imperfection or limited power due to the smalll sample size. Although the IFNL4 locus and the HLA region are known to have strong effects on HCV infection outcomes in other viral genotypes, these signals were not detected here, possibly due to viral heterogeneity. Nevertheless, I identified rs6142998, located in an enhancer of *OSBPL2*, as a potential marker influencing HCV viral load. The effect is likely mediated via virus-lipid interactions, although further validation is required. Overall, this study provides insights into host genetic factors influencing HCV susceptibility and demonstrates the practical utility of lcWGS for association analyses in genetic epidemiology.

Chapter 7 Conclusions and Discussion

This thesis explores the potential of lcWGS for advancing genetic epidemiology, particularly in diverse global populations. Investigation of lcWGS ranges from evaluating library preparation methods (Chapter 2) to capturing a broad spectrum of genomic variation, including genome-wide variants (Chapter 3), HLA alleles (Chapter 4), and SVs (Chapter 5), and culminates in a proof-of-principle GWAS in a real cohort of individuals with HCV infection (Chapter 6). The final chapter summarises the key findings, discusses the limitations and challenges of the current approaches, and outlines directions for future research.

7.1 Low-coverage whole-genome sequencing in data generation

In Chapter 2, I explored practical considerations for generating lcWGS data, focusing on library preparation methods and sequencing platforms. Using 66 NA12878 samples, I identified a high-performing library preparation condition, 100 ng DNA input with the UIIFS kit, that yielded data with low per-base error rates, minimal sequencing skew (reflecting non-uniformity of genome coverage), and low duplication rates. Although this analysis was exploratory, as the experimental design lacked multiple replicates for each combination of conditions, two notable observations emerged. First, Illumina sequencing platforms exhibited higher read 2 error rates with increasing insert size,

consistent with previous reports attributing this to cluster formation artifacts and signal crosstalk between adjacent reads [270, 272]. Second, libraries prepared with less than 10 ng input mass showed highly inconsistent sequencing performance, corroborated in an independent dataset of 91 Vietnamese samples. I also conducted *in silico* simulations of lcWGS data across different sequencing platforms and observed that imputation accuracy declines with increasing read length in long-read sequencing, highlighting a limitation of the PacBio platform in this context. Together with its higher cost, these findings suggested that low-coverage long-read sequencing is unlikely to be an efficient strategy for population-scale studies.

Two directions of further work could strengthen the current analysis. First, a more rigorous experimental design incorporating multiple replicates across library preparation conditions would enable finer optimisation of the protocol, subsequently advantageous for future research involving lcWGS. Second, while assessment of sequencing platforms and insert sizes was based on simulations, generating real datasets would allow empirical validation of these findings. This was partly supported by the results in Chapter 6, where lcWGS data generated on the AVITI platform demonstrated improved genome coverage uniformity compared to Illumina, despite slightly lower average sequencing depth.

7.2 Low-coverage whole-genome sequencing in variant discovery

7.2.1 Low-coverage whole-genome sequencing for capturing genome-wide variants

Chapters 3, 4, and 5 presented a comprehensive evaluation of lcWGS data from the GAMCC cohort, assessing its ability to capture a broad spectrum of genomic variation, including genome-wide SNPs and indels, HLA alleles, and SVs. Chapter 3 assessed genome-wide imputation performance by benchmarking imputed genotypes from lcWGS against microarray data and deep whole-genome sequencing replicates. I demonstrated that enriching the reference panel with population-specific haplotypes from the MalariaGEN dataset improved QUILT imputation accuracy, and the result can be used on inference of population structure across the four Gambian ethnic groups. To further leverage the extensive TOPMed reference panel, I devised a two-stage imputation strategy by first performing local imputation with QUILT and then re-imputing a subset of reliably imputed SNPs via the TOPMed server. This approach achieved imputation accuracy comparable to that of TOPMed-imputed microarray data, with marginally better performance for common variants. From here, I further investigated specific regions and variants that were of particular interest. Imputation accuracy remained robust across GWAS catalogue variants and blood group genes, though performance was reduced in the FCGR locus due to segmental duplications and

alignment artefacts. Finally, as the GAMCC cohort was originally established for malaria research, I successfully replicated the well-known association of the rs334 SNP with severe malaria, albeit with slightly lower statistical significance attributable to imputation imperfection.

I highlight several directions for future work arising from the findings in Chapter 3. First, a major limitation of the current QUILT implementation is its restriction to imputing only bi-allelic SNPs. Extending its capability to handle multi-allelic SNPs and indels would improve variant discovery, even though the two-stage imputation strategy provided a partial workaround. Second, these results underscored the importance of curating population-specific reference panels to enhance haplotype representation. Publicly releasing such reference resources would greatly benefit future imputation studies, particularly in the context of global populations where current data is lacking. Lastly, while the proposed two-stage imputation pipeline achieved high accuracy, it remained computationally intensive and complicated. A promising future direction would be to assess QUILT imputation performance directly using the TOPMed reference panel, ideally with a heterogeneous population dataset or different studying cohorts from global populations. Although public release of TOPMed haplotypes is constrained by ethical and privacy considerations, integrating QUILT into the TOPMed imputation server could be a feasible solution, offering improved accessibility and efficiency for future analyses. To conclude, lcWGS with imputation offers the

advantage of genome-wide, unbiased variant capture and flexibility for downstream analyses. Genotyping, in contrast, provides highly accurate calls at predefined loci and benefits from mature, standardised workflows, and low computational expenses. In practice, researchers may choose between the two approaches based on practical considerations, such as cost, laboratory expertise, and available infrastructure, depending on what best fits their study design.

Yet another way to obtain genomic data, in addition to lcWGS and microarrays, is whole-exome sequencing. It targets protein-coding regions, providing high-depth coverage that enables accurate detection of functional and particularly rare variants. This makes this approach especially valuable for studies focused on Mendelian traits or variants with strong effects on protein function. On the other hand, lcWGS complements exonic information by covering non-coding and regulatory regions, providing genome-wide coverage and capturing large SVs that may also be disease-relevant. Recently, recognising the complementary strengths of these approaches, a blended genome and exome sequencing method was proposed to maximise variant discovery while controlling costs [405]. Looking forward, flexible sequencing strategies that combine whole-genome sequencing, whole-exome sequencing, and targeted regions offer many possibilities, with researchers able to tailor choices to study design, population size, and research objectives to optimise data acquisition.

Lastly, lcWGS with imputation has already proven useful in clinical applications such as non-invasive prenatal testing (NIPT). NIPT leverages cell-free fetal DNA in maternal plasma to assess fetal genetic status (for example, aneuploidy) without the risks of invasive sampling, and lcWGS is particularly well suited for this setting because fetal DNA constitutes only a small fraction of total cell-free DNA, making high-depth sequencing inefficient while still allowing imputation to recover informative genotypes. Several studies have demonstrated the feasibility of uncovering genotype-phenotype associations using NIPT data [81, 260, 406, 407], and recent methodological advances, such as QUILT2 [67], further improve imputation in this context. As exemplified by NIPT, the utility of lcWGS is poised to expand, enabling new applications in both genetic studies and clinical practice.

7.2.2 Low-coverage whole-genome sequencing in inferring classical HLA alleles

In Chapter 4, I explored the potential of lcWGS for inferring HLA alleles, which are characterised by high genetic complexity, considerable polymorphism, and biological significance in immune-related traits. I first assessed HLA imputation performance using QUILT-HLA with the 1000 Genomes Project reference panel, benchmarking against truth HLA sequence-based typing data. Imputation accuracy was high for HLA-B, -C, and -DQB1, whereas several alleles at HLA-A and HLA-DRB1 proved challenging. To improve inference, I pursued two complementary strategies: enhancing LD-based imputation by creating a Gambian-enriched reference panel, and refining

direct read-based inference with a rigorous alignment process. For the former, I incorporated true allele calls from this cohort into the 1000 Genomes Project reference panel with a custom method to phase HLA alleles onto SNP haplotypes and applied a leave-one-out imputation test. For the latter, I replaced the original k-mer alignment approach with Wavefront aligner. I showed that these efforts jointly translate to overall better inference across all HLA loci. Finally, I compared these results with HLA imputation from microarray data and a similar two-stage imputation approach starting from lcWGS with the larger and more diverse HLA multi-ethnic reference panel via the Michigan imputation server. Performance varied by locus: QUILT-HLA outperformed for HLA-C and HLA-DQB1, was comparable for HLA-B, and underperformed for HLA-A and HLA-DRB1. By integrating the two workflows, I achieved at least 90% concordance at all loci, demonstrating the practical utility of lcWGS for HLA allele inference.

One potential limitation in this chapter concerned the Wavefront aligner process. While it achieved high alignment accuracy, it relied on a naïve pairwise alignment of sequencing reads to HLA alleles, resulting in computational overhead due to the high sequence similarity among alleles. A more efficient alternative could involve a graph-based approach, such as implemented in other HLA inference software [169, 170], which encodes variant and allele frequency information in a compact graph structure. Additionally, the final comparison of QUILT-HLA with the HLA multi-ethnic panel

was subject to several limitations. First, the two workflows relied on different reference panels, and I was unable to access the larger reference panel locally. Second, I encountered technical inconsistencies in allele coding, as the server reported HLA alleles at G-group resolution but encoded them at the 2-field level. Further evaluation of these methods across diverse populations may provide additional insights into their performance and limitations.

7.2.3 Low-coverage whole-genome sequencing in detecting large structural variants

In the final chapter on variant discovery from lcWGS, I focused on another challenging class of variants, that is, SVs. These variants often arise in complex genomic regions and serve as functional determinants for many disease outcomes. I approached the study of SVs by examining genome coverage patterns in the glycophorin region, introduced lcSV, a population-based method specifically designed applicable to lcWGS data, and demonstrated its effectiveness through both simulation studies and direct application to the glycophorin locus. To evaluate the method systematically, I conducted additional *in silico* experiments to assess its resolution with respect to variant size and allele frequency, and subsequently applied lcSV to a curated list of known SVs [343]. The results showed that the method successfully captured the majority of target SVs and provided insights into the conditions under which it succeeded or failed by examining individual examples. Finally, I compiled a list of SVs that were reliably detected with lcSV for cohorts of comparable size and sequencing depth.

Importantly, the current evaluation of genome-wide SVs focused exclusively on deletions. The method relies on accurate read alignments in the flanking regions to detect abnormal genome coverage indicative of an SV event. While a loss of coverage signals a deletion, insertions are less reliably detected, as reads may be mapped elsewhere in the genome due to different SV formation mechanisms [343]. Currently, lcSV can only detect deletions or insertions that manifest as local duplications relative to the flanking regions. Moreover, no truth dataset was generated for this cohort and thus no direct benchmarking was possible. Future work could address these limitations by performing *in silico* down-sampling from available high-coverage datasets or by cross-referencing with external SV resources, such as the Human Pangenome Reference Consortium [277], to validate and expand detection capabilities.

7.3 Low-coverage whole-genome sequencing in a genome-wide association study of Hepatitis C Virus

Finally, I conducted a GWAS in an HCV-infected cohort, illustrating the practical utility of lcWGS for biological inference in diverse populations. To our knowledge, this study is the first GWAS conducted in a predominantly HCV genotype 6 cohort. Specifically, lcWGS was performed on HCV-infected Vietnamese samples, with additional capture on the HLA region to enhance allele representation. Genome-wide variants were imputed using a two-stage strategy (QUILT imputation with the GenomeAsia-enhanced 1000 Genomes Project reference panel followed by TOPMed imputation), while HLA

alleles at G-group resolution were obtained via a consensus of three workflows: improved QUILT-HLA imputation with the 1000 Genomes Project reference panel, QUILT-HLA calls, and two-stage imputation with the HLA multi-ethnic reference panel. GWAS on viral load identified a notable SNP, rs6142998, located in an enhancer of *OSBPL2* and predicted as a regulatory variant by VEP. Colocalisation with an sQTL from the GTEx project indicated that the variant is associated with reduced *OSBPL2* splicing efficiency, potentially leading to decreased overall expression in whole blood. A possible mechanism may involve impairment of HCV replication by altered intracellular lipid homeostasis. No HLA alleles, nor the well-established *IFNL4* SNP rs12979860, showed nominal significance, potentially reflecting viral heterogeneity. Functional validation in relevant models is required to confirm these findings.

7.4 Conclusions

In summary, this thesis highlights the potential of lcWGS to advance genetic epidemiology in the context of global populations. By enabling comprehensive variant discovery, lcWGS provides a scalable framework for studying complex traits and infectious disease susceptibility across global populations. These findings emphasise the importance of expanding genomic studies beyond well-characterised cohorts, and point to future opportunities for methodological refinement, population-specific reference panels, and functional validation to fully leverage lcWGS in understanding human genetic diversity and improving global health equity.

References

1. Sanger, F., S. Nicklen, and A.R. Coulson, *DNA sequencing with chain-terminating inhibitors.* Proceedings of the National Academy of Sciences, 1977. **74**: p. 5463-5467.
2. Valencia, C.A., et al., *Sanger sequencing principles, history, and landmarks.* Next Generation Sequencing Technologies in Medical Genetics, 2013: p. 3-11.
3. Sanger, F., et al., *Nucleotide sequence of bacteriophage φX174 DNA.* Nature, 1977. **265**: p. 687-695.
4. Sanger, F. and A.R. Coulson, *A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase.* Journal of Molecular Biology, 1975. **94**: p. 441-448.
5. Maxam, A.M. and W. Gilbert, *A new method for sequencing DNA.* Proceedings of the National Academy of Sciences, 1977. **74**: p. 560-564.
6. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**: p. 860-921.
7. Dovichi, N.J., *DNA sequencing by capillary electrophoresis.* Electrophoresis, 1997. **18**: p. 2393-2399.
8. Ona, S., *Sanger Sequencing.* 2025.
9. Bentley, D.R., et al., *Accurate whole human genome sequencing using reversible terminator chemistry.* Nature, 2008. **456**: p. 53-59.
10. Clarke, J., et al., *Continuous base identification for single-molecule nanopore DNA sequencing.* Nature Nanotechnology, 2009. **4**: p. 265-270.
11. Eid, J., et al., *Real-time DNA sequencing from single polymerase molecules.* Science, 2009. **323**: p. 133-138.
12. Harris, T.D., et al., *Single-molecule DNA sequencing of a viral genome.* Science, 2008. **320**: p. 106-109.
13. Kircher, M. and J. Kelso, *High-throughput DNA sequencing–concepts and limitations.* Bioessays, 2010. **32**: p. 524-536.
14. Korlach, J., et al., *Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures.* Proceedings of the National Academy of Sciences, 2008. **105**: p. 1176-1181.
15. Margulies, M., et al., *Genome sequencing in microfabricated high-density picolitre reactors.* Nature, 2005. **437**: p. 376-380.
16. Satam, H., et al., *Next-generation sequencing technology: current trends and advancements.* Biology, 2023. **12**: p. 997.
17. Shendure, J., et al., *Accurate multiplex polony sequencing of an evolved bacterial genome.* Science, 2005. **309**: p. 1728-1732.
18. Schuster, S.C., *Next-generation sequencing transforms today's biology.* Nature Methods, 2008. **5**: p. 16-18.
19. Ost, T.W.B., et al., *Improved polymerases*. 2006.

20. Smith, G.P., et al., *Modified polymerases for improved incorporation of nucleotide analogues*. 2003.
21. Kawashima, E., L. Farinelli, and P. Mayer, *Method of nucleic acid amplification by extension of immobilized primers*. 1998.
22. Mayer, P., *Isothermal amplification of nucleic acids on a solid support*. 2001.
23. Rodriguez, R. and Y. Krishnan, *The chemistry of next-generation sequencing.* Nature Biotechnology, 2023. **41**: p. 1709-1715.
24. Modi, A., et al., *The Illumina sequencing protocol and the NovaSeq 6000 system*, in *Bacterial Pangenomics: Methods and Protocols*. 2021, Springer. p. 15-42.
25. Amarasinghe, S.L., et al., *Opportunities and challenges in long-read sequencing data analysis.* Genome Biology, 2020. **21**: p. 30.
26. Ona, S., *Next Generation Sequencing (Illumina).* 2025.
27. Deamer, D., M. Akeson, and D. Branton, *Three decades of nanopore sequencing.* Nature Biotechnology, 2016. **34**: p. 518-524.
28. Jain, M., et al., *The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community.* Genome Biology, 2016. **17**: p. 1-11.
29. Oehler, J.B., et al., *The application of long-read sequencing in clinical settings.* Human Genomics, 2023. **17**: p. 73.
30. Rhoads, A. and K.F. Au, *PacBio sequencing and its applications.* Genomics, Proteomics & Bioinformatics, 2015. **13**: p. 278-289.
31. Wang, Y., et al., *Nanopore sequencing technology, bioinformatics and applications.* Nature Biotechnology, 2021. **39**: p. 1348-1365.
32. Wenger, A.M., et al., *Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome.* Nature Biotechnology, 2019. **37**: p. 1155-1162.
33. Chen, J., et al., *Whole-genome long-read TAPS deciphers DNA methylation patterns at base resolution using PacBio SMRT sequencing technology.* Nucleic Acids Research, 2022. **50**: p. e104-e104.
34. Livak, K.J., et al., *Oligonucleotides with fluorescent dyes at opposite ends provide a quenched probe system useful for detecting PCR product and nucleic acid hybridization.* Genome Research, 1995. **4**: p. 357-362.
35. Verlouw, J.A.M., et al., *A comparison of genotyping arrays.* European Journal of Human Genetics, 2021. **29**: p. 1611-1624.
36. Hirschhorn, J.N. and M.J. Daly, *Genome-wide association studies for common diseases and complex traits.* Nature Reviews Genetics, 2005. **6**: p. 95-108.
37. Lamy, P., J. Grove, and C. Wiuf, *A review of software for microarray genotyping.* Human Genomics, 2011. **5**: p. 304.
38. Burton, P.R., et al., *Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.* Nature, 2007. **447**: p. 661-678.
39. Pe'er, I., et al., *Evaluating and improving power in whole-genome association studies using fixed marker sets.* Nature Genetics, 2006. **38**: p. 663-667.

40. Neafsey, D.E., et al., *Genome-wide SNP genotyping highlights the role of natural selection in Plasmodium falciparum population divergence.* Genome Biology, 2008. **9**: p. 1-16.
41. Price, A.L., et al., *Discerning the ancestry of European Americans in genetic association studies.* PLoS Genetics, 2008. **4**: p. e236.
42. Niu, T., *Algorithms for inferring haplotypes.* Genetic Epidemiology, 2004. **27**: p. 334-347.
43. Mlakar, V. and D. Glavac, *DNA microarrays and their use in dermatology.* ACTA DERMATOVENEROLOGICA ALPINA PANONICA ET ADRIATICA, 2007. **16**: p. 7.
44. Jiang, S., et al., *Generic Diagramming Platform (GDP): a comprehensive database of high-quality biomedical graphics.* Nucleic Acids Research, 2025. **53**: p. D1670-D1676.
45. BioRender, *DNA Microarray.* 2025.
46. Lamy, P., et al., *Genotyping and annotation of Affymetrix SNP arrays.* Nucleic Acids Research, 2006. **34**: p. e100-e100.
47. Aparna, G. and K.K. Tetala, *Recent progress in development and application of DNA, protein, peptide, glycan, antibody, and aptamer microarrays.* Biomolecules, 2023. **13**: p. 602.
48. Gibbs, R.A., et al., *The international HapMap project.* 2003.
49. Altshuler, D.M., et al., *A global reference for human genetic variation.* Nature, 2015. **526**: p. 68-74.
50. Byrska-Bishop, M., et al., *High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios.* Cell, 2022. **185**: p. 3426-3440.
51. Karczewski, K.J., et al., *The mutational constraint spectrum quantified from variation in 141,456 humans.* Nature, 2020. **581**: p. 434-443.
52. Kockum, I., J. Huang, and P. Stridh, *Overview of genotyping technologies and methods.* Current Protocols, 2023. **3**: p. e727.
53. Albrechtsen, A., F.C. Nielsen, and R. Nielsen, *Ascertainment biases in SNP chips affect measures of population divergence.* Molecular Biology and Evolution, 2010. **27**: p. 2534-2547.
54. Das, S., G.R. Abecasis, and B.L. Browning, *Genotype imputation from large reference panels.* Annual Review of Genomics and Human Genetics, 2018. **19**: p. 73-96.
55. Li, Y., et al., *Genotype imputation.* Annual Review of Genomics and Human Genetics, 2009. **10**: p. 387-406.
56. Treccani, M., et al., *A broad overview of genotype imputation: Standard guidelines, approaches, and future investigations in genomic association studies.* Biocell, 2023. **47**: p. 1225-1241.
57. Quick, C., et al., *Sequencing and imputation in GWAS: Cost-effective strategies to increase power and genomic coverage across diverse*

populations. Genetic Epidemiology, 2020. **44**: p. 537-549.
58. Li, N. and M. Stephens, *Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data.* Genetics, 2003. **165**: p. 2213-2233.
59. Alkan, C., B.P. Coe, and E.E. Eichler, *Genome structural variation discovery and genotyping.* Nature Reviews Genetics, 2011. **12**: p. 363-376.
60. Ho, S.S., A.E. Urban, and R.E. Mills, *Structural variation in the sequencing era.* Nature Reviews Genetics, 2020. **21**: p. 171-189.
61. Collins, R.L., et al., *A structural variation reference for medical and population genetics.* Nature, 2020. **581**: p. 444-451.
62. Rohland, N. and D. Reich, *Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture.* Genome Research, 2012. **22**: p. 939-946.
63. Pasaniuc, B., et al., *Extremely low-coverage sequencing and imputation increases power for genome-wide association studies.* Nature Genetics, 2012. **44**: p. 631-635.
64. Rubinacci, S., et al., *Imputation of low-coverage sequencing data from 150,119 UK Biobank genomes.* Nature Genetics, 2023. **55**: p. 1088-1090.
65. Rubinacci, S., et al., *Efficient phasing and imputation of low-coverage sequencing data using large reference panels.* Nature Genetics, 2021. **53**: p. 22.
66. Davies, R.W., et al., *Rapid genotype imputation from sequence with reference panels.* Nature Genetics, 2021. **53**: p. 1104-1111.
67. Li, Z., A. Albrechtsen, and R.W. Davies, *Rapid and accurate genotype imputation from low coverage short read, long read, and cell free DNA sequence.* BioRxiv, 2024: p. 2024.07.18.604149.
68. Santos, R., et al., *Low-coverage whole genome sequencing for a highly selective cohort of severe COVID-19 patients.* Gigabyte, 2024. **2024**: p. gigabyte127.
69. Wasik, K., et al., *Comparing low-pass sequencing and genotyping for trait mapping in pharmacogenetics.* BMC Genomics, 2021. **22**: p. 1-7.
70. Li, J.H., et al., *Low-pass sequencing plus imputation using avidity sequencing displays comparable imputation accuracy to sequencing by synthesis while reducing duplicates.* G3: Genes, Genomes, Genetics, 2024. **14**: p. 5.
71. Chat, V., et al., *Ultra Low-Coverage Whole-Genome Sequencing as an Alternative to Genotyping Arrays in Genome-Wide Association Studies.* Frontiers in Genetics, 2022. **12**: p. 9.
72. Gilly, S.A., et al., *Very low-depth whole-genome sequencing in complex trait association studies.* Bioinformatics, 2019. **35**: p. 2555-2561.
73. Martin, A.R., et al., *Low-coverage sequencing cost-effectively detects known and novel variation in underrepresented populations.* The American Journal of Human Genetics, 2021. **108**: p. 656-668.

74. Bizon, C., et al., *Variant calling in low-coverage whole genome sequencing of a Native American population sample.* BMC Genomics, 2014. **15**: p. 85.
75. Purnomo, G.A., et al., *Benchmarking Imputed Low Coverage Genomes in a Human Population Genetics Context.* Molecular Ecology Resources, 2024: p. e70007.
76. Emde, A.-K., et al., *Mid-pass whole genome sequencing enables biomedical genetic studies of diverse populations.* BMC Genomics, 2021. **22**: p. 666.
77. Reingruber, J., et al., *Assessing the suitability of formalin-fixed paraffin-embedded (FFPE) tissue for genome-wide association studies (GWAS).* BMC Research Notes, 2025. **18**: p. 254.
78. Deng, T.Y., et al., *Comparison of Genotype Imputation for SNP Array and Low-Coverage Whole-Genome Sequencing Data.* Frontiers in Genetics, 2022. **12**: p. 11.
79. Snelling, W.M., et al., *Assessment of imputation from low-pass sequencing to predict merit of beef steers.* Genes, 2020. **11**: p. 1312.
80. Liu, S., et al., *Accurate genotype imputation from low-coverage whole-genome sequencing data of rainbow trout.* G3: Genes, Genomes, Genetics, 2024. **14**: p. jkae168.
81. Liu, S., et al., *Utilizing non-invasive prenatal test sequencing data for human genetic investigation.* Cell Genomics, 2024. **4**: p. 100669.
82. Chen, D., et al., *A cost-effective, high-throughput, highly accurate genotyping method for outbred populations.* G3: Genes, Genomes, Genetics, 2025. **15**: p. jkae291.
83. Zhao, C., et al., *Towards a cost-effective implementation of genomic prediction based on low coverage whole genome sequencing in Dezhou donkey.* Frontiers in Genetics, 2021. **12**: p. 728764.
84. Gundappa, M.K., et al., *High performance imputation of structural and single nucleotide variants using low-coverage whole genome sequencing.* Genetics Selection Evolution, 2025. **57**: p. 16.
85. Cao, J., et al., *Development and evaluation of a haplotype reference panel for low-coverage whole genome sequencing genotype imputation in turbot (Scophthalmus maximus).* Aquaculture Reports, 2025. **41**: p. 102695.
86. Zhang, Z., et al., *The efficient phasing and imputation pipeline of low-coverage whole genome sequencing data using a high-quality and publicly available reference panel in cattle.* Animal Research and One Health, 2023. **1**: p. 4-16.
87. Nielsen, R., et al., *Genotype and SNP calling from next-generation sequencing data.* Nature Reviews Genetics, 2011. **12**: p. 443-451.
88. Marchini, J. and B. Howie, *Genotype imputation for genome-wide association studies.* Nature Reviews Genetics, 2010. **11**: p. 499-511.
89. Fuchsberger, C., G.R. Abecasis, and D.A. Hinds, *minimac2: faster genotype imputation.* Bioinformatics, 2014. **31**: p. 782-784.

90. Howie, B.N., P. Donnelly, and J. Marchini, *A flexible and accurate genotype imputation method for the next generation of genome-wide association studies.* PLoS Genetics, 2009. **5**: p. e1000529.
91. Rubinacci, S., O. Delaneau, and J. Marchini, *Genotype imputation using the positional burrows wheeler transform.* PLoS Genetics, 2020. **16**: p. e1009049.
92. Browning, B.L. and S.R. Browning, *A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals.* The American Journal of Human Genetics, 2009. **84**: p. 210-223.
93. Browning, B.L. and S.R. Browning, *Genotype imputation with millions of reference samples.* The American Journal of Human Genetics, 2016. **98**: p. 116-126.
94. VanRaden, P.M., C. Sun, and J.R. O'Connell, *Fast imputation using medium or low-coverage sequence data.* BMC Genetics, 2015. **16**: p. 82.
95. Ros-Freixedes, R., et al., *Accuracy of whole-genome sequence imputation using hybrid peeling in large pedigreed livestock populations.* Genetics Selection Evolution, 2020. **52**: p. 17.
96. Zheng, C., M.P. Boer, and F.A. van Eeuwijk, *Accurate genotype imputation in multiparental populations from low-coverage sequence.* Genetics, 2018. **210**: p. 71-82.
97. Spiliopoulou, A., et al., *GeneImp: fast imputation to large reference panels using genotype likelihoods from ultralow coverage sequencing.* Genetics, 2017. **206**: p. 91-104.
98. Davies, R.W., et al., *Rapid genotype imputation from sequence without reference panels.* Nature Genetics, 2016. **48**: p. 965-969.
99. Frazer, K.A., et al., *A second generation human haplotype map of over 3.1 million SNPs.* Nature, 2007. **449**: p. 851.
100. Altshuler, D.M., et al., *Integrating common and rare genetic variation in diverse human populations.* Nature, 2010. **467**: p. 52.
101. McVean, G.A., et al., *An integrated map of genetic variation from 1,092 human genomes.* Nature, 2012. **491**: p. 56.
102. Durbin, R.M., et al., *A map of human genome variation from population scale sequencing.* Nature, 2010. **467**: p. 1061.
103. McCarthy, S., et al., *A reference panel of 64,976 haplotypes for genotype imputation.* Nature Genetics, 2016. **48**: p. 1279-1283.
104. Taliun, D., et al., *Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program.* Nature, 2021. **590**: p. 290-299.
105. Wall, J.D., et al., *The GenomeAsia 100K Project enables genetic discoveries across Asia.* Nature, 2019. **576**: p. 106-111.
106. Mathias, R.A., et al., *A continuum of admixture in the Western Hemisphere revealed by the African Diaspora genome.* Nature Communications, 2016. **7**: p. 12522.

107. Gurdasani, D., et al., *The African genome variation project shapes medical genetics in Africa.* Nature, 2015. **517**: p. 327-332.
108. Hirata, M., et al., *Cross-sectional analysis of BioBank Japan clinical data: a large cohort of 200,000 patients with 47 common diseases.* Journal of Epidemiology, 2017. **27**: p. S9-S21.
109. Nagai, A., et al., *Overview of the BioBank Japan Project: Study design and profile.* Journal of Epidemiology, 2017. **27**: p. S2-S8.
110. Francioli, L.C., et al., *Whole-genome sequence variation, population structure and demographic history of the Dutch population.* Nature Genetics, 2014. **46**: p. 818-825.
111. Huang, J., et al., *Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel.* Nature Communications, 2015. **6**: p. 8111.
112. Sidore, C., et al., *Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers.* Nature Genetics, 2015. **47**: p. 1272-1281.
113. Lee, J., et al., *A database of 5305 healthy Korean individuals reveals genetic and clinical implications for an East Asian population.* Experimental & Molecular Medicine, 2022. **54**: p. 1862-1871.
114. Lee, S., et al., *Korean Variant Archive (KOVA): a reference database of genetic variations in the Korean population.* Scientific Reports, 2017. **7**: p. 4287.
115. Band, G., et al., *Insights into malaria susceptibility using genome-wide data on 17,000 individuals from Africa, Asia and Oceania.* Nature Communications, 2019. **10**: p. 19.
116. Dunkelberger, J.R. and W.-C. Song, *Complement and its role in innate and adaptive immune responses.* Cell Research, 2010. **20**: p. 34-50.
117. Dendrou, C.A., et al., *HLA variation and disease.* Nature Reviews Immunology, 2018. **18**: p. 325-339.
118. Oksenberg, J.R., et al., *Mapping multiple sclerosis susceptibility to the HLA-DR locus in African Americans.* The American Journal of Human Genetics, 2004. **74**: p. 160-167.
119. Aboulaghras, S., et al., *Meta-analysis and systematic review of HLA DQ2/DQ8 in adults with celiac disease.* International Journal of Molecular Sciences, 2023. **24**: p. 1188.
120. Noble, J.A., et al., *HLA class I and genetic susceptibility to type 1 diabetes: results from the Type 1 Diabetes Genetics Consortium.* Diabetes, 2010. **59**: p. 2972-2979.
121. Varney, M.D., et al., *HLA DPA1, DPB1 alleles and haplotypes contribute to the risk associated with type 1 diabetes: analysis of the type 1 diabetes genetics consortium families.* Diabetes, 2010. **59**: p. 2055-2062.
122. Noble, J.A. and A.M. Valdes, *Genetics of the HLA region in the prediction of*

type 1 diabetes. Current Diabetes Reports, 2011. **11**: p. 533-542.
123. Thursz, M.R., et al., *Association between an MHC class II allele and clearance of hepatitis B virus in the Gambia.* New England Journal of Medicine, 1995. **332**: p. 1065-1069.
124. Medhasi, S. and N. Chantratita, *Human leukocyte antigen (HLA) system: genetics and association with bacterial and viral infections.* Journal of Immunology Research, 2022. **2022**: p. 9710376.
125. Molberg, Ø., et al., *Tissue transglutaminase selectively modifies gliadin peptides that are recognized by gut-derived T cells in celiac disease.* Nature Medicine, 1998. **4**: p. 713-717.
126. Bodd, M., et al., *T-cell response to gluten in patients with HLA-DQ2.2 reveals requirement of peptide-MHC stability in celiac disease.* Gastroenterology, 2012. **142**: p. 552-561.
127. Fallang, L.-E., et al., *Differences in the risk of celiac disease associated with HLA-DQ2.5 or HLA-DQ2.2 are related to sustained gluten antigen presentation.* Nature Immunology, 2009. **10**: p. 1096-1101.
128. Hovhannisyan, Z., et al., *The role of HLA-DQ8 β57 polymorphism in the anti-gluten T-cell response in coeliac disease.* Nature, 2008. **456**: p. 534-538.
129. Robinson, J., D.J. Barker, and S.G. Marsh, *25 years of the IPD-IMGT/HLA D atabase.* HLA, 2024. **103**: p. e15549.
130. Amos, D., *Human histocompatibility locus HL-A.* Science, 1968. **159**: p. 659-660.
131. WHO-Nomenclature-Committee, *Nomenclature for Factors of the HL-A System.* Bull. World Health Organ., 1968. **39**: p. 483.
132. Marsh, S.G., et al., *Nomenclature for factors of the HLA system, 2010.* Tissue antigens, 2010. **75**: p. 291.
133. Erlich, H., *HLA DNA typing: past, present, and future.* Tissue Antigens, 2012. **80**: p. 1-11.
134. Erlich, H., et al., *HLA-DR, DO and DP typing using PCR amplification and immobilized probes.* International Journal of Immunogenetics, 1991. **18**: p. 33-55.
135. Santamaria, P., et al., *HLA class II "typing": direct sequencing of DRB, DQB, and DQA genes.* Human Immunology, 1992. **33**: p. 69-81.
136. Santamaria, P., et al., *HLA class I sequence-based typing.* Human Immunology, 1993. **37**: p. 39-50.
137. Van Der Vlies, S., C. Voorter, and E. Van Den Berg-Loonen, *There is more to HLA-C than exons 2 and 3: sequencing of exons 1, 4 and 5.* Tissue Antigens, 1999. **54**: p. 169-177.
138. Smith, L.K., *HLA typing by direct DNA sequencing.* Immunogenetics: Methods and Applications in Clinical Practice, 2012: p. 67-86.
139. De Santis, D., et al., *16th IHIW: review of HLA typing by NGS.* International Journal of Immunogenetics, 2013. **40**: p. 72-76.

140. Voorter, C.E., F. Palusci, and M.G. Tilanus, *Sequence-based typing of HLA: an improved group-specific full-length gene sequencing approach.* Bone Marrow and Stem Cell Transplantation, 2014: p. 101-114.
141. Liu, C., *A long road/read to rapid high-resolution HLA typing: The nanopore perspective.* Human Immunology, 2021. **82**: p. 488-495.
142. Kulski, J.K., S. Suzuki, and T. Shiina, *Human leukocyte antigen super-locus: nexus of genomic supergenes, SNPs, indels, transcripts, and haplotypes.* Human Genome Variation, 2022. **9**: p. 49.
143. Yang, J., et al., *Recent advances of human leukocyte antigen (HLA) typing technology based on high-throughput sequencing.* Journal of Biomedical Nanotechnology, 2022. **18**: p. 617-639.
144. Geo, J.A., et al., *Advancements in HLA typing techniques and their impact on transplantation medicine.* Medical Principles and Practice, 2024. **33**: p. 215-231.
145. Gabriel, C., et al., *HLA typing by next-generation sequencing–getting closer to reality.* Tissue Antigens, 2014. **83**: p. 65-75.
146. Lind, C., et al., *Next-generation sequencing: the solution for high-resolution, unambiguous human leukocyte antigen typing.* Human Immunology, 2010. **71**: p. 1033-1042.
147. Nelson, W.C., et al., *An integrated genotyping approach for HLA and other complex genetic systems.* Human Immunology, 2015. **76**: p. 928-938.
148. Hosomichi, K., et al., *The impact of next-generation sequencing technologies on HLA research.* Journal of Human Genetics, 2015. **60**: p. 665-673.
149. Bruijnesteijn, J., *HLA/MHC and KIR characterization in humans and non-human primates using Oxford Nanopore Technologies and Pacific Biosciences sequencing platforms.* HLA, 2023. **101**: p. 205-221.
150. Moutsianas, L. and J. Gutierrez-Achury, *Genetic Association in the HLA Region*, in *Genetic Epidemiology: Methods and Protocols*, E. Evangelou, Editor. 2018, Springer New York: New York, NY. p. 111-134.
151. Zheng, X., *Imputation-based HLA typing with SNPs in GWAS studies.* HLA typing: Methods and Protocols, 2018: p. 163-176.
152. Sakaue, S., et al., *Tutorial: a statistical genetics guide to identifying HLA alleles driving complex disease.* Nature Protocols, 2023. **18**: p. 2625-2641.
153. Sivaprakasam, B. and P. Sadagopan, *HLA allele type prediction: A review on concepts, methods and algorithms.* Asian Journal of Biological and Life Sciences, 2023. **12**: p. 207.
154. Abi-Rached, L., et al., *Immune diversity sheds light on missing variation in worldwide genetic diversity panels.* PLoS One, 2018. **13**: p. e0206512.
155. Major, E., et al., *HLA typing from 1000 genomes whole genome and whole exome illumina data.* PLoS One, 2013. **8**: p. e78410.
156. Gourraud, P.-A., et al., *HLA diversity in the 1000 genomes dataset.* PLoS One, 2014. **9**: p. e97282.

157. Luo, Y., et al., *A high-resolution HLA reference panel capturing global population diversity enables multi-ancestry fine-mapping in HIV host response.* Nature Genetics, 2021. **53**: p. 1504-1516.
158. Das, S., et al., *Next-generation genotype imputation service and methods.* Nature Genetics, 2016. **48**: p. 1284-1287.
159. De Bakker, P.I., et al., *A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC.* Nature Genetics, 2006. **38**: p. 1166-1172.
160. Leslie, S., P. Donnelly, and G. McVean, *A statistical method for predicting classical HLA alleles from SNP data.* The American Journal of Human Genetics, 2008. **82**: p. 48-56.
161. Dilthey, A., et al., *Multi-population classical HLA type imputation.* PLoS Computational Biology, 2013. **9**: p. e1002877.
162. Dilthey, A.T., et al., *HLA*IMP—an integrated framework for imputing classical HLA alleles from SNP genotypes.* Bioinformatics, 2011. **27**: p. 968-972.
163. Motyer, A., et al., *Practical use of methods for imputation of HLA alleles from SNP genotype data.* BioRxiv, 2016: p. 091009.
164. Zheng, X., et al., *HIBAG—HLA genotype imputation with attribute bagging.* The Pharmacogenomics Journal, 2014. **14**: p. 192-200.
165. Shen, J.J., et al., *HLA-IMPUTER: an easy to use web application for HLA imputation and association analysis using population-specific reference panels.* Bioinformatics, 2019. **35**: p. 1244-1246.
166. Jia, X., et al., *Imputing amino acid polymorphisms in human leukocyte antigens.* PLoS One, 2013. **8**: p. e64683.
167. Cook, S., et al., *Accurate imputation of human leukocyte antigens with CookHLA.* Nature Communications, 2021. **12**: p. 1264.
168. Dilthey, A., et al., *Improved genome inference in the MHC using a population reference graph.* Nature Genetics, 2015. **47**: p. 682-688.
169. Dilthey, A.T., et al., *High-accuracy HLA type inference from whole-genome sequencing data using population reference graphs.* PLoS Computational Biology, 2016. **12**: p. e1005151.
170. Dilthey, A.T., et al., *HLA*LA—HLA typing from linearly projected graph alignments.* Bioinformatics, 2019. **35**: p. 4394-4396.
171. Lee, H. and C. Kingsford, *Kourami: graph-guided assembly for novel human leukocyte antigen allele discovery.* Genome Biology, 2018. **19**: p. 1-16.
172. Maiers, M., et al., *GRIMM: GRaph IMputation and matching for HLA genotypes.* Bioinformatics, 2019. **35**: p. 3520-3523.
173. Nariai, N., et al. *HLA-VBSeq: accurate HLA typing at full resolution from whole-genome sequencing data*. in *BMC Genomics*. 2015. Springer.
174. Wang, Y.-Y., et al., *HLA-VBSeq v2: improved HLA calling accuracy with full-length Japanese class-I panel.* Human Genome Variation, 2019. **6**: p. 29.

175. Naito, T., et al., *A deep learning method for HLA imputation and trans-ethnic MHC fine-mapping of type 1 diabetes.* Nature Communications, 2021. **12**: p. 1639.
176. Tanaka, K., et al., *Efficient HLA imputation from sequential SNPs data by transformer.* Journal of Human Genetics, 2024. **69**: p. 533-540.
177. Boegel, S., et al., *HLA typing from RNA-Seq sequence reads.* Genome Medicine, 2013. **4**: p. 1-12.
178. Kim, H.J. and N. Pourmand, *HLA haplotyping from RNA-seq data using hierarchical read weighting.* PLoS One, 2013. **8**: p. e67885.
179. Bai, Y., D. Wang, and W. Fury, *PHLAT: inference of high-resolution HLA types from RNA and whole exome sequencing.* HLA Typing: Methods and Protocols, 2018: p. 193-201.
180. Szolek, A., et al., *OptiType: precision HLA typing from next-generation sequencing data.* Bioinformatics, 2014. **30**: p. 3310-3316.
181. Xie, C., et al., *Fast and accurate HLA typing from short-read next-generation sequence data with xHLA.* Proceedings of the National Academy of Sciences, 2017. **114**: p. 8059-8064.
182. Warren, R.L. and I. Birol, *Streaming long-read sequence alignments for HLA predictions using HLAminer.* Current Protocols, 2025. **5**: p. e70124.
183. Warren, R.L., et al., *Derivation of HLA types from shotgun sequence datasets.* Genome Medicine, 2012. **4**: p. 1-8.
184. Liu, C., et al., *ATHLATES: accurate typing of human leukocyte antigen through exome sequencing.* Nucleic Acids Research, 2013. **41**: p. e142-e142.
185. Ka, S., et al., *HLAscan: genotyping of the HLA region using next-generation sequencing data.* BMC Bioinformatics, 2017. **18**: p. 1-11.
186. Kawaguchi, S., et al., *HLA-HD: an accurate HLA typing algorithm for next-generation sequencing data.* Human Mutation, 2017. **38**: p. 788-797.
187. Shukla, S.A., et al., *Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes.* Nature Biotechnology, 2015. **33**: p. 1152-1158.
188. Baker, R.E., et al., *Infectious disease in an era of global change.* Nature Reviews Microbiology, 2022. **20**: p. 193-205.
189. Duggal, P., et al., *The Evolving Field of Genetic Epidemiology: From Familial Aggregation to Genomic Sequencing.* American Journal of Epidemiology, 2019. **188**: p. 2069-2077.
190. Allison, A.C., *Protection afforded by sickle-cell trait against subtertian malarial infection.* British Medical Journal, 1954. **1**: p. 290.
191. Gong, L., et al., *Biochemical and immunological mechanisms by which sickle cell trait protects against malaria.* Malaria Journal, 2013. **12**(1): p. 317.
192. Band, G., et al., *Malaria protection due to sickle haemoglobin depends on parasite genotype.* Nature, 2022. **602**: p. 106-111.
193. Rowe, J.A., et al., *Blood group O protects against severe Plasmodium*

falciparum malaria through the mechanism of reduced rosetting. Proceedings of the National Academy of Sciences, 2007. **104**: p. 17471-17476.
194. Kariuki, S.N., et al., *Red blood cell tension protects against severe malaria in the Dantu blood group.* Nature, 2020. **585**: p. 579-583.
195. Mockenhaupt, F.P., et al., *α+-thalassemia protects African children from severe malaria.* Blood, 2004. **104**: p. 2003-2006.
196. Lu, Y.-F., et al., *IFNL3 mRNA structure is remodeled by a functional non-coding polymorphism associated with hepatitis C virus clearance.* Scientific Reports, 2015. **5**: p. 16037.
197. Thomas, D.L., et al., *Genetic variation in IL28B and spontaneous clearance of hepatitis C virus.* Nature, 2009. **461**: p. 798-801.
198. Keshvari, M., et al., *The interferon lambda 4 rs368234815 predicts treatment response to pegylated-interferon alpha and ribavirin in hemophilic patients with chronic hepatitis C.* Journal of Research in Medical Sciences, 2016. **21**: p. 72.
199. Aka, P.V., et al., *Association of the IFNL4-ΔG allele with impaired spontaneous clearance of hepatitis C virus.* The Journal of Infectious Diseases, 2014. **209**: p. 350-354.
200. Bibert, S., et al., *IL28B expression depends on a novel TT/-G polymorphism which improves HCV clearance prediction.* Journal of Experimental Medicine, 2013. **210**: p. 1109-1116.
201. Xie, X., L. Zhang, and Y.-Z. Chen, *Association between IFNL4 rs368234815 polymorphism and sustained virological response in chronic hepatitis C patients undergoing PEGylated interferon/ribavirin therapy: A meta-analysis.* Human Immunology, 2016. **77**: p. 609-615.
202. Prokunina-Olsson, L., et al., *A variant upstream of IFNL3 (IL28B) creating a new interferon gene IFNL4 is associated with impaired clearance of hepatitis C virus.* Nature Genetics, 2013. **45**: p. 164-171.
203. Murakawa, M., et al., *Impaired induction of interleukin 28B and expression of interferon λ 4 associated with nonresponse to interferon-based therapy in chronic hepatitis C.* Journal of Gastroenterology and Hepatology, 2015. **30**: p. 1075-1084.
204. Ferraris, P., et al., *Cellular mechanism for impaired hepatitis C virus clearance by interferon associated with IFNL3 gene polymorphisms relates to intrahepatic interferon-λ expression.* The American Journal of Pathology, 2016. **186**: p. 938-951.
205. Schneider, W.M., M.D. Chevillotte, and C.M. Rice, *Interferon-stimulated genes: a complex web of host defenses.* Annual Review of Immunology, 2014. **32**: p. 513-545.
206. Jimenez-Sanchez, G., B. Childs, and D. Valle, *Human disease genes.* Nature, 2001. **409**: p. 853-855.
207. Shriner, D., *Overview of admixture mapping.* Current Protocols, 2023. **3**: p.

e677.
208. Phan, L., et al., *The evolution of dbSNP: 25 years of impact in genomic research.* Nucleic Acids Research, 2025. **53**: p. D925-D931.
209. Sherry, S.T., et al., *dbSNP: the NCBI database of genetic variation.* Nucleic Acids Research, 2001. **29**: p. 308-311.
210. Syvänen, A.-C., *Accessing genetic variation: genotyping single nucleotide polymorphisms.* Nature Reviews Genetics, 2001. **2**: p. 930-942.
211. Klein, R.J., et al., *Complement factor H polymorphism in age-related macular degeneration.* Science, 2005. **308**: p. 385-389.
212. Abdellaoui, A., et al., *15 years of GWAS discovery: realizing the promise.* The American Journal of Human Genetics, 2023. **110**: p. 179-194.
213. Visscher, P.M., et al., *10 years of GWAS discovery: biology, function, and translation.* The American Journal of Human Genetics, 2017. **101**: p. 5-22.
214. Uffelmann, E., et al., *Genome-wide association studies.* Nature Reviews Methods Primers, 2021. **1**: p. 21.
215. Cross, B., R. Turner, and M. Pirmohamed, *Polygenic risk scores: An overview from bench to bedside for personalised medicine.* Frontiers in Genetics, 2022. **13**: p. 1000667.
216. McCarthy, M.I., et al., *Genome-wide association studies for complex traits: consensus, uncertainty and challenges.* Nature Reviews Genetics, 2008. **9**: p. 356-369.
217. Kraft, P., E. Zeggini, and J.P. Ioannidis, *Replication in genome-wide association studies.* Statistical Science, 2009. **24**: p. 561.
218. Fabo, T. and P. Khavari, *Functional characterization of human genomic variation linked to polygenic diseases.* Trends in Genetics, 2023. **39**: p. 462-490.
219. Zhang, Z., et al., *The construction of a haplotype reference panel using extremely low coverage whole genome sequences and its application in genome-wide association studies and genomic prediction in Duroc pigs.* Genomics, 2022. **114**: p. 340-350.
220. Ewaoluwagbemiga, E.O., et al., *Genome-wide association study and regional heritability mapping of protein efficiency and performance traits in Swiss Large White pigs.* BioRxiv, 2023: p. 2023.11.28.568963.
221. Wang, X., et al., *Imputation strategies for low-coverage whole-genome sequencing data and their effects on genomic prediction and genome-wide association studies in pigs.* Animal, 2024. **18**: p. 101258.
222. Yang, R., et al., *Accelerated deciphering of the genetic architecture of agricultural economic traits in pigs using a low-coverage whole-genome sequencing strategy.* GigaScience, 2021. **10**: p. giab048.
223. Yang, R., et al., *Genome-wide association analyses of multiple traits in Duroc pigs using low-coverage whole-genome sequencing strategy.* BioRxiv, 2019: p. 754671.

224. Mehrotra, A., et al., *Genome-Wide Association Testing for Haemorrhagic Bowel Syndrome in a Swiss Large White Pig Population.* BioRxiv, 2024: p. 2024.04.05.588256.
225. Shangguan, A., et al., *Genome-wide association study of growth and reproductive traits based on low-coverage whole-genome sequencing in a Chubao black-head goat population.* Gene, 2024. **931**: p. 148891.
226. Bell, S.M., et al., *GWAS using low-pass whole genome sequence reveals a novel locus in canine congenital idiopathic megaesophagus.* Mammalian Genome, 2023. **34**: p. 464-472.
227. Murgiano, L., et al., *Low-pass whole-genome mapping and imputation reveal genetic loci associated with optic nerve hypoplasia in dogs.* Investigative Ophthalmology & Visual Science, 2025. **66**: p. 1297-1297.
228. Wang, D., et al., *Cost-effectively dissecting the genetic architecture of complex wool traits in rabbits by low-coverage sequencing.* Genetics Selection Evolution, 2022. **54**: p. 14.
229. Li, J., et al., *Genome-wide association studies for egg quality traits in White Leghorn layers using low-pass sequencing and SNP chip data.* Journal of Animal Breeding and Genetics, 2022. **139**: p. 380-397.
230. Wang, H., et al., *Genome-Wide Association Analysis of Flavor Precursor Traits in Chengkou Mountain Chicken.* Animals, 2025. **15**: p. 1726.
231. Zhu, D., et al., *GCRP: integrated global chicken reference panel from 11,951 chicken genomes.* Genomics, Proteomics & Bioinformatics, 2025: p. qzaf032.
232. Nicod, J., et al., *Genome-wide association of multiple complex traits in outbred mice by ultra-low-coverage sequencing.* Nature Genetics, 2016. **48**: p. 912-918.
233. Sun, Y., et al., *Genotype imputation-based whole-genome association study for growth-related traits in golden pompano (Trachinotus ovatus).* Aquaculture, 2025. **596**: p. 741893.
234. Garcia, B.F., et al., *Prioritized imputed sequence variants from multi-population GWAS improve prediction accuracy for sea lice count in Atlantic salmon (Salmo salar).* Aquaculture, 2024. **581**: p. 740422.
235. Song, H., et al., *Cost-effective genomic prediction of critical economic traits in sturgeons through low-coverage sequencing.* Genomics, 2024. **116**: p. 110874.
236. Beemelmanns, A., et al., *Development of SNP Panels From Low-Coverage Whole Genome Sequencing (lcWGS) to Support Indigenous Fisheries for Three Salmonid Species in Northern Canada.* Molecular Ecology Resources, 2025. **25**: p. e14040.
237. Huang, X., et al., *Optimizing genotype imputation strategies for low coverage whole genome and 2b-RAD sequencing in bivalve molluscs.* Aquaculture, 2025: p. 742713.
238. Weller, C.A., et al., *Accurate, ultra-low coverage genome reconstruction and association studies in Hybrid Swarm mapping populations.* G3: Genes,

Genomes, Genetics, 2021. **11**: p. 12.
239. Bhattarai, G., et al., *Mapping and selection of downy mildew resistance in spinach cv. whale by low coverage whole genome sequencing.* Frontiers in Plant Science, 2022. **13**: p. 16.
240. Korani, W., et al., *De novo QTL-seq identifies loci linked to blanchability in peanut (Arachis hypogaea) and refines previously identified QTL with low coverage sequence.* Agronomy, 2021. **11**: p. 2201.
241. Cai, N., et al., *Sparse whole-genome sequencing identifies two loci for major depressive disorder.* Nature, 2015. **523**: p. 588-591.
242. Li, S.M., et al., *Ultra-low-coverage genome-wide association study-insights into gestational age using 17,844 embryo samples with preimplantation genetic testing.* Genome Medicine, 2023. **15**: p. 18.
243. Kim, S., et al., *Evaluation of low-pass genome sequencing in polygenic risk score calculation for Parkinson's disease.* Human Genomics, 2021. **15**: p. 58.
244. Low-Kam, C., et al., *Whole-genome sequencing in French Canadians from Quebec.* Human Genetics, 2016. **135**: p. 1213-1221.
245. Nostaeva, A., et al., *Case–control association study between polygenic risk score and COVID-19 severity in a Russian population using low-pass genome sequencing.* Epidemiology & Infection, 2025. **153**: p. e13.
246. He, Q., et al., *A genome-wide association study of neonatal metabolites.* Cell Genomics, 2024. **4**.
247. Linthorst, J., M. Nivard, and E.A. Sistermans, *GWAS shows the genetics behind cell-free DNA and highlights the importance of p. Arg206Cys in DNASE1L3 for non-invasive testing.* Cell Reports, 2024. **43**.
248. Bhatt, I.S., et al., *Polygenic risk score-based association analysis of speech-in-noise and hearing threshold measures in healthy young adults with self-reported normal hearing.* Journal of the Association for Research in Otolaryngology, 2023. **24**: p. 513-525.
249. Al-Jumaan, M., et al., *Interplay of Mendelian and polygenic risk factors in Arab breast cancer patients.* Genome Medicine, 2023. **15**: p. 65.
250. Wünnemann, F., et al., *Validation of genome-wide polygenic risk scores for coronary artery disease in French Canadians.* Circulation: Genomic and Precision Medicine, 2019. **12**: p. e002481.
251. Bien, S.A., et al., *Enrichment of colorectal cancer associations in functional regions: Insight for using epigenomics data in the analysis of whole genome sequence-imputed GWAS data.* PLoS One, 2017. **12**: p. e0186518.
252. Dill-McFarland, K.A., et al., *Genome-wide association study in Brazil identifies genetic susceptibility to tuberculosis with single-cell gene expression effects.* MedRxiv, 2025: p. 2025.03.13.25323932.
253. Anterasian, C., et al., *Genome-Wide Association Study Identifies Novel Variant Associated with Susceptibility to Pediatric Tuberculosis.* Journal of Human Immunity, 2025. **1**: p. CIS2025abstract. 57.

254. Li, Y. and Y. Li, *Patterns of symptoms in major depressive disorder and genetics of the disorder using low-pass sequencing data*. 2013, Oxford University, UK.
255. Sun, S., et al., *Identifying risk variants for embryo aneuploidy using ultra-low coverage whole-genome sequencing from preimplantation genetic testing.* The American Journal of Human Genetics, 2023. **110**: p. 2092-2102.
256. Cai, N., et al., *RETRACTED ARTICLE: 11,670 whole-genome sequences representative of the Han Chinese population from the CONVERGE project.* Scientific Data, 2017. **4**: p. 1-14.
257. Homburger, J.R., et al., *Low coverage whole genome sequencing enables accurate assessment of common variants and calculation of genome-wide polygenic scores.* Genome Medicine, 2019. **11**: p. 74.
258. Bhatt, I.S., et al., *A genome-wide association study reveals a polygenic architecture of speech-in-noise deficits in individuals with self-reported normal hearing.* Scientific Reports, 2024. **14**: p. 13089.
259. Laflamme, R., et al., *Replication of a GWAS signal near HLA-DQA2 with AML using a disease-only cohort and external population-based controls.* Blood Neoplasia, 2025: p. 100118.
260. Liu, S., et al., *Genomic analyses from non-invasive prenatal testing reveal genetic associations, patterns of viral infections, and Chinese population history.* Cell, 2018. **175**: p. 347-359. e14.
261. Zook, J.M., et al., *Extensive sequencing of seven human genomes to characterize benchmark reference materials.* Scientific Data, 2016. **3**: p. 1-26.
262. Zook, J.M., et al., *Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls.* Nature Biotechnology, 2014. **32**: p. 246-251.
263. Eberle, M.A., et al., *A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree.* Genome Research, 2017. **27**: p. 157-164.
264. Bowden, R., et al., *Sequencing of human genomes with nanopore technology.* Nature Communications, 2019. **10**: p. 1869.
265. Ellis, P., et al., *Reliable detection of somatic mutations in solid tissues by laser-capture microdissection and low-input DNA sequencing.* Nature Protocols, 2021. **16**: p. 841-871.
266. Real, E., et al., *A single-cell atlas of Plasmodium falciparum transmission through the mosquito.* Nature Communications, 2021. **12**: p. 3196.
267. Illumina. *Illumina DNA Prep*. Jul 15, 2025; Available from: https://emea.illumina.com/products/by-brand/nextera.html.
268. Biolabs, N.E. *NEBNext® Ultra™ DNA Library Prep Kit for Illumina®*. Jul 15, 2025; Available from: https://www.neb.com/en-gb/products/e7370-nebnext-ultra-dna-library-prep-kit-for-illumina#Product%20Information_Advantages%20and%20Features.

269. Abecasis, G.R., et al., *A map of human genome variation from population-scale sequencing.* Nature, 2010. **467**: p. 1061-1073.
270. Tan, G., et al., *Long fragments achieve lower base quality in Illumina paired-end sequencing.* Scientific Reports, 2019. **9**: p. 2856.
271. Dohm, J.C., et al., *Substantial biases in ultra-short read data sets from high-throughput DNA sequencing.* Nucleic Acids Research, 2008. **36**: p. e105.
272. Wang, B., et al., *An adaptive decorrelation method removes Illumina DNA base-calling errors caused by crosstalk between adjacent clusters.* Scientific Reports, 2017. **7**: p. 41348.
273. Kircher, M., P. Heyn, and J. Kelso, *Addressing challenges in the production and analysis of illumina sequencing data.* BMC Genomics, 2011. **12**: p. 382.
274. McNulty, S.N., et al., *Impact of reducing DNA input on next-generation sequencing library complexity and variant detection.* The Journal of Molecular Diagnostics, 2020. **22**: p. 720-727.
275. Flower, B., et al. *High cure rates for hepatitis C virus genotype 6 in advanced liver fibrosis with 12 weeks sofosbuvir and daclatasvir: the Vietnam SEARCH study*. in *Open Forum Infectious Diseases*. 2021. Oxford University Press US.
276. Kruglyak, S. *Measuring the Accuracy of Element AVITI™ Sequencing Data*. Apr 22, 2025; Available from: https://www.elementbiosciences.com/blog/measuring-accuracy-element-aviti-sequencing-data.
277. Liao, W.-W., et al., *A draft human pangenome reference.* Nature, 2023. **617**: p. 312-324.
278. Mölder, F., et al., *Sustainable data analysis with Snakemake.* F1000Research, 2021. **10**: p. 33.
279. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows–Wheeler transform.* Bioinformatics, 2009. **25**: p. 1754-1760.
280. Institute, B. *Picard*. May 13, 2025; Available from: http://broadinstitute.github.io/picard/.
281. Li, H. *Seqtk*. Jul 16, 2025; Available from: https://github.com/lh3/seqtk.
282. Bonfield, J.K., et al., *HTSlib: C library for reading/writing high-throughput sequencing data.* GigaScience, 2021. **10**: p. giab007.
283. Danecek, P., et al., *Twelve years of SAMtools and BCFtools.* GigaScience, 2021. **10**: p. giab008.
284. Li, H., et al., *The sequence alignment/map format and SAMtools.* Bioinformatics, 2009. **25**: p. 2078-2079.
285. Marçais, G. and C. Kingsford, *A fast, lock-free approach for efficient parallel counting of occurrences of k-mers.* Bioinformatics, 2011. **27**: p. 764-770.
286. Band, G. *iorek*. May 13, 2025; Available from: https://github.com/gavinband/iorek.
287. Homer, N. *DWGSIM*. Jun 22, 2024; Available from: https://github.com/nh13/DWGSIM.

288. Ono, Y., K. Asai, and M. Hamada, *PBSIM: PacBio reads simulator—toward accurate genome assembly.* Bioinformatics, 2013. **29**: p. 119-121.
289. McGuire, W., et al., *Variation in the TNF-α promoter region associated with susceptibility to cerebral malaria.* Nature, 1994. **371**: p. 508-511.
290. Choudhury, A., et al., *High-depth African genomes inform human migration and health.* Nature, 2020. **586**: p. 741-748.
291. Consortium, H.A., *Enabling the genomic revolution in Africa: H3Africa is developing capacity for health-related genomics research in Africa.* Science, 2014. **344**: p. 1346.
292. Mulder, N., et al., *H3Africa: current perspectives.* Pharmacogenomics and Personalized Medicine, 2018: p. 59-66.
293. Browne, P.D., et al., *GC bias affects genomic and metagenomic reconstructions, underrepresenting GC-poor organisms.* GigaScience, 2020. **9**: p. giaa008.
294. Chen, Y.-C., et al., *Effects of GC bias in next-generation-sequencing data on de novo genome assembly.* PLoS One, 2013. **8**: p. e62856.
295. Nurk, S., et al., *The complete sequence of a human genome.* Science, 2022. **376**: p. 44-53.
296. Chen, S., et al., *A genomic mutational constraint map using variation in 76,156 human genomes.* Nature, 2024. **625**: p. 92-100.
297. Cahoon, J.L., et al., *Imputation accuracy across global human populations.* The American Journal of Human Genetics, 2024. **111**: p. 979-989.
298. Jiménez-Kaufmann, A., et al., *Imputation Performance in Latin American Populations: Improving Rare Variants Representation With the Inclusion of Native American Genomes.* Frontiers in Genetics, 2022. **12**: p. 11.
299. Yang, M.-Y., et al., *SEAD reference panel with 22,134 haplotypes boosts rare variant imputation and genome-wide association analysis in Asian populations.* Nature Communications, 2024. **15**: p. 10839.
300. Bouwman, A.C. and R.F. Veerkamp, *Consequences of splitting whole-genome sequencing effort over multiple breeds on imputation accuracy.* BMC Genetics, 2014. **15**: p. 105.
301. Hindorff, L.A., et al., *Potential etiologic and functional implications of genome-wide association loci for human diseases and traits.* Proceedings of the National Academy of Sciences, 2009. **106**: p. 9362-9367.
302. Cerezo, M., et al., *The NHGRI-EBI GWAS Catalog: standards for reusability, sustainability and diversity.* Nucleic Acids Research, 2025. **53**: p. D998-D1005.
303. Seal, R.L., et al., *Genenames. org: the HGNC resources in 2023.* Nucleic Acids Research, 2023. **51**: p. D1003-D1009.
304. Rowe, J.A., D.H. Opi, and T.N. Williams, *Blood groups and malaria: fresh insights into pathogenesis and identification of targets for intervention.* Current Opinion in Hematology, 2009. **16**: p. 480-487.

305. Loh, P.-R., et al., *Reference-based phasing using the Haplotype Reference Consortium panel.* Nature Genetics, 2016. **48**: p. 1443-1448.
306. Nimmerjahn, F. and J.V. Ravetch, *Fcγ receptors as regulators of immune responses.* Nature Reviews Immunology, 2008. **8**: p. 34-47.
307. Sepúlveda-Delgado, J., L. Llorente, and S. Hernández-Doño, *A Comprehensive Review of Fc Gamma Receptors and Their Role in Systemic Lupus Erythematosus.* International Journal of Molecular Sciences, 2025. **26**: p. 1851.
308. Frampton, S., et al., *Fc gamma receptors: Their evolution, genomic architecture, genetic variation, and impact on human disease.* Immunological Reviews, 2024. **328**: p. 65-97.
309. Band, G., et al., *Imputation-based meta-analysis of severe malaria in three African populations.* PLoS Genetics, 2013. **9**: p. e1003509.
310. Fikadu, M. and E. Ashenafi, *Malaria: an overview.* Infection and Drug Resistance, 2023: p. 3339-3347.
311. Phillips, M.A., et al., *Malaria.* Nature Reviews Disease Primers, 2017. **3**: p. 17050.
312. Band, G., et al., *A novel locus of resistance to severe malaria in a region of ancient balancing selection.* Nature, 2015. **526**: p. 253-257.
313. Bolger, A.M., M. Lohse, and B. Usadel, *Trimmomatic: a flexible trimmer for Illumina sequence data.* Bioinformatics, 2014. **30**: p. 2114-2120.
314. Xu, H., et al., *FastUniq: a fast de novo duplicates removal tool for paired short reads.* PLoS One, 2012. **7**: p. e52249.
315. Quail, M.A., et al., *SASI-Seq: sample assurance Spike-Ins, and highly differentiating 384 barcoding for Illumina sequencing.* BMC Genomics, 2014. **15**: p. 110.
316. Lowy, E., S. Fairley, and P. Flicek, *Variant calling across 505 openly consented samples from four Gambian populations on GRCh38.* Wellcome Open Research, 2021. **6**: p. 239.
317. Band, G. *QCTOOL*. May 13, 2025; Available from: https://www.chg.ox.ac.uk/~gav/qctool_v2/.
318. Van der Auwera, G.A. and B.D. O'Connor, *Genomics in the cloud: using Docker, GATK, and WDL in Terra*. 2020: O'Reilly Media.
319. Band, G. *SNPTEST*. May 13, 2025; Available from: https://www.chg.ox.ac.uk/~gav/snptest/#introduction.
320. Chang, C.C., et al., *Second-generation PLINK: rising to the challenge of larger and richer datasets.* GigaScience, 2015. **4**: p. s13742-015-0047-8.
321. Purcell, S., et al., *PLINK: a tool set for whole-genome association and population-based linkage analyses.* The American Journal of Human Genetics, 2007. **81**: p. 559-575.
322. Karimzadeh, M., et al., *Umap and Bismap: quantifying genome and methylome mappability.* Nucleic Acids Research, 2018. **46**: p. e120-e120.

323. Yates, A., et al., *The Ensembl REST API: Ensembl data for any language.* Bioinformatics, 2015. **31**: p. 143-145.
324. Goswami, R., et al., *Presence of strong association of the major histocompatibility complex (MHC) class I allele HLA-A*26:01 with idiopathic hypoparathyroidism.* The Journal of Clinical Endocrinology & Metabolism, 2012. **97**: p. E1820-E1824.
325. Andersson, G., *Evolution of the human HLA-DR region.* Front Biosci, 1998. **3**: p. d739-745.
326. Vlachopoulou, E., et al., *Evaluation of HLA-DRB1 imputation using a Finnish dataset.* Tissue Antigens, 2014. **83**: p. 350-355.
327. Cleal, K. *pywfa*. May 14, 2025; Available from: https://github.com/kcleal/pywfa.
328. Marco-Sola, S., et al., *Fast gap-affine pairwise alignment using the wavefront algorithm.* Bioinformatics, 2021. **37**: p. 456-463.
329. Frith, M.C., *How sequence alignment scores correspond to probability models.* Bioinformatics, 2020. **36**: p. 408-415.
330. Arango, M.-T., et al., *HLA-DRB1 the notorious gene in the mosaic of autoimmunity.* Immunologic Research, 2017. **65**: p. 82-98.
331. Barton, A., et al., *Sequence based HLA-DRB1, -DQB1, and -DPB1 allele and haplotype frequencies in The Gambia.* Human Immunology, 2023. **84**: p. 69-70.
332. Gorski, J., *The HLA-DRw8 lineage was generated by a deletion in the DR B region followed by first domain diversification.* Journal of Immunology, 1989. **142**: p. 4041-4045.
333. Mentzer, A.J., et al., *High-resolution African HLA resource uncovers HLA-DRB1 expression effects underlying vaccine response.* Nature Medicine, 2024: p. 37.
334. Traherne, J.A., et al., *Genetic analysis of completely sequenced disease-associated MHC haplotypes identifies shuffling of segments in recent human history.* PLoS Genetics, 2006. **2**: p. e9.
335. Yang, F., et al., *A reliable, high-resolution and high-throughput genotyping method for HLA-DRB1.* Human Immunology, 2015. **76**: p. 397-401.
336. Robinson, J., et al., *IMGT/HLA database–a sequence database for the human major histocompatibility complex.* Tissue Antigens, 2000. **55**: p. 280-287.
337. Barker, D.J., et al., *The ipd-imgt/hla database.* Nucleic Acids Research, 2023. **51**: p. D1053-D1060.
338. Carvalho, C.M. and J.R. Lupski, *Mechanisms underlying structural variant formation in genomic disorders.* Nature Reviews Genetics, 2016. **17**: p. 224-238.
339. Diskin, S.J., et al., *Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms.* Nucleic Acids Research, 2008. **36**: p. e126-e126.

340. Wang, K., et al., *Modeling genetic inheritance of copy number variations.* Nucleic Acids research, 2008. **36**: p. e138-e138.
341. Wang, K., et al., *PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data.* Genome Research, 2007. **17**: p. 1665-1674.
342. Audano, P.A., et al., *Characterizing the major structural variant alleles of the human genome.* Cell, 2019. **176**: p. 663-675. e19.
343. Ebert, P., et al., *Haplotype-resolved diverse human genomes and integrated analysis of structural variation.* Science, 2021. **372**: p. eabf7117.
344. Gahmberg, C.G., M. Jokinen, and L.C. Andersson, *Expression of the major sialoglycoprotein (glycophorin) on erythroid cells in human bone marrow.* 1978.
345. Leffler, E.M., et al., *Resistance to malaria through structural variation of red blood cell invasion receptors.* Science, 2017. **356**: p. eaam6393.
346. Hans, C., A. Dobra, and M. West, *Shotgun stochastic search for "large p" regression.* Journal of the American Statistical Association, 2007. **102**: p. 507-516.
347. BLANCHARD, D., et al., *Hybrid glycophorins from human erythrocyte membranes: Isolation and complete structural analysis of the novel sialoglycoprotein from St (a+) red cells.* European Journal of Biochemistry, 1987. **167**: p. 361-366.
348. Blumenfeld, O., A. Smith, and J. Moulds, *Membrane glycophorins of Dantu blood group erythrocytes.* Journal of Biological Chemistry, 1987. **262**: p. 11864-11870.
349. Baird, D.M., *Telomeres and genomic evolution.* Philosophical Transactions of the Royal Society B: Biological Sciences, 2018. **373**: p. 20160437.
350. Rocha, J.L., R.N. Lou, and P.H. Sudmant, *Structural variation in humans and our primate kin in the era of telomere-to-telomere genomes and pangenomics.* Current Opinion in Genetics & Development, 2024. **87**: p. 102233.
351. Lin, Y.-L. and O. Gokcumen, *Fine-scale characterization of genomic structural variation in the human genome reveals adaptive and biomedically relevant hotspots.* Genome Biology and Evolution, 2019. **11**: p. 1136-1151.
352. Kanoh, J., *Subtelomeres: hotspots of genome variation.* Genes & Genetic Systems, 2023. **98**: p. 155-160.
353. Amuzu, D.S., et al., *High-throughput genotyping assays for identification of glycophorin B deletion variants in population studies.* Experimental Biology and Medicine, 2021. **246**: p. 916-928.
354. Chaisson, M.J., et al., *Multi-platform discovery of haplotype-resolved structural variation in human genomes.* Nature Communications, 2019. **10**: p. 1784.
355. Blach, S., et al., *Global prevalence and genotype distribution of hepatitis C virus infection in 2015: a modelling study.* The lancet Gastroenterology &

Hepatology, 2017. **2**: p. 161-176.
356. Sallam, M. and R. Khalil, *Contemporary insights into hepatitis c virus: a comprehensive review.* Microorganisms, 2024. **12**: p. 1035.
357. Stasi, C., et al., *The epidemiology of chronic hepatitis C: where we are now.* Livers, 2024. **4**: p. 172-181.
358. Mbisa, J.L., et al., *Identification of 2 novel subtypes of hepatitis C virus genotype 8 and a potential new genotype successfully treated with direct acting antivirals.* The Journal of Infectious Diseases, 2024. **230**: p. e1254-e1262.
359. Marshall, A.D., et al., *Direct-acting antiviral therapies for hepatitis C infection: global registration, reimbursement, and restrictions.* The Lancet Gastroenterology & Hepatology, 2024. **9**: p. 366-382.
360. Irshad, M., D.S. Mankotia, and K. Irshad, *An insight into the diagnosis and pathogenesis of hepatitis C virus infection.* World Journal of Gastroenterology: WJG, 2013. **19**: p. 7896.
361. Nahon, P. and A. Cobat, *Human genetics of HCV infection phenotypes in the era of direct-acting antivirals.* Human Genetics, 2020. **139**: p. 855-863.
362. Walker, A.J., et al., *Host genetic factors associated with hepatocellular carcinoma in patients with hepatitis C virus infection: A systematic review.* Journal of Viral Hepatitis, 2018. **25**: p. 442-456.
363. Yan, Z. and Y. Wang, *Viral and host factors associated with outcomes of hepatitis C virus infection.* Molecular Medicine Reports, 2017. **15**: p. 2909-2924.
364. Rauch, A., et al., *Genetic variation in IL28B is associated with chronic hepatitis C and treatment failure: a genome-wide association study.* Gastroenterology, 2010. **138**: p. 1338-1345. e7.
365. Ansari, M.A., et al., *Genome-to-genome analysis highlights the effect of the human innate and adaptive immune systems on the hepatitis C virus.* Nature Genetics, 2017. **49**: p. 666-673.
366. Laidlaw, S.M. and L.B. Dustin, *Interferon lambda: opportunities, risks, and uncertainties in the fight against HCV.* Frontiers in Immunology, 2014. **5**: p. 545.
367. Kumar, V., et al., *Genome-wide association study identifies a susceptibility locus for HCV-induced hepatocellular carcinoma.* Nature Genetics, 2011. **43**: p. 455-458.
368. Miki, D., et al., *Variation in the DEPDC5 locus is associated with progression to hepatocellular carcinoma in chronic hepatitis C virus carriers.* Nature Genetics, 2011. **43**: p. 797-800.
369. Matsuura, K., et al., *Genome-wide association study identifies TLL1 variant associated with development of hepatocellular carcinoma after eradication of hepatitis C virus infection.* Gastroenterology, 2017. **152**: p. 1383-1394.
370. Zignego, A.L., et al., *Genome-wide association study of hepatitis C virus-and*

cryoglobulin-related vasculitis. Genes & Immunity, 2014. **15**: p. 500-505.
371. Hong, X., et al., *Human leukocyte antigen class II DQB1*0301, DRB1*1101 alleles and spontaneous clearance of hepatitis C virus infection: a meta-analysis.* World Journal of Gastroenterology, 2005. **11**: p. 7302.
372. Hraber, P., C. Kuiken, and K. Yusim, *Evidence for human leukocyte antigen heterozygote advantage against hepatitis C virus infection.* Hepatology, 2007. **46**: p. 1713-1721.
373. Lee, M.H., et al., *Human leukocyte antigen variants and risk of hepatocellular carcinoma modified by hepatitis C virus genotypes: a genome-wide association study.* Hepatology, 2018. **67**: p. 651-661.
374. Cooke, G.S., et al., *Treatment options to support the elimination of hepatitis C: an open-label, factorial, randomised controlled non-inferiority trial.* The Lancet, 2025.
375. McCabe, L., et al., *The design and statistical aspects of VIETNARMS: a strategic post-licensing trial of multiple oral direct-acting antiviral hepatitis C treatment strategies in Vietnam.* Trials, 2020. **21**: p. 413.
376. Arslan, S., et al., *Sequencing by avidity enables high accuracy with low reagent consumption.* Nature Biotechnology, 2024. **42**: p. 132-138.
377. Howie, B., J. Marchini, and M. Stephens, *Genotype imputation with thousands of genomes.* G3: Genes, Genomes, Genetics, 2011. **1**: p. 457-470.
378. Logsdon, G.A., et al., *The structure, function and evolution of a complete human chromosome 8.* Nature, 2021. **593**: p. 101-107.
379. Manichaikul, A., et al., *Robust relationship inference in genome-wide association studies.* Bioinformatics, 2010. **26**: p. 2867-2873.
380. Li, G. and H. Zhu, *Genetic studies: the linear mixed models in genome-wide association studies.* The Open Bioinformatics Journal, 2013. **7**.
381. Zhou, X. and M. Stephens, *Genome-wide efficient mixed-model analysis for association studies.* Nature Genetics, 2012. **44**: p. 821-824.
382. Benjamini, Y. and Y. Hochberg, *Controlling the false discovery rate: a practical and powerful approach to multiple testing.* Journal of the Royal Statistical Society: Series B (Methodological), 1995. **57**: p. 289-300.
383. Stöhr, H., et al., *A novel gene encoding a putative transmembrane protein with two extracellular CUB domains and a low-density lipoprotein class A module: isolation of alternatively spliced isoforms in retina and brain.* Gene, 2002. **286**: p. 223-231.
384. Ng, D., et al., *Neto1 is a novel CUB-domain NMDA receptor–interacting protein required for synaptic plasticity and learning.* PLoS Biology, 2009. **7**: p. e1000041.
385. McLaren, W., et al., *The ensembl variant effect predictor.* Genome Biology, 2016. **17**: p. 122.
386. Lonsdale, J., et al., *The genotype-tissue expression (GTEx) project.* Nature Genetics, 2013. **45**: p. 580-585.

387. Wang, H., et al., *ORP2 delivers cholesterol to the plasma membrane in exchange for phosphatidylinositol 4, 5-bisphosphate (PI (4, 5) P2).* Molecular Cell, 2019. **73**: p. 458-473. e7.
388. Mankouri, J., et al., *Enhanced hepatitis C virus genome replication and lipid accumulation mediated by inhibition of AMP-activated protein kinase.* Proceedings of the National Academy of Sciences, 2010. **107**: p. 11549-11554.
389. Gastaminza, P., et al., *Cellular determinants of hepatitis C virus assembly, maturation, degradation, and secretion.* Journal of Virology, 2008. **82**: p. 2120-2129.
390. Kapadia, S.B., et al., *Initiation of hepatitis C virus infection is dependent on cholesterol and cooperativity between CD81 and scavenger receptor B type I.* Journal of Virology, 2007. **81**: p. 374-383.
391. Wang, M. and S. Xu, *Statistical power in genome-wide association studies and quantitative trait locus mapping.* Heredity, 2019. **123**: p. 287-306.
392. Ge, D., et al., *Genetic variation in IL28B predicts hepatitis C treatment-induced viral clearance.* Nature, 2009. **461**: p. 399-401.
393. Thong, V.D., et al., *Hepatitis C virus genotype 6: virology, epidemiology, genetic variation and clinical implication.* World journal of Gastroenterology: WJG, 2014. **20**: p. 2927.
394. Thong, V.D., et al., *Influence of host and viral factors on patients with chronic hepatitis C virus genotype 6 treated with pegylated interferon and ribavirin: a systematic review and meta-analysis.* Intervirology, 2016. **58**: p. 373-381.
395. Akkarathamrongsin, S., et al., *Early viral kinetics during hepatitis C virus genotype 6 treatment according to IL28B polymorphisms.* World Journal of Gastroenterology: WJG, 2014. **20**: p. 10599.
396. Akkarathamrongsin, S., et al., *IFNL3 (IL28B) and IFNL4 polymorphisms are associated with treatment response in Thai patients infected with HCV genotype 1, but not with genotypes 3 and 6.* Journal of Medical Virology, 2014. **86**: p. 1482-1490.
397. Ou, G., et al., *The roles of HLA-DQB1 gene polymorphisms in hepatitis B virus infection.* Journal of Translational Medicine, 2018. **16**: p. 362.
398. Liu, C. and B. Cheng, *Association of polymorphisms of human leucocyte antigen-DQA1 and DQB1 alleles with chronic hepatitis B virus infection, liver cirrhosis and hepatocellular carcinoma in Chinese.* International Journal of Immunogenetics, 2007. **34**: p. 373-378.
399. Mokbel, A.N., et al., *Association of HLA-DQB*06 with susceptibility to systemic lupus erythematosus in Egyptians.* The Egyptian Rheumatologist, 2015. **37**: p. 17-22.
400. Kang, H., et al., *Comparison of HLA class II genes in Caucasoid, Chinese, and Japanese patients with primary Sjögren's syndrome.* The Journal of Immunology, 1993. **150**: p. 3615-3623.
401. Band, G. and J. Marchini, *BGEN: a binary file format for imputed genotype*

and haplotype data. BioRxiv, 2018: p. 308296.
402. Price, A.L., et al., *Principal components analysis corrects for stratification in genome-wide association studies.* Nature Genetics, 2006. **38**: p. 904-909.
403. Boughton, A.P., et al., *LocusZoom. js: interactive and embeddable visualization of genetic association study results.* Bioinformatics, 2021. **37**: p. 3017-3018.
404. Krassowski, M. *jupyter-locuszoom*. Jul 15, 2025; Available from: https://github.com/krassowski/jupyter-locuszoom.
405. Boltz, T.A., et al., *A blended genome and exome sequencing method captures genetic variation in an unbiased, high-quality, and cost-effective manner.* BioRxiv, 2024.
406. Dou, J., et al., *Using off-target data from whole-exome sequencing to improve genotyping accuracy, association analysis and polygenic risk prediction.* Briefings in Bioinformatics, 2021. **22**: p. bbaa084.
407. Brand, H., et al., *High-resolution and noninvasive fetal exome screening.* New England Journal of Medicine, 2023. **389**: p. 2014-2016.