The impact of antigen presentation pathway genetic variation on infectious disease phenotypes

Qijing Shen
Reuben College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Michaelmas 2025

Acknowledgements

I would like to express my sincere gratitude to everyone who has supported me throughout my three-year DPhil journey in Clinical Medicine. In particular, I wish to thank my supervisors, Azim Ansari and Gavin Band; my examiners; the Nuffield Department of Medicine; Reuben College; and my beloved family and friends.

To Gavin, thank you for your exceptional patience and mentorship. I am especially grateful for your guidance and the time you devoted to helping me improve my academic writing, technical skills, and statistical modelling, including Bayesian inference, Snakemake, and C++. Your encouragement and thoughtful suggestions were invaluable during the challenging phases of my statistical modelling work. I am truly grateful for the time and effort you invested in my development.

To Azim, I am deeply grateful for your expert guidance in advanced R programming, epidemiological methods, sequencing data analysis, and presentation skills. Your encouragement to refine my presentation abilities during combined group meetings and conferences has been especially important in helping me communicate and present my research with increasing confidence and independence.

My gratitude also extends to the Nuffield Department of Medicine and Reuben College. I am particularly grateful to Robert Gilbert for his help with my final-year funding application. I also want to thank the Reuben College staff, Stephen, Kirren, and Caroline, for their warm and steadfast support during difficult times. The vibrant community at Reuben, together with the OUISC and the painting club, greatly enriched my DPhil experience.

Lastly, I owe my deepest appreciation to my family. To my partner, Mingtao Xia, thank you for your unwavering companionship and emotional support throughout this journey. To my mother, Xiaoyan Li, I will always be grateful for your love and steadfast belief in the value of my education. I also want to express my heartfelt thanks to my father, my uncle, my maternal grandparents, and my friends for their constant encouragement and support.

Statement of Contribution

The laboratory work for this thesis, conducted within the STOP-HCV cohort, was carried out by Azim Ansari, Daisy Jennings, George Airey, Mike Xu, Chris Davis, Ana Da Silva Filipe, Andrew Hayward, Graham Cooke, Paul Klenerman, Eleanor Barnes, Emma Thomson, John McLauchlan, and Will Irving, who generated and contributed samples to the STOP-HCV study.

Data cleaning, genotype imputation, principal component analysis, covariate preparation, and clinical data processing for the STOP-HCV cohort, including analyses of HCV spontaneous clearance versus persistence, were conducted by Azim Ansari, Jocelyn Quistrebert, and Haiting Chai.

For the MalariaGEN cohort, phased genotyping data were generated by Gavin Band. I used these data and subsequently performed HLA imputation and genotype imputation using an imputation server.

The UK Biobank serological panel data were accessed and analysed through the UK Biobank Research Analysis Platform. Guidance on the structure and handling of the serological data was provided by Alexander Menzer and Amanda Chong. Genotype data for the selected subsample were phased by our group, with technical support from Gavin Band in the use of the UK Biobank Research Analysis Platform and associated analytical tools.

For the HLA-related components of this work, data on HLA features, including tapasin dependence, supertypes, and expression scores, were provided by Mary Carrington's research group. ERAP1/2 allotype and variant data were provided by Edd James and Emma Reeves.

For the statistical modelling and pathway-based framework described in Chapters 3 and 4, the interaction framework was initially proposed by Gavin Band and Azim Ansari. I subsequently developed the algorithms, conducted the modelling and simulation studies, applied the framework to the data, and built the associated software pipeline. The large-scale computational tasks, particularly the simulation studies,

were supported by the Oxford Biomedical Research Computing (BMRC) cluster and its support team.

Finally, I am deeply grateful to my supervisors, Azim Ansari and Gavin Band, for their invaluable advice, guidance, and support throughout my PhD, which fundamentally shaped this thesis.

Abstract

The antigen presentation pathway (APP) plays a pivotal role in adaptive immunity by processing and presenting pathogen-derived peptides to T cells, thereby shaping immune responses to infection. Genetic variation within the APP, including highly polymorphic human leukocyte antigen (HLA) genes, endoplasmic reticulum aminopeptidases (ERAP1 and ERAP2), transporter associated with antigen processing (TAP) genes, and proteasome subunits, has been implicated in susceptibility to infectious diseases. However, the complex genetic architecture of the APP, characterised by strong linkage disequilibrium, epistatic interactions, and population-specific variation, presents major challenges for identifying causal mechanisms.

My thesis addresses these challenges through three integrated aims. First, we develop a scalable bioinformatics pipeline to systematically extract and characterise genetic features of the APP across multiple cohorts. Second, we introduce a novel Bayesian statistical framework based on the regularised horseshoe prior, implemented via a maximum a posteriori (MAP) estimator, to enable robust fine-mapping of genetic associations in high-dimensional, correlated settings. Third, we apply this framework to jointly model the main and interaction effects among APP components in relation to infectious disease phenotypes.

By integrating pathway-wide genetic features within a flexible Bayesian framework, this work advances our understanding of how genetic variation in the APP shapes immune response diversity and infectious disease outcomes. These findings provide a foundation for future mechanistic and translational studies aimed at linking molecular diversity to immune function and clinical phenotypes.

Contents

List of Figures

List of Tables

Abbreviations

APP	Antigen Presentation Pathway
APC	Antigen-Presenting Cell
ASM	Active Subspace Method
BF	Bayes Factor
BIC	Bayesian Information Criterion
CNX	Calnexin
CRT	Calreticulin
CTL	Cytotoxic T Lymphocyte
DIC	Deviance Information Criterion
EBV	Epstein–Barr Virus
ER	Endoplasmic Reticulum
ERAP1/2	Endoplasmic Reticulum Aminopeptidase 1 and 2
ERp57	Endoplasmic Reticulum Protein 57
GWAS	Genome-Wide Association Study
HCC	Hepatocellular Carcinoma
HCV	Hepatitis C Virus
HHV	Human Herpesvirus
HIV	Human Immunodeficiency Virus
HLA	Human Leukocyte Antigen
IFN-γ	Interferon Gamma
IL-21	Interleukin-21
LD	Linkage Disequilibrium
LM	Linear Model
MAF	Minor Allele Frequency
MAP	Maximum A Posteriori
MCMC	Markov Chain Monte Carlo
MFI	Median Fluorescence Intensity
MHC	Major Histocompatibility Complex
MLE	Maximum Likelihood Estimation
NK	Natural Killer Cell
OR	Odds Ratio
PC	Principal Component
PLC	Peptide Loading Complex

PSMB	Proteasome Subunit Beta
SE	Standard Error
SNP	Single Nucleotide Polymorphism
TAP	Transporter Associated with Antigen Processing
TAPBP	TAP Binding Protein (Tapasin)
TCR	T Cell Receptor
TNF-α	Tumour Necrosis Factor Alpha
UKB	UK Biobank
VZV	Varicella-Zoster Virus

Chapter 1

Review of the Antigen Presentation Pathway and the Impact of its Genetic Variation on Human Health

1.1 Introduction

Antigen processing and presentation coordinate the detection of intracellular and extracellular proteins and the activation of T cell responses, forming a central pillar of immune surveillance in health and disease. The classical pathways mediated by the major histocompatibility complex (MHC), referred to as human leukocyte antigen (HLA) in humans, integrate proteolysis, peptide transport, peptide editing, and cell surface display to define the repertoire of peptides recognised by $CD8^+$ and $CD4^+$ T cells. Over the past decades, extensive research has emphasised the pivotal role of the antigen presentation pathway in infectious and immune-mediated diseases, underscoring the contribution of HLA molecules and endoplasmic reticulum aminopeptidases (ERAPs) in shaping the peptide landscape, modulating immune recognition, and influencing disease susceptibility and progression.

Genetic variation in antigen presentation genes represents one of the strongest and most consistent signals across human complex diseases. In particular, extensive research has focused on the highly polymorphic HLA region and the ERAP1/ERAP2 loci, given their central roles in shaping the antigenic peptide repertoire. The HLA region shows broad associations with infectious, autoimmune, and inflammatory conditions, while variation in ERAP1 and ERAP2, in either coding or regulatory regions, can modulate disease risk by altering peptide processing. However, the architecture of

this variation is highly complex: long-range and population-specific linkage disequilibrium (LD), extensive allelic diversity, and allele-specific expression can complicate analyses and obscure the true causal mechanisms. To overcome these challenges, it is essential to model haplotypes, regulatory effects, and genetic interactions within an integrated framework, thereby enabling a more accurate translation of association signals into biological mechanisms.

In this chapter, we first review the mechanisms of antigen processing and presentation, including the ubiquitin proteasome pathway and the lysosomal pathway. We then summarise current research on the key genetic components involved in these pathways, such as HLA molecules and ERAPs, and their effects on human disease. By integrating findings from recent studies, we highlight how genetic variation in these components influences antigen processing, immune recognition, and disease susceptibility. Finally, we outline the aims and rationale of the thesis, which builds upon this foundation to advance our understanding of the genetic and functional architecture of antigen presentation.

1.2 Overview of the antigen presentation pathway

In human immunology, HLA genes encode cell surface molecules that present both intracellular and extracellular peptides to T cells, playing a pivotal role in pathogen recognition and immune activation. When T cells recognise these peptides as foreign, they initiate specific immune responses to protect the host (Bjorkman et al., 1987). The ability of HLA molecules to bind and display diverse peptides is determined by their structural features, particularly the configuration of the peptide-binding groove. The amino acid composition within this groove dictates binding specificity (Figure 1.1). Owing to their essential role in adaptive immunity and their extraordinary allelic diversity, both within individuals and across vertebrate species, HLA genes have become a central focus of immunological and evolutionary research (Klein and Figueroa, 1986).

Inside the cell, HLA class I and HLA class II molecules present peptides through distinct pathways. HLA class I molecules are primarily involved in the ubiquitin proteasome pathway, whereas HLA class II molecules function mainly within the lysosomal pathway. In the following section, we introduce these two antigen processing and presentation pathways in detail, outlining their molecular mechanisms and immunological significance.

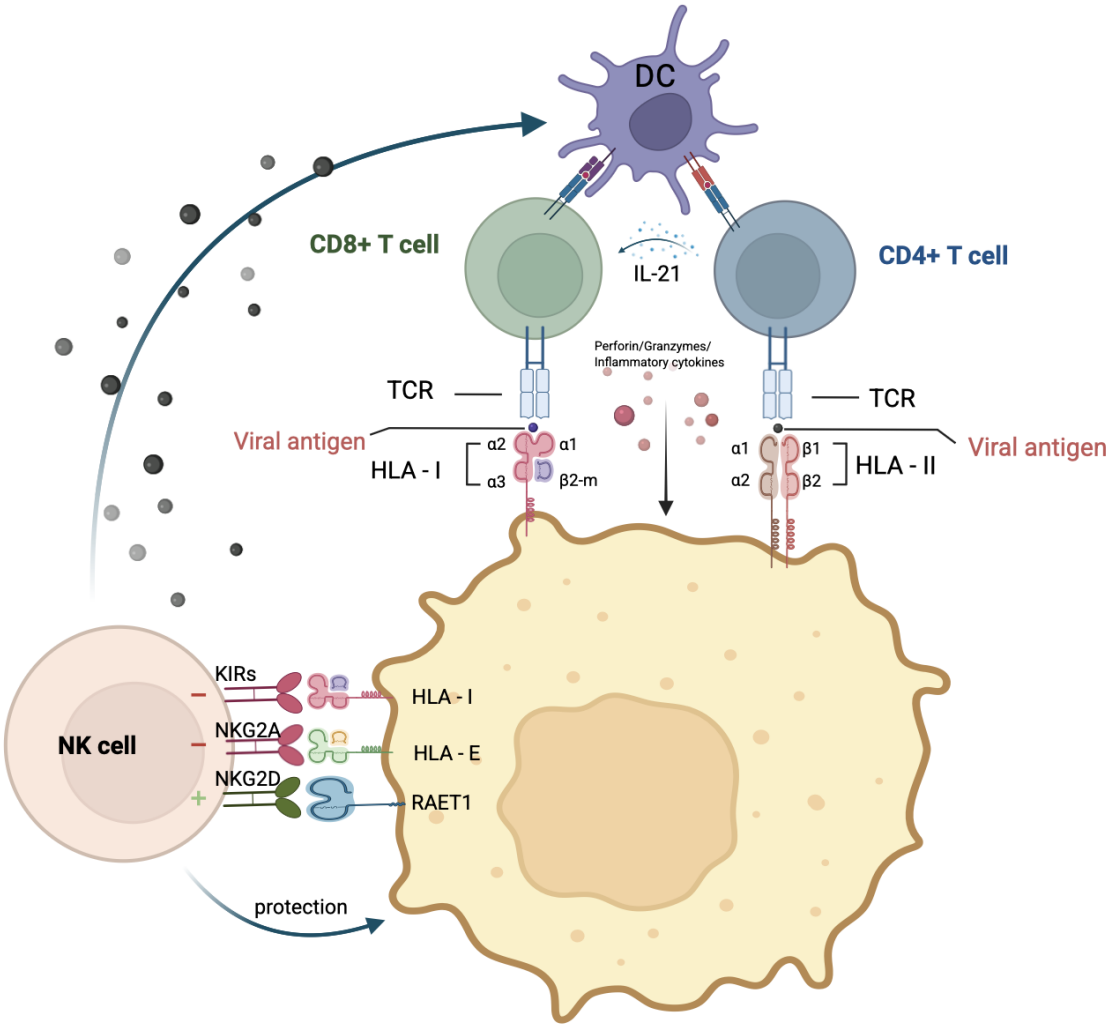


Figure 1.1: **CD4$^+$, CD8$^+$ T cells and NK cells killing of infected cell.** The simplified diagram mainly shows the interaction between cytotoxic CD8$^+$ T cells, CD4$^+$ T cells and virus-infected cells expressing viral antigen on HLA class I molecules and antigen-presenting cells expressing viral antigen on HLA class II molecules. It also highlights the potential relationship based on the IL-21 cytokine. When the T-cell receptor (TCR) recognises the foreign antigen, it triggers TCR signalling, which leads to the release of perforin (PRF) and granzymes (GzmB). During viral immune evasion, NK cells undergo activation, leading to the upregulation of activating receptors such as NKG2D, the production of IFN-γ, and the manifestation of cytotoxic activity. HLA-E is linked to inhibitory signalling through the NK cell receptor NKG2A. These substances induce apoptosis of the virally infected cells, effectively eliminating the infection. Illustration created with *BioRender.com*.

1.2.1 The ubiquitin-proteasome pathway

The process of peptide loading onto HLA class I molecules occurs within a multiprotein assembly known as the peptide-loading complex (PLC), located in the endoplasmic reticulum (ER), as illustrated in Figure 1.2. HLA class I molecules consist of a transmembrane heavy chain that associates non-covalently with β_2-microglobulin. The synthesis of the HLA class I heavy chain begins in the ER, where it is initially stabilised by chaperone proteins such as binding immunoglobulin protein (BiP) and calnexin. As the folding process progresses, BiP dissociates from the heavy chain, allowing β_2-microglobulin to associate and form the HLA class I heterodimer. During this transition, the chaperone calreticulin replaces calnexin to further stabilise the complex. The properly folded HLA class I heterodimer is then incorporated into the PLC, together with additional components including tapasin, ERp57, and the transporter associated with antigen processing (TAP) (Jackson et al., 1990; Ou et al., 1993; Pobre et al., 2019).

Different HLA class I alleles exhibit distinct dependencies on tapasin for efficient peptide loading, and thus the process can be broadly categorised into tapasin-dependent and tapasin-independent pathways. In the following sections, we describe these two mechanisms in detail, outlining their molecular interactions and functional implications for antigen presentation.

1.2.1.1 Tapasin-dependent peptide loading processing

The presentation process of HLA class I molecules is the result of a series of reactions:

1. Within the cellular cytosol, the degradation of naturally occurring proteins is primarily controlled by a specialised cellular mechanism known as the proteasome (Driscoll, 1994).

2. The transporter associated with antigen processing plays a crucial role in facilitating the movement of broken-down protein fragments, or peptides, into the interior of the ER.

3. Within the ER, the precursors are trimmed by ERAP1/ERAP2 into final peptides, which would be loaded with HLA class I. During the process of HLA class I antigen loading, ERAP1/ERAP2 play a role in ensuring the correct assembly of the PLC (Saveanu et al., 2005). ERAP1/ERAP2 could trim peptides to the optimal size for HLA class I binding, but can also over-trim and destroy HLA-I ligands.

4. The trimmed peptides are loaded onto the PLC, which includes calreticulin, ERp57, protein disulfide isomerase (PDI), and tapasin. Tapasin does not directly contact the peptides; instead, its N-terminal region interacts with the β-sheet platform that forms the floor of the HLA class I peptide-binding groove. This interaction allows tapasin to stabilise the adjacent α-helices of the groove, particularly the $\alpha 2 - 1$ helix, while also exerting mechanical force on the $\beta 7 - 8$ strands that support the $\alpha 2$ helix. Consequently, this interaction induces an outward movement of the $\alpha 1$ and $\alpha 2$ helices, effectively widening the peptide-binding groove. These dynamic interactions maintain the HLA class I groove in a peptide-receptive, open conformation, thereby stabilising the molecule and facilitating optimal peptide loading (Santos et al., 2007). Tapasin enhances the stability of HLA class I molecules and promotes the presentation of high-affinity peptides with slower dissociation rates. Increased stability at the cell surface correlates with higher peptide-binding affinity and reduced off-rate.

5. Once assembled with high-affinity peptides, HLA class I complexes are released from the ER and transported through the Golgi apparatus to the plasma membrane, where they present antigenic peptides to $CD8^+$ T cells (Colbert et al., 2020).

1.2.1.2 Tapasin-independent peptide loading processing

HLA class I molecules can also present exogenous antigens through a process known as cross-presentation, as illustrated on the right side of Figure 1.2. This mechanism likely evolved as a response to evolutionary pressures exerted by pathogens. Specifically, during cross-presentation, viral and non-classical HLA class I molecules can interact directly with classical HLA class I molecules to stabilise those that are not loaded with peptides or are not associated with β_2-microglobulin. This interaction results in the formation of open conformers of HLA class I molecules on the cell surface. These open conformers, along with molecules such as HLA-F, can be internalised from the plasma membrane into lysosomal compartments via the endosomal pathway. Within these compartments, antigen-specific peptides are generated through the degradation of internalised proteins and are subsequently reloaded onto HLA class I molecules. This process occurs independently of the traditional TAP and tapasin-mediated pathways. Consequently, even when tapasin function is impaired, certain HLA class I molecules have evolved tapasin-independent mechanisms that enable them to present viral antigens effectively. Such adaptations are crucial

for the clearance and control of pathogen dissemination and for maintaining immune surveillance (Ortmann et al., 1997; Garbi et al., 2003; Bashirova et al., 2020).

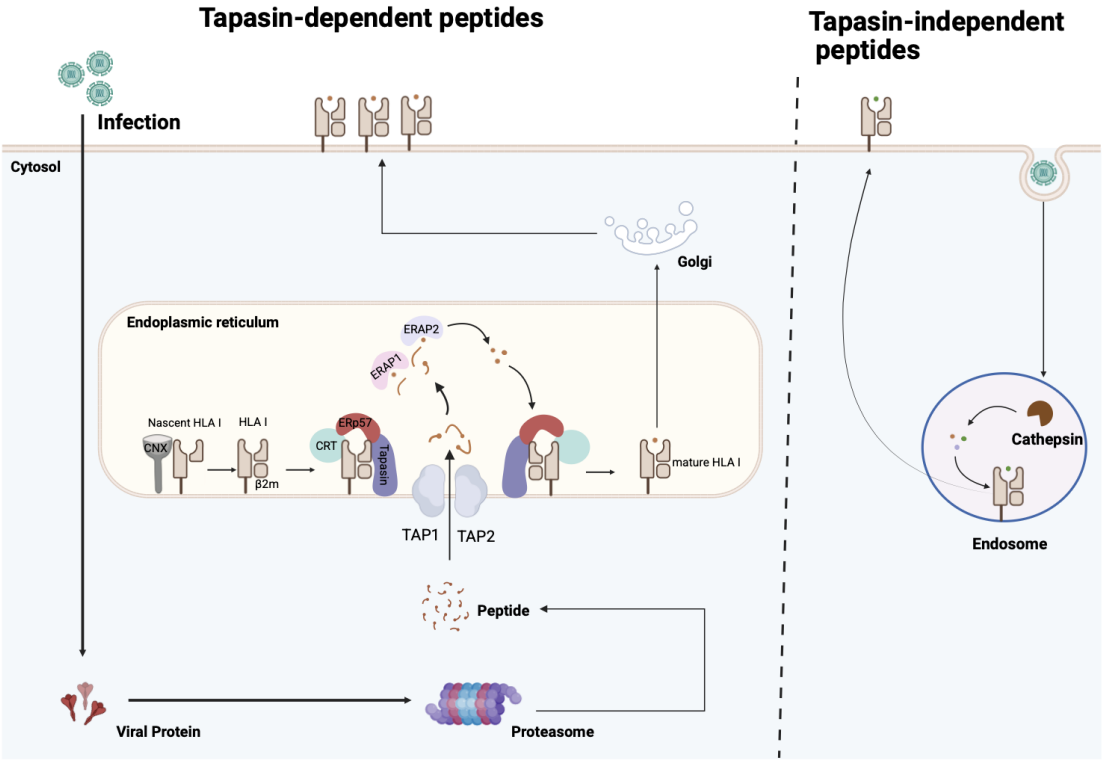


Figure 1.2: **Schematic representation of tapasin-dependent (left) and tapasin-independent (right) pathways of HLA class I peptide loading.** In the classical tapasin-dependent pathway, cytosolic proteins are degraded by the proteasome into peptides, which are transported into the ER by the TAP1/TAP2 heterodimer. Chaperones such as calnexin (CNX) and calreticulin (CRT) assist in folding and stabilisation of empty HLA class I molecules, which are then bridged to TAP by tapasin in complex with ERp57. High-affinity peptides stabilise the complex, which exits the ER for presentation to $CD8^+$ T cells. In the tapasin-independent pathway, certain HLA class I alleles can bind peptides without tapasin, often in lysosomal compartments, and reach the cell surface via alternative trafficking routes. Illustration created with *BioRender.com*.

1.2.2 The lysosomal pathway

HLA class II molecules present peptides in a manner distinct from HLA class I alleles, primarily through the lysosomal pathway. The structure of HLA class II molecules, which are integral to the immune system's antigen presentation process, comprises complex heterodimers formed by two non-covalently associated polypeptide chains: a heavy α chain with a molecular weight of approximately 30 kDa and a lighter β chain

of about 26 kDa. These molecules are predominantly expressed on the surfaces of professional antigen-presenting cells, including B cells, dendritic cells, and macrophages. The α chain contains two domains, α_1 and α_2, each contributing to the molecule's structural stability and function, while the β chain consists of two domains, β_1 and β_2. The β_2 domain interacts specifically with the $CD4^+$ co-receptor on helper T cells, ensuring effective T cell activation. In contrast, the α_1 and β_1 domains together form the peptide-binding groove, which accommodates peptides of 13–20 amino acids in length, forming the foundation for antigen recognition and the initiation of adaptive immune responses (Robinson and Delvig, 2002; Neefjes et al., 2011).

Inside the cell, the antigen is enclosed within an endosome, which subsequently fuses with a lysosome in the cytoplasm, creating endolysosomes [1]. This fusion event leads to the degradation of the foreign protein by lysosomal proteolytic enzymes, resulting in the formation of small peptides. Simultaneously, HLA class II molecules are synthesised within the endoplasmic reticulum, with the α and β chains associating with an invariant chain to prevent self-antigen binding. This complex undergoes intracellular transport from the endoplasmic reticulum to the Golgi apparatus and then to another vesicle, where the invariant chain is processed, leaving behind the class II-associated invariant chain polypeptide (CLIP) fragment. Then, the vesicle containing the HLA class II molecule fuses with a vesicle containing fragmented peptides, allowing the peptides to displace the CLIP and bind with the HLA class II molecule. This newly formed HLA class II peptide complex is subsequently transported to the cell surface, where the antigen is presented to T cells. Recognition of the peptide bound to the HLA class II molecule is facilitated by the T cell receptor, with the $CD4^+$ co-receptor binding to the β_2 domain of the HLA class II molecule, thus initiating the immune response (Neefjes et al., 2011).

[1] Endolysosomes are hybrid organelles that form when late endosomes and lysosomes fuse, acting as a major site for breaking down cellular waste and recycling materials

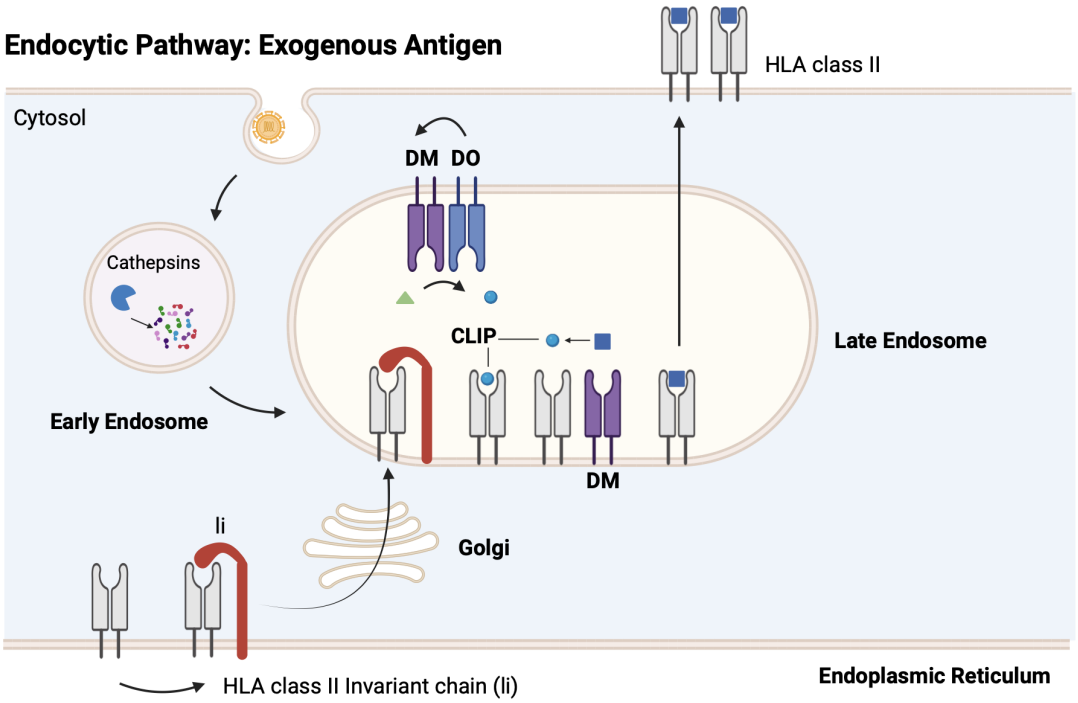


Figure 1.3: **HLA class II maturation and antigenic peptide loading.** HLA class II molecules undergo a series of intricate steps in their biosynthesis and antigen presentation. Initially, these molecules are synthesised within the ER and associated with an invariant chain (Ii). The resulting complex of HLA class II and Ii is subsequently transported through the Golgi apparatus to reach the late endosome. Within the late endosome, specialised proteases process antigens, breaking them down into shorter peptide fragments, while the Ii is also processed to yield a shorter entity known as the CLIP. Then, the interaction of a non-classical HLA-DM molecule with the HLA class II complex facilitates the exchange of CLIP with the antigenic peptide, ensuring that the peptide is presented in an optimal binding register. The resulting complex, consisting of the peptide bound to the HLA class II molecule, is then transported to the surface of the antigen-presenting cell. At the cell surface, this peptide-major histocompatibility complex (pMHC) is made available for recognition by $CD4^+$ T cells, initiating immune responses and adaptive immunity. Illustration created with *BioRender.com*.

1.3 Genetic variation in the antigen presentation pathway and its epidemiological impact

In recent years, genome-wide association studies (GWAS) have substantially advanced our understanding of human biology and disease through rigorous experimental design. The primary objective of GWAS is to investigate associations between genetic variants, particularly common single-nucleotide polymorphisms (SNPs), at specific

genomic loci and disease phenotypes within populations (Visscher et al., 2012). Although most common variants, whether individually or in combination, contribute modestly to disease risk and explain only a portion of heritability, certain conditions, such as age-related macular degeneration, are strongly influenced by variants with relatively large effects. GWAS continue to play a vital role in elucidating the genetic basis of individual variation in disease susceptibility, offering valuable insights that enhance prevention, diagnosis, and treatment (Manolio et al., 2009).

Previous genome-wide association studies have consistently demonstrated that genetic variants within antigen presentation pathways play a major role in determining susceptibility to multiple infectious diseases. In this section, we review key variants affecting proteins central to antigen presentation, including HLA molecules, ERAP1/ERAP2, TAP1/TAP2, proteasome subunits (PSMB8–10), and tapasin (TAPBP). Across these loci, SNPs, distinct allotypes,[2] and single amino acid substitutions have been shown to influence antigen processing and peptide loading. Such molecular alterations can modulate both the efficiency and specificity of antigen presentation, thereby shaping individual immune responses. Consequently, these genetic differences contribute to disease phenotypes and susceptibility patterns, underscoring the pivotal role of antigen presentation in the pathogenesis of complex immune-mediated disorders.

HLA. As discussed in the previous subsection, HLA proteins play a central role in antigen presentation through both the ubiquitin and lysosomal pathways. Building on this foundation, we now examine the genetic variation within HLA genes and its impact on disease susceptibility. The first evidence of such an association was reported in the linkage between HLA-B and Hodgkin lymphoma (Amiel et al., 1967), and since then, HLA has been recognised as one of the most influential immune-related gene families, consistently enriched across a wide range of human diseases (Trowsdale and Knight, 2013).

The HLA region is conventionally divided into three major classes, class I, class II, and class III, each characterised by distinct structural features and immunological functions. The class I region includes the highly polymorphic genes HLA-A, HLA-B, and HLA-C, as well as the less variable genes HLA-E, HLA-F, and HLA-G. Class I molecules are expressed on nearly all nucleated cells and present peptide fragments derived from intracellular proteins to $CD8^+$ cytotoxic T lymphocytes, thereby enabling immune surveillance and the elimination of virus infected or malignant cells. The class

[2]Allotypes refer to variations in a protein, particularly an immunoglobulin, that occur among members of the same species.

II region encompasses a broader array of genes, including HLA-DPA1, HLA-DPB1, HLA-DQA1, HLA-DQA2, HLA-DQB1, HLA-DQB2, HLA-DRA, HLA-DRB1, HLA-DRB2, HLA-DRB3, HLA-DRB4, and HLA-DRB5, along with the less polymorphic accessory genes HLA-DM and HLA-DO. These class II molecules are mainly expressed on professional antigen presenting cells such as dendritic cells, macrophages, and B cells, and they present peptides derived from extracellular proteins to $CD4^+$ helper T cells, orchestrating adaptive immune responses, promoting cytokine production, and modulating inflammation. In contrast, the class III region does not encode classical antigen presenting molecules but contains genes involved in immune regulation, including complement components (C2, C4), cytokines such as TNF[3], and heat shock proteins, contributing to inflammation and host defence rather than direct antigen presentation. A defining feature of the HLA locus is its dense clustering of immune relevant genes exhibiting extreme polymorphism and strong linkage disequilibrium, meaning that alleles at neighbouring loci are often inherited together more frequently than expected by chance. While this extensive diversity provides an evolutionary advantage by broadening pathogen recognition, it also complicates efforts to disentangle the precise alleles responsible for observed disease associations (Dendrou et al., 2018; Klein and Sato, 2000).

The HLA region, located on chromosome `6p21.3` within the major histocompatibility complex, is the most polymorphic region of the human genome (Horton et al., 2004). Within class I and class II loci, genes can be further categorised into classical and non-classical groups based on their level of polymorphism, expression pattern, and functional role. HLA class I genes include HLA-A, HLA-B, and HLA-C, which are highly expressed on almost all nucleated cells and are responsible for presenting endogenous peptides to $CD8^+$ T cells. These genes exhibit extensive allelic diversity, particularly within exons encoding the peptides binding groove, enabling the presentation of a broad spectrum of pathogen peptides (Bjorkman et al., 1987),(Robinson and Delvig, 2002). In contrast, non-classical HLA I genes, such as HLA-E, HLA-F, and HLA-G, display more limited polymorphism and more restricted tissue distribution. Rather than functioning primarily in conventional antigen presentation, they often exert roles in affecting the immune system. For example, HLA-E interacts with inhibitory and activating receptors on NK cells, whereas HLA-G contributes to immune tolerance at the maternal–fetal interface (Carosella et al., 2015).

[3]TNF (Tumor necrosis factor), formerly known as TNF-α, is a chemical messenger produced by the immune system that induces inflammation.

A similar distinction exists within class II genes. HLA class II genes comprise HLA-DR, HLA-DQ, and HLA-DP, each encoding α and β chains that form heterodimeric molecules expressed predominantly on professional antigen-presenting cells. Among these, HLA-DRB1 is particularly polymorphic and has been strongly implicated in susceptibility to autoimmune and infectious diseases (Trowsdale and Knight, 2013). In contrast, non-classical HLA II genes such as HLA-DM and HLA-DO do not primarily present peptides to $CD4^+$ T cells but instead regulate peptide loading within endosomal compartments. HLA-DM facilitates the removal of CLIP and promotes the binding of high-affinity peptides to classical class II molecules, thereby shaping the peptide repertoire available for immune recognition(Neefjes et al., 2011). This functional division between peptides presenting molecules and peptide-editing regulators highlights the layered organisation of the antigen presentation pathway.

The extraordinary polymorphism of classical HLA genes necessitates a standardised and high-resolution allele nomenclature system, which is maintained by the World Health Organisation (WHO) HLA Nomenclature Committee (Robinson et al., 2016). An HLA allele name begins with the gene designation, followed by an asterisk and a series of colon-separated numerical fields. For example, in the allele `HLA-A*01:01:01:01`, "HLA-A" denotes the gene. The first numeric field (`"01"`) defines the allele group, historically corresponding to serological antigen specificity. The second field (`"01"`) specifies differences that alter the amino acid sequence of the encoded protein; thus, alleles differing in the first or second field encode distinct protein sequences. The third field (`"01"`) denotes synonymous nucleotide substitutions within the coding region that do not change the amino acid sequence. The fourth field (`"01"`) identifies differences in non-coding regions, such as introns or untranslated regions. Additional suffixes may be appended to describe expression status, for example, "N" for null alleles with no detectable surface expression, "L" for low expression, or "S" for secreted molecules.

Owing to the pivotal role of HLA in presenting a broad spectrum of antigenic peptides to T cells, classical HLA class I and class II molecules are widely regarded as the primary drivers of disease associations. To date, over 15,000 distinct classical HLA class I and II alleles have been identified (Robinson et al., 2016). Extensive research has explored how these alleles influence susceptibility to and progression of infectious diseases.

In HIV infection, for instance, Philip et al. reviewed multiple cohort studies demonstrating a clear link between HLA class I profiles and the time to AIDS onset. Alleles such as HLA-B*27 and HLA-B*57 exert strong protective effects by delaying

disease progression, a finding consistently corroborated by subsequent studies showing that HLA-B*57:03 and HLA-B*13:02 are associated with slower disease progression and improved viral control. These protective effects are thought to result from more efficient presentation of HIV-derived peptides to cytotoxic T lymphocytes, enhancing the clearance of infected cells. In contrast, alleles including HLA-B*35, HLA-B*07, HLA-A*23, and HLA-A*24 have been associated with accelerated disease progression, underscoring the crucial influence of host genetics on individual susceptibility and clinical outcomes (Goulder and Walker, 2012).

HLA variation also contributes to the immune response in hepatitis infections. In hepatitis C (HCV), alleles such as HLA-DRB1*11 and HLA-DQB1*03:01 confer protection against infection. Notably, HLA-DQB1*03:01 has also been associated with responsiveness to interferon therapy in chronic hepatitis B (HBV) and with variation in vaccine responsiveness (Jiao and Wang, 2003). Moreover, HLA-DR9 displays population-specific effects: it is protective against HCV infection in Japanese individuals but confers susceptibility to chronic HBV in Korean and Chinese populations (Singh et al., 2007). These findings underscore the complex and context-dependent role of HLA polymorphisms in determining hepatitis infection outcomes.

Beyond hepatitis, HLA alleles are also implicated in the immune response to malaria. The allele HLA-B53 has been shown to mediate protection among malaria-immune Africans, where HLA-B53-restricted cytotoxic T lymphocytes recognise a conserved nonamer peptide derived from the liver stage specific antigen 1 (LSA1), although no HLA-B*53 restricted epitopes have been identified in other malaria antigens (Hill et al., 1992). This highlights the specificity of HLA-restricted immune recognition in parasitic infections.

In addition to infectious diseases, HLA variation also shapes humoral immune responses. A recent study identified multiple associations between HLA genetic variants, sociodemographic factors, and disease outcomes, with seroprevalence estimates consistent with previous reports. Notably, the variant rs6927022 within the HLA locus showed a strong association with antibody responses to the Epstein Barr virus (EBV) nuclear antigen EBNA1 (Mentzer et al., 2022).

Collectively, these findings underscore the central role of classical HLA alleles in shaping host immune responses across a diverse array of pathogens. By influencing antigen presentation, T-cell activation, and antibody production, HLA variation fundamentally determines infection outcomes, therapeutic efficacy, and vaccine responsiveness.

ERAPs. Endoplasmic reticulum aminopeptidases (ERAPs), comprising ERAP1 and ERAP2, are key enzymes that shape the HLA class I immunopeptidome[4]. Within the endoplasmic reticulum, they cooperatively trim N-terminal residues from precursor peptides to generate optimally sized antigens that fit into the HLA I binding groove. Alterations or loss of ERAPs function can markedly modify the repertoire of antigens presented by HLA I molecules, thereby influencing the activation of both NK and $CD8^+$ T cells (Woon and Purcell, 2018; Neefjes et al., 2011; Vyas et al., 2008). ERAPs belong to the M1 family of zinc-dependent aminopeptidases and are inducible by IFN-γ and TNF-α, with broad expression across human tissues (Neefjes et al., 2011). Their activities are highly coordinated: ERAP1 preferentially cleaves peptides with large hydrophobic N-terminal residues and efficiently trims 9–16 amino acid peptides into 8–9 amino acid fragments, whereas ERAP2 exhibits a preference for positively charged N-terminal residues, such as arginine and lysine, and is more efficient at processing shorter peptides that are suboptimal substrates for ERAP1 (Saulle et al., 2020).

Human ERAPs are encoded by two genes located on chromosome 5q15, arranged in opposite orientations. The ERAP1 gene spans approximately 47,379 bp and contains 20 exons, whereas ERAP2 extends over about 41,438 bp and comprises 19 exons (Cifaldi et al., 2012). Both genes are highly polymorphic, though ERAP2 exhibits fewer SNPs than ERAP1. Because of their central role in antigen processing, genetic variation in these enzymes has been extensively investigated. Numerous studies have linked ERAP1 and ERAP2 variants to altered enzyme function, contributing to the development of HLA I-associated disorders and influencing susceptibility to infectious diseases (Stamogiannos et al., 2015; Saulle et al., 2020).

ERAP1 displays extensive polymorphism, with several SNPs influencing its enzymatic activity and substrate specificity. Diseases associated variants are mainly found in the catalytic site (residues 346 and 349), the peptide-binding groove (residues 725 and 730), and regions that affect conformational rearrangements (residues 528 and 575). Additional polymorphisms occur in interdomain regions and in domain IV, which is the regulatory domain responsible for C-terminal residue binding. Collectively, these variants modulate ERAP1 peptide-trimming efficiency, thereby shaping the generation of antigenic epitopes and downstream immune responses (Kochan et al., 2011). Among the most studied variants, rs30187 (K528R) reduces peptide-trimming efficiency by impairing conformational transitions between active and inac-

[4]The immunopeptidome is the complete collection of peptides, or short protein fragments, that are presented on the surface of cells by HLA molecules.

tive states, whereas rs27044 (Q730E) alters peptide length preference and trimming specificity (Stamogiannos et al., 2015). Reeves et al. identified 13 distinct ERAP1 haplotypes that can be broadly classified as efficient, hypoactive, or hyperactive, depending on their capacity to generate antigenic epitopes (Reeves et al., 2013, 2014). Another important variant, rs10050860 (D575N), acts synergistically with K528R to further reduce enzymatic activity. A common European allotype—comprising rs2287987 (V349), rs30187 (R528), N575, rs17482078 (Q725), and E730, which produces a largely non-functional ERAP1 protein (Stamogiannos et al., 2015). Moreover, several intronic (rs2248374, rs1748133, rs149481, rs27042, rs149173) and exonic (I276, R127, N392, L848) polymorphisms have been implicated in altered disease susceptibility, particularly to infectious diseases (Yao et al., 2019).

ERAP2 exhibits fewer genetic variants than ERAP1, but several key SNPs modulate its enzymatic activity, substrate specificity, and protein expression. The most studied coding variant, rs2549782, encodes the K392N substitution, with the N392 variant being more efficient than K392 at trimming hydrophobic N-terminal residues due to structural differences in the catalytic and binding sites, resulting in inter-individual variability in antigen processing (Yao et al., 2019). The rs2248374 (A/G) polymorphism, in strong linkage disequilibrium with rs2549782, regulates ERAP2 expression. The G allele produces a transcript with an extended exon 10 containing premature stop codons, which undergoes nonsense-mediated decay (NMD). Despite this, some transcripts from the G allele generate alternative short isoforms (ERAP2/ISO3 and ERAP2/ISO4) during viral infections. These isoforms are shorter versions of ERAP2 produced via alternative splicing and may retain partial enzymatic activity or perform distinct cellular roles, suggesting a potential evolutionary advantage. Two haplotypes are defined by rs2248374: HapA (A allele) and HapB (G allele). HapB homozygotes produce little or no ERAP2 protein, whereas HapA homozygotes exhibit approximately 50% higher protein levels than heterozygotes. The persistence of both haplotypes across populations suggests potential balancing selection, possibly related to immune responses to viral infections (Evnouchidou et al., 2012; Ye et al., 2018). Additional polymorphisms also influence ERAP2 expression: rs10044354 (C/T) reduces ERAP2 levels in homozygous C individuals, while rs75862629 (A/G) in the promoter region alters both ERAP2 and ERAP1 expression, suggesting coordinated regulation of these aminopeptidases. Collectively, these variants contribute to inter-individual differences in peptide processing and immune function (Paladini et al., 2018).

ERAP1 and ERAP2 heterodimer: Previous studies have shown that full-length ERAP2 can form both heterodimers and homodimers with ERAP1, enhancing the efficiency of peptide trimming (López de Castro, 2018). It has been speculated that the expression of the two short, flu-specific ERAP2 isoforms may exert a dominant-negative effect on wild-type ERAP1 or ERAP2. This could lead to altered peptide processing, potentially providing an advantage during infection by presenting a more immunogenic antigen repertoire. Consequently, both haplotypes—and their associated transcripts, which may confer a fitness advantage under diverse environmental conditions, offer a plausible explanation for the maintenance of HapB at intermediate frequencies through balancing selection.

Given the key role of ERAPs in antigen processing and presentation, it is conceivable that these enzymes could serve as potential targets and modulators of infectious disease pathogenicity. Saulle reviewed recent research and findings (Saulle et al., 2020), discussing the role of ERAPs in modulating viral infections, summarised in the table 1.1.

Table 1.1: Selected ERAP1 and ERAP2 SNPs associated with infectious diseases

Gene	**SNP**	**Region / Variation**	**Infectious disease(s)**
ERAP1	rs30187	Exon 11 / K528R	HCV, HIV
	rs27044	Exon 15 / Q730E	HCV, HIV, HPV
	rs10050860	Exon 12 / D575N	HPV
	rs26618	Exon 5 / M276I	HCV
	rs26653	Exon 2 / P127R	HCV, HPV
	rs17481856	Exon 17 / L848L	Toxoplasmosis
	rs17481334	3' UTR / None	HCMV
	rs149481	Intron 17 / None	KD
	rs27042	Intron 16 / None	KD
	rs149173	Intron 18 / None	Toxoplasmosis
	rs17481856	Intron 17 / None	Toxoplasmosis
ERAP2	rs2549782	Exon 7 / K392N	HCV, HIV
	rs2248374	Intron 10 / None	HCV, Influenza, HIV

HLA-class I alleles and ERAPs interaction. The generation of peptides presented by HLA class I molecules is critically dependent on the activity of ERAPs. These enzymes trim precursor peptides within the endoplasmic reticulum to the optimal length required for stable binding to HLA class I molecules on antigen-presenting cells. By modulating this peptide processing step, ERAPs shape the repertoire of antigens available for immune surveillance. Consequently, variations in ERAPs expression

or enzymatic function can influence HLA class I, which is restricted antigen presentation, thereby affecting immune recognition and disease susceptibility. Notably, the ERAP1 SNP rs30187 has been reported to show allele-specific associations with HLA-B*27 and HLA-B*40:01 in ankylosing spondylitis (AS) (Cortes et al., 2015). These findings highlight how ERAPs–HLA class I interplay contributes to disease pathogenesis by altering the peptide repertoire, underscoring the epistatic relationships between antigen-processing enzymes and HLA-associated immune disorders.

TAP. Peptides are transported from the cytosol into the endoplasmic reticulum (ER) by the transporter associated with antigen processing (TAP), which is composed of TAP1 and TAP2. TAPs efficiently translocate peptides of 8–12 residues, while longer peptides are transported with reduced efficiency (Leone et al., 2013). The TAP genes are located in the HLA class II region, between the DOB and DMB loci (TAP1: chr6:32,845,209–32,853,704 TAP2: chr6:32,825,415–32,838,739). Polymorphisms within the TAP coding regions can influence the formation and stability of TAP1–TAP2 heterodimer complexes, potentially affecting the binding of antigenic peptides to HLA molecules. Recent studies have linked TAP1 and TAP2 variants to multiple disease risks. Genetic variations in TAP2 have been reported to influence immune responses to childhood vaccinations (Roby et al., 1996). Liu et al. (Liu et al., 2021) explored six SNPs in TAP1 and TAP2 in the Han Chinese population, identifying TAP2 SNPs (rs2228396 and rs241441) as potentially contributing to non-small-cell lung cancer (NSCLC) pathogenesis. Specific TAP1 alleles at amino acid positions 333 and 637, and TAP2 alleles at positions 379, 565, and 665, have been associated with the severity of systemic lupus erythematosus (SLE) (Correa Vanegas et al., 2003). Human papillomavirus (HPV) can down-regulate TAP1 expression, leading to reduced HLA class I surface levels; carriers of TAP1 I333V and D637G polymorphisms have been shown to have a lower risk of CIN III (Einstein et al., 2009). Additionally, TAP2 rs4148873 has been suggested as a potential biomarker for increased cervical cancer risk based on meta-analyses (Meng et al., 2018), and the TAP1 promoter SNP rs2071480 is associated with susceptibility to alopecia areata in the Korean population (Kim et al., 2015).

Proteasome. During the ubiquitin-proteasome system, proteasomes play an important role by mediating the activation, conjugation, and ligation of the small ubiquitin protein (a 76 amino acid polypeptide that serves as a molecular tag marking proteins) to substrate proteins in the cytoplasm through a cascade of E1, E2, and E3 enzymes (Pohl and Dikic, 2019; Seifert et al., 2010). Then, the proteasomes degrade the ubiquitin protein into smaller peptides, which will be transported to the ER via

the TAP transporter (Blum et al., 2013). Under conditions of intensified immune response, three beta-subunits (β1, β2, and β5) in the 20S proteasome may be replaced by functional counterparts LMP2 (*PSMB9*), LMP7 (*PSMB8*), and LMP10 (*PSMB10*), respectively. The resulting proteasome is called an immunoproteasome (i-proteasome), which increases the peptide supply for antigen presentation (Griffin et al., 1998). Specifically, these three subunits are induced in most cells by stimulation with type I (IFN-α and IFN-β) and type II (IFN-γ) interferons, which lead to the transcription of distinct proteasome subunits with altered catalytic characteristics. The cascade of peptides generated by i-proteasomes correlates with antigen presentation mediated by an increased peptide supply, which is essential for activating the immune system (Griffin et al., 1998).

In recent years, several studies indicated that polymorphisms in the *PSMB8* and *PSMB9* genes, which are encoded by the HLA class II region, may affect immune functions and thus be associated with the development of epidemic diseases. In 2013, Qian reported two SNPs in *PSMB9*: *rs1351383* and *rs2127675*, which may be associated with melanoma susceptibility (Qian et al., 2013). Moreover, some studies demonstrated that *rs2071543* in *PSMB8* and *rs17587* in *PSMB9* are related to HPV, HCV, and cervical cancer, respectively (Cao et al., 2005; Huang et al., 2014; Omran et al., 2013; Qian et al., 2013). Furthermore, Nasser concluded that the *PSMB9 rs17587* SNP may contribute to the risk of urothelial bladder carcinoma (UBC), whereas the *PSMB8 rs2071543* SNP showed no significant relation (Elhawary et al., 2023).

Tapasin. Tapasin is a critical component of the PLC. It bridges HLA class I molecules to TAP, acts as a chaperone that stabilises HLA-I in a peptide-receptive conformation, and facilitates the exchange of suboptimal peptides for higher-affinity ones. Downregulation or loss of tapasin function has been reported in several viral infections and cancers as a mechanism of immune evasion through disruption of the HLA-I peptides loading process. For example, loss of tapasin expression has been observed in human lung and colon cancer cells, enabling escape from tumour-associated antigen-specific cytotoxic T lymphocyte recognition (Shionoya et al., 2017). Similarly, tapasin and HLA-I dysregulation were found to correlate with survival outcomes in glioblastoma multiforme (Thuring et al., 2014). In addition to its role in antigen processing, genetic variation in the *TAPBP* gene, which encodes tapasin, has been shown to influence disease susceptibility. Carrington's group identified two single-nucleotide polymorphisms (SNPs) that regulate *TAPBP* mRNA expression in African populations: *rs111686073* (G/C), located within an AP-2α transcription factor binding site,

and *rs59097151* (A/G), located within a microRNA-44386 binding site. These SNPs were associated with reduced *TAPBP* expression, lower prevalence of *Plasmodium falciparum* parasitemia, and decreased incidence of clinical malaria (Walker-Sperling et al., 2022).

Summary of genetic variation across the antigen presentation pathway

Based on the preceding review and previous GWAS, we summarise here the key genes within the antigen presentation pathway, the principal genetic variants or allotypic forms considered in this thesis, and their reported associations with human disease.

The present work focuses primarily on proteins that directly shape the peptide repertoire available for presentation and on molecular interactions that determine peptide selection and stability, as the Table below. These include the **proteasome** (e.g., PSMB8, PSMB9), which initiates cytosolic peptide generation; the transporter associated with antigen processing, **TAP1** and **TAP2**, which regulate peptide translocation into the endoplasmic reticulum (ER); and **ERAP1** and **ERAP2**, which trim peptides to optimal lengths for HLA class I binding. The pathway also includes **tapasin** (TAPBP), which mediates peptide editing and stabilises peptide-receptive HLA class I molecules, as well as classical and non-classical **HLA class I** and **HLA class II alleles**, which ultimately present peptides to $CD8^+$ and $CD4^+$ T cells.

Gene	Protein function	Related variants	Reported GWAS associations
HLA-A, -B, -C	Peptide presentation of intracellularly derived antigens to $CD8^+$ T cells	Classical alleles (4-digit); amino acid polymorphisms; HLA from functional perspective e.g., supertypes (see 2.3 for details)	HIV viral control(Study, 2010), HCV spontaneous clearance and acute infection outcome(Duggal et al., 2013), malaria susceptibility and severity(Hill et al., 1991), and multiple autoimmune diseases including type 1 diabetes(Nejentsev et al., 2007) and ankylosing spondylitis (AS)(d'Etude Génétique des Spondylarthrites , GFEGS).
HLA-DRB1, HLA-DQB1, HLA-DPB1, HLA-DQA1	Peptide presentation of extracellularly derived antigens to $CD4^+$ T cells	Classical alleles, amino acid polymorphisms	Multiple autoimmune diseases including type 1 diabetes(Nejentsev et al., 2007), (Onengut-Gumuscu et al., 2015)), rheumatoid arthritis(Raychaudhuri et al., 2012), multiple sclerosis(int, 2011), celiac disease (Dubois et al., 2010) and systemic lupus erythematosus (Shai et al., 1999). HCV clearance and progression (Thomas et al., 2009); chronic HBV persistence(Kamatani et al., 2009); tuberculosis susceptibility (Thye et al., 2010).
ERAP1	Peptides trimming	Functional coding SNPs rs30187, rs27044, rs10050860 , rs26618, rs26653, rs17481856, intronic / regulatory SNP (rs17481334, rs149481, rs27042, rs149173) , haplotypes/allotypes	Ankylosing spondylitis(Oppermann et al., 2011); Psoriasis(Das et al., 2017); Behçet's disease(Mahmoudi et al., 2022); Type 1 diabetes(López de Castro, 2018); Inflammatory bowel disease(Castro-Santos et al., 2017); Cervical cancer(Mehta et al., 2015).
ERAP2	Peptides trimming	rs2248374, rs2549794, rs2549782, rs10044354, haplotypes, isoform	Inflammatory bowel disease(Hamilton et al., 2023); PreeclampsiaFerreira et al. (2021); Ankylosing spondylitis(Robinson et al., 2015); Severe respiratory infection(Hamilton et al., 2023); Black death(Klunk et al., 2022).
TAP1	Peptide transport from cytosol to ER	rs1057141, rs1135216, rs41551515; haplotypes	Tuberculosis(Zhang et al., 2021), Respiratory diseases(Kim et al., 2011b).
TAP2	Peptide transport from cytosol to ER	rs241447, rs2228396, rs1800454, rs67511411, rs141555015; haplotypes	Tuberculosis(Zhang et al., 2021).
TAPBP	Tapasin; peptide editing and HLA stabilisation	rs1059288, rs2071888, rs111686073	Malaria susceptibility(Walker-Sperling et al., 2022), Cervical cancer(Hu et al., 2024).
PSMB8, PSMB9	Immunoproteasome subunits; Initial peptide cleavage	rs9357155, rs17587	Parkinson(Nguyen et al., 2025).

Table 1.2: Summary of key antigen presentation pathway gene findings from GWAS studies

Several additional genes contribute to antigen presentation and immune regulation, but were not included in the primary modelling framework of this thesis. These include structural components such as *B2M*, which encodes β_2-microglobulin and is essential for HLA class I surface expression; chaperones including *CALR* (calreticulin) and *CANX* (calnexin), which assist in HLA folding; and transcriptional regulators such as *NLRC5* and *CIITA*, which control HLA class I and class II gene expression, respectively. Historically, large-scale GWAS have identified relatively few consistent and independent association signals at these loci across the diseases considered in this thesis. One possible explanation is that variation in these genes may not directly alter peptide sequence generation, trimming, transport, or editing, and therefore may exert more indirect or context-dependent effects on antigen presentation. Consequently, they were not incorporated into the primary modelling framework, which was designed to capture genetic variation that directly influences peptide repertoire composition. Nevertheless, for completeness, Chapter 2 includes an analysis of missense variation in these genes to evaluate their potential contribution within the broader antigen presentation pathway.

Future work could extend the framework developed here to incorporate transcriptional regulation and NK cell–mediated immune interactions. Genetic variation across these components may influence peptide processing efficiency, peptide length distribution, binding affinity, surface stability of peptide–HLA complexes, and ultimately T cell recognition. Previous GWAS have consistently implicated multiple APP genes in susceptibility to infectious, autoimmune, and inflammatory diseases. However, most studies have evaluated these loci individually. In contrast, this DPhil research models APP components jointly within a unified Bayesian framework to account for linkage disequilibrium, correlated effects, and potential epistatic interactions across the pathway.

1.4 Aim and rationale

In this thesis, we review the biological foundations of the antigen presentation pathway, with particular emphasis on ubiquitin-mediated and lysosomal processes. We examine the biological roles of key molecular components, including HLA, ERAPs, TAPBP, and other central elements involved in antigen processing and presentation. Previous GWAS have identified strong links between these genes and a wide range of immune-related phenotypes. However, most existing research has focused on individual factors such as specific HLA alleles or ERAPs variants, rather than on the coor-

dinated interactions among different components of the pathway. These interactions are further complicated by LD within regions such as the HLA and ERAPs loci, by regulatory effects through which one variant may influence the expression of another, and by non-additive genetic interactions across loci. As a result, the combined contribution of antigen presentation genes to disease risk cannot be adequately captured by models that consider only additive genetic effects. A more integrated framework is therefore needed to understand how variation across these interconnected pathways shapes immune-related phenotypes.

To address these challenges, this thesis has three primary aims:

1. **Feature extraction and characterisation :** Since no existing tools or pipelines can fully capture the complexity of genetic variation within the antigen presentation pathway, in Chapter 2 we develop a dedicated computational pipeline. This pipeline extracts relevant genetic features, summarises their distributions, and compares them across multiple cohorts and populations. This provides the foundation for subsequent statistical analyses.

2. **Novel statistical modelling:** Recognising the limitations of traditional GWAS models in accounting for linkage disequilibrium, gene–gene interactions, and non-additive effects, Chapter 3 presents a new Bayesian inference framework based on the regularised horseshoe prior. The model is designed to identify sparse but potentially strong genetic signals within the antigen presentation pathway and to enable flexible inference on both additive and non-additive relationships. It also explicitly considers interaction terms between key components such as HLA and ERAPs. Extensive simulation studies demonstrate the robustness, accuracy, and practical applicability of this approach.

3. **Application to phenotype association studies:** In Chapter 4, we combine the computational pipeline from Chapter 2 with the statistical model from Chapter 3 to perform association analyses between genetic variants in antigen presentation genes and clinical phenotypes. This integrated approach enables a systematic evaluation of the genetic architecture underlying the antigen presentation pathway and provides deeper insights into how genetic variation influences immune-related phenotypic outcomes.

Overall, building on these challenges and motivating observations, this thesis aims to bridge the gap between biological understanding and statistical methodology in the study of the antigen presentation pathway. By developing new computational and

statistical approaches and applying them to well-characterised datasets, this work aims to identify both established and previously unrecognised genetic determinants of antigen processing and their associations with phenotypic outcomes. The resulting integrative framework enhances our understanding of the genetic basis of antigen presentation and provides methodological advances that can be applied to other complex biological pathways.

Chapter 2

Comprehensive Analysis of Antigen Processing Pathway Genetic Variation Using a Scalable Bioinformatics Pipeline

2.1 Introduction

As outlined in Chapter 1, infectious diseases continue to impose a significant global health burden. A critical determinant of disease outcomes is the variation in individual immune responses, which largely depends on the effectiveness of the adaptive immune system, particularly the functions of $CD4^+$ and $CD8^+$ T lymphocytes. These functions rely on efficient antigen presentation, a tightly regulated process that enables the immune system to detect and respond to both intracellular and extracellular pathogens by displaying antigenic peptides on HLA molecules for T-cell recognition. Antigen presentation occurs primarily through two pathways: HLA class I and HLA class II. Each pathway involves a complex network of genes and molecular mechanisms that coordinate the processing and display of pathogen-derived peptides. Although numerous studies have investigated specific components of this system, such as individual HLA alleles or the TAP transporters, comprehensive pathway-level genetic analyses remain limited. Furthermore, most existing research has focused on isolated variants or single ancestry groups, which restricts the generalisability of findings across global populations.

In this chapter, we present a comprehensive bioinformatics framework for characterising genetic variation across the antigen processing and presentation pathway using imputed genotype array data. This work goes beyond the development of a scalable and reproducible pipeline. First, we implement dedicated software tools that

integrate variant calling, annotation, and allotype-level resolution tailored to highly polymorphic loci, such as HLA and ERAPs. Second, we perform systematic analyses of genetic variation within the pathway, focusing on both common and rare alleles, coding and regulatory variants, and their predicted functional effects. Third, we investigate the frequency and distribution of these variants across multiple population cohorts, enabling comparative insights into how immunogenetic diversity differs between ancestries and may influence disease susceptibility or therapeutic response.

Together, these contributions provide a foundation for cross-population immunogenetic research. The pipeline ensures reproducibility and scalability across large datasets, while the integrated analyses deliver the high quality, biologically meaningful outputs that can support both mechanistic studies and translational applications. Importantly, this framework enables the systematic evaluation of antigen presentation pathway variation at scale, paving the way for downstream functional experiments, risk stratification approaches, and personalised immunology.

2.2 Overview of datasets used in the analysis

In this study, we utilise four datasets to evaluate the performance and applicability of our bioinformatics pipeline across a range of infectious disease contexts. These datasets include:

1. A cohort of individuals with HCV acute infection, including participants who spontaneously cleared the virus and those who progressed to chronic infection (Jones et al., 2022).

2. A chronic HCV infection cohort from the STOP-HCV consortium

3. A severe malaria case-control dataset from the Malaria Genomic Epidemiology Network (MalariaGEN), using a subset of 17,056 identified severe malaria cases and controls across 11 countries.

4. A serological panel from the UK Biobank assessing antibody responses to multiple infectious agents

For each dataset, we provide details of the sample size and of the key phenotypic traits that are relevant to the immune response and infection status. We also provide information on the genotyping or whole-genome sequencing platforms that were used. Where applicable, we also describe the imputation and phasing strategies employed to harmonise the genetic data and ensure high-quality variant calling across cohorts.

2.2.1 Hepatitis C

HCV infection presents a major health burden, with more than 50 million people being infected worldwide, which can lead to liver failure and hepatocellular cancer in infected individuals (Mohd Hanafiah et al., 2013). During acute HCV infection, a subset of patients will spontaneously clear the virus, characterised by a continuous decline in viral load in the blood until HCV RNA becomes negative. Spontaneous clearance is influenced by demographic factors (sex, ancestry) and host genetics, most notably variation within the major histocompatibility complex. A recent trans-ancestral fine-mapping study of the HLA identified two HLA class II alleles (DQB1*03:01 and DRB1*01:01) and specific amino-acid residues in HLA-DQβ1 that are strongly associated with spontaneous HCV clearance (Jones et al., 2022). However, their analysis focused exclusively on classical HLA-II variation, without exploring other genes in the antigen-presentation pathway. We therefore obtained the Valencia et al. dataset to extend the investigation to the broader antigen-presentation machinery (Jones et al., 2022).

HCV chronic infection remains a major global health concern, leading to liver-related morbidity and mortality. We accessed and leveraged the dataset from STOP-HCV, which contained a cohort of individuals with chronic HCV infection, and the disease phenotypes included viral load, progression to liver cirrhosis, liver transplant, and Hepatocellular carcinoma (HCC).

In this study, we present two datasets related to HCV: one from the acute infection phase (Jones et al., 2022) and another from the chronic infection phase (the STOP-HCV cohort), enabling a comprehensive analysis of HCV infection across disease stages.

2.2.1.1 HCV spontaneous clearance vs chronic infection dataset

Samples The HCV spontaneous clearance vs chronic infection dataset included 3,469 individuals who actively participated in the extended HCV genetic consortium conducted in both Europe and Africa. For our analysis, we selected 3,434 individuals with clearly annotated genetic and case-control trait data, specifically indicating HCV persistence or clearance. The full sample comprises 3,434 individuals of both European and African ancestry, of whom 837 showed HCV clearance, and 1,228 showed HCV persistence. Additionally, the study encompasses 545 individuals of African ancestry, among whom 359 achieved HCV clearance, while 186 experienced HCV

persistence. Detailed cohort characteristics, including distribution by ancestry, sex, and trait status, are provided in the Table 2.1.

Table 2.1: **Characteristics of the HCV spontaneous clearance vs chronic infection dataset**

Genetically determined ancestry group	N	HCV infection persistence: clearance	Male (%)
African	545	359:186	48.6
European	2065	1228:837	68.2
Total	3434	1759:1166	59.9

Genotyping, imputation of SNPs and HLA alleles The initial dataset consisted of genotypic data obtained using the Illumina Omni1Quad BeadChip array and was processed using standard quality control protocols as required by GWAS procedures. To account for population structure and relatedness, we performed principal component analysis using the PC-AiR method implemented in the GENESIS R package.

Following quality control, SNP genotype imputation was performed using the Michigan Imputation Server, with the Haplotype Reference Consortium (HRC) panel as the reference. Imputation was carried out using Minimac4. The SNP imputation for this cohort was conducted by Dr Jocelyn Quistrebert.

After SNP imputation, we performed HLA imputation to obtain classical HLA alleles and amino acid polymorphisms within the HLA region. Unlike the original study, we applied SNP2HLA using a reference panel derived from whole-exome sequencing data of UK Biobank participants, restricted to the extended HLA region (chr6: 25–34 Mb, GRCh37). To complement imputed data, HLA alleles were also directly inferred from whole-exome sequencing reads using the HLA-HD algorithm, which aligns sequences to a curated HLA reference database for high-resolution typing. The HLA imputation and allele calling for this cohort were conducted by Dr Guillaume Butler Laporte (Butler-Laporte et al., 2023).

2.2.1.2 HCV chronic infection dataset

Samples This dataset included 3768 participants with both clinical data and human genotyping data. Participants were categorised into four self-reported ancestry groups: European, South Asian, East Asian, and African. For our study, we focused

on individuals of European (2,997) and South Asian (526) ancestry. The distribution of clinical outcomes across these populations is summarised in the Table2.3.

Table 2.3: **Characteristics of the STOP-HCV dataset**

Characteristic	**Total**	**European**	**South Asian**
Number	3768	2997	526
Age (mean ± SD)	51.1 ± 10.7	51.5 ± 10.2	47.7 ± 12.4
Gender (Male %)	70.9%	72.3%	62.8%
Virus genotype			
GT1	1038	915	31
GT3	2462	1885	478
Cirrhosis (%)	58.5%	58.7%	50.4%
HCC (%)	15.1%	16.2%	8.8%
Cirrhotic (%)	24.8%	26.4%	16.3%
Non-cirrhotic HCC (%)	1.5%	1.7%	1.2%
Virus load ($\log_{10}$)			
Total	6.01 ± 0.84	6.02 ± 0.84	5.96 ± 0.85
Cirrhotic	5.96 ± 0.85	5.98 ± 0.85	5.90 ± 0.90
Non-Cirrhotic	6.08 ± 0.81	6.09 ± 0.82	6.02 ± 0.80

HCV exhibits substantial genetic diversity, classified into seven major genotypes and numerous subtypes that differ by up to 30% at the nucleotide level (Mohd Hanafiah et al., 2013). This diversity influences disease progression, treatment response, and vaccine development, making genotype characterisation essential in clinical and research settings. To account for global differences in HCV subtype prevalence, we categorised viral genotypes of the STOP-HCV datasets into major groups based on clinical records: GT1 and GT3. Notably, genotype distribution differed markedly by region. In the South Asian cohort, GT3 was predominant (478 individuals) compared to only 31 with GT1. In contrast, the European cohort had a more balanced distribution, with GT1 (915) and GT3 (1,885) both being common. In terms of clinical outcomes, where cirrhosis is a key case-control trait which is not the prevalence in the population, it was present in approximately half of the individuals in both regions: 58.7% in Europeans and 50.4% in South Asians. We stratified virus load by cirrhosis status. Across the entire cohort, non-cirrhotic individuals had a higher average viral load (6.08 ± 0.81) than cirrhotic individuals (5.96 ± 0.85). This pattern was consistent in both European and South Asian subgroups. This finding may reflect the interplay between advanced liver damage in cirrhosis and impaired viral replication,

as previously reported, potentially due to altered immune responses or hepatocyte availability. We also observed a strong association between cirrhosis and HCC. The overall HCC prevalence was 15.1%, but among cirrhotic individuals, this rose sharply to 24.8%, compared to only 1.5% in non-cirrhotic individuals. This trend held within both the European and South Asian cohorts, highlighting cirrhosis as a primary risk factor for the development of HCC in chronic HCV infection.

Genotyping, imputation of SNPs and HLA alleles SNP imputation was conducted via the TOPMed Imputation Server, with the TOPMed reference panel and the Minimac4 algorithm. For HLA allele imputation, we employed the Michigan Imputation Server, selecting the Multi-ethnic HLA reference panel v2 (four-digit resolution). Phasing was conducted using Eagle v2.4, and the analysis was aligned to the hg19 genome build.

2.2.2 Malaria

Malaria remains one of the most significant infectious diseases worldwide, with over 200 million cases and more than 600,000 deaths reported annually, predominantly in sub-Saharan Africa (Tuteja, 2007). The disease is caused by Plasmodium parasites transmitted through the bites of infected Anopheles mosquitoes. Despite substantial progress in prevention and treatment, malaria continues to pose a major public health challenge, particularly among young children and pregnant women. Understanding the genetic basis of host resistance to malaria has been crucial for developing new strategies for disease control and for improving clinical outcomes.

2.2.2.1 Malaria Genomic Epidemiology Network

We accessed and utilised genotype data from the Malaria Genomic Epidemiology Network (MalariaGEN), a collaborative research initiative established in 2005 to explore the genetic basis of resistance to malaria. MalariaGEN integrates data from multiple international sites using harmonised protocols for phenotyping and genotyping, with a focus on equitable data sharing and building local research capacity (Band et al., 2019).

Samples For this dataset, we analysed data from the MalariaGEN severe malaria (SM) case-control cohort, which includes 17,056 individuals from 11 countries across Africa, Asia, and Oceania. The genomes were genotyped and phased at over 1.5 million SNPs, enabling GWAS of host genetic factors contributing to severe or mild

malaria. Table 4.14 summarises the number of severe malaria cases and population cases and controls included in the dataset after applying quality control procedures.

Table 2.5: **Malaria Genomic Epidemiology Network Study Samples**

Group	Cases	Controls	Total
Africa			
Gambia	2567	2605	5172
Mali	274	183	457
Burkina Faso	733	596	1329
Ghana	399	320	719
Nigeria	113	22	135
Cameroon	592	685	1277
Malawi	1182	1317	2499
Tanzania	416	403	819
Kenya	1681	1615	3296
Asia			
Vietnam	718	546	1264
Oceania			
PNG	402	374	776

Genotyping, imputation of SNPs and HLA alleles Genotype imputation was performed using the TOPMed Imputation Server, employing the TOPMed reference panel and the Minimac4 algorithm. For the imputation of HLA alleles, we used the Michigan Imputation Server with the Multi-ethnic HLA reference panel v2 (four-digit resolution).

2.2.3 UK Biobank serological panel

Antibodies serve as critical markers of prior exposure or chronic carriage of infectious agents, particularly those capable of latent or persistent infection. Previous studies have demonstrated that antibody responses align with expected seroprevalence patterns and reveal meaningful associations with host factors, including HLA genetic variants central to antigen presentation (e.g., HLA genetic variants: rs6927022 with Epstein-Barr virus EBNA1 antibodies), sociodemographic characteristics (e.g., number of lifetime sexual partners with Chlamydia trachomatis), and disease outcomes (e.g., HPV-16 seropositivity with cervical intraepithelial neoplasia, and EBV responses with multiple sclerosis) (Mentzer et al., 2022). We applied and leveraged

the UK Biobank serological panel platform to investigate the association between genetic variation in the antigen presentation pathway and relevant phenotypes.

Serology and samples

Infectious microbes are well-established causal agents in the development of several non-communicable diseases and are strongly suspected to contribute to many others (Mentzer et al., 2022). To investigate these associations, we used serological data from 9,427 randomly selected participants in the UK Biobank (UKB), accessed through the UK Biobank Research Analysis Platform. Serological measurements were conducted using Multiplex Serology at the German Cancer Research Centre (DKFZ) in Heidelberg, Germany, with all samples passing stringent quality control procedures. We focused on UKB Data-Field 23050, which provides antibody response data against 45 antigens from 20 viral, bacterial, and protozoan pathogens in Tables 2.7 and 2.9. This dataset includes both binary seropositivity calls and continuous median fluorescence intensity (MFI) values, enabling assessment of seroprevalence as well as dose-response relationships. In the UK Biobank multiplex serology assay, each participant's serum was tested on many replicate beads coated with the same antigen, each producing its own fluorescence reading. The median fluorescence intensity (MFI) is calculated as the middle value of these bead-level signals, providing a robust measure of antibody binding that is less affected by outliers. This single MFI value is reported for each antigen per individual. Besides, the panel covers a broad range of pathogens. Viral agents include herpes simplex virus types 1 and 2 (HSV-1 and HSV-2), varicella-zoster virus (VZV), Epstein-Barr virus (EBV), human cytomegalovirus (CMV), human herpesvirus 6 and 7 (HHV-6 and HHV-7), and Kaposi's sarcoma-associated herpesvirus (KSHV), along with hepatitis B and C viruses (HBV and HCV), human T-lymphotropic virus 1 (HTLV-1), human immunodeficiency virus 1 (HIV-1), BK and JC polyomaviruses (BKV and JCV), Merkel cell polyomavirus (MCV), and human papillomavirus types 16 and 18 (HPV-16 and HPV-18). In addition, bacterial and protozoan pathogens such as Chlamydia trachomatis, Helicobacter pylori, and Toxoplasma gondii are included.

Table 2.7: **Sero-prevalence Estimates for Infectious Agents in the Pilot Study**

Pathogen	**Seroprevalence (%)**
HSV-1	70
HSV-2	16
VZV	92
EBV	95
HHV-6 overall	91
HHV-6A	77
HHV-6B	79
HHV-7	95
KSHV	8.1
HBV	2.5
HCV	0.3
CMV	58.3
HIV	0.2
HTLV-1	1.6
HPV-16-I	4.4
HPV-16-II	4.6
HPV-18	2.7
BKV	95
JCV	57
MCV	67
C. trachomatis-I	21
C. trachomatis-II	2.5
H. pylori-I	18
H. pylori-II	32
T. gondii	28

Table 2.9: **Distributions of Serum Antibody Responses in UKB Before Normalisation**

Antigen	No. samples	Mean	SD	Median	Max	Min
HSV-1 (1gG)	9427	3193.306	2998.024	2981	15718	1
HSV-2 (2mgG)	9427	275.337	875.255	32	10630	1
VZV (gE/gI)	9427	980.259	1194.004	523	11635	1
EBV (VCA p18)	9427	7283.601	3415.880	7451	19010	1
EBV (EBNA-1 pep)	9427	4416.889	3109.799	4323	16969	1
EBV (ZEBRA)	9427	2349.871	1963.772	2000	13637	1
EBV (EA-D)	9427	2553.487	2588.141	1733	15853	1
CMV (pp150 Nter)	9427	1608.068	2015.932	602	13233	1
CMV (pp52)	9427	3183.264	3417.676	2030	14979	1
CMV (pp28)	9427	1347.649	1659.226	497	11120	1
HHV-7 (U14)	9427	848.189	806.655	599	10321	1
KSHV (LANA)	9427	93.432	593.821	16	12897	1
KSHV (K8.1)	9427	73.024	122.244	57	4375	1
HHV-6 (IE1B)	9427	521.504	886.696	233	12488	1
HHV-6 (IE1A)	9427	313.019	373.506	200	5678	1
HHV-6 (p101 k)	9427	149.606	463.264	21	11715	1
HBV (HBc)	9427	70.150	602.988	5	14302	1
HBV (HBe)	9427	97.112	575.780	13	13906	1
HCV (Core)	9427	31.586	318.547	3	12542	1
HCV (NS3)	9427	53.570	352.694	31	13384	1
T. gondii (p22)	9427	74.757	195.333	31	7153	1
T. gondii (sag1)	9427	128.141	153.253	89	5788	1
HTLV-1 (gag)	9427	292.142	355.512	134	2717	1
HTLV-1 (env)	9427	33.767	40.332	27	2203	1
HIV (gag)	9427	127.338	365.539	55	10819	1
HIV (env)	9427	47.088	70.572	39	3972	1
BKV (VP1)	9427	3754.102	2702.052	3419	14678	1
JCV (VP1)	9427	795.168	1165.326	315	10975	1
MCV (VP1)	9427	2210.297	2448.300	1002	10506	1
HPV16 (L1)	9427	68.016	172.519	40	4983	1
HPV16 (E6)	9427	35.633	300.897	11	11781	1
HPV16 (E7)	9427	40.515	187.527	18	9305	1
HPV18 (L1)	9427	56.926	145.125	38	5876	1
C. trachomatis (mompD)	9427	139.760	480.431	16	8022	1
C. trachomatis (mompA)	9427	80.918	303.544	19	7870	1
C. trachomatis (tarp F2)	9427	199.727	568.935	25	9199	1
C. trachomatis (tarp F1)	9427	188.078	631.489	12	9685	1
C. trachomatis (PorB)	9427	23.310	57.673	13	2599	1
C. trachomatis (pGP3)	9427	611.260	1615.517	9	13749	1
H. pylori (CagA)	4754	989.898	2266.025	72	14106	1
H. pylori (VacA)	9427	177.630	637.966	28	10515	1
H. pylori (HP1564)	9427	506.314	1189.140	43	12055	1
H. pylori (GroEL)	9427	1030.866	2233.087	21	13150	1
H. pylori (Catalase)	9427	592.704	1877.661	41	13972	1
H. pylori (UreA)	9427	440.928	1438.762	34	14085	1

Imputation, phasing data, and HLA alleles

We obtained classical HLA genotypes imputed by the UKBiobank using the HLA*IMP:02 algorithm, covering both class I (A, B, C) and class II (DRB1, DRB3, DRB4, DRB5, DPA1, DPB1, DQA1, DQB1) loci. Following UKB recommendations and prior studies, we set any allele call with the posterior probability below 0.7 to zero copies. Besides, the HLA-DR region is highly polymorphic and complex. In addition to the primary DRB1 gene, individuals may carry one of the secondary DRB genes: DRB3, DRB4, or DRB5, but never more than one. The presence or absence of these secondary genes varies between individuals and is tightly linked to specific DRB1 alleles due to strong linkage disequilibrium. The UKB appends a 99:01 suffix to indicate the absence of DRB3, DRB4, or DRB5 alleles; these were likewise set to zero.

For genome-wide imputation, we used the TOPMed reference panel (UKB `Field 21007`) in its unphased form. We employed `QCTOOL` (from UKB `app-swiss-army-knife` instance, `mem1_ssd1_v2_x16`) to extract our regions of interest from the BGEN and sample files. To generate the antigen presentation pathway allotypes ERAP1, ERAP2, TAP1, TAP2, PSMB8, and PSMB9, we extracted SNPs from chromosome 5 (ERAP1: chr5:96,674,484—97,019,703; ERAP2: chr5:96,876,500—96,919,703, extended by 10 kb upstream and downstream) and from chromosome 6 (TAP1/TAP2/PSMB8/PSMB9: chr6:32,725,415–32,959,851). We phased each extracted region using BEAGLE 4.1 (`beagle.21Jan17.6cc.jar`). Before phasing, variants with minor allele frequency $< 1\%$ and imputation quality < 0.30 were excluded. To ensure phasing accuracy, we ran BEAGLE five times per region and compared the resulting haplotypes and allotypes across replicates; only consistently phased variants were carried forward into downstream analyses. We repeated the phasing of replicates multiple times using `BEAGLE` and then assigned allotypes with our custom `C++` software (see Section 2.3 for details). We observed high consistency across phasing runs, with only a few samples ($n < 10$) showing differences in allotype assignments. Therefore, we randomly selected one for the downstream analysis.

2.3 Assimilation of antigen processing pathway genetic variation for downstream analysis

In this subsection, we describe in detail how the above datasets were processed to generate genetically derived variables for analysis, which were subsequently incorporated into our analytical pipeline along with their corresponding data sources. Specifically,

we classify these components into two categories: the HLA section, encompassing distinct HLA factors, and other pathway components, including genetic variants that define allotypes. This classification enables systematic calculation and interpretation of functional diversity within the pathway.

2.3.1 HLA factors

In Section 2.2, different HLA imputation methods were applied across cohorts. For the HCV spontaneous clearance vs. chronic infection cohort, HLA imputation and allele calling were conducted by Dr Guillaume Butler-Laporte (Butler-Laporte et al., 2023). This approach imputed both classical and non-classical HLA alleles, including loci such as HLA-F, thereby providing additional information that could be incorporated into the bioinformatics pipeline for downstream analyses. However, this method did not impute HLA amino acid polymorphisms and therefore could not capture specific amino acid variants such as the HLA-B -21 M/T dimorphism.

In contrast, the Michigan HLA Imputation Server, which was used for the STOP-HCV cohort and the MalariaGEN cohort, provides imputation at the amino acid level, enabling the analysis of specific residue variants and fine-mapping within HLA genes. However, it does not comprehensively include non-classical HLA loci. For the UK Biobank cohort, the HLA*IMP:02 algorithm was used for HLA imputation. This method provides four-digit classical HLA allele calls but does not impute non-classical loci or amino acid polymorphisms at the same resolution as the Michigan server.

These methodological differences may influence the types of HLA features that can be analysed in each cohort. Specifically, the HLA imputation approach may affect both the number and types of HLA alleles available for analysis, including whether non-classical HLA loci are imputed. Furthermore, some imputation methods do not infer HLA amino acid polymorphisms and therefore cannot capture specific variants, such as the HLA-B 21 M/T dimorphism.

Due to the central role of HLA molecules in antigen presentation, we further integrate quantitative HLA factors obtained through collaboration with Mary Carrington's laboratory. These factors include HLA alleles, HLA gene heterozygosity metrics, supertype classifications, tapasin-dependence scores, HLA A/C protein expression level estimates, and the well-studied HLA-B 21 M/T dimorphism. We analyse HLA alleles rather than individual variants or amino acids because alleles represent the complete functional genetic unit, incorporating all polymorphisms that may influence antigen presentation and phenotype. We synthesise these annotations and their known functional impacts from the literature to guide downstream analyses.

2.3.1.1 HLA alleles

HLA alleles (e.g., HLA-A*01:01) are conventionally named following the World Health Organisation (WHO) HLA Nomenclature Committee guidelines, where the first two digits denote the serological group and subsequent digits specify distinct protein or synonymous variants. In this study, we used four-digit (two-field) allele resolution, corresponding to amino acid level differences in the encoded HLA proteins. This level of resolution balances biological specificity with statistical power, as higher-resolution (three- or four-field) typing can introduce sparsity due to low allele frequencies, while lower-resolution (one-field) typing may mask functionally relevant variation.

Allele calls were derived from genotype data using a standardised HLA imputation server. For downstream analyses, we retained four-digit (two-field) HLA alleles to capture amino acid level variation and to ensure a comprehensive representation of antigen presentation diversity across the distinct cohorts.

2.3.1.2 HLA gene heterozygosity

The heterozygote advantage hypothesis posits that individuals carrying two different alleles at a given HLA gene are better equipped to resist pathogens than homozygotes. This benefit stems from the broader repertoire of pathogen-derived peptides that can be presented by heterozygous HLA molecules, thereby strengthening the adaptive immune response. For example, heterozygous individuals tend to present a wider array of HIV-1 peptides via their HLA class I alleles, and a broader peptide repertoire at the HLA-B locus has been linked to lower viral loads in people living with HIV-1 (Arora et al., 2020). We incorporated HLA gene heterozygosity into our analyses. For each HLA locus, individuals carrying two different alleles (e.g., HLA-A*01:01 and HLA-A*01:02) were classified as heterozygous, whereas those carrying identical alleles were classified as homozygous.

2.3.1.3 Tapasin dependency score

HLA class I allotypes vary in their ability to present peptides in the absence of tapasin, an essential component of the peptide-loading complex, which affects peptide selection. We included the tapasin-dependent level in our analysis with the method of quantification from Mary Carrington's group research, where they quantified the different HLA class I allotypes that require tapasin to bind peptides and applied their methodology to calculate the tapasin-dependence score for everyone (Bashirova et al., 2020). The results of the distinct allotypes tapasin-dependent (TD) score in

HLA class I are shown in Fig. 2.1. Specifically, for each individual at each locus, the tapasin-dependent score was calculated by summing the TD values of the two alleles: TD(HLA-A) = TD(A1) + TD(A2) and similarly for HLA-B and HLA-C: TD(HLA-B) = TD(B1) + TD(B2), TD(HLA-C) = TD(C1) + TD(C2). The global TD value for each individual was then calculated by summing the TD scores across the three loci: TD(HLA-A/B/C) = TD(HLA-A) + TD(HLA-B) + TD(HLA-C).

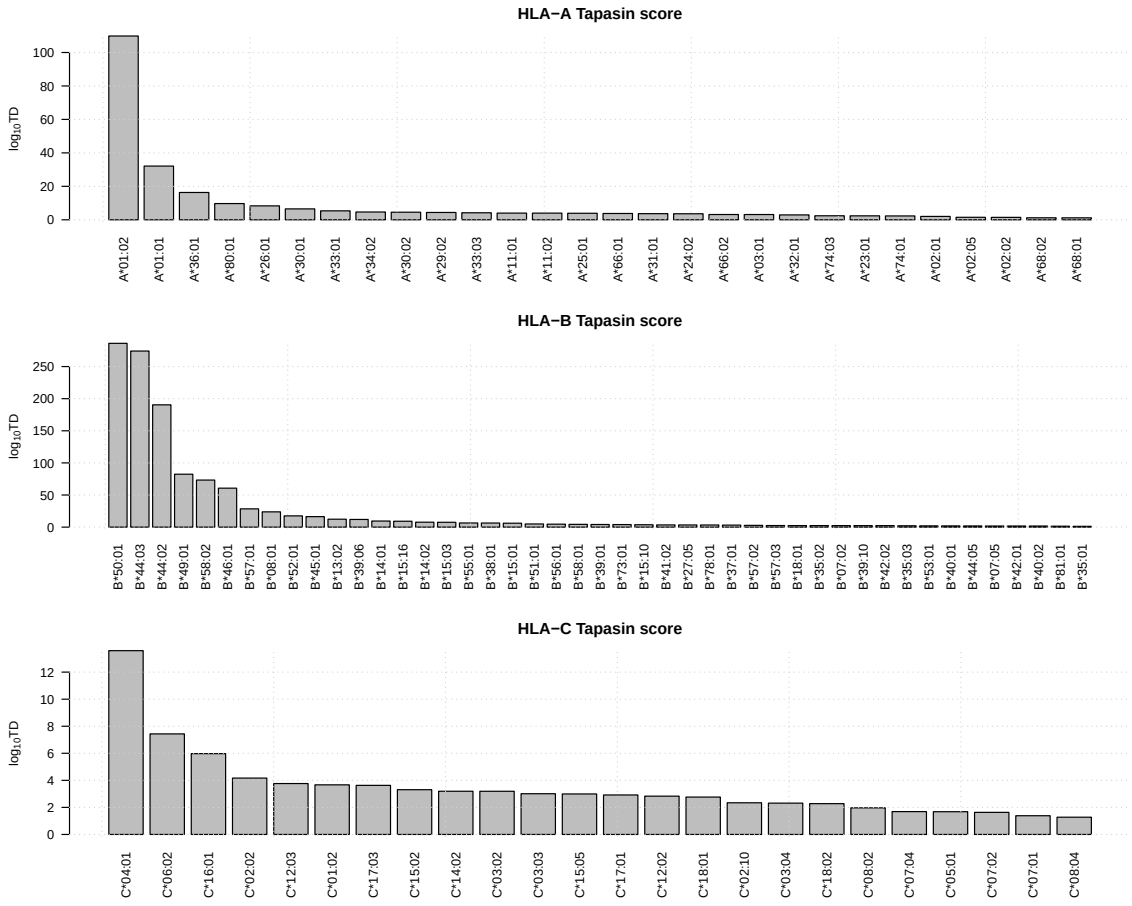


Figure 2.1: **Tapasin dependency for each HLA allotype, defined as the ratio of MFI of tapasin-positive over tapasin-negative cells, is shown in the log10 scale**

2.3.1.4 HLA supertypes

Peptide presentation by HLA molecules is determined largely by the polymorphic structure of their peptide–binding grooves. Despite extensive allelic diversity, class I and class II molecules can be grouped into supertypes that share overlapping peptide repertoires. For HLA class I, Sidney et al. analysed binding data across 945 HLA-A and HLA-B alleles and organised them into six A supertypes (A01, A01/A03,

A01/A24, A02, A03, A24) and six B supertypes (B07, B08, B27, B44, B58, B62) based on shared peptide-binding motifs (Sidney et al., 2008). Similarly, Greenbaum et al. defined seven class II supertypes from binding studies of 27 common HLA-DR, -DQ, and -DP molecules, namely Main DR, DR4, DRB3, Main DQ, DQ7, Main DP, and DP2, reflecting clustered specificity in their peptide repertoires (Greenbaum et al., 2011). In our study, we applied these established classifications to assign each HLA class I and class II allele to its corresponding supertype, facilitating comparisons of peptide presentation breadth and functional grouping across individuals.

2.3.1.5 HLA A/C allele-specific protein expression level

We incorporated allele-specific HLA-C protein expression levels. These are prior estimates of protein expression levels derived from experimental data. Higher HLA-C protein expression confers protection against HIV independently of classical HLA allelic effects (Apps et al., 2013). Elevated HLA-C levels were associated with stronger cytotoxic T lymphocyte responses and an increased rate of HIV viral escape mutations. Similarly, HLA-A alleles exhibit allotype-specific differences in expression, typically reaching 13– to 18-fold higher surface density than HLA-C and displaying greater polymorphism. Ramsuran reported that increased HLA-A protein expression impairs HIV control by enhancing the supply of HLA-A-derived signal peptides that bind HLA-E, thereby modulating engagement of the inhibitory NKG2A receptor on NK cells (Ramsuran et al., 2018). By integrating both HLA-C and HLA-A protein expression metrics into our analyses, we aim to capture the nuanced impact of protein expression variation on antigen presentation and immune regulation. In our analysis, we assigned each allele to its corresponding value from a previously generated database of protein expression levels. We then calculated the total protein expression score for each HLA-A and HLA-C locus by summing the expression values of the two alleles: PE(HLA-A) = PE (A1) + PE(A2), PE(HLA-C) = PE (C1) + PE(C2).

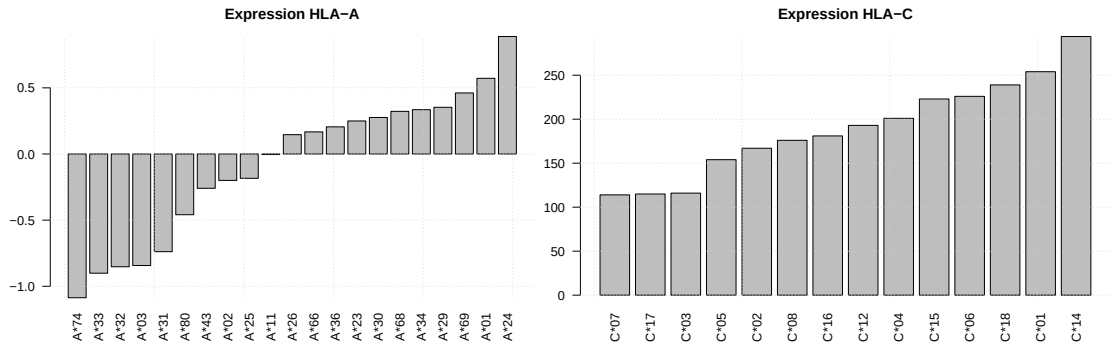


Figure 2.2: **Distribution of protein expression levels of HLA-A and HLA-C allotypes. The left panel shows HLA-A expression levels across distinct allotypes, and the right panel shows HLA-C expression levels. HLA-A expression values are presented on a standardised relative scale (z-score), whereas HLA-C expression levels are shown on the original quantitative scale (0–250), corresponding to previously reported expression estimates. These expression estimates were adapted from published studies on the role of HLA-A and HLA-C expression in HIV control (Apps et al., 2013; Ramsuran et al., 2018). The underlying data were provided by Mary Carrington's research group.**

2.3.1.6 HLA-B -21 M/T dimorphism

At position –21 of the HLA-B signal peptide (residue 2 of the mature leader sequence), a common polymorphism encodes either methionine (–21M) or threonine (–21T). In contrast, all HLA-A and HLA-C allotypes are fixed for methionine at this position. We included the HLA-B –21M/T dimorphism in our analysis because the methionine allele enhances HLA-E stabilisation and surface expression. HLA-E is a non-classical class I molecule that presents peptides derived from the leader sequences of other HLA class I proteins, including HLA-B. The –21 methionine (M) variant in the HLA-B leader peptide generates a high-affinity ligand for HLA-E, enhancing the stability of the peptide–HLA-E complex. This increased peptide binding promotes HLA-E surface expression, which can modulate immune recognition by natural killer and T cells. In contrast, the –21 threonine (T) variant produces a less effective peptide for HLA-E binding, resulting in reduced stabilisation and lower surface expression. Unlike HLA-A, whose transcriptional levels vary minimally, HLA-B expression shows broader allelic variation, and the –21M allele increases HLA-E expression in a copy-dependent manner Ramsuran et al. (2018).

2.3.2 Other key components during the antigen presentation pathway

2.3.2.1 Genetic variant characterization

Variant annotations were based on data from the Genome Aggregation Database (gnomAD), incorporating functional predictions generated using the Ensembl Variant Effect Predictor (VEP). We focused on protein-coding variants, including missense, synonymous, loss-of-function (LoF), and canonical splice-site changes with the MAF greater than 0.1% in any gnomAD population, within the genes ERAP1, ERAP2, TAPBP, TAP1, TAP2, PSMB8, PSMB9, PSMB10, CALR, CANX, PDIA3 (ERp57), and B2M (β2-microglobulin). We also included selected intronic and other non-coding variants previously shown to modulate gene expression. We incorporated rs2549794 in ERAP2, which shifts the balance between full-length and truncated transcripts and has been implicated in historical selective events such as the Black Death (Klunk et al., 2022). In total, our analysis included 132 SNPs across genes involved in the antigen presentation pathway, including *PSMB6–10*, *TAP1*, *TAP2*, *ERAP1*, *ERAP2*, *TAPBP*, *CALR*, *CANX*, and *ERp57*, excluding classical HLA alleles. As summarised in Table 2.11, the identified variants comprise coding (missense and synonymous) as well as non-coding and regulatory variants, including intronic, promoter, splice-region, stop-lost, flanking-region, and upstream variants. The detailed distribution of variant types within each gene is presented in Table 2.11. Comprehensive information for each SNP, including chromosomal position (GRCh38), rsID, reference and alternative alleles, functional annotation, and predicted amino acid consequence, is provided in Appendix B.1.

Table 2.11: **The number of targeted SNPs in each gene during the antigen presentation pathway.**

GENE/Annotation	**missense**	**synonymous**	**flanking_region**	**intron**	**promoter**	**splice_region**	**stop_lost**	**Upstream**
PSMB10	1	2	0	0	0	0	0	0
PSMB5	1	0	0	0	0	0	0	0
PSMB6	1	4	0	0	0	0	0	0
PSMB7	1	1	0	0	0	0	0	0
PSMB8	4	5	0	0	0	0	0	0
PSMB9	4	2	1	1	0	0	0	0
TAP1	14	5	0	0	1	1	0	0
TAP2	11	10	0	1	0	5	1	0
ERAP1	10	7	0	0	0	1	0	0
ERAP2	10	7	0	2	0	0	0	0
TAPBP	1	5	0	0	0	0	0	2
CALR	2	0	0	0	0	0	0	0
CANX	3	1	0	0	0	0	0	0
ERp57	1	4	0	0	0	0	0	0

2.3.2.2 Allotypes definition

In addition to the SNPs, we also focused on characterising allotypes for highly polymorphic genes, including PSMB8, PSMB9, TAP1, TAP2, ERAP1, and ERAP2, distinct protein variants encoded by combinations of multiple genetic variants within the same gene. Allotypes reflect the functional diversity arising from the specific constellation of missense mutations that collectively define a protein's structure and immunological properties. This is important in the context of infectious diseases, as different allotypes can alter antigen processing efficiency, peptide binding affinity, or immune recognition, thereby influencing individual susceptibility, disease progression, and vaccine responsiveness. To systematically capture this complexity, we developed a `C++` software tool to identify and quantify multi-variant allotypes in genes such as TAP1/2, ERAP1/2, and PSMB8/9, which harbour two or more common missense variants within the antigen presentation pathway.

2.3.3 Overall information included in the pipeline

Overall, the analysis incorporated information from 132 SNPs located in genes involved in the antigen presentation pathway, together with additional HLA-derived features (excluding classical HLA gene loci from the SNP set). For the HLA region, six categories of features were included: classical HLA alleles, gene-level heterozygosity, tapasin-dependence scores, HLA supertypes, predicted HLA-A and HLA-C surface expression levels, and the HLA-B 21 dimorphism. Thus, the model integrates SNP-level variation from 132 loci in antigen presentation pathway genes alongside multiple gene-level and functional HLA-derived features.

Classical HLA alleles and the 132 SNPs were obtained directly from the imputed genotyping data. In contrast, derived HLA features, including heterozygosity, tapasin dependence, supertypes, predicted expression levels, and the HLA-B 21 dimorphism, were computed using the software `hlafactor` (see the next section for details). Gene allotypes were also generated within this analytical pipeline and were therefore not directly available in the original dataset.

2.4 Development of a bioinformatics pipeline for large-scale genomic data and antigen presentation pathway-related software availability

To enable the systematic extraction and quantification of antigen presentation pathway (APP) related genetic features (as defined in Section 2.3), we developed two open source tools, `hlafactor` and `allotype`, together with a Snakemake-based workflow that automates the data extraction process. These tools automate the extraction of APP-related genetic variants, perform variant filtering, and compute key immunogenetic metrics, including allotypes and HLA-related factors such as heterozygosity, supertypes, HLA-B –21 M/T dimorphism, HLA-A/C alleles-specific protein expression levels, and tapasin dependence scores. Designed to efficiently process large genomic datasets, this toolkit supports downstream association and statistical analyses. In this section, we introduce the functionality and implementation of each tool.

2.4.1 HLAfactor

`HLAfactor` is a `C++` based software designed to process phased `VCF` files restricted to the extended HLA region, generating per-sample immunogenetic metrics. Specifically, `hlafactor` parses classical HLA allele calls (at two- or four-digit resolution) to assign supertypes, calculate heterozygosity scores, sum tapasin-dependence values, and retrieve allele-specific expression levels. All quantification references described in Section 2.3 are embedded as binary lookup tables within the software to ensure rapid and efficient annotation. Additionally, users can apply allele frequency filters directly within `hlafactor` by setting a minimum MAF threshold to exclude rare variants. The output is a tab-delimited summary file with one row per sample and separate columns detailing each computed HLA metric.

2.4.2 Allotype

To reconstruct multi-variant allotypes for peptide-processing genes—ERAP1/2, TAP1/2, and PSMB8/9, we developed a complementary `C++` based tool called allotype. This tool operates on dosage files generated by `QCTOOL` from region-specific `BGEN` inputs (e.g., chr5:96,674,484–97,019,703 for ERAP1/2 and chr6:32,725,415—32,959,851 for TAP/PSMB). Allotype automatically filters SNPs based on user-defined thresholds for imputation quality (R^2) and MAF, then collapses phased dosages into additive

allotype assignments. If any relevant SNPs are missing, the software will return a message indicating that the SNP is missing. The corresponding allotype (combination) will skip this SNP, and the final output will report the allotype as a string representing amino acid residues ordered from the N-terminus to the C-terminus (from low to high position). The resulting output includes text files that list per-sample allotypes identifiers alongside the corresponding variant dosages, facilitating downstream pathway analysis.

2.4.3 Bioinformatics pipeline

To streamline the analysis of genetic features during the antigen presentation pathway, we developed an integrated Snakemake-based pipeline capable of end-to-end processing of whole-genome imputed phased `VCF` files. Users can apply customisable filters for MAF and imputation quality (R^2), as well as select HLA resolution (two- or four-digit) and genetic models (additive, dominant, or recessive). Functional annotations such as missense or synonymous variants can be specified via VEP, and gene-specific analyses (e.g., ERAP1/TAP1) are supported. The workflow begins by extracting the HLA region (chr6:28,477,797–33,448,354 for GRCh37 or chr6:28,510,120–33,480,577 for GRCh38) and annotating relevant variants using the `QCTOOL` and `BCFTOOLS`. The pipeline employs `hlafactor` and `allotype` to generate dosage matrices and summarise key features for APP-related genes, including heterozygosity, supertypes, HLA-B 21 M/T dimorphism, tapasin dependency, and HLA allele-specific protein expression levels. All parameters are fully configurable via a `YAML` file, providing a flexible framework for diverse research objectives. Upon completion, the pipeline generates a unified tab-delimited report that consolidates all APP-related genetic features per sample, optimised for downstream association and statistical analyses.

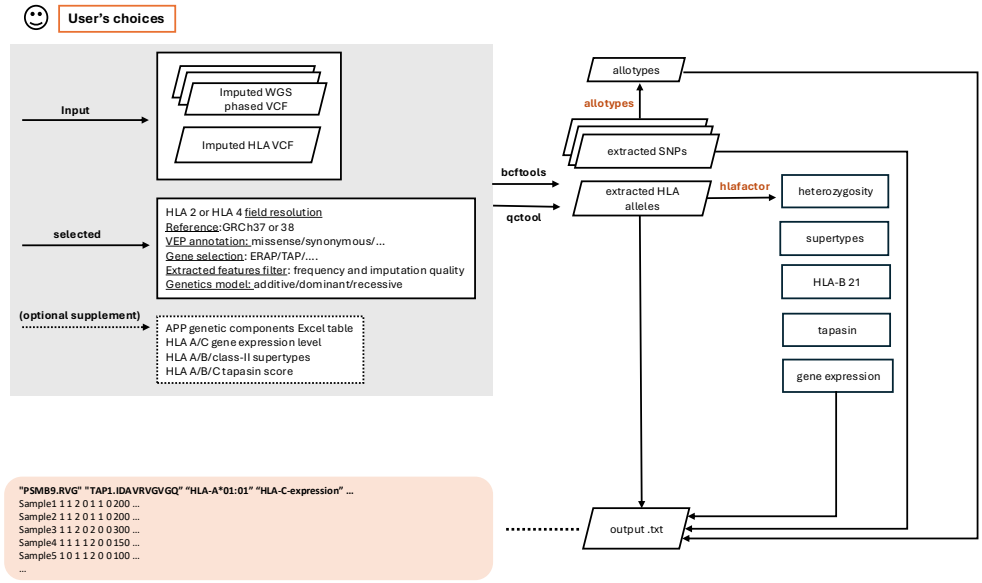


Figure 2.3: **Architecture and data flow of the Snakemake workflow for antigen presentation pathway genetic analysis**

2.4.4 Statistical comparison

In the Results section, some cohorts include individuals from different ancestry groups. To characterise heterogeneity between ancestries, selected quantitative features were compared across groups using two-sided independent-sample t-tests implemented in R (function `t.test()`). By default, R performs Welch's t-test, which does not assume equal variances between groups. When the assumption of equal variances was satisfied and explicitly specified (`var.equal = TRUE`), Student's t-test was applied instead.

Descriptive statistics are presented as mean values with corresponding standard errors (se). The standard error was calculated as

$$\mathrm{se} = \frac{\mathrm{sd}}{\sqrt{n}},$$

where sd denotes the sample standard deviation and n the sample size. Error bars in the figures represent mean$\pm 2\times$se. All tests were two-sided, and statistical significance was assessed using the corresponding p-values obtained from the tests.

URLs and software instructions

hlafactor (GitHub): https://github.com/QiJingS/hlafactor

Description: Calculate HLA gene heterozygosity (2-digit or 4-digit resolution), HLA supertypes, Tapasin Dependence Score, and HLA A/C protein expression levels.

Commands:

```
# Calculate heterozygosity (2-digit resolution)
./hlafactor -g hla.dosage -og output.txt -d2het

# Calculate heterozygosity (4-digit resolution)
./hlafactor -g hla.dosage -og output.txt -d4het

# Calculate HLA supertypes
./hlafactor -g hla.dosage -og output.txt -super

# Calculate Tapasin Dependence Score
./hlafactor -g hla.dosage -og output.txt -taps

# Calculate HLA expression levels
./hlafactor -g hla.dosage -og output.txt -expression
```

Allotype (GitHub): https://github.com/QiJingS/allotype

Description: Identify and quantify multivariant allotypes.

Command:

```
./allotype -g input.vcf -og output.txt \
           -freq 0.01 \
           -reference ../APP_summary.xlsx
```

Pipeline (GitHub): https://github.com/QiJingS/APPWAS.git

Description: Architecture and data flow of the Snakemake workflow for automated processing of whole-genome imputed VCF files to extract and summarise antigen presentation pathway genetic features.

Command:

```
snakemake -s pipeline.snakemake --cores 4
```

2.5 Results

In this section, we present the results of applying our bioinformatics pipeline (described in Section 2.4) to the four large-scale imputed genotype array datasets introduced in Section 2.2. Each dataset was processed to extract and summarise genetic variation within key components of the antigen presentation pathway. The analysis was performed in a population-stratified manner, enabling us to compare the distribution of genetic features across diverse ancestry groups. For each dataset, we produced detailed summaries of variant frequencies, HLA allele distributions, and inferred allotype diversity, which were visualised to provide a comprehensive overview of immunogenetic diversity in the APP across populations. For each cohort, allotypes are reported directly as ordered amino acid sequences. The corresponding SNPs and the resulting amino acid changes are provided in Appendix A Tables 1–4 for each cohort, with entries listed according to the order of amino acid changes. The results presented in this section served as the foundation for the association analyses described in Chapters 3 and 4, where we examined how variation in the antigen presentation pathway correlates with serological profiles and clinical outcomes.

2.5.1 HCV spontaneous clearance vs chronic infection

Using the Snakemake pipeline described in Section 2.3, we analysed APP genetic variants within the cohort of spontaneous clearance vs chronic infection. To focus on common variation, we applied the MAF threshold of more than 1%. HLA alleles were considered at four-digit resolution under an additive genetic model, and only variants with an imputation quality score (R^2) greater than 0.9 were retained. After applying these filters, the resulting dataset comprised high-confidence APP variants, as shown in Figure 2.4. Among HLA class I loci, 6 non-classical and 18 classical alleles met the inclusion criteria. For HLA class II loci, 7 non-classical and 27 classical alleles were retained. In addition, we identified 15 SNPs in PSMB5–10, 15 in TAP1/TAP2, 27 in ERAP1/ERAP2, and 3 in TAPBP.

Given the known influence of genetic ancestry on phenotypic variation, we stratified participants into European and African subgroups. As shown in Figure 2.4 B, individuals of African ancestry exhibited significantly lower allele-specific protein expression of HLA-A ($p = 6.46 \times 10^{-7}$) and higher protein expression of HLA-C ($p = 3.99 \times 10^{-4}$) compared to their European counterparts. Figure 2.4 C and D further illustrate that Africans had significantly lower tapasin-dependence scores for HLA-A ($p = 1.06 \times 10^{-13}$) and HLA-B ($p = 2.59 \times 10^{-11}$), while HLA-C scores

were significantly higher ($p = 1.75 \times 10^{-6}$). The global tapasin-dependence score was higher in Europeans than in Africans ($p = 3.38 \times 10^{-6}$), with HLA-B contributing the most to this composite score in both populations.

Figure 2.4 E demonstrates consistently high heterozygosity, more than 0.8 at classical HLA loci (A, B, C, DP, DQ, DR) across both ancestries. In contrast, non-classical loci (F, G, DM, DO) exhibited lower heterozygosity, with notable ancestry-specific differences, for instance, reduced *DPA1* heterozygosity in Europeans compared to Africans. Figure 2.4F highlights distinct patterns of HLA supertype distribution between ancestries, suggesting potential contributions to phenotypic diversity. For example, the A01A03 and A01A24 supertypes were relatively rare in both groups, although overall supertype frequencies differed markedly.

We also assess allotype frequencies of key APP genes across ancestries; see Appendix E for the corresponding SNPs extracted from the pipeline and their associated amino acids. In ERAP1, both populations share seven common allotypes; however, the `TEPIDMRDRE` allotype was more frequent in Africans, while `IEPIGMKDRQ` is prevalent in Europeans but absent in Africans. In ERAP2, two common allotypes are shared across groups, with an additional European-specific allotype, `TKQ`, absent in Africans. PSMB8 and PSMB9 allotypes are consistent across ancestries. For TAP1, Africans exhibited four common allotypes, including `VGAVQLG` and `IDAVRVC`, which are not observed in Europeans, who have only two common allotypes. In TAP2, `TVARLRA` and `AVARLRA` are shared, but three allotypes, `TVARFRA`, `TVTRLRA`, and `TIARFRA`, are unique to Africans, while `TITELRA`, `TVACLRA`, and `TIARLRA` are exclusive to Europeans.

Collectively, these findings underscore the extensive diversity of APP-related genetic variants in the spontaneous clearance cohort and highlight ancestry-specific immunogenetic differences between African and European populations. These differences may contribute to variation in immune response dynamics and outcomes.

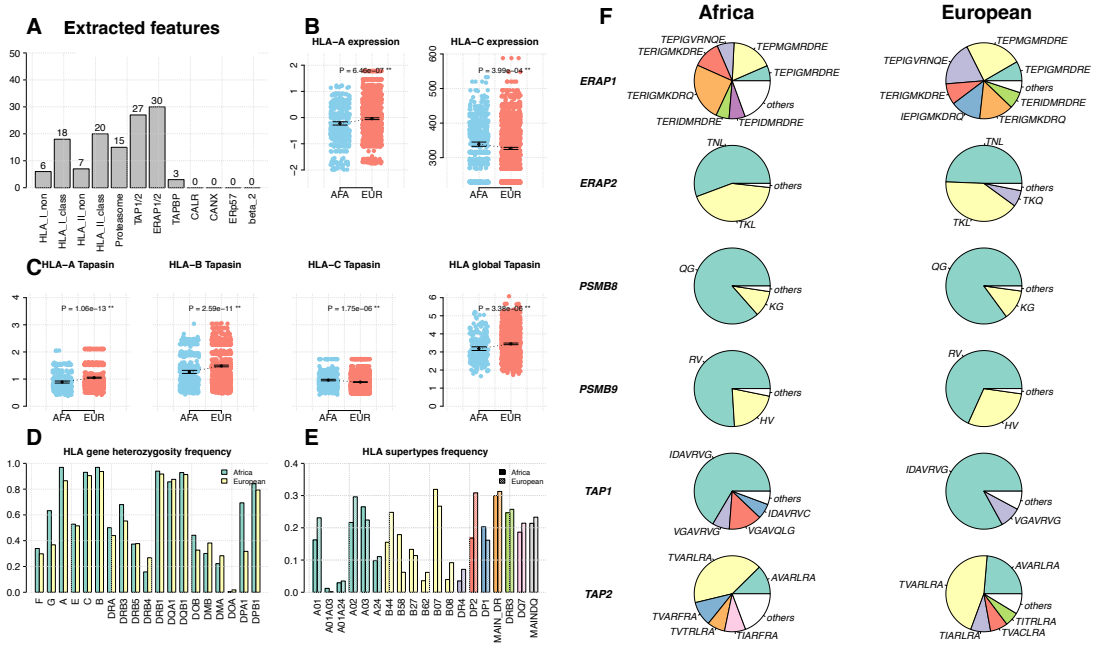


Figure 2.4: **Distribution of key APP features in the spontaneous clearance vs chronic infection cohort.** (A) Number of common ($\geq$ 1% frequency) SNPs extracted from each APP gene; HLA allele counts are shown at four-digit resolution. HLA-I includes both non-classical and classical alleles (corresponding to HLA-I_non and HLA-I_class), and the same classification applies to HLA-II. (B) Individual HLA-A and HLA-C protein expression (predicted) using previously published allele-specific expression estimates, stratified by European (blue) and African (red) ancestry. Bars represent mean $\pm$ SE; p-values from unpaired two-sided t-tests are indicated. (C) Distribution of locus-specific tapasin-dependency scores (HLA-A, HLA-B, HLA-C) and the global HLA-A/B/C score, by ancestry. (D) HLA gene heterozygosity frequency at each classical and non-classical locus. Blue bars: European; red bars: African. (E) Relative frequencies of HLA class I and class II supertypes in European (solid fill) versus African (hatched) groups; colours denote supertype families. (F) Frequencies of common ($>$ 5%) allotypes in the whole cohort, shown for each gene: left column, African ancestry; right column, European ancestry. Rows correspond to the same gene, with matching colours indicating the same allotype.

2.5.2 STOP-HCV

From the STOP-HCV dataset, we applied the bioinformatics pipeline to analyse APP genetic features. To ensure high-confidence and common variants, we filtered for MAF $>$ 1%, restricted HLA allele calls to four-digit resolution, retained only variants with

imputation quality scores ($R^2 > 0.9$), and assumed an additive genetic model. After filtering, the dataset included 16 alleles from classical HLA class I loci and 26 alleles from classical class II loci. Additionally, we identified 8 SNPs in PSMB5–10, 21 SNPs in TAP1/TAP2, 32 in ERAP1/2, and 2 in TAPBP as shown in Figure 2.5 A.

To explore ancestry-specific patterns, individuals were grouped by self-reported ancestry into European and South Asian subgroups. As shown in Figure 2.5 B, allele-specific protein expression of HLA-A was significantly higher in South Asians compared to Europeans ($p = 2.25 \times 10^{-2}$). Protein expression of HLA-C was also elevated in the South Asian group, with a mean expression level exceeding 350, compared to approximately 310 in Europeans. In terms of tapasin-dependence scores, no significant differences were observed between populations for HLA-A and HLA-B. However, HLA-C displayed higher tapasin-dependence in South Asians, approaching statistical significance ($p = 6.3 \times 10^{-2}$). The global tapasin-dependence score was slightly higher in South Asians than in Europeans.

Figure 2.5 D illustrates heterozygosity levels at classical HLA loci (A, B, C, DRB1, DQA1, DQB1, and DPB1), which were uniformly high (> 0.8) in both ancestry groups. An exception was DPA1, which exhibited markedly lower heterozygosity, below 50%, in both populations. Figure 2.5 E presents the distribution of HLA supertypes by ancestry. While some supertypes, such as A01A03 and A01A24, are rare in both groups, others displayed clear population-specific differences. The A03 supertype in HLA-A was more prevalent in South Asians, while HLA-B supertypes show minimal differences between groups, with frequency differences under 5%. In HLA class II loci, DQ7 and major DR supertypes are more common in Europeans than in South Asians.

Allotype distribution patterns also varied between the two populations; see Table A for the corresponding SNPs and associated amino acid changes. In ERAP1, both groups shared the same allotypes, though their frequencies differed. For ERAP2, two common allotypes were shared, but the *TKQ* allotype was unique to Europeans, while *TKL* was exclusive to South Asians. PSMB8, PSMB9, and TAP1 allotypes were consistent across both groups, with nearly identical frequencies. In TAP2, both populations shared the same set of allotypes, although the *TITRL* allotype was common in Europeans but not observed in South Asians.

Taken together, these results highlight the extensive immunogenetic diversity within the STOP-HCV cohort and reveal ancestry-specific variation in APP genes. These differences may have functional relevance, potentially influencing host immune

responses and contributing to differential rates of spontaneous hepatitis C virus clearance between European and South Asian populations.

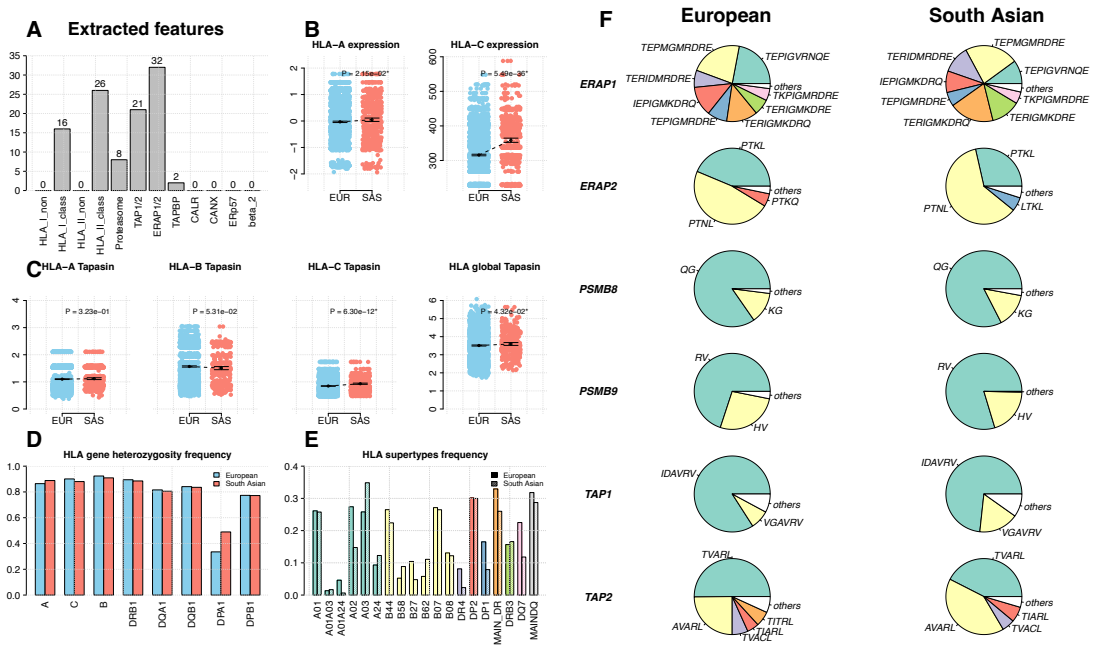


Figure 2.5: **Distribution of key APP features in the STOP-HCV cohort.** (A) Number of common (> 1% frequency) SNPs extracted from each APP gene; HLA alleles were analysed at four-digit resolution. HLA-I includes both non-classical and classical alleles (corresponding to HLA-I_non and HLA-I_class), and the same classification applies to HLA-II. (B) Distribution of HLA-A and HLA-C expression levels for each individual, stratified by European and South Asian ancestry. Bars represent mean ± SE; p-values from unpaired two-sided t-tests are indicated. (C) Distribution of locus-specific tapasin-dependence scores (HLA-A, HLA-B, HLA-C) and the global HLA-A/B/C score, by ancestry. (D) Heterozygosity frequency at each HLA locus; blue bars indicate Europeans, red bars indicate South Asians. (E) Relative frequencies of HLA-A, HLA-B, and class II supertypes; colours denote loci, with solid fill for Europeans and hatched fill for South Asians. (F) Frequencies of common (> 5%) allotypes in each APP gene, stratified by ancestry. Left column: South Asian ancestry; right column: European ancestry. Rows correspond to the same gene, with matching colours indicating the same allotype.

2.5.3 MalariaGEN

Using data from the MalariaGEN consortium, we grouped individuals based on geographic regions into Western Africa (Mali, Nigeria, Cameroon, Ghana, Burkina Faso, and The Gambia), Eastern Africa (Tanzania, Kenya, and Malawi), and a non-African group (Papua New Guinea and Vietnam), to explore regional differences in genetic diversity. Our analysis pipeline included only variants with allele frequencies greater than 1% and imputation quality above 90%, focusing specifically on HLA alleles and related immune genes. As shown in Figure2.6 B, we extracted genetic features encompassing 18 HLA class I alleles, 30 HLA class II alleles, 20 proteasome-related SNPs, 41 SNPs in TAP1/2, 31 in ERAP1/2, and 3 from the TAPBP gene. Analysis of HLA supertypes revealed marked differences between African and non-African populations. For instance, the A03 supertype was notably more frequent in non-African populations, particularly in Papua New Guinea, where it approached 60%, and in Vietnam, where it exceeded 30%. In contrast, African populations showed a more balanced distribution of HLA-A supertypes, with A03 frequencies just above 30% and slightly higher in Western compared to Eastern Africa. Papua New Guinea was characterised by a limited HLA-A diversity, dominated by A03 and A24, while Vietnam exhibited a broader range of supertypes, including higher frequencies of A01, A01A03, A01A24, and A02. However, A01 remained less common in Vietnam than in African populations. For HLA-B, the B07 supertype was most frequent in African populations, above 30% in Western Africa and above 20% in Eastern Africa, while it was considerably less common in non-African populations. Conversely, the B62 supertype was more prevalent in non-African groups, with frequencies close to 20% in both Papua New Guinea and Vietnam, but remained rare across African regions. Regarding HLA class II supertypes, DRB3 was consistently more frequent in African populations (over 10%) and nearly absent in non-African ones. While the frequencies of most class II supertypes were comparable between Western and Eastern Africa, the MainDR supertype showed a distinct regional difference, being more frequent in Eastern Africa ($\sim 30\%$) than in Western Africa ($\sim 20\%$).

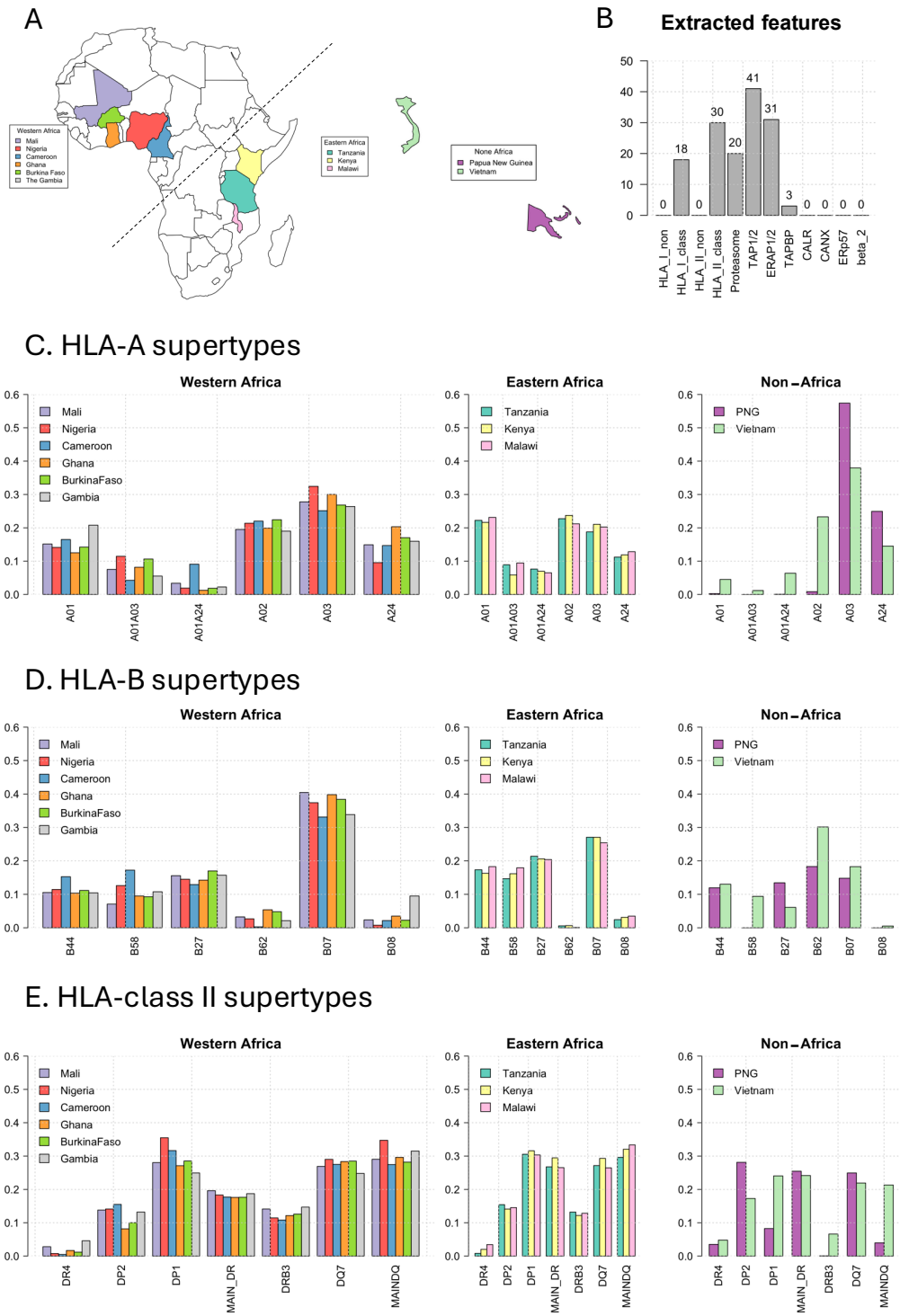


Figure 2.6: **Distribution of key APP features in the MalariaGEN cohort.** A. Map of study regions. Countries from the MalariaGEN dataset were grouped as follows: Western Africa (Mali, Nigeria, Cameroon, Ghana, Burkina Faso, Gambia), Eastern Africa (Tanzania, Kenya, Malawi), and Non-Africa (Papua New Guinea, Vietnam). B. Number of common (> 1% frequency) SNPs extracted from each APP gene; HLA alleles were analysed at four-digit resolution. C. Distribution of HLA supertypes (A, B, and Class II) across the MalariaGEN regions.

In Figure 2.7 A, which illustrates HLA expression levels, we observed regional differences across the studied populations. In Western African countries, the average expression level of HLA-A was close to zero, while in Eastern African countries, it was slightly higher, averaging around 0.5. Among the non-African populations, Papua New Guinea (PNG) exhibited marginally higher HLA-A expression levels compared to Vietnam. For HLA-C protein expression, the Eastern African countries demonstrated consistent average values around 350. In contrast, within Western Africa, Nigeria and The Gambia showed comparatively lower expression levels than the other countries in the same group. Notably, PNG displayed the highest average HLA-C protein expression among all the countries studied.

Figure 2.7 B presents tapasin scores across HLA loci. The HLA-A tapasin binding scores were relatively consistent across all countries, apart from PNG, which showed slightly elevated values. For HLA-B, Vietnam exhibited the highest tapasin score, and Eastern African countries, on average, had higher scores than those in Western Africa. Within Western Africa, Cameroon had the highest HLA-B tapasin score, standing out among its regional peers and aligning more closely with the levels observed in Eastern Africa. Regarding HLA-C, the tapasin dependent scores were generally uniform across all countries, although Vietnam exhibited a slightly lower average compared to the others. Overall, the global tapasin scores remained relatively stable across populations, with Vietnam showing a modestly higher average.

Table 9 shows the frequency distribution of various immune-related gene allotypes across different populations, including Western Africa, Eastern Africa, and non-African regions (Vietnam and Papua New Guinea); see Appendix A for the corresponding SNPs and associated amino acid changes. Genes analyses include PSMB9, PSMB8, TAP1, TAP2, ERAP1, and ERAP2, with multiple allotypes per gene. The PSMB8 *QGT* allotype is consistently high in frequency across all African populations (81.1–87.7%) and remains prevalent in non-African populations as well. In contrast, PSMB9 *HVG* shows higher variability, with a notable peak in Gambia (37.4%) compared to lower frequencies in Eastern Africa (11.1–13.1%). TAP1 allotypes such as *IDAVRVGVGQ* also display regional variation, being more common in Western Africa (up to 75.4% in Gambia) than in non-African regions. TAP2 allotypes reveal marked population-specific differences; for example, *TVARLRMAVVA* is highly frequent in Nigeria (46.6%) and Malawi (36.9%) but less so in Vietnam (35.2%) and rare in PNG (11.8%). ERAP1 and ERAP2 allotype distributions also vary significantly, with some alleles like *TERIGMKDRQ* showing high prevalence in both African (up

to 30.7% in Tanzania) and non-African populations (e.g., 39.0% in Vietnam). Finally, ERAP2 allotypes *PSVGTN* and *PSVGKT* are nearly evenly distributed in most populations, though PNG shows a distinct preference for *PSVGKT* (68.4%). Overall, the data highlight substantial genetic diversity and population specificity in antigen-processing gene allotypes.

Gene	Allotypes	Burkina Faso	Cameroon	Gambia	Ghana	Nigeria	Mali	Malawi	Kenya	Tanzania	Vietnam	PNG
PSMB9	HVG	25.2%	21.8%	37.4%	24.2%	29.4%	26.0%	11.5%	11.1%	13.1%	21.5%	18.6%
PSMB8	QGT	84.3%	87.4%	87.7%	86.2%	87.4%	81.1%	90.1%	88.5%	88.4%	81.8%	99.0%
PSMB8	KGT	13.1%	10.2%	11.6%	12.2%	11.1%	15.9%	8.3%	8.5%	8.9%	17.6%	1.0%
TAP1	IDAVRVGVGQ	59.9%	68.0%	75.4%	59.0%	64.9%	64.1%	61.8%	59.6%	58.4%	74.9%	86.0%
TAP1	VGAVQLGVGQ	12.5%	15.3%	12.8%	14.7%	8.4%	13.5%	23.0%	25.2%	26.6%	0.0%	0.1%
TAP1	IDAVRVCVGQ	8.9%	7.0%	3.3%	8.4%	12.2%	6.3%	8.1%	6.1%	6.4%	0.0%	0.0%
TAP2	TVARFRMAVVA	9.3%	8.6%	8.8%	11.5%	10.3%	11.0%	4.5%	4.6%	4.9%	1.0%	11.6%
TAP2	TIARFRMAVVA	11.3%	10.3%	24.9%	14.0%	10.3%	12.4%	9.7%	7.9%	6.7%	0.0%	0.0%
TAP2	AVARLRMAVVA	5.9%	9.4%	5.1%	7.7%	9.5%	5.8%	9.4%	12.8%	13.8%	30.4%	61.9%
TAP2	TVARLRMAVVA	45.9%	33.0%	37.6%	42.2%	46.6%	41.4%	36.9%	38.9%	38.9%	35.2%	11.8%
TAP2	TVARLRMTIVA	4.3%	1.3%	7.8%	2.8%	0.8%	7.8%	9.0%	11.3%	8.1%	0.0%	0.0%
TAP2	TVTRLRMAVVA	2.9%	16.6%	3.1%	1.4%	7.3%	3.0%	13.2%	13.5%	11.9%	3.7%	0.1%
ERAP1	TEPIDMRDRE	9.9%	9.7%	10.4%	8.8%	7.6%	8.4%	9.1%	9.4%	8.4%	0.0%	0.0%
ERAP1	TEPIGVRNQE	6.3%	2.6%	5.8%	5.6%	1.9%	6.1%	4.0%	4.2%	4.3%	4.6%	40.9%
ERAP1	TERIGMKDRE	9.2%	10.7%	15.8%	13.3%	9.2%	14.5%	15.7%	13.5%	13.1%	1.5%	5.5%
ERAP1	TEPMGMRDRE	19.2%	20.8%	15.5%	15.1%	17.2%	17.2%	16.3%	15.7%	19.1%	37.2%	37.7%
ERAP1	TERIGMKDRQ	25.8%	27.7%	27.2%	25.0%	29.8%	25.8%	30.2%	28.4%	30.7%	39.0%	11.8%
ERAP1	TKPIDMRDRE	5.7%	3.7%	5.7%	3.6%	5.7%	3.8%	5.7%	7.1%	5.0%	0.0%	0.0%
ERAP1	TERIDMRDRE	5.8%	8.7%	5.2%	9.5%	8.0%	5.3%	6.0%	8.5%	6.3%	0.2%	0.0%
ERAP2	PSVGTN	56.0%	50.9%	56.0%	58.0%	60.7%	54.7%	48.3%	45.5%	44.9%	43.7%	27.9%
ERAP2	PSVGTK	40.7%	44.2%	36.4%	38.8%	34.0%	41.7%	47.2%	46.6%	47.0%	45.2%	68.4%

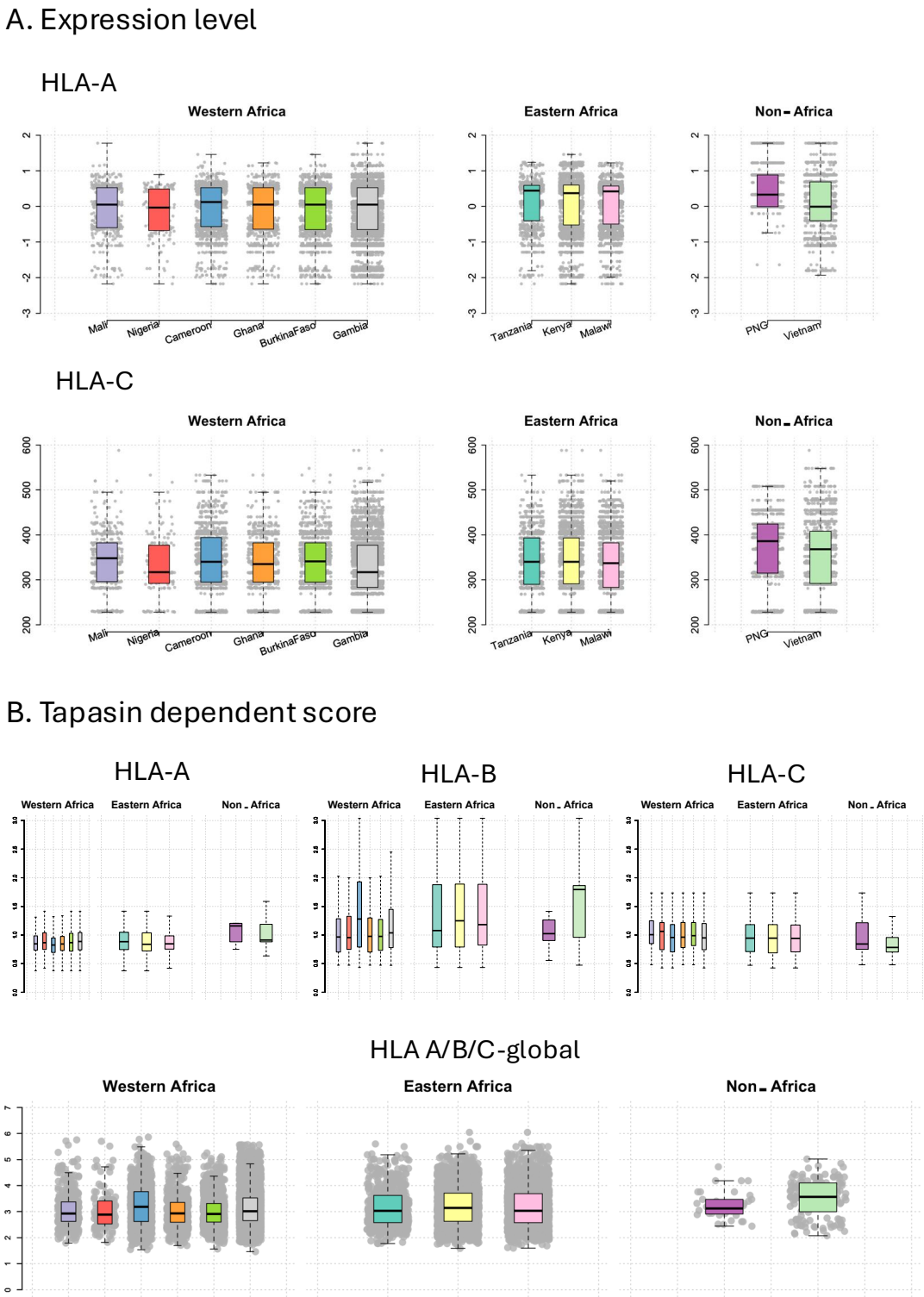


Figure 2.7: A. Allele-specific protein expression level across 11 countries in MalariaGEN. B. Tapasin dependent score across 11 countries in MalariaGEN. Box plots indicate similar distributions of both measures across all populations analysed.

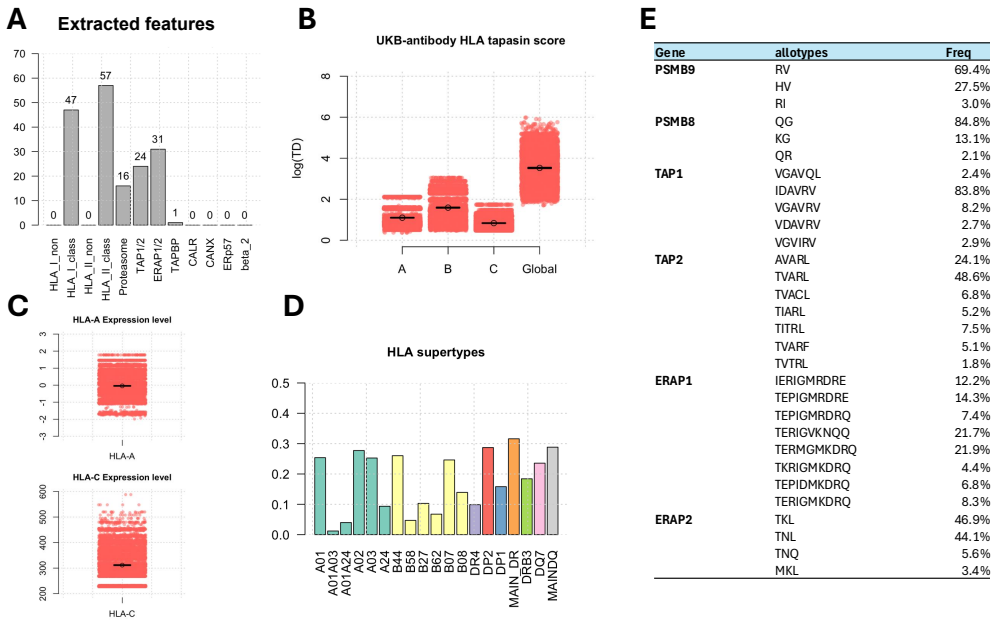
2.5.4 UKB serology cohort

In the UK Biobank serology cohort, we focus on 9,427 individuals of white British ancestry based on self-report. Using our bioinformatics pipeline, we analyse variants in key antigen processing and presentation genes. We apply a MAF threshold of $> 1\%$, consider HLA alleles at four-digit resolution, use an additive genetic model, and retain only variants with imputation $R^2 > 0.9$. The filtering process yield 16 SNPs in proteasome-related genes (PSMB5–PSMB10), 24 SNPs in TAP1/TAP2, 31 in ERAP1/ERAP2, and 1 in TAPBP (Figure 8A).

For HLA tapasin-dependence scores (Figure2.8B), HLA-B display the highest mean score among the class I loci (HLA-A, HLA-B, HLA-C) and contributed most to the global score. HLA allele-specific protein expression levels (Figure2.8C) show HLA-A protein expression centred around zero, while HLA-C exhibit a higher average expression level of approximately 300.

HLA supertype distribution (Figure2.8D) revealed that certain HLA-A supertypes, specifically A01A03 and A01A24, are relatively rare in this cohort. In contrast, HLA-B supertypes such as B07, B08, and B44 are more frequent, each occurring in more than 10% of individuals.

Finally, the distribution of APP gene allotypes is summarised in Figure 2.8 E; see Appendix A for the corresponding SNPs and associated amino acid changes. For PSMB9, *RV* and *HV* are the most common allotypes, while *QG* and *KG* are dominant in PSMB8. In TAP1, the *IDAVRV* allotype is particularly prevalent (83.8%). For TAP2, *AVARL* and *TVARL* are the major types, with frequencies of 24.1% and 48.6%, respectively. ERAP1 exhibits notable diversity, with eight distinct allotypes identified. In ERAP2, the *TKL* and *TNL* allotypes are the most common, each present in over 40% of individuals, while *TNQ* and *MKL* are much rarer.



Gene	allotypes	Freq
PSMB9	RV	69.4%
	HV	27.5%
	RI	3.0%
PSMB8	QG	84.8%
	KG	13.1%
	QR	2.1%
TAP1	VGAVQL	2.4%
	IDAVRV	83.8%
	VGAVRV	8.2%
	VDAVRV	2.7%
	VGVIRV	2.9%
TAP2	AVARL	24.1%
	TVARL	48.6%
	TVACL	6.8%
	TIARL	5.2%
	TITRL	7.5%
	TVARF	5.1%
	TVTRL	1.8%
ERAP1	IERIGMRDRE	12.2%
	TEPIGMRDRE	14.3%
	TEPIGMRDRQ	7.4%
	TERIGVKNQQ	21.7%
	TERMGMKDRQ	21.9%
	TKRIGMKDRQ	4.4%
	TEPIDMKDRQ	6.8%
	TERIGMKDRQ	8.3%
ERAP2	TKL	46.9%
	TNL	44.1%
	TNQ	5.6%
	MKL	3.4%

Figure 2.8: **Distribution of key APP features in the UKB serology cohort**. A. Number of extracted SNPs (MAF $>$ 1%) across APP genes after quality control filtering. HLA-I includes both non-classical and classical alleles (corresponding to HLA-I_non and HLA-I_class), and the same classification applies to HLA-II. B. Tapasin-dependence scores for HLA-A, HLA-B, HLA-C, and global scores, showing HLA-B as the dominant contributor. C. Allele-specific protein expression levels for HLA-A (top) and HLA-C (bottom), with notably higher mean expression for HLA-C. D. Frequency distribution of HLA supertypes, with low prevalence of A01A03 and A01A24 in HLA-A, and high frequency of B07, B08, and B44 in HLA-B. E. Distribution of APP gene allotypes with frequencies, highlighting dominant variants in PSMB8, PSMB9, TAP1, TAP2, ERAP1, and ERAP2.

2.6 Summary

In this section, we characterised the distribution of APP components across multiple cohorts and ancestries. Our analysis revealed consistent population-level patterns. In European populations, the allele-specific expression level of HLA-C was approximately 300 across all datasets. In contrast, African populations, particularly in the MalariaGEN and spontaneous clearance cohorts, show elevated HLA-C protein expression levels of approximately 320. South Asian populations, represented in the STOP-HCV study and a Vietnamese cohort within MalariaGEN, display even higher HLA-C expression, averaging around 350. Regarding the HLA tapasin-dependent score, HLA-B consistently showed higher scores than HLA-A and HLA-C in all cohorts. Tapasin scores are highest in Asian populations, followed by European, then African groups, suggesting population-level differences in peptide-loading efficiency.

We also observed differences in the distribution of HLA supertypes. For instance, A01A03 and A01A24 are among the supertypes with the lowest frequencies across all datasets. However, regional variability was prominent. The HLA-B07 allele, for example, appeared at frequencies above 30% in the African MalariaGEN cohort and is also enriched in the spontaneous clearance dataset compared to European cohorts. In terms of HLA allotypes, we identify the *IEPIGMKERQ* allotype as common in European and South Asian populations, but not in African datasets, where it is largely absent in both MalariaGEN and spontaneous clearance cohorts. Further analysis of the ERAP2 gene shows fewer common allotypes compared to ERAP1. In African populations, *TN* and *TK* types are predominant, particularly in the spontaneous clearance and MalariaGEN datasets. Conversely, in European cohorts, region-specific allotypes such as *PTKQ* are observed, while a novel type, *TNQ*, emerged in the UK Biobank cohort with a frequency slightly above 5%. South Asian populations exhibit unique ERAP2 allotypes, such as *LTKL*, distinct from both African and European profiles. Meanwhile, PSMB8 and PSMB9 show consistent allele frequencies across all datasets.

TAP1 and TAP2 genes exhibit further population-specific variation. Europeans and South Asians primarily harboured two TAP1 allotypes: *IDAVRVG* and *VGAVRVG*. In contrast, African populations display additional common allotypes, such as *IDAVRVC* and *VGAVQLG*, found in both MalariaGEN and spontaneous clearance datasets. TAP2 showed greater diversity than TAP1; for example, the *TITRL* allotype is common in European datasets (STOP-HCV) but is not present in South Asian or African

cohorts. Meanwhile, Africa-specific allotypes, including *TVARFRA*, *TVTRLRA*, and *TIARFRA*, are found in both the spontaneous clearance and MalariaGEN datasets.

The outputs generated by this bioinformatics pipeline provide a rich set of immune-genetic features and will be utilised in the next chapter for association analyses to further investigate their roles in phenotypes.

2.7 Discussion

In Chapter 2, building on the concepts introduced in Chapter 1, we review key findings from previous studies on infectious diseases and highlight the central role of the antigen presentation pathway in immune defence against viral and bacterial pathogens. To enable a more systematic exploration of this pathway and to assess the influence of genetic variation on immune responses, we developed a comprehensive bioinformatics pipeline designed for whole-genome sequencing data. The pipeline extracts detailed genetic and immunological features relevant to antigen presentation and produces a single summary file ready for downstream analysis.

Because the HLA plays a pivotal role in adaptive immunity by presenting peptide antigens to T-cell receptors, we placed particular emphasis on characterising HLA variation. To this end, we developed two `C++` software tools: `hlafactor`, which calculates immunological features such as tapasin-dependent peptide-loading scores, HLA supertypes, gene heterozygosity, and allele-specific expression, and `allotype`, which infers gene allotypes from SNP combinations. Although originally designed for antigen presentation genes, the latter is broadly applicable to other gene families. An important strength of this framework is its compatibility with diverse imputation outputs: the pipeline and software can accommodate data from different imputation platforms, including the Michigan Imputation Server, TOPMed, and other commonly used resources. This flexibility ensures consistent applicability across studies that use different data-generation strategies.

We integrated these tools into a scalable Snakemake-based pipeline and applied it to four large datasets: an early-stage HCV cohort focused on spontaneous clearance vs. chronic infection cohort, the STOP-HCV cohort, the severe malaria cohort from the MalariaGEN consortium, and an antibody serology cohort from the UK Biobank. Across these datasets, we observe marked variation in antigen presentation features, reflecting both ancestry-specific differences and disease-related diversity. Our pipeline and software provide several advantages over existing approaches. They extract both common and rare HLA variants, infer allotypes, and summarise gene-level functional

features in a scalable and generalisable framework. Importantly, unlike many prior studies that focus narrowly on HLA alleles, our approach integrates non-HLA genes and other key components of antigen processing, enabling a broader and more complete characterisation of the pathway. This wider perspective allows us to capture genetic effects that might be overlooked when analyses are limited to single loci.

Our findings should be interpreted in the broader context of previous studies showing that variation across the antigen presentation pathway is both highly population-structured and phenotypically consequential. Prior work has examined key APP components, including HLA class I supertypes and alleles, ERAP1 and ERAP2 haplotypes/allotypes, and variants in TAP1, TAP2, TAPBP, and PSMB8/PSMB9, and has consistently shown that their frequencies differ substantially across populations and may contribute to differences in immune responses and disease phenotypes.

Among APP features, HLA class I variation is the best characterised across ancestries. HLA class I supertypes, defined by shared peptide-binding motifs and anchor residue preferences (Sidney et al., 2008), provide a useful functional framework for comparing antigen presentation capacity across populations. Large population studies have shown pronounced geographic heterogeneity in classical HLA allele and haplotype frequencies, which in turn shapes supertype distributions worldwide (Solberg et al., 2008; Maiers et al., 2007; Gonzalez-Galarza et al., 2020). For example, HLA-A02 (A2 supertype) and several alleles within the B7 and B44 supertypes are broadly represented in many Eurasian populations, whereas HLA-B27 shows marked ancestry-related variation and is enriched in some European populations. This is especially relevant because HLA-B*27 has a well-established association with ankylosing spondylitis and related spondyloarthropathies (Reveille, 2012). These prior findings suggest that population differences in HLA composition are not merely descriptive but may have important implications for peptide repertoire breadth, immune responsiveness, and disease susceptibility. A similar pattern has been reported for ERAP1, where functional allotypes are determined by combinations of common missense variants, including rs30187 (K528R), rs27044 (Q730E), rs17482078 (R725Q), rs10050860 (D575N), and rs2287987 (M349V). These allotypes differ in peptide trimming efficiency and substrate specificity, providing a direct mechanistic link between genetic variation and antigen processing (Reeves et al., 2014). Notably, specific ERAP1 allotype combinations have been reported to be enriched in individuals with ankylosing spondylitis, particularly in the presence of HLA-B*27, supporting a gene–gene interaction between peptide trimming and peptide presentation (Reeves et al., 2014).

This literature is highly relevant to the interpretation of our findings because it illustrates that the phenotypic effect of APP variants may depend not only on their marginal frequency in a given population but also on the surrounding HLA background and disease context. Thus, differences observed across ancestries may reflect both population history and biologically meaningful epistatic effects.

ERAP2 provides another example in which population variation has clear functional consequences. ERAP2 haplotypes are largely defined by rs2248374 and linked variants; the rs2248374 G allele leads to loss of stable ERAP2 expression in homozygous carriers through nonsense-mediated decay, whereas the alternative haplotype permits protein expression. These haplotypes are maintained by balancing selection and vary substantially across populations (Andres et al., 2010). Functionally, ERAP2 expression status alters the peptide pool available for HLA class I loading, and has therefore been implicated in immune-mediated and infectious phenotypes (Andres et al., 2010). In relation to prior literature, this supports the view that ancestry-related differences in ERAP2 haplotype frequencies may contribute to differences in APP function between populations, even when the downstream phenotypic manifestations vary by disease setting.

Compared with HLA and ERAP genes, the literature on TAP1, TAP2, TAPBP, and PSMB8/PSMB9 is more heterogeneous and often more phenotype-specific, but still supports an important role for these loci in APP biology. TAP2 polymorphisms such as rs1800454 and rs241447 have been associated with rheumatoid arthritis risk in some populations (Dai et al., 2014), while TAP1 variants including rs1135216 and rs1057141 have been linked to asthma, allergic rhinitis, and dermatitis (Liu et al., 2017). These findings suggest that variation in peptide transport can influence a broad range of immune phenotypes, although the magnitude and consistency of associations appear to differ across ancestry groups and study designs. Similarly, TAPBP rs1059288 has been associated with cervical cancer susceptibility through effects on the regulation of tapasin expression (Hu et al., 2024), highlighting that APP variation may affect not only inflammatory disease but also tumour immune surveillance. Variants in the immunoproteasome genes PSMB8 and PSMB9, including PSMB9 rs17587, have also been associated with cancer phenotypes such as urothelial bladder carcinoma (Elhawary et al., 2023), further supporting the idea that altered peptide generation can have consequences across diverse clinical outcomes.

Compared with previous studies that focused on single components, our analysis provides a more comprehensive summary of the genetic features across the entire

pathway. In terms of allele frequency, our cohorts represent disease-specific populations, which may differ from summaries based on general population-level datasets. Regarding ERAP allotype composition, our results largely align with the haplotype structures observed in population studies, with the top three to four allotypes in each disease cohort showing similar patterns. However, we additionally incorporated the SNP rs72773968 (Thr12Ile) in the ERAP genes, which has not been considered in previous studies (Hutchinson et al., 2021). This variant shows notable population differences: it occurs more frequently in Europeans (approximately 15%) and South Asians (around 10%), but is much rarer in African populations, where the frequency is below 1%, according to the gnomAD and 1000 Genomes Project datasets. By including this SNP in our analysis, we identified a previously unreported ERAP allotype, IEPIGMKDRQ, which appears in South Asian (around 10%) and European populations (around 13%) but is absent in African populations. In addition, previous studies have not comprehensively characterised the distribution of tapasin dependence scores and HLA supertype frequencies across different ethnic groups. In this thesis, we therefore provide a systematic overview of these distributions, offering a broader perspective on population-level variation in antigen presentation–related genetic features.

Despite these strengths, some limitations should be noted. First, the pipeline does not yet incorporate structural information, such as 3D protein configurations or amino acid-level variation in HLA molecules, which could further refine functional predictions. Second, the use of imputed and phased data from the UK Biobank introduces potential inaccuracies, as phasing tools like Beagle can produce errors in haplotype inference, which may propagate into downstream analyses. Third, although our study combined multiple disease cohorts, some datasets (e.g. UK Biobank, STOP-HCV, and MalariaGEN) may share underlying imputation strategies. This raises the possibility of shared artefacts across cohorts, which could affect the generalisability of our results.

Future work should address these limitations by integrating structural bioinformatics methods, improving phasing accuracy through fully phased sequencing datasets, and expanding analyses to more diverse and underrepresented populations. Experimental validation, such as T-cell epitope mapping, would further confirm the biological relevance of our predicted associations. Incorporating multi-omics data, including transcriptomics and proteomics, would also provide deeper insights into how genetic variation influences functional immune responses.

In summary, this chapter demonstrates that genetic variation across the antigen presentation pathway contributes substantially to inter-individual differences in immune responses. By providing a flexible and scalable pipeline compatible with diverse imputation outputs, we established a comprehensive framework that unites HLA and non-HLA features into a single analytical strategy. Applied to multiple cohorts, this framework reveals both shared and ancestry-specific immunogenetic features, laying the groundwork for the association analyses presented in the next chapter, where we investigate their contribution to disease susceptibility and progression.

Chapter 3

A Bayesian Shrinkage Method to Enable Joint Analysis of Association across Antigen Presentation Pathway Variation

3.1 Introduction

In Chapter 2, we described the data processing pipeline and identified key genetic components of the antigen presentation pathway. The primary goal of this chapter is to explore how these components, including interactions between elements such as HLA class I alleles and ERAP1 and ERAP2 allotypes, influence clinical outcomes. A major challenge in this analysis arises from the strong linkage disequilibrium (LD)[1] within the HLA and ERAPs genomic regions. The resulting complex correlation structures often mask genuine genotype–phenotype associations, rendering marginal analyses insufficient to uncover underlying causal relationships.(Evseeva et al., 2010; Saulle et al., 2020).

To investigate the associations between genetic components, their interactions, and phenotypic outcomes, we employ joint association analysis. Unlike marginal analyses that assess variants individually, our objective is to model variables jointly using methods capable of capturing these complex dependencies directly. This task is complicated by LD and collinearity, which require fine-mapping approaches to separate causal variants from correlated non-causal signals. In this context, fine-mapping genetic variants within the antigen presentation pathway can be formulated as a high-dimensional sparse regression problem. We assume that only a small subset

[1]Linkage disequilibrium refers to the non-random association of alleles at different loci.

of genetic variants has biologically meaningful effects on immune response modulation.

A wide range of statistical methods has been developed for this purpose. Widely used tools include `SuSiE` (sum of single effects) (Zou et al., 2022), `FINEMAP` (Benner et al., 2016), and penalised regression techniques such as Lasso, Ridge, and Elastic Net (Ogutu et al., 2012). Bayesian approaches employ shrinkage priors, such as the Laplace, Gaussian, spike and slab (Ishwaran and Rao, 2005), and horseshoe priors (Piironen and Vehtari, 2017), to adaptively shrink small effect sizes toward zero while preserving larger, potentially causal effects. Each method offers distinct strengths and limitations. For example, `SuSiE` is not directly applicable to binary outcomes in logistic regression, and while `FINEMAP` achieves high accuracy, it can be computationally demanding. This work requires the use of individual-level genotype data rather than summary statistics to enable richer and more detailed modelling of genotype–phenotype relationships, albeit at a higher computational cost.

In this chapter, we present a novel method for the joint analysis of genetic variation across the APP using a continuous shrinkage prior, the regularised horseshoe prior. This framework enables the detection and estimation of specific genetic effects while jointly modelling variables to account for collinearity, including LD and both additive and non-additive components. Although the original horseshoe prior performs well in recovering sparse signals, high-dimensional LD structures can lead to multimodal posterior distributions, making inference through posterior means obtained by Markov chain Monte Carlo (MCMC) methods, such as `RStan`, computationally challenging. To overcome these limitations, we propose a maximum a posteriori (MAP) estimator implemented via permuted coordinate descent combined with binary search optimisation. This point-estimate approach avoids the convergence and mixing issues inherent to MCMC under strong collinearity and provides a computationally efficient and stable alternative for high-dimensional fine-mapping problems. We evaluate the performance of the proposed method through an extensive simulation study covering a range of scenarios in both linear and logistic regression. The results are compared with benchmark methods and alternative shrinkage priors to assess the advantages of the regularised horseshoe prior for identifying causal variants in both continuous and binary (case–control) outcomes.

Finally, we introduced `mapHS`, an R package implementing the permuted coordinate descent MAP estimator for the regularised horseshoe prior. The package supports both linear and logistic regression and will be applied to fine-mapping association analyses with real datasets in Chapter 4.

3.2 Bayesian inference

We aim to identify associations between a large number of genetic variables, specifically the genotype data extracted in Chapter 2, their possible interactions, and the phenotypes of interest. We assume that any number of these variables, from none to several, may truly influence the outcome. However, many of the variables are expected to be correlated with one another, either because of LD or because they originate from the same underlying genotype data, such as additive and non-additive genetic effects at the same locus. To establish this, we first consider a linear regression framework, which can be outlined as

$$\mathbf{y} = \mathbf{X}\,\boldsymbol{\beta}_{0:p} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}\big(\mathbf{0},\, \sigma^2 I_n\big)\,,$$

where $\mathbf{X}$ is an $n \times (p+1)$ design matrix, $\boldsymbol{\beta}_{0:p} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$ contains the intercept β_0 and regression coefficients $\beta_1, \ldots, \beta_p$, and σ^2 is the residual variance. Our goal is to estimate the coefficients $\boldsymbol{\beta}_{0:p}$ using Bayes' rule:

$$p(\boldsymbol{\beta}_{0:p} \mid \mathbf{X}, \mathbf{y}, \theta) = \frac{p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}_{0:p})\, p(\boldsymbol{\beta}_{0:p}, \mid \theta)}{p(\mathbf{y} \mid \mathbf{X}, \theta)} \quad \propto \quad p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}_{0:p})\, p(\boldsymbol{\beta}_{0:p} \mid \theta),$$

Where the likelihood is

$$p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}_{0:p}) = \prod_{i=1}^{n} \mathcal{N}\big(y_i \mid \mathbf{x}_i^\top \boldsymbol{\beta}_{0:p},\, \sigma^2\big),$$

and the prior can be factorised as $p(\boldsymbol{\beta}_{0:p} \mid \theta) = p(\beta_0)\, p(\boldsymbol{\beta}_{1:p} \mid \theta)$, with $\beta_0 \sim \mathcal{N}(0, 1)$. θ is the hyperparameter.

For concreteness, the model assumes that the outcome variable follows a Gaussian distribution whose mean depends on the combined effects of the genetic variables under study and a baseline mean. Each variable's contribution is governed by a parameter β, which quantifies the strength of its association with the outcome. Because a large number of parameters must be estimated simultaneously, it is often difficult to obtain stable estimates of β directly from the data. Moreover, in realistic biological settings, only a small subset of variables is expected to have strong, detectable effects. To address these challenges, we adopt a Bayesian framework that incorporates prior distributions reflecting plausible model structure. Several types of priors can be used for this purpose, but all share two key properties: they "shrink" coefficient estimates toward zero and "promote sparsity", meaning that only a subset of coefficients are typically estimated to be nonzero. These properties help prevent overfitting, which

would otherwise arise if the model were fitted directly to the data without regularisation.

Our study applies this framework to fine-mapping of genetic variants within the antigen presentation pathway to elucidate their associations with phenotypic outcomes. This constitutes a high-dimensional regression problem, typically involving thousands of genetic variables, and presents two main challenges. First, only a small proportion of coefficients β_j are likely to be truly predictive of the outcome. Second, the extensive correlation among genetic variants, particularly within regions such as HLA, complicates the identification of causal signals. The Bayesian framework, through the use of shrinkage priors, provides a principled and efficient approach for regularising such complex models.

3.2.1 Shrinkage priors

Shrinkage priors play a central role in Bayesian regression by shrinking negligible coefficients toward zero while allowing substantial effects to remain. Bayesian shrinkage approaches can be broadly categorised into two types: (i) two-component discrete mixture priors categorised as spike-and-slab priors (Mitchell and Beauchamp, 1988), and (ii) continuous shrinkage priors. In the following, we discuss these two classes separately, focusing on their prior structures and key properties.

3.2.2 Two-component discrete mixture priors

3.2.2.1 Spike and Slab

spike and slab prior is a Bayesian shrinkage prior commonly used for sparse estimation in high-dimensional regression. Each coefficient β_j is modeled as a two-component mixture: a "spike" representing near-zero values and a "slab" allowing large, nonzero effects:

$$\beta_j \mid \lambda_j, c, \varepsilon \sim \lambda_j \mathcal{N}(0, c^2) + (1 - \lambda_j)\mathcal{N}(0, \varepsilon^2), \quad \lambda_j \sim \text{Bern}(\pi), \quad j = 1, \ldots, p$$

where $\lambda_j = 1$ indicates inclusion in the model and π is the prior inclusion probability. The spike ($\varepsilon \ll c$) strongly shrinks small coefficients toward zero, and in the limiting case $\varepsilon = 0$ becomes a point mass at zero (δ_0), while the slab allows substantial nonzero effects.

The *spike and slab* prior effectively performs variable selection: coefficients are either excluded (assigned to the spike) or included (drawn from the slab). Introduced by Mitchell and Beauchamp (Mitchell and Beauchamp, 1988) and extended by George

and McCulloch (George and McCulloch, 1993), it is a foundational tool in Bayesian variable selection. Johnstone and Silverman (Johnstone and Silverman, 2004) further explored its connections to shrinkage rules and empirical Bayes methods.

3.2.3 Continuous shrinkage priors

A wide range of continuous shrinkage priors has been proposed for high-dimensional regression. A key advantage of these priors is that they enable inference through continuous techniques, thereby avoiding using high-dimensional variable search methods. In this section, we review some of the most popular continuous shrinkage priors, including the Laplace prior, the Gaussian prior, and the horseshoe prior, respectively.

3.2.3.1 Lasso regression

The ℓ_1-regularised regression, commonly known as Lasso (Least Absolute Shrinkage and Selection Operator), integrates least-squares fitting with an ℓ_1 penalty on the regression coefficients, formulated as

$$\min_{\boldsymbol{\beta}}\Big\{\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2+\lambda\sum_{j=1,}^{p}|\beta_j|\Big\},$$

Where $\lambda \geq 0$ is a tuning parameter that governs the trade-off between data fidelity and model sparsity. A larger λ induces stronger shrinkage, driving more coefficients to zero and thus performing variable selection, while a smaller λ allows more flexibility in the coefficients.

From a Bayesian perspective, Lasso corresponds to placing independent Laplace (double-exponential) priors on each coefficient:

$$\beta_j \sim \text{Laplace}(0,b)=\frac{1}{2b}\exp\Big(-\frac{|\beta_j|}{b}\Big),$$

where b controls the degree of shrinkage: smaller b implies stronger shrinkage. Setting $b=1/\lambda$ aligns the mode of the posterior distribution with the classical Lasso solution, unifying the penalised optimisation and Bayesian interpretations.

3.2.3.2 Ridge regression

Ridge regression modifies the ordinary least squares criterion by adding an ℓ_2 penalty:

$$\min_{\boldsymbol{\beta}}\Big\{\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2+\lambda\sum_{j=1}^{p}\beta_j^2\Big\},$$

Where $\lambda \geq 0$ controls the amount of shrinkage applied to the coefficients. Larger λ values result in greater shrinkage, yielding more regularised estimates.

From a Bayesian standpoint, ridge regression is equivalent to assigning independent Gaussian priors to the coefficients:

$$\beta_j \sim \mathcal{N}(0, \tau^2),$$

which leads to a Gaussian likelihood for $\boldsymbol{\beta}$. The posterior mode corresponds to the ridge estimator:

$$\hat{\boldsymbol{\beta}}_{\text{ridge}} = (\mathbf{X}^\top \mathbf{X} + \lambda I_p)^{-1} \mathbf{X}^\top \mathbf{y}.$$

3.2.3.3 Horseshoe regression

The horseshoe prior, which is the global-local shrinkage prior, exhibits strong theoretical and empirical properties (Carvalho et al., 2009). Each regression coefficient β_j is modelled as

$$\beta_j \mid \lambda_j, \tau \sim \mathcal{N}(0, \tau^2 \lambda_j^2), \quad \lambda_j \sim C^+(0, 1), \quad j = 1, \ldots, p,$$

where τ is a global shrinkage parameter controlling the overall level of shrinkage, and λ_j is a local parameter allowing each coefficient to vary individually. The heavy tails reduce bias for large coefficients, while the strong peak near zero aggressively shrinks irrelevant coefficients toward zero, effectively reducing noise.

Integrating out the local parameter λ_j, the distribution of β_j can be expressed as a scale mixture of normals:

$$p(\beta_j \mid \tau) = \int_0^\infty \mathcal{N}(\beta_j \mid 0, \tau^2 \lambda_j^2)\, p(\lambda_j)\, \mathrm{d}\lambda_j.$$

In statistical applications, the singularity of the original horseshoe prior at zero (where $p(\beta_j) \to \infty$ as $\beta_j \to 0$) can cause computational instability in estimation and sampling. To address this issue, the regularised horseshoe prior (Piironen and Vehtari, 2017) introduces controlled shrinkage through a continuous prior formulation that maintains the desirable theoretical properties of the original horseshoe while resolving numerical challenges. The regularised horseshoe prior is formally specified as:

$$\beta_j \mid \lambda_j, \tau, c \sim \mathcal{N}\big(0,\, \tau^2 \tilde{\lambda}_j^2\big),$$

where the modified shrinkage weights are defined as

$$\tilde{\lambda}_j^2 = \frac{c^2 \lambda_j^2}{c^2 + \tau^2 \lambda_j^2},$$

incorporating both the local shrinkage parameters $\lambda_j \sim C^+(0,1)$ and a hyperparameter slab width $c > 0$. Consequently, the marginal prior distribution can be written as:

$$p(\beta_j \mid \tau, c) = \int_0^\infty \mathcal{N}(\beta_j \mid 0, \tau^2 \tilde{\lambda}_j^2)\, p(\lambda_j)\, \mathrm{d}\lambda_j.$$

While the original horseshoe prior is well-behaved for its strong theoretical performance in sparse settings, its practical application can be hampered by two issues: first, a singularity at zero that can lead to computational instability during MCMC sampling, and second, the potential for unboundedly large estimates for true signals in the tails of the prior, which increases the risk of over-fitting. The primary motivation behind the regularised horseshoe prior is to engineer a solution to these problems while preserving the desirable shrinkage properties of the original. To this end, Piironen et al. introduced a regularising slab component with a hyperparameter c, which explicitly bounds the magnitude of large coefficients (Piironen and Vehtari, 2017). In their analysis, they demonstrated that modification successfully eliminated the pathological behaviour of the posterior, leading to substantially improved sampling efficiency and numerical stability. Furthermore, they showed through simulations that the regularised horseshoe provides more robust estimation and favourable predictive performance, particularly in settings that deviate from pure sparsity, by effectively preventing implausibly large coefficient estimates without sacrificing its ability to recover true signals.

3.2.4 Comparison of continuous shrinkage priors

As discussed analytically above, different priors impose distinct penalties on regression coefficients, affecting both the shrinkage of small values and retention of large values. Figure 3.1 illustrates this comparison among Laplace, Gaussian, and the horseshoe prior. In the left panel, which shows the prior densities, the horseshoe prior shrinks the small coefficients more aggressively than the Ridge and Lasso priors, while its heavy tails for large coefficients result in a more moderate penalty than the Lasso and Ridge. The right panel shows the corresponding penalty functions, highlighting these differences more clearly.

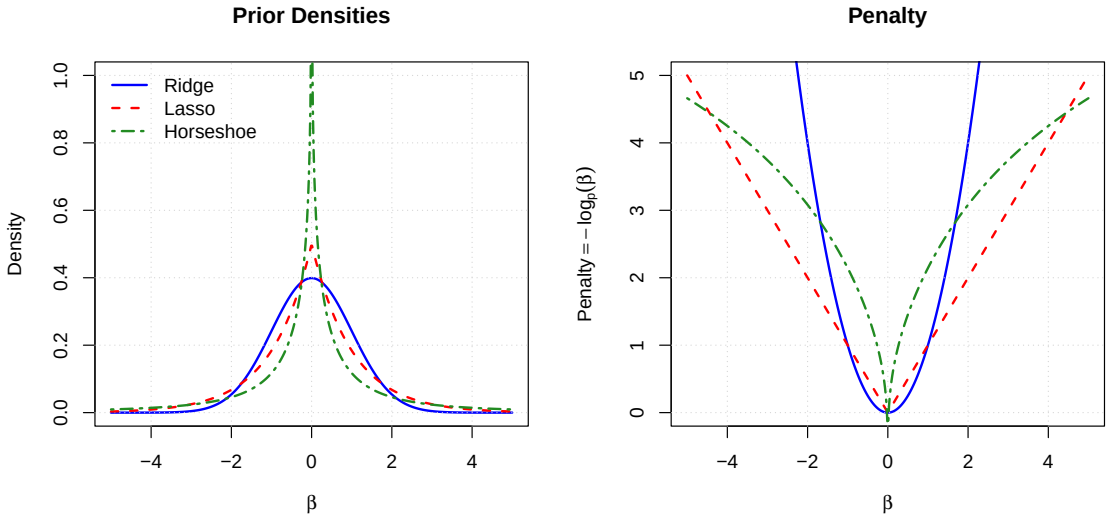


Figure 3.1: **Comparison of Laplace, Gaussian, and Horseshoe priors**. (Left) Prior densities. (Right) Corresponding penalty functions. All priors are shown with their tuning parameter set to 1 ($\lambda = 1$ for Laplace and Gaussian; $\tau = 1$ for the horseshoe).

Table 3.2 summarises the asymptotic behaviour of four widely used priors, showing how each concentrates probability near zero and how rapidly its tails decay. Small coefficients are strongly penalised by priors with high concentration at zero, while heavy-tailed priors impose moderate penalties on large coefficients. Each prior represents a different trade-off between shrinkage and variable selection. In our high-dimensional application, we aim to detect associations between genetic variations in the antigen presentation pathway and phenotypes. The ideal prior should apply strong shrinkage to small coefficients, which are likely to represent the lack of a true effect, while imposing at most penalties on true effects. Among the priors considered, the regularised horseshoe is particularly well-suited to achieve this balance, making it potentially making it a good choice for our analysis.

Having outlined the theoretical foundations of shrinkage priors, we now turn to the practical development of the MAP estimator under the regularised horseshoe prior. Section 3.3 details the optimisation framework, providing the computational machinery necessary to evaluate the estimator's performance in both simulated and real-world datasets.

Prior	Concentration at zero	Tail behaviour
Laplace	$O(1)$	$O(e^{-\lvert\beta_j\rvert/b})$
Gaussian	$O(1)$	$O(e^{-\beta_j^2/(2\sigma^2)})$
Horseshoe	$O(\log^{-1}\lvert\beta_j\rvert)$	$O(\beta_j^{-2})$
Regularised Horseshoe	$O(1)$	$O(\beta_j^{-2})$

Table 3.2: Concentration around zero and tail behaviour for Lasso, Ridge, Horseshoe, and Regularised Horseshoe priors.

3.2.5 Tuning the hyperparameter for the regularised horseshoe prior

In our analysis, we investigate the global hyperparameter τ, which controls the overall shrinkage strength in the regularised horseshoe prior. To quantify its impact, we sample 10,000 times from the prior under three values of τ: 1, 0.1, and 0.01, representing weak, moderate, and strong global shrinkage, respectively. For each value, we calculate the proportion of samples falling within the region $(-0.05, 0.05)$, which approximates the prior probability that a coefficient is effectively shrunk to zero. The corresponding probabilities for $\tau = 1$, 0.1, and 0.01 were 0.13, 0.5, and 0.98, respectively (Figure 3.2).

While a conservative value like $\tau = 0.01$ might be ideal in a standard GWAS under the assumption that most variants are irrelevant, our study context necessitates a different approach. Previous research has established that genetic components within the antigen presentation pathway, such as HLA alleles, are associated with many phenotypes. Furthermore, we aim to explore interaction terms between HLA alleles and ERAP1/2 allotypes, which are inherently rarer and may have many effect frequencies. An overly conservative global shrinkage with $\tau = 0.01$ might falsely shrink these genuine, albeit weak, interaction signals. Therefore, to balance strong sparsity assumptions with the known biology and the goal of detecting interactions, we selected a moderate prior probability of 0.5 by setting $\tau = 0.1$. This provides a 50% prior probability that any given coefficient is negligible, offering a robust default that is neither overly sceptical nor overly permissive.

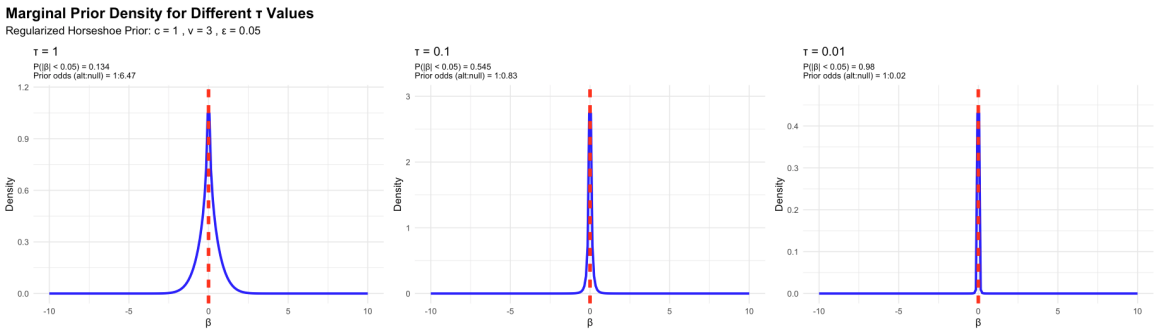


Figure 3.2: **The Global Shrinkage Hyperparameter τ for the Regularised Horseshoe Prior.** The plot illustrates the prior density and the resulting proportion of coefficients shrunk near zero for different values of τ. We compute the prior probability that a coefficient falls within $(-0.05, 0.05)$ based on 10,000 samples from the prior. The probabilities for $\tau = 1$, 0.1, and 0.01 are 0.13, 0.5, and 0.98, respectively, demonstrating how τ controls the global shrinkage and determines the prior proportion of coefficients close to zero.

3.3 Method and algorithm

3.3.1 Bayesian inference with the regularised horseshoe prior using `RStan`

We conduct a simulation study to evaluate the performance of Bayesian logistic regression with a regularised horseshoe prior, implemented using the `RStan` package. The specific Stan model code is provided in Appendix C. The simulated data consisted of 40 variables X, each encoded as 0, 1, or 2. The first two variables are highly correlated with a correlation coefficient of 0.8. The true coefficients are set to $\beta_1 = 0.2$ and $\beta_j = 0$ for $j = 2, \ldots, 40$. The response is generated from a Bernoulli distribution according to the following model:

$$y_i \sim \text{Bernoulli}\big(\sigma(\mathbf{x}_i^\top \boldsymbol{\beta}_{1:p} + \beta_0)\big),$$

with $\sigma(\cdot)$ denoting the logistic sigmoid function.

We perform MCMC sampling using the `sampling` function in `RStan` with 4 chains, 2000 iterations per chain, and a refresh rate of 500.

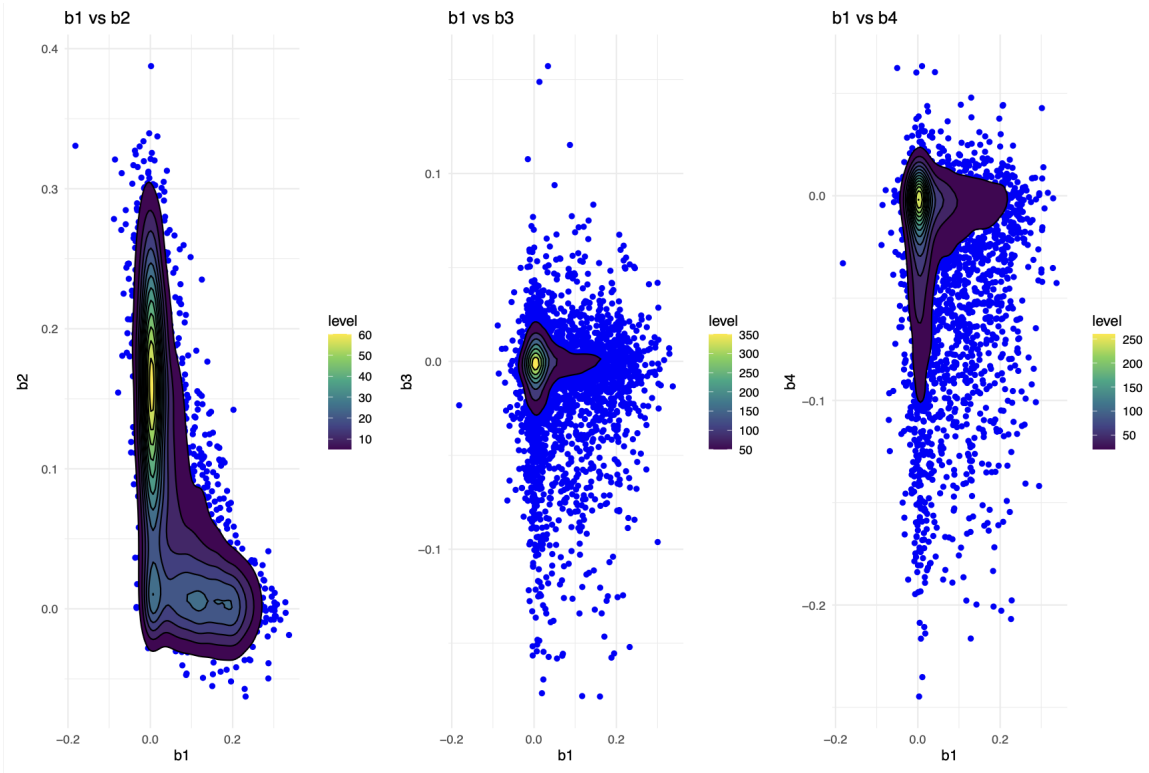


Figure 3.3: **Posterior distributions of coefficients from the regularised horseshoe prior.** Left: Joint posterior of β_1 and β_2 showing bimodality due to high correlation between the first two variables. Middle: Joint posterior of β_1 and β_3. Right: Joint posterior of β_2 and β_3.

The posterior distributions of the first three coefficients are shown in Figure 3.3. The left panel reveals a bimodal posterior distribution for β_1 and β_2, with one mode near $(\beta_1 = 0.2, \beta_2 = 0)$ and another near $(\beta_1 = 0, \beta_2 = 0.2)$. This bi-modality arises from the high correlation between the first two variables, making it difficult for the model to distinguish their individual effects. Since the `RStan` algorithm typically targets the posterior mean rather than the posterior mode, the resulting estimates may be inaccurate in such multimodal settings. One potential solution is to estimate the full posterior curve, although the heavy tails of the Cauchy distribution in the horseshoe prior can lead to instability if the number of MCMC samples is insufficient. An alternative approach is to modify the inference procedure to target the posterior mode directly.

3.3.2 Bayesian inference with the regularised horseshoe prior using maximum a posteriori estimation

The *Maximum A Posteriori (MAP) estimator* of $\boldsymbol{\beta}$ is defined as the parameter value that maximises the posterior density of $\boldsymbol{\beta}$ given the observed data:

$$\hat{\boldsymbol{\beta}}_{\text{MAP}} = \arg\max_{\boldsymbol{\beta}_{0:n}} p(\boldsymbol{\beta}_{0:n} \mid \boldsymbol{X}, \boldsymbol{Y}, \tau, c) = \arg\max_{\boldsymbol{\beta}_{0:n}} p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\beta}_{0:n})\, p(\boldsymbol{\beta}_{0:n} \mid \tau, c),$$

where $p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\beta}_{0:n})$ is the *likelihood*, and $p(\boldsymbol{\beta}_{0:n})$ is the *prior distribution*, which factorises as

$$p(\boldsymbol{\beta}_{0:n} \mid \tau, c) = p(\beta_0)\, p(\boldsymbol{\beta}_{1:n} \mid \tau, c).$$

For the intercept term, we assume the prior distribution.

$$\beta_0 \sim \mathcal{N}(0, 1).$$

For the regression coefficients $\boldsymbol{\beta}_{1:n}$, we employ the *regularised horseshoe prior*:

$$\beta_j \mid \lambda_j, \tau, c \sim \mathcal{N}\Big(0,\, \tau^2 \tilde{\lambda}_j^2\Big), \qquad \tilde{\lambda}_j^2 = \frac{c^2 \lambda_j^2}{c^2 + \tau^2 \lambda_j^2}, \qquad \lambda_j \sim C^+(0, 1), \quad j = 1, \ldots, n,$$

where $\tau > 0$ is the global shrinkage hyperparameter, $c > 0$ is a hyperparameter controlling the slab width, and $C^+(0, 1)$ denotes the standard half-Cauchy distribution. The corresponding *marginal prior* for each coefficient β_j can be expressed as

$$p(\beta_j \mid \tau, c) = \int_0^\infty p\Big(\beta_j \,\Big|\, 0, \tau^2 \tilde{\lambda}_j^2\Big)\; p(\lambda_j)\, d\lambda_j.$$

In the following, we discuss the likelihood separately for the cases of *linear regression* and *logistic regression*.

Linear regression (Gaussian likelihood). For continuous outcomes, the model is

$$Y_i = \beta_0 + X_i^T \boldsymbol{\beta}_{1:n} + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \quad i = 1, \ldots, n.$$

The corresponding log-likelihood is

$$\log p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\beta}_{0:n}) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \Big(Y_i - \beta_0 - X_i^T \boldsymbol{\beta}_{1:n}\Big)^2.$$

Logistic regression (Bernoulli likelihood). For binary outcomes $Y_i \in \{0, 1\}$, the model is

$$\Pr(Y_i = 1 \mid X_i, \boldsymbol{\beta}_{0:n}) = \sigma\big(\beta_0 + X_i^T \boldsymbol{\beta}_{1:n}\big), \qquad \sigma(z) = \frac{1}{1 + e^{-z}}, \quad i = 1, \ldots, n.$$

The corresponding log-likelihood is

$$\log p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\beta}_{0:n}) = \sum_{i=1}^{n} \Big[Y_i \log \sigma(\beta_0 + X_i^T \boldsymbol{\beta}_{1:n}) + (1 - Y_i) \log \big(1 - \sigma(\beta_0 + X_i^T \boldsymbol{\beta}_{1:n})\big) \Big].$$

3.3.3 Optimisation

Traditional gradient-based optimisers such as `BFGS` (Gerber and Furrer, 2019) could find it hard to estimate the posterior mode in the horseshoe due to the steep curvature near zero and the heavy-tailed behaviour of the posterior, which can lead to instability and slow or unreliable convergence. To efficiently compute the MAP estimate under the regularised horseshoe prior, we employ a coordinate descent algorithm combined with binary search for each parameter update, as outlined in Algorithm 1. This approach might work better for the regularised horseshoe, which induces a highly non-convex and sharply peaked posterior landscape, as we evaluate below.

Algorithm 1 MAP estimate under the regularised horseshoe prior with coordinate descent algorithm

Input: Data $(\boldsymbol{Y}, \boldsymbol{X})$; regularised horseshoe hyperparameter (τ, c); number of permutations K; stopping criteria: maximum iterations T, convergence threshold ϵ; $\beta_0 \sim \mathcal{N}(0, 1)$

Output: $\boldsymbol{\beta}^*_{0:n}$

Initialise $\boldsymbol{\beta}^*_{0:n} \leftarrow 0$, maximum posterior $\mathcal{P}^* \leftarrow -\infty$

for $k = 1$ **to** K **do**
- Randomly permute feature order in $X \rightarrow X^{(k)}$
- Initialise $\boldsymbol{\beta}_{0:n} \leftarrow 0$
- **for** $t = 1$ **to** T **do**
  - $\boldsymbol{\beta}_{\text{old}} \leftarrow \boldsymbol{\beta}_{0:n}$
  - **for** *each coordinate* j **do**
    - Define objective:
      $f(b_j) = -\log p(\boldsymbol{Y} \mid X^{(k)}, \boldsymbol{\beta}_{0:n} \text{ with } \beta_j = b_j) - \log p(\beta_j)$
    - Use binary search to minimise $f(b_j)$
    - Update β_j with the optimal value
  - **end**
  - **if** $\|\boldsymbol{\beta}_{0:n} - \boldsymbol{\beta}_{old}\| < \epsilon$ **then**
    - **break** (converged)
  - **end**
- **end**
- Reorder $\boldsymbol{\beta}_{0:n}$ back to the original feature order
- Evaluate posterior $\mathcal{P}$ with current $\boldsymbol{\beta}_{0:n}$
- **if** $\mathcal{P} > \mathcal{P}^*$ **then**
  - $\boldsymbol{\beta}^*_{0:n} \leftarrow \boldsymbol{\beta}_{0:n}$
  - $\mathcal{P}^* \leftarrow \mathcal{P}$
- **end**

end

return $\boldsymbol{\beta}^*_{0:n}$

Coordinate descent simplifies the high-dimensional optimisation problem by updating one parameter at a time while holding the others fixed. Within each update, binary search provides a robust, derivative-free method for locating the local minimum of the one-dimensional objective function, which consists of the likelihood plus the log-prior. This is especially advantageous when the prior is non-differentiable or exhibits rapidly varying gradients near zero, as is the case with the horseshoe prior.

To further mitigate the risk of convergence to local maxima, we introduce random permutations of the feature order across multiple restarts. This increases the likelihood of finding a globally optimal or near-optimal solution.

3.3.4 Efficient Uncertainty Quantification

To quantify the uncertainty of the coefficients estimated from the Bayesian posterior, we initially seek to compute the standard errors via the Fisher information matrix, which requires the analytical Hessian matrix. However, the analytical expression derived in Appendix D is conditional on the latent variables λ_i, which cannot be analytically integrated out. This analytical intractability necessitated a numerical approach for Hessian computation.

Given the high-dimensionality of our problem, the direct numerical calculation of the full Hessian matrix remains computationally prohibitive. To address this challenge, we employ the active subspace method (ASM). As the horseshoe prior induces a strong sparsity pattern in the posterior. This prior strongly shrinks the coefficients of most irrelevant variables to values near zero, while allowing a few relevant variables to retain large coefficients. Consequently, the log-posterior's curvature is dominated by a low-dimensional manifold defined by these few large-coefficient variables. The ASM efficiently identifies and exploits this structure by focusing the numerical Hessian calculation on the active subspace, the low-dimensional linear subspace where the function exhibits the most significant change, thereby reducing the computational burden without a substantive loss of information.

The implementation of the Active Subspace Method begins by identifying the sparse set of active coefficients. We define a threshold, ϵ, such that any coefficient m_i with a magnitude $|m_i| < \epsilon$ is deemed inactive. This threshold is chosen relative to the scale of the posterior estimates; a common practice is to set ϵ to a small fraction of the maximum absolute coefficient value, for instance.

The indices of the coefficients exceeding this threshold form the active set, $\mathcal{A} = i : |m_i| \geq \epsilon$. The remaining parameters are considered passive and are held fixed at their posterior mean values. The numerical Hessian is then computed only within the resulting low-dimensional subspace defined by $\mathcal{A}$. This involves calculating the partial derivatives of the log-posterior with respect to the parameters m_j for all $j \in \mathcal{A}$, while the passive variables remain constant. The resulting matrix, $\mathbf{H}_{\mathcal{A}\mathcal{A}}$, is a dense Hessian approximation of the full log-posterior but of a drastically reduced size $|\mathcal{A}| \times |\mathcal{A}|$.

Finally, the uncertainty estimates for the active coefficients are obtained by inverting this low-dimensional Hessian to approximate the relevant sub-block of the

full covariance matrix, $\Sigma_{\mathcal{A}\mathcal{A}} \approx -\mathbf{H}^{-1}_{\mathcal{A}\mathcal{A}}$. The standard errors for the coefficients in the active set are then given by the square roots of the diagonal elements of this matrix. This strategy confers a significant computational advantage, reducing the complexity of the Hessian calculation from $O(p^2)$ for the full model to $O(|\mathcal{A}|^2)$ for the active subspace.

3.4 Simulation

In this subsection, to assess the performance of the proposed method, we extend the evaluation of regression coefficient performance across different approaches under a range of genetic architectures, including single genetic models with additive effects only, models incorporating pairwise gene–gene interactions, and multi-genetic models with additive, dominant, and recessive encodings. Results are presented separately for logistic regression (binary outcomes) and linear regression (continuous traits). The simulation scenarios are summarised in Table 3.3.

Table 3.3: **Summary of Simulation Scenarios for Performance Evaluation**

Logistic Regression	**Linear Regression**
Single Genetics Model	Single Genetics Model
Single Genetic Model with Interaction Term	Single Genetic Model with Interaction Term
Multi-genetic Model: Additive/Dominant/Recessive	Multi-genetic Model: Additive/Dominant/Recessive
Real Data with MalariaGEN: Multi-genetic Effects and Interaction	/

In the simulation study, we deliberately varied several key design parameters to evaluate the behaviour and robustness of the competing statistical methods under controlled and interpretable scenarios. The simulations were designed to begin with relatively simple settings to isolate specific methodological properties before gradually increasing complexity.

To examine the impact of correlated predictors, we simulated scenarios in which one variable had a true non-zero effect on the phenotype while a second variable was highly correlated with it but had no true effect. This structure allows us to assess each method's ability to distinguish a causal predictor from a correlated non-causal variable, reflecting the linkage disequilibrium patterns commonly observed in the HLA region of genetic data. For these baseline simulations, we used 3,000 observations,

which provides adequate power to detect moderate effect sizes while maintaining a realistic sample size for many genetic studies. Similarly, when evaluating interaction effects, we constructed scenarios in which only one interaction term had a true non-zero effect. Starting with a single true interaction enables us to assess each method's ability to recover sparse interaction signals without inflating false positives. This approach ensures that differences in performance primarily reflect shrinkage behaviour and variable selection properties rather than excessive model complexity.

To investigate different modes of genetic inheritance (additive, dominant, and recessive effects), we increased the sample size from 3,000 to 10,000 observations. Dominant and especially recessive genetic models reduce the effective number of informative observations. In particular, under a recessive model, only individuals homozygous for the minor allele contribute directly to the effect estimate, which substantially reduces statistical power when allele frequencies are modest. Increasing the sample size, therefore, ensures a fair comparison across inheritance models and prevents systematic bias against methods in low-information scenarios.

Finally, we simulated variants with a MAF of approximately 10%, representing common variants typically included in genome-wide association studies. A MAF around 10% provides a balance between realism and statistical stability: it ensures sufficient observations in each genotype category for additive, dominant, and recessive encodings, while avoiding the instability and extreme sparsity associated with rare variants.

3.4.1 Simulation Study with Correlated Predictors

We design a controlled simulation to examine the impact of LD in logistic regression. The dataset consists of $n = 3{,}000$ observations and $p = 40$ predictors, each encoded as $\{0, 1, 2\}$. All genotypes are constrained to have a MAF of at least 10%, reflecting realistic distributions of common alleles.

To simulate strong LD, we induce a Pearson correlation greater than 0.8 between the first two predictors, while ensuring all other pairwise correlations remained below 0.3. The true coefficient vector is specified as

$$\boldsymbol{\beta}_{1:n} = \{0.2,\, 0,\, 0, \ldots, 0\},$$

where β_0 is the intercept and only β_1 has a nonzero effect. The binary outcomes are generated as

$$y_i \sim \text{Bernoulli}\big(\sigma(\mathbf{x}_i^\top \boldsymbol{\beta}_{0:n})\big),$$

with $\sigma(\cdot)$ denoting the logistic sigmoid function.

We then fit the regularised horseshoe model with hyperparameter $\tau = 0.1$ and $c = 1$ on a single replicate of this dataset. All coefficients are fixed at their true values except for β_1 and β_2. We evaluate the marginal log-prior, log-likelihood, and log-posterior surfaces over a grid of (β_1, β_2) values (Figure 3.4).

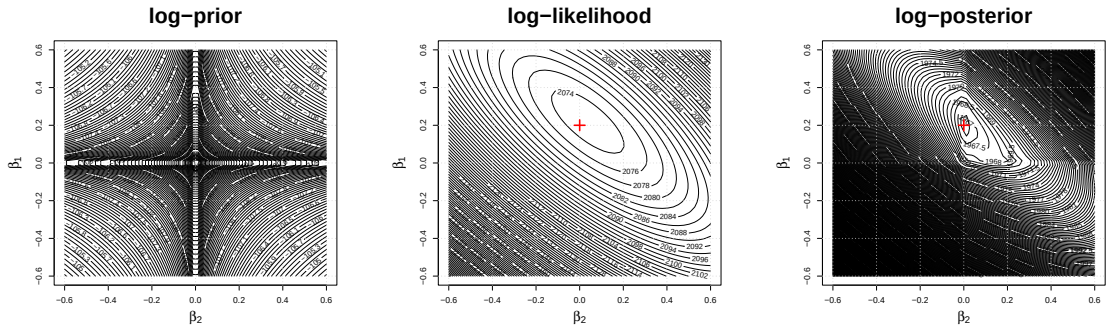


Figure 3.4: Contour plots of the marginal log-prior, log-likelihood, and log-posterior surfaces for (β_1, β_2). The red plus sign indicates the true parameter values.

The grid structure in the log-likelihood highlights the strong correlation between β_1 and β_2, and the resulting log-posterior surface exhibits two symmetric modes at $(0.2, 0)$ and $(0, 0.2)$. This multimodality poses a key challenge: gradient-based optimisers and posterior-mean estimators often perform poorly in multimodal landscapes. Standard MCMC samplers may mix slowly between modes, and gradient ascent methods could potentially converge to suboptimal local maxima.

By contrast, our implementation of a coordinate descent algorithm, combined with one-dimensional binary search for each coordinate update, successfully identifies the global MAP solution. To further mitigate convergence to local maxima, we also randomise the order of coordinate updates via a permutation strategy. Full details of the algorithm are provided in the Methods section.

3.4.2 Logistic regression

3.4.2.1 Single genetic model

We evaluate estimator performance on a single additive genetic scenario with $n =$ 3,000 samples and 40 SNPs with $p = 40$, each encoded as $\{0, 1, 2\}$ with MAF $\geq$ 10%. To induce linkage disequilibrium, SNPs x_1 and x_2 are simulated with Pearson correlation around 0.8, while all other pairwise correlations are below 0.3. The true coefficient vector is

$$\boldsymbol{\beta}_{1:n} = (0.2, 0, \ldots, 0)^T,$$

so that only the first SNP carried a non-zero effect. Binary outcomes are generated via

$$y_i \sim \text{Bernoulli}\big(\sigma(\mathbf{x}_i^\top \boldsymbol{\beta}_{1:p} + \beta_0)\big),$$

where $\sigma(\cdot)$ is the logistic link, and the entire procedure is repeated for 100 independent replicates.

We fit the simulated data using 5 approaches: marginal GLM, joint MLE, Bayesian Lasso, Bayesian ridge, and Bayesian regularised horseshoe via permuted coordinate descent. To ensure comparability, we set the Laplace prior with $b = 0.1$, the Gaussian prior with SD $= 0.1$ and the regularised horseshoe hyperparameter to $\tau = 0.1$ and $c = 1$. Results over 100 replicates are shown in Figure 3.5. The marginal GLM detects two signals, with the first coefficient around 0.2 and the second a false positive. Joint MLE avoids false discoveries but exhibits excessive variance across all coefficients. Bayesian Ridge regression over-penalised the first variable and under-penalised others, e.g., the second variable retained a coefficient of 0.05. In contrast, the Laplace prior and regularised horseshoe tightly concentrated non-zero estimates and shrank null coefficients toward zero. The regularised horseshoe provides an estimate for β_1 closest to the true value of 0.2, consistent with Section 3.2: its polynomial decay allows large coefficients to persist, while the Laplace prior's exponential decay imposes stronger shrinkage.

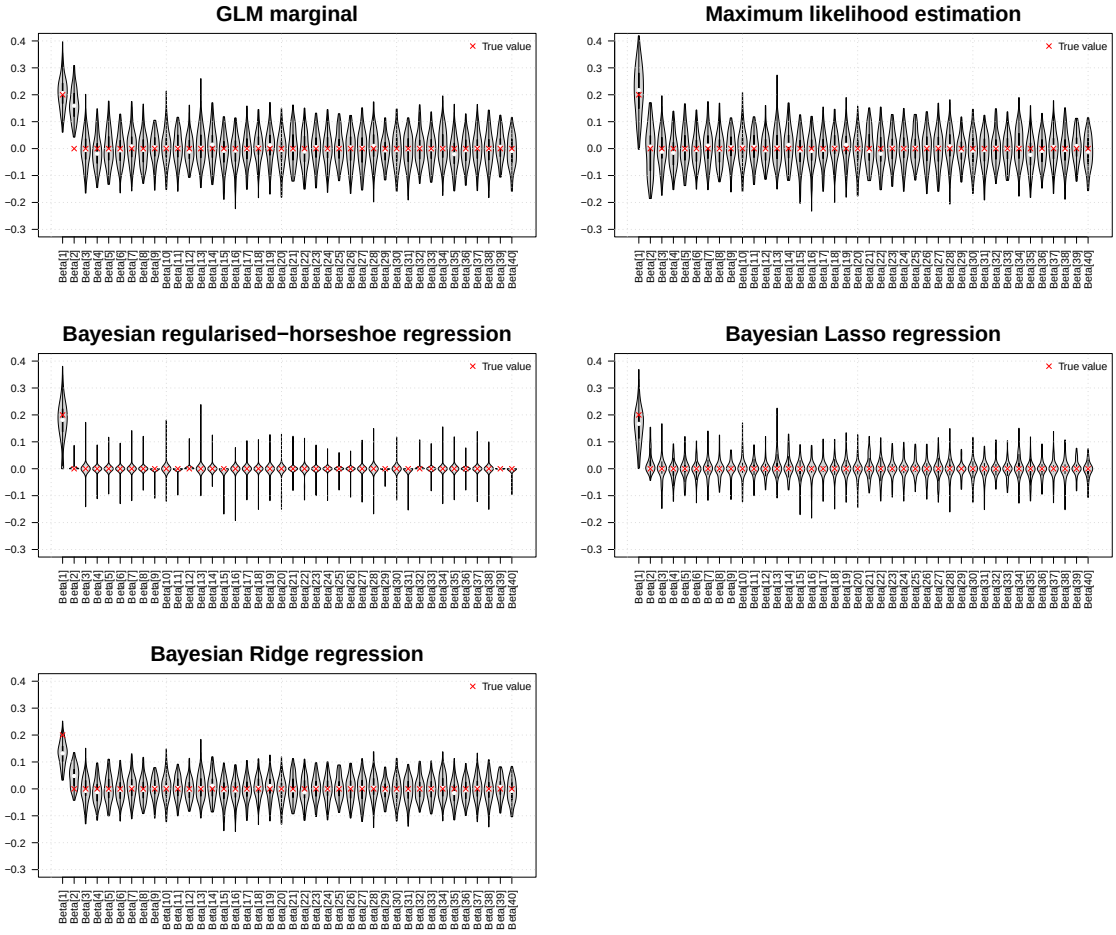


Figure 3.5: Results of coefficient estimates from a single genetic model for binary outcomes (logistic regression) across simulated replicates using five methods: marginal GLM, joint MLE, Bayesian Lasso, Bayesian Ridge, and Bayesian regularised horseshoe.

We calculate the Mean Squared Error (MSE) to assess the accuracy of coefficient estimates, averaging the squared differences between estimated values across models. The results are shown in Figure 3.6. Overall, the joint model outperformed the marginal GLM. All Bayesian approaches, Lasso, Ridge, and regularised horseshoe, further improve accuracy through shrinkage priors. Among them, the regularised horseshoe achieved the lowest MSE, followed by Lasso, highlighting their effectiveness in recovering true signals.

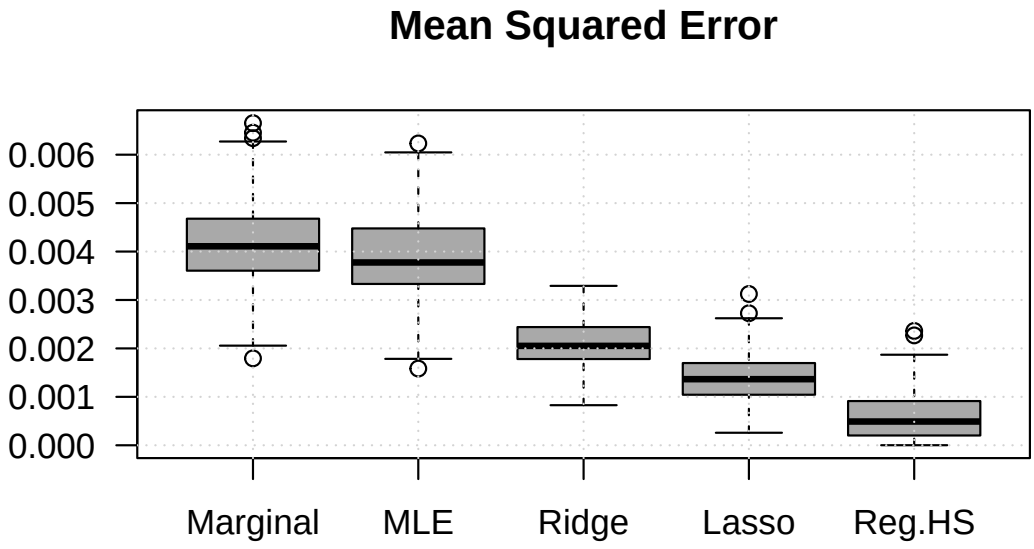


Figure 3.6: Comparison of mean squared error across simulated replicates for the single genetic model with binary outcomes using logistic regression.

To further evaluate each method's ability to detect sparse gene–gene interactions in a binary outcome setting, we simulate $n = 3{,}000$ observations on $p = 10$ SNPs (minor-allele frequency $\geq 10\%$) and included all $\binom{10}{2}$ pairwise interaction terms. Each SNP is coded additively as $\{0, 1, 2\}$, and the design matrix is expanded to include every product $x_j \times x_k$. The true coefficient vector is specified so that only the first interaction term, $\beta_{1\times 2}$, is nonzero ($\beta_{1\times 2} = 0.2$), with all other main and interaction effects set to zero. Binary responses are then generated according to

$$y_i \sim \text{Bernoulli}\left(\sigma\left(\sum_{j=1}^{p}\beta_j x_{ij} + \sum_{1\leq j<k\leq p}\beta_{j\times k}(x_{ij}x_{ik}) + \beta_0\right)\right),$$

where $\sigma(\cdot)$ denotes the logistic link function. This simulation is repeated for 100 replicates.

Again, we fit the data using five approaches: marginal GLM regression, joint MLE, Bayesian Lasso, Bayesian Ridge, and Bayesian regularised horseshoe MAP. The results are shown in Figure 3.7. The marginal GLM selects three signals, x_1, x_2, and the true interaction, introducing two false positives. The joint MLE correctly detects the true interaction but shows high variance across all coefficients. In contrast, Bayesian methods with shrinkage priors effectively penalise irrelevant variables while

estimating the interaction term. Both the Bayesian Lasso and regularised horseshoe produce estimates closer to the true value (0.2) and shrink null coefficients more strongly than Bayesian Ridge. Between them, the inference with the regularised horseshoe prior estimate is closest to the truth.

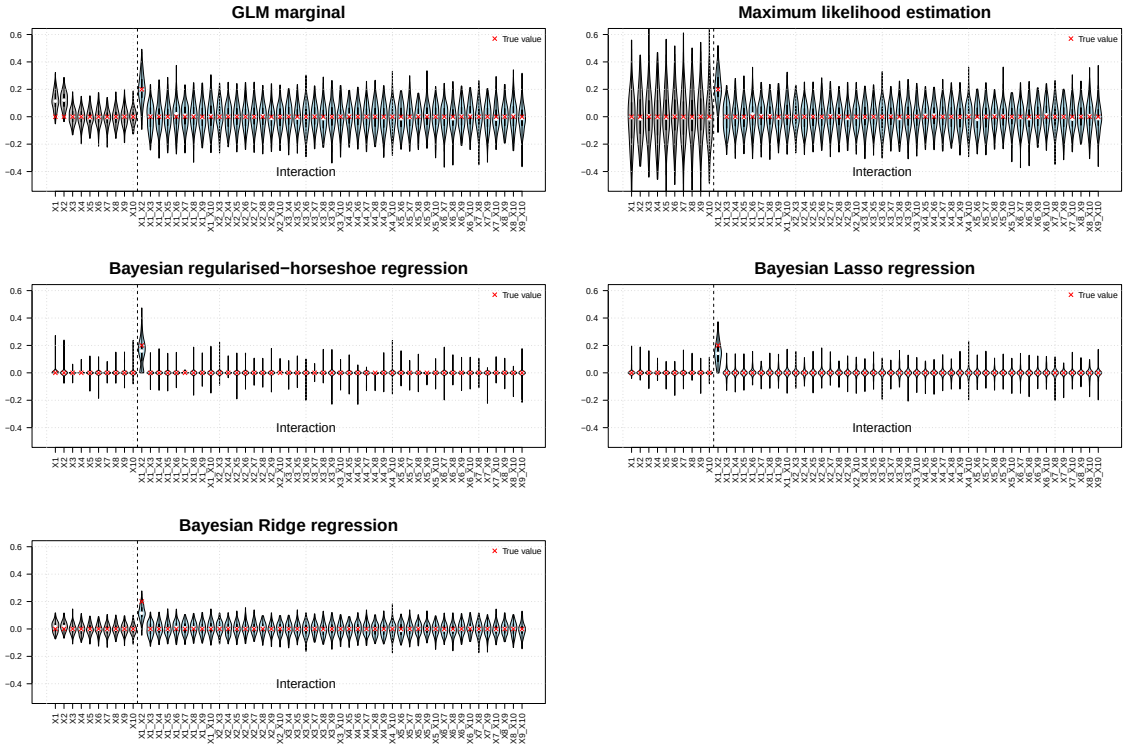


Figure 3.7: Results of coefficient estimates from a single genetic model with interaction term for binary outcomes (logistic regression) across simulated replicates using five methods: marginal GLM, joint MLE, Bayesian Lasso, Bayesian Ridge, and Bayesian regularised horseshoe.

As before, we compare the MSE across methods, as shown in Figure 3.8. The joint MLE is the worst performer, producing the highest MSE. This poor performance is likely due to substantial variability in its coefficient estimates, particularly for the first 10 variables representing the main SNP effects, as seen in Figure 3.7, which greatly inflates squared error. The marginal GLM has the second-highest MSE, reflecting its inclusion of false positives. Among the Bayesian approaches, Ridge regression performs better but still exceeds the MSE of the Bayesian Lasso, while the regularised horseshoe achieves the lowest MSE overall. The strong performance of the Bayesian shrinkage methods, especially the regularised horseshoe, arises from their ability to retain the true interaction effect while shrinking irrelevant coefficients toward zero.

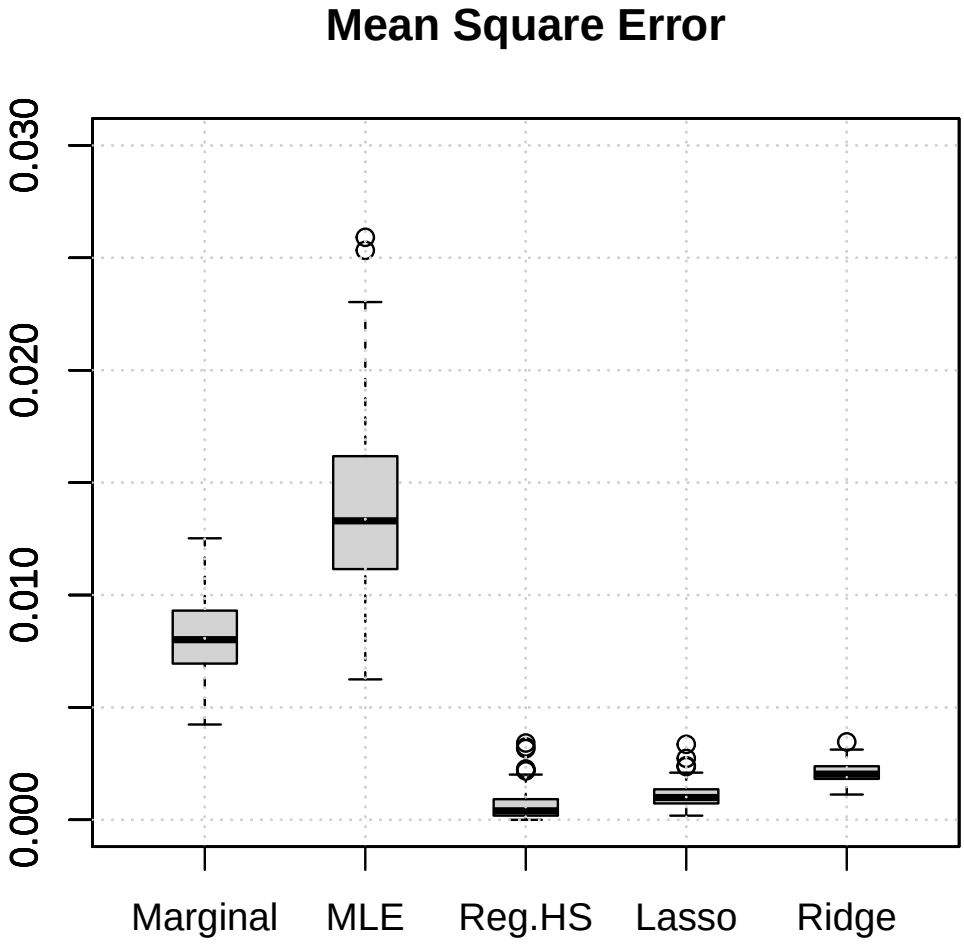


Figure 3.8: Comparison of mean squared error across simulated replicates for the single genetic model with interaction term with binary outcomes using logistic regression.

3.4.2.2 Multi-genetic Model

In practice, the relationship between genotype and phenotype may follow more complex inheritance patterns than a purely additive model. To evaluate the methods' ability to identify the most appropriate genetic encoding, we simulate a "multi-genetic" scenario incorporating additive, dominant, and recessive effects. We generate $n = 10{,}000$ observations on $p = 10$ SNPs, each represent in three coding schemes: additive $(0, 1, 2)$, dominant $\mathbb{I}\{x \geq 1\}$, and recessive $\mathbb{I}\{x = 2\}$. Here, $\mathbb{I}\{\cdot\}$ denotes the

indicator function, which equals 1 when the stated condition is true and 0 otherwise. These are concatenated to form the design matrix.

$$X_{\text{multi}} = \left[X_{\text{add}} \mid X_{\text{dom}} \mid X_{\text{rec}} \right], .$$

where the dimensionality of each block is $n \times p$. The true coefficient vector is set so that only the first dominant-coded predictor has a non-zero effect:

$$\beta_j = \begin{cases} 0.2, & j = p + 1, \\ 0, & \text{otherwise.} \end{cases}$$

Binary outcomes are then generated from

$$y_i \sim \text{Bernoulli}\left(\sigma \left(\sum_{j=1}^{3p} \beta_j x_{ij} + \beta_0 \right) \right),$$

where $\sigma(\cdot)$ is the logistic link function.

We fit the simulated data using four joint models: MLE, Bayesian Lasso regression, Bayesian Ridge regression, and Bayesian regularised horseshoe, with results shown in Figure 3.9. Due to collinearity among the additive, dominant, and recessive coding of the same SNP, the MLE identifies x_{add1}, x_{dom1}, and x_{rec1} simultaneously, but with coefficients in different directions. Ridge regularisation applies weaker shrinkage, exhibiting a similar pattern to MLE, although with smaller variance across variables. In contrast, the regularised horseshoe and Lasso perform better, generally selecting only the true effect x_{dom1}. However, in some simulation replicates, the additive encoding x_{add1} happens to fit the data better than the dominant encoding, leading both methods to select x_{add1} instead. This occasional misselection contributes to the noticeably higher variance observed for x_{add1}. Between the two, the regularised horseshoe tends to produce estimates closer to 0.2 than the Lasso when x_{add1} is selected, which can make its variance for x_{add1} appear larger in the violin plots.

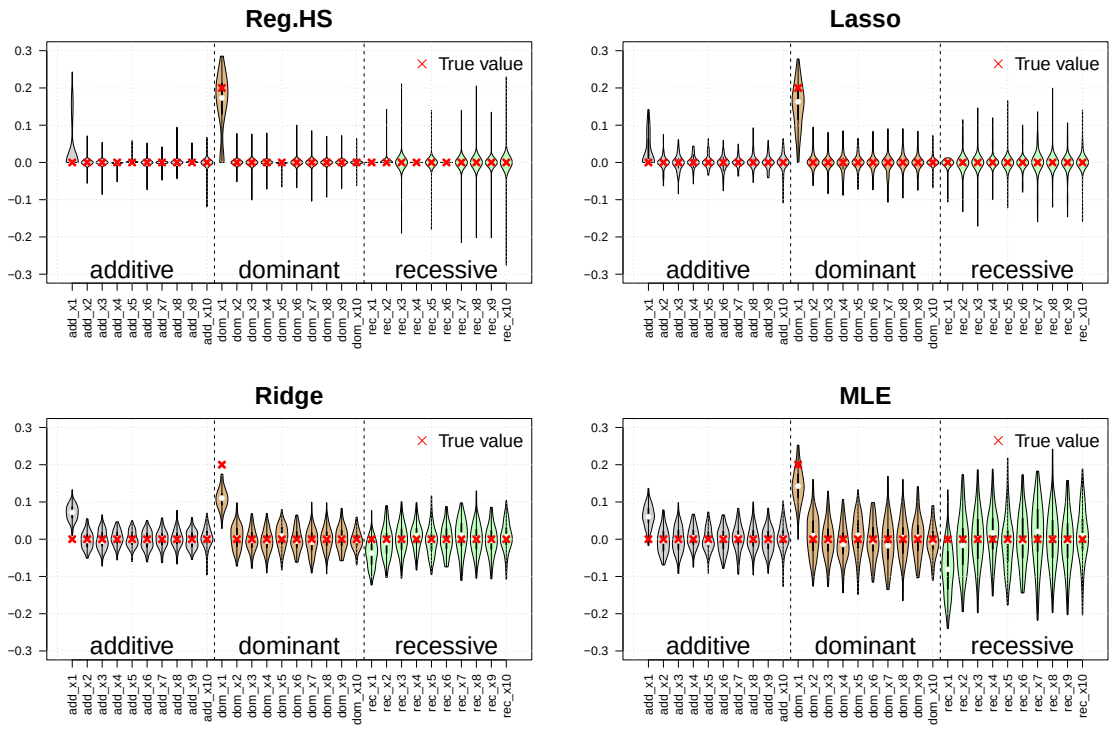


Figure 3.9: Results of coefficient estimates from a multi-genetic model for binary outcomes (logistic regression) across simulated replicates using methods: MLE, Bayesian Lasso, Bayesian Ridge, and Bayesian regularised horseshoe.

In terms of MSE (Figure 3.10), the MLE shows the poorest performance, followed by ridge regression, Lasso, and finally the regularised horseshoe, which achieves the lowest error. The improvement from ridge to Lasso and the regularised horseshoe reflects the stronger shrinkage of irrelevant effects in these methods. Although the regularised horseshoe occasionally exhibits larger variance for x_{add1}, this is not offset by its overall advantage in accuracy.

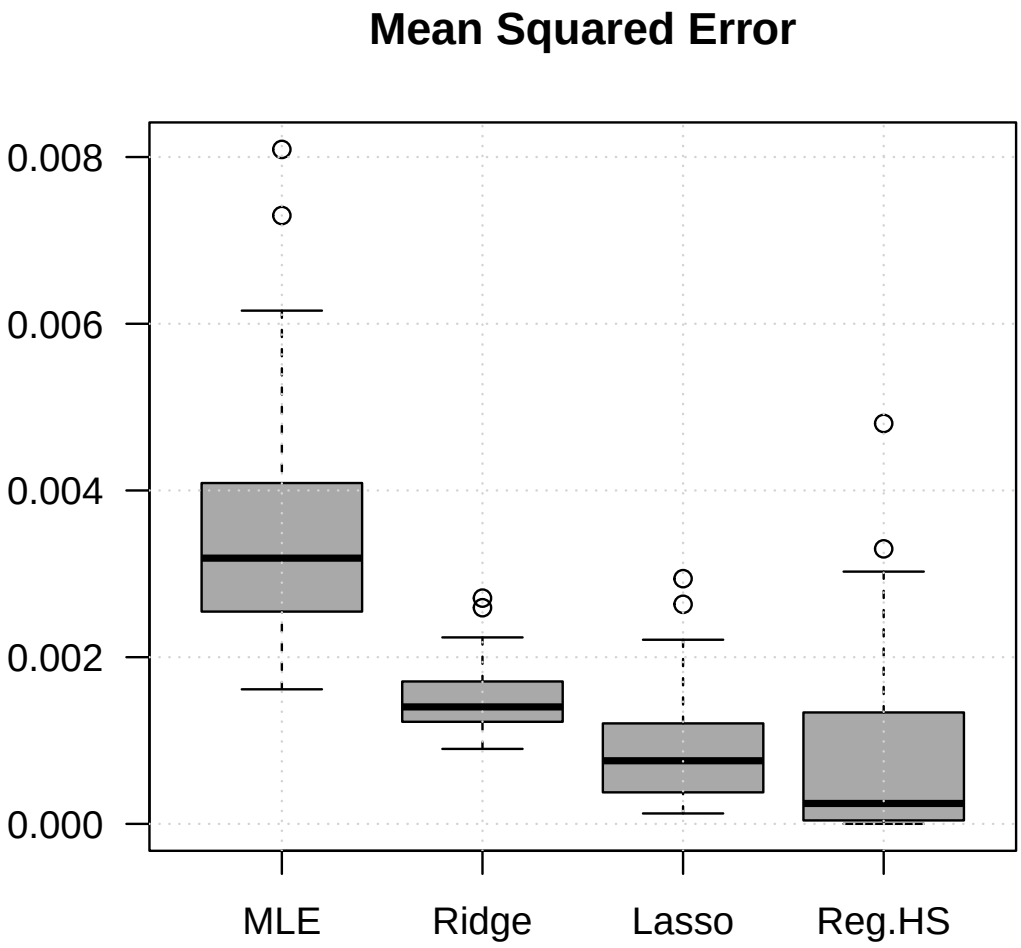


Figure 3.10: Mean squared error comparison for a multi-genetic logistic regression model across simulated replicates.

3.4.3 Linear regression

We simulate continuous phenotypes according to

$$\mathbf{y} = X\boldsymbol{\beta}_{0:p} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}\big(\mathbf{0},\, \sigma^2 I_n\big)\,,$$

where $X \in n \times (p+1)$, $\boldsymbol{\beta}_{0:p} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$ contains the intercept β_0 and regression coefficients $\beta_1, \ldots, \beta_p$, and σ^2 is the residual variance.

To calibrate the noise level such that the genetic effects explain 1% of the total phenotypic variance, we first compute the variance of the genetic component $X\boldsymbol{\beta}$

under a fixed-effect model. The residual variance σ^2 is then chosen to satisfy

$$h^2 \;=\; \frac{\mathrm{Var}(X\boldsymbol{\beta})}{\mathrm{Var}(X\boldsymbol{\beta}) + \sigma^2} \;=\; 0.01,$$

where h^2 denotes the narrow-sense heritability.

3.4.3.1 Single genetic model

For the single-gene simulation, we generate $n = 3{,}000$ samples and $p = 40$ SNPs coded $\{0, 1, 2\}$ with minor-allele frequency $\geq 10\%$. To induce linkage disequilibrium, SNPs x_1 and x_2 have Pearson correlation > 0.7, while all other pairwise correlations remained below 0.3. We fix $\beta_0 = 1$ and set

$$\boldsymbol{\beta}_{1:p} = (0.2, 0, \ldots, 0)^T,$$

so that only x_1 contribute to the genetic signal. After adding Gaussian noise calibrated for 1% heritability, we repeat this simulation 100 times.

We fit the simulated data using 5 approaches: LM marginal regression, maximum likelihood estimation, Bayesian Lasso, Bayesian ridge, and Bayesian regularised horseshoe via permuted coordinate descent. To ensure comparability, we set the Laplace prior with $b = 0.1$, the Gaussian prior with SD $= 0.1$ and the regularised horseshoe hyperparameter to $\tau = 0.1$ and $c = 1$. Results over 100 replicates are shown in Figure 3.11. The marginal GLM detected two signals, with the first coefficient around 0.2 and the second a false positive. Joint MLE avoided false discoveries but exhibited excessive variance across all coefficients. Bayesian Ridge regression over-penalised the first variable and under-penalised others, e.g., the second variable retained a coefficient of 0.05. In contrast, the Laplace prior and regularised horseshoe tightly concentrated non-zero estimates and shrank null coefficients toward zero. The regularised horseshoe provided an estimate for β_1 closest to the true value of 0.2, consistent with Section 3.2: its polynomial decay allows large coefficients to persist, while the Laplace prior's exponential decay imposes stronger shrinkage.

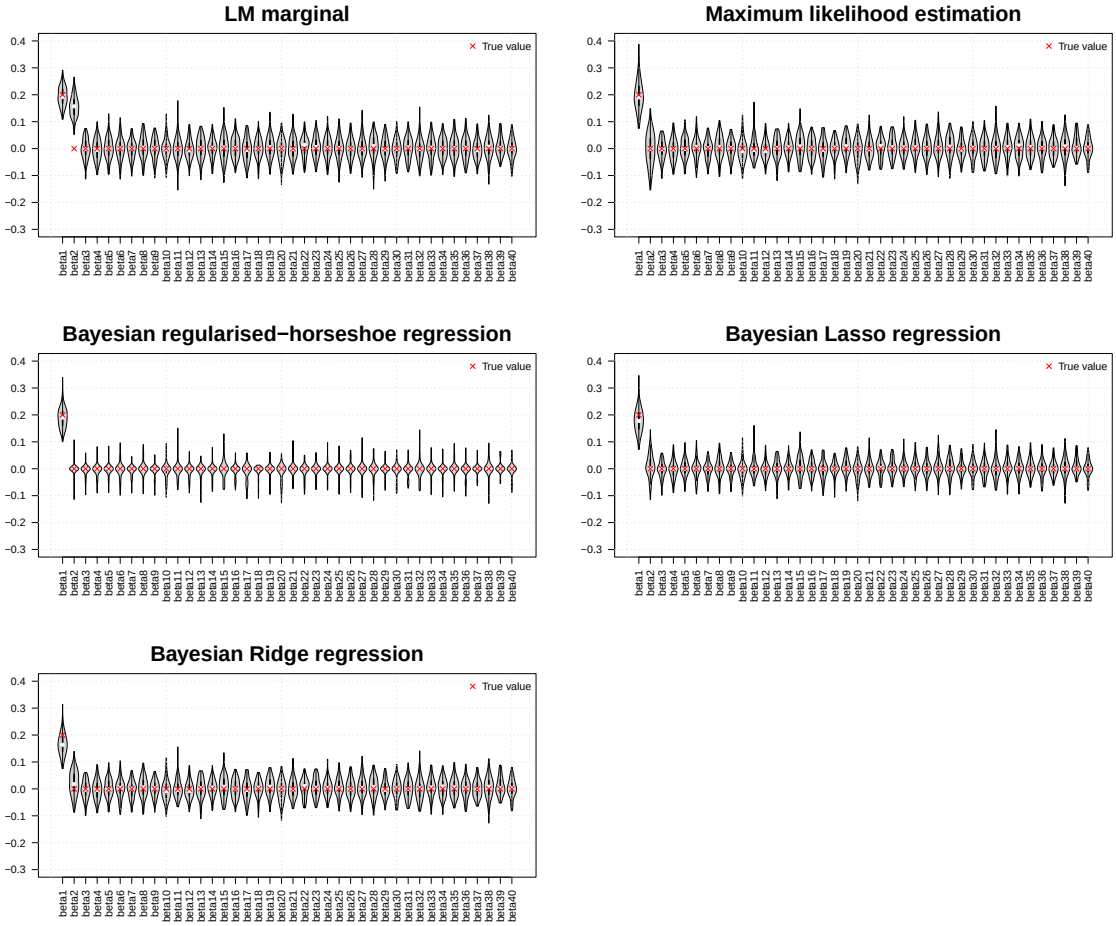


Figure 3.11: Results of coefficient estimates from a single genetic model for continuous outcomes (linear regression) across simulated replicates using five methods: marginal LM, joint MLE, Bayesian Lasso, Bayesian Ridge, and Bayesian regularised horseshoe.

A comparison of the MSE between the estimated and true values across methods (Figure 3.12) indicates that the regularised horseshoe prior yielded the most accurate estimates, as reflected by the lowest MSE. The Lasso prior showed slightly higher MSE but still outperforms the ridge prior, which in turn performs better than the joint MLE. Marginal linear regression exhibited the highest MSE among all methods.

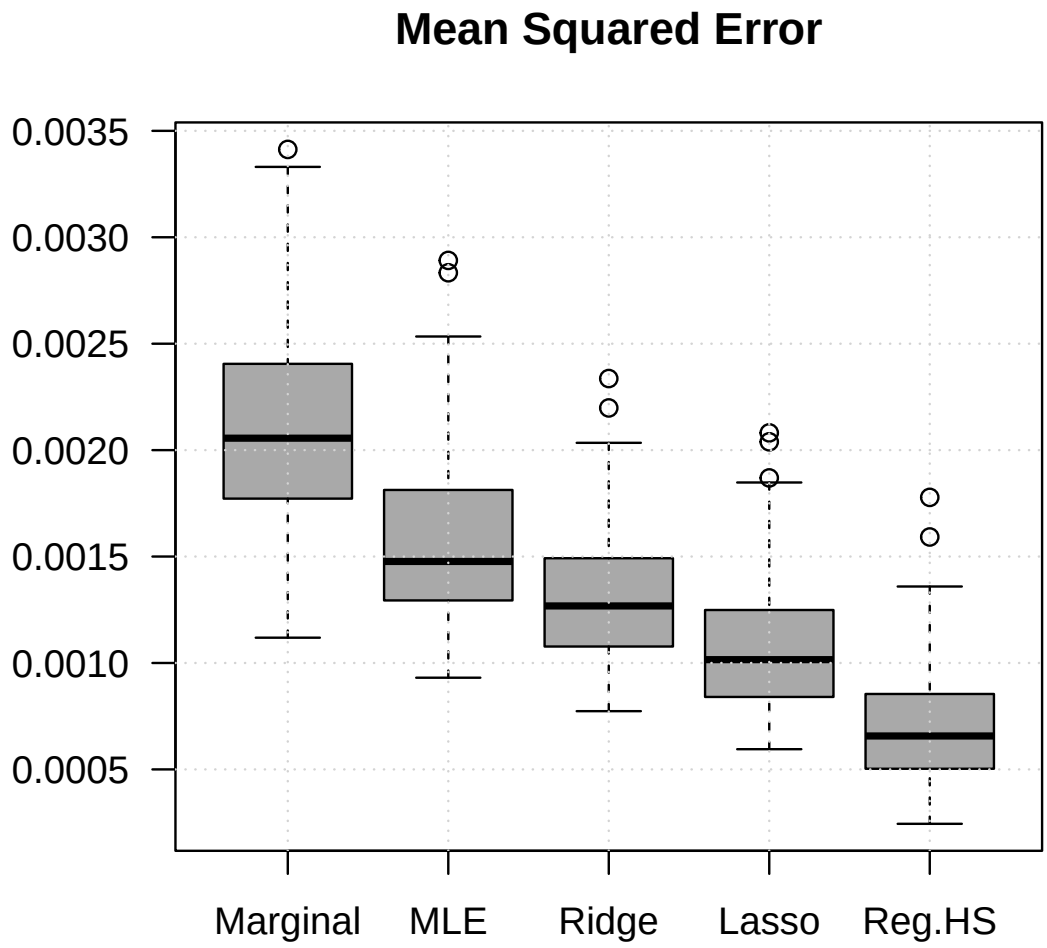


Figure 3.12: Comparison of mean squared error across simulated replicates for the single genetic model with continuous outcomes using linear regression.

To evaluate the detection of pairwise interactions, we simulate $n = 3{,}000$ samples on $p = 10$ variables with MAF $\geq 10\%$. All possible $\binom{10}{2}$ pairwise interaction terms are included in the model. The true coefficient vector $\boldsymbol{\beta}$ assigned a nonzero effect of 0.2 only to the first interaction term $x_1 \times x_2$, with all main effects and other interaction effects set to zero. After calibrating the residual variance to achieve 1% narrow-sense heritability, we generate 100 independent replicates.

The simulated model can be written as

$$y_i \;=\; \beta_0 + \sum_{j=1}^{p} \beta_j \, x_{ij} + \sum_{1 \leq j < k \leq p} \beta_{jk} \, x_{ij} x_{ik} + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}\big(0,\, \sigma^2\big) \, .$$

where $\beta_{12} = 0.2$, and $\beta_j = 0$ for all $j \neq 1, 2$, with $\beta_{jk} = 0$ for all $(j, k) \neq (1, 2)$.

We fit the simulated data using five approaches: marginal linear regression, joint MLE, Bayesian Lasso, Bayesian ridge, and Bayesian regularised horseshoe. The results are presented in Figure 3.13. The marginal linear regression selects three signals, x_1, x_2, and the true interaction, thereby introducing two false positives. The joint MLE correctly identified the true interaction but exhibited high variance across all coefficient estimates. In contrast, Bayesian methods with shrinkage priors effectively penalise irrelevant variables while accurately estimating the interaction effect. Both the Bayesian Lasso and the regularised horseshoe produce estimates close to the true value of 0.2 with smaller variance and shrank null coefficients more strongly than the Bayesian ridge. Among these, the regularised horseshoe achieves the median estimate closest to the true value.

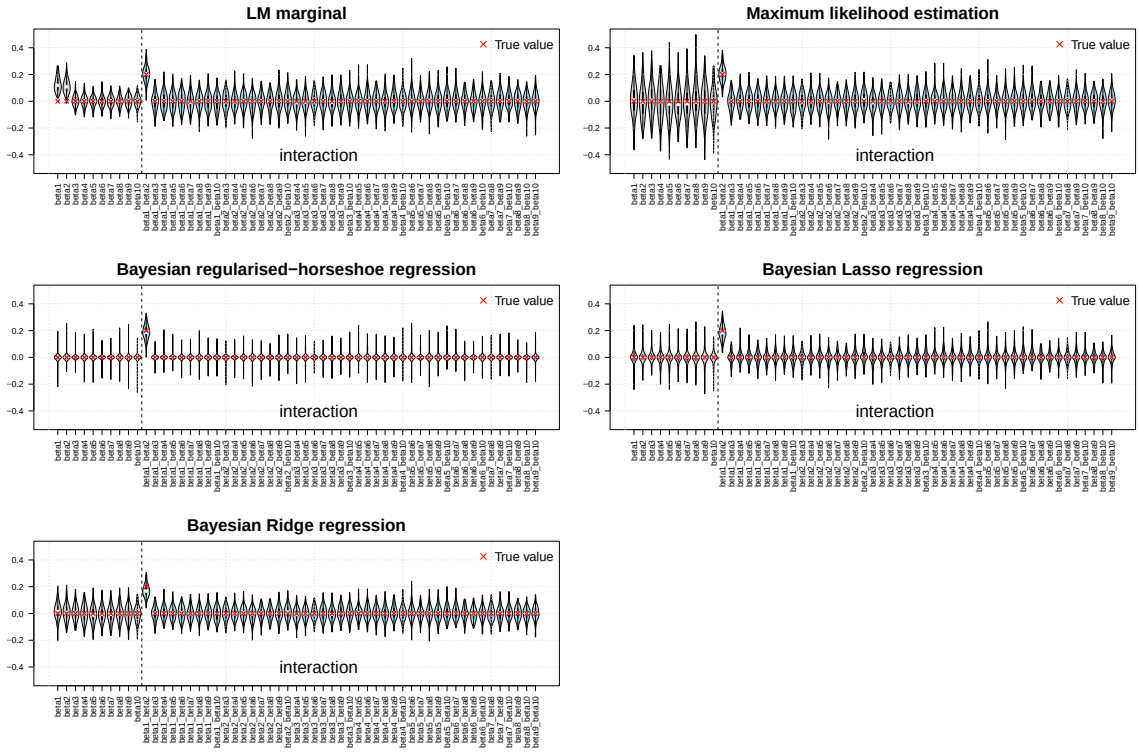


Figure 3.13: Results of coefficient estimates from a single genetic model with interaction term for continuous outcomes (linear regression) across simulated replicates using five methods: marginal LM, joint MLE, Bayesian Lasso, Bayesian Ridge, and Bayesian regularised horseshoe.

We next evaluate the MSE for each method, as shown in Figure 3.14. The joint MLE yields the poorest accuracy, with the largest MSE among all methods, primarily due to its high variability in estimating the first ten coefficients corresponding to the

main SNP effects. The marginal linear regression ranks second-worst, consistent with its selection of false positives. Within the Bayesian framework, the ridge prior performs reasonably well but still has a higher MSE than the Bayesian Lasso. The regularised horseshoe attained the best performance, delivering the smallest MSE overall. This advantage stems from its ability to preserve the true interaction effect while aggressively shrinking coefficients associated with irrelevant predictors toward zero.

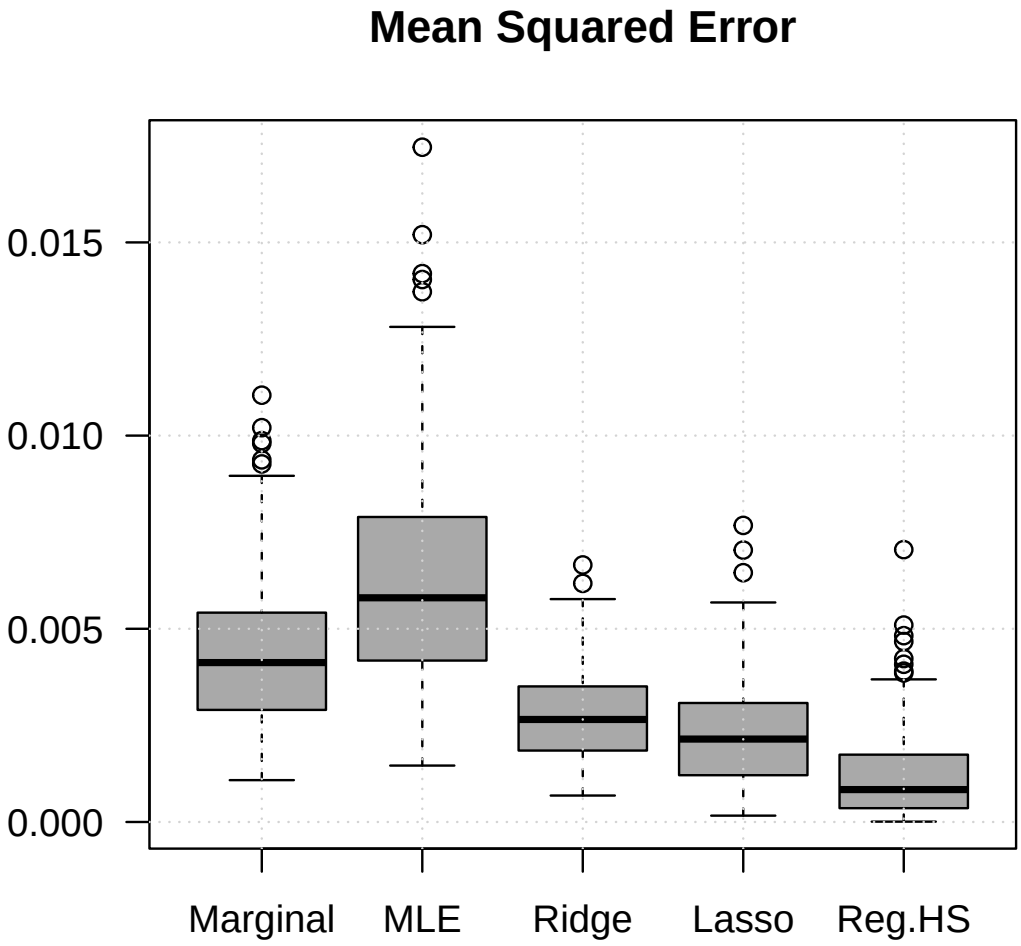


Figure 3.14: Comparison of mean squared error across simulated replicates for single-gene models including an interaction term, fitted to continuous outcomes via linear regression.

3.4.3.2 Multi-genetic model

To assess the ability of linear regression methods to identify the most appropriate genetic encoding, we simulate a "multi-genetic" scenario incorporating additive, dominant, and recessive effects. We generate $n = 10{,}000$ observations on $p = 10$ variables, each represented in three coding schemes: additive $(0, 1, 2)$, dominant $\mathbb{I}\{x \geq 1\}$, and recessive $\mathbb{I}\{x = 2\}$, where $\mathbb{I}\{\cdot\}$ denotes the indicator function (1 if the condition holds, 0 otherwise). These are concatenated to form the design matrix.

$$X_{\text{multi}} = \big[X_{\text{add}} \;\big|\; X_{\text{dom}} \;\big|\; X_{\text{rec}}\big],$$

where the dimensionality of each block is $n \times p$. The true coefficient vector is specified so that only the first dominant-coded predictor has a non-zero effect:

$$\beta_j = \begin{cases} 0.2, & j = p + 1, \\ 0, & \text{otherwise.} \end{cases}$$

Continuous outcomes are then generated according to the linear model.

$$y_i = \sum_{j=1}^{3p} \beta_j x_{ij} + \beta_0 + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2),$$

Where σ^2 is chosen to achieve 1% narrow-sense heritability, and the simulation is repeated over 100 replicates.

We analyse the simulated data using MLE, Bayesian regularised horseshoe, Bayesian Lasso, and Bayesian ridge regression, with results presented in Figure 3.15. Both MLE and ridge regression identify all three coding schemes of x_1 (additive, dominant, and recessive), but two of these are false positives, likely reflecting the correlations inherent in the genetic encodings. Ridge regression outperforms MLE, exhibiting reduced variance due to the applied penalisation. In contrast, the Bayesian Lasso and regularised horseshoe more accurately identify only the true causal variable, x_{dom1}. Among these, the regularised horseshoe produces estimates closest to the true effect size of 0.2 and applies stronger shrinkage to non-causal variables, resulting in lower variance relative to the Lasso. Occasionally, stochastic variation in the simulations leads the additive coding of x_1 to appear slightly more predictive than the dominant coding, yielding somewhat higher variance for x_{add1} compared with other null variables.

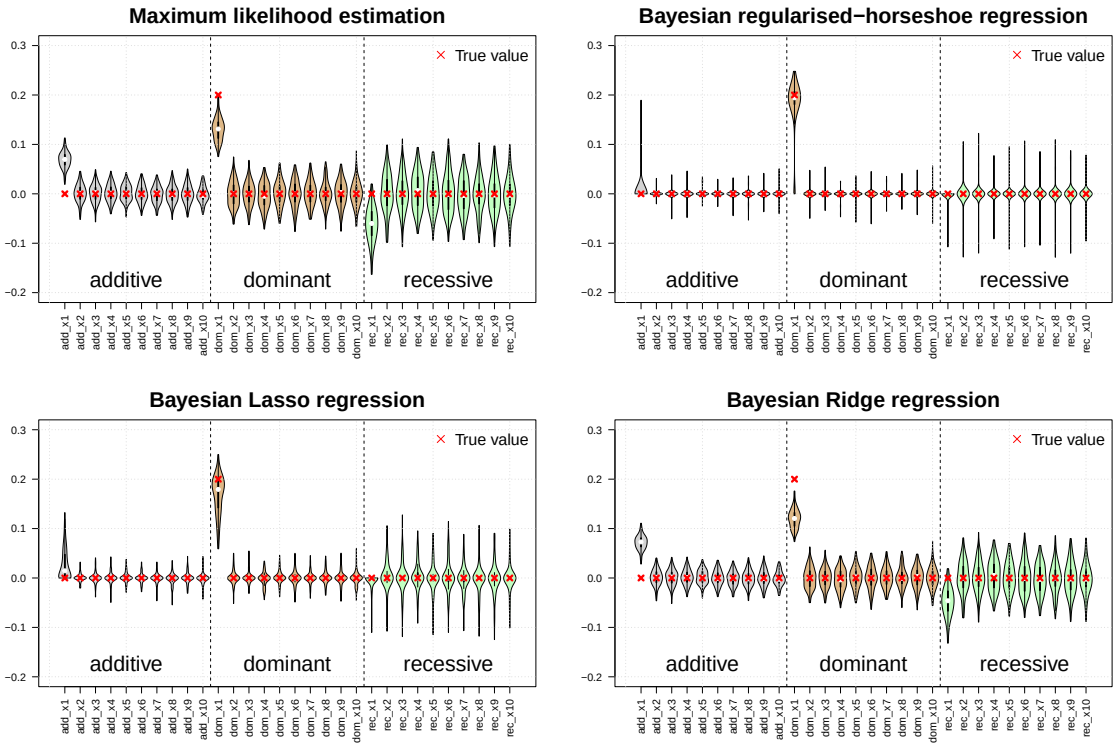


Figure 3.15: Results of coefficient estimates from a multi-genetic model for continuous outcomes (linear regression) across simulated replicates using five methods: joint MLE, Bayesian Lasso, Bayesian Ridge, and Bayesian regularised horseshoe.

We compare the MSE across the models, with results presented in Figure 3.16. The regularised horseshoe attains the lowest median MSE, followed sequentially by the Bayesian Lasso, Ridge regression, and MLE, reflecting the favourable accuracy of shrinkage-based Bayesian approaches in estimating the true effects while controlling for irrelevant variables.

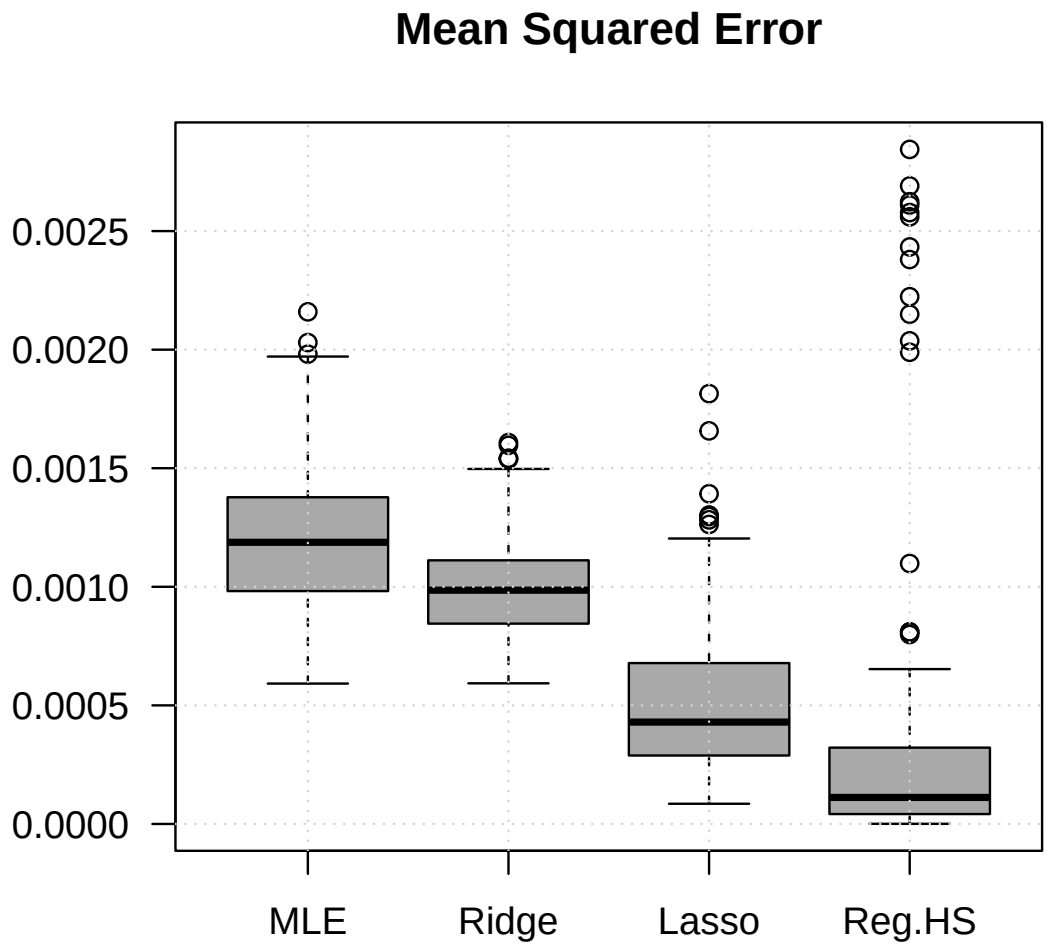


Figure 3.16: Comparison of mean squared error across simulated replicates for the multi-genetic models with continuous outcomes using linear regression.

3.4.4 Real dataset simulation

Building on the simulation studies above, which demonstrate the favourable performance of the regularised horseshoe compared with benchmark methods across single genetic models, single genetic models with interaction terms, and multi-genetic models for linear regression and logistic regression. We next explore its properties in a more complex, real-data setting. Specifically, we simulate data based on the MalariaGEN dataset introduced in Chapter 2, in order to evaluate the performance of the regularised horseshoe in regions of linkage disequilibrium, across different genetic

architectures, and in the presence of interactions.

We select HLA alleles and ERAP1/2 allotypes with a frequency greater than 5%. The model includes HLA class I and class II alleles, as well as ERAP1/2 allotypes under additive, dominant, and recessive genetic coding schemes. Additionally, we include interaction terms between HLA class I alleles and ERAP1/2 allotypes. The true coefficient vector is set as follows:

$$\beta_j = \begin{cases} -0.2, & j = \text{DQB1*05:01}, \\ 0.3, & j = \text{dominant\_DRB1*07:01}, \\ 0.3, & j = \text{ERAP1:TEPMGMRDRE-A*23:01 (interaction)}, \\ 0, & \text{otherwise.} \end{cases}$$

With all other coefficients set to zero, we design the simulation such that *DQB1*05:01* and *DQB1*03:01* are highly correlated (Pearson correlation = 0.76), allowing us to assess model performance within a region of LD across different genetic architectures, including additive and interactive effects. Case–control outcomes are generated from a Bernoulli distribution. The simulated datasets are fitted using the regularised horseshoe model. Each simulation is repeated 100 times, and within each simulation, 100 permutations are performed.

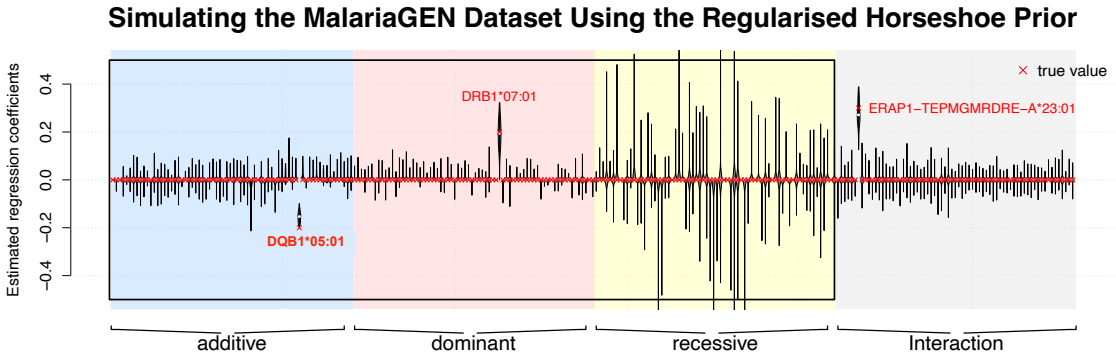


Figure 3.17: Simulation results based on the MalariaGEN dataset.

The results are shown in Figure 3.17. For clearer visualisation, the genomic region is divided into four blocks, each highlighted with a distinct colour. The regularised horseshoe accurately detected all three simulated signals, DQB1*05:01 in additive effect, DRB1*07:01 in dominant effect, and the interaction ERAP1:TEPMGMRDRE-A*23:01, while effectively shrinking all other coefficients toward zero. These results demonstrate the model's robustness and applicability to real-world genetic datasets, even in regions with linkage disequilibrium and complex genetic architectures.

3.5 Software availability and instructions

We have developed an R package `mapHS`, which is available to download from GitHub. `mapHS` implements standard and permutation coordinate descent methods for use in logistic and linear regression models. For scenarios where regions are not strongly correlated, we recommend using the standard coordinate descent method. Conversely, when dealing with multiple correlated regions, the permutation-based coordinate descent approach is preferable, as it improves the search for the optimal solution. The package can be downloaded and installed using the following commands:

```
git clone https://github.com/QiJingS/Horseshoe_project.git
cd Horseshoe_project
tar -xzvf mapHS_0.1.0.tar.gz
```

Logistic regression

Compute log-posterior value:

```
log_posterior_logistic(Y, X, parameters, tau = 0.1, c = 1)
```

Estimate coefficients using standard coordinate descent:

```
CD_HS_logistic(Y, X, init, tau = 0.1, c = 1,
               max_iter = 15, tol = 1e-5)
```

Estimate coefficients using permutation-based coordinate descent:

```
CD_HS_purm_logistic(Y, X, init, tau = 0.1, c = 1, max_iter = 30,
tol = 1e-5, K = 10, max_cores = detectCores())
```

Linear regression

Compute log-posterior value:

```
log_posterior_linear(Y, X, parameters, tau = 0.1, c = 1, sigma2 = 1)
```

Estimate coefficients using standard coordinate descent:

```
CD_HS_linear(Y, X, init, tau = 0.1, c = 1,
             max_iter = 15, tol = 1e-5)
```

Estimate coefficients using permutation-based coordinate descent:

```
CD_HS_purm_linear(Y, X, init, tau = 0.1, c = 1, max_iter = 30,
tol = 1e-5, K = 10, max_cores = detectCores())
```

Function arguments

`X` Design matrix.

`Y` Response vector.

`init` Initial values.

`K` Number of permutations (used in permutation-based methods).

`tol` Convergence threshold for stopping criterion.

`max_iter` Maximum number of iterations allowed.

`tau` Global shrinkage hyperparameter.

`c` Slab width.

`sigma2` Noise variance (used in linear regression).

3.6 Discussion

We introduced a novel fine-mapping approach based on Bayesian inference using the maximum a posteriori estimate under the regularised horseshoe prior. Through extensive simulation studies, we assessed its performance across a variety of genetic architectures, including additive, dominant, and recessive effects, as well as interaction terms, in both simulated and real datasets. To facilitate broader use, the approach has been implemented in the R package `mapHS`, which supports both linear and logistic regression models.

The regularised horseshoe prior retains the advantages of the canonical horseshoe, strong shrinkage of near-zero coefficients coupled with heavy tails that preserve large signals, while ensuring a finite density through the inclusion of a slab parameter. This continuous spike-and-slab behaviour enables more accurate recovery of true non-zero effects than classical ℓ_1 (lasso) or ℓ_2 (ridge) penalties, particularly in the presence of correlated predictors. Compared with MCMC-based implementations of the horseshoe, such as those in `RStan` or `brms`, the MAP estimator avoids tuning issues like divergent transitions and provides a single, easily interpretable point estimate, although it does not capture posterior uncertainty and may not yield genuinely sparse coefficients. Moreover, while `RStan` typically estimates posterior means, these can be unreliable in regions such as HLA/ERAPs where multimodality is present.

The motivating application in Chapter 2 concerned genetic variation and gene–gene interactions in the antigen-presentation pathway, specifically between HLA class I alleles and ERAP1/2 variants. Prior studies have shown that ERAP allotypes modulate peptide trimming in an allele-specific manner, thereby interacting with particular HLA alleles to influence immune recognition. By jointly encoding main genetic effects and interaction terms within a single regression framework, and applying the `mapHS` method to genotype–phenotype data, we were able to identify the combinations of HLA and ERAPs variants that best explain variance in antigen presentation. The variables selected by the MAP estimator correspond closely to established immunological mechanisms, underscoring the practical utility of the method for dissecting complex genetic architectures.

Nevertheless, several limitations remain. At present, different genetic encoding schemes such as additive, dominant, and recessive effects are handled by fitting separate models and selecting the MAP solution post hoc, which becomes computationally demanding as the number of SNPs and interactions increases. Although coordinate descent scales much more effectively than full MCMC, its per-iteration cost still grows linearly with the number of predictors, suggesting that additional strategies such as screening rules or warm-start heuristics could further improve scalability. In addition, we relied on a default half-Cauchy prior on the global shrinkage parameter, whereas a calibrated hyperprior on the slab width could provide better control over the degree of shrinkage and reduce the risk of over-shrinking moderate effects.

In summary, the coordinate-descent MAP estimator for the regularised horseshoe prior, as implemented in `mapHS`, offers an effective bridge between Bayesian sparsity methods and scalable optimisation. It combines adaptive, heavy-tailed shrinkage with computational tractability, producing interpretable models well suited to high-dimensional genomic applications. By applying this framework to antigen-presentation genetics, we have shown its capacity to uncover biologically meaningful main effects and interactions, paving the way for broader applications in complex trait genetics and other high-dimensional biological domains.

Chapter 4

Enhancing Inference by Jointly Modelling the Effects of Antigen Presentation Pathway Variation on Phenotypes Using a Bayesian Shrinkage Approach

4.1 Introduction

The preceding chapters have laid the essential groundwork for a comprehensive analysis of the APP. In Chapter 1, we establish the biological foundation, highlighting the critical roles of HLA molecules, ERAPs, TAPs, and proteasome subunits in immune surveillance and the substantial impact of their genetic variation on human health. Chapter 2 addresses the analytical challenge of this complexity by developing a scalable bioinformatics pipeline, enabling the systematic characterisation of APP genetic features, from HLA alleles and supertypes to multi-variant allotypes, across diverse cohorts. In Chapter 3, we confront the statistical challenges of high dimensionality and linkage disequilibrium within the APP, introducing a novel Bayesian fine-mapping approach based on the regularised horseshoe before robustly identifying sparse genetic signals amidst correlated predictors.

In this chapter, we synthesise these components to perform a unified, pathway-wide association analysis. While numerous studies have reported associations between individual HLA alleles and infectious diseases, a holistic understanding that incorporates non-HLA APP genes and, crucially, their potential epistatic interaction, remains largely unexplored. We hypothesise that a joint model, which simultaneously considers the effects of all major APP components, will provide a more powerful and

accurate picture of the pathway's genetic architecture, revealing associations that are masked in conventional single-variant analyses.

To test this hypothesis, we apply the feature set from Chapter 2 and the statistical framework from Chapter 3 to four large-scale genomic datasets: the HCV spontaneous clearance cohort, the STOP-HCV chronic infection cohort, the MalariaGEN severe malaria study, and the UK Biobank serological panel. Our primary objective is to move beyond a narrow focus on HLA by modelling the collective and interactive effects of the entire APP. Specifically, we aim to: (1) identify novel genetic determinants of infectious disease outcomes within the broader APP; (2) characterise epistatic interactions, particularly between HLA class I alleles and ERAP1/2 allotypes, which may mechanistically explain allele-specific presentation efficiencies; and (3) benchmark our joint Bayesian approach against traditional conditional analysis to evaluate its advantages in deciphering complex genotype-phenotype relationships. By integrating computational extraction, statistical innovation, and biological interpretation, this chapter seeks to deliver a paradigm shift from a gene-centric to a pathway-centric view of immunogenetics in infectious diseases.

4.2 Methods

4.2.1 Genetic predictors and data sources

For each cohort, the analysis incorporates a comprehensive set of genetic features derived from the pipeline in Section 2.3. Three main categories are included: **(i) HLA features** classical alleles (4-digit resolution), HLA gene heterozygosity, HLA supertypes, HLA tapasin-dependence scores, HLA-A/C protein expression levels, and the HLA-B -21 M/T dimorphism; **(ii) Non-HLA APP features** common variants (MAF $>$ 1%) and allotypes of key genes during the antigen presentation pathway, including *ERAP1*, *ERAP2*, *TAP1*, *TAP2*, *PSMB8*, and *PSMB9*; and **(iii) Interaction terms** interactions between HLA class I alleles (A, B, C) and ERAP1/ERAP2 allotypes, as well as between ERAP1 and ERAP2 allotypes themselves.

4.2.2 Bayesian joint regression model

We investigate the association between features of the antigen presentation pathway and relevant phenotypes using Bayesian regression models with a regularised horseshoe prior, implemented in the `mapHS` package (Section 3.5). This approach enables sparse effect estimation while robustly handling high correlations among predictors.

Model specification

For a continuous phenotype y_i (e.g., viral load, MFI) of the i-th sample, we employed a Bayesian linear regression model:

$$\begin{aligned} y_i = \beta_0 &+ \sum_{j=1}^{n_1} \left(\beta_j^{\text{add}} x_{ij}^{\text{add}} + \beta_j^{\text{dom}} x_{ij}^{\text{dom}} + \beta_j^{\text{rec}} x_{ij}^{\text{rec}} \right) \\ &+ \sum_{k=1}^{n_2} \beta_k^{\text{HLA-ERAP}} x_{ik}^{\text{HLA-ERAP}} + \sum_{l=1}^{n_3} \beta_l^{\text{ERAPxERAP}} x_{il}^{\text{ERAPxERAP}} \\ &+ \sum_{m=1}^{n_4} \beta_m^{\text{cov}} x_{im}^{\text{cov}} + \epsilon_i, \end{aligned}$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

For a binary phenotype y_i (e.g., case-control status), we use a Bayesian logistic regression model:

$$\begin{aligned} \text{logit}(P(y_i = 1)) = \beta_0 &+ \sum_{j=1}^{n_1} \left(\beta_j^{\text{add}} x_{ij}^{\text{add}} + \beta_j^{\text{dom}} x_{ij}^{\text{dom}} + \beta_j^{\text{rec}} x_{ij}^{\text{rec}} \right) \\ &+ \sum_{k=1}^{n_2} \beta_k^{\text{HLA-ERAP}} x_{ik}^{\text{HLA-ERAP}} + \sum_{l=1}^{n_3} \beta_l^{\text{ERAPxERAP}} x_{il}^{\text{ERAPxERAP}} \\ &+ \sum_{m=1}^{n_4} \beta_m^{\text{cov}} x_{im}^{\text{cov}}. \end{aligned}$$

β_0 is the intercept; x_{ij}^{add}, x_{ij}^{dom}, x_{ij}^{rec} are the genetic variant; $x_{ik}^{\text{HLA-ERAP}}$ is the HLA-ERAP interaction term; $x_{il}^{\text{ERAPxERAP}}$ is the ERAP1-ERAP2 interaction term; x_{im}^{cov} is the covariates.

Model fitting and inference

We estimate the regression coefficients $\boldsymbol{\beta}$ via their Maximum A Posteriori (MAP) values using permutation-based coordinate descent algorithms (`CD_HS_purm_linear` or `CD_HS_purm_logistic`). To quantify the uncertainty of these estimates, we compute approximate standard errors for coefficients with $|\beta_i| > 1 \times 10^{-4}$ using the active subspace method. Based on these standard errors, 95% confidence intervals are constructed as MAP $\pm$ 1.96 $\times$ SE.

4.2.3 Bayesian hypothesis testing and statistical significance

A central challenge in genetic association studies is selecting an appropriate significance threshold. Methods are solely based on the number of tests performed, such as

the Bonferroni correction, which can be overly conservative, as the threshold depends heavily on the total number of variants analysed. A more principled approach frames the problem as determining what p-value threshold yields sufficient confidence that a variable is genuinely associated with the phenotype. This determination depends on both the prior probability of association and the statistical power [1] of the study, as captured by the relation:

$$\text{posterior odds}(\text{association} \mid p < T) = \text{prior odds} \times \frac{\text{power}}{T}.$$

For a given p-value threshold T, the left-hand side gives the odds of a true association when the p-value falls below T; the corresponding probability is one minus the positive false discovery rate. Interpreting this relationship requires knowledge of the prior odds of association and the statistical power, which in turn depend on the distribution of true effect sizes and the allele frequency spectrum.

As discussed in Chapter 3, our model uses the regularised horseshoe prior with the selected hyperparameter, achieving a statistical power of approximately 0.5. If we assume that roughly 10 out of every 100 tested variants are truly associated with the trait, the prior odds are about 1/10. To obtain substantial posterior odds of association, a Bayes factor on the order of 10^3 is typically required. Even stronger evidence (e.g., 10^4) would correspond to larger effect sizes, such as odds ratios of approximately 2 or higher. Under these assumptions, p-values around 5×10^{-4} may provide compelling evidence for association, depending on allele frequencies and statistical power. However, weaker effects would require more stringent thresholds and are consequently more difficult to detect. While the conservative significance threshold established in previous GWAS research is 1×10^{-8}, the antigen presentation pathway is of particular biological relevance, and genetic components such as HLA alleles have already been shown to associate with numerous infectious disease phenotypes (Jones et al., 2007). Therefore, we adopt a broader significance threshold of $p < 5 \times 10^{-4}$ in this study.

Although Bayesian inference allows significance to be assessed via posterior credible intervals (e.g., checking whether a coefficient's interval includes zero), we also conducted formal hypothesis testing to directly quantify the evidence for the association between each genetic variant and the phenotypic outcomes. This facilitates comparison with conventional frequentist methods and provides a continuous measure of evidence strength rather than a binary significance decision.

[1] Statistical power is the probability of a study correctly detecting a real effect or relationship if one exists

We perform hypothesis testing using a Wald test based on the posterior mean $\hat{\beta}$ and standard error $se(\hat{\beta})$ from the Bayesian regression model. The test statistic (z-score) is:

$$z = \frac{\hat{\beta}}{se(\hat{\beta})},$$

under the null hypothesis $H_0 : \beta = 0$. A two-sided p-value is then computed from the standard normal cumulative distribution function Φ as:

$$p = 2 \cdot \big(1 - \Phi(|z|)\big).$$

Given the large number of variants tested, multiple testing correction is essential. Rather than relying exclusively on traditional methods (e.g., the Bonferroni adjustment), we computed Bayes Factors to quantify the evidence supporting the alternative hypothesis ($H_1 : \beta \neq 0$). We use the Wakefield Approximate Bayes Factor, which assumes a normal prior on the log-odds ratio under H_1 and offers an efficient approximation of the evidence:

$$\text{BF} = \sqrt{1 - r} \cdot \exp\left(\frac{r\, z^2}{2}\right),$$

where $r = \frac{W}{W+V}$, V is the variance of the maximum likelihood estimate, and W is the prior variance that reflects the expected magnitude of a true (non-null) effect.

4.2.4 Stepwise conditional analysis

To identify independent genetic associations while accounting for linkage disequilibrium, we perform stepwise conditional analysis using linear and logistic regression models. This iterative forward selection procedure began with an initial genome-wide scan, testing each variant individually. For continuous traits, we use the linear model:

$$Y = \beta_0 + \beta_1 G_i + \epsilon$$

For binary traits, we employed logistic regression:

$$\text{logit}(P(Y = 1)) = \beta_0 + \beta_1 G_i$$

where G_i represents the genotype of variant i.

The variant with the smallest p-value exceeding our significance threshold ($p < 5 \times 10^{-4}$) is selected as the primary association signal. We then proceed to conditional

rounds, each time including the top-associated variant as a covariate and retesting all other variants in the region. The conditional model took the form:

$$Y = \beta_0 + \beta_1 G_{\text{lead}} + \beta_2 G_i + \epsilon$$

This iterative process continued, adding each newly identified significant variant as an additional covariate in subsequent rounds. The extended conditional model became:

$$Y = \beta_0 + \beta_1 G_{\text{lead}} + \beta_2 G_{\text{secondary1}} + \cdots + \beta_k G_i + \epsilon$$

The procedure terminates when no additional variants reach the significance threshold in the conditional model.

The analysis is implemented in R using the `lm()` and `glm()` functions, ensuring model convergence and excluding variants in high linkage disequilibrium ($r^2 > 0.8$) with conditioning variants to avoid multicollinearity.

4.2.5 Covariates selection

To clarify covariate selection across cohorts, we stratified the dataset according to previously published studies: HCV spontaneous clearance versus chronic infection, MalariaGEN, and the UK Biobank serological panel. The STOPHCV dataset is a newly assembled cohort incorporating additional samples from Prof. Azim Ansari's research group. For cohorts with prior publications, we adopted the covariate adjustment strategies used in the original analyses to ensure methodological consistency and comparability. Specifically, the spontaneous clearance versus chronic infection analysis included sex and the top 10 host genetic principal components (PCs) (Jones et al., 2022). The MalariaGEN analysis included sex and the top 5 PCs (Band et al., 2019). The UK Biobank serological analysis adjusted for sex, age, and the first 20 host genetic PCs (Butler-Laporte et al., 2023). In each case, the number of PCs was determined in the original studies to adequately control for population stratification while avoiding unnecessary model complexity.

For the newly assembled STOPHCV cohort, we implemented a harmonised covariate framework. We included the top 20 host genetic PCs to account for population structure and capture finer-scale ethnic heterogeneity within this expanded dataset. Sex and age were included in all models as standard demographic covariates. For the cirrhosis phenotype, body mass index (BMI) was additionally included due to its established role as an independent risk factor for liver fibrosis progression and liver-related clinical outcomes (Nair et al., 2002). Although clinical records contained

information on injection drug use and alcohol consumption, approximately 500 samples had missing data for these variables. Furthermore, reported weekly alcohol intake was highly skewed, raising concerns about model instability and potential bias. Inclusion of these variables would therefore substantially reduce the effective sample size and statistical power. For these reasons, they were excluded from the final models.

4.3 Results

This section presents association analyses between genetic components of the APP and infectious disease phenotypes. We systematically investigate these relationships by applying the Bayesian joint modelling framework from Chapter 3 to the genetic features compiled in Chapter 2. Our primary objective is to determine whether this integrated approach could reveal associations overlooked by conventional marginal or conditional analyses. We further benchmark our results against these traditional methods, with emphasis on additive effects and pairwise interactions. Finally, we contextualise our findings within the existing literature, highlighting points of convergence and divergence, thereby providing a more cohesive biological interpretation of the observed genetic associations.

4.3.1 Bayesian joint regression analysis

The Bayesian joint model was employed to dissect the aggregate and interactive effects of antigen presentation pathway components across all cohorts. For each dataset, we model additive, dominant, and recessive genetic effects, in addition to key interaction terms. Association testing is conducted under the regularised horseshoe prior, with effect sizes estimated via maximum a posteriori estimation using the `mapHS` R package. The results of this comprehensive analysis for the HCV, MalariaGEN, and UK Biobank datasets are presented in the following subsections.

4.3.1.1 Hepatitis C

The Bayesian joint regression results for hepatitis C, as shown in Figure 4.1, include two cohorts: the *spontaneous clearance versus chronic infection* cohort and the *STOPHCV* cohort, which comprises three phenotypes: cirrhosis, viral load, and HCC. The results for each cohort and phenotype are presented individually as follows.

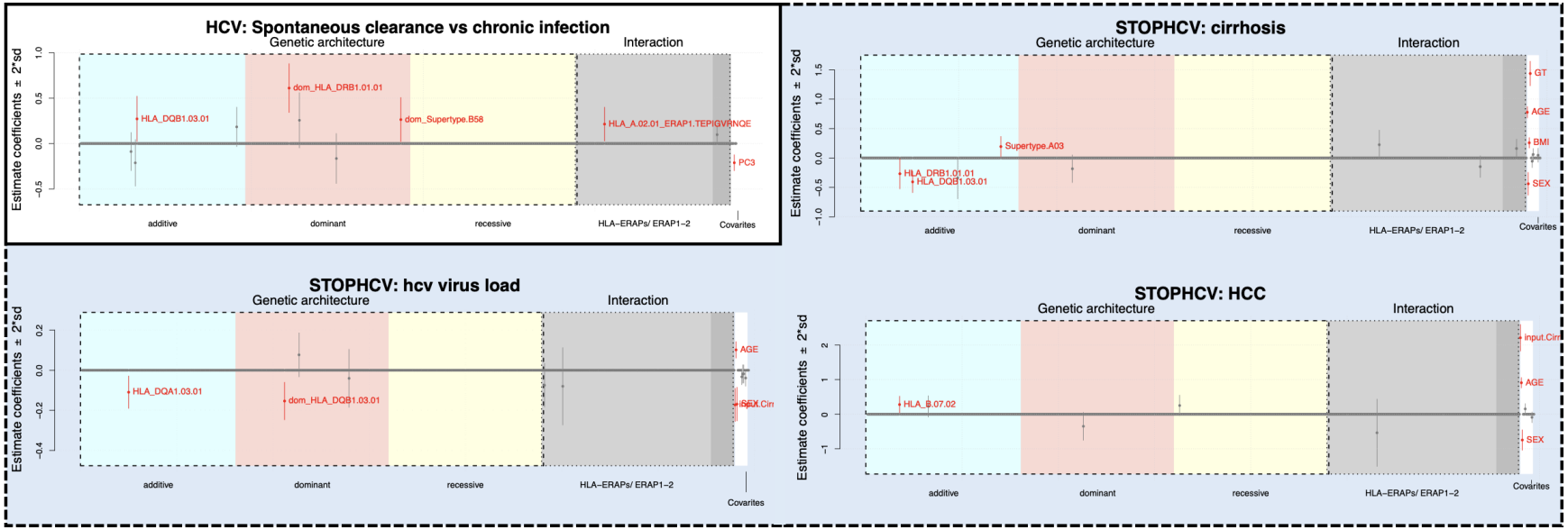


Figure 4.1: **Bayesian joint regression results for the Hepatitis C datasets**, which include two cohorts and four phenotypes in total. The cohorts are grouped into two blocks: white indicates the *spontaneous clearance versus chronic infection* cohort (containing one phenotype, top left), while light blue represents the *STOPHCV* cohort (containing three phenotypes: cirrhosis, top right; viral load, bottom left; and HCC, bottom right). Each panel comprises six blocks. Blue, red, and yellow blocks represent additive, dominant, and recessive genetic effects, respectively. Grey blocks indicate interaction terms; light grey for interactions between HLA class I alleles (A, B, C) and ERAP1/2 allotypes, and dark grey for interactions between ERAP1 and ERAP2 allotypes. White areas correspond to covariates. Error bars denote 95% confidence intervals: those crossing zero are shown in grey, while those not crossing zero are highlighted in red.

Spontaneous clearance vs chronic infection For the HCV spontaneous clearance versus persistence dataset, a total of 3,434 samples are included. The phenotype was encoded as 1 for spontaneous clearance and 0 for persistent infection. We included 10 host principal components (PCs) and gender as covariates. Genetic components are incorporated based on the defined pathway, as illustrated in **Figure 2.4**. Association analysis was performed using a logistic regression model implemented in the `mapHS` software with the `CD_HS_purm_logistic` function. The number of permutations was set to 30, `max_inter` is set to 10, and all other parameters are kept at their default values. The overall results are summarised in Figure 4.1. From the Bayesian joint regression results, 11 variables retained non-shrunk coefficients. Among them, five variables whose 95% confidence intervals (coefficient $\pm\, 2 \times$ SD) did not cross zero are highlighted in red in Figure 4.1, indicating significant effects. The remaining six variables, with intervals crossing zero, are shown in grey. For these 11 variables of interest, we computed p-values and Bayes factors.

We select variables whose 95% confidence intervals excluded zero and which have $\log_{10}(\text{BF}) > 0$ as those showing stronger evidence of association. This filtering resulted in five variables, listed in Table 4.1. Aside from the strong covariate effects, four genetic features are identified as associated with infection outcome: two HLA alleles, one HLA–ERAP1 allotype interaction, and one HLA supertype. We further examine these genetic features in this cohort to explore their biological implications and statistical associations with spontaneous clearance.

HLA-related factors. Two HLA class II alleles, *HLA-DRB1*01:01* and *HLA-DQB1*03:01*, are significantly associated with spontaneous clearance in the joint model, with p-values of 8.64×10^{-6} and 0.033, respectively (Table 4.1). *HLA-DRB1*01:01* exhibited a dominant effect, whereas *HLA-DQB1*03:01* showed an additive effect. These findings are supported by validation using real data (Table 4.1a). For *HLA-DQB1*03:01*, clearance rates increased gradually with copy number: 36.2%, 44.7%, and 56.6% for 0, 1, and 2 copies, respectively, across 1800, 721, and 76 individuals, which is consistent with an additive model. In contrast, for *HLA-DRB1*01:01*, carriers of one or two copies showed similar clearance rates (53.9% and 54.5%, respectively), compared to 37.1% among non-carriers. Given that only 11 individuals carried two copies, the data support a dominant rather than additive mode of action, which aligns with the Bayesian model results. Additionally, the HLA supertype B58 also appeared to act dominantly, as only three individuals carried two copies. Carriers of one copy showed a higher clearance rate (43.0%) than non-carriers (38.2%), suggesting a potential protective effect.

Interaction term. For the interaction between *HLA-A*02:01* and the ERAP1 allotype *TEPIGVRNQE*, we analysed the real dataset to assess its biological relevance (Table 4.1b). All nine genotype combinations (0, 1, or 2 copies of each variant) are evaluated in terms of clearance rate and sample size. Focusing on subgroups with more than 100 samples (bolded in the table), we observe that individuals carrying one copy each of *HLA-A*02:01* and *TEPIGVRNQE* have the highest clearance rate (46.2%). In contrast, those carrying only one of the two variants showed clearance rates below 40%. This pattern suggests a synergistic effect between the HLA allele and the ERAP1 allotype. Biologically, this may result from improved peptide trimming by the ERAP1 allotype *TEPIGVRNQE*, generating peptides that bind more effectively to *HLA-A*02:01*. HLA complexes may then present stable peptides on the cell surface, enhancing T-cell recognition and immune activation.

Table 4.1: Bayesian joint regression results of HCV spontaneous clearance vs. chronic Infection

	Components	**Genetic effect**	**Coeff.**	**SE**	**AF**	***p-value***	$\log_{10}$ **BF**
PC3	covariates	/	-0.21	0.045	/	2.80×10^{-6}	3.81
HLA-DRB1*01:01	HLA allele	**dominant**	**0.61**	**0.14**	**0.06**	$\mathbf{8.64 \times 10^{-6}}$	**3.42**
HLA-A*02:01 and ERAP1:TEPIGVRNQE	interaction	/	0.21	0.095	0.23/0.16	0.024	0.35
HLA-DQB1*03:01	HLA allele	additive	0.27	0.13	0.16	0.033	0.24
Supertype-B58	HLA supertype	dominant	0.26	0.12	0.08	0.034	0.094

Bold values indicate genetic components that meet the significance threshold of $p < 5 \times 10^{-4}$.
The genetic effect (additive, dominant, or recessive) shown in the table represents the model that provides the best fit for each association.

Table 4.1.a. HLA-Related factors: clearance rates and sample sizes

HLA Allele	**Clearance rate (%)** **(0 copy / 1 copy / 2 copy)**	**Sample Sizes** **(0 copy / 1 copy / 2 copy)**
HLA-DQB1*03:01	36.2 / 44.7 / 56.6	1800 / 721 / 76
HLA-DRB1*01:01	37.1 / 53.9 / 54.5	2289 / 297 / 11
Supertype-B58	38.2 / 43.7 / 33.3	2150 / 444 / 3

Table 4.1.b. Interaction Between HLA-A*02:01 and ERAP1-TEPIGVRNQE in Relation to Spontaneous Clearance Rates and Sample Sizes. The first column indicates the copy number of HLA-A*02:01, and the second column indicates the copy number of the ERAP1 allotype (TEPIGVRNQE). The third column shows the spontaneous clearance rate calculated according to the corresponding HLA allele and ERAP1 allotype combination, and the final column provides the sample size for each group.

HLA-A*02:01	ERAP1-TEPIGVRNQE	Cases Proportion (%)	Sample size
0	0	**36.4**	**398**
1	0	**39.8**	**249**
2	0	37.4	40
0	1	**39.0**	**147**
1	1	**46.2**	**123**
2	1	46.8	22
0	2	31.9	15
1	2	60.7	17
2	2	71.4	5

Stophcv For the chronic HCV infection dataset, we analyse the associations with three phenotypes: cirrhosis, hepatocellular carcinoma (HCC), and HCV viral load. The results for each phenotype are presented individually in separate figures, followed by a combined summary of all three phenotypes and their effects in a table.

Cirrhosis For the cirrhosis phenotype (a case-control trait), patients with cirrhosis are coded as 1 and those without as 0. We include 20 host PCs, gender (female = 1, male = 0), body mass index (BMI), viral genotype (GT1 = 0, GT3 = 1), and age as covariates. Association analysis was performed using the Bayesian joint regression model implemented in the `mapHS` software with the `CD_HS_purm_logistic` function, with a maximum of 30 interactions and 10 permutations.

The joint results are summarised in Figure 4.1 (top right), where 15 variables are retained while others are shrunk to zero. Among these, seven variables have coefficients whose 95% credible intervals did not cross zero. These included three genetic factors: two HLA alleles (*HLA-DQB1*03:01* and *HLA-DRB1*01:01*) associated with reduced risk of cirrhosis, and one HLA supertype (A03) associated with increased risk. The remaining four significant variables are covariates: age, BMI, sex, and viral genotype. Older age and higher BMI are associated with increased cirrhosis risk. Similarly, infection with GT1 was associated with a higher cirrhosis risk compared to GT3, while female sex was associated with a lower risk compared to male sex.

HCV viral load For the HCV viral load phenotype, a quantitative trait, we analysed the data on a logarithmic scale using the maximum viral load recorded for each patient. The association analysis included 20 host principal components (PCs), gender, body mass index (BMI), viral genotype (GT1 or GT3), age, and cirrhosis status as covariates. A Bayesian linear regression is performed within a joint modelling framework using the `CD_HS_purm_linear` function in the `mapHS` software. The results are shown in Figure 4.1 (bottom left).

In this analysis, 12 variables are retained (i.e., not shrunk to zero). Among these, five variables have 95% confidence intervals that excluded zero, indicating more robust evidence for an association. These included two HLA alleles, *HLA-DQA1*03:01* and *HLA-DQB1*03:01*, both associated with lower viral load. The remaining three variables with related associations are covariates: older age is associated with higher viral load; the presence of cirrhosis is associated with lower viral load, possibly due to

a reduced capacity for viral replication in compromised hepatocytes; and female sex is associated with a lower viral load than male sex in the chronic stage of infection.

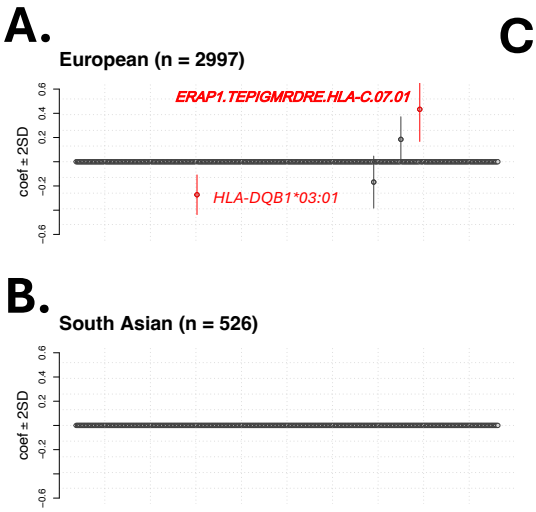
Hepatocellular Carcinoma For the HCC phenotype, a case-control trait, patients with HCC are encoded as 1 and those without as 0. We include 20 host PCs, gender, BMI, viral genotype (GT1 or GT3), age, and cirrhosis status as covariates. A Bayesian logistic regression is performed within a joint modelling framework using the `CD_HS_purm_logistic` function in the `mapHS` software. The results are presented in Figure 4.1 (bottom right). In this analysis, 10 variables are retained, suggesting their potential relevance to HCC. Among these, four variables have 95% confidence intervals that exclude zero. One of these is the HLA allele *HLA-B*07:02*, which is associated with a lower risk of HCC. The remaining three well-supported variables are covariates: older age is associated with an increased risk of HCC; the presence of cirrhosis is associated with a higher likelihood of HCC; and female patients have a lower risk of HCC than males during chronic infection.

In summary, we evaluate three phenotypes: cirrhosis, viral load, and HCC, and we summarise the corresponding signals, p-values, and Bayes factors in Table 4.2. The table includes signals with 95% confidence intervals excluding zero and $\log_{10}(\text{BF}) > 0$, indicating stronger evidence for an association. Based on our predefined criteria and excluding covariates, only one genetic feature showed compelling evidence: the *HLA-DQB1*03:01* allele is associated with a reduced risk of cirrhosis, exhibiting an additive effect. Although *HLA-DQB1*03:01* is also associated with lower viral load, the evidence ($p = 0.001$) did not meet the predefined significance threshold of $p < 5 \times 10^{-4}$. Nonetheless, other genetic features may still contribute to the progression of chronic HCV infection.

In addition to genetic factors, several covariates demonstrated consistent effects. Higher BMI is associated with an increased risk of cirrhosis, consistent with previous reports (Byrne and Wild, 2010). Older age is strongly and consistently associated with poorer outcomes across all three phenotypes, showing an increased risk of cirrhosis and HCC, as well as higher viral load. Furthermore, female sex is associated with more favourable outcomes, showing consistently lower rates of cirrhosis, HCC, and viral load compared to male sex.

Table 4.2: **Summary of Bayesian joint regression analyses for virrhosis, viral Load, and HCC in the STOP-HCV cohort**

	Components	**Genetic effect**	**Coeff.**	**SE**	**AF**	***p-value***	$\log_{10}$ **BF**
Cirrhosis							
AGE	covariates	/	0.777	0.049	/	$< 10^{-16}$	52.515
Virus-GT	covariates	/	1.436	0.106	/	$< 10^{-16}$	37.544
BMI	covariates	/	0.258	0.045	/	9.95×10^{-9}	6.157
SEX	covariates	/	-0.438	0.099	/	8.85×10^{-6}	3.407
HLA-DQB1*03:01	**HLA allele**	**additive**	**-0.405**	**0.094**	**0.13**	$\mathbf{1.50 \times 10^{-5}}$	**3.195**
A03	HLA supertype	additive	0.194	0.088	0.27	0.028	0.283
HLA-DRB1*01:01	HLA allele	additive	-0.267	0.131	0.07	0.042	0.135
ERAP1.TERIGMKDRQ_ERAP2.PTNL	interaction	/	0.161	0.081	0.14/0.48	0.046	0.100
HCC							
AGE	covariates	/	0.911	0.077	/	$< 10^{-16}$	27.990
Cirrhosis	covariates	/	2.211	0.196	/	$< 10^{-16}$	25.208
SEX	covariates	/	-0.745	0.150	/	6.97×10^{-7}	4.411
HLA-B*07:02	HLA allele	additive	0.280	0.124	0.15	0.023	0.445
PC1	covariates	/	0.154	0.078	/	0.048	0.192
Viral Load							
AGE	covariates	/	0.102	0.021	/	9.94×10^{-7}	4.120
Cirrhosis	covariates	/	-0.173	0.042	/	3.73×10^{-5}	2.626
Sex	covariates	/	-0.169	0.043	/	7.55×10^{-5}	2.339
HLA-DQB1*03:01	HLA allele	dominant	-0.154	0.048	0.13	0.001	1.215
HLA-DQA1*03:01	HLA allele	additive	-0.110	0.041	0.15	0.008	0.492



C.

ERAP1.TEPIGMRDRE * HLA-C.07.01

ERAP1	HLA	Cases/controls	Prop (%)
absent	absent	656/909	58%
present	absent	108/165	60%
absent	present	298/393	57%
present	present	45/97	68%

HLA-DQB1*03:01	Cases/controls	Prop (%)
0 copy	791/1220	61%
1 copy	284/316	53%
2 copies	32/28	47%

Figure 4.2: A. Fine-mapping results using *mapHS* in the European subgroup. B. Fine-mapping results using *mapHS* in the South Asian subgroup. C. Detailed results for the European subgroup. The top panel shows the interaction term *ERAP1.TEPIGMRDRE* × *HLA-C*07:01*, including the case/control sample sizes and the proportion of individuals without cirrhosis. The bottom panel shows the results for *HLA-DQB1*03:01*, including the case/control sample sizes and the proportion of individuals without cirrhosis.

Sensitive analysis In the analysis of cirrhosis phenotypes in the original cohort, which included both European and South Asian participants, we identified a novel genetic signal, the HLA allele *HLA-DQB1*03:01*, associated with chronic infection, related cirrhosis. To minimise potential confounding due to population structure, we conducted a sensitivity analysis by stratifying the cohort by ancestry and rerunning the analytical pipeline. The results of these subgroup analyses are shown in Figure 4.2A and B for the European and South Asian groups, respectively, where the selected genetic features are visualised. In the European subgroup, the fine-mapping model selected four genetic features. Among these, two signals, *HLA-DQB1*03:01* and the interaction term *ERAP1.TEPIGMRDRE* × *HLA-C*07:01*, which showed strong statistical support and were therefore highlighted in red in the figure. The remaining two signals had 95% confidence intervals crossing zero and were therefore considered less reliable. In contrast, the South Asian subgroup did not show evidence of any

important genetic features, as the fine-mapping model shrank nearly all coefficients toward zero.

For the European group, we further examined the two supported signals (Figure 4.2C), and the corresponding coefficients and standard deviations are summarised in Table 4.3. One signal corresponds to the HLA allele *HLA-DQB1*03:01*, which appears to have a protective effect. This finding is consistent with the result observed in the analysis of the combined cohort. Under an additive genetic model, individuals carrying this allele showed a stepwise decrease in cirrhosis prevalence: individuals with 0 copies had a cirrhosis rate of 61%, those with 1 copy had 52%, and those with 2 copies had 47%. This pattern suggests that the presence of *HLA-DQB1*03:01* may confer protection against the progression to cirrhosis.

The second signal represents a novel interaction between the *ERAP1.TEPIGMRDRE* $\times$ *HLA-C*07:01*. Compared with the baseline group, in which both variants are absent and the cirrhosis rate is approximately 58%, the presence of the ERAP1 allotype alone is associated with a cirrhosis rate of around 60%. Notably, when both the ERAP1 allotype and *HLA-C*07:01* are present, the cirrhosis rate increases to approximately 68%. One possible explanation is that the *ERAP1 allotype (TEPIGMRDRE)* may alter peptide trimming, thereby affecting peptide presentation and binding by HLA molecules, particularly *HLA-C*07:01*. This altered antigen presentation could influence immune recognition and consequently increase susceptibility to cirrhosis. Interestingly, this interaction signal was observed only in the European subgroup and not in the South Asian subgroup or the combined cohort. One possible explanation is the difference in allele frequency between populations: *HLA-C*07:01* has a frequency of approximately 17% in the European population but only around 8% in the South Asian population. The lower frequency in the South Asian cohort may reduce statistical power to detect this interaction effect.

	Coef	sd	p-value	$\log_{10}$BF
HLA-DQB1*03:01	-0.27	0.082	0.0008	2.30
ERAP1.TEPIGMRDRE $\times$ HLA-C*07:01	0.43	0.13	0.001	2.21

Table 4.3: Significant genetic features in the antigen presentation pathway associated with cirrhosis in the European subgroup.

Summary of notable signals in the HCV datasets: Spontaneous Clearance vs. Chronic Infection and the STOP-HCV Cohort Across the HCV datasets, several signals consistently appeared to influence infection outcomes and

disease progression. To provide an integrated overview, we summarise findings from both the acute phase, represented by the spontaneous clearance versus chronic infection dataset, and the chronic phase, represented by the STOP-HCV cohort, which includes cirrhosis, HCC, and viral load phenotypes. Among the genetic features observed, the alleles HLA-DQB1*03:01 and HLA-DRB1*01:01 emerge as particularly noteworthy. Although these alleles do not meet the most conservative significance thresholds in every individual phenotype, their recurrent appearance across multiple analyses and their well-established immunological roles during HCV infection make them of strong biological interest. As illustrated in Figure 4.3, both alleles are associated with a higher likelihood of viral clearance during acute infection. In the chronic phase, they are further associated with a reduced risk of developing cirrhosis, and specifically, HLA-DQB1*03:01 is also linked to lower viral load levels. While HLA-DQB1*03:01 remain a consistent protective factor across phenotypes, the effect of HLA-DRB1*01:01 on viral load is less pronounced, with its coefficient being shrunk in the joint model. Taken together, these findings suggest that certain HLA class II alleles, particularly HLA-DQB1*03:01 and HLA-DRB1*01:01, may enhance viral antigen presentation and promote effective immune clearance of HCV, thereby conferring long-term protection against disease progression.

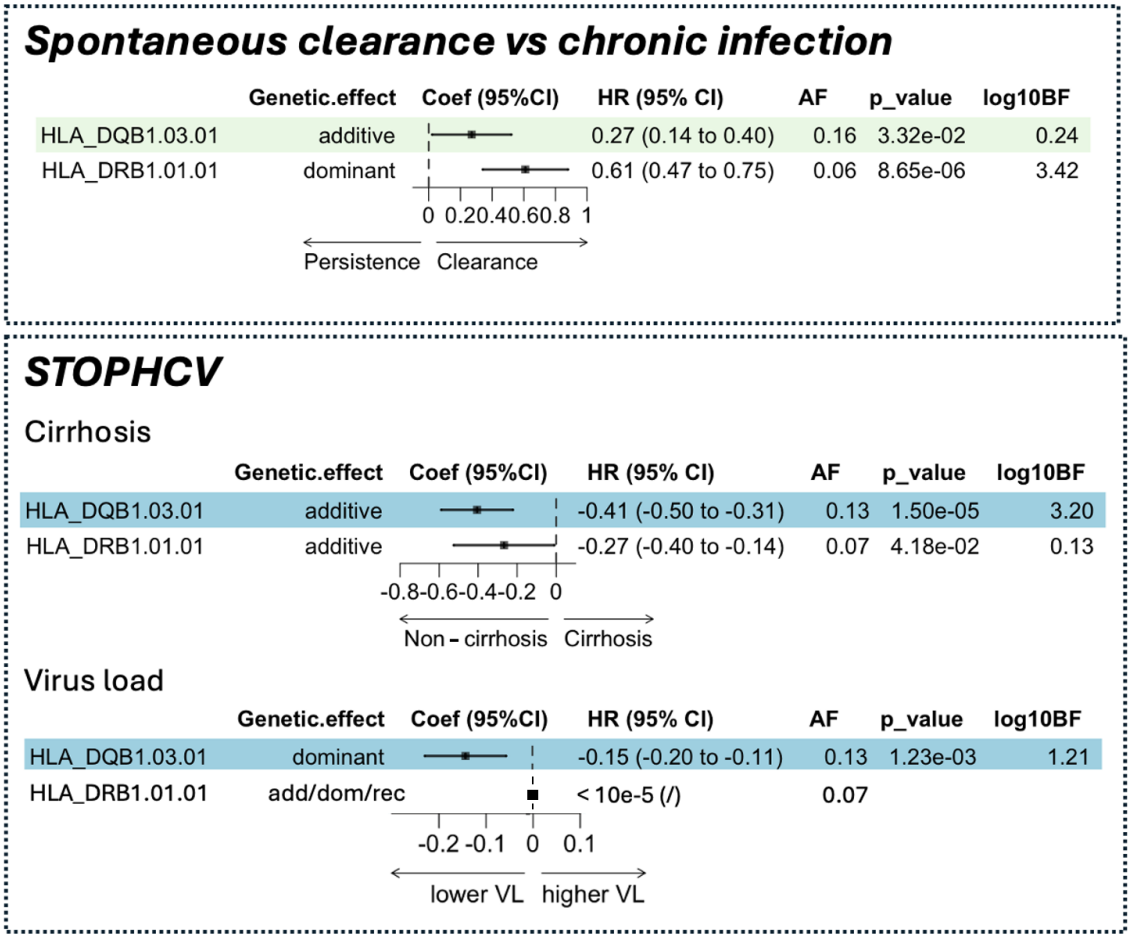


Figure 4.3: **Summary of notable HLA associations (HLA-DQB1*03:01 and HLA-DRB1*01:01) across HCV datasets.** The top panel illustrates the effect sizes of both alleles in the spontaneous clearance versus chronic infection dataset, and the bottom panel shows the corresponding associations observed in the STOP-HCV cohort.

4.3.1.2 Malaria Genomic Epidemiology Network

For the Malaria Genomic Epidemiology Network (MalariaGEN) datasets, we hypothesised that distinct Plasmodium falciparum parasite populations across Africa could exert differential genetic and immunological effects on severe malaria outcomes. To account for geographic and potential parasite strain variation, we stratified the samples into two regional groups: a western African group (Mali, Nigeria, Cameroon, Ghana, Burkina Faso, and The Gambia) and an eastern African group (Tanzania, Kenya, and Malawi). The phenotype is a case–control trait, with individuals with severe malaria encoded as 1 (cases) and those with mild or asymptomatic malaria encoded as 0 (controls). We included gender and the top 5 PCs as covariates. A

Bayesian joint analysis is performed using a logistic regression model implemented via the `CD_HS_purm_logistic` function in the `mapHS` software. The results are presented in Figure 4.4.

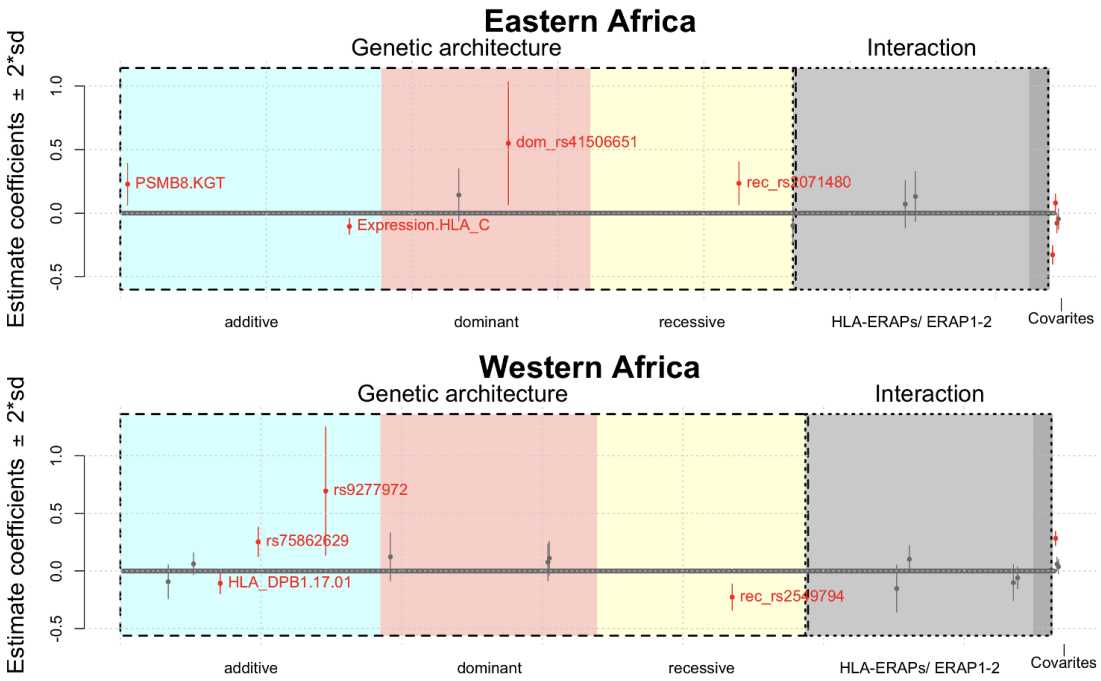


Figure 4.4: **Bayesian joint regression results of the Malaria Genomic Epidemiology Network (MalariaGEN) cohort**. The cohort is divided into two regional groups: Eastern Africa and Western Africa. The top panel shows the results for Eastern Africa, and the bottom panel presents the results for Western Africa. The figure comprises six blocks. Blue, red, and yellow blocks represent additive, dominant, and recessive genetic effects, respectively. Grey blocks represent interaction terms: light grey for interactions between HLA class I alleles (A, B, C) and ERAP1/2 allotypes, and dark grey for interactions between ERAP1 and ERAP2 allotypes. White areas denote covariates. Error bars represent 95% confidence intervals; those crossing zero are shown in grey, while those not crossing zero are highlighted in red.

Across the MalariaGEN analyses of Eastern and Western Africa, we identified several genetic components associated with severe malaria, as shown in Figure 4.4 and Table 4.4. The results summarised in the table focus on variants with 95% confidence intervals excluding zero and $\log_{10}\mathrm{BF} > 0$, indicating statistical evidence of association. In the following sections, we interpret the findings from the Western and Eastern African cohorts separately and discuss their relevance in the context of previous studies.

Western African Cohort In the Western African cohort, two SNPs met our stringent significance threshold ($p < 5 \times 10^{-4}$): *rs75862629*, located in the promoter region of the ERAP2 gene, and *rs2549794*, located in an intronic region of ERAP2. For *rs75862629*, the G allele (effect allele) was associated with increased risk of severe malaria under an additive model, indicating a higher risk per additional copy of the G allele. For *rs2549794*, individuals homozygous for the T allele (effect allele) demonstrated partial protection under a recessive model. We also detected one HLA allele, *HLA-DPB1*17:01*, which showed a modest protective association ($p = 2.2 \times 10^{-2}$) but did not reach the conservative significance threshold. Additionally, the T allele (effect allele) of *rs9277972*, an intronic SNP in the TAPBP gene, appeared to be associated with increased risk.

We further examined these two *ERAP2* SNPs in the real dataset to assess genotype–phenotype relationships (Table 4.3a). For rs75862629, individuals with one copy have a severe malaria rate of approximately 62%, and those with two copies have a rate of about 68%, compared to 56% among non-carriers. This gradient supports an additive model, where each risk allele incrementally increases susceptibility. For rs2549794, which exhibited a recessive effect, individuals homozygous for the variant have a severe malaria rate of 55%, compared to 59% among carriers of one or no copies, suggesting a protective effect in the homozygous state.

A comparison of these SNPs between the Western and Eastern African cohorts revealed consistent allele frequencies for rs75862629 (approximately 10% in both regions). In contrast, rs2549794 is more frequent in Western Africa (67%) than in Eastern Africa (58%). Population data from gnomAD indicate that rs75862629 is most prevalent in African populations (11%), compared to Europeans (5%) and South Asians (7%). rs2549794 is common and broadly distributed, with frequencies of approximately 64% in Africans, 62% in Europeans, and 73% in Asians.

Previous studies provide functional and evolutionary context for these associations. The promoter SNP rs75862629 is known to modulate the relative expression of ERAP1 and ERAP2. The minor G allele is associated with reduced ERAP2 expression and increased ERAP1 expression, altering the peptide repertoire presented by HLA class I molecules (Paladini et al., 2019). This ERAP1/ERAP2 balance influences antigen processing and may shape immune responses to *Plasmodium falciparum*. The second SNP, rs2549794, also in the *ERAP2* region, has been highlighted in evolutionary studies. Klunk et al. (Klunk et al., 2022) identified it as a top candidate for positive selection during the Black Death pandemic, possibly by enhancing immune responses to *Yersinia pestis*. Hamilton et al. (Hamilton et al., 2023) further linked

the T allele to reduced ERAP2 expression and increased susceptibility to respiratory infections, suggesting a broader immunomodulatory role maintained by balancing selection. Although the strength of selection during the Black Death has been debated, the recurrence of the *ERAP2* region in our malaria analysis underscores its immunological importance.

Eastern African Cohort In the Eastern African cohort, no genetic variant met the strict significance threshold. The top signal corresponded to HLA-C protein expression levels, where higher expression is associated with protection against severe malaria. This parallels observations in HIV infection, where elevated HLA-C expression correlates with improved viral control (Apps et al., 2013). Although no prior malaria studies have directly linked HLA-C protein expression to disease severity, this trend suggests that enhanced antigen presentation via HLA-C protein expression may play a protective role in malaria pathogenesis.

Summary of the findings in MalariaGEN datasets. For the MalariaGEN datasets, based on our broader significance threshold ($p < 5 \times 10^{-4}$), we identified two signals within the *ERAP2* region in the West African cohort. The first variant, *rs75862629*, acts as a potential risk factor under an additive model, while the second variant, *rs2549794*, appears to confer a protective effect under a recessive model. Although neither signal reaches the conventional genome-wide significance threshold of $p < 1 \times 10^{-8}$, both represent potentially meaningful associations that may influence susceptibility to severe malaria. Further validation using independent replication cohorts or functional assays will be required to confirm these findings and clarify their biological relevance.

Table 4.4: **Bayesian joint regression results from the MalariaGEN dataset across western and eastern African regions.**

	Components	Genetic effect	Effect allele	Coeff.	SE	AF	*p-value*	$\log_{10} BF$
Western Africa: Mali, Nigeria, Cameroon, Ghana, Burkina Faso, and Gambia								
PC1	covariates	/	/	0.282	0.027	/	$< 10^{-16}$	22.417
rs75862629	**ERAP2 (Promoter)**	**additive**	G	**0.252**	**0.065**	**0.10**	9.8×10^{-5}	**2.336**
rs2549794	**ERAP2 (Intron)**	**recessive**	T	**-0.227**	**0.059**	**0.670**	1.1×10^{-4}	**2.299**
rs9277972	TAPBP (Intron)	additive	T	0.694	0.284	0.01	1.45×10^{-2}	0.369
HLA-DPB1*17:01	HLA allele	additive	/	-0.107	0.0467	0.21	2.20×10^{-2}	0.214
PC2	covariates	/	/	0.060	0.0279	/	3.29×10^{-2}	0.0659
Eastern Africa: Tanzania, Kenya, and Mali								
PC1	covariates	/	/	-0.328	0.037	/	$< 10^{-16}$	15.856
HLA-C	HLA protein expression	/	/	-0.104	0.0318	/	1.00×10^{-3}	1.456
rs2071480	TAP1	recessive	T	0.235	0.0857	0.42	6.22×10^{-3}	0.747
PSMB8.KGT	PSMB8 allotypes	additive	/	0.228	0.0840	0.08	6.69×10^{-3}	0.719
rs41506651	ERAP2 (Synonymous)	dominant	T	0.550	0.246	0.01	2.57×10^{-2}	0.213
PC3	covariates	/	/	0.080	0.0360	/	2.58×10^{-2}	0.211
PC4	covariates	/	/	-0.079	0.0377	/	3.67×10^{-2}	0.0830

Table 4.3.a: **Summary of two significant SNPs identified in the Western African cohort, including case rates and sample sizes.**

SNP	Chr	Nearest gene	VEP annotation	Genetic model	Cases proportion (%)	Sample size
rs75862629	5	ERAP2	Promoter	additive (0 / 1 / 2 copies)	56 / 62 / 68	5357 / 1127 / 72
rs2549794	5	ERAP2	Intron	recessive (0 or 1 / 2 copies)	59 / 55	3574 / 2982

Table 4.3.b: **Comparison of Bayesian joint regression results for SNPs rs75862629 and rs2549794 across Western and Eastern African cohorts in the MalariaGEN dataset.**

SNP	Western Africa				Eastern Africa			
	Genetic effect	Coeff.	SE	AF	Genetic effect	Coeff.	SE	AF
rs75862629	additive	0.252	0.065	0.10	add/dom/rec	$< 1 \times 10^{-5}$	$< 1 \times 10^{-5}$	0.10
rs2549794	recessive	-0.227	0.059	0.67	add/dom/rec	$< 1 \times 10^{-5}$	$< 1 \times 10^{-5}$	0.58

4.3.1.3 UK Biobank serological panel

For the UK Biobank serological panel, we performed genetic association analyses using two approaches, depending on the type of trait: case–control and quantitative analyses. For the case–control analysis, participants are classified as seropositive or seronegative. We included only antigens with a seroprevalence greater than 15% to ensure adequate statistical power for identifying associated loci. For the quantitative analysis, we used antibody median fluorescence intensity (MFI) measurements as continuous traits. Because serological assays are susceptible to low-level cross-reactivity with non-specific antibodies that may not reflect true infection status, the quantitative analysis is restricted to samples exceeding the seropositivity thresholds recommended by UK Biobank. We aimed to identify genetic variants underlying variation in antibody-mediated immune responses within the seropositive population. Besides, due to the skewed distribution of antibody MFI values, we applied a quantile transformation before performing association tests. For both analyses (case–control and quantitative), we included sex, age, and the first 20 genetic principal components as covariates.

Given the large number of antigen-specific phenotypes, we grouped the results into four major pathogen categories to facilitate interpretation and visualisation: (1) Herpesviridae: HSV-1, HSV-2, EBV, CMV, HHV-6, HHV-7, and VZV; (2) Polyomaviridae: JCV, BKV, and MCV; (3) Bacteria: Chlamydia trachomatis and Helicobacter pylori; (4) Parasite: Toxoplasma gondii. We summarised associations with $\log_{10} \mathrm{BF} > 0$ and a 95% confidence interval excluding zero in the following tables as evidence of relevance. Detailed results for both the case–control and quantitative (MFI) analyses are presented below.

Seroprevalence in Cases-controls In the UK Biobank serological dataset, the serostatus of each pathogen-specific antigen is recorded as either "True" (seropositive) or "False" (seronegative). For analytical purposes, we encoded these values as binary variables, assigning 1 to "True" and 0 to "False".

Herpesviridae We summarised the results for all members of the *Herpesviridae* family namely CMV, EBV, HHV-6A, HHV-6B, HHV-7, HSV-1, HSV-2, and VZV in Figure 4.5 and Table 4.5. Overall, we did not identify compelling evidence for genome-wide significant genetic effects on seropositivity across this viral group. Nevertheless, several loci and interaction terms exhibited suggestive associations that

may still contribute to inter-individual variability in serological response and thus merit discussion.

For CMV, in addition to demographic covariates (age, sex, and population principal components), the allele HLA-DPB1*04:02 is associated with an increased likelihood of seropositivity, suggesting a possible role in antigen presentation or immune recognition of CMV antigens. For EBV, modest signals are detected in the PSMB8 allotype (KG variant) as well as in an interaction between HLA-A*01:01 and the ERAP1.TEPIDMKDRQ haplotype, both showing positive associations with seropositivity. In HHV-6A, two HLA alleles are implicated: HLA-DRB1*04:01, which showed a risk effect, and HLA-DQA1*03:01, which appeared protective. Additionally, interaction effects are identified between HLA-C*07:01 and ERAP1.IERIGMRDRE, as well as between ERAP1.IERIGMRDRE and ERAP2.TKL, suggesting that epistatic relationships between antigen-processing enzymes may influence viral recognition. For HHV-6B, two HLA alleles are highlighted: HLA-DQA1*01:01, associated with increased seropositivity, and HLA-DPA1*02:01, which showed a protective effect. The interaction between HLA-C*05:01 and ERAP2.TNL is also identified as a risk factor, whereas the B44 HLA supertype is associated with a protective trend. In HHV-7, we observed several modest associations, including HLA-DQB1*05:01 and ERAP1.TEPIGMRDRE allotype, both showing negative effects on seroprevalence, as well as an interaction between HLA-A*24:02 and ERAP1.TERMGMKDRQ, indicating potential joint effects on antigen processing and presentation. For HSV-1, no specific genetic components demonstrated significant effects on seroprevalence beyond age and population structure, consistent with the high baseline exposure rate in the population. In HSV-2, the HLA-DQA1*01:01 allele is identified as a risk factor under a recessive model, while the interaction between HLA-A*11:01 and ERAP1 is identified as TEPIGMRDRE, which appeared protective, suggesting genetic modulation of susceptibility to HSV-2 infection. Finally, for VZV, several HLA alleles are associated with serostatus: HLA-A*01:01 showed a protective effect, whereas HLA-DQB1*05:01, HLA-DRB1*15:01, and HLA-B*44:02 are associated with increased seropositivity. These findings may reflect allele-specific differences in T-cell recognition of VZV epitopes or immunological memory following exposure.

Overall, while few associations reached stringent significance thresholds, the recurrent involvement of antigen-processing genes (HLA, ERAP1, and ERAP2) highlights a shared immunogenetic architecture underlying humoral responses to herpesviruses.

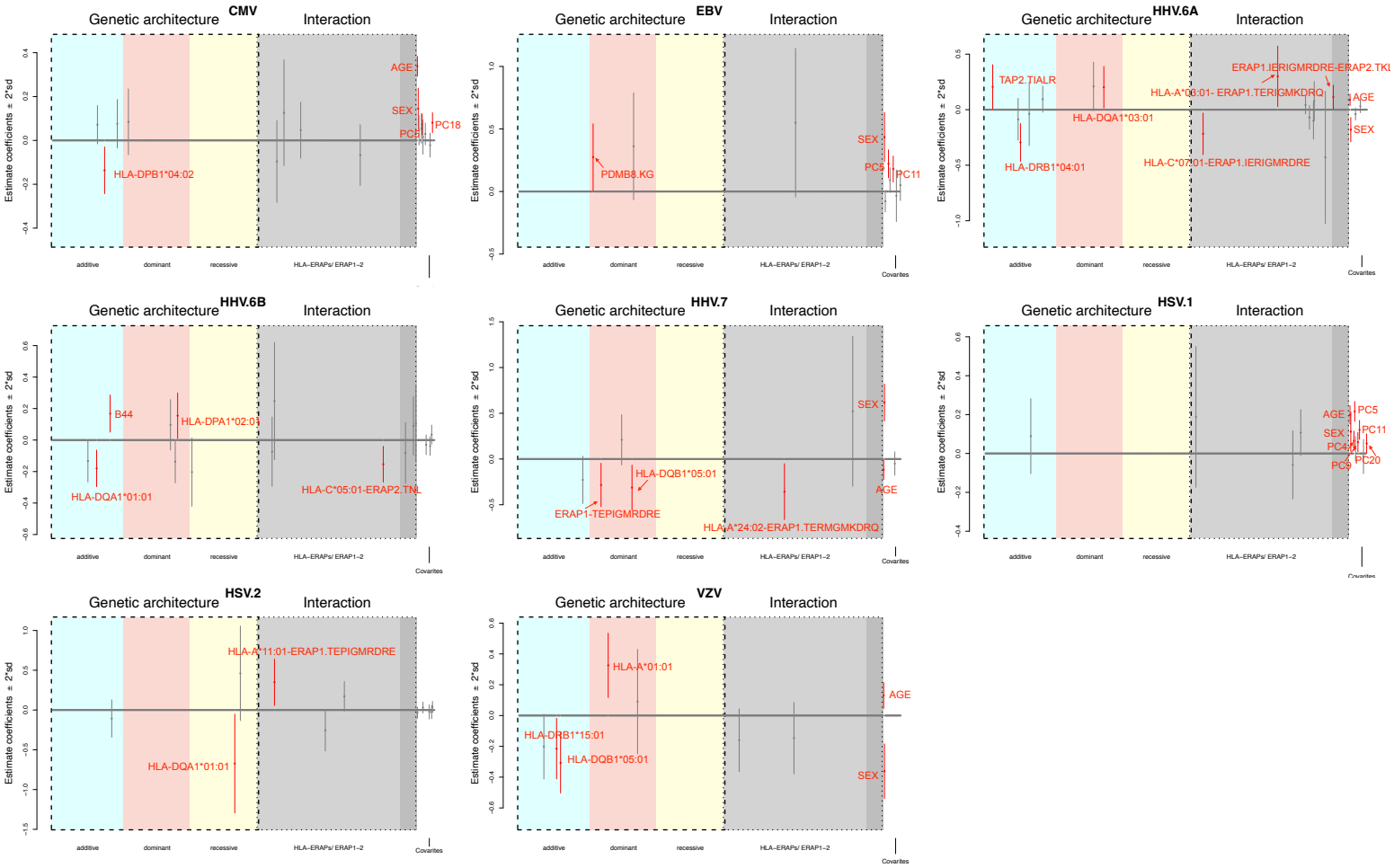


Figure 4.5: **Bayesian joint regression results from the UK Biobank serological panel for the *Herpesviridae* family.** Thanalysis includes the following pathogens: CMV, EBV, HHV-6A, HHV-6B, HHV-7, HSV-1, HSV-2, and VZV. The figure comprises six blocks. Blue, red, and yellow blocks represent additive, dominant, and recessive genetic effects, respectively. Grey b locks represent interaction terms: light grey for interactions between HLA class I alleles (A, B, C) and ERAP1/2 allotypes, and dark grey for interactions between ERAP1 and ERAP2 allotypes. White areas denote covariates. Error bars represent 95% confidence intervals; those crossing zero are shown in grey, while those not crossing zero are highlighted in red.

Table 4.5: **Bayesian joint regression results from the UK Biobank serological panel for the *Herpesviridae* family.** This group includes the following pathogens: CMV, EBV, HHV-6A, HHV-6B, HHV-7, HSV-1, HSV-2, and VZV. The table summarises associations with $\log_{10}\text{BF} > 0$.

	Components	**Genetic effect**	**Coeff.**	**SE**	**AF**	***p-value***	$\log_{10}$**BF**
CMV (58.3%)							
AGE	covariates	/	0.34	0.02	/	$< 1 \times 10^{-16}$	45.67
PC18	covariates	/	0.08	0.02	/	5.02×10^{-4}	1.64
Sex	covariates	/	0.15	0.05	/	0.002	1.05
PC5	covariates	/	0.07	0.02	/	0.003	0.98
HLA-DPB1*04:02	HLA allele	additive	-0.14	0.05	0.105	0.012	0.39
EBV (19%)							
Sex	covariates	/	0.43	0.10	/	1.55×10^{-5}	3.20
PC5	covariates	/	0.22	0.06	/	1.04×10^{-4}	2.46
PC11	covariates	/	0.18	0.05	/	6.82×10^{-4}	1.74
PSMB8.KG	PSMB8 allotype	dominant	0.27	0.14	0.13	0.044	0.21
HLA-A*01:01_ERAP1.TEPIDMKDRQ	interaction	/	0.55	0.30	0.19/0.067	0.070	0.05
PC7	covariates	/	0.09	0.05	/	0.070	0.04
PC1	covariates	/	-0.08	0.04	/	0.077	0.01
HHV.6A (77%)							
HLA-DRB1*04:01	HLA allele	additive	-0.29	0.09	0.11	0.001	1.55
AGE	covariates	/	0.09	0.03	/	0.001	1.41
Sex	covariates	/	-0.18	0.06	/	0.001	1.31
HLA-C*07:01_ERAP1.IERIGMRDRE	interaction	/	-0.22	0.10	0.18/ 0.13	0.023	0.22
HLA-A*03:01_ERAP1.TERIGMKDRQ	interaction	/	0.30	0.14	0.14/0.08	0.031	0.11
HLA-DQA1*03:01	HLA allele	dominant	0.20	0.10	0.20	0.034	0.07

Continued on next page

	Components	Genetic effect	Coeff.	SE	AF	*p-value*	$\log_{10}$ BF
ERAP1.IERIGMRDRE_ERAP2.TKL	interaction	/	0.11	0.06	0.13/0.46	0.040	0.01
HHV.6B (79%)							
HLA-DQA1*01:01	HLA allele	additive	-0.18	0.06	0.14	0.002	1.10
B44	HLA supertype	additive	0.17	0.06	0.27	0.005	0.79
HLA-C*05:01_ERAP2.TNL	interaction	/	-0.15	0.06	0.11/0.45	0.008	0.62
HLA-DPA1*02:01	HLA allele	dominant	0.15	0.07	0.14	0.036	0.06
HHV.7 (95%)							
Sex	covariates	/	0.62	0.10	/	2.05×10^{-9}	6.73
HLA-DQB1*05:01	HLA allele	dominant	-0.31	0.13	0.12	0.013	0.66
ERAP1.TEPIGMRDRE	ERAP1 allotype	dominant	-0.28	0.12	0.13	0.019	0.51
HLA-A*24:02_ERAP1.TERMGMKDRQ	interaction	/	-0.36	0.16	0.08/0.22	0.022	0.45
AGE	covariates	/	-0.12	0.06	/	0.032	0.32
HSV.1 (70%)							
PC5	covariates	/	0.22	0.03	/	2.22×10^{-16}	13.67
AGE	covariates	/	0.20	0.02	/	2.22×10^{-16}	13.59
PC11	covariates	/	0.12	0.02	/	1.18×10^{-6}	4.13
PC4	covariates	/	0.07	0.03	/	0.010	0.48
PC9	covariates	/	0.06	0.03	/	0.022	0.20
Sex	covariates	/	0.11	0.05	/	0.030	0.07
HSV.2 (16%)							
HLA-A*11:01_ERAP1.TEPIGMRDRE	interaction	/	0.35	0.15	0.06/0.14	0.019	0.35
HLA-DQA1*01:01	HLA allele	recessive	-0.67	0.32	0.15	0.034	0.14
VZV (92%)							

Continued on next page

	Components	Genetic effect	Coeff.	SE	AF	*p-value*	$\log_{10}$ BF
Sex	covariates	/	-0.36	0.09	/	6.37×10^{-5}	2.64
HLA-DQB1*05:01	HLA allele	additive	-0.31	0.10	0.12	0.002	1.28
HLA-A*01:01	HLA allele	additive	0.33	0.11	0.2	0.002	1.25
Age	covariates	/	0.13	0.04	/	0.002	1.23
HLA-DRB1*15:01	HLA allele	additive	-0.22	0.10	0.14	0.033	0.26
HLA-B*44:02	HLA allele	additive	-0.20	0.11	0.11	0.061	0.04

Polyomaviridae In the *Polyomaviridae* group, we analysed three human polyomaviruses: BK virus (BKV), JC virus (JCV), and Merkel cell polyomavirus (MCV). The results of the Bayesian logistic regression analysis are presented in Figure 4.6, and the associations with $\log_{10} \mathrm{BF} > 0$ are summarised in Table 4.6. Overall, we detected several compelling associations that met the predefined significance threshold ($p < 5 \times 10^{-4}$), indicating notable genetic effects on seropositivity within this viral family. For JCV, the allele HLA-DRB1*15:01 emerged as the most significant genetic factor, showing a strong association with increased seropositivity. Additional contributing variants included HLA-DQA1*01:03, HLA-DQB1*03:01, and HLA-DRB4*01:03, along with a signal in the antigen-processing gene TAP2 (*AVARL* allotype). These findings suggest that antigen presentation and peptide transport pathways may play a central role in shaping immune responses to JCV. In MCV, HLA-DRB1*15:01 is again identified as a highly significant risk factor, reinforcing its shared immunogenetic relevance across the polyomaviruses. Two other HLA alleles are noteworthy: HLA-DRB1*01:01, which appeared protective, and HLA-DQB1*03:02, which showed a risk effect. Moreover, HLA-B*07:02 exhibited a modest negative association with seropositivity, and an interaction between HLA-B*44:03 and the ERAP2.TNL allotype is observed, implying that peptide trimming and antigen presentation mechanisms contribute to MCV serological variability. For BKV, age and sex remained significant covariates, indicating demographic influences on seroprevalence. Two interaction terms involving HLA-C alleles and ERAP1 allotypes showed moderate effects, although they did not reach the stringent significance threshold. These interactions suggest possible antigen-processing influences that warrant further investigation.

Collectively, these results highlight consistent involvement of HLA class II alleles, particularly HLA-DRB1*15:01, and components of the antigen-processing pathway in modulating humoral immune responses to polyomaviruses.

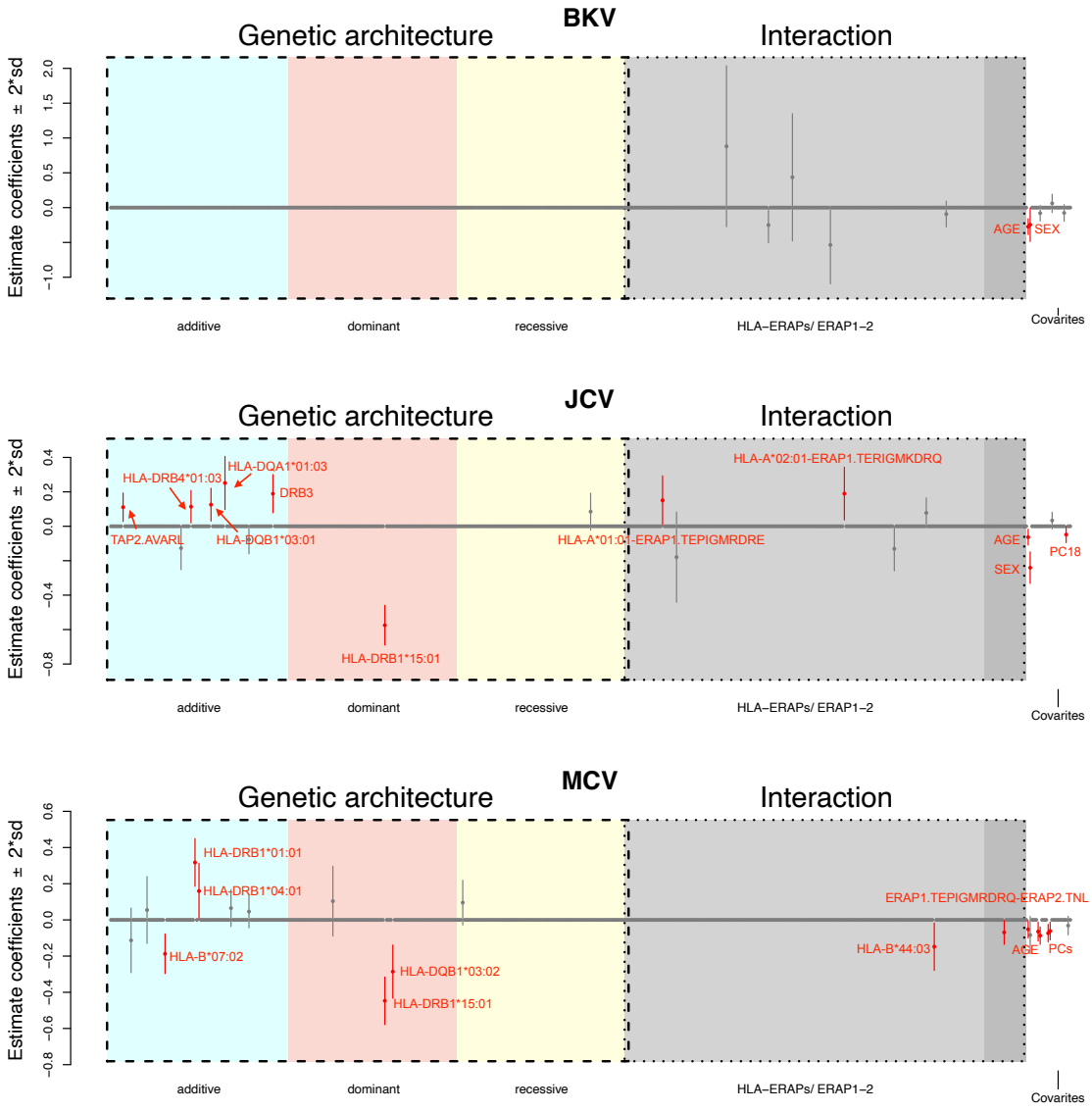


Figure 4.6: **Bayesian joint regression results from the UK Biobank serological panel for the *polyomavirus* family**. The analysis includes the following pathogens: BKV, JCV, and MCV. The figure comprises six blocks. Blue, red, and yellow blocks represent additive, dominant, and recessive genetic effects, respectively. Grey blocks represent interaction terms: light grey for interactions between HLA class I alleles (A, B, C) and ERAP1/2 allotypes, and dark grey for interactions between ERAP1 and ERAP2 allotypes. White areas denote covariates. Error bars represent 95% confidence intervals; those crossing zero are shown in grey, while those not crossing zero are highlighted in red.

Table 4.6: **Bayesian joint regression results from the UK Biobank serological panel for the *polyomavirus* family**. The analysis includes the following pathogens: BKV, JCV, and MCV. The table summarises associations with $\log_{10}\mathrm{BF} > 0$.

	Components	**Genetic effect**	**Coeff.**	**SE**	**AF**	***p-value***	$\log_{10}$ **BF**
BKV (95%)							
Age	covariates	/	-0.27	0.06	/	2.13×10^{-6}	3.97
Sex	covariates	/	-0.25	0.12	/	0.045	0.22
HLA-C*07:02_ERAP1.TERIGVKNQQ	interaction	/	-0.25	0.13	0.16/0.22	0.052	0.17
HLA-C*03:04_ERAP1.TEPIDMKDRQ	interaction	/	-0.54	0.28	0.08/0.07	0.060	0.12
JCV (57%)							
HLA-DRB1*15:01	**HLA allele**	**dominant**	**-0.57**	**0.06**	**0.14**	$\mathbf{< 1 \times 10^{-16}}$	**19.55**
Sex	covariates	/	-0.24	0.05	/	2.68×10^{-7}	4.72
DRB3	HLA heterozygosity	/	0.19	0.06	/	8.33×10^{-4}	1.44
HLA-DQA1*01:03	HLA allele	additive	0.25	0.08	0.06	0.001	1.23
Age	covariates	/	-0.06	0.02	/	0.007	0.61
TAP2.AVARL	TAP2 allotype	additive	0.11	0.04	0.24	0.009	0.53
HLA-DQB1*03:01	HLA allele	additive	0.13	0.05	0.18	0.010	0.47
HLA-A*02:01_ERAP1.TERIGMKDRQ	interaction	/	0.19	0.08	0.27/0.08	0.015	0.31
HLA-DRB4*01:03	HLA allele	additive	0.11	0.05	0.09	0.018	0.24
MCV (67%)							
HLA-DRB1*15:01	**HLA allele**	**dominant**	**-0.45**	**0.07**	**0.14**	$\mathbf{2.49 \times 10^{-11}}$	**8.60**
HLA-DRB1*01:01	**HLA allele**	**additive**	**0.32**	**0.07**	**0.09**	$\mathbf{2.49 \times 10^{-6}}$	**3.81**
HLA-DQB1*03:02	**HLA allele**	**dominant**	**-0.29**	**0.08**	**0.10**	$\mathbf{1.39 \times 10^{-4}}$	**2.17**
PC5	covariates	/	-0.09	0.02	/	3.76×10^{-4}	1.77
HLA-B*07:02	HLA allele	additive	-0.19	0.06	0.15	7.82×10^{-4}	1.48
PC9	covariates	/	-0.07	0.02	/	0.003	0.92

Continued on next page

	Components	Genetic effect	Coeff.	SE	AF	*p-value*	$\log_{10}$ BF
PC10	covariates	/	-0.06	0.03	/	0.015	0.34
PC4	covariates	/	-0.06	0.03	/	0.016	0.30
HLA-B*44:03_ERAP2.TNL	interaction	/	-0.15	0.07	0.06/0.45	0.027	0.11
Age	covariates	/	-0.05	0.02	/	0.035	0.01

Bacteria In this group, we analysed bacterial antigens from three pathogens represented in the UK Biobank serological panel: *Chlamydia trachomatis* (Definition I), *Helicobacter pylori* (Definition I), and *Helicobacter pylori* (Definition II). The Bayesian logistic regression results are illustrated in Figure 4.7, and associations with $\log_{10} \mathrm{BF} > 0$ are summarised in Table 4.7. For *C. Trachomatis*, the strongest associations are observed with demographic covariates. Sex showed a pronounced effect, with higher seropositivity rates among males, while age is negatively associated with serostatus. A modest effect of population structure, reflected by PC6, is also detected. No HLA or antigen-processing gene associations reached statistical significance, suggesting that seropositivity to *C. Chlamydia* rchomatis in this cohort is primarily driven by demographic rather than genetic factors. For *Helicobacter pylori*, two independent definitions are analysed. In both, strong associations are observed with age and PC5, indicating that older age and subtle population stratification contribute to seropositivity. Sex also showed a modest negative effect, with lower seroprevalence among females. Beyond covariates, several genetic components are implicated under Definition II: the HLA-DPB1*04:01 allele displayed a weak recessive protective effect, whereas the B62 HLA supertype appeared to increase seropositivity under a dominant model. These findings point to potential HLA-mediated modulation of the immune response to *Helicobacter pylori* exposure. Overall, while demographic factors (age, sex, and ancestry components) remained the predominant determinants of bacterial seropositivity, the presence of mild HLA associations in *Helicobacter pylori* suggests that host genetics may still play a limited but measurable role in shaping humoral responses to bacterial pathogens.

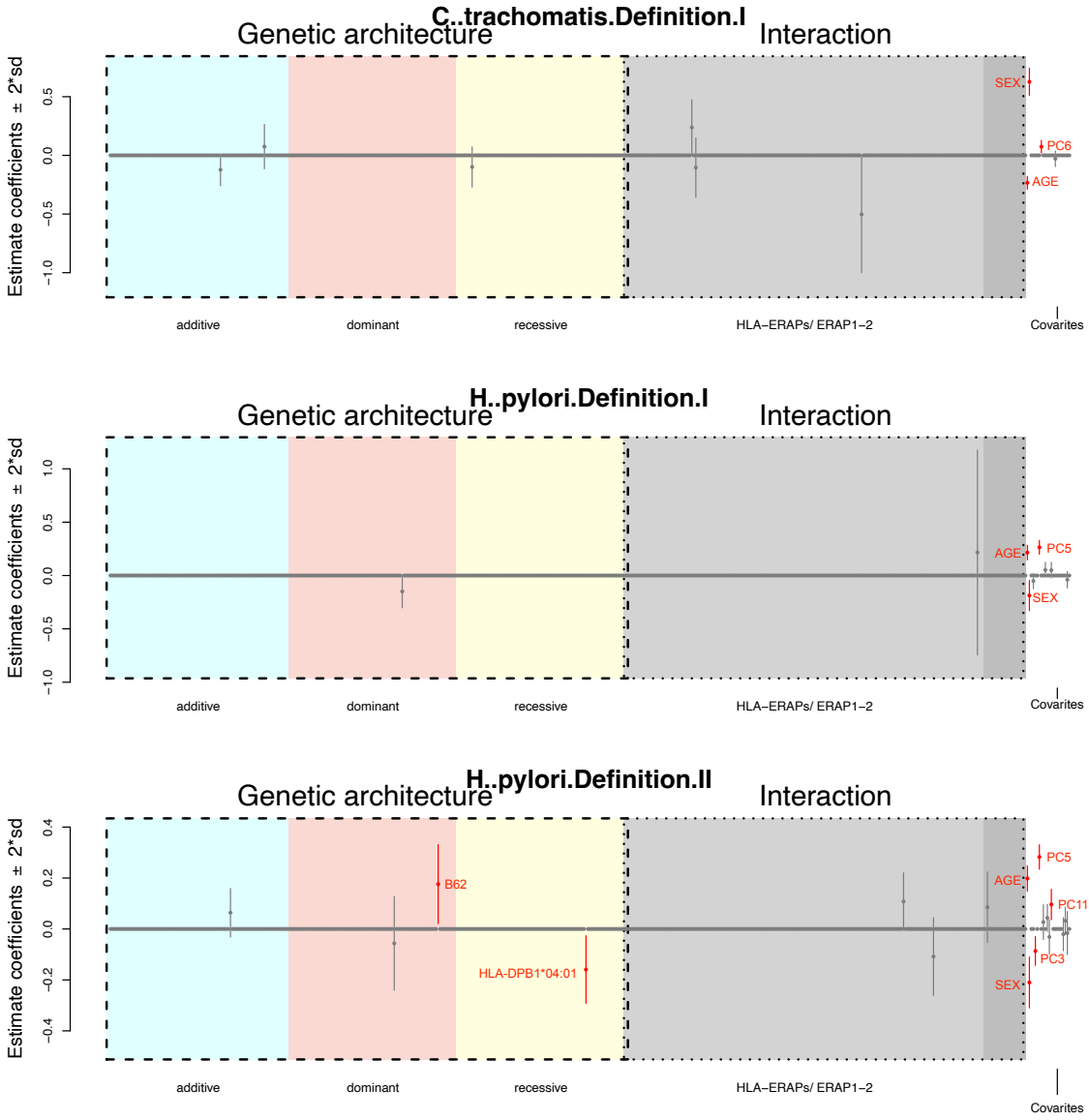


Figure 4.7: **Bayesian joint regression results from the UK Biobank serological panel for the bacteria family**. The analysis includes the following pathogens: C. trachomatous, Definition I Helicobacter pylori Definition I, and Helicobacter pylori Definition II. The figure comprises six blocks. Blue, red, and yellow blocks represent additive, dominant, and recessive genetic effects, respectively. Grey blocks represent interaction terms: light grey for interactions between HLA class I alleles (A, B, C) and ERAP1/2 allotypes, and dark grey for interactions between ERAP1 and ERAP2 allotypes. White areas denote covariates. Error bars represent 95% confidence intervals; those crossing zero are shown in grey, while those not crossing zero are highlighted in red.

Table 4.7: **Bayesian joint regression results from the UK Biobank serological panel for the bacteria family analysis include the following pathogens: C. trachomatis Definition H. pylori Definition, and H. pylori Definition II. The table summarises associations with $\log_{10}\mathbf{BF} > 0$.**

	Components	**Genetic effect**	**Coeff.**	**SE**	**AF**	***p-value***	$\log_{10}\mathbf{BF}$
C. trachomatis Definition I (21%)							
Sex	covariates	/	0.63	0.06	/	$< 1 \times 10^{-16}$	23.09
Age	covariates	/	-0.23	0.03	/	$< 1 \times 10^{-16}$	14.15
PC6	covariates	/	0.07	0.03	/	0.008	0.64
H. pylori Definition I (18%)							
PC5	covariates	/	0.26	0.03	/	2.66×10^{-15}	12.44
Age	covariates	/	0.22	0.03	/	5.92×10^{-10}	7.32
Sex	covariates	/	-0.19	0.07	/	0.009	0.64
H. pylori Definition II (32%)							
PC5	covariates	/	0.28	0.02	/	$< 1 \times 10^{-16}$	27.54
Age	covariates	/	0.20	0.03	/	6.22×10^{-15}	12.10
Sex	covariates	/	-0.21	0.05	/	0.0001	2.71
PC11	covariates	/	0.10	0.03	/	0.002	1.21
PC3	covariates	/	-0.09	0.03	/	0.003	1.00
HLA-DPB1*04:01	HLA allele	recessive	-0.16	0.07	0.44	0.019	0.27
B62	HLA supertype	dominant	0.18	0.08	0.07	0.026	0.14

Parasite For the parasite group, we focused on *Toxoplasma gondii*, a common intracellular protozoan known for its widespread prevalence and chronic infection in humans. The results of the Bayesian logistic regression analysis are presented in Figure 4.8, and associations with $\log_{10}\mathrm{BF} > 0$ are summarised in Table 4.8. Among demographic covariates, both age and sex exhibited measurable effects on seropositivity. Increasing age is associated with a higher likelihood of *T. gondii* seropositivity, reflecting the cumulative nature of exposure over time. A modest negative effect of sex is also detected, with slightly lower seroprevalence observed in females. In terms of genetic components, several interaction effects are observed between HLA class I alleles and ERAP1 allotypes, although these signals did not reach strong statistical significance. Specifically, interactions involving HLA-C*07:02 with ERAP1.TERIGVKNQQ and HLA-C*03:04 with ERAP1.TEPIDMKDRQ showed suggestive negative associations with seropositivity. These results imply that antigen processing and presentation pathways may contribute subtly to host susceptibility to *T. gondii*, but the observed effects are likely modest within this population cohort.

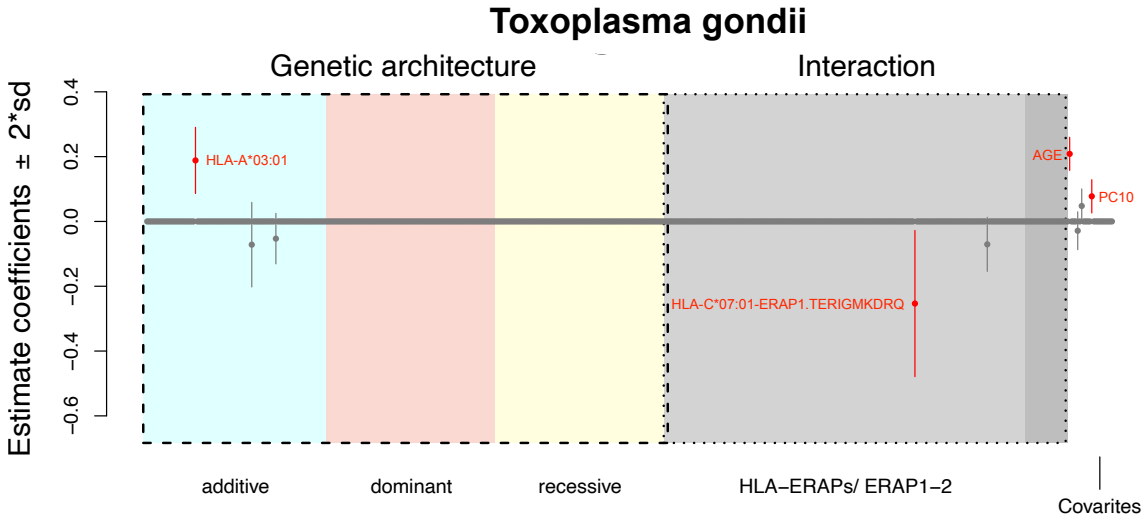


Figure 4.8: **Bayesian joint regression results from the UK Biobank serological panel for the parasite family** analysis include the Toxoplasma gondii. The figure comprises six blocks. Blue, red, and yellow blocks represent additive, dominant, and recessive genetic effects, respectively. Grey represents interaction terms: light grey interactions between HLA class I alleles (A, B, C) and ERAP1/2 allotypes, and dark grey indicates the interactions between ERAP1 and ERAP2 allotypes. White areas denote covariates. Error bars represent 95% confidence intervals; those crossing zero are shown in grey, while those not crossing zero are highlighted in red.

Table 4.8: **Bayesian joint regression results from the UK Biobank serological panel for the parasite family.** This group includes the Toxoplasma gondii. The table summarizes associations with $\log_{10}\mathrm{BF} > 0$.

	Components	**Genetic effect**	**Coeff.**	**SE**	**AF**	***p-value***	$\log_{10}$ **BF**
***Toxoplasma gondii* (28%)**							
Age	covariates	/	-0.27	0.06	/	2.13×10^{-6}	3.97
Sex	covariates	/	-0.25	0.12	/	0.045	0.22
HLA-C*07:02_ERAP1.TERIGVKNQQ	interaction	/	-0.25	0.13	0.16/0.22	0.052	0.17
HLA-C*03:04_ERAP1.TEPIDMKDRQ	interaction	/	-0.54	0.28	0.08/0.07	0.060	0.12

MFI To more sensitively capture how host genetics influence quantitative antibody responses, we also analyse the associations between genetic components within the antigen presentation pathway and antibody MFI levels. Results are grouped by antigen according to pathogen family: **Herpesviridae** (HSV-1, HSV-2, EBV, HCMV, HHV-6, HHV-7, VZV); **Polyomaviridae** (JCV, BKV, MCV); **Bacteria** (*C. trachomatis*, *H. pylori*); and **Parasite** (*T. gondii*). We employ Bayesian linear regression with a regularised horseshoe prior using the `mapHS` function with `CD_HS_purm_logistic`, and summarise results for each pathogen family as follows.

Herpesviridae The results for the infectious agent group are shown in Figure 4.9. To explore potential associations correlated with seropositive MFI levels, we summarised in Table 4.9 those results with $\log_{10} \mathrm{BF} > 0$ and 95% credible intervals that do not include zero. Associations meeting the significance threshold of $p < 5 \times 10^{-4}$ are highlighted in bold. We next describe the key findings for each pathogen.

For HSV-1 (gG-1), the allele HLA-DRB1*03:01 is identified as a protective factor, while sex and age also showed modest effects. In contrast, for HSV-2 (gG-2), no significant associations are detected. For EBV, multiple HLA class II alleles exhibited significant effects across different antigens, suggesting complex immune genetic determinants of the humoral response. Specifically, HLA-DQA1*03:01, HLA-DQA1*02:01, HLA-DPB1*03:01, and HLA-DRB1*01:01 are associated with the EA-D antigen, indicating both dominant and recessive effects. For the EBNA-1 antigen, strong associations are observed with HLA-DQB1*02:01, HLA-DQA1*05:01, HLA-DRB1*07:01, and HLA-DRB1*03:01, while additional associations are noted for HLA-DPB1*03:01, HLA-DQA1*03:01, and HLA-DRB1*13:01. HLA-DPB1*03:01 appeared consistently across multiple EBV antigens (EA-D, EBNA-1, and VCA p18), suggesting a broad modulatory role in EBV-specific antibody responses. Conversely, certain alleles such as HLA-C*04:01 and HLA-DRB1*04:01 are observed only for specific antigens (VCA p18), implying antigen-restricted genetic effects. For ZEBRA, several HLA-DQA1 alleles (HLA-DQA1*01:02, HLA-DQA1*01:01, and HLA-DQA1*01:03) and HLA-DPA1*02:01 are associated with reduced antibody responses, consistent with allele-specific regulatory effects in antigen presentation. For CMV, sex is a significant predictor across all three tested antigens (pp28, pp52, and pp150), indicating sex-related differences in CMV-specific antibody magnitude, while other genetic effects are limited. For HHV-7 (U14), HLA-DRB1*04:01 emerged as a protective factor, whereas for VZV (gE and gI), both HLA-DRB1*03:01 and HLA-DRB1*15:01 showed

significant associations, with the former acting as a protective allele and the latter as a risk allele. Together, these results highlight the strong and antigen-specific contributions of HLA class II variants in shaping antibody responses to members of the Herpesviridae family.

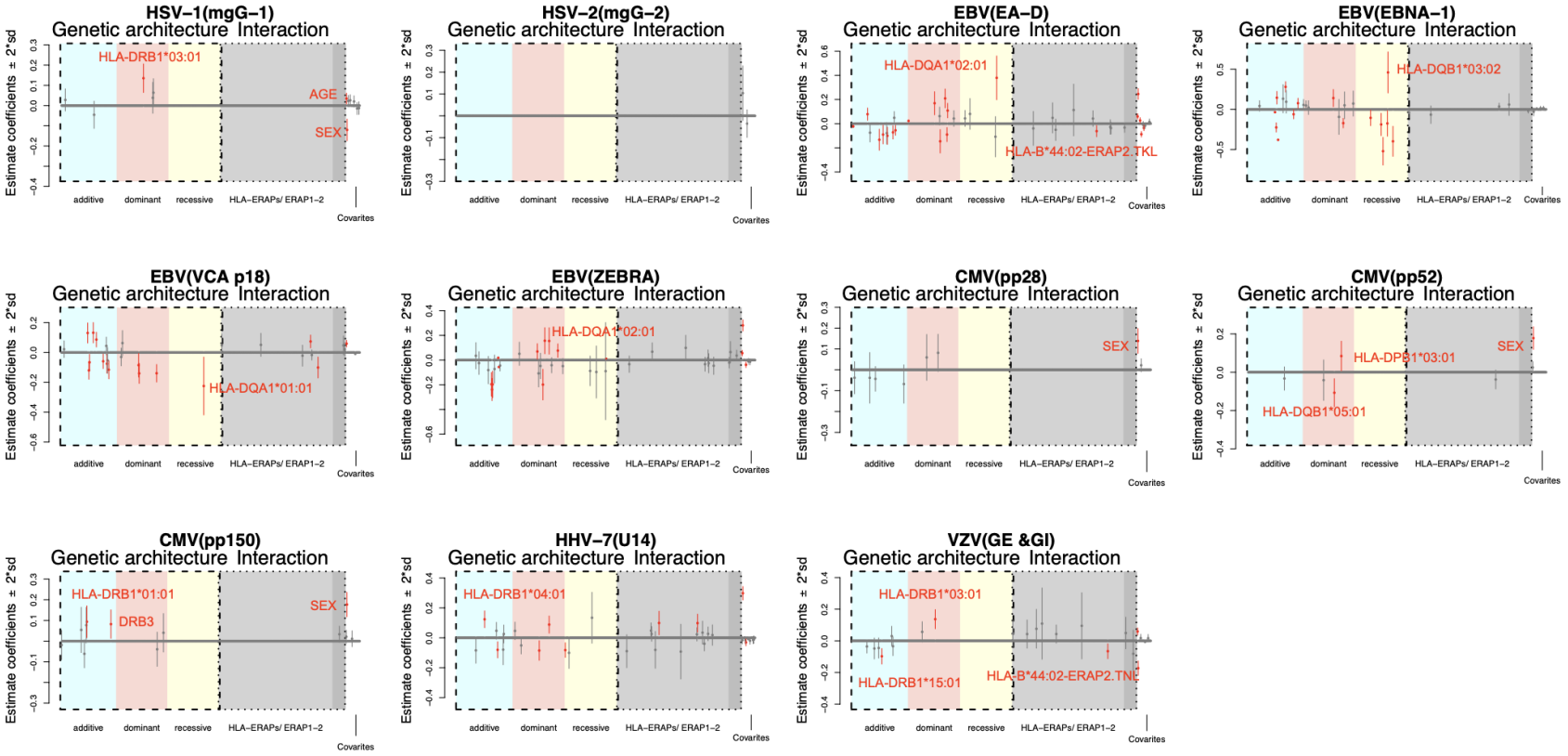


Figure 4.9: **Bayesian joint regression results from the UK Biobank serological panel for the *Herpesviridae* family among seropositive individuals.** The analysis includes the following pathogens and corresponding antigens: HSV-1 (mgG-1), HSV-2 (mgG-2), EBV (EA-D, EBNA-1, VCA p18, ZEBRA), CMV (pp28, pp52, pp150), HHV-7 (U14), VZV (GE & GI). The figure comprises six blocks. Blue, red, and yellow blocks represent additive, dominant, and recessive genetic effects, respectively. Grey blocks represent interaction terms: light grey for interactions between HLA class I alleles (A, B, C) and ERAP1/2 allotypes, and dark grey for interactions between ERAP1 and ERAP2 allotypes. White areas denote covariates. Error bars represent 95% confidence intervals; those crossing zero are shown in grey, while those not crossing zero are highlighted in red.

Table 4.9: **Bayesian joint regression results from the UK Biobank serological panel for the *Herpesviridae* family among seropositive individuals.** The analysis includes the following pathogens and corresponding antigens: HSV-1 (mgG-1), HSV-2 (mgG-2), EBV (EA-D, EBNA-1, VCA p18, ZEBRA), CMV (pp28, pp52, pp150), HHV-7 (U14), VZV (GE & GI). The table summarises associations where $\log_{10}\mathrm{BF} > 0$, and the 95% confidence interval does not include zero.

	Components	**Genetic effect**	**Coeff.**	**SE**	**AF**	***p-value***	$\log_{10}$ **BF**
HSV-1(mgG-1)							
Sex	covariate	/	-0.12	0.03	/	1.06×10^{-5}	3.02
HLA-DRB1*03:01	**HLA allele**	**dominant**	**0.14**	**0.04**	**0.16**	$\mathbf{1.73 \times 10^{-4}}$	**1.87**
Age	covariate	/	0.03	0.01	/	1.16×10^{-2}	0.20
EBV(EA-D)							
Sex	covariate	/	0.25	0.02	/	$< 1 \times 10^{-16}$	23.35
PC5	covariate	/	-0.09	0.01	/	6.37×10^{-14}	10.93
HLA-DQA1*03:01	**HLA allele**	**dominant**	**0.21**	**0.04**	**0.20**	$\mathbf{1.41 \times 10^{-7}}$	**4.75**
Age	covariate	/	0.06	0.01	/	1.88×10^{-7}	4.63
HLA-DQA1*02:01	**HLA allele**	**recessive**	**0.38**	**0.09**	**0.16**	$\mathbf{4.07 \times 10^{-5}}$	**2.40**
HLA-DPB1*03:01	**HLA allele**	**dominant**	**0.11**	**0.03**	**0.10**	$\mathbf{4.04 \times 10^{-4}}$	**1.46**
HLA-DRB1*01:01	HLA allele	dominant	0.17	0.05	0.11	5.42×10^{-4}	1.34
PC9	covariate	/	-0.04	0.01	/	2.25×10^{-3}	0.77
HLA-DPB1*02:01	HLA allele	dominant	-0.09	0.03	0.10	2.37×10^{-3}	0.75
HLA-A*03:01	HLA allele	additive	0.08	0.03	0.14	2.55×10^{-3}	0.72
HLA-DRB1*03:01	HLA allele	additive	-0.13	0.04	0.16	2.86×10^{-3}	0.68
HLA-DQB1*03:02	HLA allele	dominant	-0.15	0.05	0.10	3.19×10^{-3}	0.63
HLA-DQA1*01:02	HLA allele	additive	-0.10	0.04	0.17	5.91×10^{-3}	0.39
A02	HLA supertype	additive	-0.05	0.02	0.29	9.25×10^{-3}	0.22
HLA-B*44:02_ERAP2.TKL	interaction	/	-0.06	0.02	0.11 / 0.44	9.84×10^{-3}	0.19

Continued on next page

	Components	Genetic effect	Coeff.	SE	AF	*p-value*	$\log_{10}$ BF
HLA-DPB1*04:02	HLA allele	additive	-0.07	0.03	0.11	1.16×10^{-2}	0.13
EBV(EBNA-1)							
HLA-DQB1*02:01	**HLA allele**	**additive**	**-0.38**	**0.01**	**0.16**	$< 1 \times 10^{-16}$	**29.96**
HLA-DQA1*05:01	**HLA allele**	**additive**	**0.28**	**0.03**	**0.25**	$< 1 \times 10^{-16}$	**14.14**
HLA-DRB1*07:01	**HLA allele**	**additive**	**-0.23**	**0.03**	**0.14**	$\mathbf{2.73 \times 10^{-14}}$	**11.29**
HLA-DRB1*03:01	**HLA allele**	**recessive**	**-0.52**	**0.09**	**0.16**	$\mathbf{3.38 \times 10^{-9}}$	**6.32**
HLA-DPB1*03:01	**HLA allele**	**dominant**	**-0.17**	**0.03**	**0.10**	$\mathbf{1.33 \times 10^{-8}}$	**5.74**
HLA-DQA1*03:01	**HLA allele**	**recessive**	**-0.40**	**0.10**	**0.20**	$\mathbf{3.14 \times 10^{-5}}$	**2.50**
HLA-DRB1*13:01	**HLA allele**	**additive**	**0.14**	**0.04**	**0.05**	$\mathbf{3.42 \times 10^{-4}}$	**1.53**
HLA-DQB1*03:02	**HLA allele**	**recessive**	**0.46**	**0.13**	**0.10**	$\mathbf{4.16 \times 10^{-4}}$	**1.45**
HLA-DRB1*03:01	**HLA allele**	**additive**	**-0.04**	**0.01**	**0.16**	$\mathbf{4.53 \times 10^{-4}}$	**1.41**
HLA-DRB4*01:03	HLA allele	additive	-0.19	0.07	0.09	8.19×10^{-3}	0.27
HLA-DRB1*15:01	HLA allele	dominant	0.14	0.05	0.12	1.00×10^{-2}	0.19
MAIN_DR	HLA supertype	additive	0.08	0.03	0.32	1.56×10^{-2}	0.02
EBV(VCA p18)							
Age	covariate	/	0.06	0.01	/	2.04×10^{-7}	4.60
HLA-DPB1*03:01	**HLA allele**	**dominant**	**-0.14**	**0.03**	**0.10**	$\mathbf{3.65 \times 10^{-6}}$	**3.39**
HLA-C*04:01	**HLA allele**	**dominant**	**-0.14**	**0.03**	**0.085**	$\mathbf{4.49 \times 10^{-5}}$	**2.36**
HLA-DRB1.04.01	**HLA allele**	**additive**	**-0.12**	**0.03**	**0.12**	$\mathbf{7.76 \times 10^{-5}}$	**2.13**
HLA-DRB1*01:01	**HLA allele**	**additive**	**0.13**	**0.03**	**0.11**	$\mathbf{1.39 \times 10^{-4}}$	**1.89**
HLA-DQB1*03:02	**HLA allele**	**dominant**	**0.13**	**0.03**	**0.10**	$\mathbf{1.41 \times 10^{-4}}$	**1.89**
B08	**HLA supertype**	**additive**	**-0.12**	**0.03**	**0.15**	$\mathbf{2.15 \times 10^{-4}}$	**1.72**
HLA-DQA1*01:02	HLA allele	additive	0.09	0.03	0.17	6.20×10^{-4}	1.29
HLA-C*07:02_ERAP2.TKL	interaction	/	0.07	0.02	0.14 / 0.46	7.13×10^{-4}	1.23
HLA-C*03:03_ERAP2.TNL	interaction	/	-0.10	0.04	0.06 / 0.44	4.93×10^{-3}	0.46

Continued on next page

	Components	Genetic effect	Coeff.	SE	AF	*p-value*	$\log_{10}$ BF
EBV(ZEBRA)							
Sex	covariate	/	0.28	0.02	/	$< 1 \times 10^{-16}$	31.02
HLA-DPA1*02:01	**HLA allele**	**additive**	**-0.06**	**0.01**	**0.15**	$\mathbf{1.73 \times 10^{-8}}$	**5.63**
HLA-DQA1*01:02	**HLA allele**	**additive**	**-0.24**	**0.05**	**0.17**	$\mathbf{1.81 \times 10^{-7}}$	**4.64**
Age	covariate	/	0.06	0.01	/	9.74×10^{-7}	3.94
HLA-DQA1*01:01	**HLA allele**	**additive**	**-0.20**	**0.05**	**0.15**	$\mathbf{6.03 \times 10^{-5}}$	**2.24**
HLA-DQA1*01:03	**HLA allele**	**additive**	**-0.20**	**0.05**	**0.06**	$\mathbf{3.16 \times 10^{-4}}$	**1.56**
PC5	covariate	/	-0.04	0.01	/	1.08×10^{-3}	1.06
HLA-DQB1*02:01	HLA allele	dominant	-0.20	0.06	0.16	1.77×10^{-3}	0.87
HLA-DQB1*03:02	HLA allele	dominant	0.16	0.05	0.10	3.61×10^{-3}	0.59
HLA-DQA1*02:01	HLA allele	dominant	0.16	0.05	0.14	4.53×10^{-3}	0.50
B44	HLA supertypes	dominant	0.08	0.03	0.27	1.56×10^{-2}	0.36
CMV(pp28)							
Sex	**covariate**	**/**	**0.14**	**0.03**	**/**	$\mathbf{4.998 \times 10^{-6}}$	**3.37**
CMV(pp52)							
Sex	**covariate**	**/**	**0.18**	**0.03**	**/**	$\mathbf{3.44 \times 10^{-9}}$	**6.41**
HLA-DQB1*05:01	HLA allele	dominant	-0.11	0.04	0.13	4.22×10^{-3}	0.63
CMV(pp150)							
Sex	covariate	/	0.18	0.03	/	4.58×10^{-9}	6.29
HLA-DRB1*01:01	HLA allele	additive	0.09	0.04	0.11	1.70×10^{-2}	0.09
DRB3	heterozygosity	/	0.08	0.04	/	1.96×10^{-2}	0.04
HHV-7(U14)							
Sex	covariate	/	0.30	0.02	/	$< 1 \times 10^{-16}$	35.10

Continued on next page

	Components	Genetic effect	Coeff.	SE	AF	*p-value*	$\log_{10}$ BF
HLA-DRB1*04:01	**HLA allele**	**additive**	**0.12**	**0.03**	**0.14**	$\mathbf{1.61 \times 10^{-5}}$	**2.78**
PSMB9.RV	PSMB9 allotype	recessive	-0.08	0.03	0.68	1.05×10^{-3}	1.07
HLA-A*11:01_ERAP2.TKL	interaction	/	0.10	0.03	0.06 / 0.44	1.37×10^{-3}	0.97
HLA-DQA1*02:01	HLA allele	dominant	0.09	0.03	0.14	2.67×10^{-3}	0.70
HLA-DPB1*04:02	HLA allele	additive	-0.08	0.03	0.11	3.54×10^{-3}	0.59
PC5	covariate	/	-0.03	0.01	/	1.03×10^{-2}	0.18
HLA-DRB1*01:01	HLA allele	dominant	-0.09	0.03	0.11	1.11×10^{-2}	0.15
HLA-C*04:01_ERAP1.TERIGVKNQQ	interaction	/	0.10	0.04	0.08/0.22	1.40×10^{-2}	0.06
VZV(GE & GI)							
Sex	covariate	/	-0.17	0.02	/	1.21×10^{-13}	10.66
Age	covariate	/	0.06	0.01	/	3.13×10^{-7}	4.42
HLA-DRB1*03:01	**HLA allele**	**dominant**	**0.14**	**0.03**	**0.16**	$\mathbf{8.21 \times 10^{-6}}$	**3.06**
HLA-DRB1*15:01	**HLA allele**	**additive**	**-0.10**	**0.03**	**0.12**	$\mathbf{1.04 \times 10^{-4}}$	**2.02**

Bacteria The results for the bacterial antigen group are shown in Figure 4.10, and associations with $\log_{10} \mathrm{BF} > 0$ and 95% confidence intervals that do not include zero are reported in Table 4.10. Overall, demographic and ancestry covariates (age, sex and principal components) produced the most consistent and strongest associations, whereas HLA and interaction effects are generally modest and often antigen-specific.

For *C. trachomati*, we observed negative associations of age with responses to MOMP-A and MOMP-D, and positive associations of sex (higher MFI in females) with pGP3 and both TARP-D fragments; the effect for TARP-D F2 is particularly strong. A small number of genetic interaction terms appeared for Chlamydia antigens, including an HLA-A*24:02–ERAP1 interaction for PorB and an HLA-A*11:01–ERAP2 interaction for TARP-D F2, but these genetic signals are modest in magnitude and did not meet the significance threshold.

For *H. Pylori*, the PC5 and sex are the dominant predictors across multiple antigens (Catalase, GroEL, OMP, UreA, and VacA). A few HLA signals are detected for specific antigens: an HLA-C*04:01–ERAP2 interaction with CagA, a recessive effect of HLA-DPB1*04:01 on GroEL, HLA-B*44:03 and HLA-DQB1*03:01 associations with OMP, a dominant effect of HLA-DRB1*04:01 on UreA, and a dominant effect of HLA-DQA1*03:01 on VacA. Most of these HLA or interaction effects have modest Bayes factors ($\log_{10}$BF near zero) and *p-value* above the stringent threshold ($p < 5 \times 10^{-4}$).

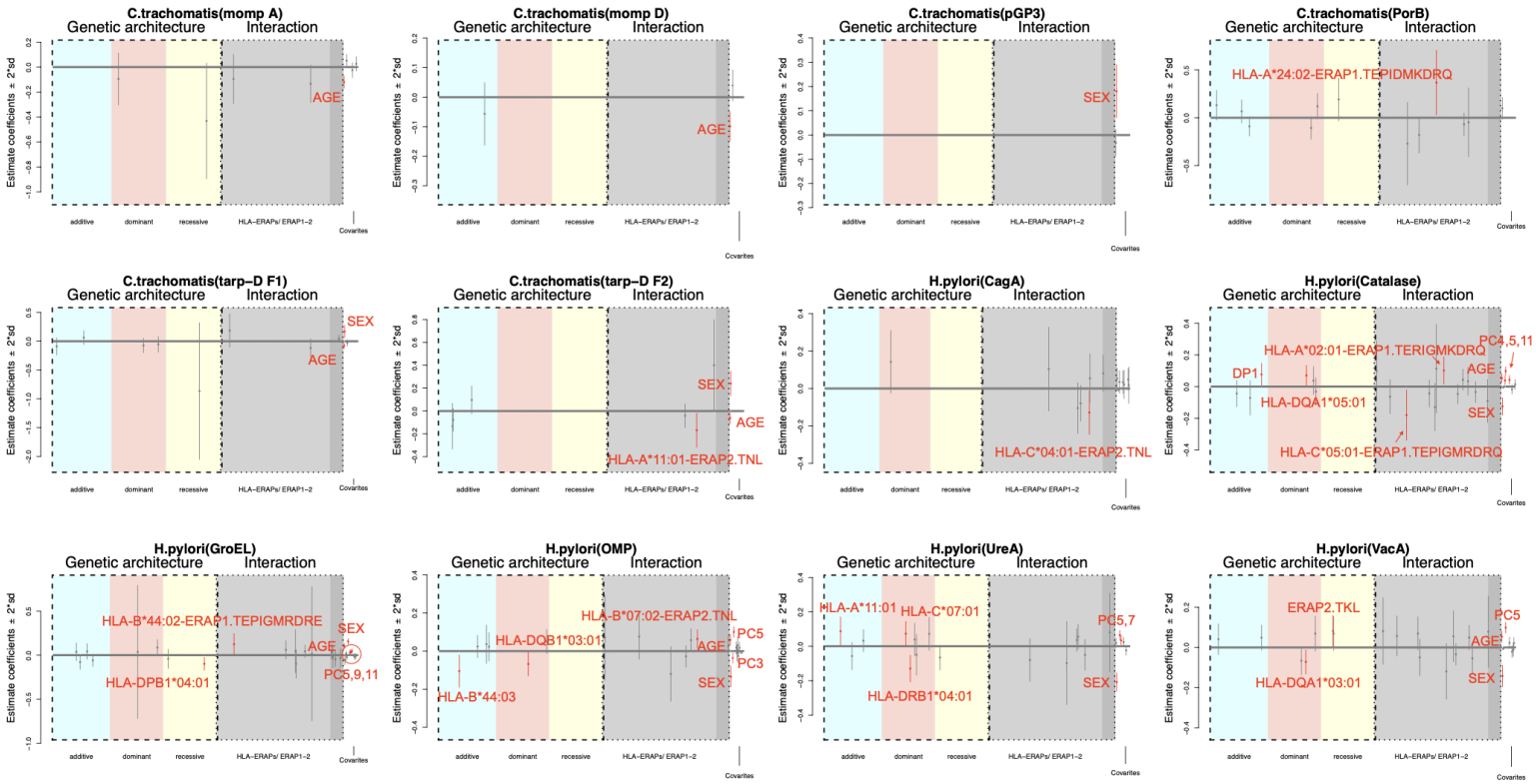


Figure 4.10: **Bayesian joint regression results from the UK Biobank serological panel for the *Bacteria* family among seropositive individuals.** The analysis includes the following pathogens and corresponding antigens: C.trachomatis (mompA, mompD, pGP3, PorB, tarp-D F1, tarp-D F2) and H. pylori (CagA, Catalase, GroEL, OMP). The figure comprises six blocks. Blue, red, and yellow blocks represent additive, dominant, and recessive genetic effects, respectively. Grey blocks represent interaction terms: light grey for interactions between HLA class I alleles (A, B, C) and ERAP1/2 allotypes, and dark grey for interactions between ERAP1 and ERAP2 allotypes. White areas denote covariates. Error bars represent 95% confidence intervals; those crossing zero are shown in grey, while those not crossing zero are highlighted in red.

Table 4.10: **Bayesian joint regression results from the UK Biobank serological panel for the *Bacteria* family among seropositive individuals.** The analysis includes the following pathogens and corresponding antigens: C.trachomatis(mompA, mompD, pGP3, PorB, tarp-D F1, tarp-D F2) and H. pylori (CagA, Catalase, GroEL, OMP). The table summarises associations where $\log_{10}\mathrm{BF} > 0$, and the 95% confidence interval does not include zero.

	Components	**Genetic effect**	**Coeff.**	**SE**	**AF**	***p-value***	$\log_{10}$ **BF**
C.trachomatis(momp A)							
Age	covariate	/	-0.120	0.025	/	1.94×10^{-6}	3.933
C.trachomatis(momp D)							
Age	covariate	/	-0.099	0.025	/	8.97×10^{-5}	2.370
C.trachomatis(pGP3)							
Sex	covariate	/	0.180	0.056	/	1.22×10^{-3}	1.326
C.trachomatis(PorB)							
HLA-A*24:02_ERAP1.TEPIDMKDRQ	interaction	/	0.368	0.173	0.073 / 0.068	3.31×10^{-2}	0.059
C.trachomatis(tarp-D F1)							
Age	covariate	/	-0.077	0.026	/	2.48×10^{-3}	1.047
Sex	covariate	/	0.167	0.056	/	2.83×10^{-3}	0.995
C.trachomatis(tarp-D F2)							
Sex	covariate	/	0.241	0.055	/	1.18×10^{-5}	3.193
Age	covariate	/	-0.060	0.026	/	2.02×10^{-2}	0.242
HLA-A*11:01_ERAP2.TNL	interaction	/	-0.168	0.076	0.064 / 0.441	2.74×10^{-2}	0.129
H.pylori(CagA)							

Continued on next page

	Components	Genetic effect	Coeff.	SE	AF	*p-value*	$\log_{10}$ BF
HLA-C*04:01_ERAP2.TNL	interaction	/	-0.128	0.060	0.088 / 0.441	3.23×10^{-2}	0.101
H.pylori(Catalase)							
PC5	covariate	/	0.098	0.014	/	1.38×10^{-12}	9.681
Sex	covariate	/	-0.125	0.028	/	6.72×10^{-6}	3.211
Age	covariate	/	0.053	0.014	/	1.04×10^{-4}	2.083
PC11	covariate	/	0.043	0.014	/	2.16×10^{-3}	0.863
PC4	covariate	/	0.039	0.014	/	5.74×10^{-3}	0.478
HLA-A*02:01_ERAP1.TERIGMKDRQ	interaction	/	0.102	0.044	0.277 / 0.087	2.17×10^{-2}	-0.032
HLA-C*05:01_ERAP1.TEPIGMRDRQ	interaction	/	-0.180	0.081	0.115 / 0.071	2.66×10^{-2}	-0.108
HLA-DQA1*05:01	HLA allele	dominant	0.070	0.033	0.247	3.14×10^{-2}	-0.170
DP1	marker	/	0.077	0.036	0.162	3.31×10^{-2}	-0.190
H.pylori(GroEL)							
PC5	covariate	/	0.156	0.016	/	0	18.979
Age	covariate	/	0.103	0.014	/	2.06×10^{-13}	10.490
PC11	covariate	/	0.046	0.014	/	1.18×10^{-3}	1.102
HLA-DPB1*04:01	HLA allele	recessive	-0.097	0.038	0.437	9.84×10^{-3}	0.269
PC9	covariate	/	0.034	0.014	/	1.61×10^{-2}	0.081
HLA-B*44:02_ERAP1.TEPIGMRDRE	interaction	/	0.125	0.062	0.113 / 0.136	4.43×10^{-2}	-0.297
H.pylori(OMP)							
PC5	covariate	/	0.100	0.014	/	6.06×10^{-13}	10.032
Sex	covariate	/	-0.134	0.028	/	1.17×10^{-6}	3.936
Age	covariate	/	0.058	0.014	/	2.19×10^{-5}	2.723
PC3	covariate	/	-0.036	0.014	/	9.59×10^{-3}	0.279
HLA-B*07:02_ERAP2.TNL	interaction	/	0.063	0.025	0.127 / 0.441	1.07×10^{-2}	0.235

Continued on next page

	Components	Genetic effect	Coeff.	SE	AF	*p-value*	$\log_{10}$ BF
HLA-B*44:03	HLA allele	additive	-0.105	0.043	0.056	1.53×10^{-2}	0.099
HLA-DQB1*03:01	HLA allele	dominant	-0.068	0.032	0.183	3.35×10^{-2}	-0.194
H.pylori(UreA)							
Sex	covariate	/	-0.208	0.027	/	3.57×10^{-14}	11.233
PC5	covariate	/	0.065	0.014	/	3.37×10^{-6}	3.496
HLA-DRB1*04:01	HLA allele	dominant	-0.129	0.040	0.117	1.11×10^{-3}	1.129
PC7	covariate	/	0.034	0.014	/	1.59×10^{-2}	0.085
HLA-A*11:01	HLA allele	additive	0.088	0.042	0.064	3.88×10^{-2}	-0.248
HLA-C*07:01	HLA allele	dominant	0.074	0.037	0.186	4.34×10^{-2}	-0.289
H.pylori(VacA)							
PC5	covariate	/	0.098	0.014	/	7.28×10^{-13}	9.954
Sex	covariate	/	-0.142	0.028	/	2.77×10^{-7}	4.536
Age	covariate	/	0.054	0.014	/	9.72×10^{-5}	2.112
HLA-DQA1*03:01	HLA allele	dominant	-0.072	0.031	0.197	1.82×10^{-2}	0.035
ERAP2.TKL	ERAP2 allotype	recessive	0.081	0.039	0.462	3.78×10^{-2}	-0.239

Polyomaviridae The Polyomaviridae group included BK virus (BKV), JC virus (JCV), and Merkel cell polyomavirus (MCV), all tested using the VP1 antigen. The results of Bayesian logistic regression are presented in Figure 4.11 and summarised in Table 4.11. Overall, genetic effects within the antigen presentation pathway, particularly involving HLA class II alleles, are prominent in shaping antibody responses to polyomaviruses. For BKV, age showed a strong negative association with antibody levels, indicating decreasing MFI with increasing age. Modest associations are also observed for HLA-DPB1*04:01 and HLA-DQB1*05:01, both acting under dominant or additive models, respectively, suggesting potential but weak contributions of these alleles to BKV antibody variation. For JCV, a strong and significant effect is identified for HLA-DRB1*15:01 under a dominant model ($p = 8.5 \times 10^{-7}$, $\log_{10} \mathrm{BF} = 4.10$), where the allele is associated with reduced antibody responses, indicating a possible risk or susceptibility role. For MCV, both HLA-DQB1*05:01 (dominant) and HLA-DRB1*15:01 (additive) are significantly associated with antibody responses, showing consistent effects with those observed for JCV. Additionally, a dominant signal for the DR4 marker is detected, suggesting further HLA-related modulation. Notably, HLA-DRB1*15:01 exhibited consistent associations across both JCV and MCV, highlighting its shared contribution to immune variation within the Polyomaviridae family, while HLA-DQB1*05:01 appeared to exert a protective effect specific to MCV.

Together, these findings underscore that alleles within the HLA class II region, particularly HLA-DRB1*15:01 and HLA-DQB1*05:01, play key roles in determining host antibody responses to polyomavirus infections, with both shared and virus-specific effects observed across the family.

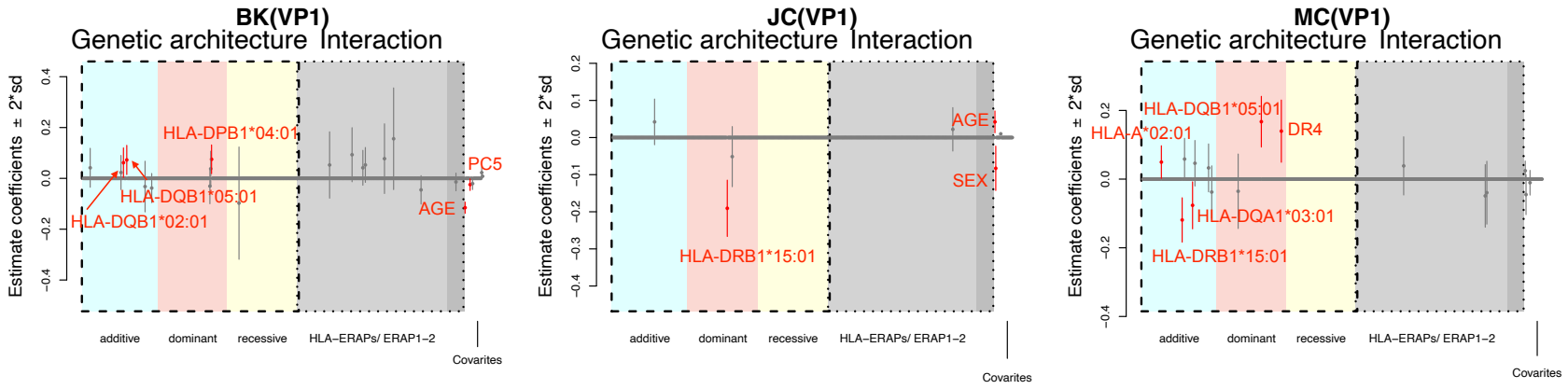


Figure 4.11: **Bayesian joint regression results from the UK Biobank serological panel for the *Polyomaviridae* family among seropositive individuals.** The analysis includes the following pathogens and corresponding antigens: BK(VP1), JC(VP1), and MC(VP1). The figure comprises six blocks. Blue, red, and yellow blocks represent additive, dominant, and recessive genetic effects, respectively. Grey blocks represent interaction terms: light grey for interactions between HLA class I alleles (A, B, C) and ERAP1/2 allotypes, and dark grey for interactions between ERAP1 and ERAP2 allotypes. White areas denote covariates. Error bars represent 95% confidence intervals; those crossing zero are shown in grey, while those not crossing zero are highlighted in red.

Table 4.11: **Bayesian joint regression results from the UK Biobank serological panel for the *Polyomaviridae* family among seropositive individuals.** The analysis includes the following pathogens and corresponding antigens: BK(VP1), JC(VP1), and MC(VP1). The table summarises associations where $\log_{10}\mathrm{BF} > 0$, and the 95% confidence interval does not include zero.

	Components	**Genetic effect**	**Coeff.**	**SE**	**AF**	***p-value***	$\log_{10}$ **BF**
BK(VP1) (95%)							
Age	covariates	/	-0.1155	0.0113	/	$< 1 \times 10^{-16}$	21.169
HLA-DPB1*04:01	HLA allele	dominant	0.0750	0.0289	0.437	9.35×10^{-3}	0.212
HLA-DQB1*05:01	HLA allele	additive	0.0725	0.0293	0.131	1.34×10^{-2}	0.073
JC(VP1)							
HLA-DRB1*15:01	**HLA allele**	**dominant**	**-0.1907**	**0.0387**	**0.120**	$\mathbf{8.50 \times 10^{-7}}$	**4.099**
Age	covariates	/	0.0422	0.0150	/	4.95×10^{-3}	0.568
Sex	covariates	/	-0.0829	0.0305	/	6.51×10^{-3}	0.462
MC(VP1) (67%)							
HLA-DQB1*05:01	**HLA allele**	**dominant**	**0.1673**	**0.0377**	**0.131**	$\mathbf{9.01 \times 10^{-6}}$	**3.092**
HLA-DRB1*15:01	**HLA allele**	**additive**	**-0.1190**	**0.0330**	**0.120**	$\mathbf{3.07 \times 10^{-4}}$	**1.647**
DR4	HLA allele	dominant	0.1397	0.0462	0.106	2.52×10^{-3}	0.804

Parasite The parasite group included *T. gondii*, with two antigens tested: p22 and SAG1. The results of Bayesian logistic regression are presented in Figure 4.12 and summarised in Table 4.12. Overall, no compelling associations are identified between genetic components in the antigen presentation pathway and antibody responses to *T. gondii*. For the p22 antigen, sex showed modest effects, while for SAG1, age exhibited a weak but significant positive association with antibody MFI levels. However, none of these associations involved HLA or other immune pathway components, suggesting that host genetic variation in classical antigen presentation loci may play a limited role in shaping humoral responses to *T. gondii* in this cohort.

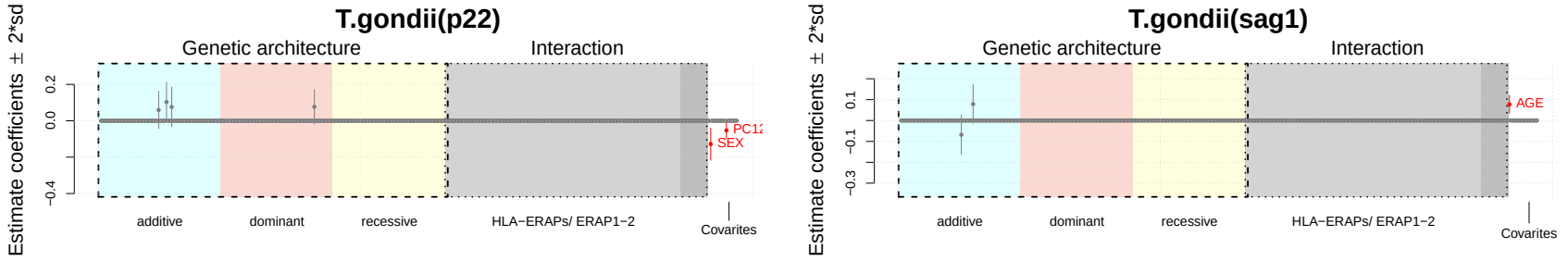


Figure 4.12: **Bayesian joint regression results from the UK Biobank serological panel for the *Herpesviridae* family among seropositive individuals.** The analysis includes T.gondii (p22 and sag1). The figure comprises six blocks. Blue, red, and yellow blocks represent additive, dominant, and recessive genetic effects, respectively. Grey blocks represent interaction terms: light grey for interactions between HLA class I alleles (A, B, C) and ERAP1/2 allotypes, and dark grey for interactions between ERAP1 and ERAP2 allotypes. White areas denote covariates. Error bars represent 95% confidence intervals; those crossing zero are shown in grey, while those not crossing zero are highlighted in red.

Table 4.12: **Bayesian joint regression results from the UK Biobank serological panel for the *Herpesviridae* family among seropositive individuals.** The analysis includes the T.gondii(p22 and sag1). The table summarises associations where $\log_{10}\mathrm{BF} > 0$, and the 95% confidence interval does not include zero.

	Components	**Genetic effect**	**Coeff.**	**SE**	**AF**	***p-value***	$\log_{10}$ **BF**
T. gondii (p22)							
Sex	covariate	/	-0.128	0.0452	/	4.6×10^{-3}	0.752
PC12	covariate	/	-0.0529	0.0223	/	1.8×10^{-2}	0.236
T. gondii (sag1)							
Age	covariate	/	0.0772	0.0219	/	4.3×10^{-4}	1.689

4.3.2 Comparison with conditional analysis and previous results

In this section, we performed a traditional conditional analysis and compared the findings with those from the Bayesian joint regression and previously reported associations in the literature.

In the marginal analysis, each genetic or interaction variable is tested independently for association with the phenotype. We used the `lm()` function for continuous traits and `glm()` for binary traits in `R`, applying the same significance threshold as in the joint analysis ($p < 5 \times 10^{-4}$). Given the strong linkage disequilibrium (LD) within the HLA region and ERAP loci, multiple significant signals from the same locus may reflect correlation rather than independent effects. To identify potential independent causal variants, we applied a stepwise conditional approach. Starting with the most significant variable from the marginal analysis, we included it as a covariate and re-ran the regression for all remaining variables. This process is iterated until no additional variable meets the significance threshold. Since dominance and recessive genetic models are more difficult to interpret in conditional analysis, we restricted these regressions to additive effects (encoded as 0, 1, 2). For interaction terms, the model is specified as:

$$\boldsymbol{Y} \sim x_1 + x_2 + x_1 \times x_2 + \text{covariates},$$

from which we extracted the coefficient and standard error of the interaction term ($x_1 \times x_2$).

We visualised the stepwise conditional analysis for a representative dataset, HCV spontaneous clearance versus chronic infection, to illustrate the sequential conditioning process and the evolution of signals across steps. For other datasets, we summarised only the final conditional results, focusing on independent genetic components and covariates that remained significant after conditioning.

Finally, we compared the results of the conditional analysis with those from the Bayesian joint model, which employed a regularised horseshoe prior and previously reported genetic associations.

4.3.2.1 Hepatitis C

Spontaneous clearance vs. chronic infection For the HCV spontaneous clearance versus chronic infection cohort, covariates included sex and the 5 PCs. We used logistic regression implemented via the `glm()` function in `R`, analysing each genetic

component individually (Figure 4.13, left). Genetic variants within the antigen presentation pathway are encoded additively (0/1/2) to represent allele dosage. The outcome variable is coded as 0 for spontaneous clearance and 1 for chronic infection.

To visualise the marginal associations, we generated a Manhattan plot with components colour-coded by type (Figure 4.13). The red dashed line indicates the significance threshold of $p < 5 \times 10^{-4}$. As shown in the left panel, several HLA alleles are initially associated with the phenotype before conditional analysis. However, due to strong linkage disequilibrium within the HLA region, we performed a stepwise conditional analysis: first conditioning on the top signal, *HLA-DQB1*03:01*, and then on the second strongest signal, *HLA-DRB1*01:01*. After conditioning on these two alleles, no additional variants exceeded the significance threshold, indicating that these two loci explain the primary genetic associations with HCV outcome in this dataset. The detailed conditional regression results are provided in Appendix Table B.1.

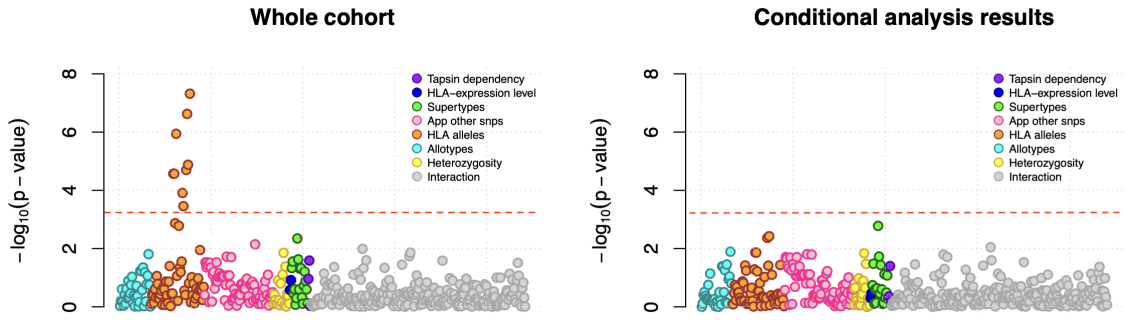


Figure 4.13: **Conditional analysis results for spontaneous clearance versus chronic infection**. The left panel shows the results before conditional analysis, and the right panel shows the results after conditioning. Colours indicate different categories of genetic components: light blue represents gene allotypes, including proteasome subunit allotypes, TAP1/2, and ERAP1/2 allotypes; orange points represent HLA alleles; pink points correspond to genetic variants within the antigen presentation pathway; yellow points indicate HLA heterozygosity; dark blue points denote HLA protein expression levels (HLA-A and HLA-C); and grey points represent interaction terms between HLA class I alleles and ERAP1/2.

STOPHCV For the STOP-HCV cohort of patients with chronic HCV infection, we also performed conditional analyses to investigate genetic associations with cirrhosis, HCC, and viral load. For cirrhosis, the phenotype is encoded as 1 for cirrhosis and 0 for non-cirrhotic individuals. The covariates included ten host principal components, sex, age, body mass index (BMI), and viral genotype. Logistic regression models are

fitted using the `glm()` function in `R`, testing each genetic variable individually against the phenotype. Only one allele, *HLA-DQB1*03:01*, showed a significant association with cirrhosis. For HCC, cases are encoded as 1 and non-HCC individuals as 0. The model includes cirrhosis status, 10 PCs, sex, age, BMI, and viral genotype as covariates, and logistic regression is again applied using `glm()`. For viral load, we use the same covariates as in the HCC model and employ linear regression with the `lm()` function in `R`. The detailed results of these analyses are provided in Appendix Table B.2.

Comparison of Conditional and Bayesian Joint Regression Analyses In summary, we integrate and compare findings from two HCV datasets: one on spontaneous clearance versus chronic infection, and the other from the STOP-HCV cohort, which includes three phenotypes (cirrhosis, hepatocellular carcinoma, and viral load). A comparison of the coefficients from the conditional analyses and the Bayesian joint regression is presented in Table 4.13. In the spontaneous clearance analysis, the Bayesian joint model identified one compelling signal, the *HLA-DRB1*01:01* allele under a dominant model, whereas the conditional analysis detected two independent alleles: *HLA-DRB1*01:01* and *HLA-DQB1*03:01*.

For the chronic infection phenotypes in the STOP-HCV cohort, the two methods yielded largely consistent results. For cirrhosis, both methods identified a protective association with *HLA-DQB1*03:01*, though the effect size is slightly larger in the conditional analysis. For viral load, the conditional analysis identified *HLA-DQB1*03:01* as a significant protective factor. The Bayesian model selected the same variant with the same direction of effect, but with a shrunken coefficient that did not meet the significance threshold.

The observed differences can be attributed to the distinct modelling frameworks. The Bayesian joint regression employs a regularised horseshoe prior, which applies statistical shrinkage to the coefficients, typically leading to more conservative effect size estimates compared to the unpenalised conditional analysis. Furthermore, the Bayesian model jointly fits additive, dominant, and recessive effects. The standard errors for dominant and recessive effects are typically larger than for additive effects, partly because the relevant genotype groups (e.g., homozygotes for recessive effects) have smaller sample sizes, leading to less precise estimates.

Table 4.13: **Comparison of conditional and Bayesian joint analysis results for the HCV datasets: Spontaneous clearance vs. chronic infection and STOPHCV.**

	(Non-Bayesian) Conditional analysis				Bayesian joint analysis				
	Coeff.	SE	*p-value*	$\log_{10}$ **BF**	Effect	Coeff.	SE	*p-value*	$\log_{10}$ **BF**
Spontaneous clearance vs. chronic infection									
HLA-DRB1*01:01	0.671	0.120	2.13×10^{-8}	5.84	dominant	0.61	0.14	8.64×10^{-6}	3.42
HLA-DQB1*03:01	0.481	0.079	1.00×10^{-9}	7.08	additive	0.27	0.13	0.033	0.24
STOP-HCV — Cirrhosis									
HLA-DQB1*03:01	-0.350	0.088	6.93×10^{-5}	2.59	additive	-0.405	0.094	1.50×10^{-5}	3.195
STOP-HCV — Viral load									
HLA-DQB1*03:01	-0.180	0.032	1.28×10^{-8}	5.97	dominant	-0.154	0.048	0.00123	1.215

4.3.2.2 MalariaGEN

For the MalariaGEN dataset, we performed conditional analysis by encoding malaria cases as 1 and controls as 0. Five host PCs and sex are included as covariates. Logistic regression models are fitted using the `glm()` function in `R`, and analyses are conducted separately for two geographic regions: Western Africa (Mali, Nigeria, Cameroon, Ghana, Burkina Faso, and Gambia) and Eastern Africa (Tanzania, Kenya, and Malawi). The detailed conditional regression results are provided in Appendix Table B.3.

When comparing the conditional and Bayesian joint analyses (Table 4.14), we observed that the variant *rs75862629* in *ERAP2* appeared in both analyses and is consistently associated with increased risk. This concordance suggests a robust genetic effect across models. In contrast, the expression-associated variable for *HLA-C* (protein expression level) showed a significant association in the conditional analysis but not in the Bayesian joint regression. Nevertheless, this variable is still selected by the Bayesian model, albeit with reduced effect size and statistical support, reflecting the conservative nature of the hierarchical shrinkage prior used in the Bayesian framework.

Table 4.14: **Comparison of conditional analysis and Bayesian joint analysis results for the MalariaGEN dataset.**

		(Non-Bayesian) Conditional analysis				**Bayesian joint analysis**				
Region / SNP	Gene	Effect size	SE	*p-value*	$\log_{10}$ **BF**	Genetic effect	Coeff.	SE	*p-value*	$\log_{10}$ **BF**
Western Africa: Mali, Nigeria, Cameroon, Ghana, Burkina Faso, Gambia										
rs75862629	ERAP2	0.25	0.06	4.54×10^{-5}	2.65	additive	0.25	0.06	9.83×10^{-5}	2.34
rs2549794			/			recessive	-0.23	0.06	0.000108	2.30
Eastern Africa: Tanzania, Kenya, Malawi										
HLA-C (protein expression)	HLA	-0.12	0.03	2.01×10^{-5}	3.02		-0.10	0.03	1×10^{-3}	1.46

4.3.2.3 UKbiobank serological panel

Review and comparison with previous study In a previous study, Guillaume et al. (Butler-Laporte et al., 2020) analysed the UK Biobank serological panel focusing specifically on the HLA region. Our study extends this work by incorporating genetic components spanning the entire antigen presentation pathway. While our analysis confirms their findings, we have identified additional genetic signals, primarily attributable to methodological differences between the two approaches.

Methodologically, our approaches diverged in two key aspects. First, while Guillaume et al. applied a logarithmic transformation to antibody MFI (median fluorescence intensity) values, we implemented a quantile-based normalisation approach. Second, to minimise false associations, Guillaume et al. employed a conservative variable selection procedure, performing 10-fold cross-validation 100 times and retaining only variables selected by Lasso in at least 95 out of 100 iterations using the 1 standard error selection rule via the *glmnet* package. In contrast, we applied the fine-mapping method described in Chapter 3 using our custom R package, *mapHS*.

To evaluate the potential conservatism of Guillaume et al.'s selection rule, we conducted a simulation study based on the linear model:

$$Y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_0,$$

where only $\beta_1 = 0.2$ is non-zero, and the heritability of x_1 varied from 1% to 5%. Applying Guillaume et al.'s selection procedure with repeated 10-fold cross-validation, we recorded the frequency with which x_1 is selected (Table 4.15).

Table 4.15: Simulation Study: Variable Selection Frequency by Heritability

Heritability (%)	Selection Count (out of 100)	Selected
1	1	No
2	45	No
3	92	No
4	100	Yes
5	100	Yes

Our simulations demonstrate that Guillaume et al.'s selection threshold exhibits considerable conservatism. Notably, a variable with 3% heritability is selected in 92% of replicates but would be excluded under their 95% threshold, suggesting their approach may fail to detect genuine medium-strength signals. Given that typical genetic heritability for many complex traits falls around 1-3%, this conservatism

could lead to substantial missing heritability. Our Bayesian inference framework, validated in Chapter 3, has demonstrated reliable detection of signals with as low as 1% heritability in simulation studies. Consequently, our identification of additional signals in real datasets might represent a more comprehensive capture of genuine genetic effects.

Comparison with stepwise conditional analysis In the UK Biobank serological panel cohort, we performed marginal and conditional analyses for both seroprevalence (case–control) and quantitative MFI phenotypes. Covariates included age, sex, and the first 20 host PCs. Logistic regression is applied to binary serostatus traits, and linear regression is used for quantitative MFI traits, both regression analyses implemented using the `glm()` and `lm()` functions in `R`, respectively. The detailed conditional analysis results are provided in Appendix Table B.4-5.

Cases and controls The comparison between the conditional and Bayesian joint analyses for the seroprevalence traits is summarised in Table 4.16. Overall, the two approaches produce highly consistent results, although modest differences are observed in the estimated coefficients, effect models, and standard errors. For JC virus (JCV), both analyses identified a strong association at HLA-DRB1*15:01, which acted as a major risk allele. The Bayesian joint model inferred a dominant effect with a coefficient of -0.575 and a standard error of 0.059, whereas the conditional analysis estimated an additive effect with a coefficient of -0.634 and a smaller standard error of 0.046. The slightly attenuated effect size and larger uncertainty in the Bayesian framework reflect its shrinkage property, which penalises extreme estimates and tends to produce more conservative results. Moreover, modelling the effect as dominant in the Bayesian analysis naturally corresponds to a lower effective allele frequency than under an additive model, which may further contribute to the smaller coefficient magnitude. For Merkel cell polyomavirus (MCV), three alleles, HLA-DRB1*01:01, HLA-DQB1*03:02, and HLA-DQB1*06:02/HLA-DRB1*15:01, are detected by at least one of the methods. Both analyses agreed on the first two associations, though the Bayesian estimates are generally smaller due to shrinkage. A notable difference appeared for the third signal: the conditional analysis identified HLA-DQB1*06:02 (coefficient -0.505, SE 0.047), while the Bayesian joint analysis highlighted HLA-DRB1*15:01 (coefficient -0.447, SE 0.067). Given the very high correlation between these two alleles (Pearson's $r = 0.96$), they likely represent the same underlying genetic signal rather than distinct associations. For T. gondii, both

analyses consistently detected HLA-A*03:01 as a significant allele. The conditional analysis estimated a coefficient of 0.218 (SE 0.050) under an additive model, whereas the Bayesian joint model yielded a slightly smaller coefficient of 0.188 (SE 0.052). Similar to the JCV results, the Bayesian shrinkage led to more conservative estimates with marginally higher uncertainty.

Table 4.16: **Comparison of conditional and Bayesian joint analysis results for the UKBiobank serological panel (cases-controls).**

		(Non-Bayesian) Conditional analysis				**Bayesian joint analysis**				
Virus	HLA	Coeff.	SE	*p-value*	$\log_{10}$ **BF**	Effect	Coeff.	SE	*p-value*	$\log_{10}$ **BF**
JCV										
	HLA-DRB1*15:01	-0.634	0.046	$< 1 \times 10^{-16}$	40.66	dominant	-0.575	0.059	$< 1 \times 10^{-16}$	19.55
MCV										
	HLA-DRB1*01:01	0.296	0.062	1.51×10^{-6}	4.02	additive	0.317	0.067	2.49×10^{-6}	3.81
	HLA-DQB1*03:02	-0.264	0.054	9.24×10^{-7}	4.22	dominant	-0.286	0.075	0.000138	2.17
	HLA-DQB1*06:02	-0.505	0.047	$< 1 \times 10^{-16}$	24.13	Shrink to $\lvert\text{coeff.}\rvert < 10^{-5}$				
	HLA-DRB1*15:01			/			-0.447	0.067	2.49×10^{-11}	8.60
T.gondii										
	HLA-A*03:01	0.218	0.050	1.19×10^{-5}	3.19	additive	0.188	0.052	0.000303	1.88

Note. The correlation between HLA-DQB1*06:02 and HLA-DRB1*15:01 is 0.96.

MFI For quantitative seropositive MFI traits, we perform a conditional analysis using linear regression via the `lm` function for each antigen-pathogen pair. Detailed results are provided in Appendix Table B.5. We compare these findings with those from the Bayesian joint analysis. A summary of all significant alleles from the conditional analysis, alongside their counterparts from the Bayesian regression, is presented in Table 4.17, with key methodological differences noted.

The primary distinction lies in the modelling approaches. The conditional analysis assumes a strictly additive genetic model, whereas the Bayesian model is flexible enough to capture more complex genetic architectures, including additive, dominant, or recessive effects. Consequently, the two models may implicate the same allele but assign it different effect modes. As expected, the Bayesian joint model, which employs shrinkage priors, yielded slightly more conservative effect size estimates, as visualised in Figure 4.14.

Furthermore, the two methods sometimes identified different, yet highly correlated alleles. For instance, for VZV (GE & GI), the Bayesian analysis highlighted *HLA-DRB1*03:01*, whereas the conditional analysis identified *HLA-DQB1*02:01*; these two alleles exhibit a Pearson correlation of 0.98. Given the current data, it remains challenging to definitively identify the true causal allele. Unlike the common practice of excluding one of a pair of highly correlated alleles in advance, our Bayesian joint regression approach retained all such variants. In this high-dimensional context, model optimisation may converge to a local rather than a global maximum. Although increasing the number of permutations could enhance confidence in identifying the global maximum and thereby the most likely causal allele, the results obtained with the current permutation settings remain statistically robust.

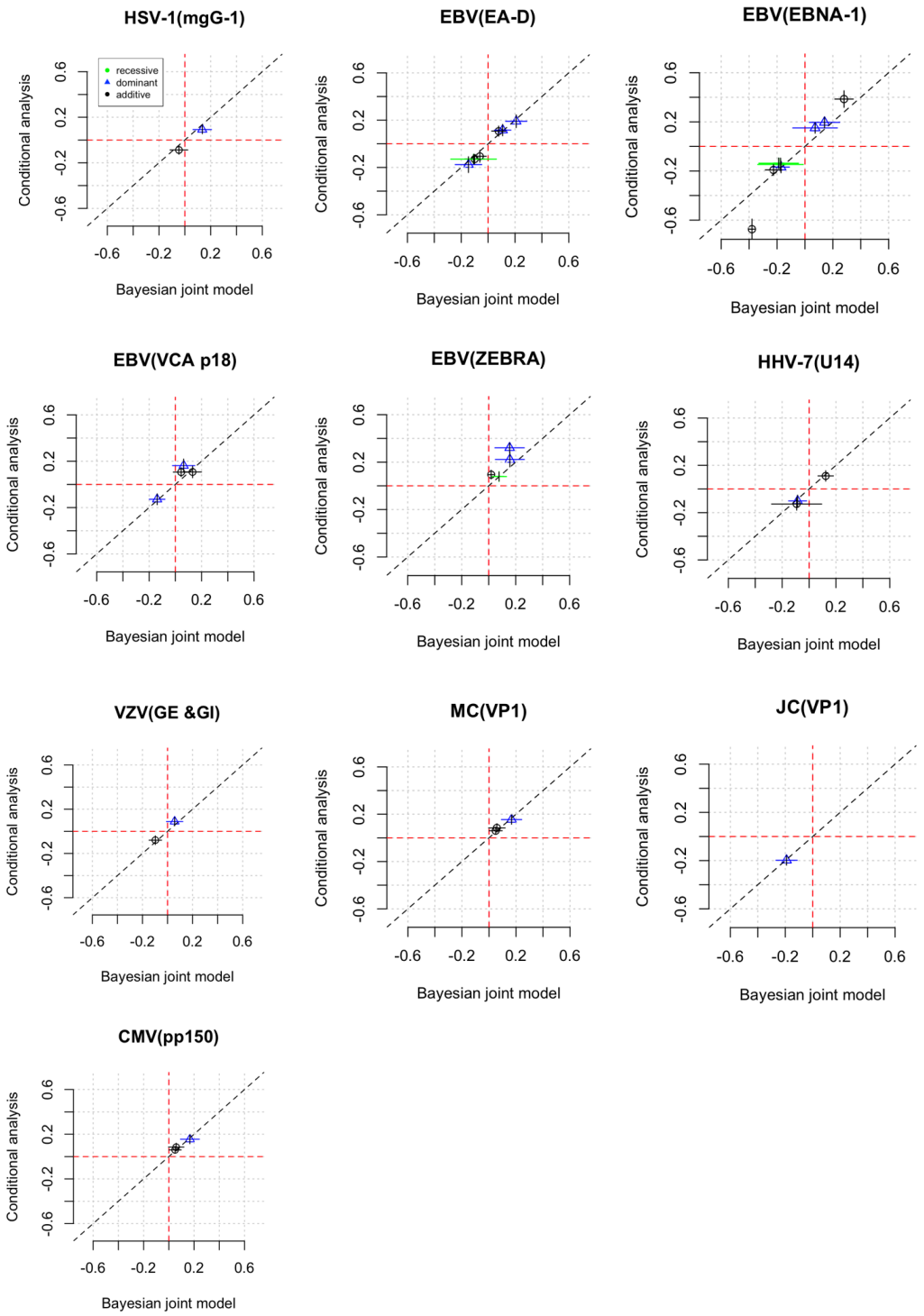


Figure 4.14: **Comparison of effect sizes from Bayesian joint and conditional analyses for UKBiobank serological MFI.** Points represent allele effect estimates, with error bars showing Coeff. $\pm$ 2 $\times$ SD . Colours indicate the genetic component (additive, dominant, or recessive) identified by the Bayesian joint model.

Table 4.17: **Comparison of conditional and Bayesian joint analysis results for the UKBiobank serological panel (MFI)**

		(Non-Bayesian) Conditional analysis				Bayesian joint analysis				
Virus	HLA	Effect size	SE	*p-value*	$\log_{10}$ **BF**	Model	Coeff.	SE	*p-value*	$\log_{10}$ **BF**
HSV-1 (mgG-1)										
	HLA-DRB1*03:01	0.092	0.018	6.52×10^{-07}	4.184	dominant	0.135	0.036	1.73×10^{-04}	1.870
	HLA-DQB1*06:02	-0.087	0.019	3.85×10^{-06}	3.443	additive	-0.045	0.034	0.184	-0.79
EBV (EA-D)										
	HLA-A*03:01	0.107	0.021	5.44×10^{-07}	4.194	additive	0.079	0.026	2.55×10^{-03}	0.723
	HLA-B*44:02	-0.106	0.025	2.34×10^{-05}	2.629	Shrink to $\lvert\text{coeff.}\rvert < 10^{-5}$				
	HLA-DQB1*02:01	-0.194	0.022	6.41×10^{-19}	15.927	dominant	-0.145	0.0494	0.00319	0.634
	HLA-DQB1*03:02	-0.176	0.033	1.49×10^{-07}	4.738	dominant	-0.146	0.0490	3.19×10^{-03}	0.634
	HLA-DQA1*01:02	-0.130	0.020	8.34×10^{-11}	7.907	additive	-0.100	0.036	5.91×10^{-03}	0.392
		/				recessive	-0.108	0.085	0.203	-0.897
	HLA-DQA1*03:01	0.191	0.027	1.25×10^{-12}	9.701	dominant	0.209	0.040	1.41×10^{-07}	4.750
	HLA-DPB1*03:01	0.116	0.024	1.43×10^{-06}	3.790	dominant	0.108	0.031	4.04×10^{-04}	1.461
	HLA-B*44:02_ERAP2.TKL			/			-0.062	0.024	9.84×10^{-03}	0.195

Note. HLA-B*44:02 is correlated with HLA-B*44:02_ERAP2.TKL with 0.76

EBV (EBNA-1)										
	HLA-DRB4*01:03	-0.137	0.021	1.30×10^{-10}	7.717	recessive	-0.188	0.071	8.21×10^{-03}	0.265
	HLA-DRB1*07:01	-0.192	0.022	6.31×10^{-18}	14.939	additive	-0.225	0.030	2.89×10^{-14}	11.266
	HLA-DRB1*15:01	0.196	0.022	8.99×10^{-19}	15.781	dominant	0.140	0.055	1.03×10^{-02}	0.178
	HLA-DQB1*02:01	-0.674	0.041	1.43×10^{-59}	57.095	additive	-0.379	0.0316	$< 1 \times 10^{-16}$	29.96

Continued on next page

Table 4.17: **Comparison of conditional and Bayesian joint analysis results for the UKBiobank serological panel (MFI)**

		(Non-Bayesian) Conditional analysis				Bayesian joint analysis				
Virus	HLA	Effect size	SE	*p-value*	$\log_{10}$ **BF**	Model	Coeff.	SE	*p-value*	$\log_{10}$ **BF**
	HLA-DQB1*03:01	-0.147	0.026	1.89×10^{-08}	5.608	recessive	-0.175	0.0816	0.032	-0.25
	HLA-DQA1*05:01	0.386	0.034	3.96×10^{-30}	27.145	additive	0.280	0.033	$< 1 \times 10^{-16}$	14.115
	HLA-DPB1*03:01	-0.168	0.023	5.00×10^{-13}	10.091	dominant	-0.172	0.030	1.34×10^{-08}	5.736
	DP1	0.151	0.022	4.18×10^{-12}	9.183	dominant	0.0724	0.0803	0.36	-1.072
	DRB3	0.116	0.027	1.93×10^{-05}	2.709	Shrink to $\lvert\text{coeff.}\rvert < 10^{-5}$				
	HLA-DRB1*03:01			/		recessive	-0.521	0.088	3.38×10^{-09}	6.317
Note. DRB3 is correlated with HLA-DRB1*03:01 with 0.61.										
EBV (ZEBRA)										
	HLA-A*03:01	0.117	0.020	1.04×10^{-08}	5.861	Shrink to $\lvert\text{coeff.}\rvert < 10^{-5}$				
	HLA-DQB1*03:02	0.309	0.023	2.66×10^{-40}	37.421	dominant	0.157	0.054	3.61×10^{-03}	0.585
	HLA-DQA1*02:01	0.308	0.021	7.34×10^{-49}	46.127	dominant	0.155	0.055	4.53×10^{-03}	0.496
	HLA-DPA1*01:03	0.101	0.019	7.35×10^{-08}	5.035	additive	0.02	0.01	0.06	-0.51
			/			recessive	0.01	0.0316	0.304	-1.01
	B44	0.095	0.019	3.59×10^{-07}	4.369	dominant	0.076	0.028	6.45×10^{-03}	0.358
	HLA-A*03:01-ERAP1.TNL			/			0.016	0.033	0.62	-1.19
Note. cor(HLA-A*03:01, HLA-A*03:01_ERAP2.TNL) = 0.74.										
EBV (VCA.p18)										
	TAP2.TVACL	0.162	0.028	1.29×10^{-08}	5.768	dominant	0.0633	0.0427	0.138	-0.772
	HLA-C*07:02	0.140	0.023	1.44×10^{-09}	6.697	Shrink to $\lvert\text{coeff.}\rvert < 10^{-5}$				

Continued on next page

Table 4.17: **Comparison of conditional and Bayesian joint analysis results for the UKBiobank serological panel (MFI)**

		(Non-Bayesian) Conditional analysis				Bayesian joint analysis				
Virus	HLA	Effect size	SE	*p-value*	$\log_{10}$ **BF**	Model	Coeff.	SE	*p-value*	$\log_{10}$ **BF**
	HLA-C*07:02_ERAP2.TKL			/			0.074	0.0316	0.191	-0.060
	HLA-DRB1*01:01	0.107	0.025	1.57×10^{-05}	2.794	additive	0.131	0.034	1.39×10^{-04}	1.893
	HLA-DRB1*15:01	0.138	0.023	3.34×10^{-09}	6.340	Shrink to $\lvert\text{coeff.}\rvert < 10^{-5}$				
	HLA-DQA1*01:02			/		additive	0.0865	0.0316	6.18×10^{-03}	0.374
	HLA-DPB1*03:01	-0.129	0.023	4.42×10^{-08}	5.249	dominant	-0.139	0.030	3.66×10^{-06}	3.392
	B44	0.106	0.019	3.88×10^{-08}	5.305	additive	0.044	0.030	0.14	-0.78
Note. cor(HLA-C*07:02, HLA-C*07:02_ERAP2.TKL) = 0.74; cor(HLA-DRB1*15:01, HLA-DQA1*01:02) = 0.83.										
HHV-7 (U14)										
	HLA-C*03:04	-0.127	0.026	1.36×10^{-06}	3.812	Shrink to $\lvert\text{coeff.}\rvert < 10^{-5}$				
	HLA-B*40:01			/		additive	-0.0843	0.0437	0.053	-0.444
	HLA-DRB1*01:01	-0.101	0.025	4.07×10^{-05}	2.401	dominant	-0.085	0.034	1.11×10^{-02}	0.146
	HLA-DRB1*04:01	0.110	0.022	7.29×10^{-07}	4.071	additive	0.124	0.029	1.61×10^{-05}	2.778
	HLA-DPB1*04:02	-0.096	0.023	2.98×10^{-05}	2.529	additive	-0.080	0.028	3.55×10^{-03}	0.592
Note. cor(HLA-C*03:04, HLA-B*40:01) = 0.80.										
JCV (VP1)										
	HLA-DRB1*15:01	-0.186	0.021	2.60×10^{-19}	16.452	dominant	-0.191	0.039	8.60×10^{-07}	4.094
MCV (VP1)										
	HLA-DQB1*03:01	0.091	0.018	1.98×10^{-07}	4.694	additive	0.058	0.031	0.058	-0.39

Continued on next page

Table 4.17: **Comparison of conditional and Bayesian joint analysis results for the UKBiobank serological panel (MFI)**

		(Non-Bayesian) Conditional analysis				**Bayesian joint analysis**				
Virus	HLA	Effect size	SE	*p-value*	$\log_{10}$ **BF**	Model	Coeff.	SE	*p-value*	$\log_{10}$ **BF**
	HLA-DQB1*05:01	0.153	0.020	1.57×10^{-14}	11.658	dominant	0.167	0.038	9.01×10^{-06}	3.090
	HLA-DQB1*06:02	-0.104	0.020	3.03×10^{-07}	4.515	Shrink to $\lvert\text{coeff.}\rvert < 10^{-5}$				
	HLA-DRB1*15:01			/		additive	-0.119	0.033	3.07×10^{-04}	1.647
Note. cor(HLA-DRB1*15:01, HLA-DQB1*06:02) = 0.97.										
VZV (GE & GI)										
	HLA-A*01:01	0.089	0.019	5.28×10^{-06}	3.251	dominant	0.0566	0.0328	0.084	-0.601
	HLA-DQB1*02:01	0.157	0.021	1.38×10^{-13}	10.648	Shrink to $\lvert\text{coeff.}\rvert < 10^{-5}$				
	HLA-DRB1*03:01			/		dominant	0.137	0.031	8.35×10^{-06}	3.054
Note. cor(HLA-DQB1*02:01, HLA-DRB1*03:01) = 0.98										

4.4 Summary

In this chapter, we applied the Bayesian joint regression model introduced in Chapter 3 to four large-scale datasets to investigate associations between genetic components of the antigen presentation pathway and diverse infectious disease phenotypes. Notably, the RHS-based joint model identified several associations that were missed by marginal and conditional analyses, particularly in settings with correlated predictors (e.g., UKB serology, ERAP × HLA interactions). This demonstrates the practical advantage of the proposed methodology.

In the HCV analyses, the allele *HLA-DRB1*01:01* (dominant effect) reached significance in the spontaneous clearance dataset, acting as a protective allele associated with an increased likelihood of viral clearance. In addition to HLA alleles, several other signals were selected by the model, although they did not reach the significance threshold. For example, the *HLA-A*02:01* and *ERAP1* allotype interaction, and the B58 supertype, both of which were suggestive of protective effects. In the STOP-HCV cohort, *HLA-DQB1*03:01* met the significance threshold and was also protective, showing a reduced risk of cirrhosis. These findings underscore the critical role of HLA class II alleles in orchestrating effective antiviral immune responses.

In the MalariaGEN study, two variants in ERAP2 rs75862629 (promoter) and rs2549794 (intronic) were significantly associated with severe malaria outcomes in West Africa. The promoter variant exhibited an additive effect that increased disease susceptibility, whereas the intronic variant showed a recessive protective effect. Additionally, rs9277972 in TAPBP was identified by the model as a potential risk factor, although it did not reach the significance threshold. In East Africa, no variants met the significance threshold; however, several model-selected signals, such as HLA-C protein expression levels and PSMB8 allotypes, appeared noteworthy. These population-specific associations, particularly those observed in West African cohorts, may also reflect differences in parasite types and local malaria transmission dynamics, suggesting adaptive immune genetic variation driven by region-specific selection pressures.

In the UK Biobank serological panel, both case–control and quantitative (MFI) analyses were performed to investigate how genetic variation in the antigen presentation pathway shapes antibody responses to 20 infectious agents. The Bayesian joint regression model identified robust and recurrent effects of HLA class II alleles, particularly *HLA-DRB1*15:01*, *HLA-DQB1*05:01*, and *HLA-DPB1*04:01*, which were consistently associated with the Herpesviridae and Polyomaviridae families. These

alleles exhibited additive or dominant effects linked to increased seropositivity and higher antibody intensity, indicating that they modulate both infection susceptibility and immune response magnitude. Among these, *HLA-DRB1*15:01* showed a shared additive risk effect for Epstein–Barr virus and JC virus, while *HLA-DQB1*05:01* exerted a dominant protective effect against Merkel cell polyomavirus infection. Beyond classical alleles, several interaction effects emerged, especially between HLA class I alleles and ERAP1/2 allotypes, although most were modest. For example, an *HLA-A*24:02–ERAP1* interaction appeared for *Chlamydia trachomatis* PorB, and an *HLA-C*04:01–ERAP2* interaction was observed for *Helicobacter pylori* CagA, suggesting potential epistatic modulation of peptide processing and antigen presentation. While these signals did not surpass the strict significance threshold, their recurrence across related bacterial antigens supports subtle pathway-level coordination between HLA and ERAPs' function. HLA supertypes also contributed to population-level variation in antibody profiles, with B07/B44 supertypes showing slightly elevated seroprevalence across multiple viral families, consistent with their broad peptide-binding repertoires. In contrast, no significant APP associations were found for *Toxoplasma gondii*, reflecting limited involvement of this pathway in anti-parasitic antibody variation.

Together, these results reveal that HLA class II alleles and HLA–ERAP interactions jointly influence both the likelihood and intensity of pathogen-specific antibody responses. The replication of these effects across case–control and MFI analyses underscores the robustness of the Bayesian joint model in uncovering pleiotropic, interaction-driven, and supertype-level immune-genetic patterns that are often obscured in conventional univariate approaches.

4.5 Discussion

In this chapter, we applied the Bayesian joint regression model and its corresponding `mapHS` implementation (developed in Chapter 3) to four distinct real-world datasets to investigate associations between genetic components of the APP and infectious disease phenotypes. By jointly modelling both main and interaction effects across APP genes, this approach provided a more integrated framework for disentangling complex genetic mechanisms than traditional single-locus or stepwise conditional analyses.

Across the four datasets, we detected a broader spectrum of significant associations, including several biologically plausible allotypes and interaction effects. These findings underscore the capability of our Bayesian joint model to capture subtle yet

meaningful patterns of genetic association that are often overlooked by conventional methods, thus offering a richer understanding of how genetic variation in the APP contributes to immune response diversity. Comparison with stepwise conditional analyses and prior studies revealed that while classical approaches tend to identify the strongest marginal effects, our Bayesian framework additionally uncovers dominant, recessive, and interaction-driven associations. This highlights the advantage of modelling the pathway's joint genetic architecture rather than testing variants individually.

We reviewed previous research examining associations between genetic variation in the antigen presentation pathway and phenotypes relating to HCV infection, malaria, and antibody responses. In the sections below, we place the findings of this thesis in the context of that literature, including prior analyses conducted in the same cohorts as well as published studies in other populations. We focus particularly on how the APP associations observed in this thesis compare with previously reported signals.

Host genetic variation within the HLA region plays a central role in determining the outcome of HCV infection, and several of the HLA signals observed in this thesis are consistent with earlier reports. Among class I alleles, *HLA-B*57* has been associated with spontaneous control of HCV infection and preservation of targeted viral epitopes, supporting a role for effective $CD8^+$ T cell–mediated immunity (Kim et al., 2011a). Similarly, subtype-specific effects of *HLA-B*27* influence $CD8^+$ T cell targeting and viral evolution, indicating that HLA-B*27-mediated immune pressure can contribute to viral control (Nitschke et al., 2014). A single-source outbreak study further demonstrated that distinct HLA class I alleles are associated with spontaneous viral clearance, reinforcing the importance of class I, which restricted antiviral responses (McKiernan et al., 2004). Class II loci have also been implicated in infection outcome: MHC class II genotype influences viral persistence versus resolution (Thursz et al., 1999), and multi-cohort genome-wide association studies confirmed that the MHC region harbours the strongest host genetic signals for spontaneous HCV clearance (Duggal et al., 2013). More recently, trans-ancestral fine-mapping identified specific amino acid residues within *HLA-DQβ1* that are independently associated with spontaneous clearance, demonstrating that structural variation within peptide-binding grooves contributes to antiviral immunity (Jones et al., 2022). Beyond classical HLA genes, variation in antigen-processing genes has been evaluated in HCV infection, although evidence has been more limited. Polymorphisms in **ERAP1** and **ERAP2** influence protein expression and isoform profiles in multiple immune-mediated diseases (Hanson et al., 2018), but direct associations between ERAP variants and HCV clinical outcomes have not been consistently established. Variation

in the transporter associated with antigen processing has also been investigated: in a Han Chinese population, combinations of **TAP** and **HLA class I** genes were associated with susceptibility to chronic HCV infection, suggesting that gene–gene interactions within the APP may influence viral persistence Tao et al. (2022). In addition, polymorphisms in **LMP7 (PSMB8)** and **TAP2** have been associated with response to interferon/ribavirin therapy in genotype 1 chronic hepatitis C Zang et al. (2017).

Compared with previous research on HCV spontaneous clearance versus chronic infection, our findings extend earlier analyses by evaluating HLA allele associations under multiple genetic models (additive, dominant, and recessive). Most previous studies fitted only an additive genetic model when assessing HLA associations. However, given the high polymorphism and functional diversity of the HLA region, it is important to determine which genetic model best reflects the biological effect of specific alleles. Under an additive model, the effect of an allele is assumed to increase in a stepwise manner with allele dosage, meaning that carrying two copies has approximately twice the effect of carrying one copy. In contrast, a dominant model implies that carrying either one or two copies of an allele produces a similar phenotypic effect. In this case, the presence of the allele alone may be sufficient to influence antigen presentation, potentially through mechanisms such as expansion of the peptide-binding repertoire or functional divergence within peptide-binding groups. A recessive model, by comparison, indicates that the phenotypic effect is only observed when two copies of the allele are present, suggesting that heterozygous carriers behave similarly to non-carriers. Using this framework, our results refine previous findings by identifying different modes of inheritance for key HLA alleles. Specifically, we observed a dominant effect for *HLA-DRB1*01:01* and an additive effect for *HLA-DQB1*03:01*. Analysis of the expanded STOP-HCV dataset further supported an additive effect of *HLA-DQB1*03:01* across the combined cohort (including both European and Asian participants). Sensitivity analyses indicated that this signal was primarily driven by the European subgroup. Notably, *HLA-DQB1*03:01* was associated with a reduced likelihood of progression to cirrhosis in chronic HCV infection, suggesting a protective role. This finding represents a novel observation in the context of chronic HCV progression and is consistent with evidence from spontaneous clearance studies implicating this allele in favourable infection outcomes. In addition, we identified interaction effects between HLA alleles and ERAP variants that have not been previously reported. These findings suggest that coordinated variation in peptide processing

(ERAP1/2) and peptide presentation (HLA molecules) may influence HCV infection outcomes by shaping the repertoire of viral peptides available for immune recognition.

In terms of malaria, a landmark case–control study conducted in The Gambia demonstrated that specific West African *HLA class I* antigens, particularly *HLA-B*53*, were associated with protection against severe *Plasmodium falciparum* malaria (Hill et al., 1991). This finding provided early evidence that HLA-B-restricted antigen presentation influences the risk of life-threatening malaria, supporting a role for cytotoxic T-lymphocyte, mediated immune responses in protective immunity. More recently, regulatory variation in **TAPBP** (tapasin), a key component of the peptide-loading complex, was shown to associate with malaria outcomes in an HLA allotype-dependent manner (Walker-Sperling et al., 2022). Walker-Sperling et al. reported that genetic variation affecting **TAPBP** expression influenced malaria outcomes depending on the intrinsic tapasin dependence of specific HLA class I allotypes.

In our analyses of the MalariaGEN cohort, we identified two SNPs, *rs75862629* and *rs2549794*, that were associated with malaria-related phenotypes in Western Africa. These variants were not reported as significant in previous analyses using the same cohort. One likely explanation is the difference in statistical thresholds: earlier studies applied a genome-wide significance threshold ($P < 1 \times 10^{-8}$), whereas our analysis focused specifically on genes within the antigen presentation pathway and therefore applied a more moderate significance threshold appropriate for a pathway-based investigation. In addition, the effect of *rs2549794* appeared to follow a recessive genetic model in our analysis, whereas previous studies focused on other statistical regression strategies. These differences in analytical strategy may therefore explain why these SNPs were not highlighted in earlier genome-wide analyses but were detectable in our targeted APP-focused analysis.

In the UK Biobank serology cohort and other population-scale serological studies, variation in HLA, particularly amino acid variation in HLA class II proteins, has been shown to influence humoral responses to common viral infections (Hammer et al., 2015; Mentzer et al., 2022). For herpesviruses, host genetic contributions to infection biology and disease manifestations have frequently been attributed to HLA-mediated immune control (Houldcroft and Kellam, 2015). Similarly, genome-wide association studies of hepatitis B virus infection identified variants in the HLA-DP locus as major determinants of chronic infection risk in Asian populations (Kamatani et al., 2009). Associations between HLA class II variation and susceptibility to other pathogens, including *Chlamydia trachomatis* (Pintea-Trifu et al., 2024), further illustrate the broad role of antigen presentation in shaping host responses to infection.

Our findings largely replicate previously reported associations, particularly the HLA signals identified in earlier analyses of antibody responses to common pathogens. In addition to confirming these established associations, our study identified additional signals beyond classical HLA alleles, including interaction effects and variation in other components of the antigen presentation pathway. By examining genetic variation across the broader APP rather than focusing solely on HLA loci, our analysis provides a more comprehensive view of how antigen processing and presentation may influence humoral immune responses. Compared with other datasets investigating genetic determinants of antibody responses, most previously reported associations have been concentrated within the HLA region. In contrast, our results suggest that variation in other APP components may also contribute to differences in serological responses. Where our findings differed from previous reports, several factors may plausibly explain these discrepancies. Population genetic background can influence allele frequencies and linkage disequilibrium patterns within the HLA region, affecting the detectability of associations. In addition, differences in pathogen exposure history and seroprevalence across cohorts may alter the strength of detectable genetic effects. Finally, methodological differences between studies, including phenotype definition, serological assay platforms, and analytical approaches, may contribute to variation in reported results.

Taken together, these considerations suggest that both concordant and divergent APP associations across cohorts are expected. Our findings, therefore, complement previous work by extending genetic analyses beyond classical HLA loci to the broader antigen presentation pathway and by highlighting potential interactions between peptide-processing and peptide-presentation components across diverse infectious disease phenotypes.

Nevertheless, several limitations should be acknowledged. Although the model can distinguish additive, dominant, and recessive genetic patterns, statistical evidence remains insufficient to conclusively determine the precise inheritance mode for each association. Visual inspection of genotype-specific case rates (homozygous versus heterozygous) can suggest patterns of dominance or additivity, but formal hypothesis testing, such as likelihood-ratio tests or Bayes-factor comparisons, would be required for definitive inference. Furthermore, the model's robustness depends on the number of permutations iterations performed; increasing these could enhance accuracy and reduce uncertainty. Future work should focus on refining the framework by incorporating more rigorous hypothesis-testing modules, exploring alternative shrinkage priors, and validating the observed associations through experimental or replication

studies, particularly for the allotype and interaction effects that appear functionally relevant.

In summary, the Bayesian joint regression framework implemented via `mapHS` demonstrates a flexible and powerful approach for analysing high-dimensional genetic data within biologically structured pathways. By integrating additive, dominant, recessive, and interaction effects into a unified model, it provides a comprehensive and interpretable view of how genetic variation in the antigen presentation pathway shapes infectious disease outcomes.

Chapter 5

Discussion and Future Work

We presented a comprehensive analysis of the genetic architecture of the antigen presentation pathway and its influence on infectious disease outcomes. Across four major chapters, we developed and applied an integrated bioinformatics and statistical framework that advances our understanding of how genetic variation within this pathway affects the phenotypes.

Chapter 1 provided the biological and immunological foundation for this research. We reviewed the key molecular components of the APP, particularly HLA molecules, ERAPs, TAPs, and proteasome subunits, and discussed their pivotal roles in shaping adaptive immunity. The chapter highlighted how genetic diversity, linkage disequilibrium, and epistatic interactions across these genes complicate causal inference, motivating the development of integrated analytical approaches.

Chapter 2 addressed this challenge by developing a scalable, modular bioinformatics pipeline to characterise APP genetic features across diverse populations and diseases. The pipeline integrates HLA alleles, supertypes, heterozygosity metrics, protein expression levels, and non-HLA allotypes (e.g., ERAPs, TAPs, and PSMBs) into a unified analytical framework. Applied to multiple large-scale datasets, including HCV, malaria, and the UK Biobank serological panel, it revealed both shared and ancestry-specific immunogenetic patterns, establishing a solid foundation for downstream statistical analyses.

In Chapter 3, we introduced a novel Bayesian fine-mapping framework based on the regularised horseshoe prior, implemented through the `mapHS` package. This model efficiently performs variable selection in high-dimensional, correlated genomic data, allowing the joint inference of additive, dominant, recessive, and interaction effects. Extensive simulations demonstrated that the method provides accurate, interpretable, and computationally tractable inference even under complex linkage disequilibrium structures, outperforming conventional penalised regression approaches.

Chapter 4 integrated the computational and statistical frameworks to jointly analyse genetic variation across the APP in four real-world datasets: two HCV cohorts, the MalariaGEN study, and the UK Biobank serology panel. By modelling the combined effects of HLA, ERAPs, TAPs, and proteasome genes—and their interactions—the analysis revealed several biologically potentially meaningful associations. Notably, HLA-DQB1*03:01 and HLA-DRB1*01:01 were consistently linked to spontaneous HCV clearance and reduced cirrhosis risk, highlighting the central role of HLA class II alleles in antiviral immunity. For malaria, ERAP2 promoter and intronic variants showed region-specific associations with disease severity in West Africa, while higher HLA-C expression appeared protective in East Africa. In the UK Biobank, recurrent signals at HLA-DRB1*15:01 for polyomavirus seropositivity and ERAP–HLA interactions illustrated the widespread influence of APP diversity on immune responses. In parallel, we also compared these findings with those obtained from the traditional conditional analysis framework commonly applied in genetic studies. Conditional analysis can be useful for identifying independent additive or dominant effects within a locus, but it remains fundamentally limited. It conditions on the most statistically significant allele, which can cause correlated or interacting variants to be masked, and it tends to detect only strong marginal associations. In addition, conditional analysis has difficulty distinguishing among additive, dominant, and recessive genetic effects on the phenotype because these modes of inheritance often overlap in their statistical signatures. Since the method is applied on a locus-by-locus basis, it yields only a partial perspective on the APP pathway and fails to capture the coordinated and biologically integrated influences of multiple genes that shape antigen presentation. Collectively, these findings demonstrate the utility of a pathway-centric, Bayesian joint modelling approach for uncovering complex and context-dependent immunogenetic associations.

Taken together, our work presented in this thesis advances both methodological and biological understanding of the antigen presentation pathway. Methodologically, the integration of a feature-rich bioinformatics pipeline with a Bayesian shrinkage framework offers a generalisable template for pathway-level genetic association analysis. This approach bridges the gap between traditional single-variant tests and systems-level models, providing a comprehensive perspective that captures additive, non-additive, and interaction effects.

Biologically, the results underscore the central role of APP variation in modulating susceptibility and the immune response to infectious diseases. The consistent involvement of HLA class II alleles in HCV clearance, the functional impact of ERAP2

variants on malaria outcomes, and the pleiotropic influence of HLA and ERAP2 interactions on serological traits collectively reveal that genetic diversity in antigen processing and presentation underpins both innate and adaptive immune variation. These findings align with evolutionary expectations that polymorphism in the APP has been shaped by pathogen-driven selection, maintaining functional diversity essential for population-level resilience.

Despite these insights, several limitations remain. The reliance on imputed and statistically phased genotypes introduces potential uncertainties in haplotype inference, particularly within highly polymorphic loci. Structural and amino acid–level variation in HLA molecules was not directly modelled, potentially obscuring finer-grained functional effects. Moreover, while the Bayesian joint regression framework distinguishes additive, dominant, and recessive inheritance modes, the current analyses lack formal statistical validation of these classifications. To confirm the true genetic architecture of observed effects, future work should employ rigorous hypothesis testing, such as likelihood-ratio tests or Bayes factor comparisons, to determine whether individual associations follow additive, dominant, or recessive models. Additionally, although the analysis incorporated multiple ancestries, limited sample sizes for underrepresented populations constrained power for detecting population specific associations. Addressing these challenges will require larger, more diverse cohorts and improved molecular resolution in future studies.

In addition, we observed that the Bayesian joint model with a regularised horseshoe prior produced somewhat more conservative effect size estimates compared with non-Bayesian conditional analyses. Several factors likely contribute to this behaviour. First, the joint model includes a very large number of predictors, often on the order of one thousand variables, and applies shrinkage to all of them simultaneously. In contrast, conditional analysis evaluates one variant at a time without shrinkage, which naturally yields larger unregularised coefficients. Second, as discussed in Chapter 3, the regularised horseshoe prior is intentionally designed to shrink noisy or weak effects toward zero while allowing only a small subset of variables to escape shrinkage. This behaviour improves robustness but may slightly attenuate true effect sizes. Third, the HLA and ERAPs regions contain extensive linkage disequilibrium. In our modelling framework, we further decomposed each allele into additive, dominant, and recessive components and included numerous interaction terms across genes. This induces substantial multicollinearity in the predictor matrix. Under such conditions, the Bayesian model distributes effect sizes across correlated predictors and applies

stronger shrinkage to stabilise estimates, which can appear conservative when compared with single variant conditional tests. Although this behaviour is expected, future refinements could explore alternative priors, dimension reduction strategies, or hierarchical structures to more accurately capture strong effects in highly correlated regions.

Several promising directions emerge from this work. First, given the central role of HLA-presented peptides in shaping $CD8^+$ T cell responses, it would be valuable to investigate how differential trimming of viral peptides by ERAP1 allotypes, together with allele-specific HLA binding, influences immune recognition and viral persistence during chronic infection. In particular, leveraging viral sequence data would allow the peptide-level consequences of ERAP1–HLA variation to be mapped directly onto viral diversity. Such analyses could identify whether specific ERAP1 allotypes preferentially generate or eliminate key epitopes, whether certain HLA alleles impose detectable selection pressure on viral sequences, and how host and pathogen co-evolution shapes the landscape of immune escape. Incorporating viral sequence information would therefore make it possible to connect host immunogenetic variation to pathogen adaptation in a mechanistic and temporally resolved manner.

Second, integrating multi-omics datasets including transcriptomics, proteomics, and epigenomics would enable the construction of dynamic, mechanistic models that link genetic variation to gene expression, protein abundance, and antigenic peptide repertoires. Such analyses could directly connect statistical associations to molecular function and bridge the gap between genotype and immunological phenotype.

Third, experimental validation of the putative epistatic interactions identified here, particularly between specific HLA alleles and ERAPs allotypes, represents a key next step. Techniques such as T cell activation assays and CRISPR based functional screens could test whether these genetic combinations alter peptide processing or T cell recognition in measurable ways.

Finally, the methodology developed here could be extended to other biological pathways, such as the HCV entry pathway. Expanding this framework to encompass multiple pathways would allow for a more comprehensive understanding of how genetic and molecular interactions shape immune responses across diverse biological contexts.

Ultimately, our work provides a unified statistical and genomic framework for interpreting APP variation in human infectious diseases. By bridging large-scale human genetics with a mechanistic understanding of antigen processing, it lays a

foundation for future translational studies and advances our broader understanding of host and pathogen interactions.

Appendix A

Summary of genetic variants and allotypes Table

Table A.1: Summary of genetic variants, excluding HLA genes, involved in the antigen presentation pathway.

Chr	Gene	Position	rsID	Ref	Alt	Annotation	Protein
6	PSMB9	32857313	rs17587	G	A	missense	Arg60His
6	PSMB9	32858456	rs20547	A	G	synonymous	Ala161Ala
6	PSMB9	32856171	rs241419	G	A	missense	Val32Ile
6	PSMB9	32858490	rs17213861	C	T	missense	Arg173Cys
6	PSMB9	32854255	rs35100697	G	A	missense	Gly9Glu
6	PSMB9	32858459	rs182700479	T	C	synonymous	Tyr162Tyr
6	PSMB9	32854492	rs1351383	A	C	Intron	–
6	PSMB9	32883073	rs2127675	A	G	3' flanking region	–
6	PSMB8	32843852	rs2071543	G	T	missense	Gln49Lys
6	PSMB8	32843868	rs2071542	A	G	synonymous	Ala43Ala
6	PSMB8	32843975	rs114772012	C	T	missense	Gly8Arg
6	PSMB8	32842188	rs41270492	C	T	synonymous	Gln161Gln
6	PSMB8	32843017	rs17220206	T	A	missense	Thr74Ser
6	PSMB8	32843015	rs116638337	G	T	synonymous	Thr74Thr
6	PSMB8	32842170	rs11540143	G	A	synonymous	Leu167Leu
6	PSMB8	32843045	rs79482999	G	A	synonymous	Asn64Asn
16	PSMB10	67936027	rs20549	A	G	synonymous	Leu107Leu
16	PSMB10	67935628	rs14178	A	G	synonymous	Gly151Gly
16	PSMB10	67936697	rs202140443	T	G	missense	Gln18Pro
14	PSMB5	23034812	rs11543947	G	A	missense	Arg24Cys
9	PSMB7	124414882	rs4574	A	G	missense	Val39Ala
9	PSMB7	124356892	rs147487944	C	T	synonymous	Val198Val
17	PSMB6	4796257	rs3169950	G	A	synonymous	Ala21Ala

Continued on next page

Chr	Gene	Position	rsID	Ref	Alt	Annotation	Protein
17	PSMB6	4797724	rs2304975	C	T	synonymous	Ser115Ser
17	PSMB6	4798107	rs7468	C	T	synonymous	Tyr177Tyr
17	PSMB6	4797698	rs2304974	C	G	missense	Pro107Ala
17	PSMB6	4798017	rs11552525	A	T	synonymous	Ser147Ser
6	TAP1	32850997	rs1057141	T	C	missense	Ile333Val
6	TAP1	32847198	rs1135216	T	C	missense	Asp637Gly
6	TAP1	32853214	rs55702652	A	G	synonymous	Val141Val
6	TAP1	32852191	rs41549617	G	A	synonymous	Gly254Gly
6	TAP1	32847125	rs41551515	C	T	synonymous	Pro661Pro
6	TAP1	32850459	rs2127679	G	A	missense	Ala370Val
6	TAP1	32848666	rs41561219	C	T	missense	Val518Ile
6	TAP1	32847165	rs1057149	C	T	missense	Arg648Gln
6	TAP1	32848995	rs41550019	C	A	missense	Val458Leu
6	TAP1	32849112	rs2228110	C	A	missense	Gly419Cys
6	TAP1	32848671	rs2228106	G	A	missense	Pro516Leu
6	TAP1	32852223	rs36229525	C	A	missense	Val244Leu
6	TAP1	32853588	rs57640466	C	G	missense	Gly17Arg
6	TAP1	32850520	rs56366814	G	A	splice region	–
6	TAP1	32845644	rs74897484	G	T	missense	Gln728Lys
6	TAP1	32853199	rs78410191	C	T	synonymous	Ala146Ala
6	TAP1	32853245	rs142907576	A	G	missense	Leu131Pro
6	TAP1	32851137	rs2228111	G	A	missense	Ser286Phe
6	TAP1	32845750	rs56337036	G	A	synonymous	Tyr692Tyr
6	TAP1	32847132	rs121917702	C	T	missense	Arg659Gln
6	TAP1	32854082	rs2071480	G	T	promoter	–
6	TAP2	32837530	rs2071466	C	T	splice region	–
6	TAP2	32828908	rs241448	A	G	stop_lost	Ter687Gl-Ter17
6	TAP2	32829520	rs241441	T	C	synonymous	Gly604Gly
6	TAP2	32828974	rs241447	T	C	missense	Thr665Ala
6	TAP2	32832447	rs2228397	C	A	synonymous	Gly386Gly
6	TAP2	32832635	rs1800454	C	T	missense	Val379Ile
6	TAP2	32830771	rs1042116	G	A	synonymous	Asn436Asn
6	TAP2	32830032	rs2228396	C	T	missense	Ala565Thr
6	TAP2	32822322	rs61021012	G	GA	splice region	–
6	TAP2	32829016	rs4148876	G	A	missense	Arg651Cys
6	TAP2	32822312	rs16870908	G	A	missense	Leu647Phe
6	TAP2	32829532	rs2229527	T	G	synonymous	Val600Val
6	TAP2	32837529	rs2071467	C	T	splice region	–
6	TAP2	32832444	rs2856992	C	T	synonymous	Val387Val
6	TAP2	32835161	rs140654840	C	T	missense	Arg313His
6	TAP2	32829996	rs2228391	T	C	missense	Met577Val

Continued on next page

Chr	Gene	Position	rsID	Ref	Alt	Annotation	Protein
6	TAP2	32832650	rs111303994	C	T	missense	Ala374Thr
6	TAP2	32830705	rs149495208	C	T	synonymous	Thr458Thr
6	TAP2	32835724	rs142794316	G	T	synonymous	Arg220Arg
6	TAP2	32830681	rs137982419	C	T	synonymous	Gly466Gly
6	TAP2	32830680	rs150253319	C	T	missense	Val467Ile
6	TAP2	32832466	rs148353836	A	G	splice region	–
6	TAP2	32832467	rs9461814	A	T	splice region	–
6	TAP2	32838190	rs55827768	A	G	missense	Val15Ala
6	TAP2	32829970	rs79098150	T	C	synonymous	Ala585Ala
6	TAP2	32838072	rs56064400	C	T	synonymous	Lys54Lys
6	TAP2	32829506	rs74770812	G	A	missense	Ala609Val
6	TAP2	32830099	rs241436	A	G	intron	–
5	ERAP1	96793809	rs27434	A	G	synonymous	Ala356Ala
5	ERAP1	96790605	rs27529	A	G	synonymous	Ser453Ser
5	ERAP1	96785820	rs469783	C	T	synonymous	Ala637Ala
5	ERAP1	96792130	rs3213809	G	A	synonymous	His417His
5	ERAP1	96781104	rs17481856	G	A	synonymous	Leu848Leu
5	ERAP1	96781851	rs61745685	C	T	synonymous	Leu763Leu
5	ERAP1	96781150	rs80088786	A	G	synonymous	Phe832Phe
5	ERAP1	96786556	rs30379	T	G	splice region	–
5	ERAP1	96803892	rs72773968	G	A	missense	Thr12Ile
5	ERAP1	96803761	rs3734016	C	T	missense	Glu56Lys
5	ERAP1	96803547	rs26653	C	G	missense	Arg127Pro
5	ERAP1	96795133	rs26618	T	C	missense	Ile276Met
5	ERAP1	96793840	rs27895	C	T	missense	Gly346Asp
5	ERAP1	96793832	rs2287987	T	C	missense	Met349Val
5	ERAP1	96788627	rs30187	T	C	missense	Lys528Arg
5	ERAP1	96786506	rs10050860	C	T	missense	Asp575Asn
5	ERAP1	96783162	rs17482078	C	T	missense	Arg725Gln
5	ERAP1	96783148	rs27044	G	C	missense	Gln730Glu
5	ERAP2	96909735	rs1056893	C	T	synonymous	Ser775Ser
5	ERAP2	96913411	rs2255546	C	T	synonymous	Leu871Leu
5	ERAP2	96901622	rs2287988	G	A	synonymous	Gln563Gln
5	ERAP2	96896438	rs2548538	T	A	synonymous	Pro435Pro
5	ERAP2	96909639	rs2549796	C	T	synonymous	Gly743Gly
5	ERAP2	96879976	rs41506651	C	T	synonymous	Ile97Ile
5	ERAP2	96909015	rs17486915	T	C	synonymous	His689His
5	ERAP2	96883857	rs3733905	C	T	missense	Pro214Leu
5	ERAP2	96889200	rs117041256	T	G	missense	Ser289Ala
5	ERAP2	96895352	rs34261036	T	G	missense	Leu411Arg
5	ERAP2	96909661	rs150892504	C	T	missense	Arg751Cys

Continued on next page

Chr	Gene	Position	rsID	Ref	Alt	Annotation	Protein
5	ERAP2	96912792	rs142362923	T	C	missense	Leu837Ser
5	ERAP2	96886656	rs80193285	T	C	missense	Val239Ala
5	ERAP2	96903530	rs61731306	G	T	missense	Gly661Val
5	ERAP2	96875556	rs75862629	A	G	intron	–
5	ERAP2	96900192	rs2248374	A	G	intron	–
5	ERAP2	96892368	rs75263594	C	T	missense	Thr347Met
5	ERAP2	96895296	rs2549782	G	T	missense	Lys392Asn
5	ERAP2	96903554	rs17408150	T	A	missense	Leu669Gln
5	ERAP2	96908845	rs2549794	C	T	intron	–
6	TAPBP	33272855	rs2071888	G	C	missense	Thr260Arg
6	TAPBP	33271953	rs144706539	G	A	synonymous	Leu418Leu
6	TAPBP	33271573	rs73410025	T	C	synonymous	Gly498Gly
6	TAPBP	33272312	rs61739590	C	T	synonymous	Gly324Gly
6	TAPBP	33271966	rs34132052	G	A	synonymous	Ser413Ser
6	TAPBP	33281505	rs45501592	C	A	synonymous	Pro58Pro
6	TAPBP	33301434	rs59097151	T	C	upstream (2kb)	–
6	TAPBP	33314158	rs111686073	G	C	upstream (2kb)	–
6	TAPBP	33308993	rs9277972	A	T	–	–
19	CALR	13054781	rs1049481	G	T	missense	Gly160Cys
19	CALR	13054615	rs143880510	A	C	missense	Glu381Ala
5	CANX	179126090	rs1134924	C	T	missense	Pro22Leu
5	CANX	179142991	rs78081978	C	T	synonymous	Pro154Pro
5	CANX	179151758	rs75423033	A	G	missense	Glu540Gly
5	CANX	179135353	rs79378421	C	A	missense	Leu140Met
15	ERp57	44038899	rs2411284	C	T	synonymous	Ala54Ala
15	ERp57	44061802	rs1053492	C	T	synonymous	His408His
15	ERp57	44063356	rs3759789	C	T	synonymous	Asn486Asn
15	ERp57	44038850	rs61734331	G	T	missense	Arg38Leu
15	ERp57	44058938	rs112260455	G	A	synonymous	Val286Val

Table A.2: Allotypes in HCV spontaneous clearance vs chronic infection cohort

Gene	SNP	Amino acids	Position
PSMB8	rs2071543	Q/K	49
	rs114772012	G/R	8
PSMB9	rs17587	R/H	60
	rs241419	V/I	32
TAP1	rs1057141	I/V	333
	rs1135216	D/G	637
	rs2127679	A/V	370

Continued on next page

Gene	SNP	Amino acids	Position
	rs41561219	V/I	518
	rs1057149	R/Q	648
	rs41550019	V/L	458
	rs2228110	G/C	419
TAP2	rs241447	T/A	665
	rs1800454	V/I	379
	rs2228396	A/T	565
	rs4148876	R/C	651
	rs16870908	L/F	647
	rs140654840	R/H	313
	rs111303994	A/T	374
ERAP1	rs72773968	T/I	12
	rs3734016	E/K	56
	rs26653	R/P	127
	rs26618	I/M	276
	rs27895	G/D	346
	rs2287987	M/V	349
	rs30187	K/R	528
	rs10050860	D/N	575
	rs17482078	R/Q	725
	rs27044	Q/E	730
ERAP2	rs75263594	T/M	347
	rs2549782	K/N	392
	rs17408150	L/Q	669

Table A.3: Allotypes in STOP-HCV cohort

Gene	SNP	Amino acids	Position
PSMB8	rs2071543	Q/K	49
	rs114772012	G/R	8
PSMB9	rs17587	R/H	60
	rs241419	V/I	32
TAP1	rs1057141	I/V	333
	rs1135216	D/G	637
	rs2127679	A/V	370
	rs41561219	V/I	518
	rs1057149	R/Q	648
	rs41550019	V/L	458

Continued on next page

Gene	SNP	Amino acids	Position
TAP2	rs241447	T/A	665
	rs1800454	V/I	379
	rs2228396	A/T	565
	rs4148876	R/C	651
	rs16870908	L/F	647
ERAP1	rs72773968	T/I	12
	rs3734016	E/K	56
	rs26653	R/P	127
	rs26618	I/M	276
	rs27895	G/D	346
	rs2287987	M/V	349
	rs30187	K/R	528
	rs10050860	D/N	575
	rs17482078	R/Q	725
	rs27044	Q/E	730
ERAP2	rs3733905	P/L	214
	rs75263594	T/M	347
	rs2549782	K/N	392
	rs17408150	L/Q	669

Table A.4: Allotypes in malariaGEN cohort

Gene	SNP	Amino acids	Position
PSMB8	rs2071543	Q/K	49
	rs114772012	G/R	8
	rs17220206	T/S	74
PSMB9	rs17587	R/H	60
	rs241419	V/I	32
	rs35100697	G/E	9
TAP1	rs1057141	I/V	333
	rs1135216	D/G	637
	rs2127679	A/V	370
	rs41561219	V/I	518
	rs1057149	R/Q	648
	rs41550019	V/L	458
	rs2228110	G/C	419
	rs36229525	V/L	244
	rs57640466	G/R	17
	rs74897484	Q/K	728

Continued on next page

Gene	SNP	Amino acids	Position
TAP2	rs241447	T/A	665
	rs1800454	V/I	379
	rs2228396	A/T	565
	rs4148876	R/C	651
	rs16870908	L/F	647
	rs140654840	R/H	313
	rs2228391	M/V	577
	rs111303994	A/T	374
	rs150253319	V/I	467
	rs55827768	V/A	15
	rs74770812	A/V	609
ERAP1	rs72773968	T/I	12
	rs3734016	E/K	56
	rs26653	R/P	127
	rs26618	I/M	276
	rs27895	G/D	346
	rs2287987	M/V	349
	rs30187	K/R	528
	rs10050860	D/N	575
	rs17482078	R/Q	725
	rs27044	Q/E	730
ERAP2	rs3733905	P/L	214
	rs117041256	S/A	289
	rs80193285	V/A	239
	rs61731306	G/V	661
	rs75263594	T/M	347
	rs2549782	K/N	392

Table A.5: Allotypes in UKBiobank serological panel

Gene	SNP	Amino acids	Position
PSMB8	rs2071543	Q/K	49
	rs114772012	G/R	8
PSMB9	rs17587	R/H	60
	rs241419	V/I	32
TAP1	rs1057141	I/V	333
	rs1135216	D/G	637
	rs2127679	A/V	370
	rs41561219	V/I	518

Continued on next page

Gene	SNP	Amino acids	Position
	rs1057149	R/Q	648
	rs41550019	V/L	458
TAP2	rs241447	T/A	665
	rs1800454	V/I	379
	rs2228396	A/T	565
	rs4148876	R/C	651
	rs16870908	L/F	647
ERAP1	rs72773968	T/I	12
	rs3734016	E/K	56
	rs26653	R/P	127
	rs26618	I/M	276
	rs27895	G/D	346
	rs2287987	M/V	349
	rs30187	K/R	528
	rs10050860	D/N	575
	rs17482078	R/Q	725
	rs27044	Q/E	730
ERAP2	rs75263594	T/M	347
	rs2549782	K/N	392
	rs17408150	L/Q	669

Appendix B

Conditional Analysis Results

B.1 HCV: spontaneous clearance vs. chronic infection

Table B.1: **Results of stepwise conditional analysis: Spontaneous clearance vs. chronic infection**

	Effect size	SE	*p-value*	$\log_{10} BF$
HLA-DRB1*01:01	0.671	0.120	2.1×10^{-8}	5.842
HLA-DQB1*03:01	0.481	0.079	1.0×10^{-9}	7.083

B.2 STOP-HCV

Table B.2: **Results of stepwise conditional analysis: STOP-HCV**

	Effect size	SE	*p-value*	$\log_{10} BF$
Cirrhosis				
HLA-DQB1*03:01	-0.350	0.088	6.9×10^{-5}	2.586
Age	0.773	0.050	7.7×10^{-55}	50.296
Sex	-0.448	0.096	2.7×10^{-6}	3.885
BMI	0.278	0.045	4.5×10^{-10}	7.421
Virus genotype	1.527	0.107	3.8×10^{-46}	41.929
HCC				
Cirrhosis	2.567	0.230	5.0×10^{-29}	24.852
Age	0.933	0.080	3.3×10^{-31}	26.886
Sex	-0.794	0.150	1.3×10^{-7}	5.075

	Effect size	SE	*p-value*	$\log_{10} BF$
Virus load				
HLA-DQB1*03:01	-0.180	0.032	1.3×10^{-8}	5.973
Cirrhosis	-0.202	0.035	8.5×10^{-9}	6.148
Age	0.111	0.017	6.1×10^{-11}	8.250
Sex	-0.187	0.034	5.2×10^{-8}	5.383

B.3 MalariaGEN

Table B.3: **Results of stepwise conditional analysis: MalariaGEN**

	Estimate	SE	*p-value*	$\log_{10} BF$
West				
rs75862629	0.249	0.061	4.5×10^{-5}	2.649
PC1	0.285	0.026	3.2×10^{-27}	24.064
East				
HLA-C (protein expression)	-0.123	0.029	2.0×10^{-5}	3.021
PC1	-0.348	0.036	1.1×10^{-21}	18.610

B.4 UKBiobank serological panel

B.4.1 Cases-controls

Table B.4: **Results of stepwise conditional analysis: UKBiobank serological panel (cases-controls)**

	Estimate	SE	*p-value*	$\log_{10} BF$
BKV				
Age	-0.297	0.057	1.7×10^{-7}	4.967
C. trachomatis (Definition I)				
Age	-0.237	0.028	1.2×10^{-17}	14.737
Sex	0.639	0.059	2.1×10^{-27}	24.196
CMV				
Age	0.343	0.023	7.8×10^{-51}	47.283

	Estimate	SE	*p-value*	$\log_{10} BF$
EBV				
Sex	0.465	0.097	1.8×10^{-6}	4.058
PC5	0.277	0.061	6.6×10^{-6}	3.539
H. pylori (Definition I)				
Age	0.225	0.034	4.8×10^{-11}	8.355
PC5	0.262	0.036	6.3×10^{-13}	10.152
H. pylori (Definition II)				
Age	0.207	0.025	3.0×10^{-16}	13.383
Sex	-0.228	0.049	4.0×10^{-6}	3.628
PC5	0.278	0.027	2.9×10^{-25}	22.165
HHV-6A				
Sex	-0.206	0.053	1.1×10^{-4}	2.293
HHV-6B				
HLA-DQB1*02:02	0.305	0.069	9.1×10^{-6}	3.326
HHV-7				
Sex	0.650	0.102	1.6×10^{-10}	7.755
HSV-1				
Age	0.208	0.024	3.9×10^{-18}	15.208
PC5	0.226	0.029	2.7×10^{-15}	12.448
JCV				
HLA-DRB1*15:01	-0.634	0.046	4.2×10^{-44}	40.663
Sex	-0.249	0.046	5.0×10^{-8}	5.419
MCV				
HLA-DRB1*01:01	0.296	0.062	1.5×10^{-6}	4.017
HLA-DQB1*03:02	-0.264	0.054	9.2×10^{-7}	4.220
HLA-DQB1*06:02	-0.505	0.047	2.9×10^{-27}	24.127
PC5	-0.104	0.026	6.2×10^{-5}	2.495
T. gondii				
HLA-A*03:01	0.218	0.050	1.2×10^{-5}	3.195
Age	0.213	0.026	2.1×10^{-16}	13.524
VZV				
Sex	-0.402	0.087	4.1×10^{-6}	3.727

B.4.2 MFI

B.4.2.1 Herpesviridae

Table B.5: **Results of stepwise conditional analysis: UKBiobank serological panel (MFI)**

	Estimate	SE	*p-value*	$\log_{10} BF$
HSV-1 (mgG-1)				
HLA-DRB1.03.01	0.0919	0.0184	6.52×10^{-7}	4.184
HLA-DQB1.06.02	-0.0873	0.0189	3.85×10^{-6}	3.443
Age	0.0386	0.0093	3.48×10^{-5}	2.531
Sex	-0.1284	0.0188	8.68×10^{-12}	8.939
HSV-2 (mgG-2)				
Sex	0.1341	0.0245	5.56×10^{-8}	5.506
EBV (EA-D)				
HLA-A.03.01	0.1073	0.0214	5.44×10^{-7}	4.194
HLA-B.44.02	-0.1056	0.0250	2.34×10^{-5}	2.629
HLA-DQB1.02.01	-0.1940	0.0218	6.41×10^{-19}	15.927
HLA-DQB1.03.02	-0.1759	0.0334	1.49×10^{-7}	4.738
HLA-DQA1.01.02	-0.1296	0.0199	8.34×10^{-11}	7.907
HLA-DQA1.03.01	0.1909	0.0268	1.25×10^{-12}	9.701
HLA-DPB1.03.01	0.1162	0.0241	1.43×10^{-6}	3.790
Age	0.0620	0.0104	2.38×10^{-9}	6.483
Sex	0.2458	0.0209	1.20×10^{-31}	28.675
PC5	-0.0904	0.0115	3.73×10^{-15}	12.191
EBV (EBNA-1)				
HLA-DRB4.01.03	-0.1369	0.0213	1.30×10^{-10}	7.717
HLA-DRB1.07.01	-0.1915	0.0222	6.31×10^{-18}	14.939
HLA-DRB1.15.01	0.1956	0.0221	8.99×10^{-19}	15.781
HLA-DQB1.02.01	-0.6739	0.0410	1.43×10^{-59}	57.095
HLA-DQB1.03.01	-0.1468	0.0261	1.89×10^{-8}	5.608
HLA-DQA1.05.01	0.3855	0.0337	3.96×10^{-30}	27.145
HLA-DPB1.03.01	-0.1682	0.0232	5.00×10^{-13}	10.091
DP1	0.1506	0.0217	4.18×10^{-12}	9.183
DRB3	0.1156	0.0270	1.93×10^{-5}	2.709
heter_DQB1	0.1575	0.0325	1.24×10^{-6}	3.851
EBV (VCA p18)				
TAP2.TVACL	0.1620	0.0285	1.29×10^{-8}	5.768
HLA-C.07.02	0.1397	0.0231	1.44×10^{-9}	6.697
HLA-DRB1.01.01	0.1071	0.0248	1.57×10^{-5}	2.794
HLA-DRB1.15.01	0.1384	0.0234	3.34×10^{-9}	6.340
HLA-DPB1.03.01	-0.1286	0.0235	4.42×10^{-8}	5.249

	Estimate	SE	*p-value*	$\log_{10} BF$
B44	0.1058	0.0192	3.88×10^{-8}	5.305
Age	0.0623	0.0101	6.87×10^{-10}	7.010
EBV (ZEBRA)				
HLA-A.03.01	0.1167	0.0204	1.04×10^{-8}	5.861
HLA-DQB1.03.02	0.3085	0.0231	2.66×10^{-40}	37.421
HLA-DQA1.02.01	0.3081	0.0208	7.34×10^{-49}	46.127
HLA-DPA1.01.03	0.1005	0.0187	7.35×10^{-8}	5.035
B44	0.0946	0.0186	3.59×10^{-7}	4.369
Age	0.0568	0.0100	1.26×10^{-8}	5.778
Sex	0.2850	0.0201	4.02×10^{-45}	42.318
CMV (pp28)				
Sex	0.1460	0.0203	7.79×10^{-13}	10.015
CMV (pp52)				
HLA-C.07.01	-0.0742	0.0176	2.46×10^{-5}	2.716
HLA-DQB1.05.01	-0.1018	0.0208	1.02×10^{-6}	4.038
Sex	0.1860	0.0193	1.01×10^{-21}	18.871
CMV (pp150)				
Sex	0.1864	0.0195	1.89×10^{-21}	18.597
HHV-7 (U14)				
HLA-C.03.04	-0.1268	0.0262	1.36×10^{-6}	3.812
HLA-DRB1.01.01	-0.1011	0.0246	4.07×10^{-5}	2.401
HLA-DRB1.04.01	0.1104	0.0223	7.29×10^{-7}	4.071
HLA-DPB1.04.02	-0.0961	0.0230	2.98×10^{-5}	2.529
Sex	0.3000	0.0200	5.39×10^{-50}	47.282
VZV (gE & gI)				
HLA-A.01.01	0.0886	0.0194	5.28×10^{-6}	3.251
HLA-DQB1.02.01	0.1572	0.0212	1.38×10^{-13}	10.648
Age	0.0610	0.0098	5.44×10^{-10}	7.114
Sex	-0.1812	0.0197	4.59×10^{-20}	17.072

B.4.2.2 Polyomaviridae

Table B.6: Polyomaviridae conditional results

	Estimate	SE	*p-value*	$\log_{10} BF$
BK (VP1)				
Age	-0.1168	0.0102	3.64×10^{-30}	27.177
JC (VP1)				

	Estimate	SE	*p-value*	$\log_{10} BF$
HLA-DRB1.15.01	-0.1858	0.0206	2.60×10^{-19}	16.452
Age	0.0463	0.0092	5.30×10^{-7}	4.310
Sex	-0.0933	0.0185	4.43×10^{-7}	4.385
MC (VP1)				
HLA-DQB1.03.01	0.0913	0.0175	1.98×10^{-7}	4.694
HLA-DQB1.05.01	0.1531	0.0199	1.57×10^{-14}	11.658
HLA-DQB1.06.02	-0.1045	0.0204	3.03×10^{-7}	4.515

B.4.2.3 Bacteria

Table B.7: Bacteria conditional results

	Estimate	SE	*p-value*	$\log_{10} BF$
BK (VP1)				
Age	-0.1168	0.0102	3.64×10^{-30}	27.177
JC (VP1)				
HLA-DRB1.15.01	-0.1858	0.0206	2.60×10^{-19}	16.452
Age	0.0463	0.0092	5.30×10^{-7}	4.310
Sex	-0.0933	0.0185	4.43×10^{-7}	4.385
MC (VP1)				
HLA-DQB1.03.01	0.0913	0.0175	1.98×10^{-7}	4.694
HLA-DQB1.05.01	0.1531	0.0199	1.57×10^{-14}	11.658
HLA-DQB1.06.02	-0.1045	0.0204	3.03×10^{-7}	4.515

B.4.2.4 Parasite

Table B.8: T. gondii conditional results

	Estimate	SE	*p-value*	$\log_{10} BF$
T. gondii (p22)				
Sex	-0.1428	0.0361	8.00×10^{-5}	2.381
PC12	-0.0529	0.0223	1.77×10^{-2}	0.236
T. gondii (sag1)				
Age	0.0864	0.0155	2.95×10^{-8}	5.676

Appendix C

Regularised Horseshoe Prior RStan Code

```
data {
  int<lower=1> n;
  int<lower=1> p;
  matrix[n, p] X;
  int<lower=0, upper=1> y[n];

  // Hyperparameters
  real<lower=0> scale;
  real<lower=0> tau;
  real<lower=0> scale_icept;
  real<lower=0> slab_scale;
  real<lower=0> slab_df;
}

parameters {
  vector[p] z;
  vector<lower=0>[p] lambda;
  real<lower=0> caux;
  real beta0;                    // Intercept
}

transformed parameters {
  vector[p] lambda_tilde;
  vector[p] beta;
  vector[n] theta;
  real c2 = slab_scale^2;
  real caux2 = square(caux);

  for (j in 1:p) {
```

```
    lambda_tilde[j] = sqrt(c2 * lambda[j]^2 /
                           (c2 + tau^2 * lambda[j]^2));
  }

  beta = z .* lambda_tilde * tau;

  theta = X * beta + beta0;
}

model {
  lambda ~ cauchy(0, scale);
  z ~ normal(0, 1);
  caux ~ inv_gamma(0.5 * slab_df, 0.5 * slab_df);
  beta0 ~ normal(0, scale_icept);     // Intercept prior

  y ~ bernoulli_logit(theta);
}
```

Appendix D

Analytical Derivation of Posterior Covariance

We consider a Bayesian linear regression model with a regularised horseshoe prior.

Likelihood:

$$\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\beta}, \sigma \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n)$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ is the vector of coefficients.

Regularised Horseshoe Prior: The prior is specified hierarchically:

$$\begin{aligned} \beta_j \mid \lambda_j, \tau, c &\sim \mathcal{N}(0, s_j), \quad \text{where} \quad s_j = \sigma^2 \tau^2 \tilde{\lambda}_j^2, \\ \tilde{\lambda}_j^2 &= \frac{c^2 \lambda_j^2}{c^2 + \tau^2 \lambda_j^2}, \\ \lambda_j &\sim \mathrm{C}^+(0, 1), \quad \text{for } j = 1, \ldots, p. \end{aligned}$$

The hyperparameter τ (global shrinkage) and c (slab scale) are treated as fixed constants. The parameter σ is the noise standard deviation.

Posterior Distribution: The posterior distribution for $\boldsymbol{\beta}$, conditional on the other parameters, is proportional to the likelihood times the prior:

$$p(\boldsymbol{\beta} \mid \mathbf{X}, \mathbf{Y}, \boldsymbol{\lambda}, \tau, c, \sigma) \propto p(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\beta}, \sigma) p(\boldsymbol{\beta} \mid \boldsymbol{\lambda}, \tau, c, \sigma).$$

Let $\ell(\boldsymbol{\beta}) = \log p(\boldsymbol{\beta} \mid \mathbf{X}, \mathbf{Y}, \boldsymbol{\lambda}, \tau, c, \sigma)$ denote the log-posterior density. Then, we have:

$$\ell(\boldsymbol{\beta}) = -\frac{1}{2\sigma^2} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) - \frac{1}{2} \boldsymbol{\beta}^\top \mathbf{S}^{-1} \boldsymbol{\beta} + \text{const.}$$

Here, $\mathbf{S}$ is the prior covariance matrix, which is diagonal:

$$\mathbf{S} = \sigma^2 \tau^2 \cdot \mathrm{diag}(\tilde{\lambda}_1^2, \tilde{\lambda}_2^2, \ldots, \tilde{\lambda}_p^2).$$

The gradient of the log-posterior (the first derivative) with respect to $\boldsymbol{\beta}$ is:

$$\nabla \ell(\boldsymbol{\beta}) = \frac{1}{\sigma^2}\mathbf{X}^\top(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) - \mathbf{S}^{-1}\boldsymbol{\beta}$$

Then, we could get the Hessian matrix $\mathbf{H}$, which is the matrix of second partial derivatives, $\mathbf{H} = \nabla\nabla^\top \ell(\boldsymbol{\beta})$. Differentiating the gradient yields:

$$\mathbf{H} = -\frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X} - \mathbf{S}^{-1}$$

The observed Fisher information matrix is defined as the negative Hessian:

$$\mathcal{I}(\boldsymbol{\beta}) = -\mathbf{H} = \frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X} + \mathbf{S}^{-1}$$

Under a Laplace approximation, the posterior distribution of $\boldsymbol{\beta}$ is approximated by a Gaussian centred at the mode $\boldsymbol{\beta}^*$ with covariance equal to the inverse of the observed Fisher information:

$$p(\boldsymbol{\beta} \mid \mathbf{X}, \mathbf{Y}, \boldsymbol{\lambda}, \tau, c, \sigma) \approx \mathcal{N}(\boldsymbol{\beta}^*, \Sigma)$$

where the covariance matrix Σ is:

$$\Sigma = (-\mathbf{H})^{-1} = \left(\frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X} + \mathbf{S}^{-1}\right)^{-1}$$

This provides a closed-form expression for the uncertainty associated with the MAP estimate $\boldsymbol{\beta}^*$, conditioned on the values of $\boldsymbol{\lambda}, \tau, c, \sigma$.

Appendix E

Details in Chapter2 figures

E.1 HCV spontaneous clearance vs. chronic infection.

Table E.1: Expression and tapasin-related summary statistics by population.

	Features	**Mean**	**SE**	**Min**	**Max**
Africa	Expression.HLA-A	−0.22	0.03	−1.99	1.22
	Expression.HLA-C	339.03	3.07	228.00	533.00
	Tapsine.HLA-A	0.89	0.01	0.37	2.05
	Tapsine.HLA-B	1.26	0.03	0.43	3.04
	Tapsine.HLA-C	0.96	0.01	0.42	1.74
	global tapasin	3.18	0.03	1.66	5.21
European	Expression.HLA-A	−0.04	0.02	−1.74	1.78
	Expression.HLA-C	326.83	1.53	228.00	588.00
	Tapsine.HLA-A	1.05	0.01	0.42	2.11
	Tapsine.HLA-B	1.48	0.01	0.47	3.06
	Tapsine.HLA-C	0.89	0.01	0.48	1.74
	global tapasin	3.45	0.02	1.73	6.08

Table E.2: Allotype frequencies by population (%).

Allotype	**Africa(%)**	**European(%)**
PSMB8.QG	86.51	85.08
PSMB8.KG	11.10	12.76
PSMB9.RV	75.96	68.26
PSMB9.HV	20.83	29.52
TAP1.IDAVRVG	66.51	83.00
TAP1.VGAVRVG	7.25	9.15
TAP1.VGAVQLG	14.13	2.83
TAP2.AVARLRA	12.39	23.66
TAP2.TVARLRA	41.38	45.71
TAP2.TIARLRA	2.02	8.47
TAP2.TVACLRA	1.65	7.46
TAP2.TVARFRA	10.46	4.62
TAP2.TITRLRA	4.22	6.00
ERAP1.TEPIGMRDRE	6.51	8.35
ERAP1.TEPMGMRDRE	17.71	24.04
ERAP1.TEPIGVRNQE	7.25	18.91
ERAP1.TERIGMKDRE	11.83	8.96
ERAP1.IEPIGMKDRQ	2.20	12.98
ERAP1.TERIGMKDRQ	24.77	14.58
ERAP1.TERIDMRDRE	5.60	7.14
ERAP2.TNL	55.69	49.37
ERAP2.TKL	42.57	40.61
ERAP2.TKQ	1.38	6.92

Table E.3: Heterozygosity by population.

Locus	Africa	European
A	0.97	0.86
B	0.97	0.94
C	0.93	0.91
E	0.53	0.52
F	0.34	0.30
G	0.63	0.37
DRA	0.50	0.44
DRB3	0.68	0.55
DRB5	0.37	0.38
DRB4	0.16	0.27
DRB1	0.94	0.92
DQA1	0.86	0.86
DQB1	0.93	0.91
DOB	0.44	0.33
DMB	0.30	0.38
DMA	0.22	0.28
DOA	0.00	0.02
DPA1	0.69	0.32
DPB1	0.84	0.79

Table E.4: Supertype frequencies by population.

Supertype	Africa	European
A01	0.16	0.23
A01A03	0.01	0.00
A01A24	0.03	0.04
A02	0.22	0.30
A03	0.27	0.22
A24	0.10	0.11
B44	0.16	0.25
B58	0.18	0.06
B27	0.13	0.11
B62	0.04	0.06
B07	0.32	0.27
B08	0.04	0.09
DR4	0.04	0.07
DP2	0.17	0.31
DP1	0.20	0.16
MAIN_DR	0.30	0.31
DRB3	0.25	0.26
DQ7	0.19	0.21
MAINDQ	0.21	0.23

E.2 STOPHCV

Table E.5: Expression and tapasin-related summary statistics by population.

	Features	**Mean**	**SE**	**Min**	**Max**
European	Expression.HLA-A	−0.03	0.01	−1.93	1.78
	Expression.HLA-C	313.73	1.28	228.00	548.00
	Tapsine.HLA-A	1.11	0.01	0.37	2.11
	Tapsine.HLA-B	1.56	0.01	0.50	3.04
	Tapsine.HLA-C	0.84	0.01	0.48	1.74
	global tapasin	3.50	0.02	1.73	5.75
South Asian	Expression.HLA-A	0.02	0.05	−1.94	1.78
	Expression.HLA-C	347.49	4.89	228.00	588.00
	Tapsine.HLA-A	1.12	0.02	0.43	2.11
	Tapsine.HLA-B	1.53	0.04	0.53	3.04
	Tapsine.HLA-C	0.93	0.02	0.48	1.74
	global tapasin	3.58	0.04	2.14	5.64

Table E.6: Allotype frequencies by population (%).

Allotype	**European(%)**	**South Asian(%)**
PSMB8.QG	84.68	84.81
PSMB8.KG	13.33	12.22
PSMB9.RV	69.85	77.41
PSMB9.HV	26.96	22.04
TAP1.IDAVRV	84.10	73.52
TAP1.VGAVRV	8.04	17.04
TAP2.TVARL	50.58	44.63
TAP2.AVARL	24.58	38.52
TAP2.TVACL	6.88	6.11
TAP2.TIARL	4.72	6.30
TAP2.TITRL	6.54	2.04
ERAP1.TEPIGVRNQE	21.84	8.89
ERAP1.TEPMGMRDRE	22.59	25.00
ERAP1.TERIDMRDRE	7.33	10.37
ERAP1.IEPIGMKDRQ	12.65	10.56
ERAP1.TEPIGMRDRE	8.21	5.19
ERAP1.TERIGMKDRQ	13.24	22.04
ERAP1.TERIGMKDRE	6.62	9.07
ERAP1.TKPIGMRDRE	5.27	5.00
ERAP2.PTKL	44.08	29.81
ERAP2.PTNL	47.13	59.63
ERAP2.PTKQ	5.53	5.00

Table E.7: Heterozygosity by population.

Locus	**European**	**South Asian**
A	0.86	0.86
B	0.92	0.86
C	0.90	0.83
DRB1	0.89	0.86
DQA1	0.82	0.78
DQB1	0.84	0.82
DPA1	0.33	0.47
DPB1	0.78	0.73

Table E.8: Supertype frequencies by population.

Supertype	**European**	**South Asian**
A01	0.27	0.29
A01A03	0.01	0.02
A01A24	0.04	0.00
A02	0.28	0.09
A03	0.26	0.37
A24	0.09	0.12
B44	0.27	0.20
B58	0.05	0.10
B27	0.10	0.04
B62	0.06	0.10
B07	0.28	0.30
B08	0.14	0.17
DR4	0.08	0.02
DP2	0.30	0.29
DP1	0.17	0.07
MAIN_DR	0.33	0.26
DRB3	0.16	0.18
DQ7	0.22	0.10
MAINDQ	0.32	0.29

E.3 MalariaGEN

Table E.9: Expression and tapasin-related summary statistics by population.

Population	Features	Mean	SE	Min	Max
BurkinaFaso	Expression.HLA-A	−0.15	0.02	−2.17	1.46
	Expression.HLA-C	335.18	1.77	228.00	533.00
	Tapsine.HLA-A	0.91	0.01	0.37	2.10
	Tapsine.HLA-B	1.05	0.01	0.47	2.52
	Tapsine.HLA-C	1.04	0.01	0.48	1.74
	global tapasin	3.00	0.02	1.56	5.25
Cameroon	Expression.HLA-A	−0.04	0.02	−2.17	1.46
	Expression.HLA-C	336.91	2.09	228.00	520.00
	Tapsine.HLA-A	0.87	0.01	0.37	2.06
	Tapsine.HLA-B	1.38	0.02	0.43	3.04
	Tapsine.HLA-C	0.98	0.01	0.42	1.74
	global tapasin	3.23	0.02	1.54	5.86
Gambia	Expression.HLA-A	−0.15	0.01	−2.17	1.78
	Expression.HLA-C	329.60	0.95	228.00	588.00
	Tapsine.HLA-A	0.97	0.01	0.37	2.15
	Tapsine.HLA-B	1.22	0.01	0.47	3.04
	Tapsine.HLA-C	0.97	0.00	0.42	1.74
	global tapasin	3.16	0.01	1.46	5.57
Ghana	Expression.HLA-A	−0.13	0.03	−2.17	1.22
	Expression.HLA-C	333.49	2.51	228.00	495.00
	Tapsine.HLA-A	0.90	0.01	0.37	2.11
	Tapsine.HLA-B	1.10	0.02	0.47	2.57
	Tapsine.HLA-C	1.03	0.01	0.48	1.74
	global tapasin	3.03	0.03	1.70	5.59
Kenya	Expression.HLA-A	0.06	0.01	−2.17	1.46
	Expression.HLA-C	331.88	1.31	228.00	520.00
	Tapsine.HLA-A	0.92	0.01	0.37	2.11
	Tapsine.HLA-B	1.33	0.01	0.43	3.04
	Tapsine.HLA-C	0.94	0.01	0.48	1.74
	global tapasin	3.20	0.01	1.59	6.05
Malawi	Expression.HLA-A	0.08	0.01	−2.17	1.22
	Expression.HLA-C	329.97	1.45	228.00	588.00
	Tapsine.HLA-A	0.90	0.01	0.37	2.05
	Tapsine.HLA-B	1.34	0.01	0.43	3.04
	Tapsine.HLA-C	0.93	0.01	0.42	1.74
	global tapasin	3.17	0.02	1.60	6.05

Population	Features	Mean	SE	Min	Max
Mali	Expression.HLA-A	−0.13	0.04	−2.17	1.16
	Expression.HLA-C	344.68	3.16	228.00	588.00
	Tapsine.HLA-A	0.92	0.02	0.37	2.07
	Tapsine.HLA-B	1.09	0.03	0.47	2.57
	Tapsine.HLA-C	1.06	0.01	0.48	1.74
	global tapasin	3.07	0.03	1.79	5.76
Nigeria	Expression.HLA-A	−0.17	0.06	−2.17	0.89
	Expression.HLA-C	329.19	6.00	228.00	533.00
	Tapsine.HLA-A	0.93	0.03	0.42	2.05
	Tapsine.HLA-B	1.12	0.05	0.47	3.04
	Tapsine.HLA-C	1.00	0.03	0.42	1.74
	global tapasin	3.05	0.07	1.82	5.70
PNG	Expression.HLA-A	0.21	0.09	−0.74	1.78
	Expression.HLA-C	390.91	17.82	228.00	508.00
	Tapsine.HLA-A	1.13	0.02	0.88	1.20
	Tapsine.HLA-B	1.15	0.07	0.56	2.46
	Tapsine.HLA-C	0.96	0.04	0.60	1.32
	global tapasin	3.23	0.09	2.44	4.72
Tanzania	Expression.HLA-A	0.07	0.03	−2.17	1.24
	Expression.HLA-C	331.00	2.67	228.00	520.00
	Tapsine.HLA-A	0.94	0.01	0.37	2.11
	Tapsine.HLA-B	1.28	0.02	0.43	3.04
	Tapsine.HLA-C	0.94	0.01	0.47	1.74
	global tapasin	3.16	0.03	1.77	5.60
Vietnam	Expression.HLA-A	−0.33	0.08	−1.93	1.78
	Expression.HLA-C	342.18	8.01	228.00	548.00
	Tapsine.HLA-A	1.06	0.02	0.64	1.59
	Tapsine.HLA-B	1.65	0.05	0.61	2.52
	Tapsine.HLA-C	0.83	0.02	0.48	1.74
	global tapasin	3.53	0.06	2.08	5.02

Table E.10: Supertype frequencies by population.

Supertype	BurkinaFaso	Cameroon	Gambia	Ghana	Kenya	Malawi	Mali	Nigeria	PNG	Tanzania	Vietnam
A01	0.13	0.17	0.21	0.12	0.22	0.23	0.16	0.15	0.01	0.22	0.09
A01A03	0.11	0.04	0.06	0.09	0.07	0.10	0.08	0.12	0.00	0.10	0.06
A01A24	0.02	0.09	0.02	0.01	0.06	0.06	0.03	0.01	0.01	0.06	0.00
A02	0.23	0.22	0.19	0.20	0.25	0.22	0.19	0.22	0.00	0.25	0.09
A03	0.27	0.26	0.26	0.30	0.22	0.21	0.28	0.32	0.81	0.20	0.53
A24	0.18	0.14	0.16	0.21	0.12	0.13	0.14	0.09	0.14	0.11	0.13
B44	0.12	0.13	0.10	0.11	0.15	0.17	0.11	0.11	0.19	0.17	0.23
B58	0.09	0.19	0.11	0.09	0.16	0.18	0.08	0.13	0.00	0.15	0.23
B27	0.16	0.12	0.16	0.13	0.22	0.21	0.16	0.14	0.17	0.23	0.06
B62	0.05	0.00	0.02	0.06	0.01	0.00	0.03	0.03	0.33	0.01	0.20
B07	0.40	0.35	0.35	0.41	0.28	0.27	0.42	0.37	0.24	0.28	0.17
B08	0.02	0.02	0.10	0.04	0.03	0.04	0.03	0.01	0.00	0.02	0.02
DR4	0.01	0.01	0.04	0.02	0.02	0.03	0.03	0.01	0.01	0.01	0.03
DP2	0.10	0.16	0.13	0.08	0.14	0.15	0.14	0.14	0.36	0.14	0.21
DP1	0.28	0.31	0.25	0.27	0.32	0.30	0.28	0.36	0.10	0.31	0.21
MAIN_DR	0.17	0.18	0.18	0.17	0.30	0.27	0.20	0.18	0.26	0.27	0.33
DRB3	0.13	0.11	0.15	0.13	0.12	0.13	0.14	0.11	0.01	0.13	0.15
DQ7	0.29	0.28	0.25	0.29	0.30	0.27	0.27	0.29	0.26	0.27	0.09
MAINDQ	0.28	0.27	0.31	0.30	0.31	0.33	0.29	0.34	0.14	0.29	0.28

E.4 UKBiobank

Table E.11: Expression and tapasin-related summary statistics in white British.

Features	**Mean**	**SE**	**Min**	**Max**
Expression.HLA-A	-0.04	0.01	−1.99	1.77
Expression.HLA-C	311.74	0.68	228.00	588.00
Tapsine.HLA-A	1.09	0.00	0.37	2.11
Tapsine.HLA-B	1.58	0.01	0.47	3.06
Tapsine.HLA-C	0.83	0.00	0.48	1.74
Global tapasin	3.52	0.01	1.70	5.98

Table E.12: Supertype frequencies British.

Supertype	**British**
A01	0.25
A01A03	0.01
A01A24	0.04
A02	0.28
A03	0.25
A24	0.09
B44	0.26
B58	0.05
B27	0.10
B62	0.07
B07	0.25
B08	0.14
DR4	0.10
DP2	0.29
DP1	0.16
MAIN_DR	0.32
DRB3	0.18
DQ7	0.24
MAINDQ	0.29

References

(2011). Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature*, 476(7359):214–219.

Amiel, J. F., Curtoni, E. S., Mattiuz, P. L., and Tosi, M. R. (1967). Study of the leucocyte phenotypes in hodgkin's disease. *Histocompatibility Testing. Copenhagen: Munksgaard*, 79.

Andres, A. M., Dennis, M. Y., Kretzschmar, W. W., Cannons, J. L., Lee-Lin, S.-Q., Hurles, M. E., Clark, A. G., Green, E. D., et al. (2010). Balancing selection maintains a form of erap2 that undergoes nonsense-mediated decay and affects antigen presentation. *American Journal of Human Genetics*, 86(6):844–857.

Apps, R., Qi, Y., Carlson, J. M., Chen, H., Gao, X., Thomas, R., Yuki, Y., Del Prete, G. Q., Goulder, P., Brumme, Z. L., et al. (2013). Influence of HLA-C expression level on hiv control. *Science*, 340(6128):87–91.

Arora, J., Pierini, F., McLaren, P. J., Carrington, M., Fellay, J., and Lenz, T. L. (2020). HLA heterozygote advantage against hiv-1 is driven by quantitative and qualitative differences in HLA allele-specific peptide presentation. *Molecular Biology and Evolution*, 37(3):639–650.

Band, G., Le, Q. S., Clarke, G. M., and Kivinen, K. (2019). Insights into malaria susceptibility using genome-wide data on 17,000 individuals from africa, asia and oceania. *Nature communications*, 10(1):5732.

Bashirova, A. A., Viard, M., Naranbhai, V., Grifoni, A., Garcia-Beltran, W., Akdag, M., Yuki, Y., Gao, X., O'hUigin, C., Raghavan, M., et al. (2020). HLA tapasin independence: broader peptide repertoire and hiv control. *Proceedings of the National Academy of Sciences*, 117(45):28232–28238.

Benner, C., Spencer, C. C., Havulinna, A. S., Salomaa, V., Ripatti, S., and Pirinen, M. (2016). Finemap: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics*, 32(10):1493–1501.

Bjorkman, P., Saper, M., Samraoui, B., Bennett, W., Strominger, J., and Wiley, D. (1987). The foreign antigen binding site and t cell recognition regions of class i histocompatibility antigens. *Nature*, 329(6139):512–518.

Blum, J. S., Wearsch, P. A., and Cresswell, P. (2013). Pathways of antigen processing. *Annual Review of Immunology*, 31(1):443–473.

Butler-Laporte, G., Farjoun, J., Nakanishi, T., Lu, T., Abner, E., Chen, Y., Hultström, M., Metspalu, A., Milani, L., Mägi, R., et al. (2023). HLA allele-calling using multi-ancestry whole-exome sequencing from the uk biobank identifies 129 novel associations in 11 autoimmune diseases. *Communications Biology*, 6(1):1113.

Butler-Laporte, G., Kreuzer, D., Nakanishi, T., Harroud, A., Forgetta, V., and Richards, J. B. (2020). Genetic determinants of antibody-mediated immune responses to infectious diseases agents: a genome-wide and HLA association study. In *Open Forum Infectious Diseases*, volume 7, page ofaa450. Oxford University Press US.

Byrne, C. D. and Wild, S. (2010). Body fat and increased risk of cirrhosis.

Cao, B., Tian, X., Li, Y., Jiang, P., Ning, T., Xing, H., Zhao, Y., Zhang, C., Shi, X., Chen, D., et al. (2005). Lmp7/tap2 gene polymorphisms and hpv infection in esophageal carcinoma patients from a high incidence area in china. *Carcinogenesis*, 26(7):1280–1284.

Carosella, E. D., Rouas-Freiss, N., Tronik-Le Roux, D., Moreau, P., and LeMaoult, J. (2015). Hla-g: an immune checkpoint molecule. *Advances in immunology*, 127:33–144.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009). Handling sparsity via the horseshoe. In *Artificial Intelligence and Statistics*, pages 73–80. PMLR.

Castro-Santos, P., Moro-García, M. A., Marcos-Fernández, R., Alonso-Arias, R., and Díaz-Peña, R. (2017). Erap1 and hla-c interaction in inflammatory bowel disease in the spanish population. *Innate immunity*, 23(5):476–481.

Cifaldi, L., Romania, P., Lorenzi, S., Locatelli, F., and Fruci, D. (2012). Role of endoplasmic reticulum aminopeptidases in health and disease: from infection to cancer. *International Journal of Molecular Sciences*, 13(7):8338–8352.

Colbert, J. D., Cruz, F. M., and Rock, K. L. (2020). Cross-presentation of exogenous antigens on mhc i molecules. *Current opinion in immunology*, 64:1–8.

Correa Vanegas, P. A., Molina Restrepo, J. F., Pinto, L., Arcos Burgos, O. M., Herrera, M., and Anaya Cabrera, J. M. (2003). TAP1 and tap2 polymorphisms analysis in northwestern colombian patients with systemic lupus erythematosus.

Cortes, A., Pulit, S. L., Leo, P. J., Pointon, J. J., Robinson, P. C., Weisman, M. H., Ward, M., Gensler, L. S., Zhou, X., Garchon, H.-J., et al. (2015). Major histocompatibility complex associations of ankylosing spondylitis are complex and involve further epistasis with erap1. *Nature Communications*, 6(1):7146.

Dai, D., Chen, Y., Ru, P., Zhou, X., Tao, J., Ye, H., Hong, Q., Tang, L., Pan, G., Lin, D., et al. (2014). Significant association between tap2 polymorphisms and rheumatoid arthritis: a meta-analysis. *Diagnostic Pathology*, 9(1):129.

Das, A., Chandra, A., Chakraborty, J., Chattopadhyay, A., Senapati, S., Chatterjee, G., and Chatterjee, R. (2017). Associations of erap1 coding variants and domain specific interaction with hla-c 06 in the early onset psoriasis patients of india. *Human immunology*, 78(11-12):724–730.

Dendrou, C. A., Petersen, J., Rossjohn, J., and Fugger, L. (2018). Hla variation and disease. *Nature Reviews Immunology*, 18(5):325–339.

d'Etude Génétique des Spondylarthrites (GFEGS), G. F., (HUNT), N.-T. H. S., of Canada (SPARCC), S. R. C., (WTCCC2), W. T. C. C. C. ., Bowness, P., Gafney, K., Gaston, H., Gladman, D. D., Rahman, P., et al. (2013). Identification of multiple risk variants for ankylosing spondylitis through high-density genotyping of immune-related loci. *Nature genetics*, 45(7):730–738.

Driscoll, J. (1994). The role of the proteasome in cellular protein degradation. *Histology and Histopathology*.

Dubois, P. C., Trynka, G., Franke, L., Hunt, K. A., Romanos, J., Curtotti, A., Zhernakova, A., Heap, G. A., Ádány, R., Aromaa, A., et al. (2010). Multiple common variants for celiac disease influencing immune gene expression. *Nature genetics*, 42(4):295–302.

Duggal, P., Thio, C. L., Wojcik, G. L., Goedert, J. J., Mangia, A., Latanich, R., Kim, A. Y., Lauer, G. M., Chung, R. T., Peters, M. G., et al. (2013). Genome-wide association study of spontaneous resolution of hepatitis c virus infection: data from multiple cohorts. *Annals of internal medicine*, 158(4):235–245.

Einstein, M. H., Leanza, S., Chiu, L. G., Schlecht, N. F., Goldberg, G. L., Steinberg, B. M., and Burk, R. D. (2009). Genetic variants in TAP are associated with high-grade cervical neoplasia. *Clinical Cancer Research*, 15(3):1019–1023.

Elhawary, N. A., Ekram, S. N., Abumansour, I. S., Azher, Z. A., AlJahdali, I. A., Alyamani, N. M., Naffadi, H. M., Sindi, I. A., Baazeem, A., Nassir, A. M., et al. (2023). Sequence variants in psmb8/psmb9 immunoproteasome genes and risk of urothelial bladder carcinoma. *Cureus*, 15(3).

Evnouchidou, I., Birtley, J., Seregin, S., Papakyriakou, A., Zervoudi, E., Samiotaki, M., Panayotou, G., Giastas, P., Petrakis, O., Georgiadis, D., et al. (2012). A common SNP in ER aminopeptidase 2 induces a specificity switch that leads to altered antigen processing. *Journal of Immunology (Baltimore, Md.: 1950)*, 189(5):2383.

Evseeva, I., Nicodemus, K. K., Bonilla, C., Tonks, S., and Bodmer, W. F. (2010). Linkage disequilibrium and age of HLA region snps in relation to classic HLA gene alleles within europe. *European Journal of Human Genetics*, 18(8):924–932.

Ferreira, L. C., Gomes, C. E. M., Duggal, P., De Paula Holanda, I., de Lima, A. S., do Nascimento, P. R. P., and Jeronimo, S. M. B. (2021). Genetic association of erap1 and erap2 with eclampsia and preeclampsia in northeastern brazilian women. *Scientific reports*, 11(1):6764.

Garbi, N., Tiwari, N., Momburg, F., and Hämmerling, G. J. (2003). A major role for tapasin as a stabilizer of the tap peptide transporter and consequences for mhc class i expression. *European Journal of Immunology*, 33(1):264–273.

George, E. I. and McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

Gerber, F. and Furrer, R. (2019). optimparallel: An r package providing a parallel version of the l-bfgs-b optimization method.

Gonzalez-Galarza, F. F., McCabe, A., Santos, E. J. M., Jones, J., Takeshita, L., Ortega-Rivera, N. D., Del Cid-Pavon, G., Ramsbottom, K., Ghattaoraya, G., Alfirevic, A., and Middleton, D. (2020). Allele frequency net database (afnd) 2020 update: gold-standard data classification, open access genotype data and new query tools. *Nucleic Acids Research*, 48(D1):D783–D788.

Goulder, P. J. and Walker, B. D. (2012). Hiv and HLA class i: an evolving relationship. *Immunity*, 37(3):426–440.

Greenbaum, J., Sidney, J., Chung, J., Brander, C., Peters, B., and Sette, A. (2011). Functional classification of class ii human leukocyte antigen (HLA) molecules reveals seven different supertypes and a surprising degree of repertoire sharing across supertypes. *Immunogenetics*, 63(6):325–335.

Griffin, T. A., Nandi, D., Cruz, M., Fehling, H. J., Kaer, L. V., Monaco, J. J., and Colbert, R. A. (1998). Immunoproteasome assembly: cooperative incorporation of interferon γ (ifn-γ)–inducible subunits. *The Journal of Experimental Medicine*, 187(1):97–104.

Hamilton, F., Mentzer, A. J., Parks, T., Baillie, J. K., Smith, G. D., Ghazal, P., and Timpson, N. J. (2023). Variation in erap2 has opposing effects on severe respiratory infection and autoimmune disease. *The American Journal of Human Genetics*, 110(4):691–702.

Hammer, C., Begemann, M., McLaren, P. J., Bartha, I., Michel, A., Klose, B., Schmitt, C., Waterboer, T., Pawlita, M., Schulz, T. F., et al. (2015). Amino acid variation in hla class ii proteins is a major determinant of humoral response to common viruses. *The American Journal of Human Genetics*, 97(5):738–743.

Hanson, A. L., Cuddihy, T., Haynes, K., Loo, D., Morton, C. J., Oppermann, U., Leo, P., Thomas, G. P., Lê Cao, K.-A., Kenna, T. J., et al. (2018). Genetic variants in erap 1 and erap 2 associated with immune-mediated diseases influence protein expression and the isoform profile. *Arthritis & rheumatology*, 70(2):255–265.

Hill, A. V., Allsopp, C. E., Kwiatkowski, D., Anstey, N. M., Twumasi, P., Rowe, P. A., Bennett, S., Brewster, D., McMichael, A. J., and Greenwood, B. M. (1991). Common west african hla antigens are associated with protection from severe malaria. *Nature*, 352(6336):595–600.

Hill, A. V., Elvin, J., Willis, A. C., Aidoo, M., Allsopp, C. E., Gotch, F. M., Ming Gao, X., Takiguchis, M., Greenwood, B. M., Townsend, A. R., et al. (1992). Molecular analysis of the association of HLA-b53 and resistance to severe malaria. *Nature*, 360(6403):434–439.

Horton, R., Wilming, L., Rand, V., Lovering, R. C., Bruford, E. A., Khodiyar, V. K., Lush, M. J., Povey, S., Talbot Jr, C. C., Wright, M. W., et al. (2004). Gene map of the extended human mhc. *Nature Reviews Genetics*, 5(12):889–899.

Houldcroft, C. J. and Kellam, P. (2015). Host genetics of epstein–barr virus infection, latency and disease. *Reviews in medical virology*, 25(2):71–84.

Hu, J., Wang, S., Zhang, X., Yan, W., Liu, H., Chen, X., Nie, Y., Liu, F., Zheng, Y., Lu, Y., et al. (2024). A genetic variant in the tapbp gene enhances cervical cancer susceptibility by increasing m6a modification. *Archives of Toxicology*, 98(10):3425–3438.

Huang, P., Dong, L., Lu, X., Zhang, Y., Chen, H., Wang, J., Zhang, Y., Su, J., and Yu, R. (2014). Genetic variants in antigen presentation-related genes influence susceptibility to hepatitis c virus and viral clearance: a case control study. *BMC Infectious Diseases*, 14(1):716.

Hutchinson, J. P., Temponeras, I., Kuiper, J., Cortes, A., Korczynska, J., Kitchen, S., and Stratikos, E. (2021). Common allotypes of er aminopeptidase 1 have substrate-dependent and highly variable enzymatic properties. *Journal of Biological Chemistry*, 296:100443.

Ishwaran, H. and Rao, J. S. (2005). Spike and slab variable selection: frequentist and bayesian strategies.

Jackson, M. R., Nilsson, T., and Peterson, P. A. (1990). Identification of a consensus motif for retention of transmembrane proteins in the endoplasmic reticulum. *The EMBO Journal*, 9(10):3153–3162.

Jiao, J. and Wang, J. (2003). Effects of hcv genotypes and HLA-drb alleles on the response of chronic hepatitis c patients to interferon alpha and libavilin. *Zhonghua Gan Zang Bing Za Zhi= Zhonghua Ganzangbing Zazhi= Chinese Journal of Hepatology*, 11(10):620–622.

Johnstone, I. M. and Silverman, B. W. (2004). Needles and straw in haystacks: Empirical bayes estimates of possibly sparse sequences. *The Annals of Statistics*, 32(4):1594–1649.

Jones, R., McArdle, W., Ring, S., Strachan, D., Pembrey, M., Clayton, D., Dunger, D., Nutland, S., Stevens, H., Walker, N., Widmer, B., Todd, J., et al. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678.

Jones, R., McArdle, W., Ring, S., Strachan, D., Pembrey, M., Clayton, D., Dunger, D., Nutland, S., Stevens, H., Walker, N., Widmer, B., Todd, J., et al. (2022). Trans-ancestral fine-mapping of mhc reveals key amino acids associated with spontaneous clearance of hepatitis c in HLA-dqβ1. *The American Journal of Human Genetics*, 109(2):299–310.

Kamatani, Y., Wattanapokayakit, S., Ochi, H., Kawaguchi, T., Takahashi, A., Hosono, N., Kubo, M., Tsunoda, T., Kamatani, N., Kumada, H., et al. (2009). A genome-wide association study identifies variants in the hla-dp locus associated with chronic hepatitis b in asians. *Nature genetics*, 41(5):591–595.

Kim, A. Y., Kuntzen, T., Timm, J., Nolan, B. E., Baca, M. A., Reyor, L. L., Berical, A. C., Feller, A. J., Johnson, K. L., Zur Wiesch, J. S., et al. (2011a). Spontaneous control of hcv is associated with expression of hla-b* 57 and preservation of targeted epitopes. *Gastroenterology*, 140(2):686–696.

Kim, H., Lee, H., Lew, B., Sim, W., Kim, Y., Lee, S., Lee, S., Cho, I., Kwon, J., and Kim, H. (2015). Association between TAP1 gene polymorphisms and alopecia areata in a korean population. *Genet Mol Res*, 14(4):18820–7.

Kim, J.-H., Park, B.-L., Pasaje, C. F. A., Bae, J. S., Park, J. S., Park, S. W., Uh, S.-T., Kim, M.-K., Choi, I. S., Cho, S. H., et al. (2011b). Genetic association analysis of tap1 and tap2 polymorphisms with aspirin exacerbated respiratory disease and its fev1 decline. *Journal of Human Genetics*, 56(9):652–659.

Klein, J. and Figueroa, F. (1986). Evolution of the major histocompatibility complex. *Critical Reviews in Immunology*, 6(4):295–386.

Klein, J. and Sato, A. (2000). The HLA system. *New England Journal of Medicine*, 343(10):702–709.

Klunk, J., Vilgalys, T. P., Demeure, C. E., Cheng, X., Shiratori, M., Madej, J., Beau, R., Elli, D., Patino, M. I., Redfern, R., et al. (2022). Evolution of immune genes is associated with the black death. *Nature*, 611(7935):312–319.

Kochan, G., Krojer, T., Harvey, D., Fischer, R., Chen, L., Vollmar, M., von Delft, F., Kavanagh, K. L., Brown, M. A., Bowness, P., et al. (2011). Crystal structures of the endoplasmic reticulum aminopeptidase-1 (erap1) reveal the molecular basis for n-terminal peptide trimming. *Proceedings of the National Academy of Sciences*, 108(19):7745–7750.

Leone, P., Shin, E.-C., Perosa, F., Vacca, A., Dammacco, F., and Racanelli, V. (2013). Mhc class i antigen processing and presenting machinery: organization, function, and defects in tumor cells. *Journal of the National Cancer Institute*, 105(16):1172–1187.

Liu, R., Chen, X., and Qi, J. (2017). Associations of tap1 genetic polymorphisms with atopic diseases: asthma, rhinitis and dermatitis. *Oncotarget*, 9(2):1553.

Liu, W., Ma, Q., Li, C., Li, Y., Liu, S., Shi, L., and Yao, Y. (2021). The association of tap polymorphisms with non-small-cell lung cancer in the han chinese population. *Human Immunology*, 82(12):917–922.

López de Castro, J. A. (2018). How erap1 and erap2 shape the peptidomes of disease-associated mhc-i proteins. *Frontiers in immunology*, 9:2463.

Mahmoudi, M., Aslani, S., Meguro, A., Akhtari, M., Fatahi, Y., Mizuki, N., and Shahram, F. (2022). A comprehensive overview on the genetics of behcet's disease. *International Reviews of Immunology*, 41(2):84–106.

Maiers, M., Gragert, L., and Klitz, W. (2007). High-resolution hla alleles and haplotypes in the united states population. *Human Immunology*, 68(9):779–788.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753.

McKiernan, S. M., Hagan, R., Curry, M., McDonald, G. S., Kelly, A., Nolan, N., Walsh, A., Hegarty, J., Lawlor, E., and Kelleher, D. (2004). Distinct mhc class i and ii alleles are associated with hepatitis c viral clearance, originating from a single source. *Hepatology*, 40(1):108–114.

Mehta, A. M., Osse, M., Kolkman-Uljee, S., Fleuren, G. J., and Jordanova, E. S. (2015). Molecular backgrounds of erap1 downregulation in cervical carcinoma. *Analytical Cellular Pathology*, 2015(1):367837.

Meng, J., Li, W., Zhang, M., Hao, Z., Fan, S., Zhang, L., and Liang, C. (2018). An update meta-analysis and systematic review of tap polymorphisms as potential biomarkers for judging cancer risk. *Pathology-Research and Practice*, 214(10):1556–1563.

Mentzer, A. J., Brenner, N., Allen, N., Littlejohns, T. J., Chong, A. Y., Cortes, A., Almond, R., Hill, M., Sheard, S., McVean, G., et al. (2022). Identification of host–pathogen-disease relationships using a scalable multiplex serology platform in uk biobank. *Nature communications*, 13(1):1818.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032.

Mohd Hanafiah, K., Groeger, J., Flaxman, A. D., and Wiersma, S. T. (2013). Global epidemiology of hepatitis c virus infection: new estimates of age-specific antibody to hcv seroprevalence. *Hepatology*, 57(4):1333–1342.

Nair, S., Mason, A., Eason, J., Loss, G., and Perrillo, R. P. (2002). Is obesity an independent risk factor for hepatocellular carcinoma in cirrhosis? *Hepatology*, 36(1):150–155.

Neefjes, J., Jongsma, M. L., Paul, P., and Bakke, O. (2011). Towards a systems understanding of mhc class i and mhc class ii antigen presentation. *Nature reviews immunology*, 11(12):823–836.

Nejentsev, S., Howson, J. M., Walker, N. M., Szeszko, J., Field, S. F., Stevens, H. E., Reynolds, P., Hardy, M., King, E., Masters, J., et al. (2007). Localization of type 1 diabetes susceptibility to the mhc class i genes hla-b and hla-a. *Nature*, 450(7171):887–892.

Nguyen, H. D., Tran, H. B. T., Nguyen, T. T., Kwak, I. H., Kim, Y. J., Ma, H.-i., Kim, H.-J., and Kim, Y. E. (2025). Pathological role of immunoproteasome psmb8 in parkinson's disease: a link between α-synuclein aggregation and immune activation. *EBioMedicine*, 121.

Nitschke, K., Barriga, A., Schmidt, J., Timm, J., Viazov, S., Kuntzen, T., Kim, A. Y., Lauer, G. M., Allen, T. M., Gaudieri, S., et al. (2014). Hla-b 27 subtype specificity determines targeting and viral evolution of a hepatitis c virus-specific cd8+ t cell epitope. *Journal of hepatology*, 60(1):22–29.

Ogutu, J. O., Schulz-Streeck, T., and Piepho, H.-P. (2012). Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. In *BMC Proceedings*, volume 6, page S10. Springer.

Omran, M. H., Fotouh, B. E., Youssef, S. S., Ibrahim, N. E., Nabil, W., Mahdy, E.-S. M., Shosha, W. G., and El-Awady, M. K. (2013). Association between low molecular polypeptide 7 single nucleotide polymorphism and response to therapy in hepatitis c virus infection. *World Journal of Hepatology*, 5(3):97.

Onengut-Gumuscu, S., Chen, W.-M., Burren, O., Cooper, N. J., Quinlan, A. R., Mychaleckyj, J. C., Farber, E., Bonnie, J. K., Szpak, M., Schofield, E., et al. (2015). Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. *Nature genetics*, 47(4):381–386.

Oppermann, Udo, D. M., Spencer, C. C., Pointon, J. J., Su, Z., Harvey, D., Kochan, G., Oppermann, U., Dilthey, A., et al. (2011). Interaction between erap1 and hla-b27 in ankylosing spondylitis implicates peptide handling in the mechanism for hla-b27 in disease susceptibility. *Nature genetics*, 43(8):761–767.

Ortmann, B., Copeman, J., Lehner, P. J., Sadasivan, B., Herberg, J. A., Grandea, A. G., Riddell, S. R., Tampe, R., Spies, T., Trowsdale, J., et al. (1997). A critical role for tapasin in the assembly and function of multimeric mhc class i-tap complexes. *Science*, 277(5330):1306–1309.

Ou, W.-J., Cameron, P. H., Thomas, D. Y., and Bergeron, J. J. (1993). Association of folding intermediates of glycoproteins with calnexin during protein maturation. *Nature*, 364(6440):771–776.

Paladini, F., Fiorillo, M. T., Tedeschi, V., Cauli, A., Mathieu, A., and Sorrentino, R. (2019). Ankylosing spondylitis: a trade off of HLA-b27, erap, and pathogen interconnections? focus on sardinia. *Frontiers in Immunology*, 10:35.

Paladini, F., Fiorillo, M. T., Vitulano, C., Tedeschi, V., Piga, M., Cauli, A., Mathieu, A., and Sorrentino, R. (2018). An allelic variant in the intergenic region between

erap1 and erap2 correlates with an inverse expression of the two genes. *Scientific Reports*, 8(1):10398.

Piironen, J. and Vehtari, A. (2017). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, 11(2):5018–5051.

Pintea-Trifu, M.-L., Vică, M. L., Bâlici, S.-, Leucuța, D.-C., Coman, H. G., Nemeș, B., Trifu, D.-M., Siserman, C.-V., and Matei, H.-V. (2024). Hla-dr and hla-dq polymorphism correlation with sexually transmitted infection caused by chlamydia trachomatis. *Medicina*, 60(5):808.

Pobre, K. F. R., Poet, G. J., and Hendershot, L. M. (2019). The endoplasmic reticulum (er) chaperone bip is a master regulator of er functions: Getting by with a little help from erdj friends. *Journal of Biological Chemistry*, 294(6):2098–2108.

Pohl, C. and Dikic, I. (2019). Cellular quality control by the ubiquitin-proteasome system and autophagy. *Science*, 366(6467):818–822.

Qian, J., Liu, H., Wei, S., Liu, Z., Li, Y., Wang, L.-E., Chen, W. V., Amos, C. I., Lee, J. E., investigators, G., et al. (2013). Association between putative functional variants in the psmb 9 gene and risk of melanoma–re-analysis of published melanoma genome-wide association studies. *Pigment Cell & Melanoma Research*, 26(3):392–401.

Ramsuran, V., Naranbhai, V., Horowitz, A., Qi, Y., Martin, M. P., Yuki, Y., Gao, X., Walker-Sperling, V., Del Prete, G. Q., Schneider, D. K., et al. (2018). Elevated HLA-a expression impairs hiv control through inhibition of nkg2a-expressing cells. *Science*, 359(6371):86–90.

Raychaudhuri, S., Sandor, C., Stahl, E. A., Freudenberg, J., Lee, H.-S., Jia, X., Alfredsson, L., Padyukov, L., Klareskog, L., Worthington, J., et al. (2012). Five amino acids in three hla proteins explain most of the association between mhc and seropositive rheumatoid arthritis. *Nature genetics*, 44(3):291–296.

Reeves, E., Colebatch-Bourn, A., Elliott, T., Edwards, C. J., and James, E. (2014). Functionally distinct erap1 allotype combinations distinguish individuals with ankylosing spondylitis. *Proceedings of the National Academy of Sciences*, 111(49):17594–17599.

Reeves, E., Edwards, C. J., Elliott, T., and James, E. (2013). Naturally occurring erap1 haplotypes encode functionally distinct alleles with fine substrate specificity. *The Journal of Immunology*, 191(1):35–43.

Reveille, J. D. (2012). The genetic basis of ankylosing spondylitis. *Current Opinion in Rheumatology*, 24(4):332–338.

Robinson, J., Soormally, A. R., Hayhurst, J. D., and Marsh, S. G. (2016). The ipd-imgt/HLA database–new developments in reporting HLA variation. *Human Immunology*, 77(3):233–237.

Robinson, J. H. and Delvig, A. A. (2002). Diversity in mhc class ii antigen presentation. *Immunology*, 105(3):252–262.

Robinson, P. C., Costello, M.-E., Leo, P., Bradbury, L. A., Hollis, K., Cortes, A., Lee, S., Joo, K. B., Shim, S.-C., Weisman, M., et al. (2015). Erap2 is associated with ankylosing spondylitis in hla-b27-positive and hla-b27-negative patients. *Annals of the rheumatic diseases*, 74(8):1627–1629.

Roby, K., Gershon, D., and Hunt, J. (1996). Expression of the transporter for antigen processing-1 (tap-1) gene in subpopulations of human trophoblast cells. *Placenta*, 17(1):27–32.

Santos, S. G., Campbell, E. C., Lynch, S., Wong, V., Antoniou, A. N., and Powis, S. J. (2007). Major histocompatibility complex class i-erp57-tapasin interactions within the peptide-loading complex. *Journal of Biological Chemistry*, 282(24):17587–17593.

Saulle, I., Vicentini, C., Clerici, M., and Biasin, M. (2020). An overview on erap roles in infectious diseases. *Cells*, 9(3):720.

Saveanu, L., Carroll, O., Lindo, V., Del Val, M., Lopez, D., Lepelletier, Y., Greer, F., Schomburg, L., Fruci, D., Niedermann, G., et al. (2005). Concerted peptide trimming by human erap1 and erap2 aminopeptidase complexes in the endoplasmic reticulum. *Nature Immunology*, 6(7):689–697.

Seifert, U., Bialy, L. P., Ebstein, F., Bech-Otschir, D., Voigt, A., Schröter, F., Prozorovski, T., Lange, N., Steffen, J., Rieger, M., et al. (2010). Immunoproteasomes preserve protein homeostasis upon interferon-induced oxidative stress. *Cell*, 142(4):613–624.

Shai, R., Quismorio Jr, F. P., Li, L., Kwon, O.-J., Morrison, J., Wallace, D. J., Neuwelt, C. M., Brautbar, C., Gauderman, W. J., and Jacob, C. O. (1999). Genome-wide screen for systemic lupus erythematosus susceptibility genes in multiplex families. *Human Molecular Genetics*, 8(4):639–644.

Shionoya, Y., Kanaseki, T., Miyamoto, S., Tokita, S., Hongo, A., Kikuchi, Y., Kochin, V., Watanabe, K., Horibe, R., Saijo, H., et al. (2017). Loss of tapasin in human lung and colon cancer cells and escape from tumor-associated antigen-specific ctl recognition. *Oncoimmunology*, 6(2):e1274476.

Sidney, J., Peters, B., Frahm, N., Brander, C., and Sette, A. (2008). HLA class i supertypes: a revised and updated classification. *BMC Immunology*, 9:1.

Singh, R., Kaul, R., Kaul, A., and Khan, K. (2007). A comparative review of HLA associations with hepatitis b and c viral infections across global populations. *World Journal of Gastroenterology: WJG*, 13(12):1770.

Solberg, O. D., Mack, S. J., Lancaster, A. K., Single, R. M., Tsai, Y., Sanchez-Mazas, A., and Thomson, G. (2008). Balancing selection and heterogeneity across the classical human leukocyte antigen loci: a meta-analytic review of population studies. *Tissue Antigens*, 71(5):397–410.

Stamogiannos, A., Koumantou, D., Papakyriakou, A., and Stratikos, E. (2015). Effects of polymorphic variation on the mechanism of endoplasmic reticulum aminopeptidase 1. *Molecular Immunology*, 67(2):426–435.

Study, I. H. C. (2010). The major genetic determinants of hiv-1 control affect hla class i peptide presentation. *Science*, 330(6010):1551–1557.

Tao, Y., Han, X., Liu, N., Shi, L., Shi, L., Liu, S., and Yao, Y. (2022). Association study of tap and hla-i gene combination with chronic hepatitis c virus infection in a han population in china. *International Journal of Immunogenetics*, 49(3):169–180.

Thomas, D. L., Thio, C. L., Martin, M. P., Qi, Y., Ge, D., O'hUigin, C., Kidd, J., Kidd, K., Khakoo, S. I., Alexander, G., et al. (2009). Genetic variation in il28b and spontaneous clearance of hepatitis c virus. *Nature*, 461(7265):798–801.

Thuring, C., Geironson, L., and Paulsson, K. (2014). Tapasin and human leukocyte antigen class i dysregulation correlates with survival in glioblastoma multiforme. *Anti-Cancer Agents in Medicinal Chemistry-Anti-Cancer Agents)*, 14(8):1101–1109.

Thursz, M., Yallop, R., Goldin, R., Trepo, C., and Thomas, H. C. (1999). Influence of mhc class ii genotype on outcome of infection with hepatitis c virus. *The Lancet*, 354(9196):2119–2124.

Thye, T., Vannberg, F. O., Wong, S. H., Owusu-Dabo, E., Osei, I., Gyapong, J., Sirugo, G., Sisay-Joof, F., Enimil, A., Chinbuah, M. A., et al. (2010). Genome-wide association analyses identifies a susceptibility locus for tuberculosis on chromosome 18q11. 2. *Nature genetics*, 42(9):739–741.

Trowsdale, J. and Knight, J. C. (2013). Major histocompatibility complex genomics and human disease. *Annual review of genomics and human genetics*, 14(1):301–323.

Tuteja, R. (2007). Malaria- an overview. *The FEBS journal*, 274(18):4670–4679.

Visscher, P. M., Brown, M. A., McCarthy, M. I., and Yang, J. (2012). Five years of gwas discovery. *The American Journal of Human Genetics*, 90(1):7–24.

Vyas, J. M., Van der Veen, A. G., and Ploegh, H. L. (2008). The known unknowns of antigen processing and presentation. *Nature Reviews Immunology*, 8(8):607–618.

Walker-Sperling, V., Digitale, J. C., Viard, M., Martin, M. P., Bashirova, A., Yuki, Y., Ramsuran, V., Kulkarni, S., Naranbhai, V., Li, H., et al. (2022). Genetic variation that determines tapbp expression levels associates with the course of malaria in an hla allotype-dependent manner. *Proceedings of the National Academy of Sciences*, 119(29):e2205498119.

Woon, A. P. and Purcell, A. W. (2018). The use of proteomics to understand antiviral immunity. In *Seminars in Cell & Developmental Biology*, volume 84, pages 22–29. Elsevier.

Yao, Y., Liu, N., Zhou, Z., and Shi, L. (2019). Influence of ERAP1 and ERAP2 gene polymorphisms on disease susceptibility in different populations. *Human Immunology*, 80(5):325–334.

Ye, C. J., Chen, J., Villani, A.-C., Gate, R. E., Subramaniam, M., Bhangale, T., Lee, M. N., Raj, T., Raychowdhury, R., Li, W., et al. (2018). Genetic analysis of isoform usage in the human anti-viral response reveals influenza-specific regulation of ERAP2 transcripts under balancing selection. *Genome Research*, 28(12):1812–1825.

Zang, F., Yao, Y., Liu, M., Fan, H., Yue, M., Chen, M., Wang, J., Yu, R., and Huang, P. (2017). The association of lmp7 and tap2 gene polymorphisms with treatment response to interferon/ribavirin in patients with genotype 1 chronic hepatitis c. *International Journal of Molecular Medicine*, 40(6):1983–1990.

Zhang, M., Wang, X., Zhu, Y., Chen, S., Chen, B., and Liu, Z. (2021). Associations of genetic variants at tap1 and tap2 with pulmonary tuberculosis risk among the chinese population. *Epidemiology & Infection*, 149:e79.

Zou, Y., Carbonetto, P., Wang, G., and Stephens, M. (2022). Fine-mapping from summary data with the "sum of single effects" model. *PLoS Genetics*, 18(7):e1010299.