

**Identification and Characterisation of the
Genetic Determinants of Variable Response to
Antigens from Infectious Agents**



Alexander John Mentzer

Keble College

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2016/17

Abstract

Despite the success of vaccines in routine use worldwide, there are substantial challenges hampering our ability to develop vaccines against extant diseases including malaria and tuberculosis. Novel approaches are urgently required to help us understand immunological correlates of protection against disease and facilitate our understanding of the impact of human genetic variation on the success of diverse vaccines.

To identify host genetic factors responsible for variation in antibody responses against vaccine antigens delivered routinely to infants worldwide I performed a genome-wide association study (GWAS) involving 2,499 infants recruited from three diverse sites across Africa. I identified strong genetic associations between variants in the class II major histocompatibility complex (MHC) locus and responses against five antigens: pertussis toxin (PT), filamentous haemagglutinin (FHA) and pertactin; diphtheria toxin (DT); and hepatitis B surface antigen.

To characterise these associations at the gene and allelic level I developed a large, high-resolution (6-digit ‘G’) population-specific human leukocyte antigen (HLA) imputation reference panel including 697 individuals from the vaccine GWAS typed at 11 genes, highlighting the diversity of HLA across the African continent. Using this panel I imputed HLA into the remaining GWAS dataset to fine-map the associations to specific HLA alleles, amino acid and single nucleotide polymorphism sites; some of which were found to be African specific.

I then used these HLA association findings observed with PT response to correlate, through genetics, this trait with susceptibility to whooping cough in an independently recruited and analysed set of cohorts from the UK. I further used these genetic correlations to demonstrate the relevance of levels of PT-specific circulating follicular helper T-cells and *TRBV29-1* T-cell receptor gene expression levels in the development of this protective immune response against PT. By using HLA-peptide binding studies I also demonstrate the diversity of mechanisms that are involved in HLA-disease association, showing that the breadth and affinity of DT-peptide binding are increased with HLA-DRB1 alleles associated with increased DT antibody responses.

Taken together, these data represent the first comprehensive genetic association study of multiple vaccine responses undertaken in African infants. These results highlight the importance of human genetics in modulating protective responses against vaccine antigens and demonstrate how such associations can be harnessed to understand biological mechanisms of protective efficacy in greater detail that may in turn facilitate future vaccine development.

“In the long history of humankind (and animal kind, too) those who learned to collaborate and improvise most effectively have prevailed.”

- Charles Darwin

Acknowledgements

It would take more than a thesis to thank those people I need to here and so, first and foremost, I will apologise to any that I have forgotten.

Thank you to all those participants who have made this work possible. I can only hope that this work will have future relevance both for your health and to worldwide humanity.

I must thank Professor Adrian Hill, my principal supervisor, for the trust and confidence he had in me before even knowing me. His belief spurred me forwards and gave me opportunities to go places and meet people I could never have dreamed of doing without support. To my second supervisor Professor Georg Holländer, yet again I apologise for our initial ideas not coming to pass. I will always have fond memories of the discussions that we had, enlightening me to complex immunological ideas and I am sure I will be seeking similar inspiration in the years to come. I must thank my surrogate supervisors: Professors Alison Elliott, Gil McVean, Paul Klenerman, and Dr Manj Sandhu for their eternal patience with my never-ending questioning and requests for help from a technical, statistical, ethical, logistic or personal perspective. I learnt a great deal from them and they taught me the true value of effective collaboration.

I am grateful to my lab colleagues: Kate(s), Anna, James, Vivek, Tom, Tara and Anne who encouraged me to think laterally and exposed my shortcomings. The years of 2012-2017 were awe-inspiring for me and this was largely due to those individuals who came together into the group at the same time.

I have too many other people to thank in all the labs I have been lucky to work with around the world: Katie, Nick, Julie, Georgie, Danny, Rachel, and Babatunde at the Jenner; Tommy, Deepti and Cristina (to name a few) at the Sanger; Shabir, Clare, Nadia, Siobhan, Peter at RMPRU; Fiona, Gaby and Pieter at RIVM; Alessandro, Bjoern, John, Cecilia, Ricardo, Chelsea and everyone else at La Jolla; never mind Julian Knight, Alex Dilthey, Luke Jostins, Ben Fairfax, Bill Petri and everyone I have worked with at UKB, LSHTM, DKFZ and PHE; these individuals taught me so much beyond what I imagine a 'normal' PhD student could expect.

I am indebted to those at Keble College, the Oxford BRC, the ERC and the Wellcome Trust for their financial support and educational, academic and clinical guidance.

I am grateful to the streets of Headington and the winding paths of Shotover and Horspath for allowing me to pound you down whilst working out the best way to tackle that script I had to fix or complex E-mail I had to send. You kept me sane.

I want to extend my sincerest thanks to my mother and father for giving me every opportunity I could ever ask for and your patience with my never-ending search for education is to be applauded. Sam, our science chats have always been inspiring - long may they continue. Melissa and Rowan; I'm sorry I've been distant at times but I hope in the future you will look back and think it was worth it, even if you don't understand a word beyond this page. And finally, to Caroline; my wife and long-term inspiration. Thank you for your never-ending support and patience even when the going got rough. It was a life-line knowing you have always believed in me. I could not have done this without you.

Attributions

This thesis is a report of research work carried out at the Wellcome Trust Centre for Human Genetics, University of Oxford, between September 2014 and October 2017. The work was funded by the Wellcome Trust and the European Research Council. All concepts, analysis and interpretation of the results presented in the thesis are original and those of the author.

Much of the research presented in the thesis was made possible owing to work undertaken by other groups.

The Entebbe Mother and Baby Study was directed by Professor Alison Elliott in Entebbe, Uganda supported by the Medical Research Council and the London School of Tropical Medicine and Hygiene. The study was running for many years before I became involved but many individuals contributed directly to the data generated that contributed to this thesis including: Allan Muhwezi, Swaib Lule, Harriett Mpairwe, Florence Akello, Mirriam Akello, Hellen Akurut, Dennison Kizito, Joy Kabagenyi, Lawrence Lubyayyi, Moses Kizza, Stephen Cose, Anne Wajja, Gyaviira Nkurunungi and Patrice Mawa.

The Soweto Vaccine GWAS study was directed by Professor Shabir Madhi and supported by many including Clare Cutland, Nadia van Niekerk, Siobhan Trenors, Peter Adrian, Linh Diep and Razia Hassan.

The VAC050 ME-TRAP vaccine trial was directed by Professor Adrian Hill and Dr Sodiomon Sirima and supported by a number of individuals who also collected

vaccine related data including Alfred Tiono, Katie Ewer, Georgie Bowyer, Danny Wright, Philip Bejon and Nick Edwards.

The genotyping was performed by the Wellcome Trust Sanger Institute Sample Management and Genotyping Core including Cristina Pomilla, Siobhan Austin-Guest, Lucy Hildyard, Nicola Corton and Hannah McLaren.

Luminex-based antibody assays were performed at the National Institute for Public Health and the Environment in Bilthoven, The Netherlands, directed by Dr Fiona van der Klis and supported by Gaby Smits and Pieter van Gageldonk. Hepatitis B surface antigen antibody assays were performed at the Microbiology laboratories at the John Radcliffe Hospital, part of Oxford University Health Foundation NHS Trust directed by Katie Jeffery and supported by Lorraine Clark and Susan Wareing.

HLA typing was performed, in part at Addenbrooke's NHS Foundation Trust by Sarah Peacock and Sarah Maxwell. The remainder was undertaken by Histogenetics directed by Nezih Cereb and supported by HwaRan Kim and SungKyu Kim.

I am particularly grateful for the support and guidance provided by Drs Alexander Dilthey and Loukas Moutsianas and Professor Gil McVean for HLA imputation, and I am eternally grateful for the hospitality and generosity of Professor Alex Sette and his huge lab and the team in La Jolla for their support whilst I was in California undertaking flow and binding assays in 2016.

Associated Publications

Mentzer AJ, O'Connor D, Pollard AJ, Hill AV. 2015; Searching for the human genetic factors standing in the way of universally effective vaccines; *Philos Trans R Soc Lond B Biol Sci.* 19;370

Dilthey AT, Gourraud PA, **Mentzer AJ**, Cereb N, Iqbal Z, McVean G. 2016; High-Accuracy HLA Type Inference from Whole-Genome Sequencing Data Using Population Reference Graphs; *PLoS Comput Biol*; 12(10):e1005151

Nash S, **Mentzer AJ**, Lule S, Kizito D, Smits G, van der Klis F, Elliott A. 2017; The impact of prenatal exposure to parasitic infections and to anthelmintic treatment on antibody responses to routine immunisations given in infancy: secondary analysis of a randomised controlled trial; *PLOS NTD*; 11:e0005213

Mentzer AJ et al.; Understanding Cellular Mechanisms of Vaccine Efficacy through High-Resolution HLA Association Profiling in African Infants; *manuscript in preparation* (34 authors)

Additionally, I have contributed to work in the following manuscripts that have not been documented in this thesis:

Bergström A, Oppenheimer SJ, **Mentzer AJ**, Auckland K, Robson K, Alpers M, Koki G, Pomat W, Siba P, Xue Y, Sandhu MS, Tyler-Smith C. 2017; A Neolithic Expansion and Strong Genetic Structure in the Independent History of New Guinea; *Science*; 357(6356):1160-1163

Parks T et al. 2017; Association between a common immunoglobulin heavy chain allele and rheumatic heart disease risk in Oceania; *Nature Commun.*; 8:14946 (author 7 of 19)

Malaspinas AS, Westaway MC, et al. 2016; A genomic history of Aboriginal Australia; *Nature*; 538(7624):207-214 (author 47 of 75)

Mentzer A, Nayee S, Omar Y, Hullah E, Taylor K, Goel R, Bye H, Shembesh T, Elliott TR, Campbell H, Patel P, Nolan A, Mansfield J, Challacombe S, Escudier M, Mathew CG, Sanderson JD, Prescott NJ. 2016; Genetic Association Analysis Reveals Differences in the Contribution of NOD2 Variants to the Clinical Phenotypes of Orofacial Granulomatosis; *Inflamm Bowel Dis*;22(7):1552-8

Thorne L*, Nalwoga A*, **Mentzer AJ**, de Rougemont A, Hosmillo M, Webb E, Nampijja M, Muhwezi A, Carstensen T, Gurdasani D, Hill AV, Sandhu MS, Elliott AE, Goodfellow I; The First Norovirus Longitudinal Seroepidemiological Study from Sub Saharan Africa Reveals High Seroprevalence of Diverse Genotypes Associated with Host Susceptibility Factors; *in press*

Gurdasani D, Carstensen T et al; The Uganda Genome Project creates a roadmap for GWAS in Africa; *submitted* (author 21 out of 51)

Homberger J et al; Genomic Insights into the Ancestry and Human Demography of Remote Polynesia; *manuscript in preparation* (third from last author)

Mentzer AJ*, Brenner N* et al; A scalable platform for assessing the seroprevalence of multiple infections important in the development of chronic disease; *manuscript in preparation*

Mentzer AJ et al; Stratifying the efficacy of malaria vaccines through high-resolution HLA typing; *manuscript in preparation*

Mentzer AJ et al; Understanding the impact of HLA diversity on outcomes of patients with Acute Respiratory Distress Syndrome; *manuscript in preparation*

Jostins L*, **Mentzer AJ*** et al; Cytokine quantitative trait loci in a spectrum of cellular models of innate immune activity; *manuscript in preparation*

Contents

Abstract.....	i
Acknowledgements.....	iii
Attributions	v
Associated Publications	vii
List of Figures	xviii
List of Tables	xxiii

Chapter 1 1

1 Introduction 1

1.1 Vaccines.....1

1.1.1 Expanded Programme on Immunisation 2

1.1.2 Live attenuated vaccines.....5

1.1.3 Inactivated vaccines.....7

1.1.4 Inactivated toxoid vaccines 8

1.1.5 Capsular polysaccharide vaccines 13

1.1.6 Recombinant subunit vaccines 14

1.1.7 Other vaccines 15

1.2 The human immune system 16

1.2.1	The innate immune system	17
1.2.2	The adaptive immune system	19
1.2.2.1	Dendritic cells	19
1.2.2.2	T cells	20
1.2.2.3	B-cells and antibody production	25
1.2.2.4	Follicular helper T-cells	29
1.2.3	Vaccine-targeted immunity	30
1.3	Human genetics.....	32
1.3.1	The genetics of human populations in relation to disease.....	33
1.3.2	Genetic association studies.....	35
1.3.3	HLA in complex disease	38
1.3.4	Genetic association studies of vaccine responses.....	41
1.3.4.1	Hepatitis B.....	41
1.3.4.2	Other EPI vaccines	46
1.4	Conclusions and outline of thesis	48
Chapter 2		51
2	Methods	51
2.1	The cohort populations	51

2.1.1	Uganda: The Entebbe Mother and Baby Study	53
2.1.2	South Africa: The Soweto Vaccine Response Study	55
2.1.3	Burkina Faso: The VAC050 ME-TRAP Malaria Vaccine Trial	55
2.2	Laboratory methods	57
2.2.1	Whole blood sample preparation	57
2.2.2	DNA extraction and quantification	57
2.2.3	Antibody concentration measurements	58
2.2.3.1	Multiplex immunoassay	58
2.2.3.2	HBsAg ELISA and other vaccine responses	61
2.2.4	Genome-wide genotyping	62
2.3	Analytical methods	62
2.3.1	Statistical power	62
2.3.2	Genotype quality control	63
2.3.2.1	Variant mapping	64
2.3.2.2	Sample quality control per individual	64
2.3.2.3	Sample quality control per marker	66
2.3.2.4	Population structure	68
2.3.2.5	Genotype imputation	68

2.3.3	Antibody phenotype normalisation and covariate modelling.....	71
2.3.4	Association testing.....	72
2.3.5	Quality control following association analysis.....	75
2.3.6	Meta-analysis.....	76
2.3.7	Heterogeneity assessment.....	77
2.3.8	Multi-ethnic meta-analysis	78
Chapter 3		80
3	Genome-Wide Association Study of Multiple Vaccine Antibody Responses in Genetically Diverse African Infants	80
3.1	Background.....	80
3.2	Cohorts and study design.....	81
3.3	Genotype data quality control.....	84
3.4	Population structure within populations	85
3.5	Genotype imputation.....	87
3.6	Measured vaccine antibody responses	89
3.7	Other predictors of vaccine response	95
3.8	Intra-population vaccine response GWAS	101
3.9	Fixed effects meta-analysis.....	104

3.10	Non-MHC associations previously reported with vaccine-related traits ..	109
3.11	Conclusions.....	112
Chapter 4		118
4	Framework and Population-Specific Resource Development for Fine-Mapping MHC Associations with Vaccine Responses in African Infants....	118
4.1	Background.....	118
4.2	Methods	121
4.2.1	HLA typing.....	121
4.2.1.1	Sanger HLA typing	121
4.2.1.2	MiSeq HLA typing.....	1222
4.2.2	HLA nomenclature and comparisons between platforms	1233
4.2.3	HLA diversity characterisation across African populations.....	1266
4.2.4	HLA imputation.....	1288
4.2.5	Pooled LMM association testing, validation and multivariate association analyses	13030
4.2.6	Fine-mapping HLA associations	13131
4.3	Results.....	1344
4.3.1	Concordance between HLA typing platforms	1344
4.3.1.1	HLA-B.....	1355

4.3.1.2	HLA-C.....	1366
4.3.1.3	HLA-DPA1	1377
4.3.1.4	HLA-DPB1	1377
4.3.1.5	HLA-DQA1.....	1388
4.3.1.6	HLA-DQB1.....	1388
4.3.1.7	HLA-DRB4	1399
4.3.2	HLA typing 697 individuals from three African populations	14040
4.3.3	African-specific HLA imputation reference panel validation	1588
4.3.4	Development of a framework for conditional association analyses in structured populations	16363
4.3.5	Determining the genetic architectures of multiple vaccine responses	165
	65	
4.3.5.1	pLMM analysis of genotypes across the extended MHC locus 16565	
4.3.5.2	Multivariate association analysis of all vaccines administered simultaneously	1677
4.3.6	Building a model of genetic architecture of vaccine responses	1688
4.3.7	Understanding between-study heterogeneity	1777
4.4	Conclusions.....	18181

Chapter 5	19090
5 Characterising the Molecular and Cellular Effects of HLA Diversity on Vaccine Efficacy	19090
5.1 Background	19090
5.2 Methods	19292
5.2.1 Whooping-cough GWAS	19292
5.2.2 Determining correlation of variant effect estimates between studies 19494	
5.2.3 Antigen Inducible Marker assay	19595
5.2.4 Peptide binding assays	1988
5.2.5 T-cell receptor expression dataset	200
5.3 Results	202
5.3.1 Correlation between PT response and susceptibility to WC	202
5.3.2 Measuring the significance of correlation of amino acid residue associations between independent GWAS studies	21010
5.3.3 Identifying cellular mechanisms correlating with protection against pertussis	215
5.3.4 HLA-DRB3 affecting PT response	2177

5.3.5	Identifying cellular mechanisms correlating with protection against diphtheria.....	220
5.3.6	Peptide binding studies.....	222
5.3.6.1	Pertussis toxin	222
5.3.6.2	Diphtheria toxin	2277
5.3.6.3	Validation of in silico methods using in vitro binding assays ...	231
5.3.7	Using the HLA residue association signal to correlate antibody responses to T-cell receptor expression	232
5.4	Conclusions.....	234
Chapter 6		244
6 Discussion		24444
Limitations and future directions		2488
References.....		25252
Supplementary Tables.....		278

List of Figures

Figure 1.1: Laboratory confirmed cases of pertussis in England by age group and month: January 2012 to February 2017.	12
Figure 1.2: Types of memory T cells.	22
Figure 1.3: Types and functions of T-cell subsets.	24
Figure 2.1: National guidelines for timings and combinations of vaccines delivered as part of the Expanded Programme on Immunisation (EPI) and approximate representation of sampling times in the respective study groups.	54
Figure 2.2: Concepts of the Luminex assay.	60
Figure 2.3: Regression line through continuous normally distributed data that is associated with the SNP genotype.	73
Figure 3.1: Principal components one to four (PC1-PC4) of the three populations of vaccinated infants.	86
Figure 3.2: Mean imputation quality score per set of SNPs on chromosome 6.	88
Figure 3.3: Distributions of vaccine responses in individual cohorts following transformation using two methods.	91
Figure 3.4: Correlation between normalised vaccine antigen responses across study populations.	92
Figure 3.5: Correlation between vaccine responses against MV and TT in Entebbe individuals as measured using Luminex and ELISA assays.	94
Figure 3.6: Relationship between normalised antibody concentration against DT and time between last vaccination and sampling within each cohort.	96

Figure 3.7: Relationship between normalised antibody concentrations and sex in three populations.	97
Figure 3.8: Relationship between normalised antibody concentrations and HIV infection status in Entebbe and Soweto populations.	98
Figure 3.9: Relationship between normalised measles antibody concentrations and time between sampling and vaccination and maternal measles concentrations at delivery.	99
Figure 3.10: Relationship between normalised antibody concentrations and asymptomatic malaria parasitaemia in two populations.	100
Figure 3.11: Quantile-quantile plots of association statistics for all measured vaccine traits in all three populations.	102
Figure 3.12: Manhattan plots of association statistics for all measured vaccine traits in all three populations.	103
Figure 3.13: Manhattan plots of measured vaccine response fixed effects meta-analysis.	105
Figure 4.1: Nomenclature of the human leukocyte antigen alleles.	124
Figure 4.2: Diversity of HLA-A alleles by population.	144
Figure 4.3: Diversity of HLA-B alleles by population.	145
Figure 4.4: Diversity of HLA-C alleles by population.	146
Figure 4.5: Diversity of HLA-DPA1 alleles by population.	147
Figure 4.6: Diversity of HLA-DPB1 alleles by population.	148
Figure 4.7: Diversity of HLA-DQA1 alleles by population.	149
Figure 4.8: Diversity of HLA-DQB1 alleles by population.	150
Figure 4.9: Diversity of HLA-DRB1 alleles by population.	151

Figure 4.10: Diversity of 2-digit HLA-DRB3 alleles by population.	152
Figure 4.11: Diversity of 2-digit HLA-DRB4 alleles by population.	152
Figure 4.12: Diversity of 2-digit HLA-DRB5 alleles by population.	153
Figure 4.13: Allelic richness of HLA alleles at 2 and 4-digit resolution across the three African populations calculated using bootstrap resampling.	154
Figure 4.14: Concordance between HLA alleles imputed with HLA*IMP:02 or HLA*IMP:02G and typed using MiSeq.	162
Figure 4.15: Correlation between significance values observed across the MHC when testing for association between variants and each trait comparing the pooled LMM and FE meta-analysis approaches	164
Figure 4.16: Regional plots of association between SNP variants and antibody response traits with and without inclusion of top associated variants as covariates.	166
Figure 4.17: Manhattan plot of a multivariate association between genetic variants and seven correlated vaccine responses.	167
Figure 4.18: The genetic architecture of associations between class II HLA variants and responses to multiple vaccine antigens.	175
Figure 4.19: Credible set of variants across the class II HLA region associated with response to PT as determined using all three cohorts or excluding the Soweto population.	180
Figure 4.20: Credible set of variants across the class II HLA region associated with response to FHA as determined using all three cohorts or excluding the Soweto population.	181
Figure 5.1: Gating strategy to identify AIM non-naïve CD4 ⁺ T-cells and the T _{FH} (CXCR5 ⁺) subset.	197

Figure 5.2: Correlation of SNP beta effect estimates between WC GWAS and GWAS of PT, FHA and PRN responses.	208
Figure 5.3: : Correlation of biallelic HLA amino acid beta effect estimates between WC and PT, FHA and PRN response GWA studies.	209
Figure 5.4: Permutation distributions (100,000 permutations) measuring the significance of correlations between effect estimates of HLA amino acids pruned by LD ($r^2 < 0.35$) from WC and vaccine response GWAS.	212
Figure 5.5: Correlation between LD pruned HLA amino acid effect estimates for WC susceptibility between 1958BC and ALSPAC studies and the corresponding permutation distribution (100,000 permutations) to measure significance.	213
Figure 5.6: Permutation distributions (100,000 permutations) testing significance of correlations between HLA amino acid effect estimates for WC and PT vaccine responses in the three individual tested cohorts.	214
Figure 5.7: Ratio of circulating PT:TT specific TFH in donors of known HLA-DRB1 type.	217
Figure 5.8: Ratio of circulating PT:TT specific TFH in donors with known numbers of copies of HLA-DRB3*02 or HLA-DRB3*03 (A) or HLA-DRB3*01 (B) alleles.	219
Figure 5.9: Ratio of circulating DT:TT specific TFH in donors of known HLA-DRB1 type.	222
Figure 5.10: IEDB predicted affinities for PT-derived peptides predicted to bind to all IMGT HLA-DRB1 alleles divided into groups dependent on likely effect on PT antibody responses.	225
Figure 5.11: IEDB predicted affinities for DT-derived peptides predicted to bind to all IMGT HLA-DRB1 alleles divided into groups dependent on likely effect on DT antibody responses.	228

Figure 5.12: Correlation between predicted (IEDB) and measured binding affinities for 28 peptides derived from DT. 231

Figure 5.13: Permutation distributions (100,000 permutations) measuring the significance of correlations between effect estimates of HLA amino acids pruned by LD from TCR V gene and vaccine response GWAS. 233

List of Tables

Table 1.1: Dates of introduction of some first generation vaccines for use in humans.	4
Table 1.2: Terms used in defining responses to vaccines.	5
Table 1.3: Properties of immunoglobulin isotypes and subclasses.	27
Table 2.1: Clinical and demographic data sought from recruited participants and their parents in the three populations.	52
Table 3.1: Descriptive characteristics of the individuals recruited into the study from three African populations.	83
Table 3.2: Summary of genotype QC steps for each population.	84
Table 3.3: Number of variants imputed into individual datasets divided between autosomes and the X-chromosomes.	89
Table 3.4: Association statistics for the index associated MHC variants with each measured vaccine response trait.	106
Table 3.5: Association statistics for SNPs in the MHC found associated with variable responses to HBsAg in publications including two Asian populations.	107
Table 3.6: Association of SNP rs7965675 with antibody response to Hib in the three studied populations and in the final fixed effects meta-analysis.	108
Table 3.7: Summary statistics from meta-analysis testing association of <i>ANO4</i> rs7965675 SNP with all vaccine response traits.	108
Table 3.8: Variant meta-analysis association statistics with vaccine traits for those SNPs found associated with vaccine-related traits in previous studies.	110

Table 4.1: Exons targeted for Sanger-sequence based HLA typing in 90 Ugandan individuals.	122
Table 4.2: Exons targeted for MiSeq-sequence based HLA typing.	123
Table 4.3: Concordance between allele types generated using Sanger and MiSeq based sequencing using samples from 47 individuals from Entebbe.	134
Table 4.4: Discordant alleles typed at <i>HLA-B</i> locus using Sanger and MiSeq based typing.	136
Table 4.5: Discordant alleles typed at <i>HLA-C</i> locus using Sanger and MiSeq based typing.	137
Table 4.6: Discordant alleles typed at <i>HLA-DQB1</i> locus using Sanger and MiSeq based typing.	139
Table 4.7: New HLA alleles detected through MiSeq across the three tested populations.	141
Table 4.8: Allelic richness estimates at all HLA loci calculated using the rarefaction method at 2 and 4-digit levels of resolution.	155
Table 4.9: Heterozygosity estimates based on 2-digit allele data across all loci.	156
Table 4.10: Heterozygosity estimates based on 4-digit allele data across all loci.	157
Table 4.11: Concordance between classical types determined using MiSeq-based typing and imputed with HLA*IMP:02 for individuals from Entebbe.	159
Table 4.12: Concordance between classical types determined using MiSeq-based typing and imputed with HLA*IMP:02 for individuals from Soweto.	160

Table 4.13: Concordance between classical types determined using MiSeq-based typing and imputed with HLA*IMP:02 for individuals from Banfora.	161
Table 4.14: Results from manual and automated step-wise modelling of class II HLA variants with five vaccine responses including principal components and time between sampling and vaccination as covariates.	170
Table 4.15: The variance of HLA associated vaccine responses described as amino acid or classical allele variation compared using the BIC.	173
Table 4.16: Testing for additive or dominant models of inheritance in variants associated in final models of vaccine response traits.	174
Table 4.17: Class II HLA signals of association with three vaccine response traits demonstrating significant evidence of heterogeneity.	179
Table 5.1: Flow antibody panel used for the antigen inducible marker assay.	196
Table 5.2: Cell lines and radiolabelled ligands used as binding controls in competitive HLA-peptide binding assays.	200
Table 5.3: Summary statistics for variants most significantly associated with WC susceptibility or PT vaccine antibody.	204
Table 5.4: Summary statistics for index variants associated with WC susceptibility or PT vaccine antibody response identified through cohort or stage of study.	205
Table 5.5: Summary statistics for HLA-DRB1 alleles associated with WC susceptibility or PT vaccine antibody response defined through cohort or stage of study.	206
Table 5.6: Summary statistics for HLA-DQA1 and HLA-DQB1 alleles associated with WC susceptibility or PT vaccine antibody response defined through cohort or stage of study.	207

Table 5.7: Characteristics of donors and PBMC samples used for PT-specific T _{FH} assay.	216
Table 5.8: Characteristics of donors and PBMC samples used for DT-specific T _{FH} assay.	221
Table 5.9: Differences in affinities of the peptides predicted to bind most tightly to HLA-DRB3 alleles comparing alleles associated with increased (DRB3-74Arg) or decreased PT antibody responses.	224
Table 5.10: Differences in affinities of the peptides predicted to bind most tightly to HLA-DRB1 alleles comparing alleles associated with increased (DRB1-233Thr) or decreased PT antibody responses.	224
Table 5.11: Breadth of PT-derived peptides binding to associated HLA-DRB3 alleles.	226
Table 5.12: Breadth of PT-derived peptides binding to associated HLA-DRB1 alleles.	227
Table 5.13: Breadth of DT-peptides binding to associated HLA-DRB1 alleles.	229
Table 5.14: Differences in HLA-DRB1 DT-peptide binding affinity.	230
Table S1: Non-genetic factors explaining the greatest proportion of variance for each vaccine response trait in each population.	279
Table S2: List of class II alleles included in amended MiSeq 6-digit ‘G’ groups.	281

Chapter 1

1 Introduction

The overarching aims of the work described in this thesis are to identify and characterise the mechanisms of human genetic variants impacting on immune responses and protective efficacy of vaccines delivered in early infancy. In this chapter I describe, firstly, the history and concepts relating to the vaccines studied in this thesis, followed by a summary of our current understanding of immune mechanisms that are important for vaccine response. Finally, I describe the field of human genetics and how it has been applied to try to improve our understanding of the mechanisms underlying variable immunogenicity, efficacy and effectiveness of vaccines in use around the world.

1.1 Vaccines

Vaccines arguably represent one of the greatest public health interventions in modern history. Historical records frequently report the observations by modern man of protection against disease afforded by prior exposure to the same disease. 11th Century texts from China, for example, depict early attempts at variolation

against smallpox¹. Individuals insufflated variola scabs thus exposing themselves to unaltered infectious material in order to protect themselves against future infection with the same microorganism. Although this method did reduce risks of succumbing to future episodes of smallpox there were significant risks of contracting smallpox acutely from the scabs². It was not until Edward Jenner demonstrated protection against the same devastating disease through exposure to a relatively harmless poxvirus family member that vaccination became plausible³.

Through the subsequent advances in microbiology^{4,5}, immunology^{6,7}, bio-manufacturing^{8,9} and molecular biology¹⁰ it has now become possible to mass develop, and deliver multiple vaccines that have been successful in preventing, and in some cases eradicating, diseases that historically caused substantial levels of morbidity and mortality. Despite these successes, our understanding of the molecular mechanisms underlying efficacy of vaccines in preventing these diseases remains limited. This in turn hinders our ability to develop new vaccines against infections that continue to cause a substantial burden of morbidity and mortality such as malaria, human immunodeficiency virus (HIV) and tuberculosis (TB)¹¹. Improving our mechanistic understanding of vaccines that are already in widespread use globally, should arguably lead to advances in vaccine development for infections that are proving more difficult to prevent through vaccination.

1.1.1 Expanded Programme on Immunisation

Following Jenner's earliest observations, the ensuing scientific developments led to the sequential introduction of vaccines helping to reduce the incidence of several diseases (**Table 1.1**). By the mid-twentieth century, several diseases were

notably less prevalent owing, most likely, to increased vaccine availability and uptake. However, these benefits were largely limited to developed countries¹². In an effort to build upon the successful implementation of the 1958 Programme for Eradication of Smallpox by the World Health Organisation (WHO), the same organisation launched the Expanded Programme on Immunisation (EPI) in 1974. The aim of this programme was to improve vaccination coverage in every country around the world against the six target diseases: diphtheria, pertussis, tetanus, measles, poliomyelitis and tuberculosis especially in young infants who were most at risk of disease¹³. Hepatitis B virus (HBV), yellow fever and *Haemophilus influenzae* and, in some countries, pneumococcal and meningococcal disease have since been added to this historical list of diseases to target. This successful Programme continues today to ensure that national vaccination schedules are maintained and surveillance systems are in place to ensure that over 90% of the world's infants receive the necessary vaccinations against these, and other important infections.

Table 1.1: Dates of introduction of some first generation vaccines for use in humans in any country worldwide. Adapted from Bland and Clements¹².

Year	Vaccine-preventable disease targeted
1798	Smallpox
1885	Rabies
1897	Plague
1923	Diphtheria
1926	Pertussis
1927	Tuberculosis
1927	Tetanus
1935	Yellow fever
1955	Injectable polio vaccine
1962	Oral polio vaccine
1963	Measles
1967	Mumps
1969	Rubella
1981	Hepatitis B

Given the range of vaccines in use today, any comprehensive understanding is best afforded through a description of the vaccines and diseases organised by subtype¹⁴. In order to understand how some of these vaccines became introduced into the EPI schedule a number of terms require definition as outlined in **Table 1.2**.

Table 1.2: Terms used in defining responses to vaccines. Adapted from ¹⁵⁻¹⁷

Term	Definition
Vaccine efficacy	The reduction of incidence of vaccine-targeted disease in a vaccinated group compared to a non-vaccinated group under ‘ideal’ conditions – usually measured in a double-blind randomised controlled trial
Vaccine effectiveness	The reduction of incidence of disease at the population level – usually determined following introduction of a vaccine on a national scale with subsequent analysis of disease reporting and adverse incidence data
Protection against infection	Successful stimulation of a long-lived immune response that interferes with colonisation or replication of the targeted pathogen in the vaccinated individual
Protection against disease	Successful stimulation of a long-lived immune response that prevents the development of disease from the targeted pathogen when colonisation or replication of the pathogen may still occur
Primary vaccine failure	Development of disease in a vaccinated individual due to failure to mount an appropriate immune response against the vaccine-targeted pathogen
Secondary vaccine failure	Development of disease in a vaccinated individual despite a documented appropriate immune response against the vaccine-targeted pathogen
Correlate of protection	An immunological measurement that is statistically interrelated and <i>may</i> be responsible for protection; either against infection or disease

1.1.2 Live attenuated vaccines

The smallpox vaccine is considered to be one of the first examples of a live attenuated vaccine, although it does not fit the strictest of definitions. The complete definition of an attenuated vaccine is the use of a nonvirulent strain of pathogen to induce protection against future challenge with other virulent strains¹⁸. The smallpox vaccine instead consists of another species of poxvirus rather than a non-virulent form of the same species¹⁹. Nevertheless, principles learnt from the development of the smallpox vaccine stimulated some of the earliest successes in vaccinology. Pasteur’s realisation of culturing microorganisms in the presence of environmental insults such as high

temperature, oxygen and chemicals rendering the organism ‘a-‘virulent or attenuated⁵, led to the repeated passage of *Mycobacterium bovis* to produce an attenuated tuberculosis vaccine by Camille Guérin and Leon Calmette^{4,20}. As the ability to culture viruses in cells grown *in vitro* became realised, a spate of attenuated vaccines against pathogenic viruses were similarly developed. Oral polio vaccine was developed by passage through human kidney cell cultures in work including that was started by Hilary Koprowski and followed up by Albert Sabin^{21,22}. In contrast, a successful combination of human kidney and amnion cell cultures followed by embryonated hens’ eggs and chick embryo cell cultures were required to develop the measles vaccine (MV)^{23,24}.

After the development of the vaccine it was important to demonstrate, firstly that the vaccines were safe for delivery into human subjects, and secondly that they did not result in acute presentation of the disease targeted by vaccination, and that ultimately the vaccines provided some level of protection against disease when recipients were later exposed to the wild-type agent. Achieving an understanding of the mechanisms underlying vaccine success was not the main focus of the research and most work at that time was limited to observing an increase in neutralisation or complement-fixing antibodies in vaccine recipients that was, of course, ground-breaking for this era^{24,25}.

Irrespective of the putative mechanisms of protection afforded by these vaccines, epidemiological evidence fell strongly in support of the use of attenuated vaccines for preventing disease that led to their inclusion in the EPI schedules. The introduction of polio vaccines into national programmes led to the sequential and independent disappearance of poliomyelitis; acute flaccid paralysis or

encephalomyelitis, from these countries²⁶. Furthermore, multiple studies investigating MV have demonstrated effectiveness estimates of reduction in morbilliform rash, fever and associated sequelae of between 84 and 94% depending on the number of doses and the age at which the vaccine was received²⁷.

Several other attenuated vaccines have been developed and deployed either directly as part of EPI or in more focussed national or travel-related schemes. Bacille Calmette-Guérin (BCG), the vaccine developed to prevent cases of tuberculosis, is perhaps most prominent amongst these. The effectiveness of BCG in preventing tuberculosis is well noted to vary substantially by distance from the equator²⁸ but the mechanisms underlying this observation and immunological correlates of protection remain poorly understood²⁹. Identifying such correlates are the active focus of many other research individuals and groups^{30,31} and therefore any further description or discussion of vaccines targeting tuberculosis fall outside the scope of this thesis.

1.1.3 Inactivated vaccines

Around the same time as the discovery of attenuation, investigators realised that careful killing of bacteria through heat or chemical treatment could still confer immunogenicity without inflicting disease¹⁴. This work resulted in early vaccines against typhoid, cholera and the plague. However, it is the whole-cell vaccine, which protects against whooping cough caused by *Bordetella pertussis*, that has persisted in vaccination schedules since its historical development. Madsen was the first to report the efficacy of formaldehyde treated bacteria in reducing the incidence of whooping cough in case series and trials from Denmark and the

Faroe Islands³². However it was the work of Kendrick and Eldering that led to the large-scale manufacture and development of a safer thimerosal-inactivated form of the vaccine³³. Once developed, their field trials followed by the seminal Medical Research Council trials in the 1940s demonstrated significant reductions in pertussis incidence in vaccinees compared to unvaccinated individuals that contributed to the vaccine being introduced into international vaccination schedules^{34,35}.

Equivalent efforts were undertaken for viruses. Fuelled by the cell culture revolution discussed earlier, Salk undertook one of the fastest translational developments in history generating formalin-inactivated forms of poliovirus³⁶ and completed one of the largest clinical trials ever undertaken^{37,38} ensuring licensure and deployment of the vaccine. Alongside the Sabin oral vaccine this has undoubtedly contributed to the near-global eradication of polio.

Efforts to develop vaccines for other viruses based on inactivation have not been so successful and in some cases have met with unfortunate consequences. An inactivated form of the measles virus was discovered not to induce long-lived immunity and instead, when recipients were exposed to wild-type virus, they developed severe atypical measles infections²⁴.

1.1.4 Inactivated toxoid vaccines

In cases where disease could be attributed to specific proteinaceous products of the infectious agent, advances in identifying these components and ‘detoxifying’ them in a manner to render them non-injurious yet immunogenic led to the development of an increasing number of successful vaccines. Glenny and Hopkins

were the first to report the production of a diphtheria ‘toxoid’ (the toxin – DT – deactivated through chemical means) using formalin treatment³⁹. These early preparations were improved upon by Ramon resulting in the ‘anatoxine-Ramon’ product that was administered into children very soon after development⁴⁰ and found to reduce the incidence of cases of laryngeal and tracheal oedema and sequelae due to diphtheria. A similar principle was applied to tetanus toxin¹ to give the tetanus toxoid (TT) that again was found to reduce the incidence of clinical tetanus following vaccination. These scientific developments and early administration into volunteers meant that formal trials were never undertaken to test the efficacy of either toxoid vaccine. However, subsequent epidemiological studies have demonstrated high levels of effectiveness alongside immunological anti-toxin responses (now well known to be antibody mediated) warranting their international deployment⁴¹. The effectiveness of DT is estimated to be in the region of 95% following three doses of the vaccine delivered in infancy. Outbreaks of diphtheria in Russia and Laos have been associated with reduced vaccination coverage of the population usually attributed to challenges in keeping the vaccine preparation temperature chilled in the field (a breakdown in the cold-chain)^{42,43}. Interestingly, one study in Laos observed correlations between reduced DT antibody responses and ethnicity that had not been reported previously⁴². Following three doses with a booster dose ten years after the final dose, TT is considered to be highly effective with very little variation in effectiveness observed worldwide⁴¹.

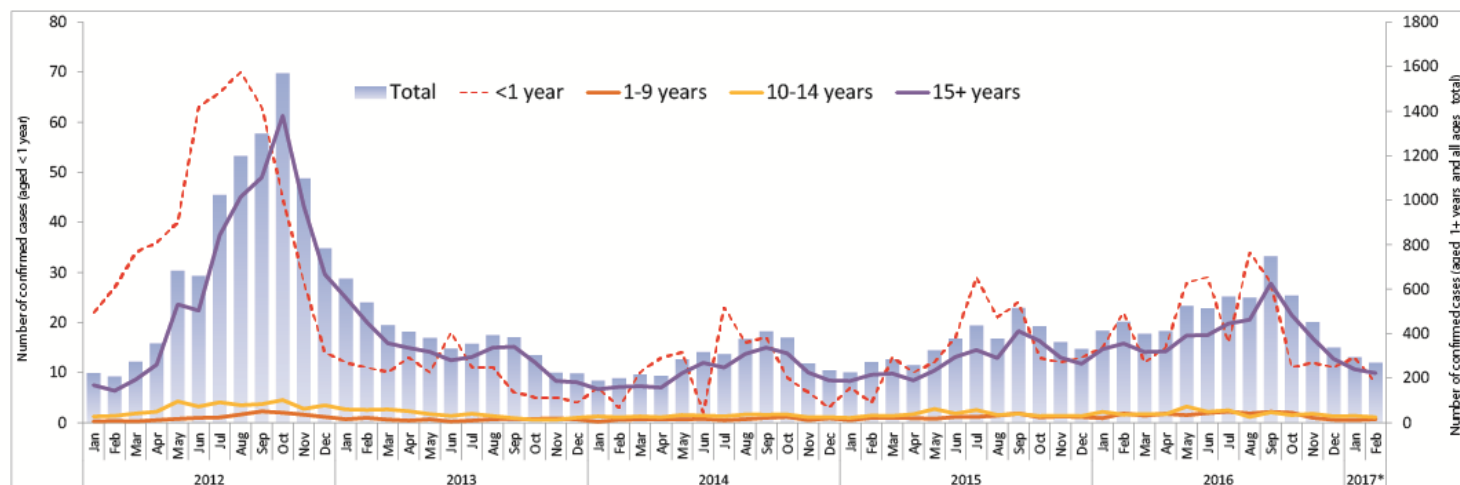
DT and TT components were combined with the inactivated pertussis vaccine into a single vaccine by Kendrick in 1942⁴⁴ giving rise to the DTP vaccine (or DTwP to represent the ‘whole’-cell pertussis organism). The subsequent trials that were

undertaken to assess the efficacy of this combination also became the first systematic assessment of the effect of alum additives into the vaccine preparations that became the future mainstay adjuvant used in vaccines.

By 1975, however, there was significant clinical and public opposition to the use of the whole cell pertussis vaccine owing to low perceived effectiveness and an increased risk of unwanted effects including febrile seizures with concomitant risks of permanent neurological disability⁴⁵. This led to the variable introduction of a modified version of the pertussis vaccine including inactivated products of specific antigens believed to be immunodominant (including the toxin (PT), filamentous hemagglutinin antigen (FHA), pertactin (PRN) or fimbriae (FIM)) excluding the toxic endotoxin to result in the acellular form of the vaccine (aP). aP was then combined with TT and DT to give DTaP and ultimately a number of other vaccines including those targeting polio, *Haemophilus influenzae*, and hepatitis B were combined to result in 'five-in-one' or 'six-in-one' vaccine preparations used widely around the world today. An alternative to the DTaP preparation is the TdaP that contains different concentrations of diphtheria toxoid.

The numerous combinations possible with each of the pertussis vaccine components resulted in multiple trials being undertaken comparing aP with wP vaccines. The overall conclusions of these trials was that wP vaccines were marginally more likely to protect against clinical and microbiological diagnoses of whooping cough despite demonstrating equivalent levels of antibody mediated immunogenicity against the retained vaccine components⁴⁶. This improved efficacy comes at a cost of increased reactogenicity including an increased risk of fever and febrile seizures. These findings led to an increase in uptake of aP

vaccines into national vaccination schedules. However, in recent years there have been a number of outbreaks of pertussis in aP-vaccinated populations including episodes in England (**Figure 1.1**) raising concerns that long-term aP use may not be sustainable for controlling pertussis without further boosters in older adolescence or young adulthood⁴⁷. The reasons for these differences remain unclear although there is growing evidence that T-cell responses, as measured both in humans and non-human primate models, may be fundamentally different between aP and wP priming that may affect the differential memory responses^{48,49}. The key impact of these differences appears to be a faster waning of immunity to pertussis following aP priming compared to wP. It remains unclear, however, what immunological measures constitute the true correlates of protection against pertussis^{15,50} which makes understanding these differences particularly challenging.



*2017 are provisional data

Figure 1.1: Laboratory confirmed cases of pertussis in England by age group and month: January 2012 to February 2017.

Number of cases of pertussis in infants less than 1 year of age (red dashed line and left y-axis) or other age stratified groups of children and adolescents (continuous lines and right y-axis) with totals shown in binned bars categorised by month (right y-axis). Adapted from PHE⁵¹

1.1.5 Capsular polysaccharide vaccines

Early developments in bacteriology determined that a limited number of bacterial species were surrounded by polysaccharide capsules that were targeted by an effective immune system and furthermore that the immunological targeting against such components could induce protection against colonisation and disease following vaccination⁵². This strategy was first applied to *Streptococcus pneumoniae* serotypes that frequently caused invasive pneumococcal diseases including pneumonia, meningitis and bacteraemia associated with shock⁵³. The same principles were subsequently applied to *Haemophilus influenzae* type b (Hib) organisms which used to be a significant cause of meningitis, pneumonia, epiglottitis, bacteraemia and septic arthritis⁵⁴; and *Neisseria meningitidis* serogroups that were also capable of causing fulminant meningococcal meningitis and septic shock⁵⁵. The devastating nature of diseases caused by Hib in particular led to early attempts to vaccinate individuals using the polysaccharide subunit alone; a polyribosyl ribitol phosphate (PrP). However early attempts in infants were unsuccessful due to rapid waning of the antibody (determined early on to be the correlate of protection) and infants became quickly re-susceptible to disease⁵⁶.

To circumvent this problem in infants peptide proteins were conjugated to the polysaccharides to enhance the immunological responses⁵⁷. Conjugation of the Hib PrP to TT or to a genetically detoxified form of DT (CRM197) was widely demonstrated to show efficacy in several clinical trials although there were differences between populations⁵⁶. Despite these differences, these vaccines have undoubtedly been effective, contributing to the near-disappearance of Hib disease in all countries that achieve sustained high levels of vaccination⁵⁸. Moreover, the

effectiveness of Hib vaccination has persisted despite observations that the immunogenicity of Hib vaccines may be impaired when combined with DTaP combinations as opposed to DTwP suggesting that these immunological differences may not have an effect on clinical effectiveness of the vaccine⁵⁹.

Similar conjugation efforts have been applied to the causative agents of pneumococcal and meningococcal disease but these vaccines have only been introduced into the schedules of multiple countries in very recent years. These vaccines also, therefore, fall out of the scope of this thesis.

1.1.6 Recombinant subunit vaccines

Hepatitis B virus (HBV) is a virus that can be spread either vertically from mother to child, or horizontally through sexual or blood-borne routes. HBV is capable of causing a spectrum of disease spanning acute and chronic hepatitis, cirrhosis, hepatocellular carcinoma (HCC) and liver failure. Despite the recognised burden of this infection on healthcare systems worldwide, it was some time before a vaccine against HBV was possible owing to the challenges of culturing the virus *in vitro*⁶⁰. However, advances in virology led to the understanding that the outer capsid protein: hepatitis B surface antigen (HBsAg) was capable of preventing infection when sourced from the serum of infected individuals and subjected to heating at 98°C⁶¹. Concerns about the safety of this approach in light of the potential for transmission of other infections drove the field to harness the growing field of genetic engineering to clone, reproduce and assemble multimeric HBsAg viral capsids in yeast⁶². These recombinant vaccines were successfully evaluated in clinical subjects demonstrating immunogenicity equivalent to plasma derived vaccines with no demonstrable *Candida spp.* seropositivity⁶³ that led to

early calls to introduce the vaccine into the EPI schedules. Such an introduction was quicker in some countries⁶⁴ than in others and arguably most successful in Taiwan^{65,66} and The Gambia⁶⁷⁻⁶⁹ where vaccination started between 1984 and 1986. The close epidemiological surveillance maintained in these countries allowed effective estimates to be calculated for each of: immunogenicity as measured through antibodies against HBsAg; efficacy in terms of prevention of seroconversion against the virus (measured by core antibody levels – different to the surface antigen used for vaccination); and ultimately effectiveness at reducing population levels of HCC. Overall, the vaccine is reported to be highly efficacious and effective, which explains its continued inclusion into EPI schedules worldwide where incidence is sufficiently high. Routine HBsAg immunisation has not traditionally been delivered in the United Kingdom⁷⁰ although as of mid-2017 it has now been introduced into the routine infant immunisation schedule⁷¹.

1.1.7 Other vaccines

Given the success of many of the vaccines listed above there has been an increasing drive to develop new vaccines against a host of other infections and diseases including cancer. The complexity of this task for some pathogens, especially malaria, HIV and TB, has led to increasingly intricate vaccine designs. Such vaccines will not be discussed further in this thesis as these are not part of the EPI schedule. The aim of the work presented here is instead to generate an enhanced mechanistic understanding of responses to multiple vaccines through first appreciating those of more ‘simple’ vaccine types that have been in use for many years.

1.2 The human immune system

The goal of modern vaccinology is to administer a compound that induces durable, ideally lifelong, protection against possible infection or colonisation by the target agent (sterilising immunity) in all recipient individuals. For some infectious agents it may be too difficult to reduce the chances of infection. In these cases the aim instead is to nullify the risk of sequential disease resulting from physiological derangement following infection. The historical development of vaccines detailed above led to a significant improvement in understanding of immunology and, in particular, mechanisms of immunological memory developed against infectious agents following initial exposure. What now follows is a summary of the current understanding of the steps involved in the immune response against foreign antigens, most adequately represented by those presented as part of an infectious agent. It should be noted that many of the insights into the immunological mechanisms as described below are based on experiments undertaken in animal models. Testing the activity of pathways in the timed presence or absence of infectious agents or toxic stimuli, and accessing crucial lymph tissue (for example) in humans is ethically unacceptable except in increasing cases of human vaccine challenge models⁷². Phenotypes such as cell protein marker surface expression and human cell behaviour *in vitro* may then be correlated with the findings from experiments undertaken in animal models to demonstrate that mechanisms uncovered in such animal models are also likely to be relevant in humans.

1.2.1 The innate immune system

Activation of the innate immune system is a fundamental part of a successful vaccine response. Any breach of a physical barrier such as the skin or respiratory or gut mucosa results in the earliest immune activation through the release of chemical mediators including cytokines and chemokines from damaged tissue or resident immune cells including, classically, macrophages or dendritic cells. In some cases, these steps alone may be sufficient to kill and clear the invading pathogen in which case the inflammation will subside. Alternatively, inflammation may progress to resulting in an activation of the adaptive arm of the immune system⁷³.

The earliest steps in the innate immune response include the activation of the complement cascade⁷⁴. This complex proteinaceous amplification system may be triggered either by exposure to the pathogen itself through either the lectin or 'alternative' pathways, or through the complexing of antibodies with foreign antigen (where previous exposure to the antigen would have been necessary to induce antibodies)⁷³. The conclusion of complement activation includes one or several of the following actions: coating of the invading agent in proteins that augment phagocytosis through opsonisation; the formation of a membrane attack complex that can introduce lethal pores into the pathogen; or the production of chemoattractant molecules that encourage immune cell recruitment. Complement proteins including C3a and C5a induce rapid expression of P-selectin on endothelial cells that trigger the recruitment of neutrophils and other phagocytic cells and induce the cells to release proteases and generate oxidants⁷⁴.

Macrophages and dendritic cells are activated through exposure to foreign antigen using Toll-like receptors (TLRs) and other pattern recognition receptors (PRRs). These receptors recognise conserved repetitive biological units (denoted as pathogen-associated molecular patterns; PAMPs) present on the outer membranes or exposed surfaces of bacteria, viruses or parasites. Once bound to their foreign ligand they are capable of inducing immune cell activation through intracellular signalling^{73,75}. There is now good evidence that re-exposure to some PAMPs can induce a heightened response compared to first exposure suggesting an element of innate memory; better defined as ‘trained’ immunity. This process is not just limited to cells expressing TLRs but also to cells expressing other PRRs, natural killer (NK) cells and innate lymphoid cells (ILCs) that are all designated part of the innate immune system⁷⁶. The mechanisms underlying this training remain unclear but appear to be related to epigenetic effects including chromatin remodelling and altered speed of gene expression. One situation where such activity is believed critical is in the changing dynamic state of the neonate. In such a scenario the opposing mechanism is in fact desired; reduced response to PAMPs whilst the infant becomes accustomed to microbial colonisation with bacterial strains. If this training were not present, bacteria containing products such as lipopolysaccharide would potentially induce an overwhelming innate immune response whilst attempting to clear the resident microbiota in the neonate⁷⁷.

In an innate immune response there is classically a release of cytokines including tumour necrosis factor- α (TNF- α), interferon- γ (IFN- γ), interleukin (IL)-1 β , IL-4, IL-6, IL-10, IL-12, IL-18, CC-motif chemokine ligand 4 (CCL4, also known as macrophage inflammatory protein; MIP-1 β) and tumour growth factor beta (TGF- β)⁷⁸. It is increasingly recognised that the identity and levels of such expressed

agents are highly time, cell and context dependent which together orchestrate subsequent steps in the immune response.

Alongside complement activation, cellular recruitment and cytokine upregulation a critical component of innate activation is the recruitment and activation of professional antigen presenting cells (APCs) that are typically dendritic cells (DCs). Upon activation, these cells avidly phagocytose surrounding debris or foreign material, upregulate their major histocompatibility complex (MHC) class I and II molecule expression and present foreign antigen alongside the co-stimulatory molecules CD80 and CD86 on their surface⁷³. These activated and antigen-carrying cells then migrate through the lymph circulation to lymph nodes where they interact directly with cells of the adaptive system to activate that arm of immunity.

1.2.2 The adaptive immune system

1.2.2.1 *Dendritic cells*

Although the immune system is classically divided into adaptive and innate arms, it is clear that these two complexes function concurrently with constant bidirectional signal interaction shaping the dynamic response of each other. DCs are perceived to sit poised between these two arms orchestrating this coordination. Activated through PRR and cytokine mechanisms, these cells phagocytose foreign material or debris and upregulate CC-chemokine receptor 7 (CCR7); the homing signal for lymphatic endothelial expressed CCR21 directing the cell to the T-cell rich lymph node paracortex⁷⁹. As is becoming increasingly realised in many arms of biological research, such a descriptive concept is very likely to be an

oversimplification. Dendritic cells can be distinguished into different subsets, in animal models at least, and cellular activation varies by organ site and immunological challenge⁷⁹. Nevertheless, the core objective appears to remain the same: increased contact with T-cells in the lymph circulation.

1.2.2.2 *T cells*

The primary contacts of DCs are T-cells. Similar to DCs, T-cells can be subdivided into multiple cell types. They are characterised by their development in the thymus (hence ‘T’) and their expression of a T-cell receptor (TCR)⁷³. As part of their development the germline TCR undergoes a complex rearrangement system whereby multiple variable (‘V’), joining (‘J’) and diversity (‘D’ in the case of the TCR β and δ chain) gene segments are rearranged in a cell-specific manner with an upper bound of 10^{18} possible different combinations⁸⁰. This vast number of possible TCRs is reduced in number through active methods of selection; first positively (ensuring TCR can recognise MHC complexed with self-derived peptides with *moderate* affinity) and then negatively, removing T-cells with TCR that recognise self-antigen too strongly⁷³. Before commencing positive selection, T-cells express both cluster of differentiation (CD) co-receptor 4 (CD4) and CD8 (termed double positive). As they interact with the class I and II MHC during positive selection, the variable affinity for the TCR with either class of MHC determines whether the T-cell becomes a single positive CD4 or CD8 expressing cell⁸¹. The final result is a set of $\alpha\beta$ TCR expressing T-cell co-expressing either CD4 or CD8 that are capable of binding MHC-bound peptide complexes. These processes refer to the majority of T-cells expressing heterodimeric $\alpha\beta$ TCR. A

minority of T-cells are found to express $\gamma\delta$ chains that are limited in distribution to the gut mucosa, skin, lungs and uterus with, as yet, poorly defined roles⁸².

T-cells expressing CD8 recognise MHC class I complexed with peptide bound on APCs. These CD8⁺ T-cells become activated if they detect non-self antigen derived either from intracellular infection of the APC itself (secondary to a viral infection or intracellular living bacteria for example), or following phagocytosis of another infected cell by the APC, or through an increasingly recognised mechanism of sharing of MHC-peptide complexes through ‘cross-dressing’⁸³. CD4⁺ T-cells become activated through the recognition of non-self antigen bound to APCs that are derived almost exclusively through phagocytosis of extracellular pathogen⁷³. The end result of MHC-peptide-TCR complexing is clonal proliferation of antigen specific T-cells and maturation into both effector cells mediating the acute host response, and self-renewing memory cells that are capable of producing long-lived protection⁸⁴. Irrespective of their co-expressed molecule, the memory T-cells (expressing CD45RO and lacking CD45RA) may either be a central type (T_{CM}; CCR7 expressing) or effector type (T_{EM}; non CCR7 expressing). T_{CM} are located in lymphoid tissue and are capable of swift and sustained proliferation alongside IL-2 production. T_{EM} are found in non-lymphoid tissue and are capable of robust and rapid cytokine responses becoming effector T-cells (T_{Eff}) upon re-challenge: IFN- γ and TNF- α in the case of CD4⁺, and cytotoxic proteins in CD8⁺⁸⁵ (**Figure 1.2**). Recently, another class of T-cells has been identified that remains limited to specific tissue sites: the tissue-resident T-cell (T_{RM})⁸⁶. T_{RM} are similar to T_{EM} but with more tissue-specific activities⁸⁷. Even more recently a further classification of T-cell, the stem cell-like memory T-cell (T_{SCM}) has been described. This naïve-like T-cell is capable of very rapid and

substantial proliferation and can differentiate into any of the other types of memory T-cell⁸⁸.

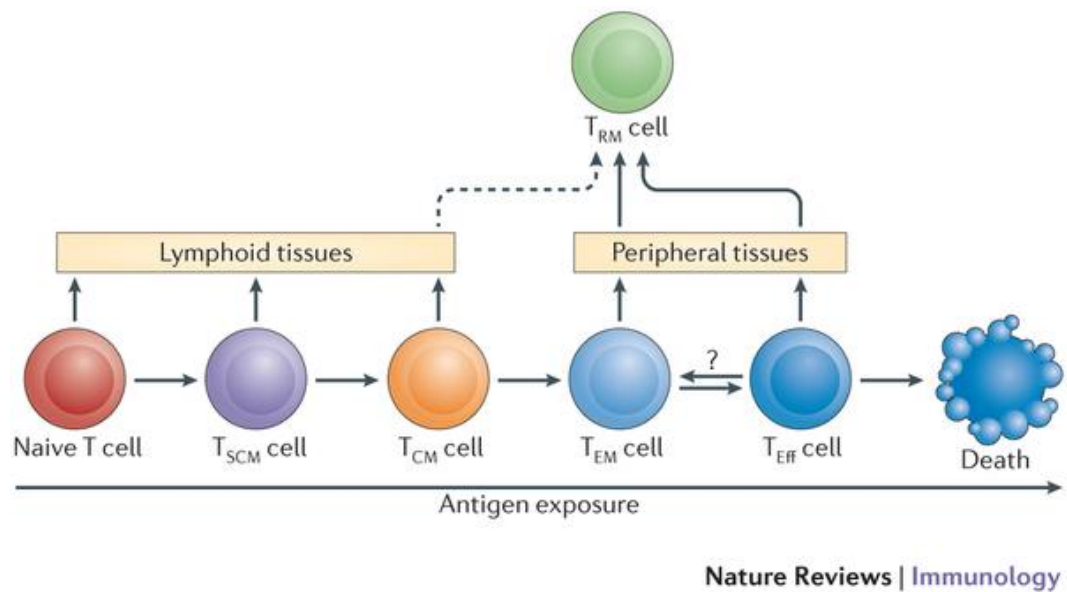


Figure 1.2: Types of memory T cells.

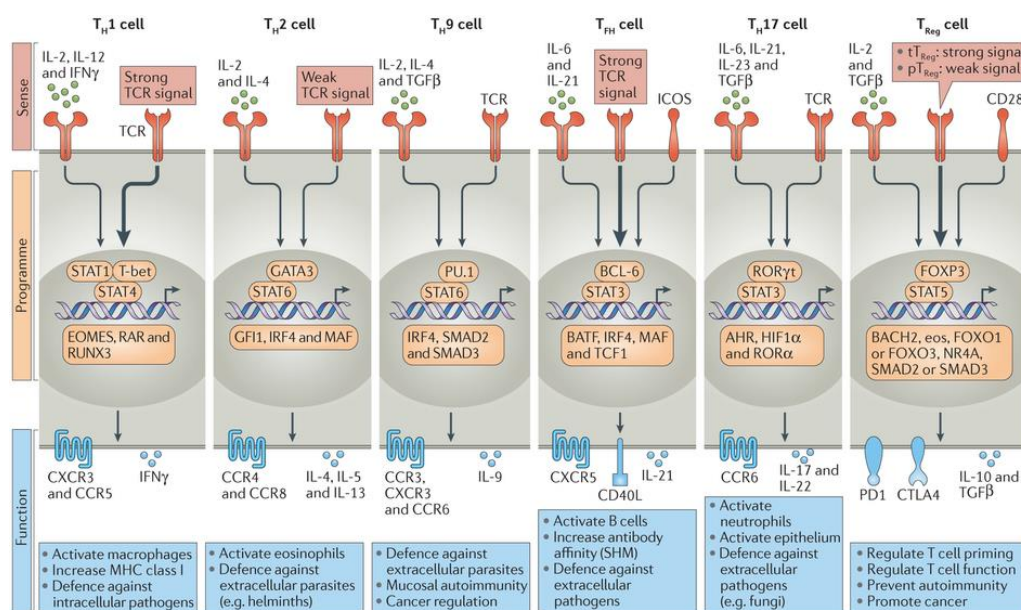
Reproduced with permission from Nature Publishing Group (License no. 4171290836747)⁸⁵

The effector state of either class of T-cell (CD4⁺ or CD8⁺) can only arise through appropriate cell-surface molecule co-stimulation. These markers include CD27 on the T-cell (complexing to CD70 on the APC), CD28 (CD80 and CD86), CD40 (CD40L), OX40 (OX40L) and 4-1BB (4-1BBL)^{73,86}. Appropriate class I MHC-peptide and CD8⁺ stimulation with co-stimulation for approximately 4-5 days will, through the action of IL-2, induce clonal proliferation and terminal differentiation of a proportion of the cells into cytotoxic T lymphocytes (CTLs)⁸⁶. CTLs alter their adhesion molecule expression from a lymph-homing (L-selectin) to a more inflammation site-homing (CD2, lymphocyte function-associated antigen-1 (LFA1) and very late antigen-4 (VLA4)) profile. In these inflamed sites

the CTLs act in an antigen-specific way to target infected, or indeed oncogenic, class I MHC expressing foreign-antigen bound cells without the need for co-stimulation. Once detected, the CTLs release perforin and granzymes, that alongside potential triggering of the Fas (on the surface of the target cell) and Fas ligand (FasL on the surface of the CTL) is capable of triggering target cell death through apoptosis⁷³. There is increasing evidence that CTLs can be subdivided into subsets (TC1 with increased IFN- γ expression and TC2 with increased IL-4 expression for example) although the importance and relevance of this remains unclear⁸⁹.

CD4⁺ cells will also expand in a co-stimulation IL-2 dependent manner but, in contrast to CD8⁺ cells, these cells demonstrate a more established balance and diversity of subset differentiation. The earliest subsets to be identified were T_H1 and T_H2, polarised dependent upon concentrations of IL-12 and IFN- γ ; or IL-4 respectively^{90,91}. However, since the earliest definitions of these subsets there has been an increased identification of other subsets including T_H9, T_H17, follicular helper (T_{FH}) and regulatory (T_{Reg}). These cells each have well defined roles and are differentiated through the cell contact and cytokine microenvironment (**Figure 1.3**) although some plasticity is ever-present⁹¹. The T_H1 and T_H2 subsets support cell-mediated and humoral mediated immunity respectively. T_H1 cells are capable of upregulating macrophage activation through the production of cytokines such as IFN- γ and increased CD40L surface expression alongside CTL IFN- γ expression⁷³. T_H2 cells produce more IL-4, IL-5 and IL-10, which alongside CD40L, stimulate and support B-cell development down-regulate the cell-mediated arm. T_H9 cells are characterised by IL-9 production linked with enhanced T_H2 responses demonstrating an IgE bias and hence linked with asthma

or helminth infections⁹². T_H17 cells are more established neutrophil-enhancing cells putatively involved in inflammatory conditions such as rheumatoid arthritis and psoriasis⁹³. Conversely, the complex sets of T_{Reg} cells differentiate both centrally and peripherally to stem excessive immune activation and activity that are mechanisms deemed important in reducing the risk of autoimmunity and rejection following organ transplantation⁹⁴⁻⁹⁶. T_{FH} cells meanwhile are becoming recognised as the key subset responsible for B-cell help and efficient antibody development as discussed in more detail below.



Nature Reviews | Immunology

Figure 1.3: Types and functions of T-cell subsets.

Reproduced with permission from Nature Publishing Group (License no. 4171291367351)⁹¹

1.2.2.3 *B-cells and antibody production*

The production of antibody is a key component of the adaptive immune response responsible primarily for the protection against extracellular pathogens through either neutralisation of foreign antigen, increasing the likelihood of phagocytosis through opsonisation or triggering of the classical complement cascade⁷³. Antibody is produced by antibody secreting cells (ASCs) that are comprised of terminally differentiated plasmablasts and plasma cells. Stem cells develop into immature B-cells in the bone marrow, undergoing a receptor rearrangement system very similar to that observed for T-cells. Immunoglobulin (Ig) genes undergo reorganisation whereby V-D-J recombination occurs, resulting in surface expression of the heterodimeric IgM in a tightly regulated process. Following recombination, a vast number of clones are deleted through the triggering of apoptosis following the failure to form dimers. Once such dimers are successfully formed the immature B-cells undergo negative selection whereby, similar to the T-cell selection process, they receive apoptosis signals if they react too strongly to self antigen⁷³. Since Ig production does not require concurrent MHC binding there is a lower estimate of Ig protein diversity compared to TCRs with a number prior to selection in the region of 10^{13} ⁹⁷.

Following selection, immature B-cells leave the bone marrow and home to the lymphoid organs where they can mature into a variety of subsets. The majority become follicular B-cells located in the follicles of lymphoid organs. Smaller numbers of cells become B-1 cells (located in peripheral tissues such as the gut and lung) and marginal zone (MZ) B-cells located in the eponymous region of the spleen and lymph nodes.

Follicular B-cells constantly circulate through the lymph and tissues scouting for antigen with their immunoglobulin antigen receptor. They may therefore encounter antigen either in the peripheral tissue early in the inflammation cascade or circulating free in the blood or lymph^{73,98}. Dendritic cells derived both from the infected tissue and located in the follicles (follicular dendritic cells; FDCs) present opsonised, non-degraded antigen on their surface for B-cell detection^{99,100}. Elegant dynamic microscopy studies have demonstrated the rapid speed of antigen diffusion into the lymph circulation either passively or actively. The B-cell immunoglobulin acts as an antigen-specific receptor triggering the early proliferation of the B-cells. Thereafter the immunoglobulin internalises and the cell digests antibody-bound antigen to present on class II MHC molecules at which point the B-cell becomes an APC. Cell activation also promotes the transient secretion of large amounts of antibody that facilitates early functions critical for clearance of the invading pathogen. This may involve secretion of IgM, or the process of class switch recombination may result in the production of other antibody isotypes (including IgG or IgA) that may be more relevant to the invading pathogen (**Table 1.3**). These steps function to increase the speed of B-cell antigen detection that facilitates rapid pathogen clearance and, perhaps more crucially, the regional localisation of cell-antigen contact that is permissive of additional T-cell recruitment and contact.

Table 1.3: Properties of immunoglobulin isotypes and subclasses. From Schroeder and Cavacini¹⁰¹

Isotype	Serum (%)	Complement Fixation	Opsonisation	Cross Placenta
IgG	75	+	+++	+
<i>IgG1</i>	67% IgG	Yes	Yes	+
<i>IgG2</i>	22% IgG	Yes	Yes	+
<i>IgG3</i>	7% IgG	Yes	Yes	+
<i>IgG4</i>	4% IgG	No	No	+
IgM	10	+++	+	—
IgA	15	—	—	—
<i>IgA1</i>		—	—	—
<i>IgA2</i>		—	—	—
IgD	<0.5	—	—	—
IgE	<0.01	—	—	—

MZ B-cells are, in contrast, more innate-like in their function⁹⁸. Unlike other follicular B-cells and more akin to T-cells, these cells appear to develop with a degree of positive selection. MZ cells recognise self antigen or antigen that is constantly being exposed to the immune system including that derived from the microbiota. Together these cells appear to express antibody that is intrinsically more broad in its specificity and can respond to a range of foreign antigen including, in particular, the capsules of some bacteria^{73,102}. Activation of MZ B-cells is classically independent of T-cell help (hence the antigens are referred to as T-cell independent (TI)). These cells therefore act more at the intersection of the innate and adaptive systems being able to produce broad non-specific antibody quickly without the need for antibody maturation undertaken by follicular B-

cells¹⁰³. Their absence partly explains why asplenic individuals are so susceptible to infection with encapsulated bacteria¹⁰².

Concurrent antigen detection by both B and T-cells alongside cellular activation in lymphoid tissue coordinated by APCs allows contact between these two lymphocyte types recognising the same antigen. T_H2 CD4⁺ T-cell activation induces IL-4 secretion supporting and accentuating the B-cell proliferation and antibody secretion. This first phase of antibody production occurs in the T-cell zones of the lymph node⁷³. This 'primary focus' only persists for a few days to produce sufficient quantities of antibody that may help clear the pathogen. During this time, some of B and T-cells may migrate to the primary lymphoid follicle and continue active proliferation around the specialised FDCs that continue to present antigen recognised by the cells compartmentalised into so-called germinal centres (GCs).

GC formation appears to fulfil multiple roles. Crucially, B-cells undergo unprecedented proliferation and a unique mechanism of Ig V region diversification occurs through the process of somatic hypermutation. This process results in a number of clones with antibodies that have a range of affinities for the antigen that the cells are specific for. These cells may differentiate to become professional ASCs producing high affinity antibody that may be more proficient at clearing or attacking the invading pathogen compared to the primary response, or alternatively they may become memory cells that enter a long-lived state ready to be activated and produce antibody rapidly following subsequent re-exposure¹⁰⁴. It is becoming increasingly recognised that the specialised T_{FH} cell is critical in supporting these antibody refining mechanisms.

1.2.2.4 *Follicular helper T-cells*

As described above, T_{FH} cell differentiation is triggered through activated DC-antigen priming with concurrent TCR, IL-6 and IL-2 signalling. Thereafter the cell migrates towards the follicle. T_{FH} cells can interact with equivalent antigen-specific B-cells either within the primary focus or in the follicle. Irrespective of location, synergistic cellular differentiation ensures GC formation and maintenance¹⁰⁵. T_{FH} express CD40L and secrete IL-4 and IL-21 which induces B-cell proliferation and somatic hypermutation. The synergistic relationship occurring in the GC requires close physical contact mediated through mechanisms including self-ligand signalling lymphocyte activation molecule F (SLAMF) structures and chemoattractant CXC-chemokine receptor 5 (CXCR5) and CXC-chemokine ligand 13 (CXCL13) expression¹⁰⁶. Other co-stimulatory survival signals involved in the T_{FH}-B-cell GC development are proposed to include OX40, programmed cell death-1 (PD-1), and inducible T-cell co-stimulator (ICOS-1) expressed on the T-cell, binding OX40L, PD-L1 and ICOSL respectively on the activated B-cell^{107,108}. PD-1 expression is proposed to deliver an inhibitory signal to the T_{FH} cell although these molecules appear to be expressed either transiently or on different subtypes or developmental stages of the T_{FH} cell.

Traditional methods used to study antigen-specific T_{FH} cells depended on intracellular staining techniques (seeking IL-21, TNF, IFN- γ or CD40L expression and production or secretion). These experiments frequently demonstrated very low numbers of cells present in the GC or led to the detection of non-T_{FH} cells owing to overlapping cytokine production profiles. A recent set

of studies have demonstrated that activated antigen-specific T_{FH} rapidly express OX40 and CD25 that together are highly sensitive and specific for T_{FH} cell presence in the GCs of individuals challenged with *Streptococcus pyogenes* or pertussis vaccine^{109,110}. Furthermore, using this technique it has been possible to detect such cells circulating in the peripheral blood of individuals as predominantly TEM-type presumably poised to support rapid GC development following re-exposure to the same antigen¹⁰⁹. Recently these cells, through marginally different methods, have been found to be correlated with protective efficacy following Ebola vaccine administration supporting their critical involvement in protective immune responses¹¹¹.

1.2.3 Vaccine-targeted immunity

The processes of immune activation following infectious agent exposure described above are all important to consider in the context of vaccination. The Holy Grail of vaccinology is to understand which of these complex immunological mechanisms are necessary and sufficient to induce lifelong protection against agent-induced infection or disease, and how to trigger these responses in all individuals whilst minimising unwanted side-effects^{15,112}. For those vaccines in current use targeting predominantly extracellular pathogens, protective mechanisms are believed to be antibody-mediated supported by CD4⁺ help. Evidence for these claims come largely from observations of protection against vaccine-preventable diseases mediated by post-exposure antibody infusion or disease susceptibility in the context of inherited immunodeficiency states selectively affecting humoral or cell-mediated pathways^{15,112}. In contrast,

protection against intracellular pathogens such as tuberculosis or malaria are believed to require a greater degree of cell-mediated immunity¹¹³.

Live vaccines are capable of inducing more sustained antibody responses, presumably as a result of improved antigen persistence. Toxoid vaccines are noted to induce highly variable responses. For example, TT induces a greater magnitude of response compared to DT¹¹⁴. However, it is not just vaccine type that influences the magnitude or efficacy of response. The properties of the vaccine preparation have been repeatedly demonstrated to have an effect on vaccine response depending on, for example, the route of administration and quality of storage of the vaccine¹¹⁵. The dose of vaccine delivered is critically important. Too low a dose may increase competition for FDC-associated antigens and result in reduced selection of high-affinity B-cells¹¹², whereas too high a dose may increase the risks of unwanted effects. Most vaccines require repeated dosing to induce effective responses. Single dosing rarely results in antibody of sufficient levels or affinity to induce protection even if protection is mediated through antibody. Repeated dosing permits repeated GC formation and recurrent attempts to refine the antibody response⁹⁹. The repeated exposure must be carefully timed, however, to ensure that the initial antibody response from earlier exposures has waned sufficiently to allow antigen to be presented effectively, hence explaining the widely-used minimum four-week interlude between vaccinations. In some cases there is evidence that this interval may optimally need to be six months¹¹⁶. Timing is equally of critical importance for some pathogens including measles where maternal IgG delivered to the infant trans-placentally may interfere with protective responses and therefore sufficient time must be allowed for maternal antibody to clear sufficiently before vaccination. A balance must be struck in

these cases to catch children before they become too susceptible to this highly contagious virus and the debate continues as to whether MV should be delivered at 6, 9 or 13 months of age¹¹⁷.

Similarly, factors integral to the individual have been shown to affect the quality of the immune response including the age of the individual and their general health¹¹⁸. Extremes of age are recognised to impair effective responses but the mechanisms underlying these observations remain unclear¹¹². There have also been multiple reports of differences in responses based on sex although this appears to occur more in response to viral vaccines such as influenza¹¹⁹ and some pneumococcal and meningococcal vaccines¹²⁰ where females induce higher antibody responses than males. Again, the mechanisms underlying these observations remain poorly understood.

It has been apparent for some time that vaccine responses can vary substantially between individuals of differing ethnic backgrounds¹²¹. These observations have led to the suggestions that human variation at the genetic level may impact responses to some vaccines.

1.3 Human genetics

Since the first description of heritable transmission of traits and disease in plants¹²² and humans¹²³, our ability to identify the underlying factors explaining this segregation in monogenic Mendelian conditions has developed alongside enhancements in technology. These successes have predominantly been driven by our appreciation of the genetic principle of linkage, the elucidation of the structure of deoxyribonucleic acid (DNA)¹²⁴, the availability of the polymerase chain

reaction¹²⁵, the ability to rapidly distinguish regions of variation in human genomes^{126,127}, and developments in statistical frameworks to test the relationships between variants and disease in sets of related individuals¹²⁸. These developments bore much fruit in terms of understanding the molecular basis of disease through genetic variant identification and correlation with biological pathways in conditions that were clearly heritable. However, such approaches could not be so readily applied to complex traits where the underlying cause was likely a result of variably penetrant genetic variation at multiple loci interacting with the environment to lead to disease in a less predictable manner. Although these conditions are more difficult to study, they are inherently more prevalent in the population and so there was substantial incentive to develop equivalent methods to apply to diseases including diabetes and inflammatory bowel disease.

1.3.1 The genetics of human populations in relation to disease

At the turn of the millennium there was a concerted and internationally coordinated shift towards testing for differences in allele frequencies between disease cases and healthy controls at the population level¹²⁹ rather than relying on linkage within families for studying complex disease. The publication of the sequence of the human genome^{130,131} demonstrated a proof of concept that sequencing of the three billion base-pair diploid genome was possible through consortia efforts. This initiative fuelled equivalent parallel efforts providing a series of crucial insights into the structure of the human genome. Firstly, variation between individuals at the genetic level is enormous. It appears that the more we look, the more variation we observe. The largest published whole-genome reference panel to date includes genomes from 32,488 individuals and lists

39,235,157 single nucleotide polymorphism (SNP) sites with an alternate allele frequency of 0.05 or greater¹³². However, this data is derived from whole genome sequence data that covers each base with relatively few reads (2-4x coverage on average) which means that there is a reasonable probability that only one chromosome of the two chromosomes in every diploid cell is sequenced at any one locus¹³³. This intermittently low coverage results in uncertainty at some SNP sites. Furthermore, these values do not take small or large structural variation into account since existing sequencing methods are ineffective at detecting such variation. Equivalent large-scale characterisation of genetic variation in the protein-coding region of the genome (that constitutes approximately 1.5% of the total coding sequence; the exome) has identified 7,404,909 variants of high certainty generated from sequence data with a depth of coverage of 10-100x¹³⁴. Given that mutations should be less tolerated in protein coding regions, the potential for genetic variation across the whole genome is appreciably large.

The second key insight from advances in genetic sequencing, equivalent to principles underlying linkage, is that variation observed at any one site is not independent of neighbouring variation. Variants occurring in close genetic proximity are more likely to be correlated with each-other although the distance over which this correlation exists varies markedly across the genome¹³⁵⁻¹³⁷. This linkage disequilibrium (LD) exists within blocks – or haplotypes – separated by well-defined hotspots of recombination. Other factors including natural selection and genetic drift also determine the correlations between variants¹³⁸. Since the variation occurs in these blocks of haplotypes, it is hypothetically not necessary to type all variants across any one individual's genome to understand their complete diversity. One may focus instead on typing smaller sets of SNPs that best capture

this variation in haplotypes of interest using the principle of tagging¹³⁹. This concept is a crucial feature of modern genetic studies. A great deal of variation is rare (less than 1% minor allele frequency) and these can arise independently of other, more common, haplotypes and therefore such methods do not apply so reliably to rare variants. Furthermore, the importance of considering the ancestral origin of individuals included in genetic studies before assuming SNPs tag the same haplotypes across populations has become increasingly clear.

The 1000 Genomes (1000G) Project has markedly improved our understanding of population diversity through undertaking sequencing and genotyping of over 2,500 individuals from 26 populations across the world^{140,141}. This project and others¹⁴² have confirmed earlier observations that populations from Africa have a larger number of haplotypes and consequently exhibit shorter spans of LD on average compared to other non-African populations. This characterisation of haplotypes in worldwide populations helps highlight potential advantages and disadvantages in including individuals of differing ancestry in studies investigating the mechanisms of disease processes as outlined in more detail below¹⁴³.

1.3.2 Genetic association studies

These advancements in our understanding of genetics have led to an explosion of literature describing the association of genetic variants with disease states through case-control studies with sample sizes approaching tens or hundreds of thousands of cases and controls¹⁴⁴. The studies involve typing variants (usually SNPs) spread across the genome and testing each SNP for association with the phenotype of interest requiring that the association surpasses a stringent significance value

(usually with a p-value $\leq 5 \times 10^{-8}$ as defined using the Bonferroni correction when accounting for the presence of 1,000,000 haplotypes across any genome). These studies have identified variants occurring in a host of related pathways that are providing mechanistic insights and identifying translational targets in the mechanisms responsible for complex diseases including Crohn's disease, psoriasis, ankylosing spondylitis¹⁴⁵, multiple sclerosis¹⁴⁶ and type 2 diabetes¹⁴⁷. Similar initiatives are being applied to biological traits that can be measured more simply or may be a single component in the diagnostic criteria of some of the diseases above. Human height, for example, has been interrogated in a study of over 700,000 individuals¹⁴⁸ and the diversity of white cell populations has been tested for association with genetics in over 170,000 individuals¹⁴⁹. The size of these initiatives allows investigators to test for evidence of replication between independent cohorts that minimises chances of observing associations due to type I error (or false positive findings). They also permit the identification of variants that may either be low in frequency (generally in the region of 1-5% minor allele frequency if the variant typing is undertaken through genotyping), or those that are estimated to have a low likelihood of effect. These observations are in contrast to those associations identified in more historical sample sets of lower number where associations had, by definition, to incur larger effect estimates. Many biological traits are measured as quantitative traits rather than as dichotomous variables that increases power to identify causative variants, not only through increased samples size, but also through the nature of the association models used¹⁵⁰.

Other methods used to increase power to identify novel signals of association in modern genetic association studies include the use of variant imputation and

leveraging genetic diversity observed between populations (or population structure). Variant imputation uses a converse approach to the observation described above of haplotypes occurring in populations tagged by SNPs, whereby the identity of untyped variants can be determined using statistical methods. The process depends on having access to a reference panel of individuals where the majority of common variants are known, such as the 1000G dataset described above. The genotyped set of variants of a test set of individuals are compared to those of the reference set and by assessing potential haplotype structures, a high quality collection of population specific variation that may be up to tens of millions of variants may be generated having only directly typed a few hundred thousand variants¹⁵¹. These variants may then all be tested for association that may uncover the presence of association signals with variants not directly typed. Undertaking such analyses in populations with a larger number of haplotypes, such as those of African descent, would potentially mean that more variants would have to be directly typed to acquire the same extent of information¹⁴³.

Whilst LD is helpful in minimising the number of SNPs required to be tested for association, it can also make identification of the causative variant challenging to identify. Many of the associations observed through the efforts described above are still to be understood in full mechanistic detail because the variant driving the association, the so-called causative variant, remains unidentified. In those cases where the signals have been characterised, subtle effects on levels of gene expression are usually found to be more relevant as opposed to the effect of protein coding change that was originally proposed to be the mechanism of importance^{152,153}. Including populations from diverse ancestries in genetic association studies can increase the likelihood of identifying the likely causative

variant. Harnessing diversity in LD and haplotype structure between populations should facilitate the identification of the causative variant if it is causative across populations since the candidate region should be narrower compared to analyses including only one population. This approach has been successful in traits including type 2 diabetes¹⁵⁴, renal disease¹⁵⁵ and adiposity¹⁵⁶. Inferring the likelihood of variants being present in some populations but not others is inherently more challenging.

1.3.3 HLA in complex disease

One region of the genome that is found associated with disease more frequently than any other is the MHC¹⁵⁷. Located on chromosome 6 this is the most gene-dense (with approximately 260 genes within 4 megabases (Mb)) and polymorphic region of the genome containing the *HLA* genes, a proportion of the complement genes and complexes of several other immune-related genes including TNF and lymphocyte antigen 6 family members¹⁵⁸. Some of the earliest genetic association studies undertaken for complex diseases including type I diabetes¹⁵⁹, coeliac disease¹⁶⁰ and rheumatoid arthritis¹⁶¹ identified the MHC region as being very important. In the vast majority of cases these signals of association have been determined to be driven by variation in the *HLA* genes and these associations have large effect sizes meaning that if carried, the risks of disease are substantially increased.

In humans, the *HLA* gene products, encoded in the MHC region of the genome, are critical components of the antigen presentation process for detection by CD8⁺ (in the case of class I *HLA* genes), and CD4⁺ T-cells (in the case of class II) as detailed in the Immunology section above. *HLA* genes are organised and named

based on their classes in man: *HLA-A*, *HLA-B* and *HLA-C* in class I; and *HLA-DPA1*, *HLA-DPB1*, *HLA-DQA1*, *HLA-DQB1* and *HLA-DRB1* in class II. Other so-called minor genes and paralogues (such as *HLA-DRB3*, *HLA-DRB4* and *HLA-DRB5*) also exist on selected *HLA-DRB1* allele carrying haplotypes making the region even more complex. The class I gene products form heterodimers with the conserved beta-2 microglobulin gene product and the *HLA-DRB1* product likewise forms a heterodimer with the relatively conserved *HLA-DRA* gene product. However, the alpha and beta subunits of the *DP* and *DQ* gene products form heterodimers that are more diverse as these subunits are each polymorphic⁷³. *HLA-B* has upwards of 2,000 distinct alleles being reported differing by at least one amino acid position that is far in excess of any other locus of the genome. This observation has led to the theories of this region being a focus of diversifying and balancing selection in order to maximise the chances of antigen detection from foreign agents across worldwide populations¹⁶². This diversity has led to challenges in nomenclature although generally only a proportion of the alleles are observed frequently within populations. Some alleles do appear more specific to different populations¹⁶³. Furthermore, some alleles appear to pass preferentially from one generation to the next as part of extended haplotypes either within the class I or II genes, or even across both classes¹⁵⁷. This results in very large haplotype blocks. The reasons for this co-inheritance affecting some alleles of *HLA* genes remain unclear.

These properties of the MHC region and the *HLA* gene region in particular provide several challenges to genetic studies. Firstly, designing probes to type SNPs or other regions of variation is particularly difficult since the region is so polymorphic and therefore ensuring conserved stretches of the 15 to 20 bases

necessary for annealing primers is challenging. Moreover, if an association signal is identified, then fine mapping the causative variant is difficult as the haplotype block may cover several kilobases (kb), if not Mb, that can include several tens of genes requiring investigators to make assumptions about the underlying causative genes. To complicate analyses even further, there are frequently observations of multiple independent associations across multiple *HLA* genes (both class I and II) that can be identified statistically using conditional regression¹⁶⁴, but there are concerns that these methods may be susceptible to increased risks of false positive associations as a result of the complex LD patterns¹⁶⁵. Therefore, more complex model testing methods have been employed instead to test these associations including the use of the Bayesian Information Criterion (BIC)^{165,166}. The high rates of polymorphisms across the MHC similarly make measuring expression of *HLA* genes challenging. Hence the expression profiles and roles of the individual gene products remain poorly understood. Any existing understanding remains limited to cellular studies with antibody staining of HLA proteins that lack the resolution of molecular-based methods.

Despite these challenges, advancements in fine-mapping signals of association are being realised for a variety of complex traits. Unsurprisingly, HLA signals are observed in genetic studies of association investigating susceptibility to infectious diseases including tuberculosis¹⁶⁷, HIV¹⁶⁴ and *Salmonella typhi*¹⁶⁸. However, pathogenic variation is likely to dilute the chances of observing genetic signals at this locus unless this diversity is accounted for. The greatest advancements in understanding HLA-mediated effects on disease are being realised for autoimmune conditions. Large genotyped case-control sample sets exist and have been used to tease out causative signals through rare allele combination, whilst

also testing for unusual patterns of inheritance or statistical interaction where adequate power can only be achieved through the availability of sample-sets numbering over 5,000 individuals^{165,166,169,170}. These studies have also implemented elegant methods to understand the allelic composition of the *HLA* genes in individuals using genotype data using imputation. Similar to genotype imputation, these methods depend on the availability of a reference panel where samples have both genotype and classical HLA allele type available¹⁷¹. These methods have been shown to be effective in diverse populations although they have not been rigorously tested in those of African descent¹⁷².

Despite the advancements in fine-mapping, the mechanisms underlying these associations in humans remain poorly understood. In the case of rheumatoid arthritis altered binding and presentation of citrullinated peptides such as those derived from vimentin and aggrecan or even particular HLA alleles themselves are believed to be responsible for the observed associations^{173,174}. For type I diabetes, HLA-DQ instability is predicted to play a role¹⁷⁵ whereas in HIV a mechanism of altered HLA-C expression is predicted to be involved¹⁷⁶.

1.3.4 Genetic association studies of vaccine responses

1.3.4.1 *Hepatitis B*

Surprisingly few large-scale studies have been completed trying to identify the effect of human genetic variation on response to vaccination, either when measured immunologically, or in terms of disease susceptibility post-vaccination. The few studies that have been undertaken have predominantly focused on HBV. This bias is for several reasons. Firstly, antibodies against HBsAg are thought to

be an adequate correlate of protection against HBV¹⁵. Furthermore, this vaccine has a rather unusual aspect of ‘non-response’ in that a proportion of individuals do not respond to the vaccine no matter how many doses they receive. However, the more doses that are administered, the more likely any one individual is to respond¹⁷⁷. Non-response is generally referred to as an HBsAg antibody level <10mIU/ml at least six weeks after the most recent (usually following three) dose of vaccine. The proportion of individuals suffering non-response are noted to vary between populations with up to 5% frequency in some Asian groups¹⁷⁸. This observation was quickly hypothesised to be the result of human genetic variation, particularly within the MHC region and a series of studies comparing ‘responders’ (generally ≥10mIU/ml) to ‘non-responders’ (<10mIU/ml) have added substantial weight to this theory¹⁷⁹⁻¹⁸⁵. There is now good evidence to support the fact that individuals carrying HLA-DRB1*03 or HLA-DRB1*07 alleles are more likely to be non-responders, compared to those carrying HLA-DRB1*01, HLA-DRB1*13 or HLA-DRB1*15 alleles. Some other studies have suggested that the association signal is, in fact, driven more by variation in the complement *C4A* gene in a neighbouring region of the MHC but this has not been replicated¹⁸¹. This idea of associations being driven by genes other than *HLA* in the MHC is, however, intriguing and complement genes have been found to explain MHC signals in numerous diseases including, most recently and notably, schizophrenia¹⁸⁶.

There have been some attempts to identify the molecular reasons for the observations of ‘non-response’ associating with MHC genes but overall the studies have been in disagreement with each other and thus inconclusive. APCs from peripheral blood mononuclear cells of non-responders have been

demonstrated to stimulate DR-matched T_H cells from responders *in vitro* as effectively as APCs from high responding individuals. These findings have been interpreted to suggest that a deficit in antigen processing and presentation is unlikely^{182,187-189}. Conversely, these studies found that the T-cells from non-responders could not be activated by either their non-responding APC counterpart or HLA-DR-matched APCs from responders. The theory most favoured on the basis of these findings is, therefore, that there is a ‘hole’ in the T-cell repertoire associated with these particular HLA alleles resulting in reduced antibody levels. Such a theory has not been supported by any other research to date and, based on our knowledge of *HLA* and TCR development and interaction, it is unclear why such a scenario would result in consistent strong genetic association signals being observed across the MHC. Similarly, the mechanisms underlying the increased responses observed with some HLA alleles has not been explained definitively but the consensus is that it occurs as a consequence of improved antigenic epitope binding¹⁹⁰.

Irrespective of the underlying mechanism some investigators have noted that a smaller proportion of variance of response to HBsAg is explained by human genetic variation in later childhood when compared to infants¹⁹¹. Such observations echo suggestions from earlier work investigating responses to TT¹⁹². Observations of vaccine response heterogeneity being less dependent on human genetics were, however, based on studies involving small numbers of individuals and since they were performed in older children they did not account for the anamnestic response. This response is a robust memory-driven antibody generation where levels of antibody return very quickly to immediate post-vaccine levels even if they have waned significantly post exposure owing to a long time

since last antigen exposure. The afore-mentioned studies had not taken these effects into account or differentiated those individuals who had waned antibody responses but could still mount a sufficient response following re-exposure therefore the conclusions require some further justification.

All of the HBsAg studies described above can be considered as candidate *HLA* gene studies. The authors proposed a strong prior hypothesis of HLA involvement and tested only for association at a number of HLA alleles. This approach is highly valid in this setting and relaxes the requirement to achieve the very small significance values necessary in genome-wide association studies ($p \leq 5 \times 10^{-8}$ as described above). Two GWAS studies have been undertaken assessing the impact of human genetic variation on HBsAg antibody responses. The first GWAS was undertaken in a cohort of 2,057 individuals in Indonesia where individuals were stratified into groups dependent on whether they induced a low (0-99.9 IU/L), medium (100-1000 IU/L) or high (>1000 IU/L) antibody response six months following a third dose of HBsAg vaccine. Using ordinal regression to test for association the authors identified three independent genome-wide significant associations across the MHC region that they subsequently replicated in 2,156 individuals; one localising to *HLA-DRB1*, one to *HLA-DPB1* and one to the so-called class III MHC region that contains complement genes including *C4A*¹⁹³. The authors did not undertake any HLA imputation and so the classical alleles underlying these associations remain unknown.

The second GWAS was undertaken in a Chinese Han population where high responders (>1000 mIU/ml after three doses of vaccine) were compared to non-responders (<10 mIU/ml after 6 doses). The investigators identified SNPs in *HLA*-

DRB1 and another gene in the MHC called *BTNL2* (encoding the Butyrophilin-Like 2 protein with a poorly defined role in T-cell activation) to be significantly associated with non-response. The authors used a complex approach of dividing individuals between discovery (total 382 individuals) and replication cohorts (two cohorts of 1,496 and 260 individuals respectively). Some direct typing of HLA was undertaken in both cases and controls (as opposed to the use of imputation) identifying HLA-DRB1*07:01 to be associated with the risk of non-response¹⁹⁴.

Despite the increasing evidence for a role of class II (and possibly class III) MHC diversity in mediating variable responses to HBsAg, there is evidence from twin studies that variants outside HLA contribute substantially to the heritability of responses. Substantial heritability (77%) for responses against HBsAg was observed in an elegant study including 207 twins from The Gambia. Accounting for HLA-DRB1 variation suggested that two thirds of the heritability was in fact due to non-DRB1 variation¹⁹⁵. Similar conclusions were drawn from an earlier twin study performed in German adults¹⁹⁶. The Gambian study also observed similar patterns for responses against TT and DT although heritability estimates were lower than those for HBsAg (44 and 49% respectively). No non-MHC associations were identified to statistical significance in the above GWAS studies although two independent candidate gene studies have proposed that variation in *FOXP1* and *ITGAL* genes may be associated with differential HBsAg responses in studies undertaken in The Gambia^{180,197} and Indonesia respectively. These findings, by definition, lie within plausible candidates but in contrast to the HLA candidate approach used above, this genome-wide approach of candidate testing is more susceptible to false positive associations and so the findings should be interpreted with caution. The associations were not validated between studies

although this lack of association could be the result of differential SNP frequencies between populations.

1.3.4.2 Other EPI vaccines

Other than the HBsAg vaccine, few other EPI vaccines have been comprehensively assessed to determine the impact of human genetic variation on vaccine responses. MV has been interrogated in multiple published studies although many of these reported findings have been derived from overlapping collections of cohorts¹⁹⁸⁻²⁰³. The predominant conclusions from these studies are that variants in both *HLA* and non-*HLA* candidate genes are associated with levels of neutralising antibody and T-cell responses against MV although many signals did not demonstrate consistent replication. Most recently, many of these cohorts have been combined into a single GWAS format demonstrating variants on chromosome 1 around the *CD46* and interferon-induced protein 44 like (*IFI44L*) genes²⁰⁴ associated with neutralising antibody levels. The *CD46* protein is an excellent candidate since it is the main protein target for cellular entry for measles virus²⁰⁵. *IFI44L* is part of the type I interferon signalling cascade and is upregulated in measles virus infection²⁰⁶. Interestingly both of these signals were also associated with the risk of seizures following vaccination with the measles-mumps-rubella (MMR) vaccine in a GWAS undertaken a cohort of over 9,000 individuals from a Danish biobank²⁰⁷, further substantiating the validity of these loci. Similar variants in *CD46* have also been found associated with risk of measles in an earlier candidate gene study undertaken in Mozambican children who had failed MV; i.e. suffered measles despite vaccination²⁰⁸. However, the mechanisms underlying the link between increased levels of neutralising

antibodies and MMR-related febrile seizures (as determined by common direction of effects at the top associated SNPs between these studies) is not yet clear.

No variants within immunoglobulin genes or in *HLA* were found to be significantly associated with neutralising antibody levels in the above-mentioned GWAS²⁰⁹. Some suggestive associations at HLA-B*57:01, HLA-DQB1*06:02 and HLA-DRB1*15:05 alleles were identified²¹⁰.

The only other genetic studies undertaken relating to other EPI vaccines include a study investigating risk of invasive *Haemophilus influenzae* disease post-vaccination²¹¹ and self-reported whooping cough (WC) across two population based studies²¹². Both of these studies were designed to test for genetic factors associated with vaccine failure. The former tested this explicitly as the small number of cases had to have documented evidence of Hib vaccination whereas it was more implied in the latter expecting that the vast majority of the large cohort had received their EPI vaccines including wP as children. The Hib failure study was designed as a candidate gene study and identified variants in the TIR domain containing adaptor protein (*TIRAP* also known as *MAL*) and *IL-10* gene associated with increased risk of Hib vaccine failure. The *TIRAP* finding is compelling given the role of the gene product in TLR1 signalling. These findings have, however, not been replicated. However, there are now so few new cases of invasive *Haemophilus influenzae* disease following the introduction of vaccination that collecting a cohort suitable for such replication is challenging thus highlighting the uniqueness of the cohort used in the analysis²¹¹. The WC study was a large-scale GWAS that found a strong signal across the class II MHC region. Despite undertaking imputation of classical HLA alleles and amino acids and finding

many strong associations the authors could not reliably identify a single allelic or amino acid signal that was capable of explaining the mechanism underlying differential susceptibility to WC²¹².

1.4 Conclusions and outline of thesis

Despite the success of modern vaccines there are still several challenges impairing our ability to design vaccines that are universally effective against all infectious agents. For example, the mechanisms underlying the faster waning of protection observed with aP compared to wP are poorly understood. Similarly, the mechanisms underlying the non-response observed with HBV vaccine are increasingly apparent to be driven by variation in HLA between individuals but understanding how we can translate these observations into improved effectiveness remains unclear. Furthermore, as molecular mechanisms associated with immunological response and memory formation following priming and boosting becomes clearer, it is apparent that there is huge scope for increasing complexity and heterogeneity of mechanisms across populations and vaccine types. This heterogeneity appears to be driven, in part, by variation at the genetic level in human vaccinees within the various components of diverse immunological pathways.

In this thesis, I attempt to advance on observations highlighted above aiming to improve our understanding of how human genetic variation impacts on responses to EPI vaccines delivered using data and samples from three cohorts of African infants. By undertaking the largest study of multiple vaccine responses undertaken in African infants I identify genetic variants determining vaccine response

heterogeneity. By undertaking further molecular and cellular assessments of these findings through a range of follow up studies I provide increased insight into the mechanisms of vaccine success and failure, especially in relation to pertussis.

Chapter two presents a general discussion of methods and concepts employed in modern genetic epidemiology studies, with particular reference to techniques used throughout this thesis. Detailed experimental and analytical methods are provided in the relevant results chapters (chapters three to five).

Chapter three presents the findings from the GWAS of EPI vaccine responses in three African populations finding novel associations with multiple vaccine antigen responses localising to the MHC region. Within this chapter I further present an assessment of loci highlighted in other published studies testing for replication of signals within my new data from Africa at the genotype level.

In chapter four I present the first HLA imputation reference panel for populations of African descent. By undertaking high-resolution classical typing of 697 individuals from across the three African populations under study I describe the frequencies of alleles in one of the largest descriptive HLA studies undertaken in Africa and use this data to infer the HLA alleles of the other unknown individuals. Using this data I present the results of the reanalysis of the GWAS data taking this imputed allele information into account and the methods used to fine-map the signals of association.

In chapter five I investigate the correlation between the HLA association signature observed for the three pertussis antigens, and susceptibility to WC in a UK population providing strong evidence of the identity of the correlate of protection

for pertussis. I demonstrate the importance of describing HLA variation at the amino acid level rather than using classical allele descriptors. I further use this description to demonstrate further correlations with publically available data of TCR expression that begins to shed light onto the molecular mechanisms involved in differential pertussis antigen antibody responses. Furthermore, I present a cellular immunology approach to characterise the impact of human genetic variation across HLA on levels of circulating T_{FH} in EPI vaccine recipients. Using a novel assay for detecting antigen specific T_{FH} in the peripheral circulation I discover that HLA variation correlates with levels of these cells specific for PT providing novel insight into the pathways involved in the differential protection afforded by the pertussis vaccine. I finally present some HLA-peptide binding results providing some evidence that peptide binding affinity may explain differential responses to DT but not for PT.

In chapter six I provide a discussion of the key findings presented in this thesis. I further discuss the limitations of the work, and the opportunities these present for future work.

Chapter 2

2 Methods

In this chapter I describe the principles of recruitment of individuals from three African cohorts and the downstream methods applied to samples acquired from the participants to identify and confirm the genetic associations observed with antibody responses against multiple vaccine antigens. These methods and the findings derived from their application form the basis for the analyses undertaken in chapters three, four and five. Further methods specific to the work undertaken in these chapters are described in more detail within each relevant chapter.

2.1 The cohort populations

Participants included in this study were recruited from the three African countries selected partly due to their geographic dispersal across the continent and partly due the availability of high quality metadata and relevant biological samples. These sites were in Uganda, South Africa and Burkina Faso. Individuals from each of the cohorts were included in this study if their dates of birth, vaccination and blood sampling were available and if it was confirmed that they had received

three doses of vaccines including DT, TT, pertussis, Hib, and HBsAg and a single dose of MV. The receipt of vaccines was confirmed through referencing the vaccination cards of infant participants. Beyond exclusion criteria involved in preliminary recruitment of the individuals, participants were not further excluded based on gender, ethnicity, HIV exposure or status or other health status. A range of clinical and demographic data was sought for the three cohorts (**Table 2.1**). What now follows is a more detailed description of each of the populations in turn.

Table 2.1: Clinical and demographic data sought from recruited participants and their parents in the three populations.

Metadata Captured
Ancestry of parents
Parental tribal affiliation
Parental occupation
Child date of birth
Child gender
Consenting metadata*
Vaccination history
Child illnesses [†]
Pregnancy details [‡]
Birth weight, length and head circumference
Lower respiratory tract infection episode frequency at six and twelve months of age
Diarrhoeal episode frequency at six and twelve months of age
Upper respiratory tract infection episode frequency at six and twelve months of age
Number of infant hospitalisation episodes
Number of co-habiting individuals at home
Breast feeding and infant formula details including ages started and stopped and age of weaning
Clinical examination findings at six and twelve months of age
Medication history
*: metadata includes: language consenting undertaken in, use of interpreter, opportunity for questioning, consent for sample storage and genetic analyses
[†] : illnesses including bleeding disorder, HIV infection or other cause of immunosuppression
[‡] : pregnancy details including gestational age, date, place and mode of delivery, and birth complications

2.1.1 Uganda: The Entebbe Mother and Baby Study

The Entebbe Mother and Baby Study (EMaBS) is a prospective birth cohort that was originally designed as a randomised control trial to test whether anthelmintic treatment during pregnancy and early infancy was associated with differential response to vaccination or incidence of infections such as pneumonia, diarrhoea or malaria (<http://emabs.lshtm.ac.uk/>)²¹³.

EMaBS originally recruited 2,356 women and 2,507 livebirths and 2,115 children were still enrolled at 1 year of age. Pregnant women in the second or third trimester were enrolled at Entebbe Hospital antenatal clinic if they were resident in the study area, planning to deliver in the hospital, willing to know their HIV status and willing to take part in the study. They were excluded if they had evidence of possible helminth-induced pathology (severe anaemia, clinically apparent liver disease, bloody diarrhoea), if the pregnancy was abnormal, or if they had already enrolled during a previous pregnancy. The mothers and infants underwent intensive surveillance during the first year of infant life and blood samples were taken and stored from both around the time of birth and at the infant age of one year. Further samples were also taken and stored annually from the infants who are currently aged between 12 and 14 years²¹⁴. All infants under follow up had a sample of whole blood collected annually on or around their birthday (2-5 ml depending on the age) that was subsequently divided into plasma and red cell pellets as described in more detail below. The national vaccination schedule is described in **Figure 2.1**. Infants were included in the present study if 1) receipt of three doses of DTwP/Hib/HBsAg and one dose of MV could be confirmed from their vaccination card 2) DNA could be extracted from stored red

cell pellets 3) plasma samples were available from the 12 month age point of sampling. Ethical approval was provided locally by the Uganda Virus Research Institute (reference GC/127/12/07/32) and Uganda National Council for Science and Technology (MV625), and in the UK by London School of Hygiene and Tropical Medicine (A340) and Oxford Tropical Research (39-12 and 42-14) Ethics Committees. The Ugandan study will hereafter be referred to as the Entebbe study.

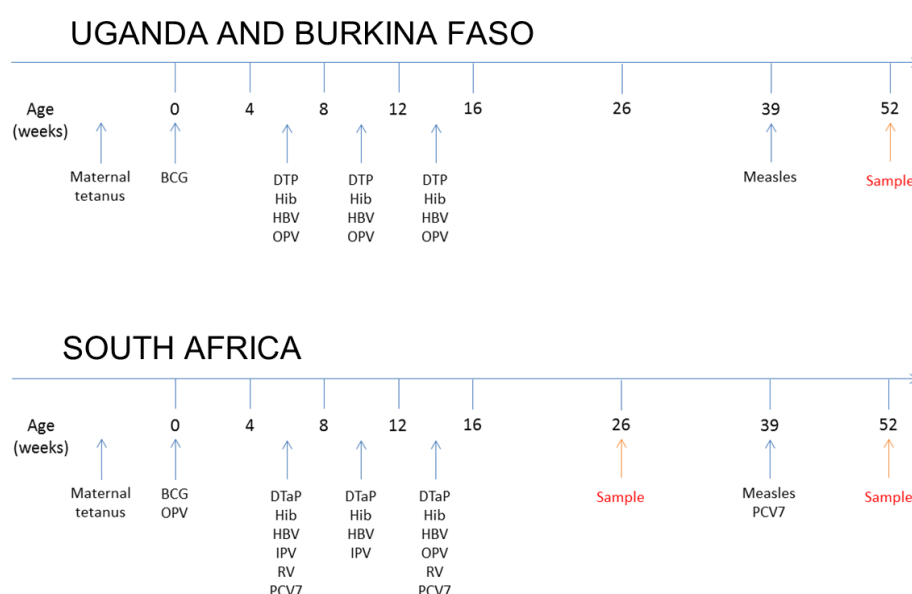


Figure 2.1: National guidelines for timings and combinations of vaccines delivered as part of the Expanded Programme on Immunisation (EPI) and approximate representation of sampling times in the respective study groups.

Timings are provided for Uganda, South Africa and Burkina Faso. BCG: bacille Calmette-Guérin; DTP: diphtheria, tetanus and whole cell pertussis; DTaP: diphtheria, tetanus and acellular pertussis; Hib: *Haemophilus influenzae* type B; HBV: hepatitis B virus vaccine; OPV: oral polio vaccine; IPV: inactivated polio vaccine; RV: rotavirus; PCV7: conjugated pneumococcal vaccine-7 serotypes.

2.1.2 South Africa: The Soweto Vaccine Response Study

The South Africa study was designed to replicate the Entebbe study. Six-month infants born in Chris Hani Baragwanath Hospital living in the Soweto region of Johannesburg, South Africa were identified from screening logs and databases of participants involved in vaccine clinical trials²¹⁵ coordinated by the Respiratory and Meningeal Pathogens Unit (<http://www.rmpru.com/>). Mothers of the infants were approached if the infants had received all of their vaccines up to six months (**Figure 2.1**). After receiving information about the study the mothers were consented in accordance with ethical approval from the University of Witwatersrand Human Research Ethics Committee (reference M130714) and the Oxford Tropical Research Ethics Committee (1042-13 and 42-14). The infants were sampled prospectively at six, and at 12 months of age following receipt of MV at nine months. Whole blood samples were collected at each visit and prepared using a similar protocol to that used in Entebbe. The South African study will hereafter be referred to as the Soweto study.

2.1.3 Burkina Faso: The VAC050 ME-TRAP Malaria Vaccine Trial

Infants between the ages of 6 and 18 months living in the Banfora region of Burkina Faso were recruited into a Phase 1/2b clinical trial to test the safety, immunogenicity and efficacy of an experimental heterologous viral-vectored (chimpanzee adenovirus 63 (ChAd63) multiple epitope-thrombospondin-related adhesion protein (ME-TRAP) and modified vaccinia Ankara (MVA) ME-TRAP prime-boost blood-stage malaria vaccine²¹⁶. It was expected that the recruited infants would receive their EPI vaccines as part of the usual national schedule

(Figure 2.1). Infants were precluded from participating in the trial if they were found to have clinical or haematological (venous haemoglobin < 8 g / dL) evidence of severe anaemia, history of allergic or neurological disease or malnutrition. Of a total of 730 infants that were recruited into the study, samples suitable for extraction of DNA were collected and stored from 400 infants (350 vaccine recipients and 50 recipients of a control rabies vaccine). Samples of plasma were available from the infants at multiple time-points following the experimental vaccine receipt. Samples from individuals taken at time points as close to the 12-month age as possible were prioritised for EPI vaccine response measurements. The infants and parents underwent intensive clinical history and examination during screening and follow-up as part of the vaccine trial but the study was not primarily designed as a genetic study and so data points differ in terms of overlap and consistency of reporting compared to the Entebbe and Soweto studies. Ethnicity information in particular was not requested during the trial and instead was gathered after the trial had concluded resulting in difficulty in collecting such data from some participants. The mothers of the participating infants provided consent for their children to be enrolled in the clinical trial and for subsequent genetic studies to be undertaken for all vaccines received in accordance with ethical approval from the Ministère de la Recherche Scientifique et de l'Innovation in Burkina Faso (reference 2014-12-151) and the Oxford Tropical Research Ethics Committee (41-12). The Burkinabe study will hereafter be referred to as the Banfora study.

2.2 Laboratory methods

2.2.1 Whole blood sample preparation

Whole blood was sampled into vacutainer tubes (BD, Becton Dickinson and Company, New Jersey, USA) containing ethylenediaminetetraacetic acid (EDTA; for the Entebbe and Soweto studies) or lithium heparin (Banfora) as an anticoagulant in the three populations. Following centrifugation the samples were separated into their constituent parts (plasma, buffy coat and red cell / erythrocyte layers) and stored at -80°C until downstream analysis in batches. DNA was extracted from the erythrocyte layer in the Entebbe study and from the buffy coat in Soweto and Banfora studies.

Whole blood was sampled into serum separator tubes (SST; BD, New Jersey USA) in the Entebbe study and serum was isolated and stored according to the recommended protocols.

2.2.2 DNA extraction and quantification

DNA was extracted from the relevant samples using Qiagen QIAamp DNA Mini or Midi Kits (Qiagen, Hilden, Germany) using recommended protocols. DNA was extracted from Entebbe samples in Entebbe, Uganda with the assistance of a single research assistant. DNA was extracted from the Soweto samples solely by technicians in Johannesburg. DNA from Banfora was extracted from samples unaided in Oxford, UK. Two methods were used to quantify all extracted DNA: Nanodrop (Thermo Fisher Scientific Inc., Waltham, Massachusetts, USA) that depends on absorption spectrophotometry of unaltered extracted DNA; and PicoGreen (Life Technologies, Carlsbad, California, USA) that uses fluorescence

detection of a compound that only fluoresces when bound within double stranded DNA. The latter method is believed to be less susceptible to over-estimation of DNA concentration due to ribonucleic acid (RNA) or degraded DNA and was used preferentially.

2.2.3 Antibody concentration measurements

Serum (Entebbe) and / or plasma (all three cohorts) were sent to the laboratory of Fiona van der Klis at the National Institute for Public Health and Environment in the Netherlands (<http://www.rivm.nl/en>) and to the Microbiology laboratory at the John Radcliffe Hospital in Oxford, UK for antibody assays.

2.2.3.1 *Multiplex immunoassay*

A validated multiplex immunoassay was used to measure antibody concentrations against DT, TT, PT, PRN, FHA, Hib polysaccharide and measles whole virus using published protocols^{217,218}. All work was undertaken by staff in Dr van der Klis' laboratory in the Netherlands. This method measures total IgG against each respective antigen and does not measure functional (e.g. neutralising) antibody concentrations. The immunoassay uses the Luminex technology (Luminex Corporation, Austin, Texas, USA) that depends on conjugation of commercially available antigens (PT from Netherlands Vaccine Institute, FHA from Kaketsuken (Kumamoto, Japan), PRN expressed and purified from an *E. coli* construct as previously described²¹⁹, DT and TT from Sigma-Aldrich (St. Louis, MO), Hib polysaccharide (Chiron, Siena, Italy) and measles virus strain Edmonston (GenWay, San Diego, CA, USA)) to fluorescently labelled carboxylated microspheres (**Figure 2.2**). The conjugation used by Dr van der Klis' laboratory

involved a two-step carbo-diimide reaction to covalently link a single antigen to a uniquely fluorescing bead-set (Bio-Rad laboratories, Hercules, CA, USA). Multiple different bead-sets with different bound antigens were incubated with the same plasma or serum sample with a phosphate buffered saline (PBS)/Tween-20/3% bovine serum albumin (Sigma, St Louis, MO, USA), mixture simultaneously to allow the binding of any antibody present in the medium whilst minimising background in a manner similar to a monoplex solid-phase enzyme-linked immunosorbent assay (ELISA). The bead-antigen-antibody complexes were then separated from remaining plasma or serum through the use of a vacuum manifold before washing with PBS and incubating with an anti-human IgG antibody conjugated to R-phycoerythrin (R-PE from Jackson ImmunoResearch Laboratories Inc. (Westgrove, PA, USA)), and washing again prior to detection in the Luminex flow cytometer. The cytometer depends on two lasers; the first detects the identity of the fluorescently labelled bead (and therefore antigen bound), and the second detects the fluorescence intensity of R-PE (related to the concentration of primary antibody in solution) bound to each bead passing through the detection channel²¹⁸.

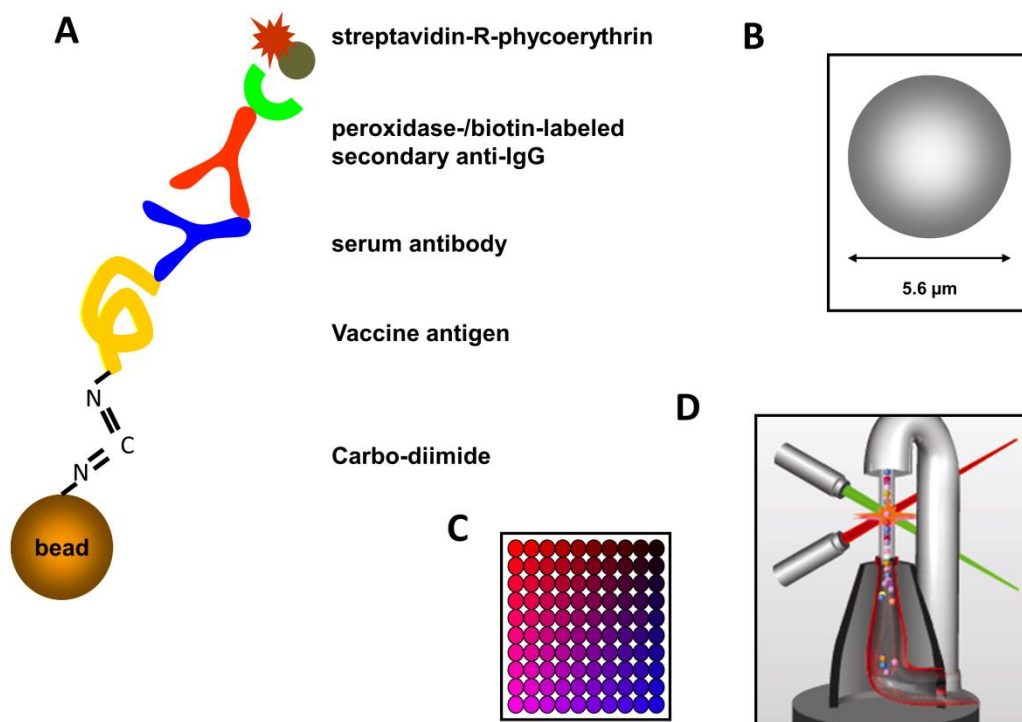


Figure 2.2: Concepts of the Luminex assay.

(A) Antigen (yellow) is covalently bound to a fluorescently labelled microsphere (B) that can have primary antibody (blue) bound before detection with a R-PE labelled secondary antibody. In an ELISA format the antigen would instead be complexed to the side of a well. (C) a range of fluorescently labelled beads are available that can each have different antigens bound that are then detected passing through a dual-laser flow cytometer (D). Some images modified, with permission, courtesy of T. Waterboer, Heidelberg, Germany.

The final concentration of bound antibody was calculated by determining the median fluorescence intensity of the antigen-specific beads using diluted standards to convert measurements into concentrations in international units standardised for each antigen. This assay has been used and validated by multiple groups although the method of antigen production or sourcing and covalent

linkage to the bead may vary^{220,221}. ELISA results were available for MV and TT antibody responses from a subset of the Entebbe participants as performed as part of the early investigation undertaken in the EMaBS cohort²¹³.

2.2.3.2 *HBsAg ELISA and other vaccine responses*

HBsAg responses were measured by staff in the John Radcliffe Microbiology department using the anti-HBs kit on the ABBOTT Architect i2000 using recommended protocols (Abbott Laboratories, Chicago IL, USA).

Levels of measles and tetanus-specific IgG were available from the original EMaBS protocol as measured by commercial ELISA kits (Enzygnost, Germany) performed using blood sampled from the mother at delivery and the infants at their first birthday.

Antibody responses against polio were initially prioritised for measurement as part of this proposed project but methods for quantifying these at the time of project commencement were unsustainable with regards to consumption of biological material (requiring >150µL plasma or serum per individual serotypes 1-3) and hence measurements for polio were not included in this study. Similarly, methods for measuring responses against the heptavalent pneumococcal and rotavirus vaccines were equally consuming and given that these were only administered in the SA cohort with no potential for replication these assays were not prioritised. Furthermore, as discussed in chapter one, the correlate of protection for tuberculosis through BCG administration is unknown and so no measurements of immunogenicity were undertaken for BCG.

2.2.4 Genome-wide genotyping

Genotyping of SNPs at approximately 2,500,000 sites across the genome was undertaken using the Illumina HumanOmni 2.5M-8 ('octo') BeadChip array version 1.1 (Illumina Inc., San Diego, USA), performed by the Genotyping Core facilities at the Wellcome Trust Sanger Institute (WTSI). Genomic DNA underwent whole genome amplification and fragmentation before hybridisation to locus specific oligonucleotides bound to 3µm diameter silica beads. Fragments were extended by single base extension to interrogate the variant by incorporating a labelled nucleotide enabling a two-colour detection²²². Genotypes were called from intensities using two clustering algorithms (Illuminus and GenCall) in GenomeStudio (Illumina Inc., San Diego, USA) incorporating data from proprietary pre-determined genotypes. The Illuminus calls were used preferentially owing to higher overall call rates.

2.3 Analytical methods

The bioinformatic processes applied to the variant data generated as part of this project is described below. All base pair positions provided in this thesis refer to the Genome Reference Consortium Human Build 37 (Build 37, February 2009).

2.3.1 Statistical power

Statistical power is the probability of rejecting the null hypothesis (H_0) when the alternative hypothesis (H_1) is true, or alternatively the probability that we do not make a type II error (that is a false negative). Calculating the power of studies to detect genetic associations with quantitative traits depends on calculating the

noncentrality parameter (NCP) that represents the degree to which the mean of the distribution of the trait in the tested sample departs from the mean when the null is true¹⁵⁰. The NCP (ϖ) in a random population sample is given by:

$$\varpi = N \times \frac{\beta^2 \text{Var}(x)}{\text{Residual variance of } y}$$

Where y is the measured trait and x is the dosage of an allele at a genetic locus and β is the effect estimate of y on x . Using the assumption that genetic loci in polygenic traits do not explain a significant proportion of the overall variance of the trait, the NCP can be given by the proportion of trait variance explained by the quantitative trait locus (V_A) multiplied by N , the sample size. Error may be induced through either measurement of the trait or imperfect capture of the associated variant through LD or failed typing. Power calculations were undertaken using the Genetic Power Calculator (<http://zzz.bwh.harvard.edu/gpc/>)²²³.

2.3.2 Genotype quality control

Errors in genotype calling have the potential to systematically introduce bias into genetic association analysis potentially leading to false-positive and false-negative associations. The large number of SNPs typed on modern arrays increases the probability of observing these errors due to chance but they also offer the opportunity to identify these errors through increased power²²⁴.

The quality control (QC) of the genotype variant data for the three cohorts studied in this thesis was undertaken in each cohort separately. The steps used were identical for each cohort and were undertaken using either standard Unix

commands, PLINK (version 1.7)²²⁵ or EIGENSOFT (version 4.2)²²⁶. QC pipelines were validated and tested on independent Ugandan cohorts by Tommy Carstensen and Deepti Gurdasani from Dr Manjinder Sandhu's group at the WTSI¹⁴² before being applied independently to the cohorts studied as part of this work. Deviations from the orders of steps in this pipeline were found to potentially alter the final number of SNPs or individuals retained for any one population.

2.3.2.1 Variant mapping

Genotype calls were provided by WTSI Genotyping Core facilities in binary PLINK format based on data available from Illumina Inc. aligned to Build 37. A list of probes used to generate the calls was available for every typed SNP. To ensure the data was consistent with Build 37, files generated by Dr Will Rayner from the Wellcome Trust Centre for Human Genetics were used to determine whether SNPs called using the probes defined by Illumina Inc. would be expected to align with the Build 37 genome and whether the coordinates were consistent with those supplied (<http://www.well.ox.ac.uk/~wrayner/strand/>). Using these files, SNPs that were predicted to map to multiple regions in the genome or those that were predicted to not map to the genome were removed²²⁷. Furthermore, any duplicate SNPs or those with non-informative variant calls were removed.

2.3.2.2 Sample quality control per individual

Call rates of all SNPs in all individuals were calculated in PLINK and following inspection of missingness distributions, samples with less than 97% of SNPs called were excluded. Heterozygosity of SNPs was also calculated in PLINK using the following equation:

$$H_o = \frac{(N - O)}{N}$$

Where H_o is the observed heterozygosity, N is the number of non-missing genotypes and O is the number of observed heterozygous loci. To identify individuals with excessive rates (potentially indicating sample cross-contamination) or reduced levels (suggesting excessive inbreeding potentially biasing association analyses) of heterozygosity, the mean and standard deviation measured within each population was calculated and any individual lying outside of three standard deviations were excluded²²⁴. To identify samples that may have been mislabelled, the sex of each individual was calculated in PLINK using homozygosity rates across non-pseudoautosomal SNPs on the X chromosome and compared to those expected based on reported sex. Due to the presence of a single X chromosome in males, they would be expected to have homozygosity rates of 1, compared to females, who would be expected to have rates less than 0.2. Individuals reported as male with X chromosome homozygosity values < 0.2 and females > 0.8 were excluded.

SNPs called in less than 97% of the remaining individuals were then excluded before calculating the identity-by-descent (IBD). Genetic association studies classically depend on including samples from unrelated individuals since, otherwise, genotypes within families may be overrepresented biasing estimates across populations²²⁴. To identify duplicate and related individuals, identity by state (IBS) may be calculated for every pair of individuals (within each population) based on the average proportion of alleles shared at genotyped SNPs. PLINK was used to calculate identity-by-descent (IBD) across all tested loci by calculating the probabilities that zero ($P(IBD = 0)$), one ($P(IBD = 1)$) or two

($P(IBD = 2)$) alleles were IBD. The proportion of alleles with shared IBD is thus calculated:

$$\hat{\pi} = P(IBD = 2) + \frac{P(IBD = 1)}{2}$$

Identical twins would therefore be expected to have $\hat{\pi}$ values near to 1, siblings in the region of 0.5 and first cousins in the region of 0.125²²⁸. In the association model used in this thesis, relatedness was explicitly accounted for and so these calculations were only used to identify samples that were likely duplicates or cross-contaminants (that would also have elevated estimates of heterozygosity). PLINK was used to calculate IBD using regions of the chromosome not recognised to be under extended ranges of LD²²⁹ and after pruning SNPs by LD. Pruning was performed in windows of 50 SNPs removing SNPs correlated with a coefficient of determination (r^2) more than 0.5, sliding 5 SNPs sequentially. Samples with $\hat{\pi}$ greater than 0.9 who were not known to be twins were excluded. Only one individual from each related pair was excluded, keeping the individual with the lowest SNP missingness rates.

2.3.2.3 *Sample quality control per marker*

The probability that each of the SNPs deviated from Hardy-Weinberg equilibrium (HWE) was calculated²³⁰. This equation states that the amount of genetic variation in a population remains constant from one generation to the next in absence of disturbing features and as such the expected genotype frequencies at a given locus can be represented as:

$$p^2 + 2pq + q^2 = 1$$

Where p and q are the allele frequencies at a bi-allelic locus. Departures from this equilibrium can be tested using the Pearson's χ^2 test with one degree of freedom:

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

Where O_i are observed genotype counts and E_i are expected counts. Departures from HWE may be observed in situations of migration, non-random mating, genetic drift or selection. It was hypothesised that any SNPs deviating from HWE may be susceptible to genotyping error or may increase bias. Due to the large number of tests performed and the likelihood of identifying statistically significant deviation from HWE due to chance any SNPs deviating from HWE with a $p < 1 \times 10^{-8}$ were excluded. Only individuals defined as founders ($\hat{\pi} < 0.05$) were used for these calculations. Recently it has been demonstrated that association analysis is robust to most but the most extreme of those deviating from HWE. These can be identified by defining a very stringent cut-off ($p < 1 \times 10^{-50}$)²³¹. This approach was not applied in this thesis.

Finally, only SNPs with a minor allele frequency (MAF) greater than 1% were included in downstream analyses to avoid including SNPs that may have been called as heterozygous by chance in clustering algorithms and to reduce the risk of observing false positive associations driven by a small number of individuals with large effect sizes²²⁴.

SNPs on the X chromosome were also passed through similar QC but HWE and MAF were only calculated in female founders.

2.3.2.4 *Population structure*

GWAS studies can be confounded by population stratification where systematic ancestry differences may correlate with the measured phenotype. Newer association study modelling approaches can account for population structure and cryptic relatedness but it is still important to identify and exclude major population outliers who may induce spurious results perhaps resulting from low quality data passing other tests during QC²³². Population structure can be modelled using principal components analysis (PCA) of genome-wide genotype data²²⁶. Genotypes are defined as a matrix of genotypes across all individuals and independent linear variables (or principal components – PCs) are defined through multivariate modelling. Those PCs explaining the greatest proportion of variance are ordered first and frequently explain population structure. However, other effects may also induce variation, such as batch effects and so it is important to test for these concurrently. This can be done by merging datasets with data from populations equivalent to those in question and ensuring the PCs overlap as expected based on self-reported ethnicity¹⁴². PCs in the genetic analyses presented in this thesis were calculated using the EIGENSTRAT software in EIGENSOFT calculated only using founders (IBD < 0.05) and using only pruned SNPs that were not in regions of long-range LD. PCs were calculated within populations and merged with other African-specific datasets including the African Genome Variation Project (AGVP)¹⁴² and 1000G¹⁴¹.

2.3.2.5 *Genotype imputation*

A major goal of the work in this thesis was to identify and fine map the likely causative genetic variants associated with vaccine traits common across diverse

populations. In order to increase the power to identify such variants, the identity of missing variants were inferred through genotype imputation. Pre-phasing of the directly typed genotype data (i.e. probabilistic determination of the chromosomal positioning and ordering of alleles) was undertaken to generate likely haplotypes. Subsequently, missing variants were imputed into these haplotypes using allele frequencies and haplotypes determined from independent reference datasets²³³. Both steps required a suitable reference dataset representative of the populations under study.

A combined reference dataset comprising low-coverage (4x) whole genome sequence data from 4,802 individuals of varied descent were used for both phasing and imputation. This dataset included individuals sequenced during phase 3 of 1000G (n = 2,504, European, Asian, African and South American descent)¹⁴¹, 320 individuals from AGVP (derived from Ethiopia, Somalia, South Africa (Zulu) and Uganda)¹⁴², and a further 1,978 individuals of diverse ethnicity from rural Uganda (Ug2G). The Ug2G dataset has not yet been published but was generated by Dr Manjinder Sandhu's group at the WTSI. These complete reference datasets were designed and built through phasing of whole genome sequence reads and merged by the same group.

A variety of software packages offer phasing functions. Those determined to be most accurate depend on a hidden Markov model (HMM), modelling sequential ordered genotypes as an imperfect mosaic of the true haplotypes determined from the reference dataset²³⁴. The algorithm underlying the segmented haplotype estimation and iteration tool (SHAPEIT) has been shown to offer superiority both in terms of the number of reference haplotypes it can assess, and its computational

performance scaling linearly rather than quadratically with the number of reference haplotypes. It has therefore become the gold standard for phasing although new algorithms continue to be developed in the face of genotype datasets numbering hundreds of thousands of individuals²³⁵. Similarly, multiple software packages are available to perform imputation once the genotypes are phased. IMPUTE2 offers enhanced computational performance though assessing haploid chromosomes rather than diploid^{151,233,236}. IMPUTE2 is streamlined with SHAPEIT and therefore this is currently the preferred pipeline for imputation. These algorithms have been tested using diverse reference panels and using genotype data from populations of African descent and have shown to be robust. Furthermore, the inclusion of multiple individuals of diverse ancestry in the reference has been shown to improve imputation accuracy²³⁷. This observation has encouraged the use of over 60,000 samples in some of the newer reference panels¹³².

The imputation and pre-phasing were both done within populations separately. Pre-phasing with SHAPEIT2 (version 2.790) was performed using standard settings including an effective population size of 17,469 as recommended in the software for African populations. IMPUTE2 (version 2.3.2) was used for imputation using recommended settings and imputation performance was assessed using the mean of the imputation metric outputted from the IMPUTE2 software (r^2 of 'type 0' SNPs).

The final output from imputation performed using these two algorithms was a list of probabilities for all genotyped and imputed variants inferring whether each individual is homozygous or heterozygous for that variant. The final number of

variants will depend on the number available in the reference panel. In addition to using these probabilities in assessing the relevance of the SNP in association with the tested phenotypes, the imputation software outputs a number of metrics that may be used to assess the performance of the imputation. The post-imputation information score (info metric) represents, for any imputed SNP, the relative statistical information about the allele frequency compared to directly typed SNPs in the dataset. Values range between 0 where there is little confidence in SNP genotypes, to 1 where there is maximal confidence. A cut-off was selected to exclude SNPs with an info metric less than 0.3 in the present study. A further measure of performance is available in the IMPUTE2 output in the form of the correlation between directly observed genotypes and the imputed genotypes at these same sites when the true genotypes are masked (r^2 of 'type 0' SNPs). The performance of imputation is well recognised to be impaired with low frequency SNPs since these are poorly tagged by common variants and therefore only SNPs with $MAF > 0.01$ were included in downstream analyses²³⁶.

2.3.3 Antibody phenotype normalisation and covariate modelling

All of the available antibody traits were normalised using both logarithmic and inverse normalisation in R (using the `qqnorm()` function). In order to determine the most appropriate covariates to control for in the association regression modelling analysis, normalised traits were manually tested for association with a variety of available metadata endpoints. Including covariates correlating strongly with quantitative traits has been shown to significantly increase the power to identify genetic association signals²³⁸. Tested endpoints included time between vaccination and sampling, sex, age, number of illnesses, socio-economic status,

HIV status (if known) and asymptomatic malaria parasitaemia. Not all variables were available for all populations. Further exploration of variables affecting vaccine responses was undertaken using step-wise forward and reverse model testing calculating Akaike's information criterion (AIC) to test between models. Automated regression modelling was performed using the step() function in R.

2.3.4 Association testing

Most modern methods of testing association between genotyped or imputed genetic variants and relevant traits depend on regression modelling. Such modelling methods offer benefits over other association tests in that one can include covariates to protect against spurious associations generated through confounding (classically driven by population structure but other causes are common and trait specific), and also to increase power to see true associations through controlling for non-confounding covariates as described above²³⁸. Regression modelling also allows the testing of more complex models investigating non-independence between SNPs that may be in LD or testing for interaction between loci²³⁹. Quantitative traits are tested for association through linear regression which depends on the assumption that the traits are normally distributed. If traits are not normally distributed then they may be transformed to approximate normality using log or inverse normal transformation as further discussed above. Linear regression consists of finding the best fitting linear relationship between the trait defined as the measured or outcome variable (y , that can be defined as a vector of values across multiple individuals), and the genotypes as explanatory variables (x , that may equally be a vector) usually coded

as additive dosages of the minor or rare allele. The equation of the regression line is:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Where β_0 is the intercept through the y axis (when $x = 0$), β_1 is the slope of the line referring to the effect estimate of the SNP on the trait (**Figure 2.3**) and ε is the residual error after fitting the model that may include other fixed effect covariates each with their own effect on the outcome variable. Estimating β_0 and β_1 is derived from the least squares method.

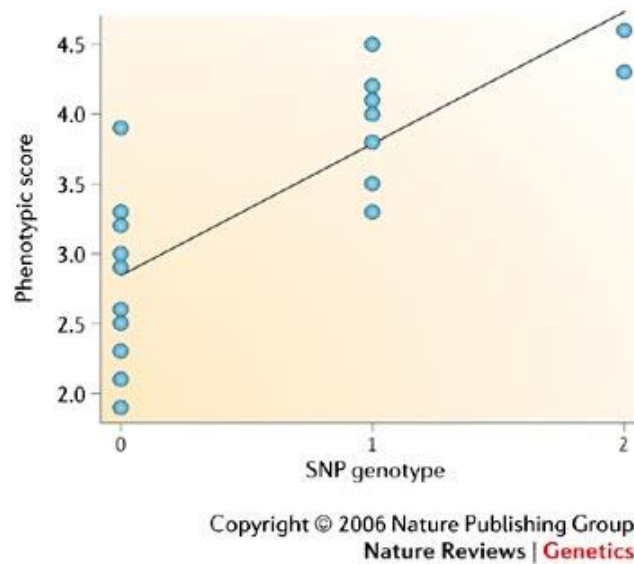


Figure 2.3: Regression line through continuous normally distributed data that is associated with the SNP genotype.

The SNP is coded as a series of additive dosages of the rare variant. In this example β_1 will be positive and the variant increases the likelihood of the trait. Reproduced with permission from Nature Publishing Group (License no. 4171380766135)²³⁹

For binary traits such as case-control analyses the trait is not normally distributed and so a generalised linear model is fitted with the *logit* function that transforms the risk of an individual developing the outcome (π) into the log odds. The *logit* function takes the form:

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1 - \pi}\right)$$

The log odds can then be fitted to the same linear predictor as in linear regression described above.

Traditionally, controlling for population structure and cryptic relatedness in GWAS has involved, firstly, excluding individuals with IBD > 0.1875 followed by calculation and inclusion of anywhere between 3 and ten PCs to control for population structure. Individuals who are obvious outliers based on PC analysis would first be removed and often, if structure is apparent, association analyses will be undertaken within populations defined by PC categorisation. Each of these steps reduces power and could therefore conversely increase the risks of observing false positive associations. Newer methods aim to explicitly account for both population structure and cryptic relatedness. Mixed linear models (MLM) also termed linear mixed models (LMM used hereafter) are becoming the method of choice for such simultaneous analyses and depend on building a genetic relatedness matrix (GRM) that models genome-wide sample structure²⁴⁰. Algorithms that can undertake such association analyses estimate the contribution of the GRM to phenotypic variance when treated as a random effect before reporting the association statistics. The models are termed as ‘mixed’ because the

model may also include other covariates that are treated as conventional fixed effects. The LMM model therefore becomes:

$$y = C\beta_C + x^{(l)}\beta_l + \vartheta + \varepsilon$$

Where C is a matrix containing all of the fixed effect covariates and the intercept with an associated effect (β_C) and β_l is the effect of SNP $x^{(l)}$ and ϑ is a random effect modelling the GRM $\mathbf{R}$:

$$\vartheta | (\eta, \sigma^2) \sim N(0, \eta, \sigma^2 \mathbf{R})$$

Early software restrictions made calculating the effects through this model unfeasible but with the use of modern genotyping arrays with very dense genotype data permitting fast and effective calculation of the GRM, combined with the assumption that the contribution of SNPs to a polygenic trait are negligible, $\mathbf{R}$ remains constant making computation accessible^{241,242}. There has been a recent explosion of algorithms providing a LMM option but few early programs offered the opportunity to include dosages derived from imputation probabilities. This option was available in the software GEMMA^{243,244} and so to include imputation probabilities alongside maximising power through including structured and related individuals all association analyses in this thesis were undertaken using GEMMA (version 0.94).

2.3.5 Quality control following association analysis

Genome-wide association results are classically inspected using a quantile-quantile (QQ) and Manhattan plots. The Manhattan plot demonstrates the statistical significance of each SNP with each tested trait on the vertical axis

designated as the negative logarithm of the p-value, with SNPs ordered by chromosomal base pair position along the genome on the x-axis. Associations with small p-values will therefore be seen as peaks with index SNPs at the ‘top’ and surrounding SNPs with lower significance in LD with the index SNP below. QQ plots demonstrate the ordered distribution of p-values compared to an expected distribution determined under the null hypothesis (H_0) generated from a χ^2 distribution with one degree of freedom. The genomic inflation factor is calculated as the ratio of the median observed χ^2 test statistic over the expected. Inflation factor values over 1.05 are usually considered to reflect some confounding variation inducing bias, typically caused by population stratification, although high values may also be observed when many variants are truly associated in either single or multiple loci across the genome. Repeating the calculations excluding such associated regions permits a re-calculation of the inflation factor for the remainder of the genome²⁴⁵. Manhattan and QQ plots were generated in the R statistical software package using the ‘qqman’²²⁴ package or using customised R scripts with the ‘ggbio’ package²⁴⁶.

2.3.6 Meta-analysis

To identify variants associated with traits common across populations, the evidence of association for each variant was combined using fixed effects (FE) meta-analysis. Random effects (RE) meta-analysis is recommended in situations where there may be heterogeneity of effect between populations but is notably conservative and FE therefore remains the method of choice²⁴⁷. FE meta-analysis assigns a weight to individual studies determined through inverse weighting. The effect sizes (θ) from i studies with variance (σ^2) may be pooled using:

$$\theta_{meta} = \frac{\sum \frac{\theta_i}{\sigma_i^2}}{\sum \frac{1}{\sigma_i^2}}$$

And the pooled variance may be calculated using:

$$\sigma_{meta}^2 = \frac{1}{\sum \frac{1}{\sigma_i^2}}$$

FE meta-analyses were undertaken using the METASOFT software²⁴⁸ and the individual study effect estimates and errors were plotted as Forest plots generated in R.

2.3.7 Heterogeneity assessment

An assessment of heterogeneity of association statistics between study populations is important in order to understand and attempt to identify the underlying reasons for the observed heterogeneity that may be biologically meaningful. It also helps inform the appropriate method for comparing and combining association statistics through meta-analysis²⁴⁷. The Q statistic²⁴⁹ has been proposed as a measure to test for heterogeneity and is given by:

$$Q = \sum \frac{(\theta_i - \theta_{meta})^2}{\sigma_i^2}$$

Using the same notations as for the inverse variance method described above. The Q statistic follows a χ^2 distribution. The I^2 index²⁵⁰ has furthermore been used to quantify the extent of heterogeneity and is given by:

$$I^2 = \begin{cases} \frac{Q - (i - 1)}{Q} & \text{for } Q > (i - 1) \\ 0 & \text{for } Q \leq (i - 1) \end{cases}$$

2.3.8 Multi-ethnic meta-analysis

A Bayesian multi-ethnic meta-analysis approach (MANTRA) has recently been developed in order to facilitate fine-mapping of causal variants using populations of varying ancestry²⁵¹. This method tests whether effect estimates of variants with common traits tested across populations of differing ancestry are consistent, taking genetic differences between the populations into account. The method defines the genetic distance (consistent with F_{ST}) between populations before undertaking the meta-analysis. This approach has the advantage of allowing for heterogeneity in effect estimates between population groups allowing for the differential patterns of linkage disequilibrium observed between them, assuming a shared underlying causal variant²⁵². A substantial benefit to using this analysis for meta-analysis is the opportunity to define a 99% likely credible interval (similar to a frequentist-based confidence interval) for the ‘causative’ variant. The boundaries of the credible intervals will depend on the presence or absence of signals of association with each variant and the degree of LD between variants. In this thesis a 1Mb region of association (500kb either side of the top associated variant for each phenotype) was determined before calculating the posterior probability that the j^{th} SNP was causal among those reported in the meta-analysis:

$$\varphi_j = \frac{BF_j}{\sum_k BF_k}$$

Where BF_j donates the Bayes' Factor (BF) in favour of the j^{th} SNP and the denominator is the summation of BF s for all other variants in the region. The 99% credible set was then constructed by ranking all SNPs according to their BF and combining ranked SNPs until their cumulative probability exceeded 0.99.

Chapter 3

3 Genome-Wide Association Study of Multiple Vaccine Antibody Responses in Genetically Diverse African Infants

3.1 Background

Individuals respond variably to vaccination and, as a result, no vaccine is 100% effective at preventing colonisation, infection or disease across multiple populations. There is increasing evidence that variation in vaccine response, measured either immunologically or clinically, can be explained by both host genetic and environmental variation^{193-196,204,211}. However, a description and understanding of the mechanisms underlying this variation has not been prioritised. An improved understanding of the processes underlying vaccine response heterogeneity should facilitate the development of universally effective vaccines including against pathogens that are more challenging to implement such

as those targeting malaria and HIV¹¹. Modern large-scale genetic association methods permit a characterisation of how host genetics influence traits across populations.

This chapter describes the results of a GWAS combining data from whole-genome genotyping and immunological responses against multiple vaccines delivered routinely to infants. The studies were undertaken in African infants firstly in order to ensure that any findings could be readily translated to those populations with substantial risk of future infectious exposure, and secondly in order to facilitate fine-mapping of the causative variants since populations in Africa are known to be some of the most genetically varied populations in the world¹⁴². The quality control and population diversity of the genetic data and the effect of other factors on vaccine responses in the three cohorts are presented first. Thereafter the results of the GWAS are presented for each cohort independently including imputed variants before undertaking meta-analyses identifying novel genetic signals of association with multiple vaccine antigen responses.

3.2 Cohorts and study design

This study was designed in order to be sufficiently powered to identify genetic variants associated with multiple vaccine responses common across three diverse African populations. Although diversity across the populations was expected, the study proposed here aimed to identify specific signals that were predicted to demonstrate minimal heterogeneity between populations. On this basis, and assuming that the causative variant itself or a variant in close LD would be typed or accurately imputed and that the trait would follow an additive model of

inheritance, a sample size of 2,500 individuals was determined to have 94.7% power to identify variants explaining 2% of the variance with a p-value threshold of observing an association due to chance (α) of 1×10^{-8} (the lowest available in the Genetic Power Calculator software²²³). This calculation assumed that we would include the causal variant, which we considered to be likely, given our imputation approach. The same sample size would provide 99.8% power to identify variants explaining over 3% of the variance. HLA alleles have previously been shown to explain 2.5 to 7.2% of the variance of antibody traits in population studies²⁵³ and some non-HLA variants affecting height have been demonstrated to explain up to 2% of variance²⁵⁴ increasing the likelihood that this proposed study would identify relevant SNPs driving at least some of the observed variation in vaccine responses.

Whole-genome genotyping was undertaken in a total of 2,499 infants recruited from cohorts in Entebbe, Uganda (n = 1,391); Soweto, South Africa (n = 755); and Banfora, Burkina Faso (n = 353) using the HumanOmni 2.5M-8 BeadChip array. A summary of the descriptive characteristics of individuals included from each cohort is provided in **Table 3.1**.

Table 3.1: Descriptive characteristics of the individuals recruited into the study from three African populations.

	Entebbe	Soweto	Banfora
Final number	1,391	755	353
Number of Males (%)	706 (50.8)	384 (50.1)	177 (50.1)
Number exposed to HIV (%)	130 (9.3)	431 (57.1)	NA
Number infected with HIV (%)	19 (1.3)	1 (0.1)	NA
Mean weeks of age at serum/plasma sampling (s.d.)	54.1 (3.9)	26.4 (3.7)	87.4 (23.4)
Number of twin / triplet sets (%)	15 (1.1)	25 (3.3)	0 (0)
<u>Maternal Ethnicity</u>			
Baganda number (%)	707 (50.8)		
Banyankole number (%)	131 (9.4)		
Bunyarwanda number (%)	76 (5.5)		
Busoga number (%)	40 (2.8)		
Batooro number (%)	65 (4.7)		
Luo number (%)	78 (5.6)		
Other Ugandan number (%)	294 (21.1)		
Zulu number (%)			278 (36.8)
Sotho number (%)			186 (24.6)
Xhosa number (%)			112 (14.8)
Tsonga number (%)			54 (7.2)
Tswana number (%)			51 (6.8)
Venda number (%)			22 (1.6)
Other South African number (%)			72 (5.2)
Gouin number (%)			167 (47.3)
Karaboro number (%)			65 (18.4)
Turka number (%)			19 (5.4)
Peulh number (%)			6 (1.7)
Other Burkinabe number (%)			8 (2.3)
Unknown number (%)	0	2 (0.3)	89 (25.2)

s.d.: standard deviation; NA: not available

3.3 Genotype data quality control

A total of 2,391,739 genetic variants were genotyped. A mean of 2,184,676 SNPs across the three populations remained following QC (**Table 3.2**).

Table 3.2: Summary of genotype QC steps for each population. The number of individuals and variants removed within each population during sequential steps of QC.

		Entebbe	Soweto	Banfora
Pre-QC	Total number sent for genotyping	1,416	784	360
	Number of variants typed and mapping to Build 37	2,328,340	2,328,340	2,328,340
Individual and autosomal variant QC	Individuals failing call rate (<97%)	2	2	1
	Individuals with extreme heterozygosity (>3s.d. around the mean)	13	13	5
	Individuals failing sex-check	2	2	1
	Individuals IBD > 0.9 and not twins / triplets	8	12	0
	Autosomal variants with call rate <97%	73,540	66,464	79,400
	Autosomal variants in Hardy-Weinberg disequilibrium ($p < 10^{-8}$)	27,002	19,629	11,647
Chromosome X QC	Chromosome X variants with call rate <97%	1,552	1,244	1,599
	Chromosome X variants in Hardy-Weinberg disequilibrium ($p < 10^{-8}$)	45	29	8
Post-QC	Total individuals remaining	1,391	755	353
	Total autosomal variants remaining	2,176,695	2,191,144	2,186,190
	Total chromosome X variants remaining	45,725	47,292	46,868

s.d.: standard deviation

3.4 Population structure within populations

The populations studied here are under-represented in existing published GWAS studies. Given the African descent of all participants, it was deemed likely that significant population structure would be observed within each sample set. Genetic principal components were calculated and correlated with reported ethnicity (according to the mother's reported tribal affiliation) demonstrating a small degree of structure in all three populations (**Figure 3.1**). Although small in number, individuals with reported maternal Peulh ethnicity from Banfora were observed to cluster away from the other ethnic groups along PC1. These individuals represent part of the Fulani community who live across West Africa. Such individuals have frequently been noted to be genetically different and cluster away from other ethnic groups, especially noted in studies from The Gambia investigating malaria susceptibility²⁵⁵. The Luo group in the Entebbe population is a Nilotic group that similarly clusters away from other groups along PC1. Otherwise, most individuals clustered together suggesting either that either they were more genetically similar than expected (being largely of Bantu descent) or that there was a substantial degree of admixture occurring between groups in the predominantly urban settings where recruitment for these studies took place, or that there was a combination of both of these effects. Accordingly, little stratification was observed in PC3 and PC4.

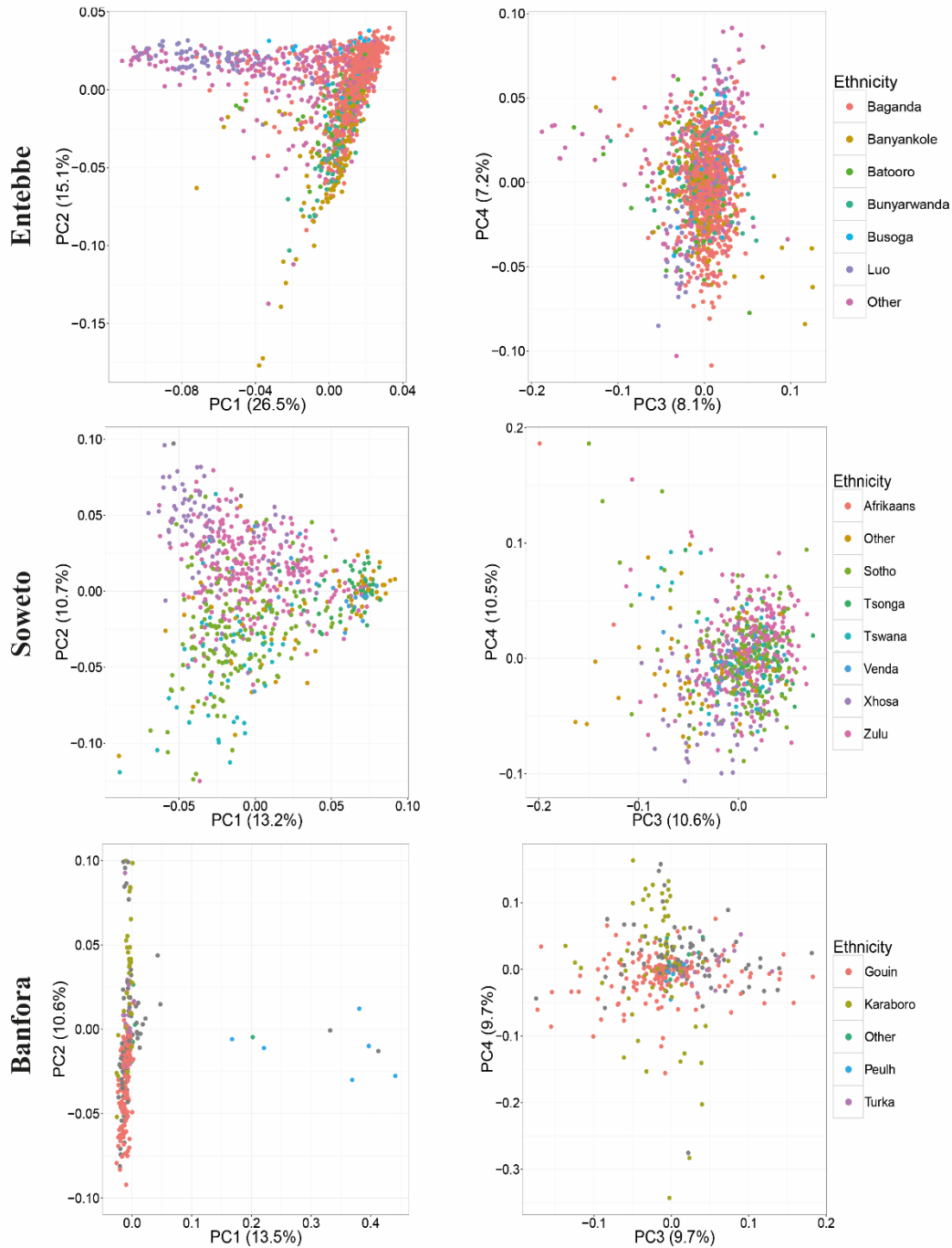


Figure 3.1: Principal components one to four (PC1-PC4) of the three populations of vaccinated infants.

Individuals represented as points coloured by maternal reported ethnicity. 12 individuals of mixed European ancestry were removed from the Soweto analysis prior to calculation of PCs to achieve a greater resolution of population-specific structure. The proportion of variance explained by each principal component is shown in the axes titles for each ethnicity.

3.5 Genotype imputation

A large African-specific reference panel was used to impute missing genotype variants (including individuals of Asian, American, European and African ancestry from 1000G; individuals from Ethiopia, Somalia, South Africa (Zulu) and Uganda from the AGVP set; and individuals of diverse ethnicity from rural Uganda from the Ug2g).

Comparing the concordance between imputation and genotype calls ('type 0' SNPs masked during imputation) on chromosome 6 between Entebbe and Soweto populations binned into categories defined by the MAF and stratified by imputation reference panel demonstrated improvements in imputation performance for common variants using the merged 1000G/AGVP/Ug2g reference panel compared to 1000G alone or 1000G/AGVP (**Figure 3.2**). An equivalent level of performance was observed for other chromosomes tested individually. The Banfora population performed in a near-identical fashion to the Entebbe population (excluded from the figure for simplicity), Performance was observed to be lower for the Soweto population that is diverse and distinct from the other two African populations and only referenced by a single collection of 100 Zulu individuals in the imputation reference panel.

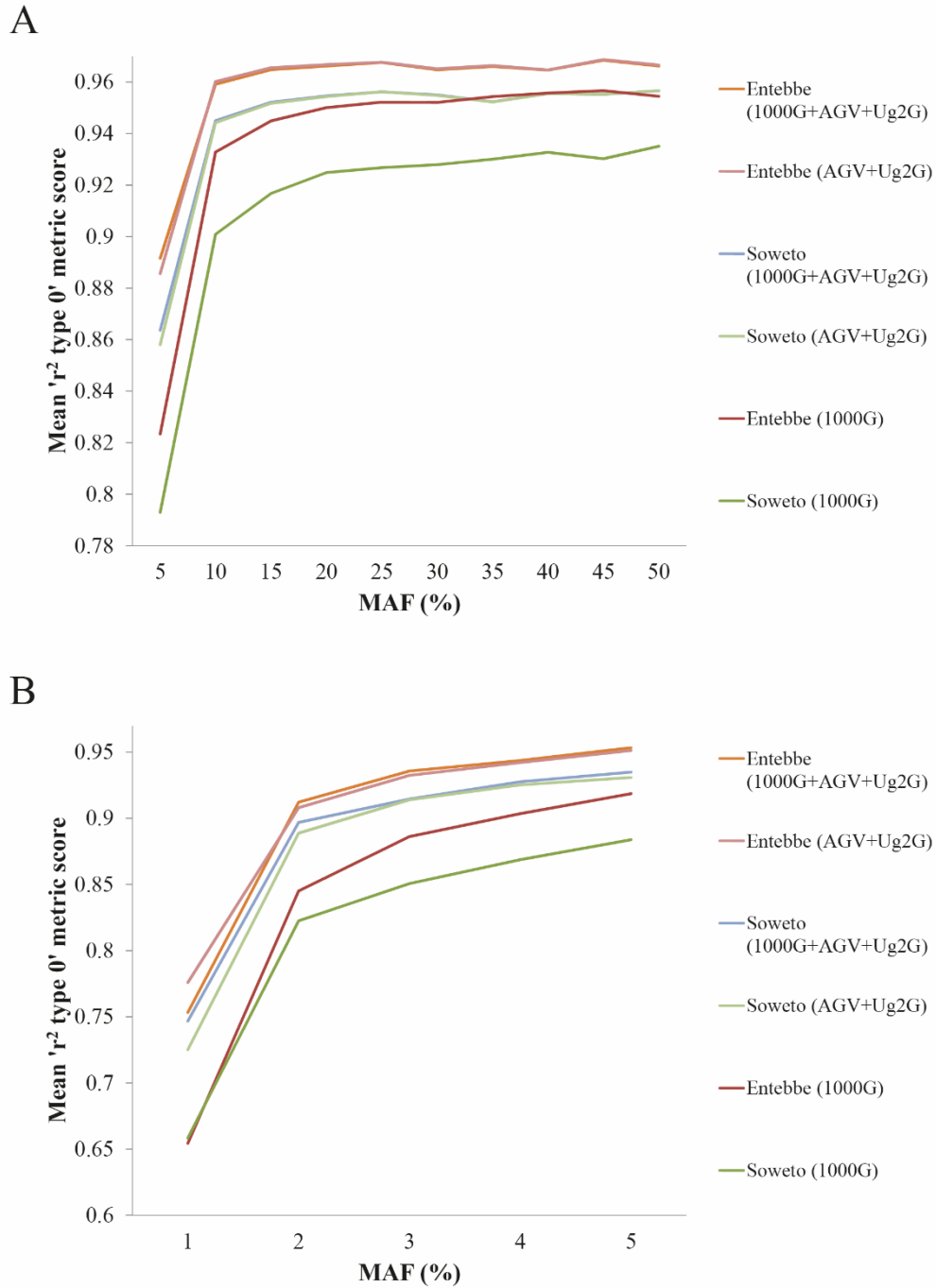


Figure 3.2: Mean imputation quality score per set of SNPs on chromosome 6.

Directly genotyped variants were masked and imputed and r^2 calculated between imputation and genotype calls with results plotted binned by minor allele frequency (MAF) focussing independently on common (5 – 50%; A) and rare (1 – 5%; B) variants. Lines are coloured by population and by reference panel used for imputation. 1000G: 1000 genomes; AGV: African Genomes Variation; Ug2G: 2000 Ugandan genomes

The global panel (1000G+AGV+Ug2g) was therefore used in further downstream analysis for all chromosomes (autosomes and X) in the three populations. A summary of the number of variants imputed in each population is provided in **Table 3.3**.

Table 3.3: Number of variants imputed into individual datasets divided between autosomes and the X-chromosome.

Autosomal variants post-imputation			
Cohort	Total imputed (n)	Total high quality* (n)	Total tested for association† (n)
Entebbe	103,992,756	49,820,138	17,812,003
Soweto	103,948,407	39,414,539	20,351,192
Banfora	103,543,613	31,941,883	18,537,816
X-chromosome variants post-imputation			
Entebbe	4,257,294	1,792,278	724,497
Soweto	4,257,257	1,354,576	757,584
Banfora	4,257,217	1,140,538	682,994

*: High quality SNPs are those only with an info score greater than 0.3

†: SNPs tested for association were only those of high quality with a minor allele frequency (MAF) greater than 0.01

3.6 Measured vaccine antibody responses

A multiplex Luminex-based immunoassay was used to measure antibody responses against DT, TT, PT, FHA, PRN, Hib and MV in all three populations using plasma or serum samples available from infants at defined time-points (**Figure 2.1**). The distributions of responses following transformation using either a logarithmic or inverse normal transformation function are shown in **Figure 3.3**. Following logarithmic transformation significant differences were observed

between the populations. No PRN was detectable in the Soweto population as would be expected since it was known that these infants received an aP preparation containing no PRN. Similarly the infants recruited from Soweto were sampled at 6 months of age prior to MV administration; hence responses were lower than those observed in Entebbe or Banfora. A small distribution of infants from Entebbe exhibited a similar distribution. It had been recorded that these infants had all received their MV. Otherwise the distributions of responses to other vaccines were generally higher in the Soweto group compared to the other two groups, most likely owing to the earlier age of sampling. Some responses were observed to be more stable between populations and therefore perhaps over time (TT and Hib for example), whereas other responses waned more rapidly. HBsAg responses in particular appeared to wane significantly as has been previously reported⁶⁸. Inverse normal transformation within populations resulted in sufficient normalisation of responses between populations. Responses against PRN and MV in the Soweto group were not included in further downstream analyses.

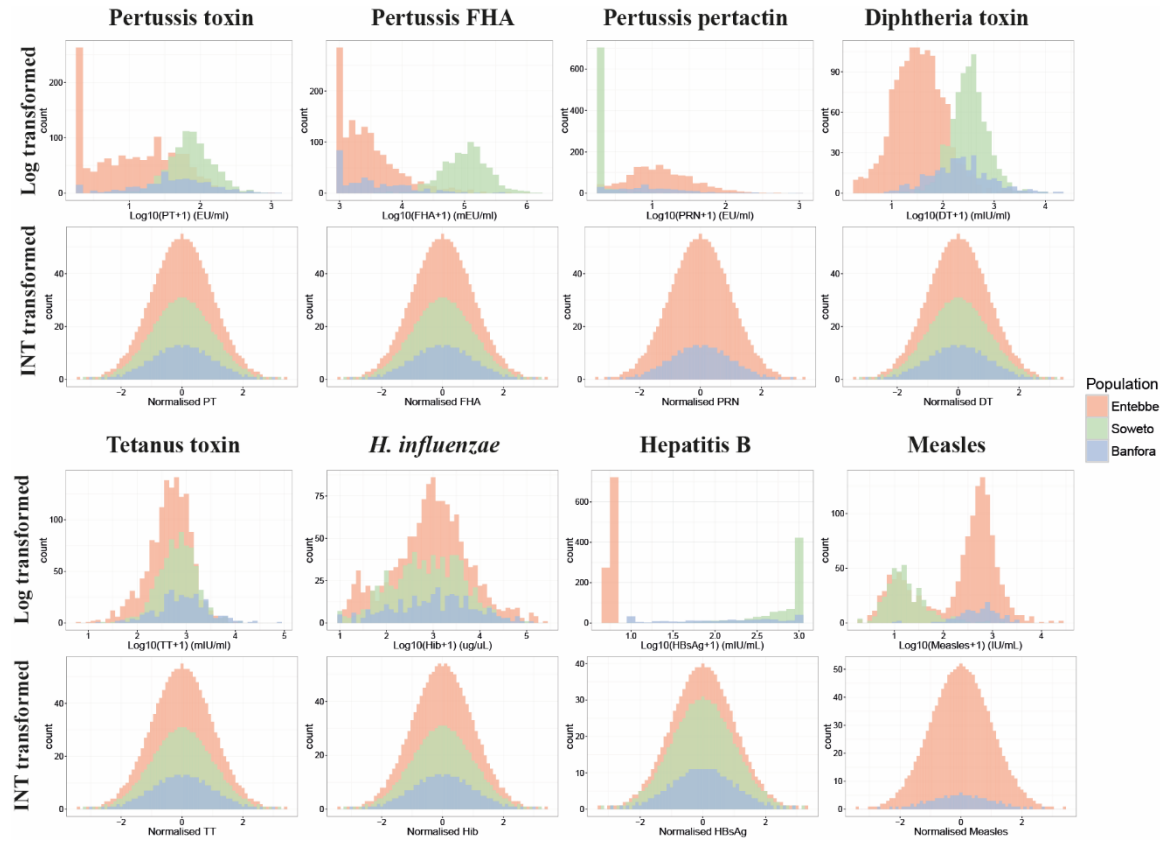


Figure 3.3: Distributions of vaccine responses in individual cohorts following transformation using two methods. Log is logarithmic and INT is inverse normal transformation to a distribution with mean 0 and standard deviation of 1 performed within each cohort individually. INT was not performed for the pertactin or MV traits in the Soweto population owing to the antigens not being delivered by the time of blood sampling.

Many of the vaccine responses were correlated (Pearson's $r > 0.4$), especially those administered together as part of DTP/Hib/HBsAg (**Figure 3.4**).

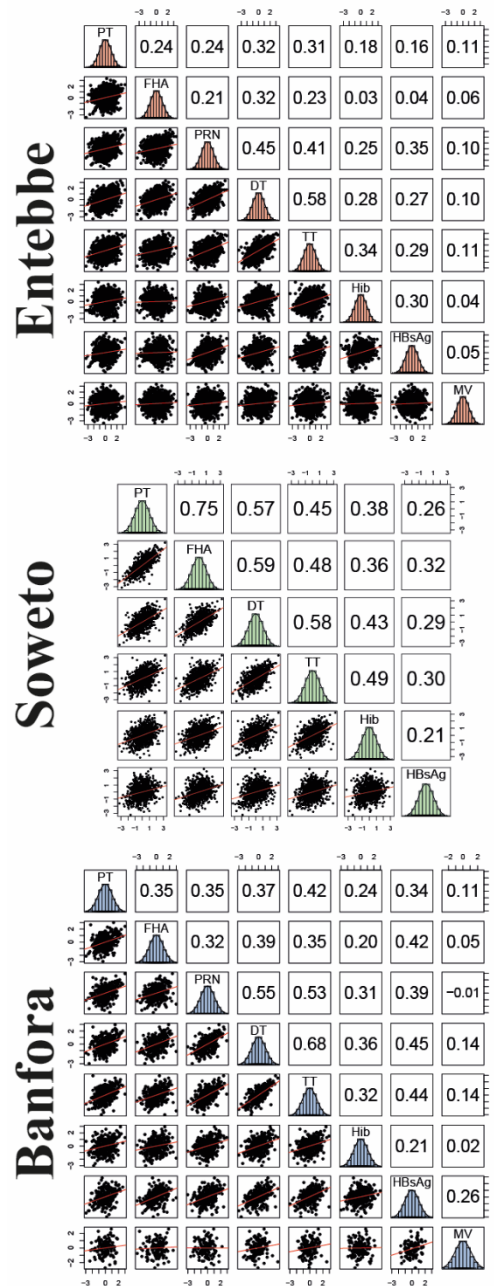


Figure 3.4: Correlation between normalised vaccine antigen responses across study populations.

Histograms of individual responses are shown on the diagonal for each cohort individually with scatterplots of pairs of responses on the lower panels and Pearson's r provided in the upper panel for each combination of vaccine responses in each population.

A positive correlation was observed for all pairs of vaccine antigen responses across the three populations with one exception (MV vs PRN in Banfora; Pearson $r = -0.01$). The range of observed correlations were higher in Soweto (Pearson r range 0.21-0.75) compared to Entebbe (0.03 – 0.58) and Banfora (-0.01 – 0.68). The greatest correlations were observed between responses in components against the toxoid vaccines. In Entebbe and Banfora the greatest correlations were observed between DT and TT (r of 0.58 and 0.68 respectively), whereas in Soweto the greatest correlation was observed for PT and FHA (0.75). MV responses were generally less correlated with the other responses in Entebbe and Soweto groups consistent with the administration of this vaccine at a different time point compared to the others.

A high inter-platform correlation was observed between TT (Pearson $r = 0.72$; $n = 1,104$) and MV (0.83; $n = 942$) vaccine responses measured on Luminex compared to ELISA platforms (**Figure 3.5**). The cut-off for detection of TT responses was appreciably lower using the ELISA compared to the Luminex method that is recognised to have a lower limit of detection²⁵⁶.

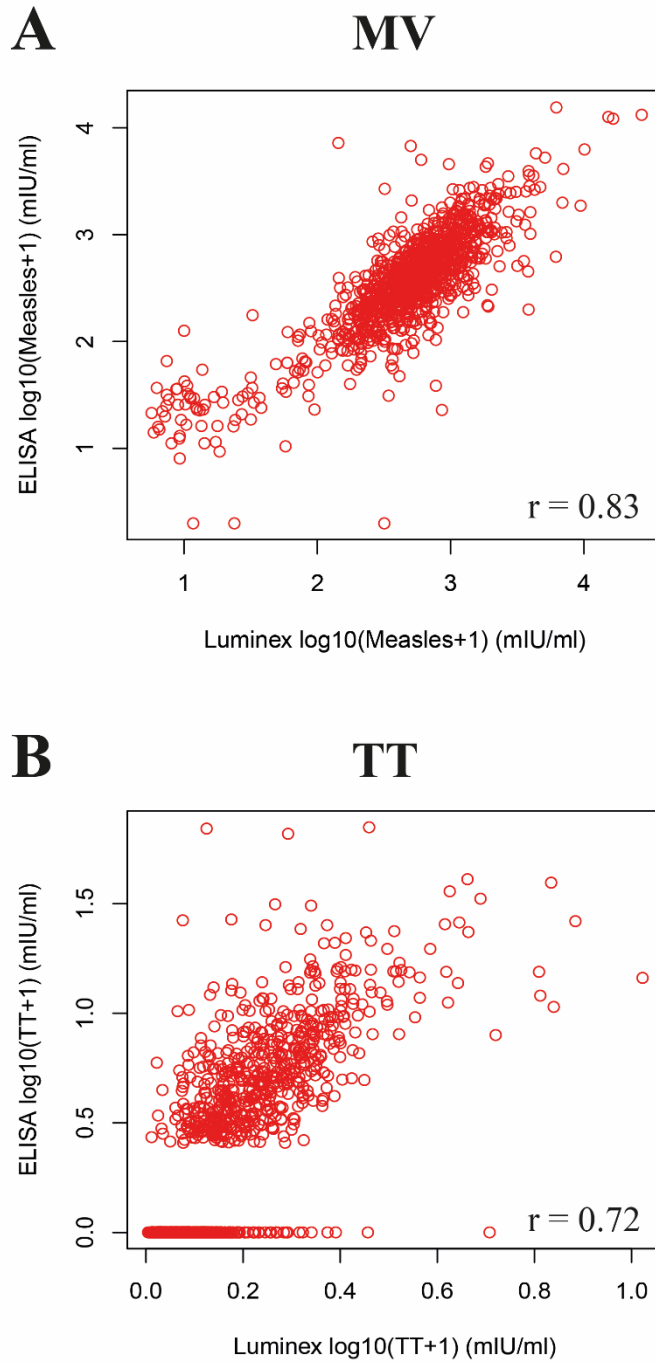


Figure 3.5: Correlation between vaccine responses against MV and TT in Entebbe individuals as measured using Luminex and ELISA assays.

Log transformed concentrations were compared for each individual where Luminex immunoassay (read as part of this thesis) and ELISA results (generated as part of the initial trial) were available for individuals in Entebbe.

3.7 Other predictors of vaccine response

Variables such as sex, time between vaccination and sampling, breast feeding, HIV diagnosis in the infant, ethnicity, weight-for-length Z score and asymptomatic malaria parasitaemia were explored for their relationship with each of the antibody responses (**Supplementary Table S1**). Not all variables were available in all cohorts and there were differences in how each were measured. A linear association between the time between vaccination and sampling was observed across nearly all traits across all cohorts. This association was particularly pronounced for the DT trait (**Figure 3.6**) whereas for MV there was only a marginal association ($p = 0.01$ across both Entebbe and Banfora populations and 0.06 in the Entebbe population alone). It is likely that the true relationship between time since vaccination and antibody levels (or antibody decay) would be best modelled as a negative exponential but in the measured time span (60 weeks), a linear fit appeared most appropriate. No significant associations for any responses (given the multiple number of tests performed) were observed with sex (**Figure 3.7**). A reduction in antibody levels was observed for those infants with confirmed HIV infection in Uganda but no differences were observed for those infants with exposure to HIV *in utero* which has been hypothesised to reduce humoral responses to EPI vaccines²⁵⁷ (**Figure 3.8**). The numbers of infants vertically infected with HIV in Soweto were too low to allow any replication of the observations in Entebbe.

It was noted in the Entebbe population that the concentration of measles IgG measured in the mother at delivery was significantly associated with the measles concentration measured in the infant at 1 year of age ($p = 8.2 \times 10^{-5}$; **Figure 3.9**).

This relationship may, in part, explain the reduced effect of time between sampling and vaccination observed for MV. Despite observing some associations between asymptomatic malaria parasitaemia and responses to PT and measles in Entebbe and Banfora where malaria is endemic, these associations were marginal and not replicated with consistent directionality across the two tested populations (**Figure 3.10**).

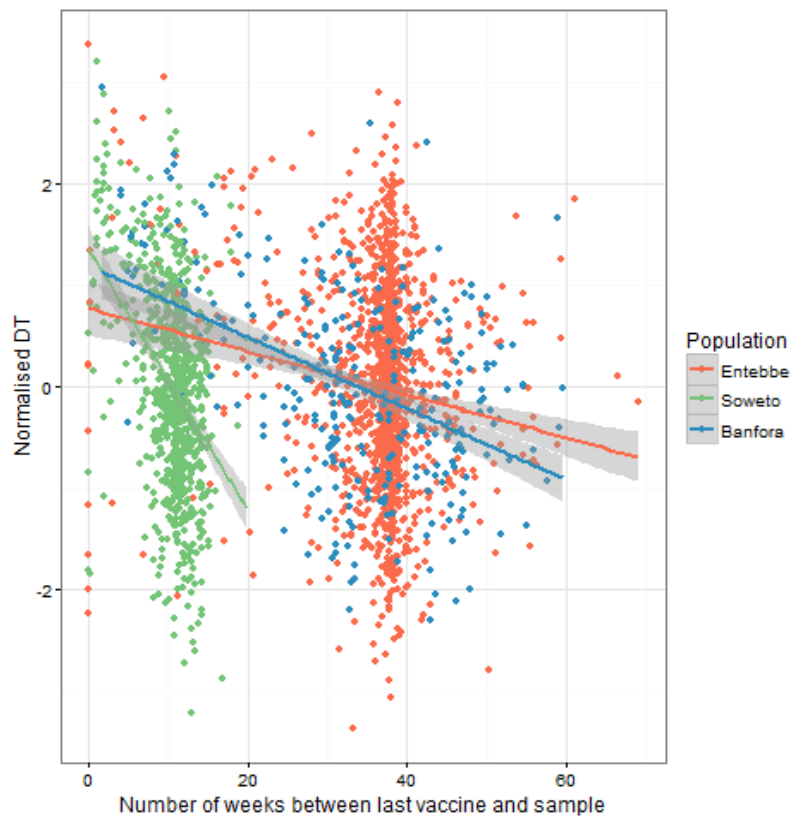


Figure 3.6: Relationship between normalised antibody concentration against DT and time between last vaccination and sampling within each cohort.

Inverse normalised concentrations of DT are presented for each individual plotted by time measured between vaccination and sampling. Each point represents an individual coloured by population. A linear line of best fit with 95% confidence interval is shown for each population.

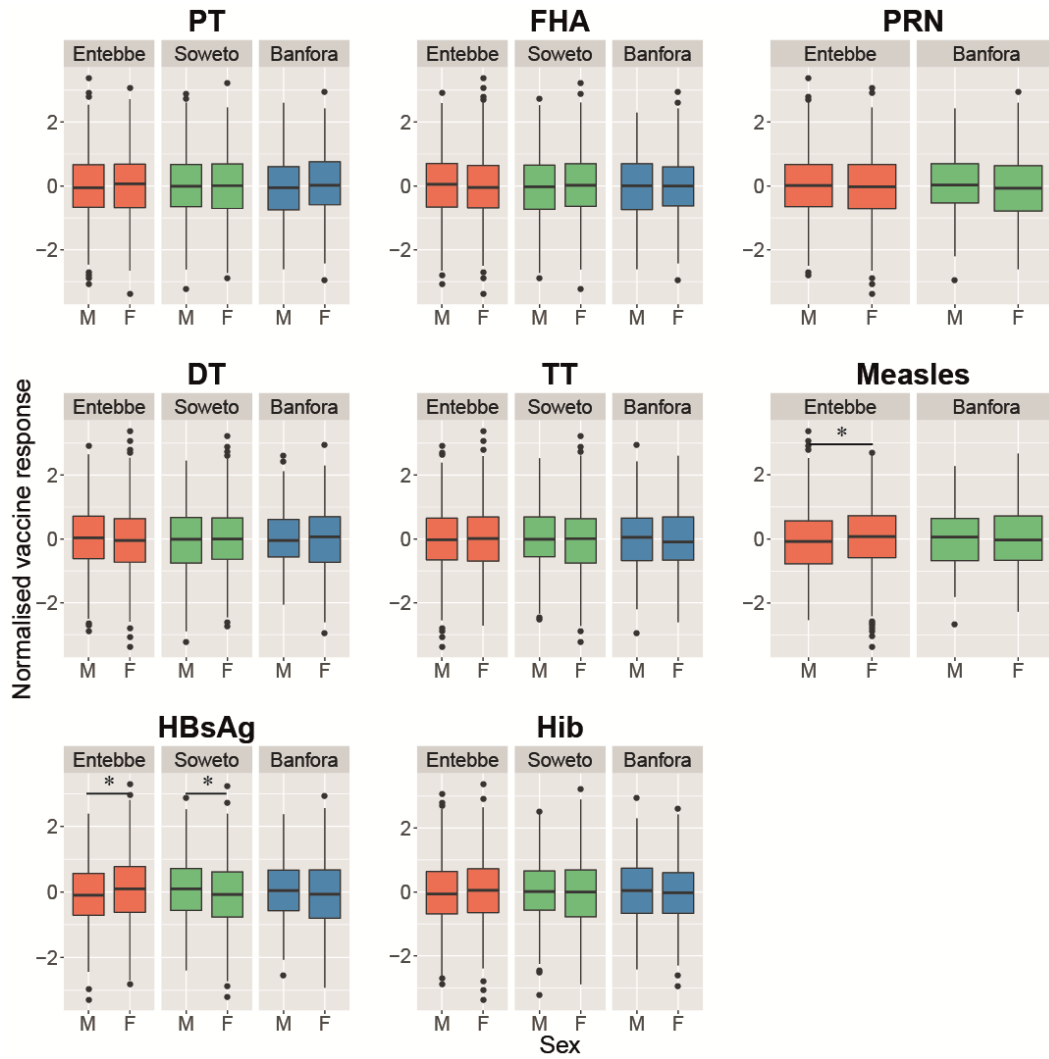


Figure 3.7: Relationship between normalised antibody concentrations and sex in three populations.

Inverse normalised distributions of each measured vaccine response trait are presented by sex within each population. Significant differences were only observed for measles response in Entebbe and HBsAg responses in Entebbe and Soweto. F: females; M: males; *: Student's t-test $p < 0.05$.

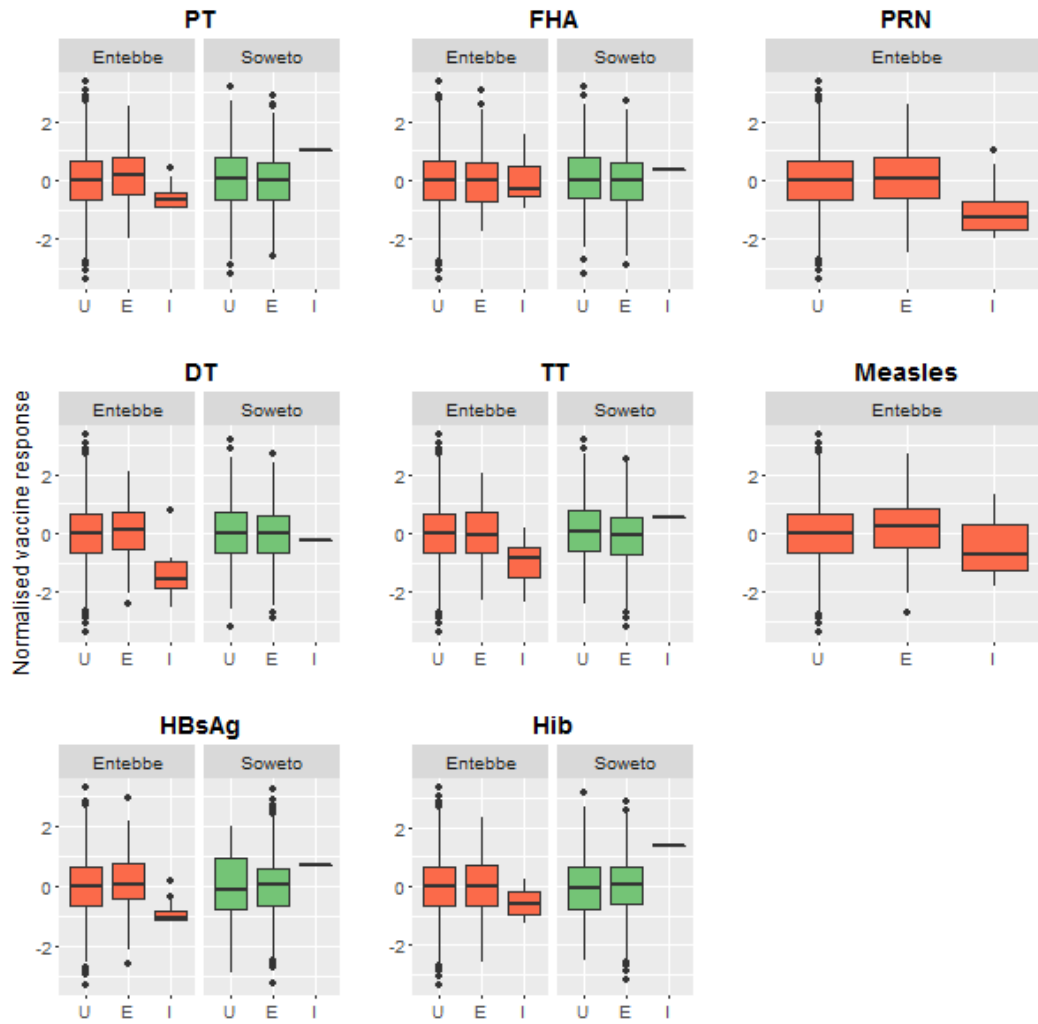


Figure 3.8: Relationship between normalised antibody concentrations and HIV infection status in Entebbe and Soweto populations.

Inverse normalised distributions of each measured vaccine response trait are presented by HIV status at 12 (Entebbe) or 6 months (Banfora) within each population. A significant ($p < 0.01$) reduction in responses against all vaccines except FHA were observed for individuals infected with HIV in Entebbe. No differences were observed for the single HIV infected individual included in the Soweto study. U: child uninfected with HIV; E: child exposed to HIV *in utero* but uninfected at birth (i.e. mother infected); I: child vertically infected by HIV.

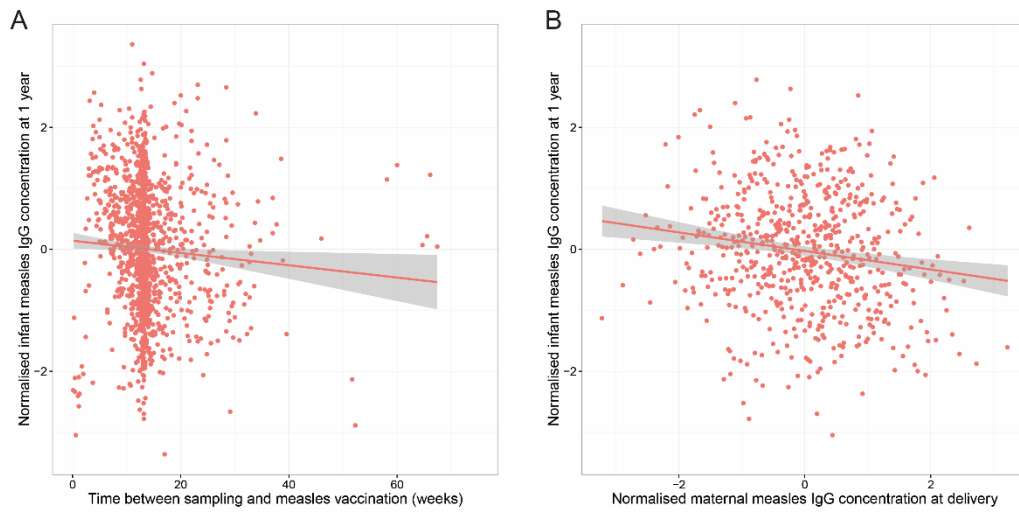


Figure 3.9: Relationship between normalised measles antibody concentrations and time between sampling and vaccination (A) and maternal measles concentrations at delivery (B) in the Entebbe individuals.

Inverse normalised concentrations of MV are presented for each individual from Entebbe plotted by time measured between vaccination and sampling (A) or measured and normalised concentration of measles antibody measured in the cord blood from their mother (B). Each point represents a single individual. A linear line of best fit with 95% confidence interval is shown for each scatter plot.

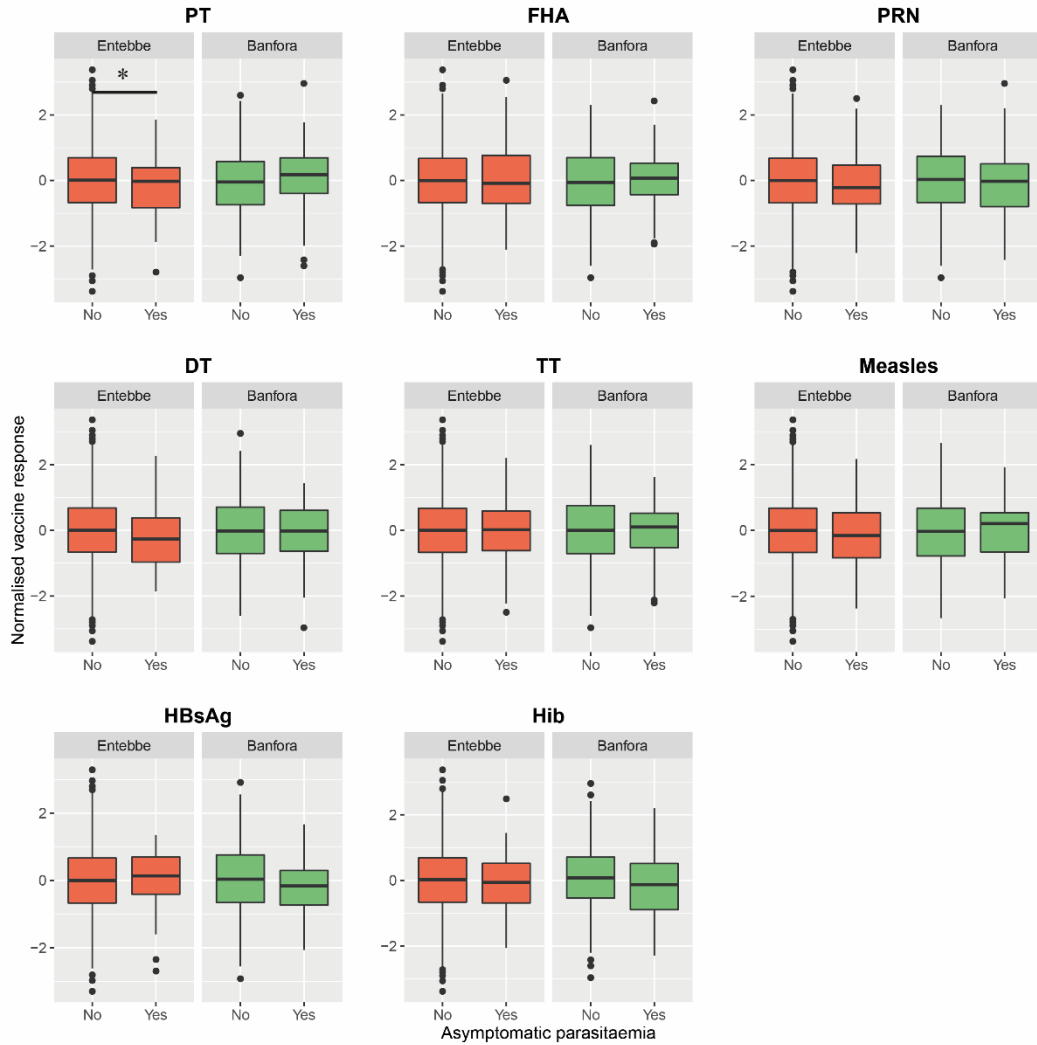


Figure 3.10: Relationship between normalised antibody concentrations and asymptomatic malaria parasitaemia in two populations.

Inverse normalised distributions of each measured vaccine response trait are presented by whether asymptomatic malaria parasitaemia was detected at the time of sampling for individuals within each population. A significant difference was only observed for PT response in Entebbe.

*: Student's t-test $p < 0.05$

3.8 Intra-population vaccine response GWAS

Following quality control, genotyped and imputed SNPs in each of the three populations were tested for association with the normalised vaccine response traits (8 traits in Entebbe and Banfora and 6 in Soweto) using a LMM testing an additive model of inheritance including time between vaccination and sampling (in weeks) as a continuous covariate. Inspection of the QQ plots and genomic control inflation factors (**Figure 3.11**) demonstrated adequate correction for population structure and cryptic relatedness ($\lambda < 1.05$) in all traits and populations tested except PT ($\lambda = 1.13$), PRN ($\lambda = 1.05$) and DT ($\lambda = 1.12$) in the Entebbe cohort. Inspection of the Manhattan plots (**Figure 3.12**) demonstrated that peaks of significant ($p < 5 \times 10^{-9}$) associations were observed on chromosome 6 in multiple traits in multiple populations. These associations all fell within the MHC region (position 25.5-34 Mb) and re-calculation of the genomic control inflation factor excluding this region in those traits with raised statistics resulted in satisfactory control (PT $\lambda = 0.99$; PRN $\lambda = 0.99$; DT $\lambda = 1.00$).

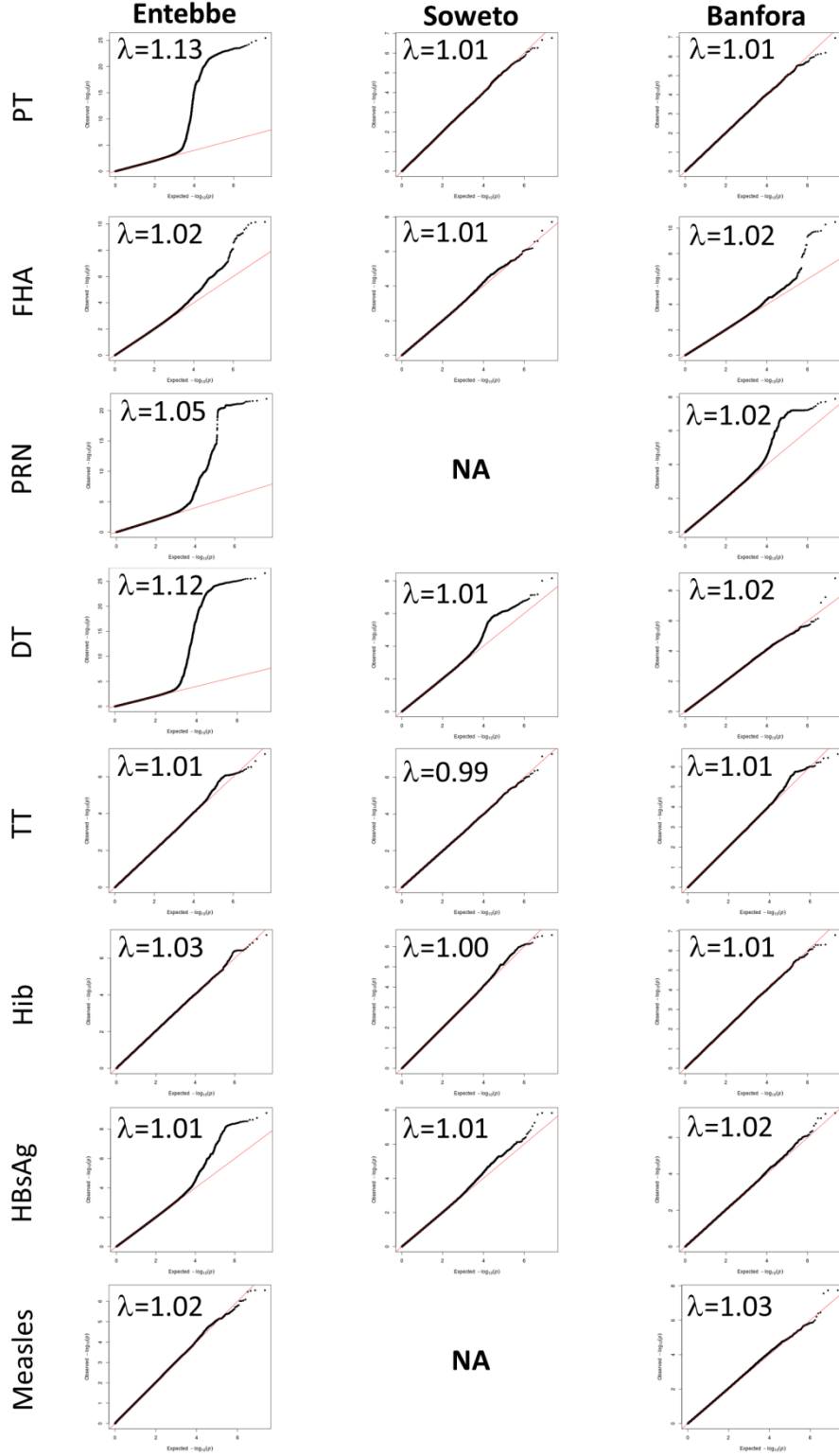


Figure 3.11: Quantile-quantile plots of association statistics for all measured vaccine traits in all three populations. λ : Genomic control inflation factor including all loci genome-wide; x-axis is the expected p-value; y-axis is the measured p-value.

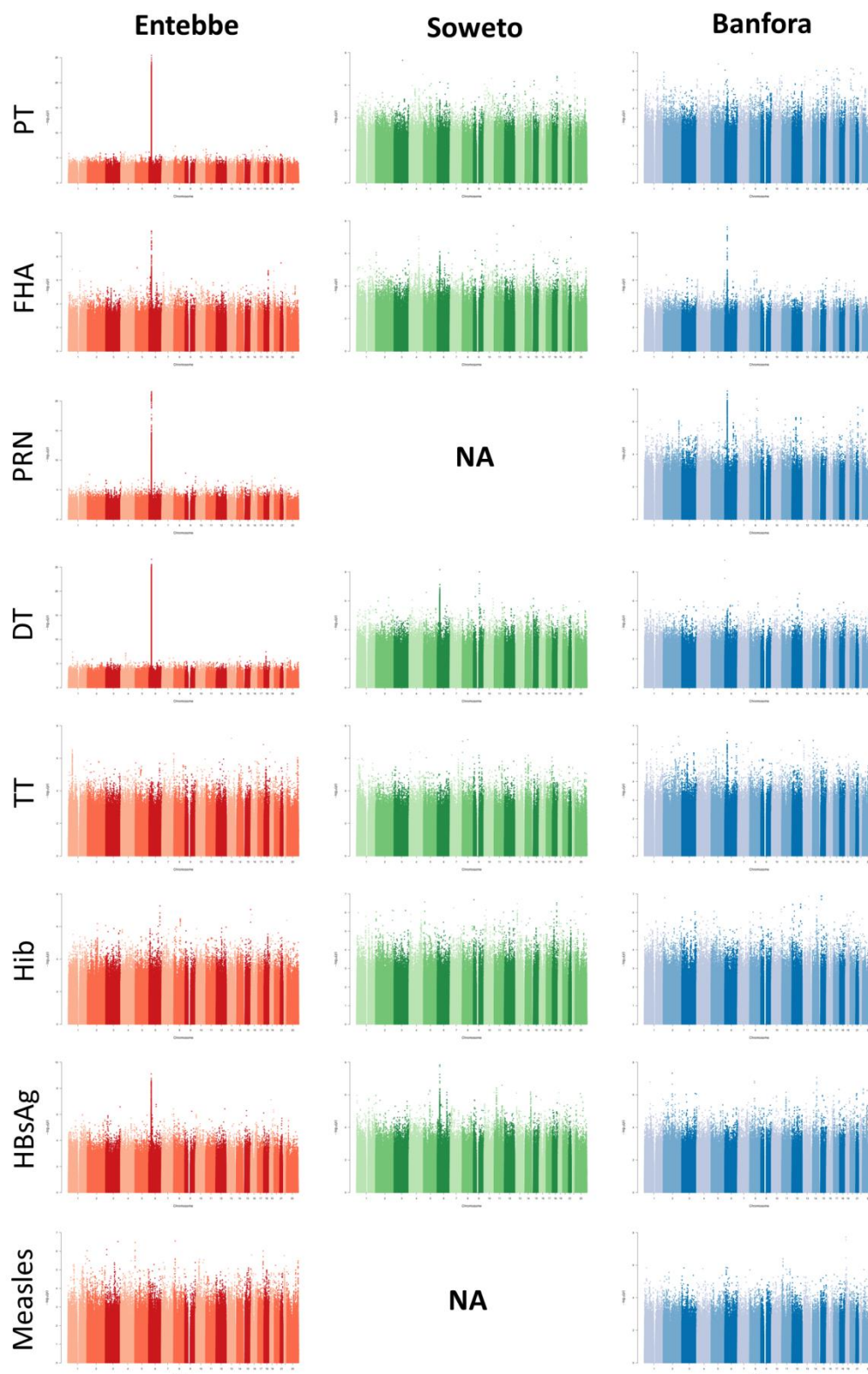


Figure 3.12: Manhattan plots of association statistics for all measured vaccine traits in all three populations.

3.9 Fixed effects meta-analysis

Having demonstrated that the LMM approach controlled appropriately for population structure in all cohorts, the results were combined through a fixed effects meta-analysis to search for genetic associations with shared effects across all populations. Statistically significant associations were observed across the MHC region with responses against PT, FHA, PRN, DT and HBsAg (**Figure 3.13**). The SNPs with the most significant evidence of association from each GWAS (smallest p-value) were defined as the index SNP. The direction of effects were consistent for all index SNPs in all traits in all three cohorts (**Table 3.4**). No overlap in association statistics was observed between the results from this analysis and those index SNPs defined through two recent GWAS studies of HBsAg response undertaken in Asian populations (**Table 3.5**).

Other than variants within the MHC, one other SNP was found to be significantly associated with response to Hib (rs7965675; chromosome 12; bp 101,450,995; $p_{FE} = 1.6 \times 10^{-9}$; **Table 3.6**). This SNP (ancestral G to T) was determined to have minor allele frequencies in each population (Entebbe 0.36; Soweto 0.29; Banfora 0.34) consistent with those predicted through the 1000G population African populations (Luhya in Kenya 0.40 and Yoruba in Nigeria 0.30). The SNP had very high imputation performance statistics (in terms of reported info score), and is located in an intron of the anoctamin 4 (*ANO4*) gene that encodes a chloride channel. This SNP demonstrated evidence of reducing levels of responses against PT, PRN, TT, Hib and HBsAg in the final meta-analyses (**Table 3.7**).

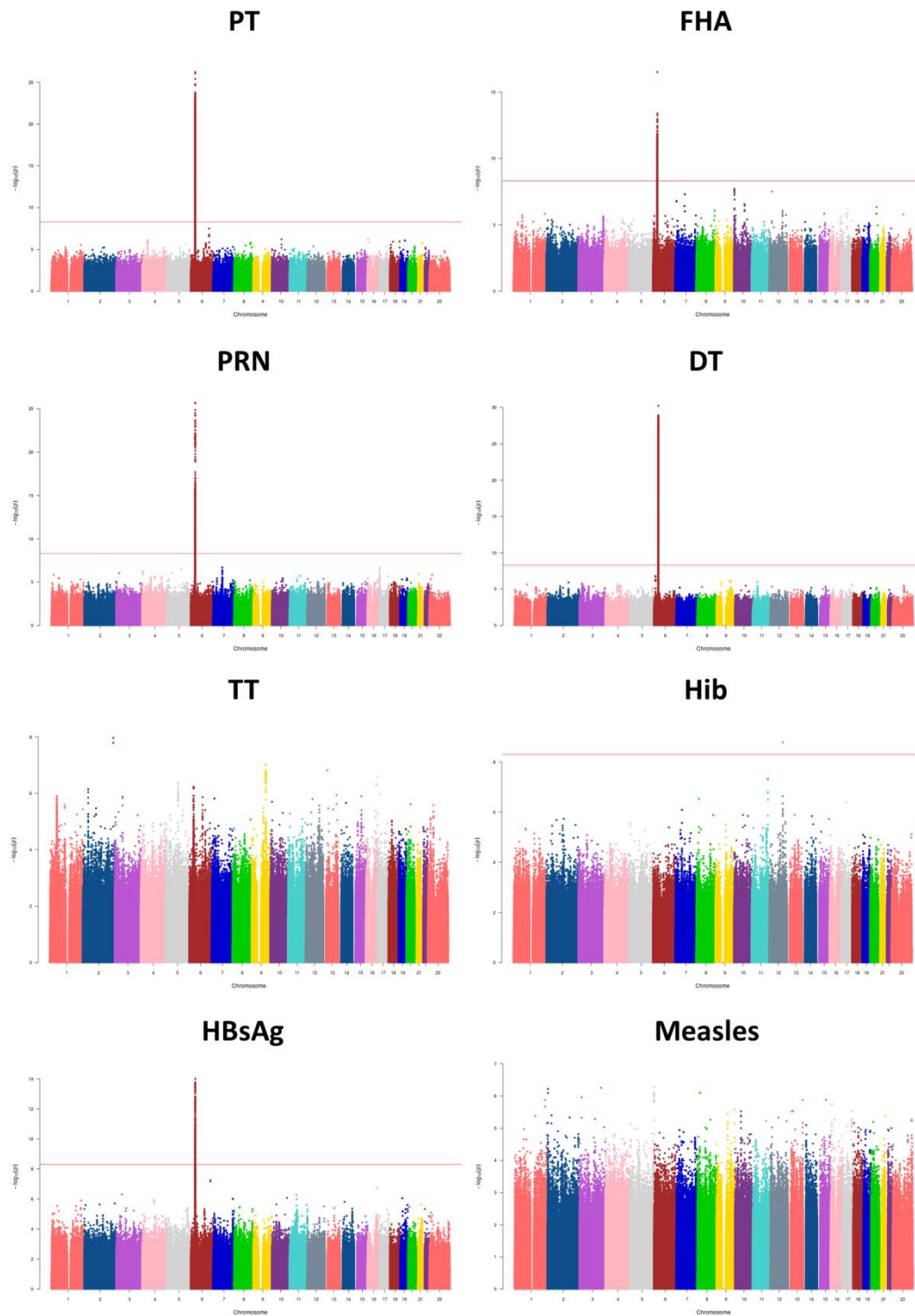


Figure 3.13: Manhattan plots of the results from the meta-analyses of GWAS performed for each vaccine response in two or three populations.

Meta-analyses significant associations were considered above the red horizontal line representing $p \leq 5 \times 10^{-9}$.

Table 3.4: Association statistics for the index associated MHC variants with each measured vaccine response trait. Only variants with plausible ($p < 1 \times 10^{-5}$) associations are presented with GWAS significant ($p_{FE} < 5 \times 10^{-9}$) associations highlighted in bold.

Trait	BP (b37)	SNP*	Info [†]	Gene [‡]	Allele	Entebbe			Soweto			Banfora			Meta
						MAF [¶]	Beta	s.e.	MAF [¶]	Beta	s.e.	MAF [¶]	Beta	s.e.	p _{FE}
PT	32,497,431	rs73727916	0.83-0.91	<i>HLA-DRB1</i>	T	0.60	0.43	0.04	0.64	0.16	0.06	0.49	0.24	0.08	6.1x10⁻²⁷
FHA	32,677,592	-	0.90-0.92	<i>HLA-DQB1</i>	C	0.24	0.30	0.04	0.20	0.30	0.04	0.30	0.30	0.04	3.1x10⁻¹⁷
PRN	32,599,601	rs147857322	0.97-0.99	<i>HLA-DQA1</i>	G	0.40	0.38	0.04	-	-	-	0.29	0.34	0.08	2.0x10⁻²⁶
DT	32,522,103	rs201150249	0.88-0.95	<i>HLA-DRB4</i>	C	0.14	0.60	0.06	0.21	0.31	0.06	0.16	0.25	0.10	5.9x10⁻³¹
TT	32,410,716	rs9268658	NA	<i>HLA-DRA</i>	G	0.46	0.14	0.03	0.63	0.14	0.03	0.53	0.15	0.04	5.9x10 ⁻⁷
Hib	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
HBsAg	32,590,364	rs6914950	0.96-0.98	<i>HLA-DQB1</i>	T	0.12	-0.36	0.05	0.18	-0.36	0.05	0.12	-0.36	0.06	9.9x10⁻¹⁵
Measles	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

*: Variants derived from either the AGVP or Ug2G reference panels do not yet have 'rs' identifiers as determined through dbSNP.

†: Imputation performance score range across the three populations. Value is 'NA' if variant was originally genotyped.

‡: Nearest protein-coding gene within +/- 25 kb

¶: MAF: minor allele frequency (may be greater than 0.5 despite being classified as 'minor')

||: s.e.: standard error.

Table 3.5: Association statistics for SNPs in the MHC found associated with variable responses to HBsAg in publications including two Asian populations.

BP (b37)*	SNP	Refs. [†]	Allele	O.R. [‡]	Entebbe				Soweto				Banfora			
					MAF	Beta	s.e. [§]	p	MAF	Beta	s.e. [¶]	p	MAF	Beta	s.e. [¶]	p
31,870,856	rs9267665	¹⁹⁴	A	2.05	12.4	0.01	0.07	0.88	12.6	-0.02	0.08	0.81	17.5	0.05	0.12	0.74
32,389,648	rs3135363	¹⁹³	A	1.53	18.1	0.10	0.06	0.10	28.7	0.06	0.06	0.29	11.7	0.04	0.13	0.75
32,569,691	rs477515	¹⁹⁴	G	3.59	13.5	0.07	0.07	0.31	15.9	0.03	0.07	0.68	19.0	0.08	0.10	0.45
33,054,861	rs9277535	¹⁹³	A	0.72	18.0	-0.02	0.06	0.75	21.6	-0.03	0.06	0.69	14.2	0.004	0.12	0.99

*: BP: base-pair position in build 37 (b37)

†: Refs: publication references

‡: O.R. published odds ratios for the associated SNPs. Minor allele frequencies were not available for the SNPs in the published studies.

¶: s.e.: standard error for the association beta in each of the African populations

Table 3.6: Association of SNP rs7965675 with antibody response to Hib in the three studied populations and in the final fixed effects meta-analysis (Meta).

Entebbe			Soweto			Banfora			Meta
Info	Beta	s.e.*	Info	Beta	s.e.*	Info	Beta	s.e.*	p _F E
0.98	-0.18	0.03	0.97	-0.18	0.03	0.97	-0.18	0.03	1.6x10 ⁻⁹

*: s.e: standard error.

Table 3.7: Summary statistics from meta-analysis testing association of ANO4 rs7965675 SNP with all vaccine response traits.

Trait	Beta _F E	s.e. _F E	p _F E
PT	-0.08	0.03	6.0x10 ⁻³
FHA	-0.04	0.03	0.25
PRN	-0.10	0.04	3.9x10 ⁻³
DT	-0.09	0.03	4.1x10 ⁻³
TT	-0.10	0.03	1.0x10 ⁻³
Hib	-0.18	0.03	1.6x10 ⁻⁹
HBsAg	-0.08	0.03	0.01
Measles	-0.03	0.04	0.42

3.10 Non-MHC associations previously reported with vaccine-related traits

The results from the meta-analyses of the vaccine related traits in the three African populations were inspected for evidence of association across loci previously reported to be associated with other related traits (**Table 3.8**). No previously associated variants were noted to have a $p_{FE} < 0.05$. Variants within a 10 kb window of 5 variants were found to have a $p_{FE} < 1 \times 10^{-2}$. These variants lay within the *FOXP1* (n = 1), *IF144L* (n = 2) and *CD46* (n = 2) genes.

Table 3.8: Variant meta-analysis association statistics with vaccine traits for those SNPs found associated with vaccine-related traits in previous studies. SNPs in a 10kb region of those SNPs also represented.

	CHR	BP (b37)	SNP*	Gene	Minor Allele	Present in data	MAF	p _{FE}	Number Variants [†]	Minimum p _{FE} [†]	Refs
Hib	1	206,944,233	rs1554286	<i>IL10</i>	T	Yes	0.40-0.45	0.79	31	0.09	211
	11	126,160,687	rs1893352	<i>TIRAP</i>	G	Yes	0.08-0.10	0.71	27	0.10	211
HBsAg	1	218,568,829	rs2799090	<i>TGFB2</i>	G	Yes	0.22-0.30	0.36	42	0.11	180
	3	71,001,255	rs6789153	<i>FOXP1</i>	C	Yes	0.39	0.32	26	8.1x10⁻³	180
	5	55,294,328	rs6870870	<i>IL6ST</i>	C	No	-	-	35	0.38	180
	8	144,235,008	rs7461107	<i>LY6H</i>	T	Yes	0.15-0.20	0.12	42	0.12	180
	9	117,558,703	rs6478108	<i>TNFSF15</i>	C	No	-	-	18	0.84	180
	9	123,762,676	rs1978270	<i>C5</i>	T	Yes	0.47-0.60	0.92	43	0.02	180
	10	54,525,509	rs2506	<i>MBL2</i>	C	No	-	-	45	0.94	180
	12	9,985,915	rs2232548	<i>KLRP1</i>	T	No	-	-	39	0.07	180
	14	76,427,780	rs3917210	<i>TGFB3</i>	A	Yes	0.20-0.28	0.42	27	0.03	180
	16	30,518,041	rs2230433	<i>ITGAL</i>	C	No	-	-	53	0.23	197
	17	34,323,944	rs854692	<i>CCL15</i>	C	No	-	-	49	0.86	180
	19	55,174,213	rs1654668	<i>LILRB4</i>	T	No	-	-	69	0.02	180

*: SNPs present in post-filtered data (info > 0.3 and MAF > 0.05).

[†]: Number of variants within a +/- 5kb window

Table 3.8: continued

Measles	1	79,093,818	rs273259	<i>IFI44L</i>	G	Yes	0.34-0.40	0.06	26	6.0x10⁻³	207
	1	79,094,081	rs1333973	<i>IFI44L</i>	T	Yes	0.29-0.32	0.08	26	6.0x10⁻³	204
	1	208,014,922	rs1318653	<i>CD46</i>	C	Yes	0.18-0.20	0.09	42	5.7x10⁻³	207
	1	207,930,203	rs2724384	<i>CD46</i>	A	Yes	0.06-0.07	0.10	32	0.7x10⁻³	204

*: SNPs present in post-filtered data (info > 0.3 and MAF > 0.05).

†: Number of variants within a +/- 5kb window

3.11 Conclusions

Through undertaking the first GWAS of responses to multiple vaccines delivered to infants in three diverse African populations, this study has identified statistically significant associations between genetic variants across the class II MHC region and antibody responses against PT, FHA, PRN, DT and HBsAg. These signals, which with the exception of HBsAg are all novel, were identified through the use of a LMM association analysis approach, a validated sensitive set of antibody assays and sets of well characterised cohorts.

The identification of genetic variants across the class II MHC being associated with vaccine responses is consistent with previous observations from twin studies looking at responses to hepatitis B, diphtheria and pertussis vaccines^{195,196} and GWAS studies looking at HBsAg responses^{185,193,194}. No overlap was observed, however, between those SNPs with the most significant associations (index SNPs) observed for HBsAg response in published studies of HBsAg response or non-response in Asian populations^{189,190} compared to the results from the African populations described here. There are multiple possible explanations for this observation. Firstly, differences in observed genetic association signals may result from differences in schedules of vaccination between continents demonstrating itself as between-study environmental effect heterogeneity. The individuals included in both Asian studies were adults receiving between 2 and 6 doses of vaccine with a range of 1 to 6 months between the final vaccine and the sampling for response measurements. Furthermore, the analyses in the Asian populations were comparing vaccine responders (<10mIU/ml) against non-responders

(>1000mIU/ml). The study performed in Indonesia did include an intermediate category to increase power of detection through ordinal regression. Such a comparison would not have been possible in the cohorts included in this study given the heterogeneity in responses between study populations owing to differences in timing between the last vaccination and blood sampling (**Figure 3.3**). Despite these notable differences in both timings and distributions of vaccine responses the index variants and the effect estimates for these variants associated with HBsAg response in the African populations were remarkably consistent with little evidence of heterogeneity (**Table 3.4**) especially when compared to the other vaccine response associations. These observations suggest that the top associated genetic variant for HBsAg response identified in all three African populations is either itself the causative variant or it is in close LD with another variant that explains a significant proportion of the variance of response distribution. Little is known about this particular SNP (rs6914950) in terms of regulatory effects. Irrespective of the underlying mechanism the effect is clearly consistent across African populations and therefore if the signal of association was common across populations of differing ancestry, the signal would be expected to be observed independent of the method used to define response to HBsAg.

A second possible explanation for the observed differences is one of genetic heterogeneity²⁴⁸. Polymorphism rates are exceptionally high across the MHC locus and resultant alleles are well documented to occur in a population specific manner. Together, these combined observations of high polymorphism rates and multi-allelic states are taken as evidence in support of the theory of selection occurring across the locus¹⁵⁷. Thus it is plausible that the variants found associated in each of the two sets of studies (African and Asian) are tagging

shared common variants that have not been compared between populations owing to the unavailability of summary statistics from the Asian studies that would normally facilitate meta-analyses. Based on genetic associations investigating the effect of MHC variation on diseases including HIV disease progression¹⁶⁴ and multiple autoimmune conditions^{165,169}, it is likely that these variants are HLA alleles. Investigating this possibility and the possible mechanisms underlying these associations will be the subjects of subsequent chapters.

The absence of a significant genotype association signal across the MHC with measles vaccine response is consistent with a recent meta-analysis of GWAS studies undertaken in cohorts recruited in the USA²¹⁰. Although surprising, it is important to note that MV, being a live virus, represents a completely different type of vaccine compared to the others tested in this thesis. Taken together these findings suggest that association signals observed in earlier candidate HLA gene studies²⁵⁸ may represent false positive reports. It is still possible, however, that any HLA alleles that are associated have not been appropriately tagged by those SNPs typed in the African populations included in this thesis. These effects will again be explored in the subsequent chapters.

The non-MHC findings presented for MV response in this thesis do however provide some evidence that genetic heterogeneity may explain the apparent lack of association signal at the top index variants *CD46* and *IF144L* genes identified in other association studies of MV response and febrile seizures following MMR administration^{207,210}. While the index SNPs do not demonstrate evidence of association in the African populations, there is a suggestion that SNPs within a 10kb window are significantly associated. This may again be a result of the truly

causative SNP being best tagged by different SNPs in populations of differing ancestries. It is possible that the association signals may become more consistent if equivalent phenotypes were used such as neutralising antibody levels since the MV neutralising antibody response is regarded as a more definitive correlate of protection¹⁵. It would be expected, therefore, that measuring total IgG may diminish the ability to measure specific antibody responses that correlate more appropriately with protection. Replicating those genetic association signals reported for the neutralisation antibody and a defined MV phenotype (febrile seizures) may therefore be more challenging. Although planned for in the early stages of the project, the measurement of MV neutralising antibodies was not possible in this thesis owing to limited sample availability and the resource-consuming volumes required for such assays.

The meta-analyses presented in this chapter also identified one variant in an intron of the *ANO4* gene demonstrating evidence of association with Hib antibody response (**Figure 3.13** and **Table 3.6**). The anoctamins are a family of ten proteins that function as calcium ion activated chloride channels²⁵⁹. The first family members to be discovered (*ANO1* and *ANO2* genes) have been associated with epithelial chloride secretion, smooth muscle peristalsis and olfactory signal transduction. There is some evidence that the anoctamins may be associated with cellular volume regulation linked to cell migration and immunity and a genetic signal of association has been observed between a variant in *ANO10* and *Borrelia burgdorferi* seropositivity^{260,261}. However, these effects have not been replicated or investigated in detail. The *ANO4* transcript has recently been discovered to be highly expressed in the zona glomerulosa of the adrenal gland and the gene product is associated with secretion of aldosterone²⁶². The potential mechanism

underlying calcium ion activated chloride channel activity and antibody responses against Hib are not clear but may hypothetically involve some form of innate cell volume regulation. In testing whether this mechanism may be common to other vaccine responses it was observed that the same SNP was also found associated in the same direction with responses against PT, PRN, DT, TT and HBsAg. Although interesting, these *ANO4* associations require more investigation to, firstly, attempt to replicate the findings in other independently recruited cohorts and secondly, understand the effects of the variant in cellular or animal models.

Finally, the datasets presented in this thesis offered an exceptional opportunity to test the effect of non-genetic effects on vaccine responses. Notably, sex was found to have very little consistent effect on vaccine responses despite multiple prior observations of an increase in antibody responses against measles, HBsAg and influenza in females²⁶³⁻²⁶⁶. Furthermore, contrary to earlier reports²⁵⁷, exposure to HIV *in utero* did not appear have any effect on antibody vaccine responses in infants either 6 or twelve months following birth. Together these observations suggest that any true effects of sex or HIV exposure *in utero* may either be limited to specific vaccine types or specific immunological responses not measured in the African populations presented here.

In summary, the findings from the meta-analyses presented in this chapter demonstrate strong evidence in favour of an association between genetic variants across the class II MHC and responses against five vaccine antigens delivered in infancy: PT, FHA, PRN, DT and HBsAg. There is more limited evidence in support of an association between an intronic variant in the *ANO4* gene and response to Hib. No significant evidence of association was observed for variants

either inside or outside of the MHC region with any of the other measured traits and therefore the architecture and implications of MHC variation will be the focus of the following chapters.

Chapter 4

4 Framework and Population-Specific Resource Development for Fine-Mapping MHC Associations with Vaccine Responses in African Infants

4.1 Background

The observations of strong genetic signals of association between variants across the MHC and antibody responses to PT, FHA, PRN, DT and HBsAg in African infants confirm the importance of the complex region, and the class II locus in particular, in the development of immune responses that may be protective against future infectious challenge. Such associations are consistent with those observed for other infectious and inflammatory conditions where diversity at the *HLA* gene level is postulated to explain the majority of the variation^{164,267}. The diversity of *HLA* genes in the African populations studied in this thesis is poorly understood

and much of the currently available data is limited to specific populations with particular emphases on class I loci²⁶⁸. Furthermore, methods for imputing classical HLA alleles using genotype data have infrequently been tested in African populations and the reference panels of individuals with genotype and classical HLA type data available for use in such algorithms are poorly representative of individuals of African descent¹⁷².

In order to characterise the diversity of *HLA* genes in the African populations studied in this thesis, classical HLA typing was undertaken in just over one quarter of studied individuals at all 11 major *HLA* genes. A novel method based on exon-targeted next-generation sequencing (using MiSeq) of HLA genes was used to achieve high throughput and reduced levels of ambiguity through improved phase resolution^{269,270}. Since no literature is available validating these methods in populations of African descent, the data from this platform was compared to data from conventional Sanger sequence generated from 47 individuals from Entebbe with paired MiSeq data available. This resource offered the opportunity to understand levels of diversity and richness of alleles present at each gene locus in the three African populations located on different extremes of the African continent. The final resource including HLA types from 697 individuals was used to firstly, test the accuracy of imputation of HLA alleles using existing HLA*IMP:02 software and reference panels¹⁷² and subsequently to re-perform imputation using dataset as a population-specific reference panel. The HLA*IMP algorithm was further customised to output amino acid variants for association testing in addition to the standard HLA alleles¹⁶⁴⁻¹⁶⁶.

Meta-analysis frameworks are not suitable for testing for independent signals of association across the HLA when individual genotype data is available²⁷¹. Therefore a pooled linear mixed model (pLMM) approach was validated to facilitate cross-population conditional analyses taking individual LD between variants, population structure and relatedness into account. The validation of this approach further permitted the undertaking of a multivariate GWAS approach of testing for association between genetic variants and multiple correlated phenotypes simultaneously²⁴⁴. This approach is proposed to increase power to identify variants associated with either one or several of the tested traits when compared to equivalent association testing in single univariate testing frameworks.

Finally, the results from the univariate pLMM analysis were combined with a manual modelling method employed in a recent study of multiple sclerosis susceptibility¹⁶⁶ to achieve five goals for each trait. These goals were: 1) to define the most likely gene locus driving the primary observed association; 2) to investigate whether other *HLA* gene loci demonstrated evidence of independent associations; 3) to understand whether HLA alleles or amino acid residues represented the causal allele at these loci; 4) to assess whether the alleles or amino acid residues explained these associations most parsimoniously at these loci; 5) to ultimately generate a likely model of genetic architecture for each of the observed HLA-associated vaccine responses.

4.2 Methods

4.2.1 HLA typing

4.2.1.1 *Sanger HLA typing*

90 unrelated ($IBD < 0.25$) individuals with post-QC genotype data available were selected at random from the Entebbe cohort and were typed at 11 *HLA* gene loci through exon sequencing (included exons listed in **Table 4.1**). Typing was performed by the Accredited Tissue Typing Laboratory at Addenbrooke's Hospital, Cambridge University Hospitals NHS Foundation Trust using the proprietary uTYPE software version 7 (Fisher Scientific, Pittsburgh, USA). The list of possible ambiguous calls were minimised by using the 'allele pair' export function in this software which lists all possible and permissible allele pair possibilities for each locus for each individual. Alleles were defined using the IMGT/HLA Release: 3.22.0 2015 October 10. Best-call allele pairs for each locus in each individual were determined based on local guidelines prioritising alleles that were 'Common and Well-Documented' (CWD)²⁷². Six duplicate samples were typed to test the within-assay concordance of this typing platform.

Table 4.1: Exons targeted for Sanger-sequence based HLA typing in 90 Ugandan individuals.

Gene	Exons sequenced
<i>HLA-A</i>	2,3,4
<i>HLA-B</i>	2,3,4
<i>HLA-C</i>	2,3,4
<i>HLA-DPA1</i>	2,3
<i>HLA-DPB1</i>	2,3,4
<i>HLA-DQA1</i>	2,3
<i>HLA-DQB1</i>	2,3
<i>HLA-DRB1</i>	2,3
<i>HLA-DRB3</i>	2
<i>HLA-DRB4</i>	2
<i>HLA-DRB5</i>	2

4.2.1.2 *MiSeq HLA typing*

To harness improved phase resolution and lower costs of next-generation sequencing, further unrelated individuals from all three African populations underwent exon focussed MiSeq (Illumina, California, USA) sequencing by Histogenetics (Ossining, New York, USA). DNA libraries from individual DNA were prepared according to MiSeq protocols with two amplification rounds tagging adaptor and index sequences followed by sequencing on a MiSeq machine according to manufacturer protocols. The resultant fastq files were processed and typed using proprietary HistoS and HistoTyper softwares (Histogenetics, New York, USA)²⁷³. Typing was performed in two batches. The first batch constituted 47 samples from Entebbe overlapping with those processed with Sanger-sequence based typing, and the second batch constituted samples from all three populations. The number of exons used for typing differed between batches (**Table 4.2**). In both batches alleles were called using IMGT release 3.20.0(Apr 2015).

Table 4.2: Exons targeted for MiSeq-sequence based HLA typing.

Gene	Batch 1 Exons sequenced	Batch 2 Exons sequenced
<i>HLA-A</i>	2,3,4	1,2,3,4
<i>HLA-B</i>	2,3,4	1,2,3,4
<i>HLA-C</i>	2,3,4,7	1,2,3,4,7
<i>HLA-DPA1</i>	2	2
<i>HLA-DPB1</i>	2	2,3
<i>HLA-DQA1</i>	2	2
<i>HLA-DQB1</i>	2,3	2,3
<i>HLA-DRB1</i>	2,3	2,3
<i>HLA-DRB3</i>	2	2
<i>HLA-DRB4</i>	2	2
<i>HLA-DRB5</i>	2	2

4.2.2 HLA nomenclature and comparisons between platforms

Throughout this thesis HLA alleles are classified according to the World Health Organisation Nomenclature Committee for Factors of the HLA System²⁷⁴ (**Figure 4.1**). All alleles have an ‘HLA’ prefix followed by a hyphenated gene name and a subsequent star separator. Two or three digit fields of between two to three digits in length each then follow separated by colons. The first field (frequently two digits in length in modern nomenclature and therefore referred to as ‘2-digit’ level of resolution) refers to the serological group. The second field (two or three digits long, often referred to as ‘four-digit’ level of resolution) refers to resultant amino acid sequences giving rise to the HLA protein. Alleles that differ in the second field must have at least one identified non-synonymous difference between them. The third field (usually two digits in length) identifies sequences divergent due to synonymous variation (equivalent to ‘6-digit’ resolution). The fourth field refers to variation within the introns (‘8-digit’ resolution). The allele may be suffixed by

a set of letters which in this thesis will only constitute a ‘N’ which denotes known ‘null’ expression of the gene product.

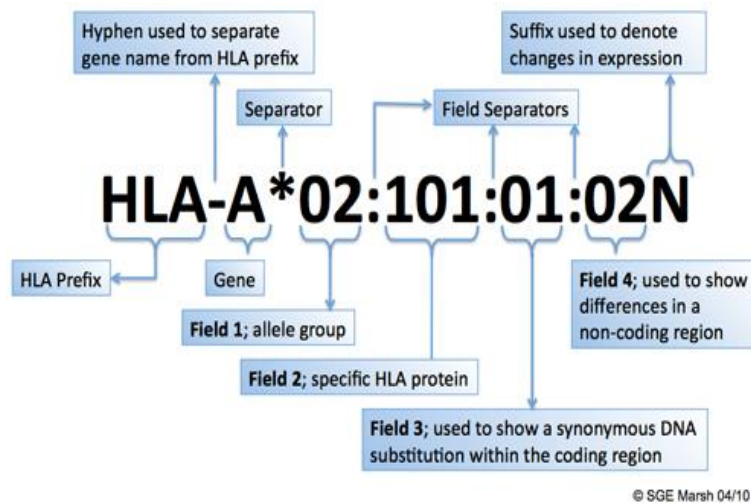


Figure 4.1: Nomenclature of the human leukocyte antigen alleles. Adapted with permission from Marsh²⁷⁴.

Traditionally, only the peptide binding domains have been targeted for defining HLA alleles (exons 2 and 3 for class I and exon 2 for class II). Therefore, true sequence diversity across all exons, and indeed introns for each gene, is relatively unknown (best appreciated through browsing the IMGT databases²⁷⁵). As a consequence of this observation, a ‘6-digit G’ level of resolution has been determined whereby alleles can be suffixed by a ‘G’ denoting that the sequence of the exons encoding the peptide binding domain of that gene would be consistent with a ‘group’ of alleles. These groups of alleles are defined according to a list maintained by the IMGT/HLA working group (http://hla.alleles.org/wmda/hla_nom_g.txt accessed on 20 April 2016). In the typing performed as part of this thesis using both Sanger and MiSeq-based

methods, improved resolution and reduced ambiguity compared to these existing ‘G’ lists was achieved through sequencing additional exons. Therefore, in some cases, strings of potential alleles not fitting the classical ‘G’ lists may be referred to in this thesis, or in other cases a single 6-digit allele call may be reported if the call could be made unambiguously. A list of class II gene ‘G’ code allele lists is provided in **Supplementary Table S2**.

Comparisons between typing platforms and imputed or typed calls were undertaken at the 4-digit level of resolution. If ambiguities were present, only the first called alleles from both platforms (almost always CWD) were used for comparison. When comparing Sanger and MiSeq-based typing methods, Sanger typing was considered the ‘Gold Standard’ and MiSeq the ‘test’ platform that allowed determination of allele calls. By comparing each individual allele in turn it was possible to define calls of the test platform that were true positives (TP), false positives (FP; called by the test platform as that allele when it was in fact another allele according to the Gold Standard), false negatives (FN; called by the test platform as another allele when it was in fact this allele), and true negatives (TN). Subsequently various metrics could be calculated. Sensitivity was defined as:

$$TP / (TP + FN)$$

Specificity was defined as:

$$TN / (TN + FP)$$

Positive predictive value (PPV) was defined as:

$$TP / (TP + FP)$$

Negative predictive value (NPV) was defined as:

$$TN / (TN + FN)$$

Accuracy was defined as:

$$(TP + TN) / (TP + FP + FN + TN)$$

Finally concordance was calculated at the level of the locus. To calculate this concordance, for every pair of chromosomes with data available in both the Gold-Standard and test platforms the number of identical allele calls between platforms was calculated and divided by the total number of alleles. Any individual with missing alleles on either or both chromosomes on either platform were excluded from these calculations.

4.2.3 HLA diversity characterisation across African populations

Wright's F_{ST} statistic is frequently used throughout genetics to describe variation between subpopulations both at the level of an individual locus and across entire genomes²⁷⁶. The fixation index is commonly derived as

$$F_{ST} = \frac{(H_T - H_S)}{H_T}$$

Where H_T is the level of heterozygosity at the locus in the entire population and H_S is the level of heterozygosity at the subpopulation level²⁷⁷. Estimates vary between 0.0 and 1.0 with 0 representing no differences in allele frequencies between populations and 1 representing fixation for alternate alleles. Calculating this statistic may be biased when population sizes are small and so estimators are frequently used as those proposed by Cockerham and Weir²⁷⁸. Such statistics are

not appropriate for multi-allelic loci where heterozygosity is invariably high and estimates will always, therefore, be lower than any theoretical maximum making their interpretation less straight forward. The G_{ST} statistic was the first variation of F_{ST} that could be generalised to multi-allelic sites specifically with the HLA locus in mind²⁷⁹ but problems including underestimation of differentiation persist which have led to the development of alternative statistics. Jost's D statistic has been proposed as an alternative that more accurately estimates diversity and differentiation. This method depends on measuring the fraction of allelic variation among populations²⁸⁰ however there still remains some uncertainty about how to interpret the outputs from such statistics²⁸¹. The 'diveRsity' software available in R allows effective calculation of each of these statistics at multi-allelic loci to enable comparison²⁸². Therefore both G_{ST} and Jost's D were calculated at all HLA gene loci in the work in this thesis describing the alleles at both the 2 and 4-digit levels of resolution. Levels of heterozygosity at each locus were thus calculated and compared to published estimates for African populations²⁸³. Furthermore, in considering multi-allelic loci it is also possible to describe diversity in terms of allelic richness. This measure reports the number of alleles observed at each loci corrected for sample size. Two methods were used to generate estimates of allelic richness at each locus using 2 and 4-digit resolution allele calls. Both methods have been developed to account for the different sizes of samples and the greater likelihood of observing rare or private alleles in populations of larger sample size. The first method depends on bootstrap sampling ($n = 1000$ times) with replacement to estimate the average number of alleles observed with standard errors. A second method depends on the rarefaction framework down-sampling all population sample-sets to that of the smallest size (in this case, Banfora) and

estimating the likelihood of observing multiple alleles in each population given that smaller sample size given the observed number of alleles in the larger original sample sets²⁸⁴. All calculations were undertaken excluding novel alleles, defining them as missing.

4.2.4 HLA imputation

The HLA*IMP:02 software was used to impute classical HLA alleles at all 11 loci in all samples with available genotype data¹⁷². HLA*IMP:02 was used preferentially in place of other software including SNP2HLA²⁸⁵ and HIBAG²⁸⁶ because of the inclusion of individuals of West African ancestry in the reference panel of HLA*IMP:02 and reported accuracies of imputation of individuals from diverse population backgrounds¹⁷². Imputation was performed by Dr Alexander Dilthey using the March 2016 release of the HLA*IMP:02 reference panel using default settings but all other downstream analyses were performed independently. All alleles were imputed to both 2- and 4-digit resolution and the ‘best-guess’ call was defined for each diploid allele in every individual using the output from the algorithm in the presence or absence of an imposed threshold for calling using the posterior probability of 0.7. Imposing this threshold is proposed to improve the quality of the total number of calls at the expense of reducing the total number of available calls.

Measures of sensitivity, specificity, NPV, PPV and accuracy were determined for each allele and concordance determined at the locus level for all 697 individuals from across the three populations with classical HLA data available.

Subsequently, the 697 individuals with genotype and classical HLA data available were used to generate an African-specific reference panel using the same methods applied to make other HLA*IMP:02 reference panels¹⁷². Instead of using the 4-digit HLA call, the 6-digit 'G' code call derived from MiSeq typing was used. In order to measure the locus-specific concordance of imputation with this new panel, samples with genotype and classical HLA allele data were randomly assigned to either a reference panel (two thirds of the dataset) or validation panel (one third of the dataset) and imputation was performed in the validation panel using the reference panel. Concordance was assessed between imputed and true types in the validation panel. This process was repeated five times.

Finally, all samples with genotype and classical HLA data available were merged into a single reference panel used to impute classical HLA alleles into all samples (including those samples contributing to the reference panel) to generate the final imputed set for all individuals from all three populations. Using these imputed allele calls the IMGT/HLA reference database was used to assign posterior probabilities for each amino acid variant in all alleles for each haplotype of each individual at all loci. Both imputed alleles and amino acid residues with associated posterior probabilities were used in downstream association testing. The probability distributions over the dosages of each variant (either amino acid or allele) were used to test association with each phenotype in the LMM implemented in GEMMA using the same GRMs calculated for the individual and pooled analyses described in chapter three. Again, any novel alleles defined through MiSeq typing were defined as missing for the purposes of concordance testing or imputation reference panel development.

4.2.5 Pooled LMM association testing, validation and multivariate association analyses

The use of a pooled analysis (also known as mega-analysis) as opposed to a meta-analysis approach offers multiple benefits including, most importantly, the opportunity to undertake multi-population conditional analyses simultaneously²⁴⁷. This approach involves pooling, or merging, individual level genetic data (if available for all populations as is the case here) and then calculating association statistics based on all individuals simultaneously rather than merging the summary statistics from the each population calculated individually. Although caution has traditionally been applied to this method owing to the need to consider intra-population differences that may introduce bias, two observations suggested that a mega-analysis approach may be effective in the African populations under study here. Firstly the LMM association testing frameworks used in the analyses presented in chapter three were observed to control effectively for population structure and relatedness and furthermore these methods are recognised for their ability to control for other non-genetic differences that may otherwise introduce systematic biases²⁴⁰. Secondly the quantitative phenotypes could readily be normalised within-population (**Figure 3.3**) making cross-comparisons highly effective. The individual level genetic data from individuals in all three populations were therefore merged using PLINK and then genetic associations between SNPs and intra-population normalised phenotypes were tested using a GRM calculated from the merged dataset including only directly genotyped SNPs. The resultant p-value association statistics were then compared to output from the fixed effects meta-analysis approach determined using METASOFT (version

2.0.0) using the Pearson correlation coefficient and showing correlations graphically using the R software.

Association statistics were calculated for SNPs after including the index SNP dosages as fixed effect covariates in the model to assess for evidence of interdependence whilst taking differential LD patterns into account across all populations.

The pLMM approach was also used to undertake a multivariate LMM analysis testing for association between genetic variants and correlated vaccine responses simultaneously. The vaccine responses tested included those delivered together: PT, FHA, PRN, DT, TT, Hib and HBsAg. Only directly genotyped variants were tested for association using the multivariate function available in the GEMMA software (version 0.94)²⁴⁴.

4.2.6 Fine-mapping HLA associations

To define the most likely locus driving the primary observed associations with each vaccine response phenotype, the pLMM approach was used to define statistically significant independent peaks of association incorporating all genotyped and imputed SNPs, imputed HLA alleles and imputed amino-acid residues in all individuals simultaneously. This step addressed the first aim of five outlined in the Background section above; aiming to fine-map the MHC associations. Stepwise conditional modelling was then performed in GEMMA for each trait to identify independently associated loci achieving a significance threshold of $p \leq 5 \times 10^{-9}$ (aim 2). This approach resulted in a range of SNPs, HLA alleles or amino acids that were determined to be independently associated with

each trait, frequently spanning multiple loci across the class II region. The gene origins of these independent index variants were determined (SNP or amino acid residues in HLA-DRB1 for example) and the dosages of all variants were then incorporated in a manual modelling approach to achieve the third to fifth aims of this fine-mapping exercise.

An approach similar to that used by Moutsianas and colleagues investigating the effect of HLA in multiple sclerosis¹⁶⁶ was used to compare and contrast the results of both manual and automated step-wise linear modelling approaches restricting the interpretations guided by the evidence of association from the pooled LMM analyses described above. A refined number of unrelated individuals ($IBD < 0.2$) were selected and models of association were tested using additive dosage probabilities for imputed genotype, classical allele and bi-allelic amino acid residues across all 11 loci with a population average minor allele frequency (MAF_{AV}) greater than 0.01. Null models were defined for each trait by including the first five principal components and the ‘time between sampling and most recent vaccination’ covariate. Independent index variants (i.e. those with the most significant, smallest, p-value) discovered through the pLMM analyses with a $p \leq 5 \times 10^{-9}$ were assessed both in *univariate* (i.e. single SNP, HLA allele or bi-allelic amino acid residue variable) models or *multivariable* (i.e. defining more than one single SNP, HLA allele or amino acid residue) models. Models were rationally tested and compared based on the known associations between amino acid residues and classical alleles. For example, an arginine at position 74 in the HLA-DRB1 protein (designated DRB1-74Arg) is only found in alleles in the 2-digit HLA-DRB1*03 allele group. Using the 6-digit ‘G’ resolution the only allele groups therefore containing DRB1-74Arg include HLA-DRB1*03:02:01 and

HLA-DRB1*03:01:01G. Each model defined using this framework was tested and compared. Using the given example univariate models comparing the DRB1-74Arg and HLA-DRB1*03 variants, and a conditional model including both HLA-DRB1*03:02:01 and HLA-DRB1*03:01:01G would be compared. All models included the same principal components and time covariates as defined in the null model for each trait. The models were compared to the null using the likelihood ratio test (LRT), if the models were nested, or using the Bayesian Information Criterion (BIC) otherwise although BIC values were reported throughout. Models with lower BIC values or those with log likelihoods closer to 0 were interpreted to explain the variance in the observed data most parsimoniously.

Finally, any prior knowledge from the associations derived from the LMM associations were removed and automated bidirectional stepwise model selection based on the BIC was undertaken to test whether models incorporating amino acid residues or classical alleles best explained each trait at each locus (aim 4) and also whether any other variants should be considered in a final model other than those identified using the manual approach above. A consensus model was then determined based on the results of the manual and automated approaches for each trait (aim 5). Manual and automated modelling steps were performed in R.

Given the relatively small size of the dataset compared to existing efforts for other diseases including multiple sclerosis¹⁶⁶ and inflammatory bowel disease¹⁶⁵ only additive models of association were tested. Once the final models were generated, the identified associations were tested to determine whether a dominant model of inheritance resulted in a better fit compared to the additive model. Due to

proposed lack of power, any further tests of deviation from additivity or interaction between HLA variants were not undertaken.

4.3 Results

4.3.1 Concordance between HLA typing platforms

In the 47 individuals from Entebbe tested, 100% concordance between the classical HLA calls determined using the Sanger and MiSeq-based typing platforms was observed at the *HLA-A*, *HLA-DRB1*, *HLA-DRB3* and *HLA-DRB5* loci. More variable concordance estimates were observed at the remaining loci varying from 67% for *HLA-DPB1* to 99% for *HLA-DQB1* (Table 4.3).

Table 4.3: Concordance between allele types generated using Sanger and MiSeq based sequencing using samples from 47 individuals from Entebbe.

Gene	Number Validated [†]	4-digit Concordance
<i>HLA-A</i>	94	1
<i>HLA-B</i>	94	0.97
<i>HLA-C</i>	90	0.93
<i>HLA-DPA1</i>	94	0.93
<i>HLA-DPB1</i>	94	0.67
<i>HLA-DQA1</i>	94	0.79
<i>HLA-DQB1</i>	94	0.99
<i>HLA-DRB1</i>	94	1
<i>HLA-DRB3</i>	92	1
<i>HLA-DRB4</i>	94	0.98
<i>HLA-DRB5</i>	94	1

[†]: The number of alleles (n = 2 per individual) compared with non-missing data.

Of 6 duplicate samples sent for Sanger-based HLA typing there was 100% concordance between calls at all loci. With the exception of one allele in *HLA-*

DPA1 all differences observed between MiSeq and Sanger-based typing methods could be explained by either reduced ambiguity made possible through strand phasing with the MiSeq method or a difference in exon coverage. No differences were observed that were attributable to differences in IMGT/HLA data releases used. A detailed description of the discrepancies observed between the two platforms now follows.

4.3.1.1 *HLA-B*

Three instances of discrepant HLA calls were observed at *HLA-B* between the two typing platforms (**Table 4.4**). One individual had discrepancies on both of their chromosomes (HLA-B*45:01 / HLA-B*53:01 by MiSeq and HLA-B*35:01 / HLA-B*44:18 by Sanger). The HLA alleles determined using MiSeq were reported unambiguously. In contrast, the allele calls determined using Sanger typing were listed along with other potential alternative calls. The increased numbers of alternative calls generated from the Sanger-based typing method were a result of an inability to resolve cis-/trans- phase which was possible using the MiSeq typing method. One individual was found to have a novel allele (a non-synonymous mutation in exon 3 at AA-97 AGG(R)->TGG(W) for which the closest allele is HLA-B*15:03:01G) using MiSeq. This substitution was not detected using the Sanger typing method and HLA-B*15:03 was called instead.

Table 4.4: Discordant alleles typed at *HLA-B* locus using Sanger and MiSeq based typing.

Individual	Sanger		MiSeq	
	Allele 1*	Allele 2*	Allele 1*	Allele 2*
Entebbe_a	35:01 /42/57, 45:01	44:18 , 53:01	45:01	53:01
Entebbe_b	15:03 /103/131/220	45:01 /04	15:XX [†]	45:01

: Alleles are provided at the 4-digit (2 field) resolution omitting the HLA-B prefix. Ambiguities are demonstrated either by separating the second fields with a '/' or, in the case of serological ambiguities with a ','. Allele calls used for comparison between platforms are highlighted in bold.

[†]: XX: denotes new allele.

4.3.1.2 *HLA-C*

Two individuals failed typing on MiSeq for technical reasons and therefore comparisons were not possible on four chromosomes. Five instances of calls discrepant between platforms were observed at this locus (**Figure 4.5**). Four individuals were observed to have discordant calls occurring on single chromosomes. In all such cases MiSeq reported HLA-C*02:10 unambiguously whereas the Sanger-based method reported HLA-C*02:02 as being most likely with HLA-C*02:10 listed as at least one potential alternative call. HLA-C*02:10 could be called unambiguously using MiSeq owing to the single molecular MiSeq sequencing method allowing phase resolution. One individual had alleles reported inconsistently between typing platforms on both chromosomes (HLA-C*06:02 / HLA-C*08:02 using MiSeq and HLA-C*05:01 / HLA-C*06:11 using Sanger). Using the phase resolution possible using MiSeq the allele on one chromosome could be called unambiguously and this allele was listed as one of many possible alternatives using the Sanger method. Although the true identity of the allele is difficult to formally test using either platform, the lack of ambiguity present in the MiSeq data suggests that the calls using MiSeq are correct.

Table 4.5: Discordant alleles typed at *HLA-C* locus using Sanger and MiSeq based typing.

Individual	Sanger		MiSeq	
	Allele 1*	Allele 2*	Allele 1*	Allele 2*
Entebbe_c	02:02 /05/10/27	16:01 /02/07/63	02:10	16:01
Entebbe_d	05:01 /20, 06:02/03 /06/28/30/67/83 /127/129/138	06:11 /12, 08:02/05/12/15 /28/29/37/94/103	06:02 /83	08:02
Entebbe_e	02:02 /10/12 /14/16/27	07:01 /06/09/18 /22/26/210/343 /419	07:01 /06/18 /343/419	02:10
Entebbe_f	02:02 /10	04:01 /09N/30/82	04:01 /82	02:10
Entebbe_g	02:02 /05/10/14	06:02 /03/83/127	06:02 /83	02:10

: Alleles are provided at the 4-digit (2 field) resolution omitting the HLA-C prefix. Ambiguities are demonstrated either by separating the second fields with a '/' or, in the case of serological ambiguities with a ','. Allele calls used for comparison between platforms are highlighted in bold.

4.3.1.3 *HLA-DPA1*

Discrepant reported HLA allele calls were observed for four individuals. In three individuals a novel mutation was observed in exon 3 using Sanger typing that was not sequenced with MiSeq (heterozygous guanine/cytosine sequenced at position 444 as opposed to consensus homozygous cytosine/cytosine). For one other individual unambiguous calls were made using both platforms but were discordant (HLA-DPA1*01:03 / HLA-DPA1*03:01 by MiSeq and HLA-DPA1*01:03 / HLA-DPA1*01:03 by Sanger). These differences could not be reconciled.

4.3.1.4 *HLA-DPB1*

24 alleles were called using the MiSeq platform as HLA-DPB1*04:02 (with HLA-DPB1*105:01 listed as a single alternative) compared to the Sanger method that called the HLA-DPB1*105:01 allele on the same chromosomes unambiguously. The difference between these 2 alleles is a cytosine at DNA position 619 in exon 3 in HLA-DPB1*105:01 rather than adenine in HLA-

DPB1*04:02. It was not possible to distinguish these alleles in this batch of MiSeq typing since exon 3 was not sequenced. A similar issue of differences in exon coverage employed between the two methods was responsible for three other ambiguity issues: three individuals where HLA-DPB1*104:01 was called unambiguously by Sanger and HLA-DPB1*03:01 was called by MiSeq with HLA-DPB1*104:01 listed as a less likely alternative; one individual with HLA-DPB1*19:01 called instead of HLA-DPB1*106:01 which was called unambiguously by Sanger; and one final case where HLA-DPB1*131:01 was called unambiguously by Sanger but the call HLA-DPB1*17:01 was made instead by MiSeq.

4.3.1.5 *HLA-DQA1*

15 chromosomes had allele calls of HLA-DQA1*05:01 made using the MiSeq platform (listing a range of potential alternative calls including HLA-DQA1*05:05) whereas HLA-DQA1*05:05 was called unambiguously by Sanger. This reduction in ambiguity achieved with the Sanger method was achieved through the additional sequencing of exon 3. Similarly, 4 alleles were called using the MiSeq platform as HLA-DQA1*03:01 with other alleles listed as an alternatives, whereas through the sequencing of an additional exon, this allele was demonstrated to be more likely to be DQA1*03:02 with an alternative being DQA1*03:03 using the Sanger method.

4.3.1.6 *HLA-DQB1*

Only one discrepancy between calls on different platforms was noted at this locus resulting from improved phase resolution with MiSeq (**Table 4.6**).

Table 4.6: Discordant alleles typed at *HLA-DQB1* locus using Sanger and MiSeq based typing.

Individual	Sanger		MiSeq	
	Allele 1*	Allele 2*	Allele 1*	Allele 2*
Entebbe_h	05:01	02:01/02/47	02:02	05:01

: Alleles are provided at the 4-digit (2 field) resolution omitting the HLA-DQB1 prefix. Ambiguities are demonstrated by separating the second fields with a '/'. Allele calls used for comparison between platforms are highlighted in bold.

4.3.1.7 *HLA-DRB4*

Two individuals were called unambiguously by MiSeq as possessing the null expressed HLA-DRB4*01:03N allele whereas this allele could not be resolved using Sanger sequencing which instead classified this as being most likely HLA-DRB4*01:01 with HLA-DRB4*01:03N being listed as a potential alternative. This difference was due to different primer selection with the exon 2 primer used in the MiSeq method being capable of detecting the acceptor splice site variant present in the null-expressed allele²⁸⁷.

4.3.2 HLA typing 697 individuals from three African populations

The MiSeq-based platform was used to type classical HLA alleles of all 11 genes of 697 individuals from the Entebbe (n = 330), Soweto (n = 200) and Banfora (n = 167) populations including the sequence data from an additional exon in *HLA-DPB1* (exon 3) to further improve resolution at this polymorphic locus. 13 new alleles were detected in 28 individuals (with some alleles being observed more than once; **Table 4.7**). With the exception of one individual, all individuals who carried a new allele only had a single new allele at one locus. Two alleles (reported types HLA-DPA1*02:XX and HLA-DQA1*04:XX) were observed in multiple unrelated individuals in two populations (Soweto and Entebbe). No new alleles were observed in individuals from Banfora and only one new allele was detected in the class I genes.

Table 4.7: New HLA alleles detected through MiSeq across the three tested populations.

Locus	Sample ID	Reported Type	Mutation details	Mutation Type	Closest allele
C	Entebbe_1	C*07:43P [†]	AA-99 TAT(Y)->TAC(Y)	Synonymous	C*07:43
DPA1	Soweto_1	DPA1*01:XX	AA-84: AAT(N)->AGT(S)	Non-Synonymous	DPA1*01:05
	Soweto_2*	DPA1*01:XX	AA-15: ACG(T)->ACC(T), AA-50: CAA(Q)->CGA(R)	Non-Synonymous	DPA1*01:06:01
	Entebbe_2	DPA1*01:03P	AA-10 GCC(A)->GCT(A)	Synonymous	DPA1*01:03:01G
	Soweto_3	DPA1*02:XX	AA-76 CGT(R)->TGT(C)	Non-Synonymous	DPA1*02:01:01
	Soweto_4				
	Soweto_5				
	Soweto_6				
	Entebbe_3				
	Entebbe_4				
	Soweto_7	DPA1*02:XX	AA-83: GCC(A)->ACC(T), AA-84: AAT(N)->AAC(N)	Non-Synonymous	DPA1*02:02:02
	Soweto_8	DPA1*03:XX	AA-50: CAA(Q)->CGA(R), AA-83: ACC(T)->GCC(A)	Non-Synonymous	DPA1*03:02
DPB1	Soweto_9	DPB1*19:XX	AA-75: AGG(R)->GGG(G)	Non-Synonymous	DPB1*19:01:01G
	Entebbe_8	DPB1*158:XXX	AA-69 AAG(K)->GAG(E)	Non-Synonymous	DPB1*158:01

*: Individuals with more than one new allele are highlighted in red.

[†]: A 'P' suffix represents a set of alleles with identical peptide sequences across the peptide-binding domain, but differing nucleotide sequences.

Table 4.7 continued

DQA1	Entebbe_9	DQA1*01:XX	AA-87 AAT(N)->ACT(T)	Non-Synonymous	DQA1*01:02:01G
	Entebbe_10				
	Entebbe_11	DQA1*04:XX	AA-51: CTC(L)->CAC(H)	Non-Synonymous	DQA1*04:01:01G
	Soweto_10				
	Soweto_11				
	Soweto_12				
	Soweto_13				
	Soweto_2*				
	Soweto_14				
	Soweto_15				
DQB1	Entebbe_12	DQB1*02:XX	AA-21 ACA(T)->GCA(A)	Non-Synonymous	DQB1*02:01:01G
DRB1	Soweto_16	DRB1*03:XX	AA-56: CCT(P)->CAT(H)	Non-Synonymous	DRB1*03:02:01

*: Individuals with more than one new allele are highlighted in red.

Estimates of differentiation between populations describing HLA alleles at the 2-digit level of resolution were consistently higher for the class I HLA alleles compared to class II with distances estimated to be greater between Entebbe and Banfora compared to other pairwise comparisons at all loci (**Figures 4.2 to 4.13**). Differentiation was particularly striking at the *HLA-B* locus. Estimates of genetic differentiation calculated using G_{ST} and Jost's D using 2-digit descriptors were equivalent across all loci. The differentiation persisted through to the 6-digit 'G' descriptors as demonstrated with the site-frequency spectra of the more common alleles (greater than 0.05 in any single population). Average measures of allelic richness correcting for sample size and the presence of rare alleles calculated using bootstrap resampling (**Figure 4.13**) were equivalent to estimates generated through the rarefaction method (**Table 4.8**). The bootstrap method offered the advantage of calculating confidence intervals. These intervals demonstrated a significant increase in allelic diversity in the Entebbe population at *HLA-B* compared to Soweto at the 2-digit level of resolution, and compared to both Soweto and Banfora at the 4-digit level. The Entebbe and Soweto populations were similarly observed to have significantly greater richness of alleles at the *HLA-A* locus when described at the 4-digit level of resolution. Otherwise there were few differences between populations across other loci. With heterozygosity estimates in the region of 90-95% for class I alleles in all populations (**Table 4.9 and Table 4.10**) estimates at the 4-digit level of resolution are comparable to previous reports of those in other diverse African populations²⁸³ supporting the occurrence of diversifying selection at these polymorphic loci.

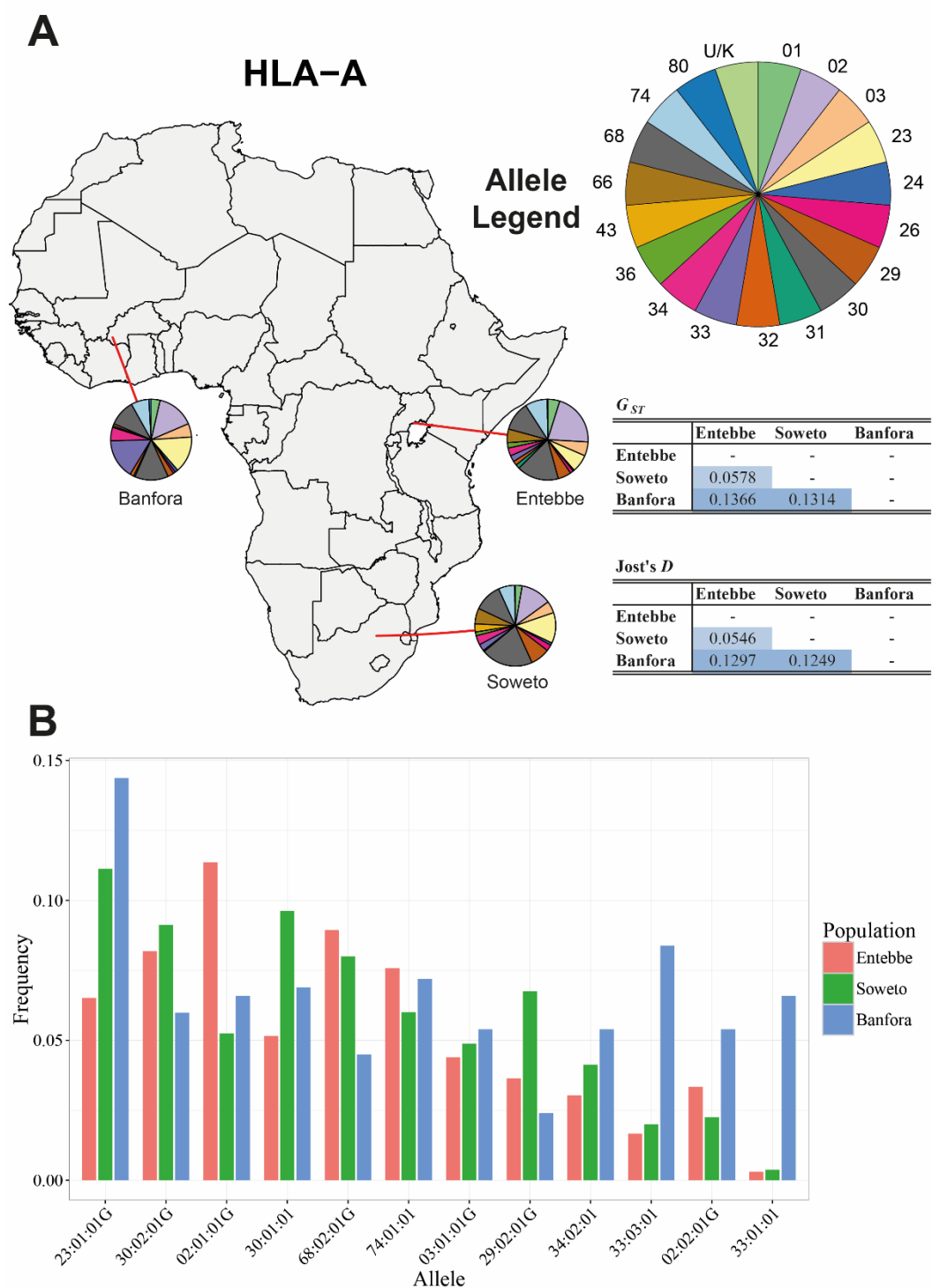


Figure 4.2: Diversity of HLA-A alleles by population. A) Frequencies of all directly typed HLA-A alleles at 2-digit resolution represented as pie charts with tabulated estimates of G_{ST} and Jost's D shaded in blue categories of increasing hue: 0 – 0.05 – 0.1 – 0.2 – 0.3. U/K: unknown. B) Site frequency spectrum of 6-digit 'G' alleles with a frequency of at least 0.05 in a single population ordered in decreasing frequency averaged across all three populations.

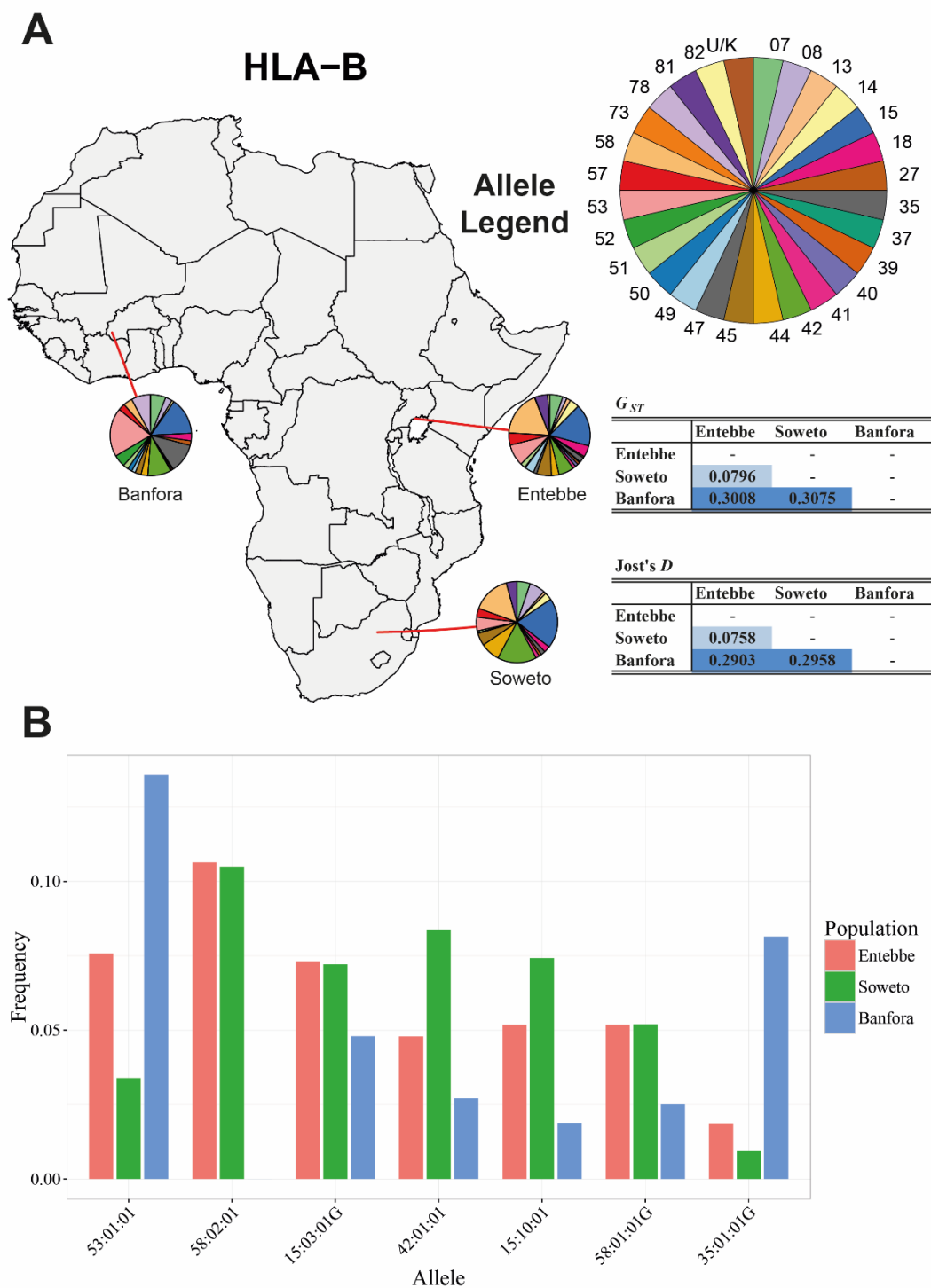


Figure 4.3: Diversity of HLA-B alleles by population. A) Frequencies of all directly typed HLA-B alleles at 2-digit resolution represented as pie charts with tabulated estimates of G_{ST} and Jost's D shaded in blue categories of increasing hue: 0 – 0.05 – 0.1 – 0.2 – 0.3. U/K: unknown. B) Site frequency spectrum of 6-digit 'G' alleles with frequency of at least 0.05 in a single population ordered in decreasing frequency averaged across all three populations.

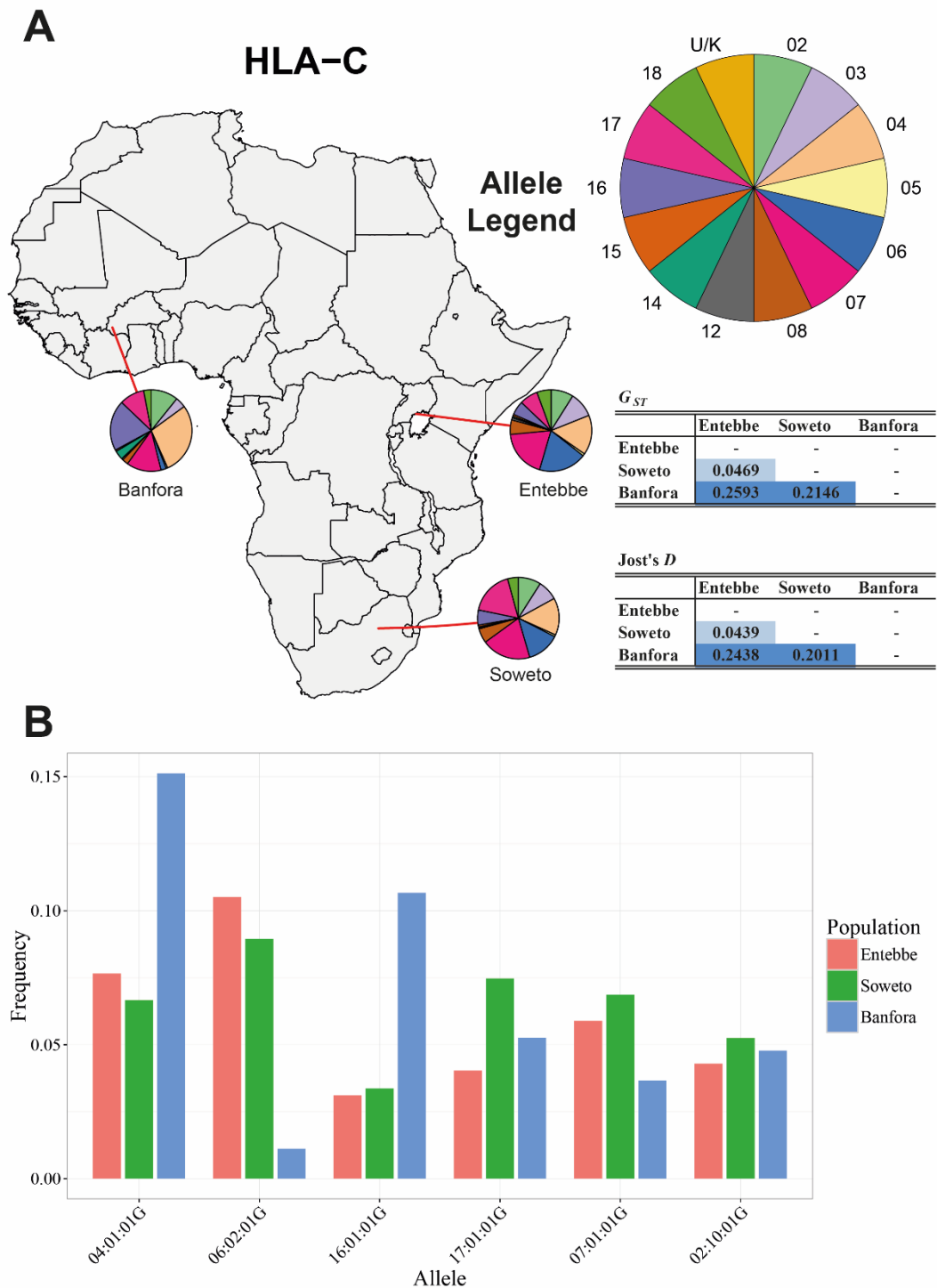


Figure 4.4: Diversity of HLA-C alleles by population. A) Frequencies of all directly typed HLA-C alleles at 2-digit resolution represented as pie charts with tabulated estimates of G_{ST} and Jost's D shaded in blue categories of increasing hue: 0 – 0.05 – 0.1 – 0.2 – 0.3. U/K: unknown. B) Site frequency spectrum of 6-digit 'G' alleles with a frequency of at least 0.05 in a single population ordered in decreasing frequency averaged across all three populations.

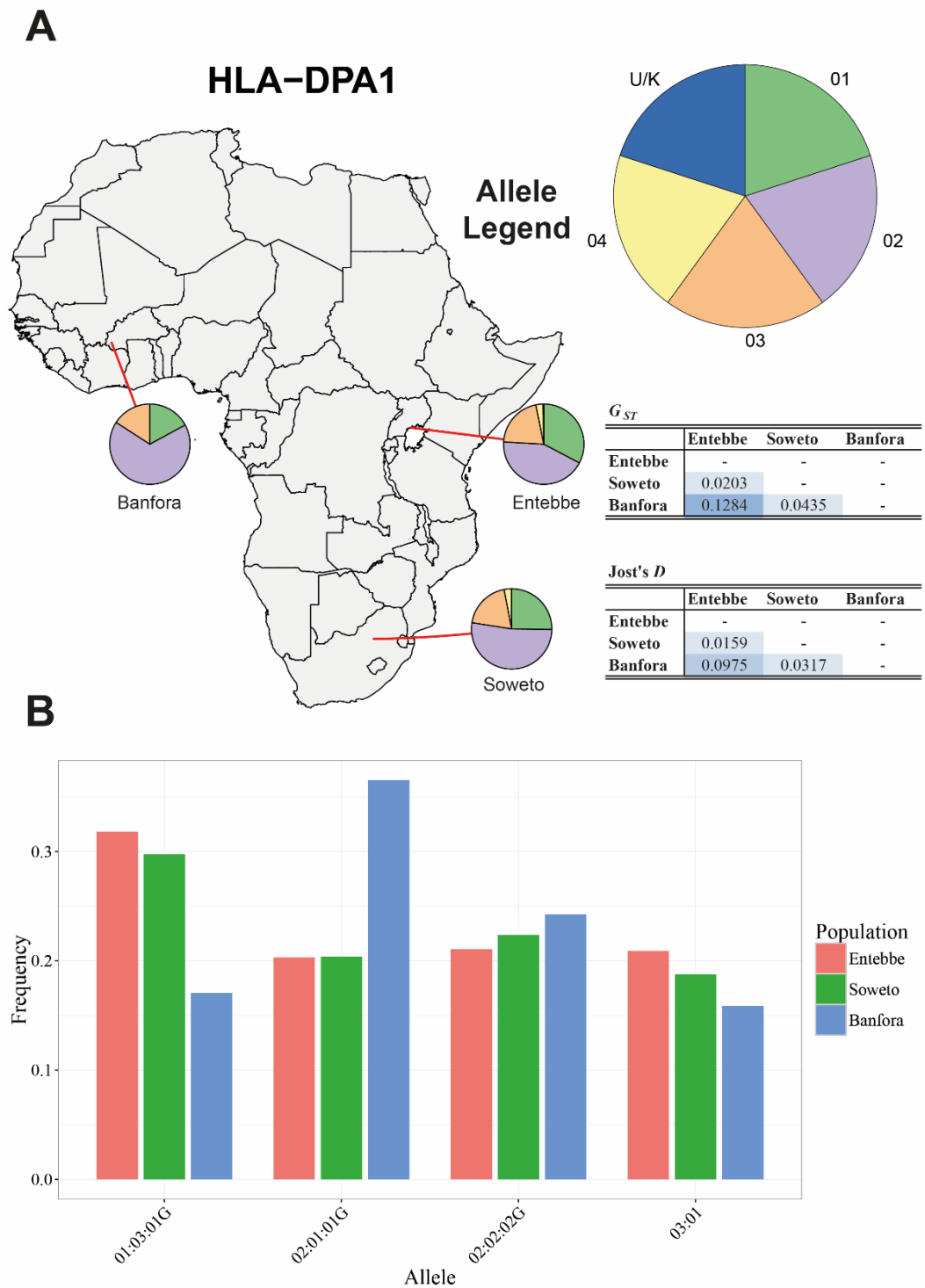


Figure 4.5: Diversity of HLA-DPA1 alleles by population. A) Frequencies of all directly typed HLA-DPA1 alleles at 2-digit resolution represented as pie charts with tabulated estimates of G_{ST} and Jost's D shaded in blue categories of increasing hue: 0 – 0.05 – 0.1 – 0.2 – 0.3. U/K: unknown. B) Site frequency spectrum of 6-digit 'G' alleles with a frequency of at least 0.05 in a single population ordered in decreasing frequency averaged across all three populations.

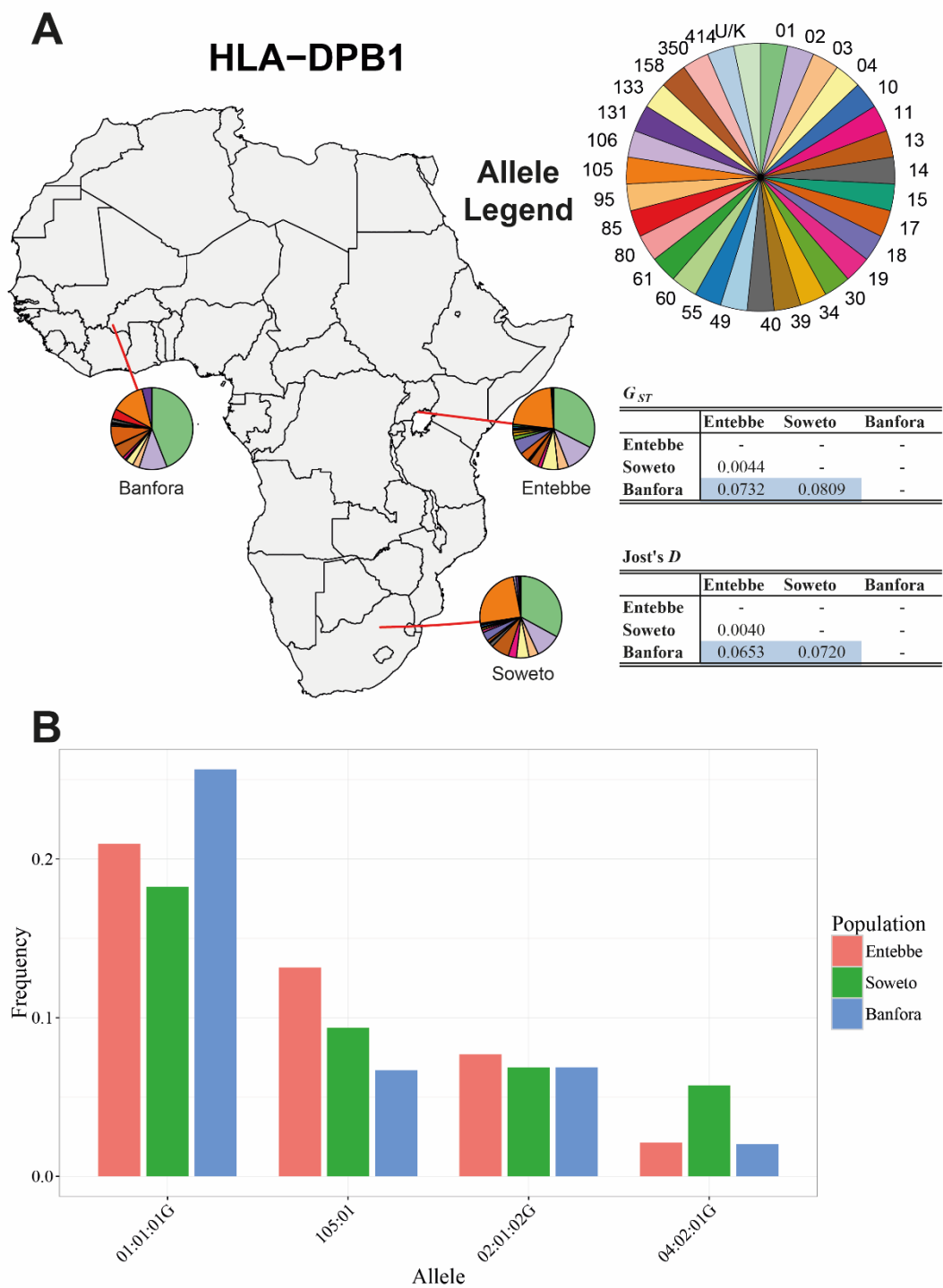


Figure 4.6: Diversity of HLA-DPB1 alleles by population. A) Frequencies of all directly typed HLA-DPB1 alleles at 2-digit resolution represented as pie charts with tabulated estimates of G_{ST} and Jost's D shaded in blue categories of increasing hue: 0 – 0.05 – 0.1 – 0.2 – 0.3. U/K: unknown. B) Site frequency spectrum of 6-digit 'G' alleles with a frequency of at least 0.05 in a single population ordered in decreasing frequency averaged across all three populations.

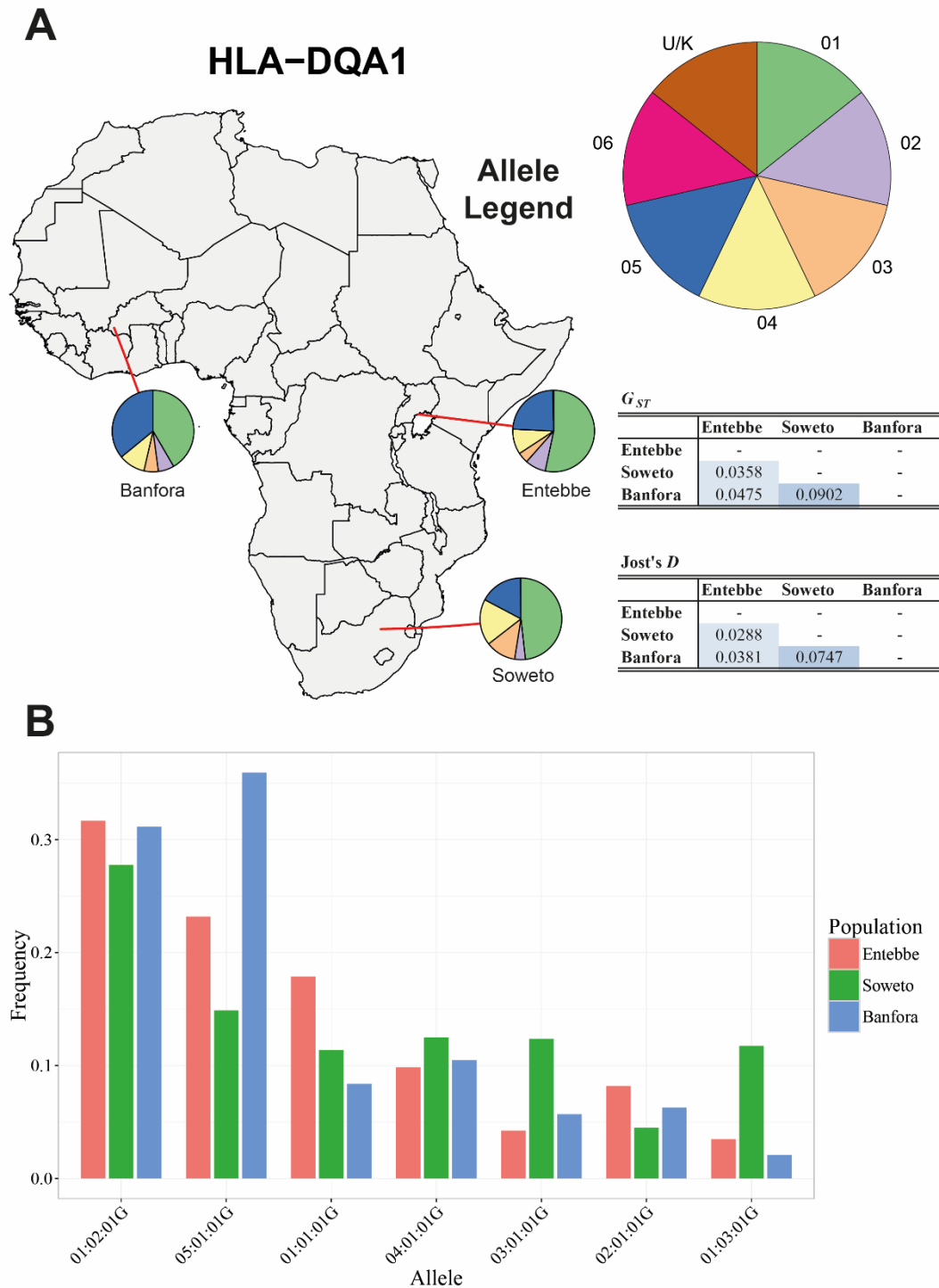


Figure 4.7: Diversity of HLA-DQA1 alleles by population. A) Frequencies of all directly typed HLA-DQA1 alleles at 2-digit resolution represented as pie charts with tabulated estimates of G_{ST} and Jost's D shaded in blue categories of increasing hue: 0 – 0.05 – 0.1 – 0.2 – 0.3. U/K: unknown. B) Site frequency spectrum of 6-digit 'G' alleles with a frequency of at least 0.05 in a single population ordered in decreasing frequency averaged across all three populations.

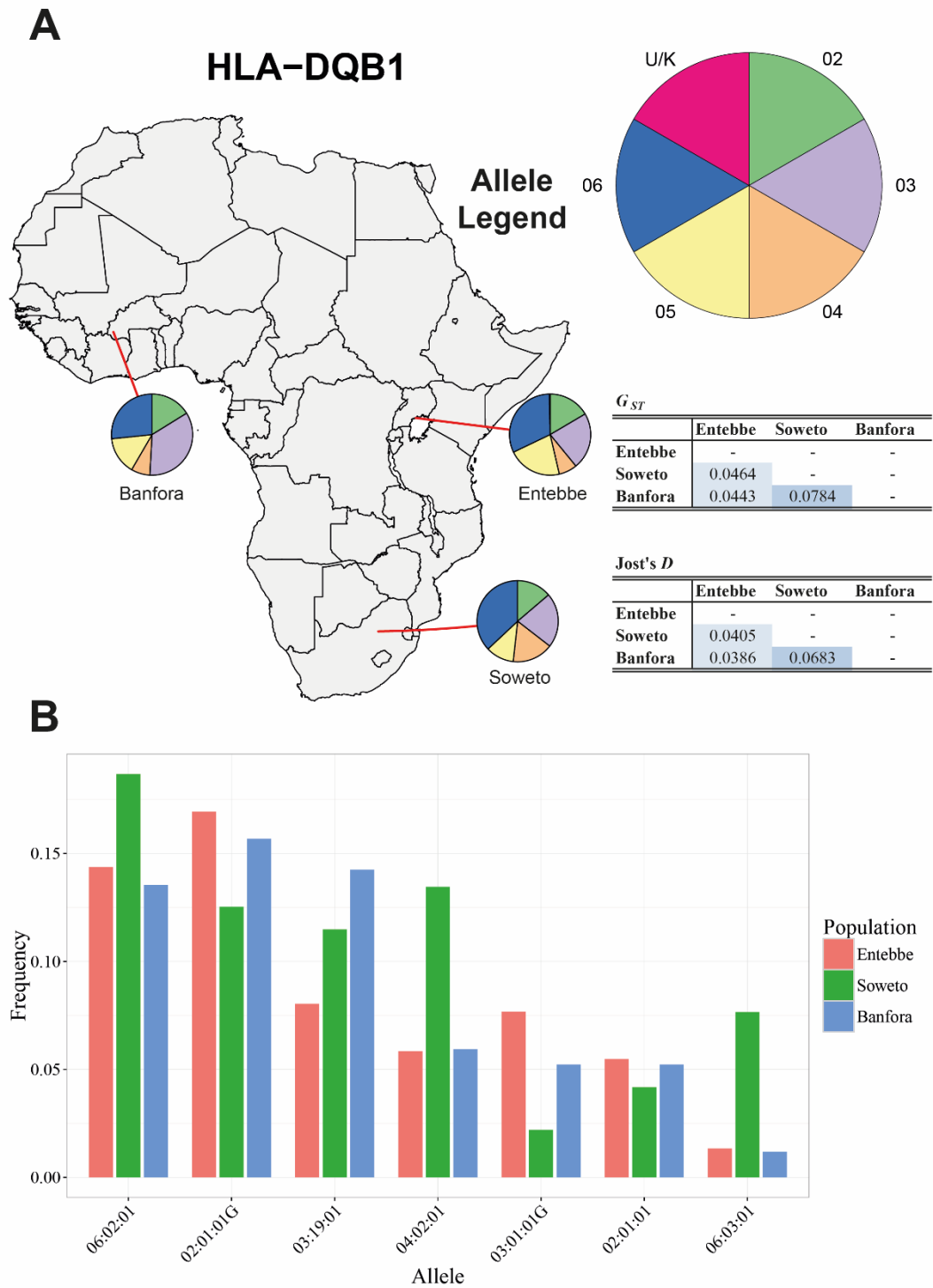


Figure 4.8: Diversity of HLA-DQB1 alleles by population. A) Frequencies of all directly typed HLA-DQB1 alleles at 2-digit resolution represented as pie charts with tabulated estimates of G_{ST} and Jost's D shaded in blue categories: 0 – 0.05 – 0.1 – 0.2 – 0.3. U/K: unknown. B) Site frequency spectrum of 6-digit 'G' alleles with a frequency of at least 0.05 in a single population ordered in decreasing frequency averaged across all three populations.

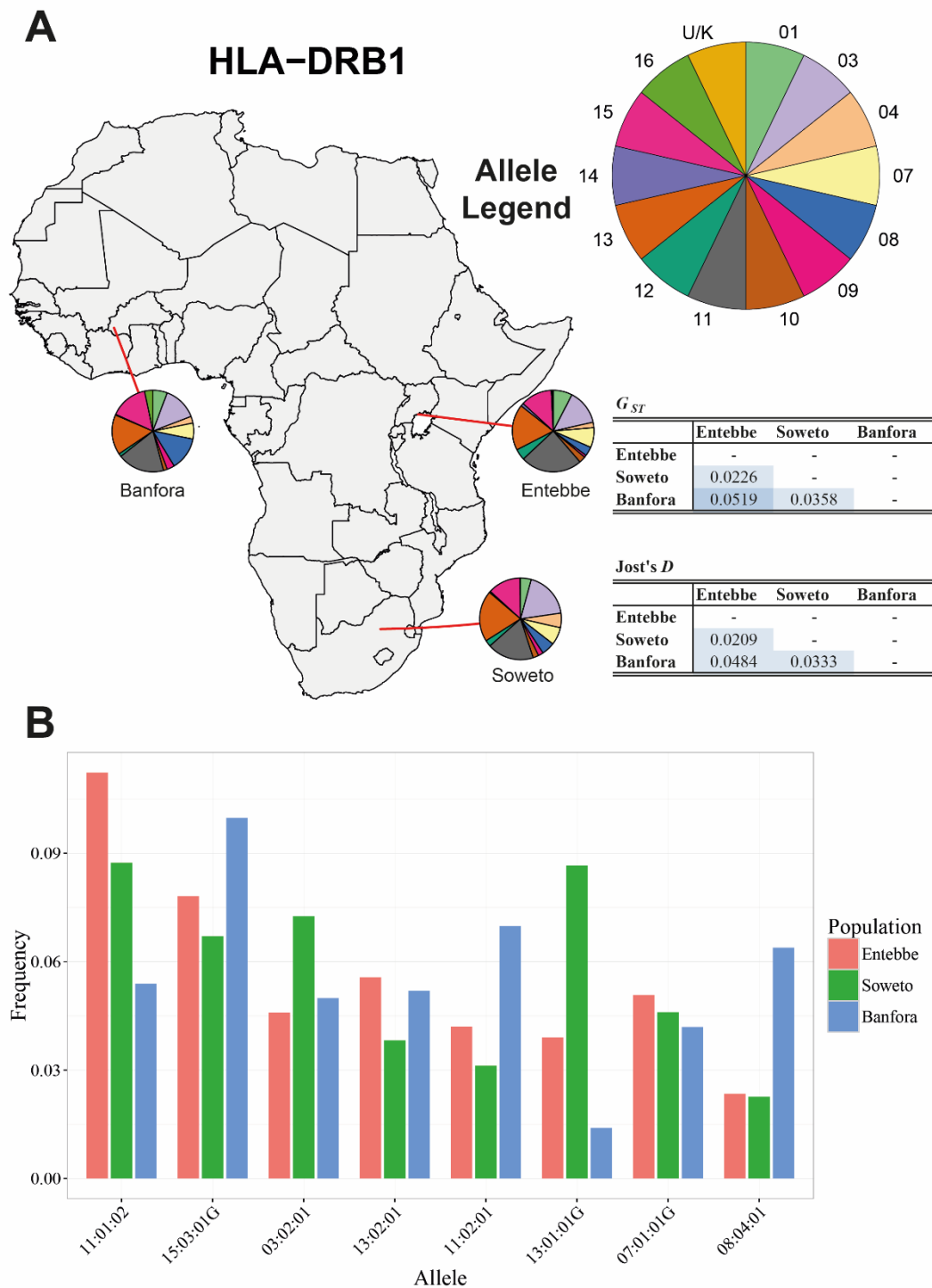


Figure 4.9: Diversity of HLA-DRB1 alleles by population. A) Frequencies of all directly typed HLA-DRB1 alleles at 2-digit resolution represented as pie charts with tabulated estimates of G_{ST} and Jost's D shaded in blue categories: 0 – 0.05 – 0.1 – 0.2 – 0.3. U/K: unknown. B) Site frequency spectrum of 6-digit 'G' alleles with a frequency of at least 0.05 in a single population ordered in decreasing frequency averaged across all three populations.

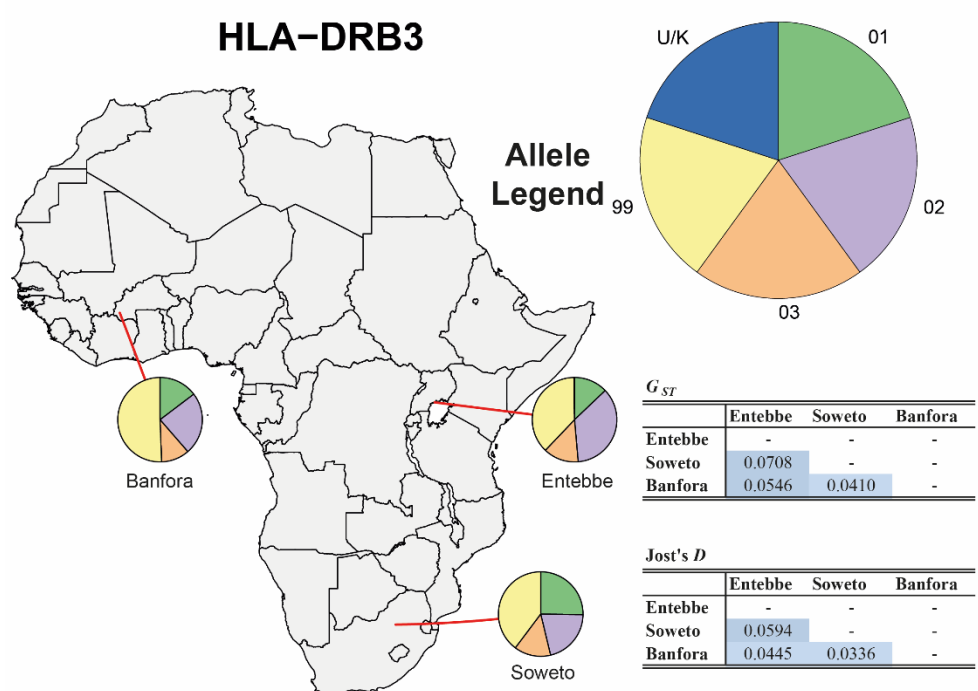


Figure 4.10: Diversity of 2-digit HLA-DRB3 alleles by population. Tabulated estimates of G_{ST} and Jost's D shaded in blue categories: 0 – 0.05 – 0.1. U/K: unknown. 99: absent on haplotype.

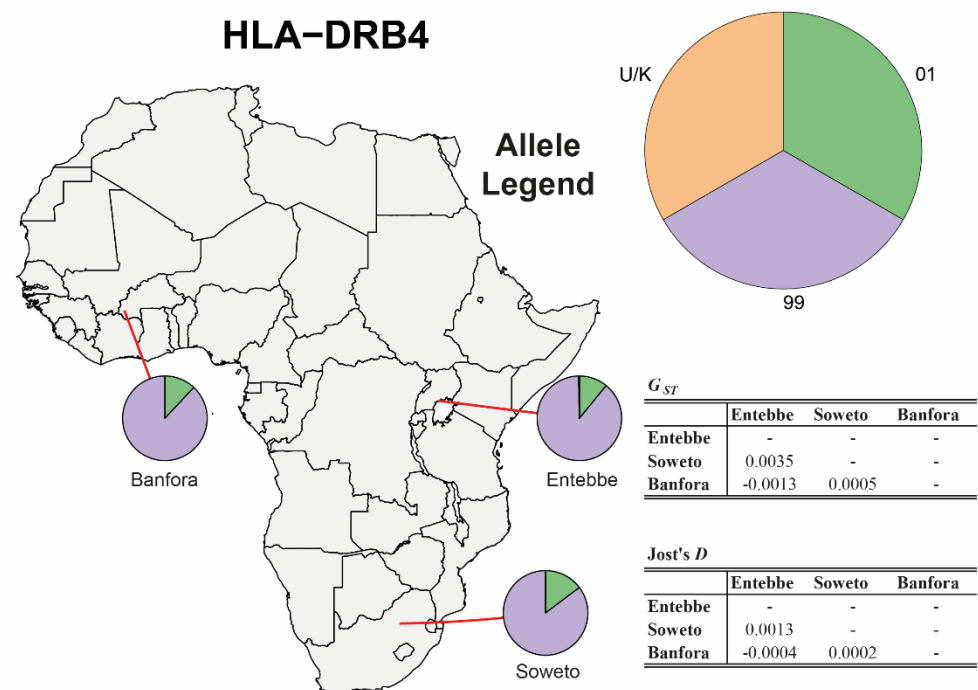


Figure 4.11: Diversity of 2-digit HLA-DRB4 alleles by population. Tabulated estimates of G_{ST} and Jost's D shaded in blue categories: 0 – 0.05 – 0.1. U/K: unknown. 99: absent on haplotype.

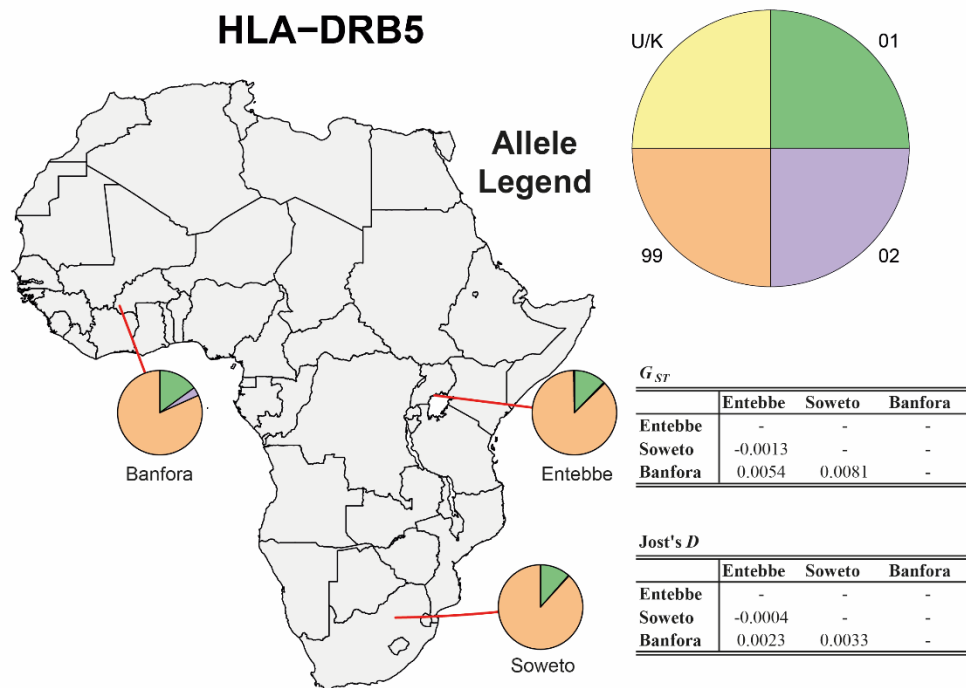


Figure 4.12: Diversity of 2-digit HLA-DRB5 alleles by population. Tabulated estimates of G_{ST} and Jost's D shaded in blue categories: 0 – 0.05 – 0.1. U/K: unknown. 99: absent on haplotype.

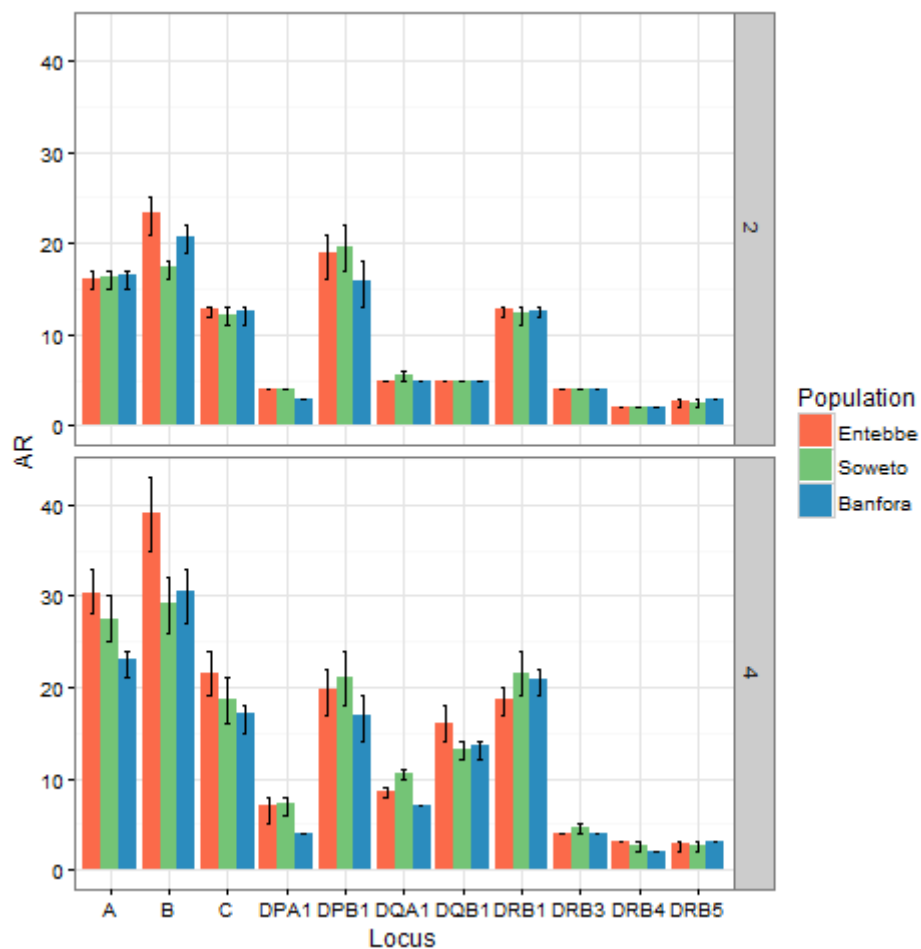


Figure 4.13: Allelic richness of HLA alleles at 2 and 4-digit resolution across the three African populations calculated using bootstrap resampling. An estimate of allelic richness was calculated for each genic locus in each of the three studied African populations using 6-digit ‘G’-resolution HLA types refined to either 2- (top panel) or 4-digit (bottom) resolution. Calculations were performed using the Dive’R’sity package accounting for rare alleles and calculating confidence intervals using bootstrap resampling (n=1000).

AR: allelic richness.

Table 4.8: Allelic richness estimates at all HLA loci calculated using the rarefaction method at 2 and 4-digit levels of resolution. Point estimates of allelic richness are provided per gene locus and population level using 6-digit ‘G’ types refined to either 2- (upper table) and 4-digit (lower) level of resolution using the rarefaction method implemented in the Dive’R’sity package.

2-digit											
Population	A	B	C	DPA1	DPB1	DQA1	DQB1	DRB1	DRB3	DRB4	DRB5
Entebbe	16.16	23.36	12.71	4.00	18.82	5.00	5.00	12.74	4.00	2.00	2.79
Soweto	16.33	17.38	12.23	4.00	19.74	5.55	5.00	12.37	4.00	2.00	2.57
Banfora	16.53	20.71	12.59	3.00	15.87	5.00	5.00	12.63	4.00	2.00	3.00
4-digit											
Entebbe	30.44	39.02	21.49	7.09	19.85	8.69	15.97	18.8	4.00	2.99	2.77
Soweto	27.57	29.25	18.69	7.37	21.05	10.51	13.09	21.50	4.55	2.56	2.55
Banfora	23.04	30.56	17.03	4.00	16.77	7.00	13.72	21.01	4.00	2.00	3.00

Table 4.9: Heterozygosity estimates based on 2-digit allele data across all loci. Estimates of heterozygosity at each loci calculated using 6-digit ‘G’ types refined to 2-digit resolution calculated using the Dive’R’sity package.

Population	N*	HLA-A		HLA-B		HLA-C		HLA-DRB1		HLA-DQB1		HLA-DPB1	
		K†	H‡	K†	H‡	K†	H‡	K†	H‡	K†	H‡	K†	H‡
Entebbe	330	17	0.86	25	0.89	13	0.87	13	0.88	5	0.76	22	0.82
Soweto	200	17	0.89	18	0.89	13	0.87	13	0.86	5	0.74	23	0.78
Banfora	167	17	0.9	22	0.87	13	0.81	13	0.84	5	0.77	18	0.75

*: N: number of individuals typed at all loci

†: K: number of distinct alleles

‡: H: heterozygosity estimate corrected for sample size

Table 4.9: Continued

Population	N*	HLA-DPA1		HLA-DQA1		HLA-DRB3		HLA-DRB4		HLA-DRB5	
		K†	H‡	K†	H‡	K†	H‡	K†	H‡	K†	H‡
Entebbe	330	4	0.64	5	0.61	4	0.70	2	0.20	3	0.22
Soweto	200	4	0.6	6	0.69	4	0.68	2	0.24	3	0.22
Banfora	167	3	0.5	5	0.71	4	0.7	2	0.18	3	0.32

*: N: number of individuals typed at all loci

†: K: number of distinct alleles

‡: H: heterozygosity estimate corrected for sample size

Table 4.10: Heterozygosity estimates based on 4-digit allele data across all loci. Estimates of heterozygosity at each loci calculated using 6-digit ‘G’ types refined to 4-digit resolution calculated using the Dive’R’sity package.

Population	N*	HLA-A		HLA-B		HLA-C		HLA-DRB1		HLA-DQB1		HLA-DPB1	
		K†	H‡	K†	H‡	K†	H‡	K†	H‡	K†	H‡	K†	H‡
Entebbe	330	33	0.92	48	0.94	25	0.90	20	0.92	18	0.88	23	0.82
Soweto	200	30	0.90	33	0.94	21	0.87	24	0.91	14	0.85	25	0.78
Banfora	167	24	0.93	34	0.90	18	0.83	22	0.91	14	0.89	19	0.75

*: number of individuals typed at all loci

†: number of distinct alleles

‡: heterozygosity estimate corrected for sample size

Table 4.10: Continued

Population	N*	HLA-DPA1		HLA-DQA1		HLA-DRB3		HLA-DRB4		HLA-DRB5	
		K†	H‡	K†	H‡	K†	H‡	K†	H‡	K†	H‡
Entebbe	330	8	0.76	9	0.79	4	0.70	3	0.20	3	0.22
Soweto	200	8	0.75	11	0.84	5	0.68	3	0.25	3	0.22
Banfora	167	4	0.71	7	0.76	4	0.70	2	0.18	3	0.32

*: number of individuals typed at all loci

†: number of distinct alleles

‡: heterozygosity estimate corrected for sample size

4.3.3 African-specific HLA imputation reference panel validation

A high concordance (> 0.9) between 4-digit HLA allele calls determined using imputation (using HLA*IMP:02) and classical typing was only demonstrated for all three cohorts at those loci demonstrating low levels of diversity and allelic richness (*HLA-DRB3*, *-DRB4* and *-DRB5*; **Tables 4.11 to 4.13**). Loci with increased levels of diversity and richness demonstrated substantially lower levels of concordance (*HLA-B* and *-DPB1* in particular). The only exception to these observations was for *HLA-A* in the Banfora population which demonstrated a very high concordance compared to the other populations (0.97 with no threshold applied compared to 0.84 and 0.86 in Entebbe and Soweto respectively). This difference may partly be explained by the observation that individuals listed as being of African descent in the HLA*IMP:02 reference panel are actually of Western African origin (Yoruba from Nigeria¹⁷²). As has been reported for the HLA*IMP:02 algorithm previously, concordance could be marginally increased (at the expense of reduced numbers of calls) by applying a posterior probability threshold of 0.7 in all tested populations.

In contrast, using the new HLA*IMP:02G algorithm and reference panel, a five-fold cross validation demonstrated concordance to be over 0.9 at all loci except *HLA-B* where concordance was estimated at 0.88 (**Figure 4.14**).

Table 4.11: Concordance between classical types determined using MiSeq-based typing and imputed with HLA*IMP:02 for individuals from Entebbe. The concordance between directly typed (6-digit ‘G’ refined to 2- or 4-digit resolution) and imputed (2- or 4-digit) types are provided for those haplotypes available from Entebbe. A comparison is shown between concordances with no confidence threshold being applied against those when applying a posterior probability confidence threshold of 0.7 generated for the imputed calls by HLA*IMP:02 (resulting in fewer alleles being called by the imputation algorithm).

Gene	Number Validated	No threshold applied				Threshold (0.7) applied			
		2-digit		4-digit		2-digit		4-digit	
		Call rate	Concordance	Call rate	Concordance	Call rate	Concordance	Call rate	Concordance
A	660	1	0.90	1	0.84	0.97	0.91	0.95	0.86
B	660	1	0.75	1	0.69	0.9	0.77	0.87	0.72
C	660	1	0.83	1	0.80	0.97	0.83	0.96	0.81
DPA1	658	1	0.72	1	0.68	0.98	0.73	0.96	0.70
DPB1	658	1	0.43	1	0.42	0.69	0.53	0.65	0.54
DQA1	658	1	0.89	1	0.82	0.95	0.91	0.91	0.84
DQB1	658	1	0.88	1	0.76	0.98	0.89	0.90	0.78
DRB1	658	1	0.83	1	0.78	0.94	0.85	0.92	0.80
DRB3	658	1	0.90	1	0.90	0.98	0.91	0.98	0.90
DRB4	658	1	0.98	1	0.95	0.99	0.98	0.98	0.95
DRB5	658	1	0.95	1	0.95	1	0.95	1	0.95

Table 4.12: Concordance between classical types determined using MiSeq-based typing and imputed with HLA*IMP:02 for individuals from Soweto. The concordance between directly typed (6-digit ‘G’ refined to 2- or 4-digit resolution) and imputed (2- or 4-digit) types are provided for those haplotypes available from Entebbe. A comparison is shown between concordances with no confidence threshold being applied against those when applying a posterior probability confidence threshold of 0.7 generated for the imputed calls by HLA*IMP:02 (resulting in fewer alleles being called by the imputation algorithm).

Gene	Number Validated	No threshold applied				Threshold (0.7) applied			
		2-digit		4-digit		2-digit		4-digit	
		Call rate	Accuracy	Call rate	Accuracy	Call rate	Accuracy	Call rate	Accuracy
A	394	1	0.90	1	0.86	0.99	0.90	0.96	0.87
B	394	1	0.85	1	0.73	0.9	0.86	0.87	0.74
C	394	1	0.90	1	0.87	0.96	0.92	0.96	0.89
DPA1	394	1	0.73	1	0.69	0.98	0.74	0.96	0.71
DPB1	394	1	0.40	1	0.40	0.62	0.48	0.60	0.49
DQA1	394	1	0.91	1	0.82	0.96	0.93	0.94	0.84
DQB1	394	1	0.92	1	0.80	0.96	0.92	0.85	0.82
DRB1	394	1	0.84	1	0.77	0.9	0.86	0.84	0.82
DRB3	394	1	0.89	1	0.88	0.95	0.90	0.94	0.88
DRB4	394	1	0.97	1	0.93	0.99	0.97	0.97	0.93
DRB5	394	1	0.97	1	0.97	0.98	0.97	1	0.97

Table 4.13: Concordance between classical types determined using MiSeq-based typing and imputed with HLA*IMP:02 for individuals from Banfora. The concordance between directly typed (6-digit ‘G’ refined to 2- or 4-digit resolution) and imputed (2- or 4-digit) types are provided for those haplotypes available from Entebbe. A comparison is shown between concordances with no confidence threshold being applied against those when applying a posterior probability confidence threshold of 0.7 generated for the imputed calls by HLA*IMP:02 (resulting in fewer alleles being called by the imputation algorithm).

		No threshold applied				Threshold (0.7) applied			
		2-digit		4-digit		2-digit		4-digit	
Gene	Number Validated	Call rate	Accuracy	Call rate	Accuracy	Call rate	Accuracy	Call rate	Accuracy
A	334	1	0.99	1	0.97	1	0.99	0.99	0.97
B	334	1	0.67	1	0.60	0.79	0.76	0.77	0.70
C	334	1	0.87	1	0.84	0.98	0.88	0.98	0.86
DPA1	334	1	0.80	1	0.73	0.99	0.80	0.98	0.74
DPB1	334	1	0.44	1	0.44	0.78	0.48	0.75	0.49
DQA1	334	1	0.84	1	0.82	0.99	0.85	0.96	0.84
DQB1	334	1	0.88	1	0.77	0.96	0.88	0.88	0.79
DRB1	334	1	0.82	1	0.77	0.93	0.83	0.91	0.80
DRB3	334	1	0.90	1	0.90	0.98	0.91	0.98	0.91
DRB4	334	1	0.93	1	0.90	0.99	0.93	0.99	0.90
DRB5	334	1	0.96	1	0.96	0.99	0.96	0.99	0.96

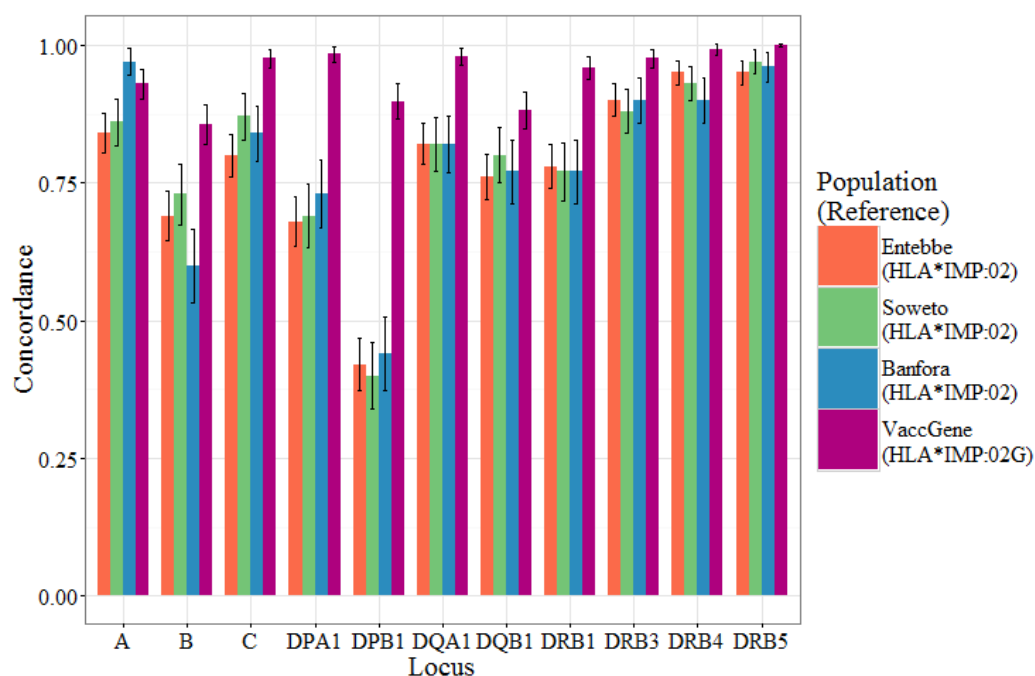


Figure 4.14: Concordance between HLA alleles imputed with HLA*IMP:02 or HLA*IMP:02G and typed using MiSeq.

Concordance values were determined using 4-digit level of resolution calls for the HLA*IMP:02 comparisons within populations (Entebbe, Soweto and Banfora) and at the 6-digit 'G' resolution using the HLA*IMP:02G algorithm (comprising five cross-validations integrating all populations). No posterior probability thresholds for imputed HLA allele calls were applied for any comparison.

4.3.4 Development of a framework for conditional association analyses in structured populations

In order to facilitate conditional fine-mapping of signals of association across the MHC, a pLMM approach was applied to the pooled datasets from all three populations. A strong correlation (Pearson's $r > 0.99$) was observed between significance p-values calculated across the MHC region determined using the pLMM and the FE meta-analyses (**Figure 4.15**). Variants where deviations in significance values were observed between association testing formats were often in the lower bounds of significance ($p > 1 \times 10^{-5}$) in either association testing approach used.

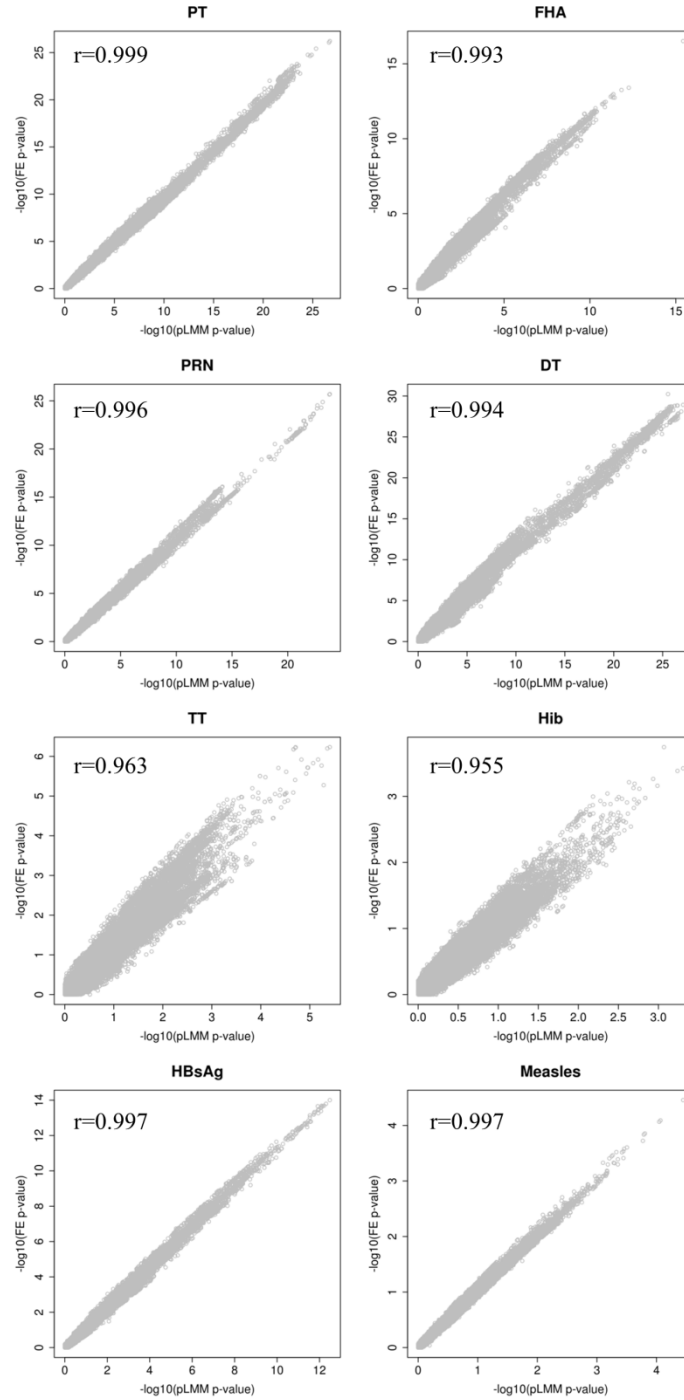


Figure 4.15: Correlation between significance values observed across the MHC when testing for association between variants and each trait comparing the pLMM and FE meta-analysis approaches. Scatter plots of p-values derived from both association analysis approaches are shown for each trait across the extended MHC (25.5–34Mb on chromosome 6, build 37). Pearson correlation coefficient (r) are provided for each trait.

4.3.5 Determining the genetic architectures of multiple vaccine responses

4.3.5.1 *pLMM analysis of genotypes across the extended MHC locus*

Conditional analyses employing the pLMM structure testing only the merged set of genotypes identified independent peaks of association ($p < 5 \times 10^{-9}$) across the class II MHC region for responses to FHA, PRN and HBsAg (**Figure 4.16**). After conditioning on the top, index SNP only one independent peak was observed for both FHA and PRN, whereas two peaks could be observed for HBsAg. These analyses demonstrated peaks of variant association occurring in and around all class II genes. The patterns of SNP association were noted to be differentiated between traits.

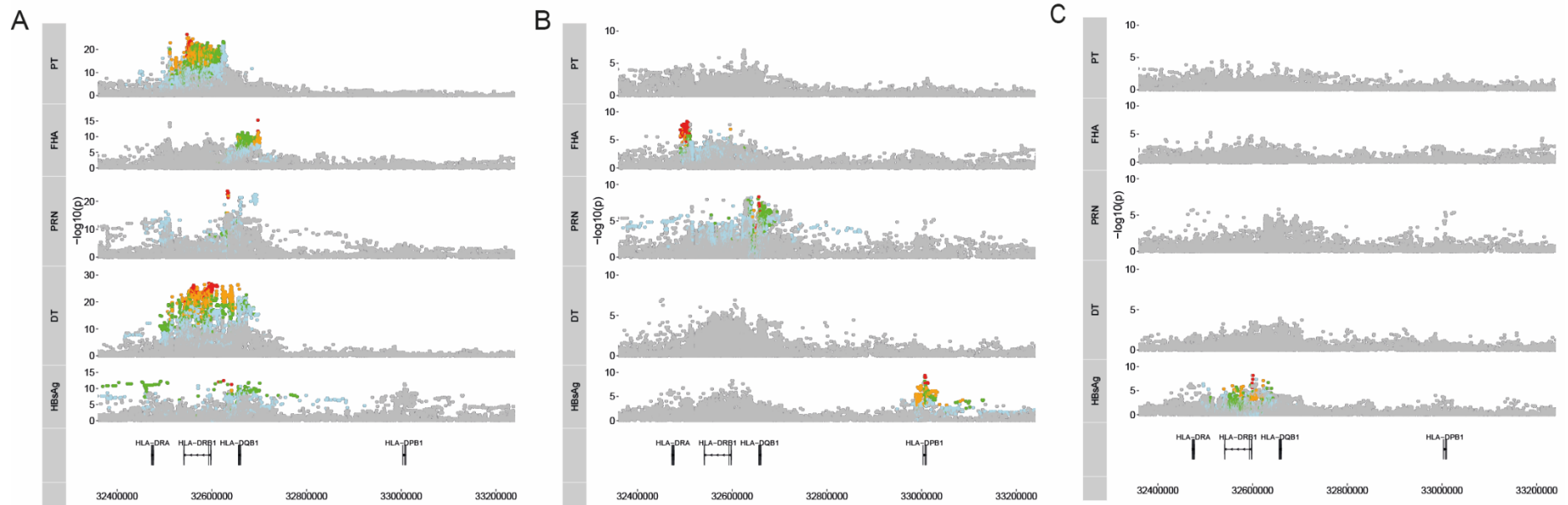


Figure 4.16: Regional plots of association between SNP variants and antibody response traits with and without inclusion of top associated variants as covariates.

Unconditional analyses (A); analyses conditioning on the top associated variant (B); and analyses including both variants associated from unconditional and first round of conditional analyses (C) are shown for each associated vaccine response trait. Wherever the index (most significant p-value) variant demonstrated significance p-values $< 5 \times 10^{-9}$, association peaks have SNPs coloured by LD (r^2) with top associated variant (red 0.8-1; orange 0.6-0.8; green 0.4-0.6; blue 0.2-0.4; grey < 0.2), otherwise all points are coloured in grey.

4.3.5.2 *Multivariate association analysis of all vaccines administered simultaneously*

Undertaking a multivariate association analysis of the correlated antibody responses (Pearson's $r > 0.3$) administered simultaneously (PT, FHA, PRN, DT, TT, Hib, HBsAg) revealed no regions of the genome demonstrating evidence of association outside of the MHC region (**Figure 4.17**).

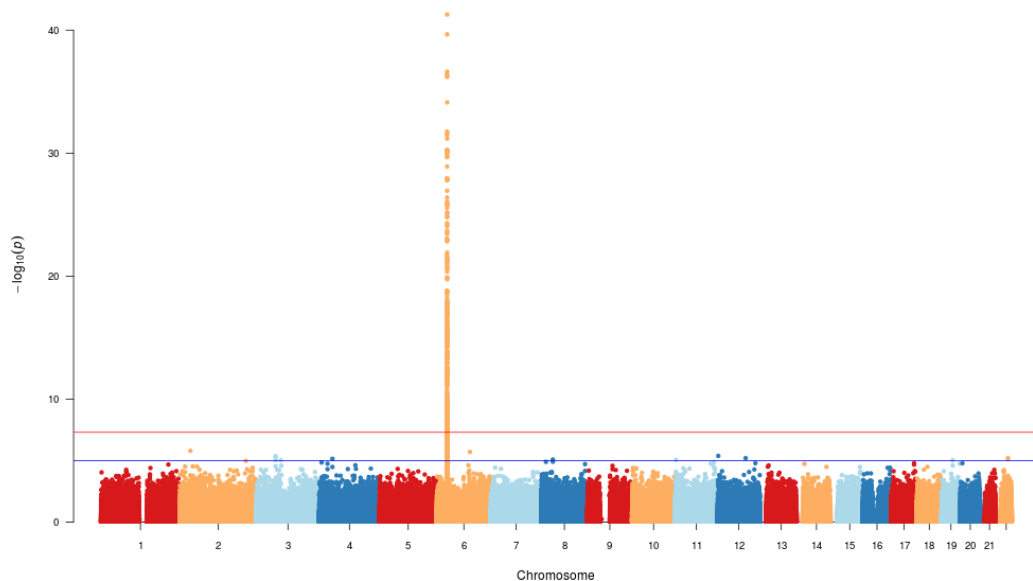


Figure 4.17: Manhattan plot of a multivariate association between genetic variants and seven correlated vaccine responses. A multivariate regression model was undertaken in the GEMMA software using only directly genotyped variants in all three populations including a pooled genetic relatedness matrix and including the following vaccine responses simultaneously: PT, FHA, PRN, DT, TT, Hib and HBsAg.

4.3.6 Building a model of genetic architecture of vaccine responses

Multiple models of association were tested with each of the MHC-associated traits using variants derived from direct genotyping and SNP imputation and HLA allele and amino acid imputation using a pLMM framework combined with an adapted manual modelling method (including an IBD-pruned cohort with PCAs included as covariates; **Table 4.14**). Variants spanning nearly every gene across the class II HLA region (except *HLA-DRB4*) were found to be associated with at least one of the traits (**Figure 4.18**). Combinations of either classical HLA alleles or amino acid residues were found to explain the greatest proportion of variance.

Some traits were best explained by alleles alone (FHA and DT for example) whereas some were best explained by amino acid variants alone (HBsAg) and others explained by a combination spread over multiple genes. With the exception of the response to PT, individual SNPs were not, in themselves, observed to explain the majority of the variance. The SNP rs73727916 was observed to explain a significant proportion of the variance of PT response ($R^2 = 0.076$). Conditioning on this SNP substantially reduced the significance of the top associated *HLA-DRB1* variant (HLA-DRB1-233Thr $p_{\text{unconditional}} = 4.3 \times 10^{-24}$; $p_{\text{conditional}} = 6.3 \times 10^{-4}$) but did not reduce the significance of the top HLA-DRB3 variant to the same extent (HLA-DRB3-74Gln $p_{\text{unconditional}} = 2.6 \times 10^{-24}$; $p_{\text{conditional}} = 4.4 \times 10^{-6}$). This particular SNP is located in an intron of *HLA-DRB1*. It has not been found associated as an index SNP in any genetic association studies of other diseases or phenotypes. This SNP is predicted to overlay a CCCTC-binding factor (CTCF) binding site that is active in immune cells including macrophages and

monocytes but the state or cell-specific effects of this variant are poorly described^{288,289}.

With the exception of HLA-DRB3 in PT response and HLA-DQB1 in PRN, variation was explained most parsimoniously when variation was described at the amino acid level as opposed to classical alleles at each locus as determined using automated step-wise modelling (**Table 4.15**).

Testing for either a dominant or additive model of inheritance for those variants found independently associated with each trait revealed an excess of associations that were more appropriately modelled as additive inheritance (**Table 4.16**). However a striking excess of associations in HLA-DRB1 were observed to fit more of a dominant model (four out of the total of six detected).

Table 4.14: Results from manual and automated step-wise modelling of class II HLA variants with five vaccine responses including principal components and time between sampling and vaccination as covariates.

	Method	Variant 1*	Variant 2*	Variant 3*	Variant 4*	p _{LMM}	p _{uni} [†]	p _{multi} [‡]	BIC [¶]
PT	Manual for HLA-DRB1	rs73727916	-	-	-	1.9x10 ⁻²⁷	8.1x10 ⁻³⁰	-	6442.55
		DRB1-233Thr	-	-	-	-	2.0x10 ⁻²⁷	-	6453.50
	Manual for HLA-DRB3	HLA-DRB3*02:02:01G [§]	HLA-DRB3*03:01:01G [†]	-	-	-	-	1.76x10 ⁻³⁶	6460.35
		DRB3-74Gln [§]	-	-	-	2.6x10 ⁻²⁴	1.6x10 ⁻²⁷	-	6453.00
	Final	rs73727916	DRB3-74Gln	-	-	-	-	3.6x10 ⁻³⁴	6424.94
FHA	Manual for HLA-DRB1	HLA-DRB1*08:04:01 [§]	-	-	-	4.8x10 ⁻¹⁵	3.9x10 ⁻¹⁷	-	6487.19
		DRB1-74Leu [§]	-	-	-	4.9x10 ⁻¹⁵	4.3x10 ⁻¹⁷	-	6487.39
	Automated	DRB5-72Arg [§]	-	-	-	-	1.7x10 ⁻¹⁰	-	6517.21
		HLA-DRB5*01:01:01 [§]	-	-	-	-	6.0x10 ⁻¹⁰	-	6519.67
	Final	HLA-DRB1*08:04:01	DRB5-72Arg	-	-	-	-	1.8x10 ⁻²²	6465.66

*: Variants associated with each trait through identified through the pLMM or univariate modelling approaches

†: The univariate model tested only on individuals restricted by IBD (< 0.2)

‡: Where multiple variants are tested in a conditional multivariate model the omnibus p-value is shown (p_{multi})

¶: The Bayesian Information Criterion was calculated for all models. The most parsimonious model will have a BIC closer to 0. The model with the lowest BIC for each phenotype at each locus is shown in red and the BIC of the final model is shown in bold.

§: For each locus and each trait any associated HLA amino acid was also tested in a univariate or multivariate model as described by classical HLA alleles that contain that residue.

Table 4.14 continued

PRN	Manual for HLA-DQB1	DQB1-74Ser [§]	-	-	-	1.3x10 ⁻²²	5.4x10 ⁻²⁵	-	4472.86
		DQB1*05:01:01G [§]	HLA-DQB1*04:02:01 [§]	-	-	-	-	3.7x10 ⁻²⁷	4465.17
	Manual for HLA-DRB1	HLA-DRB1*11:02:01	-	-	-	8.4x10 ⁻¹⁶	5.2x10 ⁻¹⁷	-	4509.18
		HLA-DRB1*03:02:01	-	-	-	-	3.6x10 ⁻¹⁶	-	4513.02
DT	Final	DQB1*05:01:01G	HLA-DQB1*04:02:01	HLA-DRB1*11:02:01	HLA-DRB1*03:02:01	-	-	8.1x10 ⁻³⁶	4431.07
	Manual for HLA-DRB1	DRB1-4Arg [§]	-	-	-	5.4x10 ⁻²⁸	6.2x10 ⁻³¹	-	6345.77
		HLA-DRB1*07:01:01G [§]	HLA-DRB1*09:01:02G [§]	-	-	-	-	2.5x10 ⁻²⁸	6360.18
	Final	DRB1-4Arg				5.4x10 ⁻²⁸	6.2x10 ⁻³¹		6345.77

[§]: The Bayesian Information Criterion was calculated for all models. The most parsimonious model will have a BIC closer to 0. The model with the lowest BIC for each phenotype at each locus is shown in red and the BIC of the final model is shown in bold.

[§]: For each locus and each trait any associated HLA amino acid was also tested in a univariate or multivariate model as described by classical HLA alleles that contain that residue.

Table 4.14 continued

HBsAg	Manual for HLA-DRB1	DRB1-74Arg [†]	-	-	-	2.45x10 ⁻¹⁴	2.45x10 ⁻¹⁸	-	5276.86
		HLA-DRB1*03:02:01	-	-	-	-	3.0x10 ⁻¹⁵	-	5290.65
		HLA-DRB1*03 [†]	-	-	-	-	2.8x10 ⁻¹⁸	-	5276.86
		HLA-DRB1*03:02:01 [§]	HLA-DRB1*03:01:01G [§]	-	-	-	-	4.5x10 ⁻¹⁸	5280.553
	Manual for HLA-DPB1	DPB1-85Gly [§]	-	-	-	9.5x10 ⁻⁹	9.5x10 ⁻¹³	-	5294.74
		HLA-DPB1*105:01	-	-	-	-	2.6x10 ⁻⁷	-	5326.42
		HLA-DPB1*02:01:02 [§]	HLA-DPB1*04:01:01G [§]	HLA-DPB1*18:01 [§]	-	-	-	5.6x10 ⁻¹¹	5321.68
	Automated	DRB1-67Phe [§]	-	-	-	-	4.2x10 ⁻¹¹	-	5309.38
		HLA-DRB1*08:04:01 [§]	HLA-DRB1*09:01:02G [§]	HLA-DRB1*11:01:02 [§]	-	-	-	2.2x10 ⁻⁹	5324.74
		DPB1-35Tyr	-	-	-	-	1.4x10 ⁻¹³	-	5298.23
	Final	DRB1-74Arg	DPB1-85Gly	DRB1-67Phe	DPB1-35Tyr	-	-	3.4x10 ⁻³¹	5223.14

[§]: The Bayesian Information Criterion was calculated for all models. The most parsimonious model will have a BIC closer to 0. The model with the lowest BIC for each phenotype at each locus is shown in red and the BIC of the final model is shown in bold.

[†]: For each locus and each trait any associated HLA amino acid was also tested in a univariate or multivariate model as described by classical HLA alleles that contain that residue.

Table 4.15: The variance of HLA associated vaccine responses described as amino acid or classical allele variation compared using the BIC. The fit of models determined using stepwise regression containing all variants described as either Models with BIC values closer to 0, considered a better fit, are highlighted in red.

	Locus	Variant Category	BIC
PT	DRB1	Alleles	6445.67
		Amino Acids	6422.78
	DRB3	Alleles	6437.08
		Amino Acids	6442.69
FHA	DRB1	Alleles	6467.76
		Amino Acids	6463.73
PRN	DQB1	Alleles	4439.73
		Amino Acids	4440.92
	DRB1	Alleles	4430.53
		Amino Acids	4424.41
DT	DRB1	Alleles	6358.29
		Amino Acids	6325.12
HBsAg	DRB1	Alleles	5245.41
		Amino Acids	5241.46
	DPB1	Alleles	5313.72
		Amino Acids	5298.74

Table 4.16: Testing for additive or dominant models of inheritance in variants associated in final models of vaccine response traits.

Trait	Variant	LogLik _{add} *	p _{add}	LogLik _{dom} *	p _{dom}	BIC _{add} †	BIC _{dom} †	Best Model
PT	rs73727916	-3186.4	8.1x10 ⁻³⁰	-3193.1	7.2x10 ⁻²⁷	6442.5	6456	Additive
	DRB3-74Gln	-3191.6	1.6x10 ⁻²⁷	-3227.4	8.2x10 ⁻¹²	6453	6524.5	Additive
FHA	HLA-DRB1*08:04:01	-3208.7	3.9x10 ⁻¹⁷	-3206.9	6.5x10 ⁻¹⁸	6487.2	6483.6	Dominant
	DRB5-72Arg	-3223.7	1.7x10 ⁻⁷	-3224.4	3.3x10 ⁻⁷	6517.2	6518.5	Additive
PRN	DQB1*05:01:01G	-2237.2	5.3x10 ⁻¹⁰	-2241.3	3.5x10 ⁻⁸	4540.9	4549	Additive
	HLA-DQB1*04:02:01	-2223.2	3.5x10 ⁻¹⁶	-2224.7	1.5x10 ⁻¹⁵	5412.9	4515.9	Additive
	HLA-DRB1*11:02:01	-2221.4	5.2x10 ⁻¹⁷	-2219.7	9.3x10 ⁻¹⁸	4509.2	4505.8	Dominant
	HLA-DRB1*03:02:01	-2223.3	3.6x10 ⁻¹⁶	-2224.9	1.9x10 ⁻¹⁵	4513	4516.3	Additive
DT	DRB1-4Arg	-3142.9	8.1x10 ⁻²⁹	-3138	5.9x10 ⁻³¹	6355.5	6345.7	Dominant
HBsAg	DRB1-74Arg	-2604.4	2.8.10 ⁻¹⁸	-2609.2	3.7x10 ⁻¹⁶	5276.9	5286.5	Additive
	DPB1-85Gly	-2617	9.5x10 ⁻¹³	-2628.5	1.3x10 ⁻⁷	5301.9	5325	Additive
	DRB1-67Phe	-2620.7	4.2x10 ⁻¹¹	-2616.8	7.7x10 ⁻¹³	5309.4	5301.5	Dominant
	DPB1-35Tyr	-2615.1	1.4x10 ⁻¹³	-2622.7	3.4x10 ⁻¹⁰	5298.2	5313.4	Additive

*: LogLik is the log likelihood derived from testing either an additive (add) or dominant (dom) model of inheritance where the better fitting model is closer to 0.

†: BIC Is the Bayesian information criterion from testing an additive (add) or dominant (dom) model of inheritance. Best fitting model BIC highlighted in bold.

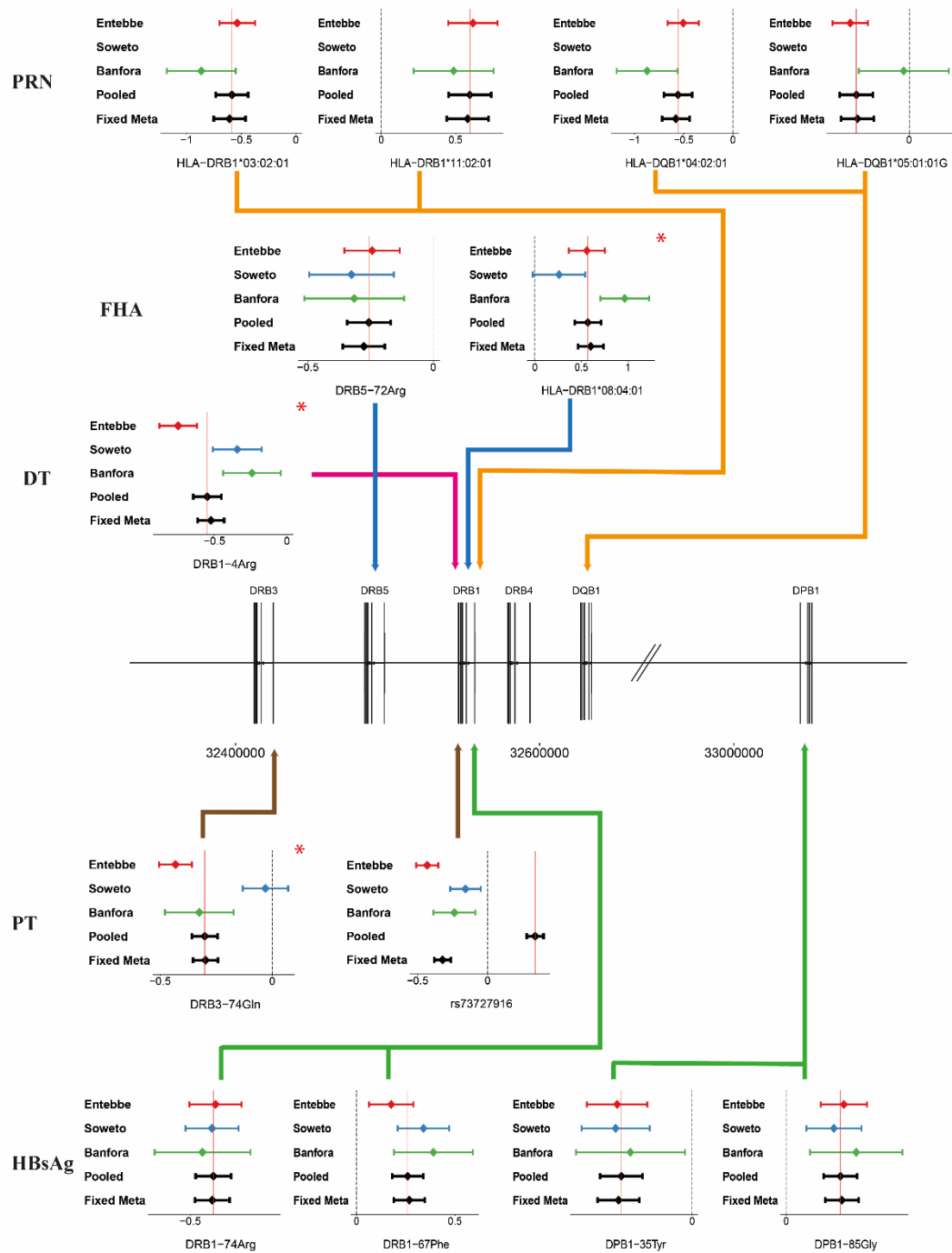


Figure 4.18: The genetic architecture of associations between class II HLA variants and responses to multiple vaccine antigens.

Forest plots of the beta effect estimates (points) with 95% confidence intervals coloured by population and corresponding distributions for the pLMM ('Pooled' – solid black horizontal line) and FE meta-analysis ('Fixed Meta'). Dashed vertical black lines represent no effect (beta = 0) and solid vertical red lines cross the beta estimate of the pLMM as a reference. Each independent

association identified through a combined manual and automatic modelling approach for each trait is shown and the originating locus of association is represented by solid arrowed lines coloured by trait indicating the relevant region on chromosome 6.

* Associations demonstrating significant evidence ($p_Q \leq 1 \times 10^{-3}$) of heterogeneity.

4.3.7 Understanding between-study heterogeneity

Significant evidence of heterogeneity ($p_Q \leq 1 \times 10^{-3}$) was observed at three associated loci (**Table 4.17**). The allele with the most evidence of heterogeneity (DRB3-74Gln in PT response) was noted to have a significant difference in allele frequencies between populations (Entebbe-Soweto $p_{\chi^2} = 1.66 \times 10^{-7}$; Entebbe-Banfora $p_{\chi^2} = 1.44 \times 10^{-9}$). These differences could have been expected based on the differences in genetic differentiation at this locus (**Figure 4.10**). However the consistent observed effect of this measured variant between Entebbe and Banfora populations (those with more significant differences in allele frequencies) and absent effect observed with the Soweto population makes it unlikely that the allelic differences are the cause of the heterogeneity.

There was very little evidence for a difference in allele frequencies of DRB1-4Arg for DT response (Entebbe-Banfora $p_{\chi^2} = 0.03$). Therefore a purely genetic effect of heterogeneity was again unlikely at this locus. For FHA, however, the difference in frequencies of the HLA-DRB1*08:04:01 allele was consistent with an observed difference in allele frequencies (Entebbe-Banfora $p_{\chi^2} = 2.00 \times 10^{-8}$).

A Bayesian multi-ethnic meta-analysis approach (MANTRA)²⁵¹ was used to further assess whether the heterogeneity observed at those loci was due to genetic heterogeneity. Using this approach it was reasoned that if the observed heterogeneity was a result of significant allelic differences in the one cohort that demonstrated reduced genetic association signals, the addition of this cohort to the meta-analysis would be expected to contribute very little to the range over which the credible variants were found and the range would be expected to remain the same or decrease by a small margin. If, however, the difference was not due to

genetic effects, and the LD structure around the variants (either those defined as causal or in LD with the casual variant) were preserved, the apparent lack of signal in the single cohort would lead to an increase in the interval over which the credible variants would be hypothesised to be present²⁵².

Although it was difficult to account for multiple independent association signals using this method, an increase in the range of the credible interval was observed for the DRB3-74Gln signal observed with response to PT (**Figure 4.19**) when the Soweto population was added to the analysis. This unusual effect provided increased evidence in favour of the observed heterogeneity being unlikely to be occurring due to differences in allelic effect and more likely to be due to other, possible environmental effects. As a comparison, the same analysis was undertaken for the HLA-DRB1 association with FHA where the converse was true demonstrating differences in line with the FHA heterogeneity being due to genetic differences and the PT signal being due to other effects (**Figure 4.20**).

Table 4.17: Class II HLA signals of association with three vaccine response traits demonstrating significant evidence of heterogeneity.

		Entebbe			Soweto			Banfora					
Variant		Beta	s.e.	MAF*	Beta	s.e.	MAF*	Beta	s.e.	MAF*	P _{FE} [†]	P _{RE} [‡]	p _Q [¶]
PT	DRB3-74Gln	-0.43	0.04	0.48	-0.03	0.05	0.40	-0.33	0.08	0.36	6.06x10 ⁻²⁷	0.05	2.4x10 ⁻⁹
FHA	HLA- DRB1*08:04:01	0.56	0.1	0.04	0.26	0.14	0.04	0.96	0.13	0.08	8.81x10 ⁻¹⁸	1.31x10 ⁻³	1.3x10 ⁻³
DT	DRB1-4Arg	-0.75	0.07	0.11	-0.34	0.09	0.11	-0.24	0.10	0.15	1.72x10 ⁻²⁹	7.00x10 ⁻³	6.0x10 ⁻⁶

*: minor allele frequency

†: p-value calculated from a fixed effects meta-analysis combining the effects across the three studied populations

‡: p-value calculated from a random effects meta-analysis combining the effects across the three studied populations

¶: p-value of evidence of heterogeneity between the three studied population calculated from the Cochran's Q statistic

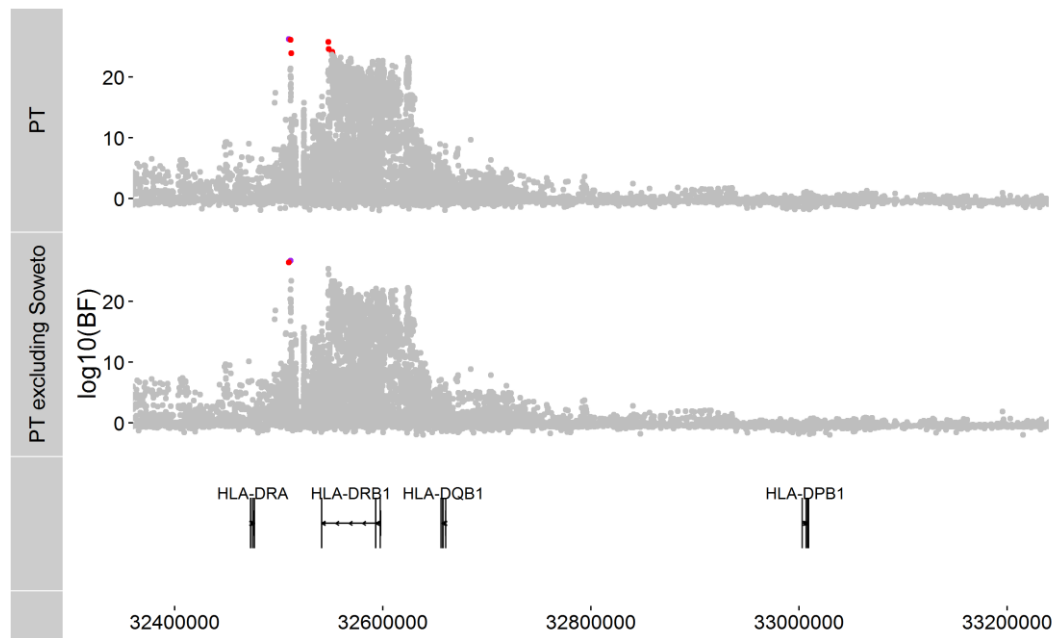


Figure 4.19: Credible set of variants across the class II HLA region associated with response to PT as determined using all three cohorts or excluding the Soweto population.

All variants including genotyped and imputed SNPs, HLA alleles and amino acid variants are included with the index variant (largest $\log_{10}(BF)$) filled in purple and all other variants within the 99% credible interval filled in red. All other variants are filled in grey.

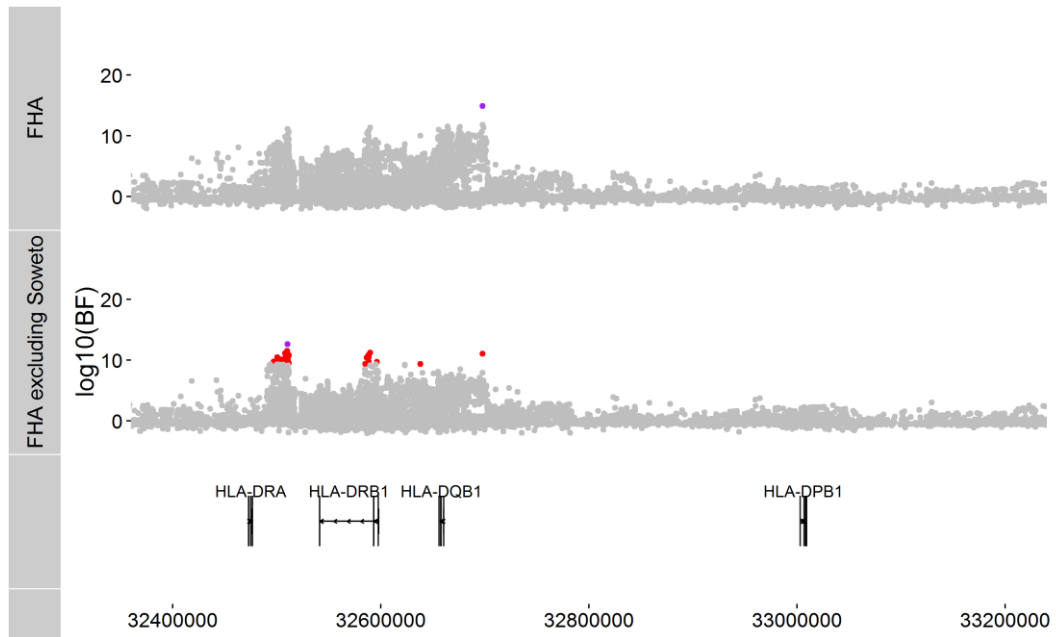


Figure 4.20: Credible set of variants across the class II HLA region associated with response to FHA as determined using all three cohorts or excluding the Soweto population.

All variants including genotyped and imputed SNPs, HLA alleles and amino acid variants are included with the index variant (largest $\log_{10}(BF)$) filled in purple and all other variants within the 99% credible interval filled in red. All other variants are filled in grey.

4.4 Conclusions

The collection of high resolution, low ambiguity HLA type calls from three distinct African populations covering 11 *HLA* genes described here offers one of the most comprehensive multi-site characterisations of African HLA types ever performed. The large number of individuals included in the analysis permitted the calculation of statistics with a high accuracy whilst using methods that corrected for unequal sample sizes. Furthermore, this work represents the first time that classical HLA imputation has been carried out to a six-digit level of resolution in

structured populations using a pLMM method of association testing. The consistency of significance values both between genotyped SNPs and imputed HLA variants and across populations increases the likelihood that the methods are robust and accounting for population structure and that the majority of associations are unlikely to represent false positive associations.

Much of the work undertaken to date investigating HLA diversity undertaken in populations of African descent have focussed on class I genes owing to their recognised association with a range of diseases including malaria²⁹⁰ and HIV²⁹¹. At the class I loci, the findings from this study are in keeping with previous research, demonstrating high diversity and differentiation at *HLA-B* in particular although we do not observe (to the level of resolution achieved here) any new alleles in any of the three included populations. Interestingly using both the G_{ST} and Jost's D statistic less differentiation was observed between the Soweto and Entebbe groups, opposed to the West African Banfora population when compared to either of the other two populations. This appears to suggest a more major East-West Africa divide and spectrum of differentiation rather than other geographical divides. However more populations across Africa would need to be studied to support this theory.

In comparison to class I genes, the class II genes were less differentiated across the three African populations. However, some loci, *HLA-DPB1* especially, were substantially more diverse than expected from previous studies. Furthermore, multiple new alleles were identified in these genes, likely pertaining to the fact that alleles from these genes are underrepresented in existing databases of HLA alleles^{268,292,293} although overall these new alleles were still observed at low

frequency ($MAF_{AV} < 0.01$). HLA-DP alleles have traditionally been rarely typed for tissue matching in cases preparing for tissue transplantation meaning that this locus is poorly represented in many population reference panels²⁹². In keeping with observations from substantially larger studies undertaken in the USA¹⁶³ and in China²⁹⁴, the alleles identified in this study represent only a fraction of the total number of alleles identified and recorded to date (13,190 alleles in the IMGT/HLA database 3.20). This finding supports the theory that many of the alleles present in the reference databases are specific to very small numbers of individuals from isolated populations.

The high observed levels of diversity and allelic richness specific to African populations are likely to be responsible for the suboptimal imputation of classical HLA alleles using existing methods and reference panels implemented in HLA*IMP:02 (concordance < 0.85 at 4-digit resolution). Surprisingly, with the exception of the *HLA-A* gene, imputation performance was not seen to be of greater concordance in the West African Banfora population who are likely to have less genetic distance between them and other West African or African-American individuals forming the African component of existing reference panels¹⁷². Performance at *HLA-B* and *HLA-DPB1* genes was particularly poor likely owing to the uncharacterised diversity at both loci and low number of representative alleles in the reference panel for *HLA-DPB1*. Accordingly, when the new African-specific reference panel was used instead for imputation, using a five-fold cross-validation method, concordance improved substantially. Only statistics for *HLA-B* remained below 0.9 (4-digit concordance 0.88), again likely owing to the large number of low frequency alleles observed differentially across populations.

The use of imputed variants (encoded as both alleles and amino acids) in downstream association models resulted in patterns of association consistent with other similar studies where HLA imputation has been used to fine-map association signals^{164,165,168-170,231}. In regions where significant variant associations (at the level of the genotyped or imputed SNP) were observed, corresponding associations at the classical allele or amino acid variant level were also observed. Frequently, these associations were noted to have greater significance values than the original genotype variants. Similarly, in the absence of associations at the genotyped or imputed SNP level, no associations were observed following classical allele or amino acid imputation.

In agreement with previous fine-mapping exercises, the index amino acid variants frequently lie in, or very near, the peptide binding groove. It is, however, important to note that due to the nature of the typing based on sequencing of exons covering the peptide binding region, these variants will implicitly be called with greater accuracy. In further agreement with previous studies, the architecture of HLA association appears to be diverse and complex encompassing multiple genes and alleles. Even when describing variation at the SNP level alone, there is evidence from the conditional analysis that other peaks of association across the HLA may exist at levels of significance below the stringent threshold set in this study (**Figure 4.16**). These peaks may become more apparent when sample size is increased further.

There is currently substantial debate relating to the relevance of observing stronger evidence of association at HLA amino acid residue variant sites as opposed to classical HLA alleles alone. Studies of HIV-1 disease progression and

rheumatoid arthritis susceptibility have claimed that single amino acid residues (described as bi-allelic variants as in this study), or even multi-allelic residue position sites explain the risk of disease most parsimoniously in frequentist models (compared to a null model using the LRT)^{164,170}. However, a recent study of inflammatory bowel disease susceptibility noted that the describing multi-allelic variants in a ‘residue position’ model did not account appropriately for the complex LD in the region that may result in an over-inflation of test statistics¹⁶⁵. This observation has led to the more recent use of the BIC for model testing that can cope more robustly with spurious relationships and the method has been applied to a large meta-analysis of multiple sclerosis to characterise the diverse genetic architecture of HLA associations with the disease. A combination of the above methods was used in this study to account for all of these recent developments although these methods were only applied to bi-allelic variants and multi-allelic position sites were not included in analyses. Even when applying the more robust BIC-based testing method it was found that the associations were better explained with variants coded as amino acid variants compared to alleles. However, by undertaking fine-scale manual and automated modelling the exact variant associated does, with the data available here, seem to suggest that the variance of some phenotypes are explained most parsimoniously by variants described either as alleles or amino acids or even as SNPs.

The identification of the SNP rs73727916 being independently strongly associated with PT response, and the known CTCF-binding domain location of this SNP, would make this a compelling candidate for modulating effects through altered gene expression, as has been documented for *HLA-C* and control of HIV-1¹⁷⁶. Measuring the expression of *HLA* genes is challenging, particularly for class II

genes owing to the levels of polymorphism and the difficulty in designing primers to reliably measure the expression of all alleles²⁹⁵. Therefore there is currently a lack of suitable data available that can be used to test whether the identified SNP variant may influence expression of neighbouring genes such as *HLA-DRB1* that would be a reasonable mechanism to influence vaccine responses against PT.

In the case of FHA, the most significantly associated variant was a classical allele (HLA-DRB1*08:04:01) as opposed to the amino acid (DRB1-74Leu). Interestingly, the HLA-DRB1*08:04:01 allele is relatively specific to Africa being observed in higher frequencies in populations of predominantly West African ancestry²⁶⁸. Similarly, the HLA-DRB1 variant most significantly associated with HBsAg response almost exclusively tags HLA-DRB1*03:02:01 that is also a relatively African-specific allele. Interestingly, there was no evidence for an association with HLA-DRB1*07 alleles and HBsAg response that have classically been associated with non-response, but consistent with previous reports from Asia there was significant evidence for an effect of HLA-DP variants. These differences are worthy of further exploration especially given the poor understanding or appreciation of *HLA-DP* and the newly characterised diversity of these loci in this work, although the differences in genetic signals occurring across populations will require careful consideration. It is important to note that for many of the traits the differences between models testing different variants was, as would be expected, marginal and it is impossible to reliably favour one model over the other. Indeed, the observed differences between many of the models for many of the tested traits were small. It is therefore impossible to reliably infer causality with any of the variants identified to contribute to the genetic architecture of the vaccine response traits owing to the extensive LD existing

across the HLA region. Furthermore, the fact that other genes (non-HLA) in the MHC were not typed means that it is not possible to rule out the possibility of other non-HLA variants may explain the variation more robustly. Comprehensive typing of all MHC genes at scale is not practically possible through current applications although developments in both short²⁹⁴ and long-read²⁹⁶ sequencing technology may make these tractable options in the future.

The noted predisposition of HLA-DRB1 associations towards a dominant model of inheritance is interesting and may be highly relevant to vaccine design. Dominant models of inheritance have long been observed for autoimmune conditions and interestingly these have been most prominent for HLA-DRB1 associations^{297,298}. However the mechanisms of these associations have not been confirmed. Indeed, the relative contributions of specific HLA-DR, or HLA-DQ or indeed HLA-DP alleles to peptide presentation are poorly understood²⁹⁹. The data presented here may suggest that in the presence of an HLA-DR effect, whatever the underlying mechanism of that association, this is sufficient to dominate the eventual downstream immunological response. For other loci, it appears that a more additive model may be responsible although other models such as recessive or interactive were not formally tested here. These interesting insights warrant further investigation using larger cohorts of vaccine-related traits to understand the interaction of HLA molecules in terms of peptide presentation.

The absence of associations of between HLA variants and other measured vaccine responses is surprising given the strength of associations observed with the five main traits studied here. The absence of signals for TT is particularly notable given that it is a toxoid vaccine equivalent to four other antigens demonstrating

associations (PT, FHA, PRN and DT). It is possible that this lack of signal is due to imperfect tagging of truly associated HLA alleles or variants in the African populations. This possibility is supported by the observation of a small but notable peak in the MHC from the genotype data in the Banfora population (**Figure 3.12**) although the lack of replication and the deviation of this signal from the main *HLA* genes makes such a feature unlikely. Furthermore, the peptide sequence that gives rise to the tetanus toxin has been described as containing a ‘universal’ epitope³⁰⁰ and is thus often used as a control in experiments testing the HLA restriction of other peptides. Therefore it can be reasoned that if this is the case then there will be little variation between how individuals present and respond to tetanus epitopes. Furthermore, if these HLA signals are driven by pressures of positive or negative selection then it is possible that selection against *C. tetani* occurs independently of toxin exposure since levels of the toxin acquired through natural infection are rarely sufficient to raise protective levels of antibody³⁰¹.

The absence of association between HLA and measles responses was again observed following imputation. Comparing the association results discovered in the analysis presented in this thesis compared to those from the most recently published GWAS²¹⁰ found that two of the alleles reported to be associated at suggestive levels of statistical significance (HLA-B*57:01 and HLA-DRB1*15:05) were not observed in the African populations and the remaining allele (HLA-DQB1*06:02) demonstrated no significant association ($p_{LMM} = 0.17$). As discussed in chapter three, these findings could be in keeping with either genetic heterogeneity between populations of European and African descent, the difference in phenotype measures (total IgG in the present study compared to

neutralising IgG in the USA study) or alternatively the earlier reports may represent false positive association findings.

Taken together, the data presented here are strongly in support of the hypothesis that the SNP associations observed with vaccine related traits in chapter three are due to variants in *HLA* genes. In line with other reports of associations between HLA variants and biological traits, multiple independent associations were observed and the variants best capturing the variation may be one of either amino acid, classical allele or SNP descriptors. Understanding the mechanisms underlying these associations using specific functional assays is the focus of the following chapter.

Chapter 5

5 Characterising the Molecular and Cellular Effects of HLA Diversity on Vaccine Efficacy

5.1 Background

The identification of multiple strong, independent signals of genetic association between variants across the class II MHC region and responses to multiple vaccine antigens is consistent with other reports of association between HLA and disease. The mechanisms underlying many of these historical reports of HLA-disease association remain poorly understood although they are generally believed to involve altered HLA-peptide binding and presentation³⁰². Confirming such a model in conditions such as autoimmune disease (which constitute the majority of such association signals) has been challenging due to the fact that the peptide epitope driving the signal is largely unknown. Using vaccine responses as a

platform to understand these relationships in greater detail offers a novel approach to this problem given that the sequences of the delivered peptides are explicitly known. In this chapter, the HLA association signals observed for two vaccine response traits (PT and DT using others as controls where applicable) were tested to determine whether they involved immunological pathways long hypothesised to be involved in antigenic response and memory through animal model studies. These pathways included differential HLA-peptide binding, TCR activation and T-cell clonal activation, and T_{FH}-B cell axis mediated antibody production and maturation.

As a first step towards achieving these goals, the associations observed across the class II MHC region for responses against multiple pertussis antigens was compared to class II MHC association signals recently reported from a case-control analysis of individuals self-reporting whooping cough²¹². A novel method of testing for correlation between traits at the genetic level of HLA was developed and the results from these analyses demonstrated important links between pertussis vaccine failure and antibody responses against a single vaccine antigen.

Subsequently, after discovering that an antibody response against a single vaccine antigen was related to vaccine protection, other pertussis-related immunological measures were explored to determine whether the same genetic stratification could be observed for such traits. First, the downstream targets of class II HLA peptide presentation, the T_{FH} cells, were assessed to determine whether there were measurable differences in levels of circulating antigen specific cells between individuals stratified by HLA using a newly reported flow cytometry-based technique¹⁰⁹. Secondly, differences in peptide binding profiles were interrogated

to determine whether particular PT peptides were bound more tightly by specific HLA alleles – the mechanism largely hypothesised to drive many HLA-disease associations³⁰³⁻³⁰⁵. Finally, the data from a recently published TCR gene expression dataset was used to test for evidence of genetic correlation between the expression levels of specific TCR genes and vaccine responses. Using the same genetic correlation approach applied to the WC dataset, the results of this TCR analysis provided further evidence in support of common immunological pathways linking vaccine antibody responses to efficacy, each impacted by variation in the human genetic code.

5.2 Methods

5.2.1 Whooping-cough GWAS

The association statistics (odds ratios, beta estimates and standard errors) from a meta-analysis of two GWAS of self-reported whooping cough are publically available²¹². The cohorts in this published analysis included the 1958 Birth Cohort (1958BC) and the Avon Longitudinal Study of Children and Parents (ALSPAC) both recruited in the United Kingdom. The 1958BC was originally designed to detect and measure circumstances and outcomes around birth, recruiting 17,000 participants born in a single week in 1958. This cohort became a longitudinal study with recruited individuals resurveyed at various ages. DNA samples were collected from participants in ongoing follow-up at age 45. Although vaccination histories were not directly available for these participants, vaccine coverage with pertussis was high in 1958; the year that national vaccination began (>85%)³⁰⁶, and therefore most birth cohort participants would be expected to have received

the wP pertussis vaccine. A history of whooping cough in the genotyped participants was determined through the use of a routine questionnaire sent to all parents when the participants were between 7 and 8 years of age. Part of the questionnaire asked the parents which infectious diseases the cohort participant had suffered during their early life.

ALSPAC was a longitudinal birth cohort with mothers identified through expected birth dates with 14,541 pregnant women and children recruited in 1992. Mothers completed a self-report questionnaire at approximately the twelfth week of pregnancy that included the question: “Have you ever had any of the following infections” and the mothers were presented with a list of common infections including whooping cough. They could respond either as a ‘Yes’, “No never” or “I don’t know” and for the purposes of the GWAS “Yes” respondees were cases compared to “No never” controls. Again, no formal record of immunisation was available for ALSPAC but the mothers were aged between 16 and 45 at the time of recruitment therefore they would have received their childhood vaccinations between the years of 1947 and 1976. Since aP was not delivered routinely until 2003 any vaccinated mothers would all have received wP vaccine. There remains a possibility, however, that some mothers may not have received vaccines given a fall in coverage between 1972 and 1977 owing to safety fears³⁰⁶.

The genetic analyses undertaken in the published WC GWAS included comparing cases (self-reported whooping cough) to controls using genetic variant data generated from direct genotyping and imputation of both genotypes and classical HLA alleles and amino acids (imputed using SNP2HLA²⁸⁵ and the Type 1 Diabetes Genomics Consortium reference panel). A fixed effects meta-analysis of

all genotyped and imputed SNPs and HLA alleles and amino acids between the ALSPAC and 1958BC individuals was then performed demonstrating GWAS significant associations across the HLA region. Summary statistics for the SNP variants were only available from the authors for those SNPs with p-values $< 1 \times 10^{-5}$.

5.2.2 Determining correlation of variant effect estimates between studies

In order to test for genetic correlation between WC susceptibility and antibody responses against multiple vaccine antigens targeting both pertussis and other agents, the beta estimates for bi-allelic SNP variants, and imputed HLA alleles and amino acids were compared separately between African vaccine response and UK whooping cough susceptibility GWAS studies. Estimates were compared either independently on a single variant basis, or compared across multiple variants simultaneously using the Pearson's correlation coefficient calculation. In cases where multiple variants were tested simultaneously, a measure of statistical significance of the final measure of Pearson's r was determined using permutation. The null distribution was calculated by randomly assigning different SNP identities to the calculated beta coefficients from the WC GWAS and recalculating Pearson's r up to 100,000,000 times dependent on the significance values calculated at thresholds starting at 100,000 permutations. The p_{perm} value was calculated as the frequency at which a Pearson's r value calculated from permutation was observed to surpass the r from the true data. These calculations were undertaken using both complete variant datasets and datasets pruned by LD (keeping only SNPs with $r^2 < 0.35$).

5.2.3 Antigen Inducible Marker assay

A novel Antigen Inducible Marker (AIM) method was used to measure and compare proportions of circulating antigen-specific T_{FH} cells in the circulating blood of donors defined by HLA-DRB1 allele carriage. The AIM assay uses flow-cytometry to detect proportions of antigen-specific T_{FH} cells defined as co-expressing CD25, OX40 and CXCR5 markers following *ex-vivo* antigen stimulation of PBMCs¹⁰⁹.

1 x 10⁶ PBMCs from stored samples collected from consenting participants recruited into studies coordinated by the laboratory of Professor Alessandro Sette were selected based on HLA-DRB1 allele type. The donor samples used were derived from studies investigating immunodominant peptides associated with responses against pertussis⁴⁹, tuberculosis³⁰⁷, dengue³⁰⁸, and IgE allergy³⁰⁹. The samples were thawed and cultured with 30µg/ml PT (Reagent proteins, USA), 5µg/ml DT (Reagent proteins, USA), 5µg/ml TT List Biological Laboratories Inc., Campbell, CA), 10 µg/ml phytohaemagglutinin (PHA, Sigma, St Louis, MO, USA), or toxoid diluent (water) at 37°C for 24 hours. The cells were then washed, labelled with an antibody panel (**Table 4.1**) for 15 minutes at 4°C before being fixed with paraformaldehyde (Sigma, St Louis, MO, USA) and acquired on an LSRII (Becton, Dickinson and Company, New Jersey, USA). Data derived from the gating strategy (**Figure 5.1**) was analysed using FlowJo Software version 10 (FlowJo LLC, Oregon, USA) and either one- or two-tailed Mann-Whitney U tests or linear regression statistical tests were performed in R. All participating donors were known either to have received DTaP as part of a vaccine study undertaken in the Sette lab, or self-reported having received EPI vaccines during childhood.

Table 5.1: Flow antibody panel used for the antigen inducible marker assay.

Antibody	Clone	Conjugation	Supplier
Anti-CD25	M-A251	Fluorescein (FITC)	Biolegend, San Diego, USA
Anti-CD45RA	HI100	eFluor450	eBioscience, San Diego, USA
Anti-CD14	M5E2	V500	BD Biosciences, San Jose, USA
Anti-CD19	H1B19	V500	BD Biosciences, San Jose, USA
Anti-CCR7	G043H7	Peridinin-Chlorophyll-protein (PerCp)-Cy5.5	Biolegend, San Diego, USA
Anti-CD8	RPA-T8	V500	BD Biosciences, San Jose, USA
Anti-OX40	Ber-ACT35	Phycoerythrin (PE)-Cy7	Biolegend, San Diego, USA
Anti-CD3	UCHT1	Alexa Fluor (AF) 700	BD Biosciences, San Jose, USA
Anti-CD4	RPA-T4	Allophycocyanin (APC)- eFluor780	eBioscience, San Diego, USA
Anti-CXCR5	J252D4	Brilliant Violet (BV) 605	Biolegend, San Diego, USA
LIVE/DEAD Aqua stain	-	-	Thermo Fisher Scientific Waltham, USA
Phosphate buffered saline (PBS)	-	-	Gibco, Thermo Fisher Scientific Waltham, USA

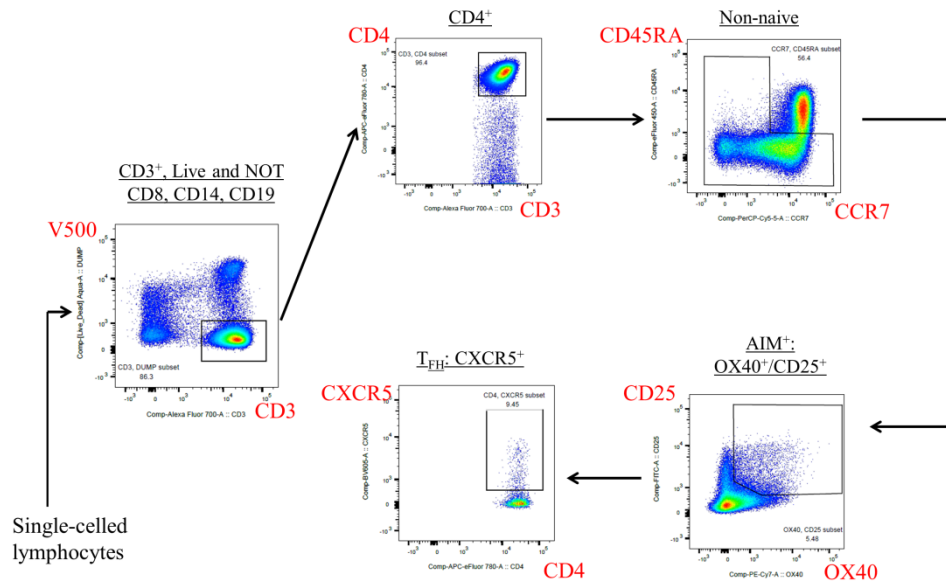


Figure 5.1: Gating strategy to identify AIM non-naïve CD4⁺ T-cells and the T_{FH} (CXCR5⁺) subset.

It was hypothesised that individuals who carried HLA-DRB1 alleles associated with either increased PT or DT antibody responses were more likely to possess increased proportions of circulating antigen-specific T_{FH} cells compared to individuals carrying HLA alleles associated with lower antibody responses. Since no association between HLA and levels of TT antibody were observed in the African infant GWAS and since TT is almost always co-administered with PT (as part of TdaP or DTaP), TT response was used to control for differential time between vaccination and sampling in the studied donors by analysing results as a ratio of PT vs TT-specific T_{FH} cell proportions. Based on the genetic association findings a ratio difference of 1.4 was expected between groups. Assuming an equal standard deviation between groups (0.5) it was calculated that there would

be 80% power to see a difference with an α of 0.05 including 13 donors in each group. Between 15 and 17 individuals homozygous for HLA-DRB1 alleles were selected depending on availability within the Sette donor cell repository. All HLA types of samples were concealed during the experimental phase and samples were randomised across batches. Unblinding only occurred after the final data had been generated and curated prior to final analysis.

5.2.4 Peptide binding assays

A variety of approaches were used to test whether HLA alleles associated with differential antibody responses against PT or DT were likely to bind antigen-specific peptides with a greater breadth or affinity. The Immune Epitope Database (IEDB)³¹⁰ collates results of published HLA allele-peptide binding assays and provides online *in silico* HLA-peptide binding prediction methods including NetMHCIIpan³¹¹ and others³¹². The output from the binding prediction algorithm included a binding affinity prediction (IC₅₀ - measured in nM) and a percentile rank for each possible 15-mer from any peptide sequence binding to the allele of interest. The percentile rank was generated by comparing the predicted IC₅₀ for each HLA allele against scores of 5,000,000 random 15-mers selected from the SWISSPROT database³¹².

To compare the breadth and affinity of binding of peptides to different HLA alleles the percentile rank scores of 15-mer peptides derived from PT (GenBank accession ALH76457), DT (BAL14546) and TT (WP_011100836) were compared. The highest affinity peptide per protein and allele was defined as the peptide with the smallest percentile score. The breadth of antigen-specific peptide binding by class II HLA alleles was defined by measuring the proportion of 15-

mer peptides predicted to bind within the top 5th percentile of all peptides from each peptide for each allele of interest, compared across antigens.

A modified analysis was performed in order to increase power to see differences between groups of HLA alleles including those not observed within either African or WC GWAS studies. All HLA-DRB1 alleles present in the IMGT database were divided into groups dependent on their sequences and whether they possessed an excess of amino acid residues associated with either increased (defined as ‘Good’ alleles) or decreased (defined as ‘Bad’ alleles) antibody responses. The definition of these alleles for both PT and DT vaccine responses was undertaken as follows. Firstly, the number of residue positions found to be significantly ($p < 0.05$) associated with either PT ($n = 39$) or DT ($n = 23$) responses were determined. Alleles were then defined as to whether they had an excess ($>1.5x$) of residues associated with either a positive beta (defined ‘Good’) or those with an excess ($>1.5x$) of negative beta effect estimates (defined ‘Bad’). The distributions of affinities of the top-predicted binding peptides for each of the alleles classified as such were then compared and tested for differences using a two-tailed Mann-Whitney U test.

To validate the findings from these *in silico* approaches, competitive binding assays were performed using an iodine¹²⁵-labeled standard probe peptide (**Table 5.2**) using a monoclonal antibody assay capture system³¹³. 0.1-1 nM of radiolabelled peptide was co-incubated at room temperature with 1 μ M to 1 nM of purified MHC-II in the presence of a cocktail of protease inhibitors and 1.6% (v/v) NP-40 detergent and PBS (Scripps Laboratories). Following a two-day incubation, MHC-II bound radioactivity was determined by capturing MHC-

II/peptide complexes on LB3.1 (anti-HLA class II monoclonal antibody, American Type Culture Collection, Manassas, VA, USA)-coated Lumitrac 600 plates (Greiner Bio-one, Frickenhausen, Germany), and bound radioactivity was measured using the TopCount (Packard Instrument Co., Meriden, CT) microscintillation counter. The concentration of peptide yielding 50% inhibition of the binding of the radiolabelled peptide was calculated. Under the conditions utilised, where $[\text{label}] < [\text{MHC}]$ and $\text{IC}_{50} \geq [\text{MHC}]$, the measured IC_{50} values were reasonable approximations of the true dissociation constant (K_D) values. Each competitor peptide was tested at six different concentrations covering a 100,000-fold dose range, and in three or more independent experiments. As a positive control, the unlabelled version of the radiolabelled probe was also tested in each experiment.³¹⁴

Table 5.2: Lymphoblastoid cell lines homozygous for the alleles of interest and radiolabelled ligands used as binding controls in competitive HLA-peptide binding assays. Adapted from Sidney et al.³¹³

Allele	Cell Line	Sequence	Source
DRB1*0701	PITOUT	YATFFIKANSKFIGITE	<i>C. tetani</i> (tetanus toxin 828)
DRB1*0901	HID	TLSVTFIGAAPLILSY	<i>H. sapiens</i> (PSA 9)
DRB1*1501	L466.1	PVVHFFKNIVTPRTPPY	<i>H. sapiens</i> (MBP 78)

5.2.5 T-cell receptor expression dataset

Recently, through the use of expression quantitative trait loci (eQTL) mapping, Jonathan Prichard and colleagues have shown that SNP variants encoding

multiple class II HLA amino acid variants are in *trans*-association with levels of TCR variable (V) gene expression³¹⁵. Using genotype and next-generation ribonucleic acid sequencing (RNA-Seq) data available from a cohort of 922 adults recruited as part of the Depression Genes and Networks Project³¹⁶, amino acid variants in HLA were associated with differential expression of multiple alpha (TRAV) and beta (TRBV) TCR V genes. These residues are proposed to be in direct or close contact with the TCR or peptide binding contact sites. Noting that the participants in the RNA-Seq dataset were sampled from the broad population with an absence of common immune exposure, the authors suggested that many of the TCR V genes that they measured the expression of were regulated through thymic selection during development in an HLA-dependent manner. Summary statistics for the RNA-seq GWAS were provided by the corresponding authors.

Using principles detailed earlier for assessing evidence of genetic correlation between effect estimates measured through association testing in independently performed GWAS, the correlation between effect estimates measured at the HLA amino acid residues was compared between the African vaccine antibody response and the TCR RNA-Seq studies. The same methods of measuring and testing the significance of correlation were employed as for the WC GWAS comparison including the measurement of the Pearson correlation coefficient and the use of permutation to calculate a significance value.

5.3 Results

5.3.1 Correlation between PT response and susceptibility to WC

The top associated biallelic variant from the WC GWAS was a SNP (rs9271768). This SNP did not demonstrate GWAS significant association with any measured vaccine response trait in the African infant pLMM analysis (**Table 5.3**). A significant association was observed, however, between the top associated variant from the PT response GWAS (rs72851029) and the WC susceptibility GWAS observed to have an opposite direction of effect between studies. No equivalent overlap was observed for the top variants from the other two pertussis vaccine responses FHA and PRN. On closer inspection, there was noted to be an overlap in significance levels between the index associated variants from each of the two studies included in the WC GWAS (1958BC, ALSPAC) and the PT antibody pLMM analysis performed in African infants presented in this thesis (**Table 5.4**). The index variants in the WC GWAS were noted to vary between the independent study populations. Similarly, the index HLA-DRB1 amino acid residue identified through the PT pLMM analysis was also found to be significantly associated in the WC GWAS (**Table 5.4**). In all cases except rs9271768 the association of the variants were in opposite directions between WC and PT response GWAS. There was also an overlap in the strengths of associations of the imputed HLA-DRB1 (**Table 5.5**) and HLA-DQ alleles (**Table 5.6**). In the case of HLA-DRB1, different 4-digit alleles of the same serological class demonstrated equivalent strengths of association in accordance with those alleles more common in individuals of European or African ancestry. The directions of effects were again consistently in

opposite directions between WC and PT response GWAS. Imputed alleles for HLA-DRB3, HLA-DRB4 and HLA-DRB5 were not available for the WC GWAS.

In order to test for evidence of correlation between the genetic associations observed with the pertussis antibody responses and WC susceptibility the beta effect estimates were compared at the SNP level between the WC and PT-response GWAS (**Figure 5.2**). Effect estimates were only available from the WC GWAS for those SNPs with a meta-analysis $p < 1 \times 10^{-5}$. For those SNPs with available association statistics a strong negative correlation was observed between the PT-response and WC GWAS (Pearson's $r = -0.86$) that was not observed with FHA response (0.19) or PRN (0.38).

Since a comparison between the studies was not possible at the 4-digit level of resolution of HLA alleles (due to population differences in HLA-DRB1 allele frequencies as demonstrated in **Tables 5.5** and **5.6**), the effect estimates of the bi-allelic amino acid residues between studies were instead compared (**Figure 5.3**). A strong negative correlation between the WC and PT response GWAS results was again observed (Pearson $r = -0.83$; **Figure 5.3A**) that persisted following pruning of the residues by LD (pruned by $r^2 < 0.35$; Pearson $r = -0.69$; **Figure 5.2D**). After pruning, no correlation was observed between the GWAS estimates for WC and responses to either FHA (-0.02; **Figure 5.2E**) or PRN (-0.01; **Figure 5.2F**).

Table 5.3: Summary statistics for variants most significantly associated with WC susceptibility or PT vaccine antibody. Variants were identified from the WC GWAS meta-analysis or pooled LMM pertussis vaccine antibody GWAS studies in African infants (PT, FHA and PRN).

	Variant	BP (b37)	WC Meta-analysis				PT pLMM				FHA pLMM			PRN pLMM		
			MAF _{AV} *	Beta	s.e. †	p‡	MAF _{AV} *	Beta	s.e. †	p‡	Beta	s.e. †	p‡	Beta	s.e. †	p‡
WC Meta	rs9271768	32,594,188	0.46	-0.39	0.04	1.21⁻¹⁸	0.22	-0.05	0.04	0.23	-0.04	0.04	0.27	0.13	0.04	5.1x10 ⁻³
PT Response	rs72851029	32,501,639	0.56	-0.33	0.05	2.9x10⁻¹¹	0.34	0.34	0.03	4.6x10⁻²⁶	-0.01	0.03	0.77	0.02	0.04	0.58
FHA Response	HLA- DRB1*08	-	0.02	-0.09	0.56	0.41	0.05	0.06	0.07	0.36	0.55	0.07	4.9x10⁻¹⁵	-0.05	0.08	0.57
PRN Response	rs114671076	32,669,817	0.87	0.29	0.06	4.9x10⁻⁶	0.75	-0.10	0.04	4.8x10 ⁻³	-0.01	0.03	0.96	0.41	0.04	2.4x10⁻²³

*: MAF_{AV}: average minor allele frequency calculated across populations – either 1958BC or ALSPAC in the WC GWAS, or Entebbe, Soweto and Banfora in the African GWAS

†: s.e.: standard error.

‡: p-values with significant ($p < 1 \times 10^{-5}$) highlighted in bold.

Table 5.4: Summary statistics for index variants associated with WC susceptibility or PT vaccine antibody response identified through cohort or stage of study.

			ALSPAC				1958BC				WC Meta			pLMM			
	Variant	BP (b37)	MAF*	Beta/ OR†	s.e.	p‡	MAF*	Beta/ OR†	s.e.	p‡	Beta	s.e.	p‡	MAF _{AV}	Beta	s.e.	p‡
WC Meta	rs9271768	32,594,188	0.48	-0.43	0.06	1.1x10⁻¹³	0.44	-0.33	0.07	3.2x10⁻⁷	-0.39	0.04	1.2x10⁻¹⁸	0.22	-0.05	0.04	0.23
ALSPAC	DRB1-9Trp	-	0.41	<i>0.63</i>	0.06	1.2x10⁻¹⁵	0.40	<i>0.80</i>	0.06	2.1x10 ⁻⁴	-0.34	0.04	1.0x10⁻¹⁶	0.29	0.29	-0.03	1.1x10⁻¹⁷
1958BC	rs111696694	32,589,122	0.35	0.32	0.06	1.9x10⁻⁸	0.35	0.37	0.06	4.5x10⁻¹⁰	0.34	0.04	5.3x10⁻¹⁷	0.60	-0.28	0.03	6.6x10⁻²⁰
PT Response	rs72851029	32,501,639	0.56	-0.38	0.07	1.7x10⁻⁷	0.57	-0.29	0.07	2.3x10 ⁻⁵	-0.33	0.05	2.9x10⁻¹¹	0.34	0.34	0.03	4.6x10⁻²⁶
PT Response	DRB1- 233Thr	-	0.35	<i>1.38</i>	0.06	7.8x10⁻⁹	0.37	<i>1.39</i>	0.06	2.0x10⁻⁸	0.33	0.04	8.5x10⁻¹⁶	0.61	-0.33	0.04	4.3x10⁻²⁴

*: MAF: minor allele frequency; MAF_{AV}: average minor allele frequency calculated across populations

†: Odds ratios (OR) are shown for amino acid residues inferred in the ALSPAC and 1958BC populations (also highlighted in *italics*), otherwise the beta statistic is presented

‡: s.e.: standard error.

¶: p-values with significant ($p < 1 \times 10^{-5}$) highlighted in bold.

Table 5.5: Summary statistics for HLA-DRB1 alleles associated with WC susceptibility or PT vaccine antibody response defined through cohort or stage of study.

4-digit alleles frequently differ between individuals of European and African ancestry at the same 2-digit serological type and therefore the respective 4-digit types are provided by population (European: ALSPAC, 1958BC and WC meta-analysis; or African: pLMM)

Allele (2-digit)*	European 4-digit	African 4-digit	ALSPAC				1958BC				WC Meta-analysis			pLMM			
			MAF [†]	OR	s.e. [‡]	p [¶]	MAF [†]	OR	s.e. [‡]	p [¶]	Beta	s.e. [§]	p [¶]	MAF _{AV} [†]	Beta	s.e. [‡]	p [¶]
DRB1*01	01:01	01:02	0.11	0.64	0.10	7.2x10⁻⁶	0.12	0.82	0.09	3.4x10 ⁻⁴	-0.31	0.07	3.8x10⁻⁶	0.06	0.28	0.06	1.5x10⁻⁶
DRB1*03	03:01	03:01	0.14	1.40	0.07	2.5x10⁻⁶	0.15	1.26	0.08	2.00x10 ⁻³	0.29	0.05	3.1x10⁻⁸	0.06	-0.33	0.06	7.2x10⁻⁷
DRB1*11	11:01	11:02	0.10	1.37	0.09	6.8x10 ⁻⁴	0.10	1.10	0.09	0.33	0.20	0.07	1.67x10 ⁻³	0.07	-0.38	0.07	7.8x10⁻⁹
DRB1*13	13:01	13:02	0.08	1.08	0.10	0.46	0.05	1.34	0.13	0.02	0.16	0.08	0.04	0.08	-0.20	0.06	4.9x10 ⁻⁴
DRB1*15	15:01	15:03	0.14	0.71	0.09	7.2x10⁻⁶	0.13	0.75	0.09	1.9x10 ⁻³	-0.31	0.06	7.8x10⁻⁷	0.14	0.31	0.05	4.8x10⁻¹²

*: Alleles only shown if $p \leq 0.05$ in both sets of studies (WC and PT antibody response) and if MAF > 0.01 in all cohorts. The most common 4-digit allele often differed between European and African populations and hence only the statistics for the more common alleles in each population were compared.

[†]: MAF: minor allele frequency; MAF_{AV}: average minor allele frequency calculated across populations

[‡]: s.e.: standard error

[¶]: p-values with significant ($p < 1 \times 10^{-5}$) highlighted in bold.

Table 5.6: Summary statistics for HLA-DQA1 and HLA-DQB1 alleles associated with WC susceptibility or PT vaccine antibody response defined through cohort or stage of study.

Allele*	ALSPAC				1958BC				Meta			Pooled LMM			
	MAF [†]	OR	s.e. [‡]	p [¶]	MAF [†]	OR	s.e. [‡]	p [¶]	Beta	s.e. [‡]	p [¶]	MAF _{AV} [†]	Beta	s.e. [‡]	p [¶]
DQA1*01:01	0.37	0.75	0.06	3.7x10⁻⁷	0.37	0.95	0.06	0.33	-0.18	0.04	1.7x10 ⁻⁵	0.16	0.18	0.04	2.3x10 ⁻⁵
DQA1*05:01	0.26	1.44	0.06	9.9x10⁻⁹	0.27	1.19	0.06	6.9x10 ⁻³	0.27	0.04	3.5x10⁻¹⁰	0.22	-0.21	0.04	6.0x10⁻⁹
DQB1*02:01	0.15	1.41	0.08	1.5x10 ⁻⁵	0.14	1.17	0.08	0.05	0.25	0.06	7.5x10⁻⁶	0.06	-0.33	0.06	9.8x10⁻⁸
DQB1*05:01	0.11	0.62	0.11	1.1x10 ⁻⁵	0.14	0.85	0.10	9.9x10 ⁻²	-0.30	0.07	3.4x10 ⁻⁵	0.17	0.14	0.04	3.7x10 ⁻⁴
DQB1*06:02	0.18	0.74	0.08	3.3x10 ⁻⁴	0.15	0.80	0.09	1.2x10 ⁻²	-0.27	0.06	1.5x10 ⁻⁵	0.22	0.21	0.04	4.9x10⁻⁸

*: Alleles only shown if demonstrated $p \leq 0.05$ in both sets of studies (WC and PT antibody response) and if MAF > 0.01 in all cohorts.

†: MAF: minor allele frequency; MAF_{AV}: average minor allele frequency calculated across populations

‡: s.e.: standard error

¶: p-values with significant ($p < 1 \times 10^{-5}$) highlighted in bold.

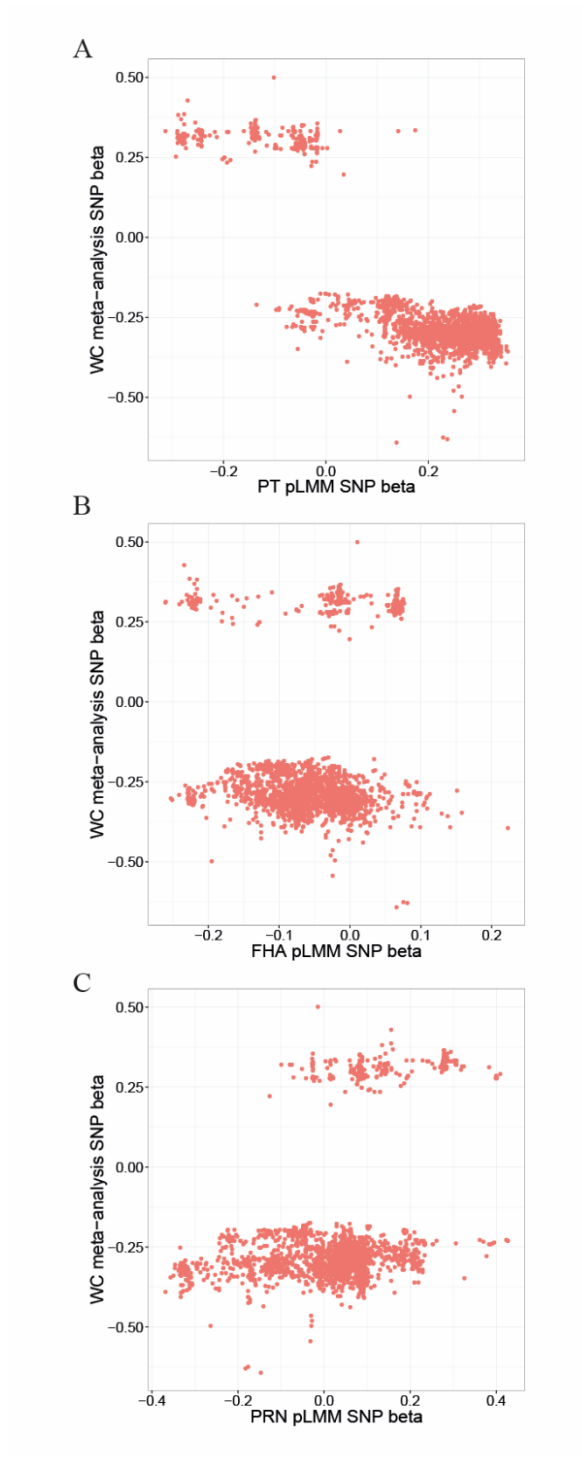


Figure 5.2: Correlation of SNP beta effect estimates between WC GWAS and GWAS of PT (A), FHA (B) and PRN (C) responses.

Beta effect estimates were not available for WC GWAS for SNPs with $p > 1 \times 10^{-5}$.

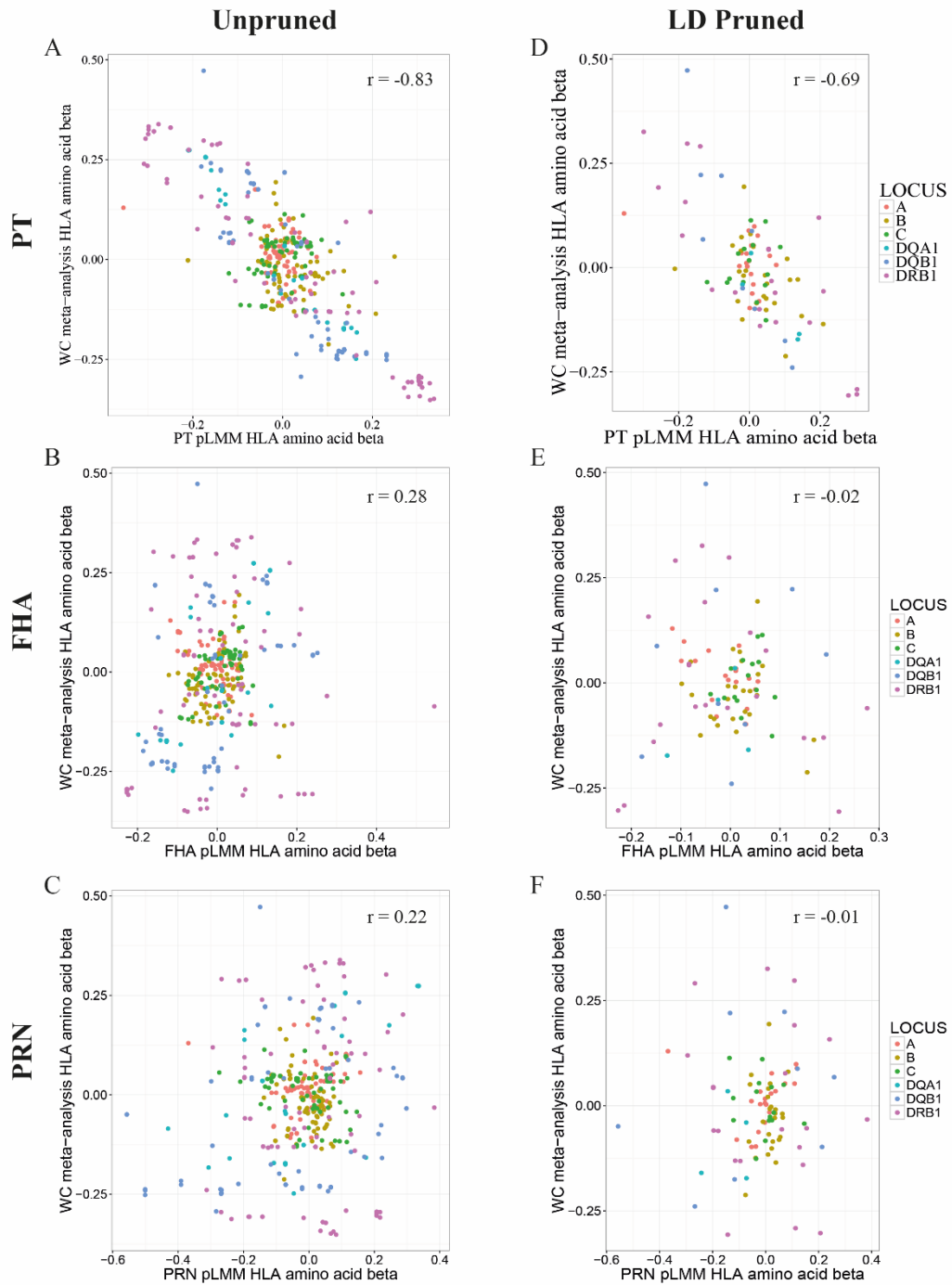


Figure 5.3: Correlation of biallelic HLA amino acid beta effect estimates between WC and PT, FHA and PRN response GWA studies.

Correlations are shown with no correction for LD (A-C) and pruning by LD ($r^2 < 0.35$; D-F) with Pearson correlation coefficients $\textcircled{r}$ represented for each trait.

5.3.2 Measuring the significance of correlation of amino acid residue associations between independent GWAS studies

Since the significance p-value calculated directly from the Pearson correlation coefficient is liable to over-inflation with large numbers of data points³¹⁷, permutation analysis was used to calculate the significance of the observed correlations between amino acid effect estimates between studies. Using the LD pruned datasets and comparing effect estimated between the PT vaccine response and the WC GWAS, no Pearson r value as prominent as that observed with the true dataset was observed after 100,000 permutations (**Figure 5.4**). The correlation observed with PT ($p_{\text{perm}} < 1 \times 10^{-5}$) was far in excess of that observed for any other trait (FHA $p_{\text{perm}} = 0.57$; PRN $p_{\text{perm}} = 0.91$; DT $p_{\text{perm}} = 0.08$; TT $p_{\text{perm}} = 7.8 \times 10^{-3}$; HBsAg $p_{\text{perm}} = 0.01$). Even after running 100,000,000 permutations, no level of correlation as strong as that seen with the true data for PT was observed ($p_{\text{perm}} < 1 \times 10^{-8}$) and further increases in numbers of permutation performed were not deemed necessary.

To test that this method was robust at detecting both positive and negative correlations through likely shared genetic architecture and that these results were not affected by any bias attributable to differences between European and African populations, the beta effect estimates for amino acid residues across the HLA were compared between the two WC studies: 1958BC and ALSPAC. A strong positive correlation ($r = 0.52$) was observed that was highly significant ($p_{\text{perm}} < 1 \times 10^{-5}$; **Figure 5.5**). Furthermore, the negative genetic correlation between the WC GWAS and those from the three independent vaccine response study sites was still observed although the association was marginally less significant for the

Soweto study ($p_{\text{perm}} = 1.0 \times 10^{-3}$) compared to Entebbe (3.4×10^{-5}) and Banfora (2.7×10^{-5}) populations (**Figure 5.6**).

Together with the similar patterns of opposing directions of effect at the HLA allele and SNP level between PT and WC GWAS these data provided strong evidence in favour of a shared genetic architecture between antibody responses raised against PT following vaccination in infancy and susceptibility to WC following vaccination. The negative correlation supported the notion that an increased antibody response was protective against WC making it likely that, following wP at least, PT antibody responses are likely to be a correlate of protection against WC. Although this test does not allow the identification of any specific variant being causally associated with the two traits (WC susceptibility and PT response), the higher effect estimates observed for HLA-DRB1 residues and alleles would strongly support an effect mediated through the HLA-DR locus. Since statistics for the HLA-DRB1 paralogues were not provided for the WC GWAS it was not possible to test whether variants in these genes were also correlated between studies, and whether they may have even greater effect estimates than those in HLA-DRB1.

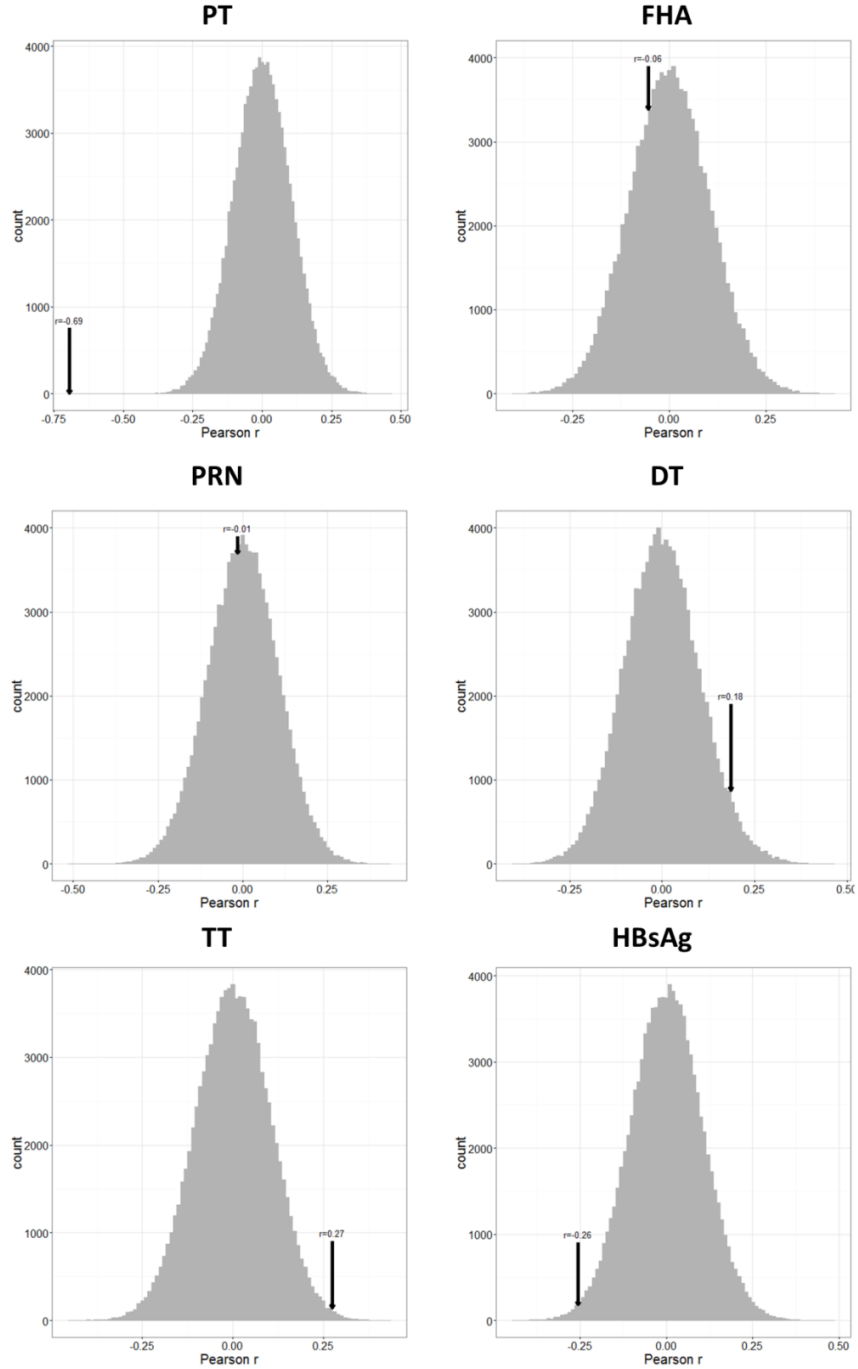


Figure 5.4: Permutation distributions (100,000 permutations) measuring the significance of correlations between effect estimates of HLA amino acids pruned by LD ($r^2 < 0.35$) from WC and vaccine response GWAS.

The distributions were determined through counts of Pearson correlation coefficients observed after relabelling of the WC variants generating the null distribution. The correlation coefficients measured between the true datasets are represented with vertical arrows.

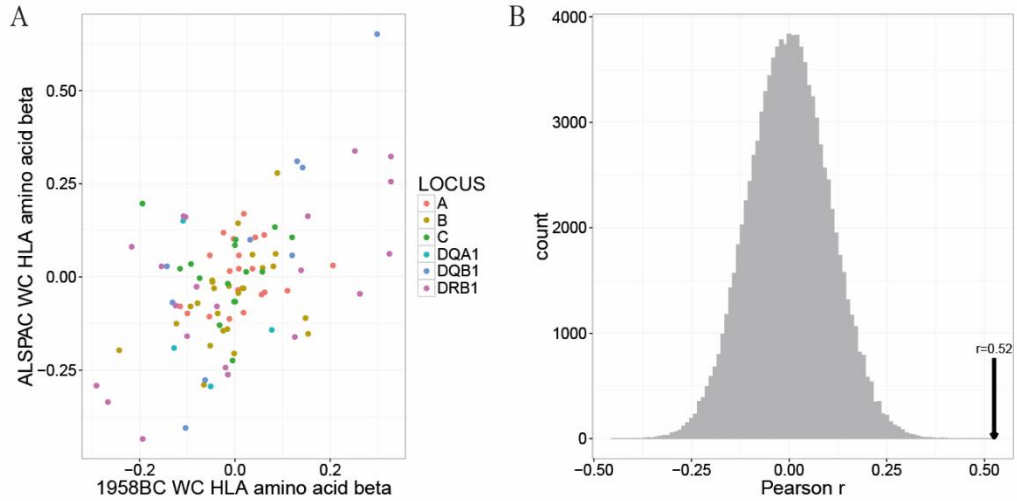


Figure 5.5: Correlation between LD pruned HLA amino acid effect estimates for WC susceptibility between 1958BC and ALSPAC studies (A) and the corresponding permutation distribution (100,000 permutations) to measure significance (B).

The correlation coefficient measured with the true dataset is represented with a vertical arrow.

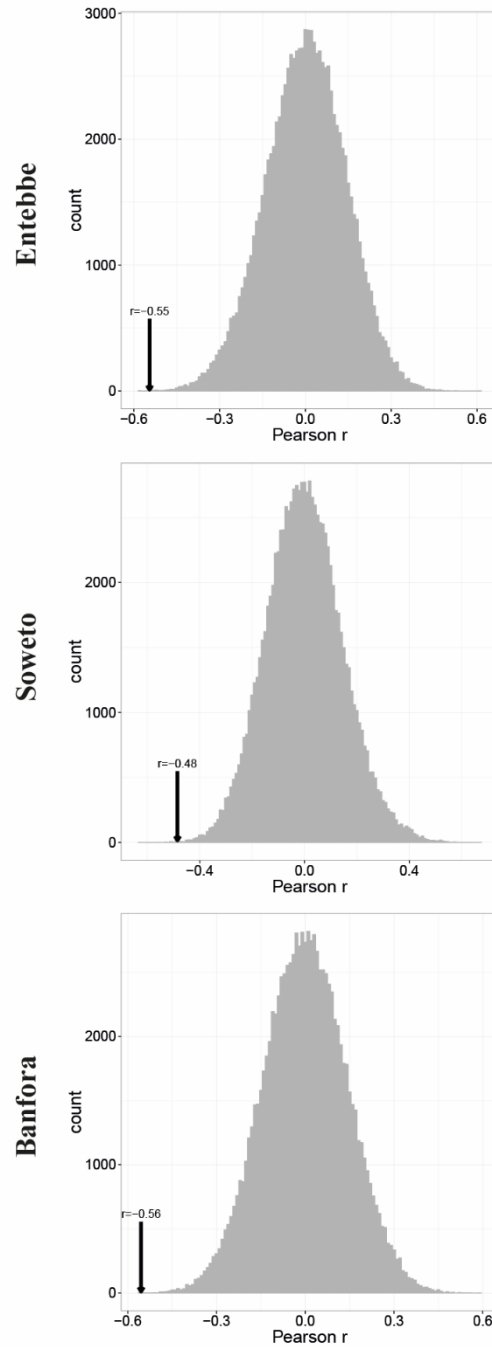


Figure 5.6: Permutation distributions (100,000 permutations) testing significance of correlations between HLA amino acid effect estimates for WC and PT vaccine responses in the three individual tested cohorts.

The distributions were determined through counts of Pearson correlation coefficients observed after relabelling of the WC variants generating the null distribution. The correlation coefficients measured between the true datasets are represented with vertical arrows.

5.3.3 Identifying cellular mechanisms correlating with protection against pertussis

All individuals in the Sette donor cell repository used to interrogate the mechanisms responsible for responding to PT and DT had HLA-DRB1 types available. In contrast, only a limited number had DRB paralogue types (including HLA-DRB3) available. Therefore only HLA-DRB1 signals were assessed as part of the primary analysis. The AIM assay was used to test for differences in levels of circulating antigen-specific T_{FH} between individuals carrying different HLA-DRB1 amino acid residue alleles associated with differential PT vaccine response and WC susceptibility: DRB1-233Thr (associated with increased antibody response and reduced WC susceptibility) or DRB1-233Arg. These HLA-DRB1 variants were observed at approximately a MAF of 50% in both African and European populations and therefore the individuals studied as part of the AIM assay were divided into groups either carrying two copies of HLA-DRB1*03, HLA-DRB1*11 or HLA-DRB1*13 (each containing DRB1-233Arg), or two copies of HLA-DRB1*01 or HLA-DRB1*15 (containing DRB1-233Thr). Characteristics of the individuals included in the final cell analysis are provided in **Table 5.7**.

The ratio of circulating PT- to TT-specific T_{FH} cells was on average 1.2 fold greater in individuals carrying two copies of DRB1-233Thr compared to those carrying two copies of DRB1-233Arg (**Figure 5.7**; Mann-Whitney U $p = 0.007$).

Table 5.7: Characteristics of donors and PBMC samples used for PT-specific T_{FH} assay.

	DRB1-233Arg	DRB1-233Thr
Final number	15	14
Number of Males (%)	7 (46.7)	6 (42.9)
Mean age at sampling in years (s.d.)	46.1 (23.7)	43.2 (24.3)
Mean storage time of PBMCs in years (s.d.)	3.87 (2.1)	3.81 (2.7)
<u>Participant Ethnicity</u>		
Asian number (%)	0	1 (7.1)
Pacific Islander number (%)	1 (6.7)	3 (21.4)
Black number (%)	6 (40.0)	0
Hispanic / Latino number (%)	3 (20.0)	0
White number (%)	4 (26.7)	9 (64.2)
Unknown number (%)	1 (6.7)	1 (7.1)
<u>4-Digit HLA-DRB1 Alleles</u>		
01:01 number (%)	0	9 (32.1)
01:02 number (%)	0	2 (7.1)
03:01 number (%)	4 (13.3)	0
03:02 number (%)	1 (3.3)	0
03:17 number (%)	1 (3.3)	0
11:01 number (%)	7 (23.3)	0
11:02 number (%)	4 (13.3)	0
11:04 number (%)	4 (13.3)	0
13:01 number (%)	4 (13.3)	0
13:02 number (%)	1 (3.3)	0
13:03 number (%)	2 (6.7)	0
13:04 number (%)	2 (6.7)	0
15:01 number (%)	0	9 (32.1)
15:02 number (%)	0	7 (25.0)
15:09 number (%)	0	1 (3.6)

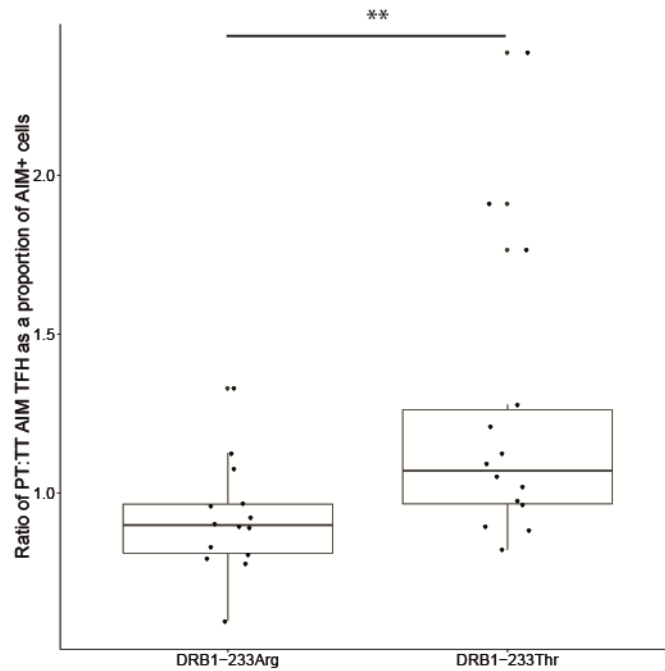


Figure 5.7: Ratio of circulating PT:TT specific T_{FH} in donors of known HLA-DRB1 type.

** : Mann-Whitney U $p < 0.01$

5.3.4 HLA-DRB3 affecting PT response

Individuals carrying DRB1-233Arg were expected to also carry HLA-DRB3 alleles. The top HLA variant associated with PT response was DRB3-74Gln which divided individuals into those carrying HLA-DRB3*02:01:01G and HLA-DRB3*03:01:01G (both with DRB3-74Gln) associated with lower levels of PT antibody, compared to individuals with HLA-DRB3*01:01:01 (DRB3-74Arg). A secondary analysis was performed using the samples selected by HLA-DRB1 to test whether individuals carrying HLA-DRB3*02:01:01G or HLA-DRB3*03:01:01G induced lower levels of PT-specific T_{FH} compared to individuals either with HLA-DRB3*01:01:01 or those with HLA-DRB3 absent on

the haplotype. Although not all individuals in the complete Sette database had types available for the HLA-DRB1 paralogue genes, by chance, all individuals selected based on HLA-DRB1 status for the above experiments did have HLA-DRB3 types available. Consistent with the independent HLA-DRB3 signal observed in the GWAS, individuals carrying either HLA-DRB3*02 or *03 alleles had a 1.28 fold difference in proportions of PT-specific T_{FH} ratios compared to individuals carrying either HLA-DRB3*01 alleles or no HLA-DRB3 (Mann-Whitney U $p = 0.04$; **Figure 5.8A**). However, individuals carrying HLA-DRB3*01 alleles also had lower levels of T_{FH} suggesting that the effect was unlikely to be specific to specific HLA-DRB3 alleles or amino acid residues and that the effect was instead more likely to be due to linkage with other variants; that may include those within HLA-DRB1 on the HLA-DRB3 carrying haplotypes (**Figure 5.8B**).

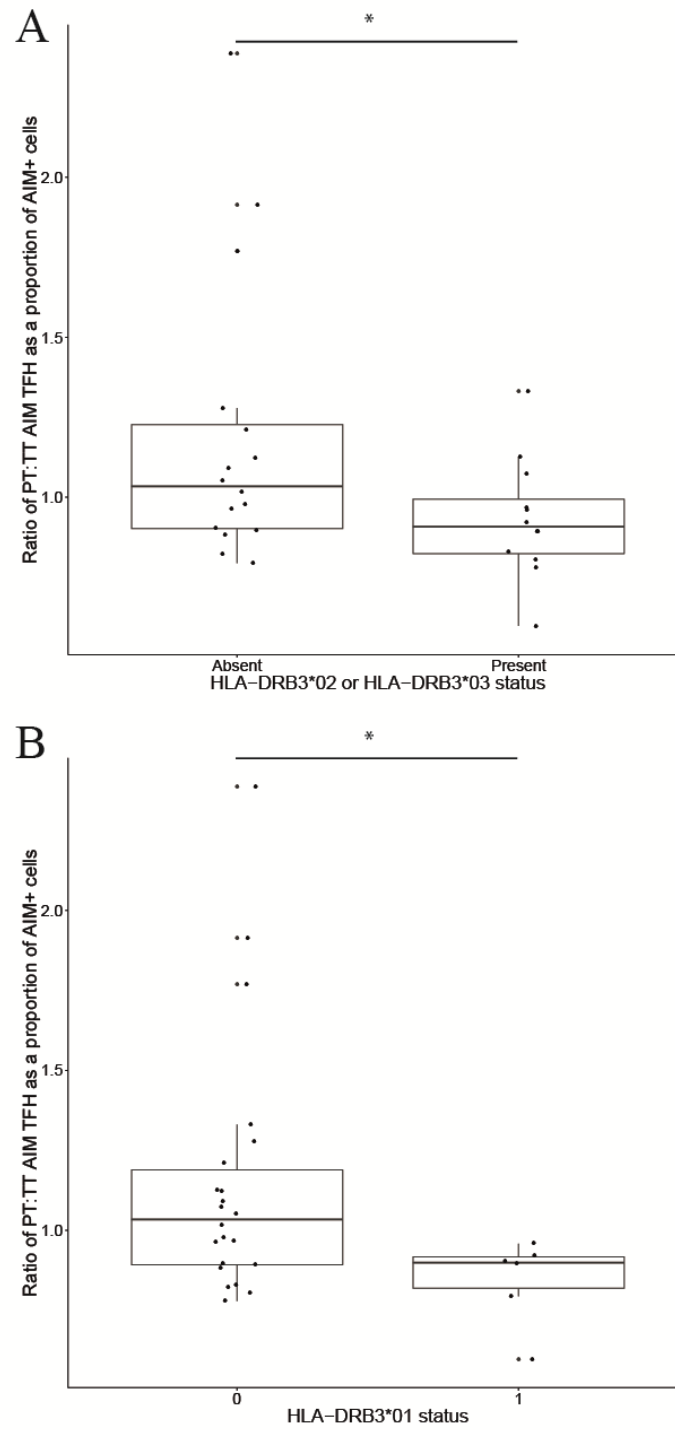


Figure 5.8: Ratio of circulating PT:TT specific T_{FH} in donors with known numbers of copies of HLA-DRB3*02 or HLA-DRB3*03 (A) or HLA-DRB3*01 (B) alleles.

*: Mann-Whitney U $p < 0.05$

5.3.5 Identifying cellular mechanisms correlating with protection against diphtheria

The same T_{FH} assay methodology was applied to other donors stratified by HLA-DRB1 alleles associated with differential response to DT. The characteristics of the donors are provided in **Table 5.8**. A non-significant (Mann-Whitney U $p = 0.173$) difference in mean storage time of the PBMC samples was observed between the groups. No difference was observed in ratios of DT-specific T_{FH} compared to TT-specific cells ($p = 0.28$; **Figure 5.9**). A trend in the opposite direction to that which was expected was instead observed; a lower ratio of DT:TT T_{FH} in the individuals carrying DRB1-4Gln (associated with increased antibody levels) compared to those carrying DRB1-4Arg.

Table 5.8: Characteristics of donors and PBMC samples used for DT-specific T_{FH} assay.

	DRB1-4Gln	DRB1-4Arg
Final number	17	19
Number of Males (%)	5 (33.3)	7 (38.9)
Mean age at sampling in years (s.d.)	43.3 (25.8)	43.2 (24.3)
Mean storage time of PBMCs in years (s.d.)	2.98 (2.17)	4.31 (2.6)
<u>Participant Ethnicity</u>		
Asian number (%)	2 (11.8)	1 (5.3)
Pacific Islander number (%)	2 (11.8)	3 (15.8)
Black number (%)	1 (5.9)	1 (5.3)
Hispanic / Latino number (%)	2 (11.8)	1 (5.3)
Sri Lankan	1 (5.9)	0
White number (%)	7 (41.2)	12 (63.2)
Unknown number (%)	2 (11.8)	1 (5.3)
<u>4-Digit HLA-DRB1 Alleles</u>		
01:01 number (%)	0	10 (26.3)
01:02 number (%)	0	3 (7.9)
01:03 number (%)	0	2 (5.3)
07:01 number (%)	21 (61.8)	0
09:02 number (%)	13 (38.2)	0
15:01 number (%)	0	14 (36.8)
15:02 number (%)	0	7 (18.4)
15:03 number (%)	0	1 (2.6)
15:09 number (%)	0	1 (2.6)

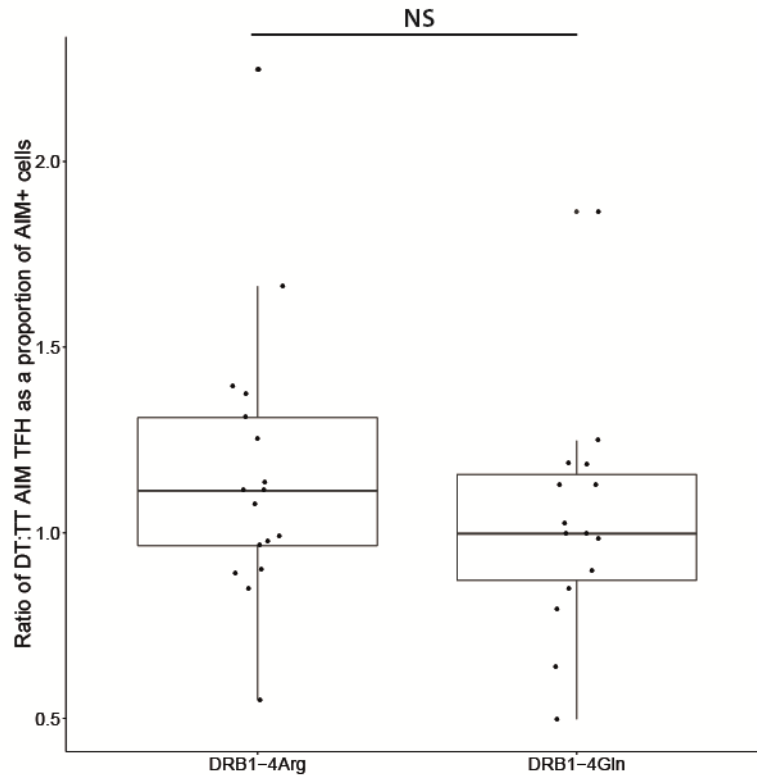


Figure 5.9: Ratio of circulating DT:TT specific T_{FH} in donors of known HLA-DRB1 type.

NS: not significant.

5.3.6 Peptide binding studies

In silico HLA-peptide binding approaches were used to test whether PT and DT derived peptides bound differentially to associated HLA alleles both in terms of affinity and breadth.

5.3.6.1 *Pertussis toxin*

HLA-DRB3*01 alleles bound their top predicted PT-derived peptides with significantly greater affinity than HLA-DRB3*02 or *03 alleles ($p = 0.007$; **Table 5.9**). In contrast, no differences in levels of affinity were observed for those

peptides predicted to bind with the greatest affinity to HLA-DRB1 alleles when comparing those alleles demonstrating increased or decreased PT antibody responses in the African GWAS (including only those alleles with MAF_{AV} greater than 0.01; **Table 5.10**).

In order to improve power to observe a difference based on HLA amino acid sequences, all HLA-DRB1 alleles present in the IMGT database were divided into groups dependent on their sequences and whether they possessed an excess of residues associated with either increased or decreased antibody responses (as detailed in the Methods section in this chapter). No differences were observed between these groups divided by effects on PT antibody responses. Furthermore, no differences in terms of breadth of PT-derived peptide binding were observed at either the HLA-DRB3 (**Table 5.11**) or HLA-DRB1 (**Table 5.12**, **Figure 5.10**) loci when compared with TT-derived peptides. A greater breadth of TT-derived peptides were predicted to bind to all HLA-DRB1 alleles tested but no differences were observed between alleles grouped by association status (DRB1-233Thr or DRB1-233Arg).

Table 5.9: Differences in predicted affinities of the peptides predicted to bind most tightly to HLA-DRB3 alleles comparing alleles associated with increased (DRB3-74Arg) or decreased PT antibody responses.

	Allele	PT-derived sequence	Predicted Affinity*	p-value†
DRB3-74Arg	HLA-DRB3*01:01	GDLCVFTRDGNPVV	0.32	
	HLA-DRB3*01:05	LCAVFTRDGNPVVQA	0.79	
	HLA-DRB3*01:11	LCAVFTRDGNPVVQA	0.74	
	HLA-DRB3*01:12	LCAVFTRDGNPVVQA	0.74	
	Mean (s.d.)		0.65 (0.22)	
DRB3-74Gln	HLA-DRB3*02:02	LETLRQGLPLRVYV	1.88	
	HLA-DRB3*02:10	LETLRQGLPLRVYV	1.88	
	HLA-DRB3*02:24	ETLRQGLPLRVYVD	1	
	HLA-DRB3*03:01	LETLRQGLPLRVYV	2	
	Mean (s.d.)		1.69 (0.46)	0.007

*: Percentile rank calculated in IEDB

†: Comparing DRB3-74Gln and DRB3-74Arg allele groups, Mann-Whitney U test

Table 5.10: Differences in predicted affinities of the peptides predicted to bind most tightly to HLA-DRB1 alleles comparing alleles associated with increased (DRB1-233Thr) or decreased PT antibody responses.

	Allele	PT-derived sequence	Predicted Affinity*	p-value†
DRB1-233Thr	HLA-DRB1*01:02	MKRISPRAGWLAASL	1.77	
	HLA-DRB1*15:03	ASLLCLLLPGARAWA	0.52	
	Mean (s.d.)		1.15 (0.88)	
DRB1-233Arg	HLA-DRB1*03:01	KRARLADIRGLRDHF	0.54	
	HLA-DRB1*11:01	RESYATLRRRALETLR	0.63	
	HLA-DRB1*11:02	LADIRGLRDHFCLAS	0.39	
	HLA-DRB1*13:02	MRELSGNEPWPIHGL	4.19	
	Mean (s.d.)		1.44 (1.84)	0.24

*: Percentile rank calculated in IEDB

†: Comparing DRB1-233Arg and DRB1-233Thr allele groups, Mann-Whitney U test

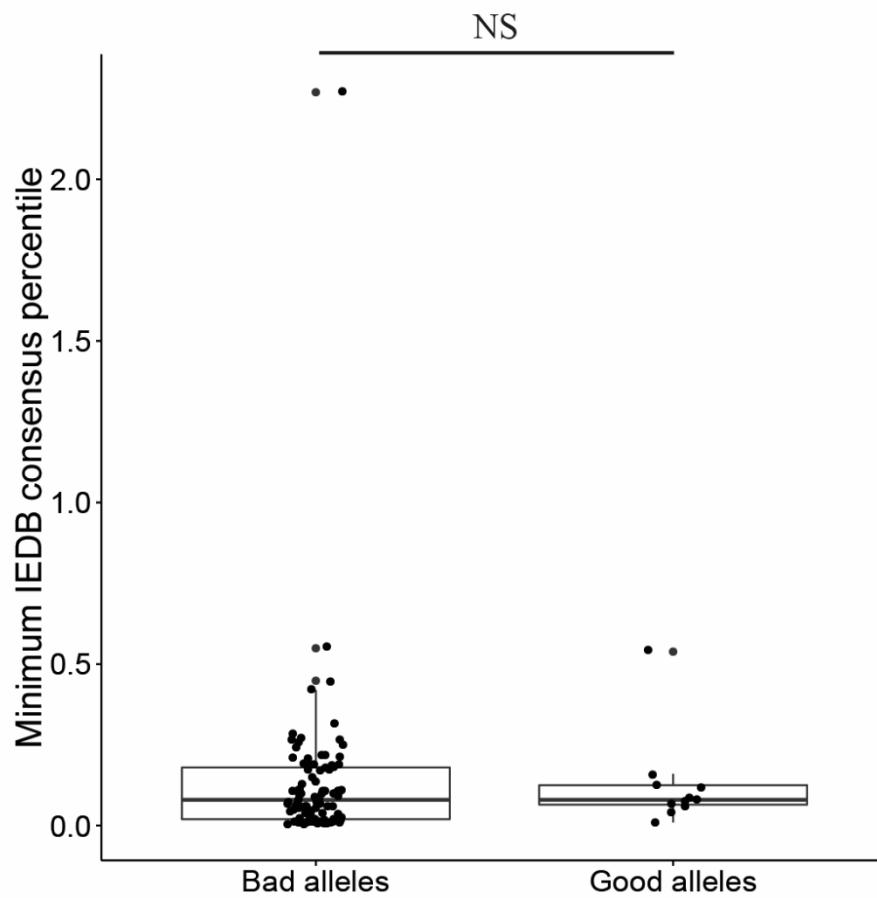


Figure 5.10: IEDB predicted affinities for PT-derived peptides predicted to bind to all IMGT HLA-DRB1 alleles divided into groups dependent on likely effect on PT antibody responses.

HLA alleles with more than 1.5x ratio of residues increasing PT antibody responses ('Good alleles') were compared to alleles with more than 1.5x ratio of residues associated with reduced PT responses ('Bad alleles'). NS: not significant.

Table 5.11: Breadth of PT-derived peptides binding to associated HLA-DRB3 alleles.

	Allele	PT Breadth*	TT Breadth*	p-value
DRB3-74Arg	HLA-DRB3*01:01	6.17	14.83	
	HLA-DRB3*01:05	3.96	9.15	
	HLA-DRB3*01:11	3.96	9.22	
	HLA-DRB3*01:12	3.96	9.22	
	Average	4.5 (+/- 1.1)	10.6 (+/- 2.8)	0.003
DRB3-74Gln	HLA-DRB3*02:02	3.08	7.38	
	HLA-DRB3*02:10	3.08	7.38	
	HLA-DRB3*02:24	3.08	4.77	
	HLA-DRB3*03:01	1.32	5.69	
	Average	2.6 (+/- 0.9)	6.3 (+/- 1.3)	0.002

*:Values represent % of peptides derived from the corresponding toxin that are predicted to bind with high affinity (+/- standard deviation) calculated according to the 5th percentile rank in IEDB.

Table 5.12: Breadth of PT-derived peptides binding to associated HLA-DRB1 alleles.

	Allele	PT Breadth*	TT Breadth*	p-value
DRB1-233Thr	DRB1*01:01	3.1	4.6	
	DRB1*01:02	24.2	26.3	
	DRB1*15:01	3.5	11.3	
	DRB1*15:03	3.1	8.1	
	Average	8.5 (+/- 10.5)	12.6 (+/- 9.5)	NS
DRB1-233Arg	DRB1*03:01	22.03	26.7	
	DRB1*11:01	9.3	6.5	
	DRB1*11:02	0.4	7.5	
	DRB1*13:02	12.3	22.4	
	Average	11.0 (+/- 8.9)	15.8 (+/- 10.3)	NS

*:Values represent % of peptides derived from the corresponding toxin that are predicted to bind with high affinity (+/- standard deviation) calculated according to the 5th percentile rank in IEDB.

5.3.6.2 *Diphtheria toxin*

When applying the method of testing for differences in affinities of DT-derived peptides across all IMGT alleles, significant evidence was found in favour of DT-derived peptides binding with a greater affinity to ‘Good’ alleles compared to ‘Bad’ (**Figure 5.11**). Furthermore HLA-DRB1 alleles encoded by DRB1-4Gln (HLA-DRB1*07:01:01G and HLA-DRB1*09:01:02G; associated with increased DT antibody responses) were observed to bind a greater breadth of DT-derived peptides to compared to other alleles observed in the African dataset (**Table 5.13**). These two alleles associated with increased antibody responses were found to have strikingly similar breadths of peptide binding (equivalent to TT peptides at the same alleles) whereas every other tested allele demonstrated a reduced breadth compared to TT. Testing for differences in affinity comparing only those alleles observed in the African GWAS was likely underpowered owing to the small

number of alleles available for testing between the increased and decreased responder groups (**Table 5.14**).

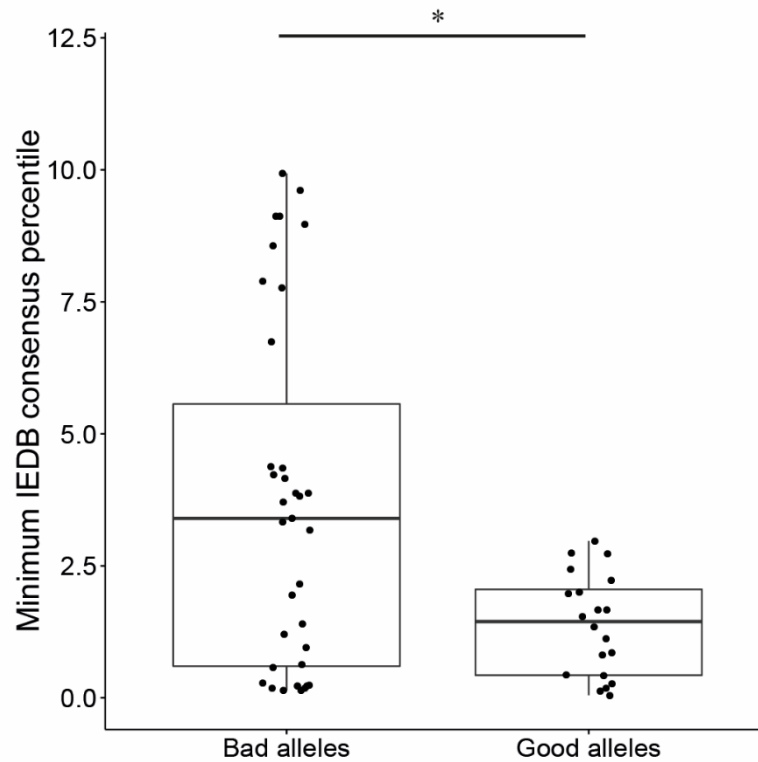


Figure 5.11: IEDB predicted affinities for DT-derived peptides predicted to bind to all IMGT alleles divided into groups dependent on likely effect on DT antibody responses.

HLA alleles with more than 1.5x ratio of residues increasing DT antibody responses ('Good alleles') were compared to alleles with more than 1.5x ratio of residues associated with reduced DT responses ('Bad alleles').*: Mann-Whitney U $p < 0.05$

Table 5.13: Breadth of DT-peptides binding to associated HLA-DRB1 alleles.

Association	Allele	DT Total*	TT total*	p
DRB1-4Gln	DRB1*07:01	9.3	7.7	
	DRB1*09:01	9.9	9.7	
	Average	9.6 (+/- 0.4)	8.7 (+/- 1.5)	0.23
DRB1-4Arg	DRB1*01:01	2.4	4.6	
	DRB1*01:02	18.2	26.2	
	DRB1*03:01	15.5	20.8	
	DRB1*03:02	2.0	6.7	
	DRB1*08:04	19.7	31.3	
	DRB1*10:01	1.5	7.3	
	DRB1*11:01	13.2	22.8	
	DRB1*11:02	23.4	26.7	
	DRB1*11:04	17.7	26.6	
	DRB1*12:01	3.1	5.5	
	DRB1*13:01	18.4	29.7	
	DRB1*13:02	7.1	7.6	
	DRB1*13:03	2.2	6.1	
	DRB1*14:01	0.7	5.1	
	DRB1*15:03	0	8.5	
	Average	9.7 (+/- 8.5)	15.7 (+/- 10.5)	5.5 x 10⁻³

*:Values represent % of peptides derived from the corresponding toxin that are predicted to bind with high affinity (+/- standard deviation) calculated according to the 5th percentile rank in IEDB.

Table 5.14: Differences in HLA-DRB1 DT-peptide binding affinity.

	Allele	DT-derived sequence	Predicted Affinity*	p-value†
DRB1-4Gln	HLA-DRB1*07:01	ANLHVAFHRSSEKI	0.13	
	HLA-DRB1*09:01	HVAFHRSSEKIHSN	0.02	
	Mean (s.d.)		0.08 (0.08)	
DRB1-4Arg	HLA-DRB1*01:01	IALSSLMVAQAIPLV	3.49	
	HLA-DRB1*01:02	SIALSSLMVAQAIPL	0.28	
	HLA-DRB1*03:01	LSCINLDWDVIRDKT	0.08	
	HLA-DRB1*03:02	HVAFHRSSEKIHSN	0.52	
	HLA-DRB1*08:04	SVNGRKIRMRCRAID	0.02	
	HLA-DRB1*10:01	TEEIVAQSIALSSLM	2.35	
	HLA-DRB1*11:01	EKTTAALSILPGIGS	1.66	
	HLA-DRB1*11:02	HHNTEEIVAQSIALS	0.14	
	HLA-DRB1*11:04	NYAAWAVNVAQVIDS	0.3	
	HLA-DRB1*12:01	RKLFASILIGALLGI	2.25	
	HLA-DRB1*13:01	HHNTEEIVAQSIALS	0.23	
	HLA-DRB1*13:02	LTKVLALKVDNAETI	0.61	
	HLA-DRB1*13:03	LHVAFHRSSEKIHS	1.56	
	HLA-DRB1*14:01	YPGLTKVLALKVDNA	3.03	
	HLA-DRB1*15:03	SIINLFQVVHNSYNR	5.91	
	Mean (s.d.)		1.50 (1.68)	0.24

*: Percentile rank calculated in IEDB

†: Comparing DRB1-4Gln alleles to all others with $MAF_{AV} \geq 0.01$ in African DT response GWAS

5.3.6.3 Validation of *in silico* methods using *in vitro* binding assays

The correlation between the predicted and measured binding affinities for 28 15-mer peptides derived from DT predicted to bind most tightly to three HLA alleles was measured (two antibody increasing alleles: HLA-DRB1*07:01 and HLA-DRB1*09:01; and one intermediate: HLA-DRB1*15:01). The affinities calculated through estimation and direct assay were correlated (**Figure 5.12**) although they were more tightly correlated for the two ‘increasing’ alleles (Pearson *r* of 0.68 and 0.60 respectively) compared to HLA-DRB1*15:01 (0.33).

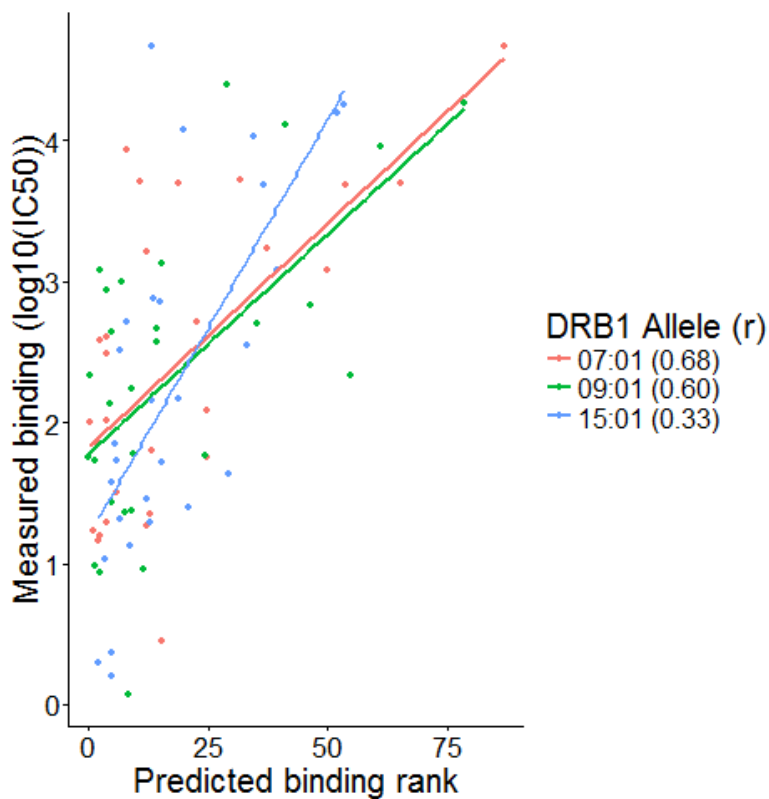


Figure 5.12: Correlation between predicted (IEDB) and measured binding affinities for 28 peptides derived from DT.

The Pearson's *r* correlation coefficient is shown in the legend for each allele.

5.3.7 Using the HLA residue association signal to correlate antibody responses to T-cell receptor expression

Levels of circulating levels of TCR V gene expression have recently been found to be associated with variation in HLA in a study of human adults³¹⁵. To test the hypothesis that levels of specific TCR V genes may be differentially upregulated in the presence of particular HLA variants and consistent vaccine antigen exposure between individuals, the approach used to measure genetic correlation between PT response and WC susceptibility was applied to the summary statistics from the TCR expression dataset. Association results from a total of 84 V genes (44 TRAV and 40 TRBV) were available for testing and positive correlations were observed between the HLA amino acid residue association beta statistics calculated for each of the vaccine antibody responses measured in the African infants and at least one TCR V gene. However, by using the same permutation method used with the WC analysis the observed correlation was only found to be significant for PT response ($r = 0.55$, **Figure 5.13**). 10,000,000 permutations were required to calculate a p-value for this correlation observed between PT vaccine response and *TRBV29-1* associations ($p_{\text{perm}} = 4 \times 10^{-7}$). In line with the observed positive correlation, the index HLA-DRB1 residue associated with PT response (DRB1-233Arg) was associated in the same direction as that observed with *TRBV29-1* expression. A less significant but still positive correlation between vaccine response and TCR V gene expression was observed with HBsAg and *TRAV3* expression ($r = 0.47$; $p_{\text{perm}} = 5 \times 10^{-5}$).

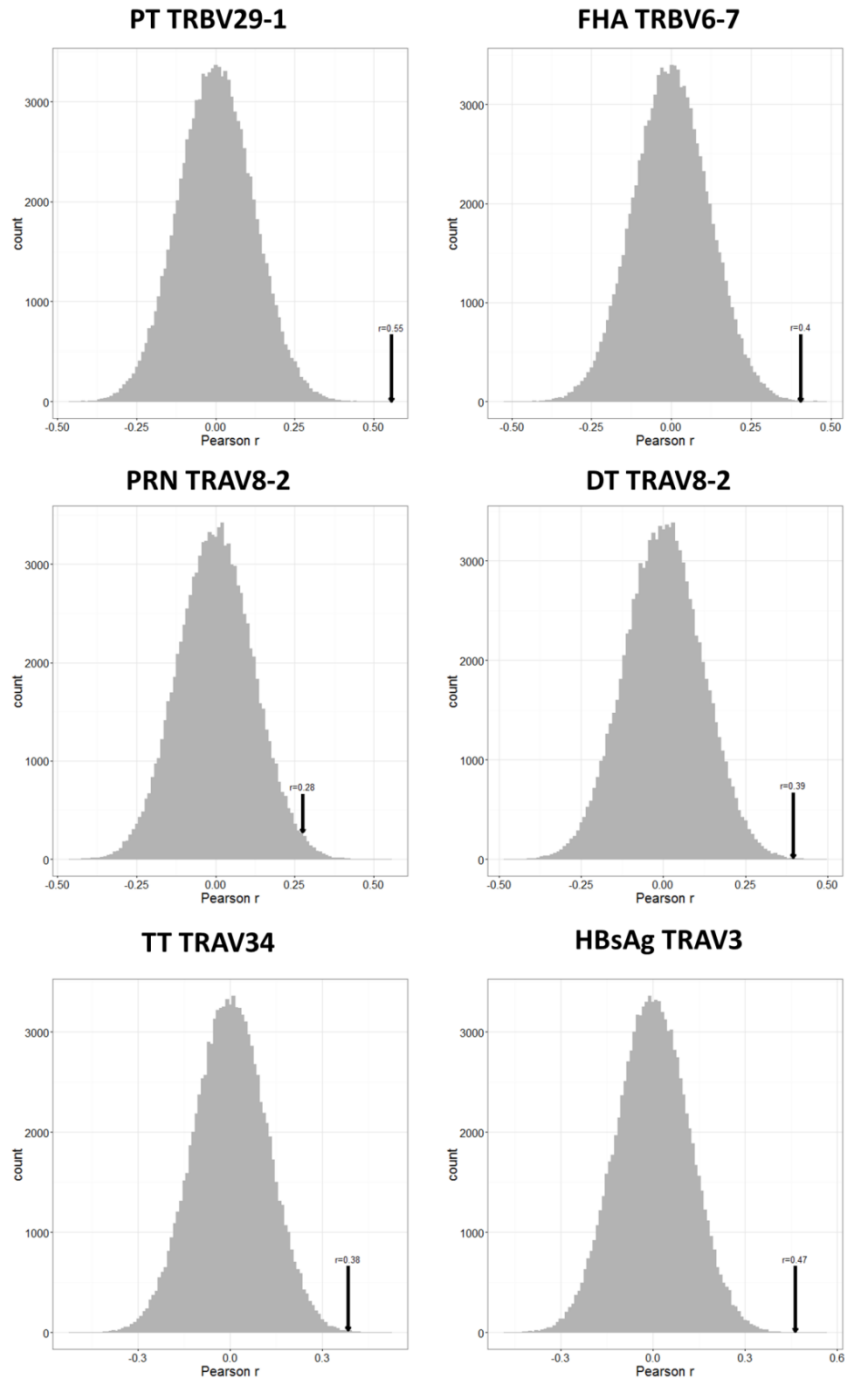


Figure 5.13: Permutation distributions (100,000 permutations) measuring the significance of correlations between effect estimates of HLA amino acids pruned by LD from TCR V gene and vaccine response GWAS.

The distributions were determined through counts of Pearson correlation coefficients observed after relabelling of the TCR study SNPs generating the null distribution. The correlation coefficients determined from testing correlation between the true datasets are represented with vertical arrows.

5.4 Conclusions

The findings presented in this chapter represent the first time that significant evidence of genetic correlation across the MHC region has been reported between two related but distinct traits tested in two independent GWAS studies. Furthermore, through identifying these relationships, the findings and methodologies applied from these analyses were expanded to identify similar stratifications in terms of antigen-specific T_{FH} (in the case of PT) and HLA allele-peptide binding (in DT) as well as specific TCR gene expression levels (for PT and HBsAg). These findings have multiple important implications for future vaccine design and development strategies.

Firstly the approach of testing for genetic correlation allowed an interpretation of the directionality of the relationship between tested traits. The striking negative correlation observed between the HLA genetic association signal effect sizes derived from PT antibody responses compared to those derived from self-reported WC would be consistent with a model of increased PT antibody levels correlating with reduced risk of self-reported WC. Although it is not possible to formally test causality using this analysis, these findings support the hypothesis that PT antibodies represent at least one correlate of protection against WC, relevant across populations of at least two ancestries: European and African as tested here.

Traditionally, testing for correlation or sharing of genetic architectures between different traits has been undertaken using methods such as meta-analysis (assuming that the direction of effects between traits are consistent)³¹⁸, or testing for pleiotropy on a locus-by-locus basis³¹⁹. More recently, methods have been developed whereby correlations between genotype array-based estimates of

heritabilities are measured³²⁰. However these methods depend on the availability of summary statistics from across the genome in all datasets and calculation of LD patterns based on populations of a single ancestry, and furthermore these calculations frequently exclude the MHC region owing to the complex LD patterns observed across the locus¹⁴⁵. Another potential alternative for testing for the presence of a relationship between two traits would be to apply the method of Mendelian randomisation. A key requirement of this method, however, is for the variant used as the instrumental variable to not demonstrate evidence of pleiotropy. For HLA alleles, especially those of *HLA-DRB1*, it is difficult to assume lack of pleiotropy given the frequent reports of associations of such alleles with multiple traits either through predicted direct causation or LD. Therefore the approach used here was simplified to test for correlation between traits using only measured effect estimates across the MHC with prior removal of those variants in high LD followed by permutation to test for statistical significance. The strength of this method depends, in part, on the complex patterns of association that will occur across the region exacerbated by the long-range LD, and the likelihood of multiple independent signals being present and being distinctive to individual traits. In such a scenario, even if two unrelated traits were associated with the same single allele (a *HLA-DRB1* allele for example) as a result of pleiotropy, the presence or absence of independent associations with even one other allele in one of the traits would result in a loss of correlation between at least some residue effect estimates when using the method applied here. Although it is still possible that the correlation of associations observed between PT response and WC are a result of pleiotropy not related to shared mechanism, the chances of this occurring with two such related phenotypes is very small. To add further evidence in favour

a biological link, the availability of the other antigen responses (including PRN and FHA as well as the other traits: TT, Hib and HBsAg) allowed for their efficient use as controls where such correlations would not be expected. Another benefit to the method applied here was that the analysis did not depend on any HLA classification or nomenclature system. A strong relationship between traits measured in populations of differing ancestries was observed at the amino acid residue level whereas equivalent comparisons would not have been possible using the 4-digit classification system as detailed in **Table 5.5**.

The correlate of protection against pertussis has been debated for many years and although PT antibodies have always been regarded as the most likely candidate, roles for other antigen responses have also been postulated⁵⁰. Confirming the identity of such a correlate is challenging and classically would require large-scale, long term follow up studies to measure vaccine response and measure disease incidence in the context of low prevalence disease states. The methods here therefore offer a new, more cost and time-effective set of systems to identify the correlates. The lack of genetic correlation between PRN and FHA responses and the WC GWAS suggest that these other antigen responses are less likely to represent correlates of protection against WC as compared to PT responses. It is possible that the absence of correlation may be due to the genetic associations observed with the antibody responses being population specific and that the genetic architectures of the same responses are completely different in European populations, or even that they are absent. Such a hypothesis may be especially true for the FHA signal given the African-specific nature of the HLA-DRB1*08:04 association but it would be less likely for the PRN HLA-DQ associations where the DQ alleles have similar frequencies across European and

African populations. Ultimately, acquiring definitive evidence in support or against this theory would only be possible through the undertaking of vaccine response GWAS in a set of European populations equivalent to those presented in this thesis for African population.

The observation of a correlation between MHC genetic association signals and both PT vaccine response and self-reported WC led to efforts to test for similar correlations with immunological pathways to substantiate and conceptualise mechanisms associated with varied vaccine responses. The pathways tested in this chapter represented the axes most frequently associated with differential antibody responses: those of APC presenting peptides (through *HLA* gene products) to TCR, in turn inducing T-cell clonal proliferation and ultimately antibody production and maturation through, in part, T_{FH} -B cell interaction. Stratifying individuals by their carriage of the index HLA-DRB1 amino acid residue demonstrated differential proportions of PT-specific T_{FH} in PT recipients and different expression levels of a specific TCR V gene: *TRBV29-1*. Despite these striking observations, no differences in affinity or breadth of PT-derived peptide binding to these alleles were observed. In contrast, groups of HLA-DRB3 alleles associated with differential PT responses demonstrated significant differences in affinities of HLA-peptide binding even though no robust differences in proportions of PT-specific T_{FH} were observed between individuals carrying different HLA-DRB3 alleles. Moreover, a significant increase in affinity and breadth of peptide binding was observed with those HLA-DRB1 alleles associated with increased DT response over and above all other alleles, despite no significant differences being observed for either DT-specific T_{FH} proportions or TCR V gene expression.

When taken together with the genetic correlation findings for PT response and WC susceptibility, the PT-specific T_{FH} and TCR V gene expression data findings are consistent with a model whereby individuals carrying specific variants mount increased, more persistent levels of PT-specific antibody associated with increased expression of particular TCR V genes and circulation of PT-specific T_{FH} cells. These increased responses appear to offer the most protection against WC. Precisely identifying the causative variants driving these associations remains challenging due to the classic problem of LD across the HLA region¹⁵⁷. However interpreting the comparisons of T_{FH} using HLA-DRB1 or HLA-DRB3 allele descriptors would suggest the causative variants are closely related to the HLA-DRB1 alleles. Antigen-specific T_{FH} have only recently been experimentally linked with vaccine efficacy owing largely to challenges in identifying them in the circulation due to low levels of specific cytokine response and poorly characterised cell marker expression¹⁰⁹. However the growing appreciation of their critical role in plasmablast and antibody maturation and production¹⁰⁵ make them strong candidates for being the main orchestrators for effective and protective vaccine responses. The first reported example of such an important role comes from a recent study using an investigational vaccine rVSV-ZEBOV where a different flow-based method to that employed in this thesis demonstrated that levels of antigen-specific T_{FH} were correlated with antibody responses against Ebola (that have been deemed as correlated of protection through primate models). In the Ebola vaccine study, once stimulated, levels of T_{FH} appeared to be maintained up to 56 days post-vaccination which was the limit of measurement applied by the investigators¹¹¹. The findings presented in this thesis are in keeping with these Ebola vaccine observations and suggest that levels of antigen-

specific T_{FH} may remain increased in the circulation for many years following vaccination. Maintaining higher levels of specialised antigen-specific memory cells in the circulation would permit a more rapid recruitment of such cells to the tissue or lymph sites necessary upon antigen re-exposure, facilitating more rapid and specific adaptive immunity that may increase the likelihood of protection. These steps would all be critical in ensuring adequate protection.

The lack of consistency between T_{FH} and HLA-peptide binding experimental observations for vaccine antibody traits is interesting and may have important relevance. It is possible that the observed discrepancies may be the result of multiple independent associations. For example, the PT-*HLA-DRB1*-TCR- T_{FH} association may represent one mechanistic pathway, whereas the PT-*HLA-DRB3* and DT-*HLA-DRB1* binding differences may represent other sets. However, it is important to realise the limitations of the peptide binding studies before confirming this model. Although popular in studies trying to understand the functional effects of HLA allele diversity, HLA-peptide binding analyses are well recognised to only represent a part of the complex interactions involved in APC-T-cell co-stimulation and furthermore methodologies for generating and interpreting data have not been validated or agreed upon. Observations of differences in quality of antigen-specific binding profiles for *HLA-DRB3* alleles with PT, and indeed *HLA-DRB1* alleles with DT, may be encouraging, but an absence of differences with *HLA-DRB1* alleles for PT peptides is not necessarily indicative of a lack of association. Immunodominant effects of peptides are known to occur with HLA-peptide binding values of up to 1000nM^{321} which are far lower than those used as cut-offs in the experiments described in this chapter. Furthermore, other factors such as TCR expression or binding and expression

levels of the HLA alleles are not explicitly considered in such binding experiments. Moreover, for the HLA-DRB3 associations reported with PT response, the limited number of reported HLA-DRB3 alleles and the obvious similarity in binding profiles between some groups of alleles may make a false-positive association finding more likely. Nevertheless, it is interesting to note that those phenotypes with the most evidence of effects seem to occur through different mechanisms. That is, the most striking differences in T_{FH} and TCR V gene expression associated with PT were observed in the absence of clear binding differences between candidate alleles, whereas, conversely, differences in DT-derived binding appeared to occur between groups of alleles where no differences in T_{FH} proportions or TCR expression levels were observed. It is important to note, however, that based on the absence of GWAS studies of diphtheria susceptibility it has not been possible to confirm that DT antibodies are as important in protecting against clinical episodes of diphtheria as PT antibodies appear to be for WC. Thus it is not possible to rule out the possibility of other immune measures being more important for DT protection which are in turn more correlated with different levels of T_{FH} which are indeed true functional correlates of protection.

The novel observation of a positive genetic correlation between PT antibody response and *TRBV29-1* gene expression supports the importance of this gene in WC protection. Little work has been published characterising this particular gene amongst other TCR V genes although levels have been found to be overabundant in the T-cell repertoires of individuals with chronic idiopathic neutropenia³²² and psoriasis³²³. Since the TCR V gene expression data was generated from whole blood it is not possible to determine whether expression levels are representative

of particular T-cell subsets although it would be assumed that they would be driven by T_H or T_{Reg} cells that make up the abundance of lymphocytes in the blood. Given this observation, it would be tempting to speculate that a combination of DRB1-233Thr carriage and PT exposure is more likely to induce clonal expansion of *TRBV29-1* expressing T-cells. These mechanisms could be more likely to induce protection against WC following PT challenge, possibly because of improved antibody production and maturation pathways in the presence of such molecules. The availability of the large repository of cells from known vaccinee donors used in this thesis offers future opportunities to test this hypothesis. The antigen-specific T-cells may be isolated before undergoing TCR sequencing to determine the V gene usage of the clonal proliferation in the same individuals stratified by HLA types.

Overall, the immunological mechanisms tested for correlated association with vaccine response in this chapter only represent a fraction of those hypothesised to play a role. For example, in the case of class I loci both CD8 TCR and NK cell binding and interactions are involved in differential downstream interaction and disease development, alongside complex allele-specific expression effects³²⁴. Furthermore, newer insights into HLA-E and NK cell involvement³²⁵ highlight the growing complexity of simultaneous multi-cellular multi-receptor interactions. It is likely that equally complex interactions exist for the class II alleles. For example, class II expression levels are known to vary by allele although the full extent of this variation is still poorly characterised³²⁶. *DR4* linked HLA-DRB1*07 and *09 alleles have been reported have lower basal expression³²⁷. This is particularly notable in that these are the alleles associated with increased DT response and therefore it is quite possible that increased peptide binding affinity

may be critical to compensate for the expression levels to mediate differential effects. The observation of the rs73727916 association with PT response only serves to further highlight the potential complexity of these associations. Untangling these observations will ultimately require an improved comprehensive understanding of differential expression levels and multi-receptor interaction profiles with multiple antigens which are resources that are not currently available for class II genes across the MHC but are clearly sorely needed.

Altogether, the findings presented in this chapter highlight a multitude of functional mechanisms that are likely be responsible for driving, or at least correlating with, the observed HLA-vaccine trait associations. The findings support a model for the immunological mechanisms underpinning response to PT, and thereby a reduction in odds of self-reported WC, whereby the carriage of specific MHC variants (most likely but not definitively linked to *HLA-DRB1*) act to increase levels of PT-specific antibodies through an HLA-peptide-TCR mediated complex that is associated with an increased durable circulation of PT-specific T_{FH}. These mechanisms appear to occur in the absence of a measurable difference in HLA-peptide binding differences although this may be a result of the detected TCR differences that have not been accounted for in the assays used in this thesis and other potential deficiencies in the assays used rather than a true representation of mechanism. In contrast, there is little evidence for equivalent mechanisms being responsible for differential DT response whereas there are rather striking differences in the binding profiles of DT-derived peptides based on grouping of alleles. Understanding whether these observations are truly relevant or of specific functional importance will require further replication and independent functional interrogation using, for example, multiple samples

collected prospectively and longitudinally from individuals vaccinated with PT and DT. Providing evidence of such pathways and differences in mechanism would have important relevance for understanding how to improve vaccination testing and development strategies. Using the examples described above it would be important to understand how to increase levels of circulating antigen-specific T_{FH} in all PT vaccinees, whereas for DT, the emphasis may instead be on finding and delivering immunodominant peptides that bind tightly to all HLA-DRB1 alleles rather than just those encoded by DRB1-4Arg.

Chapter 6

6 Discussion

As described in chapter one of this thesis, vaccines have been hugely successful in reducing the prevalence of vaccine preventable diseases worldwide. However, the timing of their development and deployment over the last few centuries, often without the requirement for rational drug trialling or deep understanding of mechanisms has made any detailed understanding of mechanisms of protection associated with their use elusive. These problems are compounded by the fact that it is now unethical to withhold any routine vaccinations due to be administered to children and undertaking any longitudinal studies involving comprehensive sampling for immunophenotyping of children is equally challenging from an ethical point of view where excessive bloodletting may be detrimental to young infants. Such longitudinal studies are also inherently costly making them unattractive funding opportunities regardless of their potential for understanding vaccine correlates of protection.

Therefore, novel methods are required to improve our understanding of the mechanisms underlying variable vaccine efficacy, protection or failure. Arguably it is important to understand these processes first in those vaccines that are considered ‘simple’ before moving towards those vaccines that are sorely needed, but are inherently tackling more ‘complex’ organisms such as malaria, HIV or tuberculosis.

In this thesis, I have developed a series of approaches that improve our insight into potentially protective mechanisms applied to some vaccine antigens delivered to every infant world-wide, focussing in particular on the vaccine targeting pertussis.

In chapter three I used novel genetic association approaches in well characterised cohorts of African infants to identify strong, consistent associations between variants within the class II MHC and responses to five vaccine antigens. Four of these associations are novel whilst the observed HBsAg associations, although expected, overlapped poorly at the genotype level with those observed previously in Asian populations. Nevertheless, the consistent associations observed between HLA variants and HBsAg response serves as a useful positive control validating the underlying analyses and stressing the importance of understanding HLA diversity in a population dependent manner.

In chapter four I present a resource for characterising HLA allelic diversity across Africa to a resolution and comprehensiveness, which to our knowledge, has not been previously reported. The resource highlights the diversity and differentiation observed across multiple genes but also demonstrates how class II genes appear to be less differentiated across Africa than may have initially been expected.

Together these differences are likely to have helped facilitate the replication of associations across the included populations. I subsequently used this resource to undertake imputation of HLA alleles in the African infant dataset to a high resolution and accuracy and I further used this data to fine-map the HLA association signals demonstrating the relevance of amino acid residues, complete functional allelic units and SNPs in these association signals. In keeping with multiple other HLA-disease associations I found an important role for *HLA-DRB1* variation, but I also found that variation in other *HLA-DRB* paralogues may be equally important alongside *HLA-DQ* alleles as well as, importantly, the relatively poorly understood *HLA-DPB1* locus. The absence of genetic association signals with some traits is perhaps as interesting as the presence of others. A key example of this phenomenon exists for TT where there was little evidence of a strong genetic contribution to variability. This may be due to population differences or deficiencies in causative allele tagging and ultimately, a deeper understanding will only come from further studies including additional populations. Furthermore, the dominant inheritance pattern observed only with *HLA-DRB1* variants is highly relevant for vaccine design and further validates the use of these approaches for understanding and tackling modern vaccine design issues.

In chapter five I used these association findings to demonstrate a striking genetic correlation between PT antibody response and WC susceptibility. This provided strong evidence in support of PT antibody response being the main correlate of protection against WC, highlighting the importance of human genetic variation in influencing the effectiveness of vaccination. I then harnessed these findings to demonstrate that these genetic correlations traced through multiple immunological mechanisms including differential circulating T_{FH} proportions and expression

levels of *TRBV29-1* TCR V genes. Furthermore, using HLA-peptide binding studies I found evidence that the *HLA-DRB3* signal with PT antibody response and the *HLA-DRB1* signal with DT response may be mediated through improved breadth or affinity of binding. The T_{FH} linked *HLA-DRB1* PT signal may occur independently of the strength of individual allele-peptide binding, although important limitations in HLA-peptide binding assays must be considered before substantiating such claims.

My work highlights, yet again, the critical importance of the MHC region in differential responses against infectious agents and shows how, in the context of known antigens, these signals can be used to understand mechanism even in the absence of identifying true causative alleles. The results show that individuals do respond differently to vaccines based on genetic variation and that this variation in turn influences the efficacy of vaccination. I demonstrate the utility of genetic correlations in proposing and potentially understanding immunological correlates of protection against infection and highlighting the importance of T_{FH} cells in responses against some, but not all, vaccine antigens. Furthermore, the results from the analyses suggest that different mechanisms may be more critically important for some responses compared to others (e.g. TCR-T_{FH} for PT and HLA-peptide binding for DT). The work presented in this thesis therefore offers a range of approaches and resources to begin to understand vaccine responses in children to a level not previously achieved. It is hoped that these methods may be used as a framework towards undertaking more comprehensive studies, perhaps offering a degree of cost-effectiveness, in equivalent cohorts of infants enabling the use of more cross-sectional study design rather than more life-long longitudinal studies, to help build a comprehensive picture of how such vaccines mediate their effect

and protection. Only in light of such studies can mechanisms be understood to help facilitate the development of universally effective vaccines against multiple infections and diseases.

Limitations and future directions

There are several limitations to the work presented in this thesis that are important to highlight that may direct future research in the area. Firstly, although the sample sizes used in the genetic association study were adequate to identify strong, consistent HLA signals, it is possible that the study remains underpowered to identify association signals outside of the HLA locus. Furthermore, given the complexity and diversity of HLA association signals it would be of interest to test for deviations from purely additive models of association, assessing further for dominant or recessive models or, indeed, interactions between variants. Such analyses would only be possible with larger sample sizes. The work presented in this thesis presents only the first phase in our proposed efforts to characterise and identify the genetic variants determining differential response to vaccination. Sample and data collation and vaccine response measurements are planned or ongoing for a variety of other populations including those from Kenya, The Gambia and Bangladesh. Furthermore, it may even be possible to use the Luminex-based methods to measure vaccine responses in large cohort studies such as UK Biobank (including 500,000 adults in the UK) where knowledge of timing of vaccines may not be necessary. Using such approaches to increase sample size will not only allow a better understanding of the complexity of the signals occurring across the MHC, but they will also increase the power to

identify small effect signals occurring outside of this polymorphic region. The work presented here does however provide a validated framework for undertaking similar analyses in populations of other ancestries facilitating further fine-mapping initiatives both inside and outside of the MHC using the pLMM approach. Furthermore the HLA reference panel presented in this thesis will be expanded upon to not only improve our understanding of HLA diversity in other African populations but also those in other populations of Asian and other, previously poorly characterised, descent including those from multiple countries across the Pacific.

The inherent differences in methodologies applied to each of the sample collections used in this thesis does mean that attributing potentially interesting observations of heterogeneity in genetic association signals to genetic or non-genetic causes is challenging. Arguably, the most interesting example in this thesis is the observation of the apparently absent PT genetic association in the Soweto population being due to differences in vaccine preparation. Confirming this finding will be difficult in the absence of strict harmonisation of study protocols. Achieving such harmonisation across multiple developing country sites is highly challenging from a practical perspective but is something that our collaborative network is intent on achieving to allow appropriate comparisons between sites. The observation of a difference in genetic association signals for PT response potentially being attributable to differences in vaccine type (aP being given to the Soweto group compared to wP in the others) is highly relevant given the noted differences in long term effectiveness between these different preparations of the pertussis vaccine. These observations have led to plans to specifically account for variation in human genetics in a set of independent studies

aiming to understand the mechanistic reasons for these differences in vaccinated individuals followed longitudinally, hopefully both in African and European populations.

Finally, the functional work I have presented correlating the genetic signals of observed with PT response with effects through T_{FH} , TCR gene and peptide binding processes only represent a fraction of the potential mechanisms that may be responsible for the observed differences. Given this diversity and complexity of proposed molecular and cellular mechanisms associated with HLA, it was not possible to comprehensively assess such systems here especially in relation to all of the genetic signals observed. The presence of some genetic association signals such as rs73727916 with PT responses provides evidence in support of the potential role of expression differences that may serve to modulate these more direct mechanisms. It will therefore be important to characterise mechanisms such as differential allelic expression, and proportions of other T-cell subsets including T_{Reg} cells and B-cells in future studies. Other immunoglobulin subsets or subtypes will also be increasingly important to characterise especially for those agents where mucosal invasion are important, such as pertussis. The binding experiments presented in this thesis did not consider the potential for post-translational modification of peptide sequences, such as citrullination, which are known to be relevant in certain autoimmune conditions. Ultimately a full understanding of these mechanisms will require large, well characterised populations of vaccinated individuals with detailed immuno-phenotype data available to understand the underlying complexities. The longitudinal aP/wP study described above may offer several opportunities to help improve our understanding of the relationship of such mechanisms to each other in the context of human genetic variation.

However, given the suggestion of multiple distinct molecular mechanisms being associated with differential efficacy or effectiveness of different vaccine types it will be important to replicate similar longitudinal studies for other subtypes including diphtheria and hepatitis B. For this latter pathogen in particular, there is a wealth of data already available describing the effects of human genetic variation on progression and sequelae of chronic hepatitis B infection. Our group is in the advanced stages of securing access to the relevant GWAS datasets of multi-ethnic multi-phenotype analyses relating to HBV with a view to applying similar methods to those employed in this thesis to sample sizes in the region of tens of thousands of individuals. Understanding such relationships between genetics, vaccine response and subsequent disease state will no doubt be important to help support improvements in existing vaccine design based on how these genetic findings may play a role in reducing future incidence of disease. Thus, in line with the centuries of research already undertaken in vaccinology, it is clear that despite the exciting discoveries and opportunities that the findings in this thesis offer, there is still much work to be done before rationalisation of vaccine design based on human genetics becomes a reality.

References

- 1 Plotkin, S. A. Vaccines: past, present and future. *Nature medicine* **11**, S5-11, doi:10.1038/nm1209 (2005).
- 2 Plotkin, S. A. History of Vaccine Development Introduction. *History of Vaccine Development*, 1-4, doi:10.1007/978-1-4419-1339-5_1 (2011).
- 3 Jenner, E. *An inquiry into the causes and effects of the variolæ vaccinæ, a disease discovered in some of the Western Counties of England ... known by the name of the cow pox. [With plates.]*. (Printed for the author, 1798).
- 4 Calmette, A., Guérin, C. & Breton, M. Contribution a l'etude de la tuberculose experimental du cobaye (infection et etasis de vaccination par la voie digestive). *Ann. Inst. Pasteur Paris* **21**, 401-416 (1907).
- 5 Pasteur, L. De L'attenuation du virus cholera des poules. *C. R. Acad. Sci. Paris* **991**, 673-680 (1880).
- 6 Ehrlich, P., Dale, H. H. S. G. B. E., Himmelweit, F. & Marquardt, M. *The Collected Papers of Paul Ehrlich ... Including a complete bibliography. Compiled and edited by F. Himmelweit ... with the assistance of Martha Marquardt. Under the editorial direction of Sir Henry Dale, etc. [With a portrait.]*. (London & New York; Vienna printed : Pergamon Press, 1956-).
- 7 Mechnikov, I. Immunity in infective diseases. By Il'ia Il'ich Mechnikov, 1905. *Reviews of infectious diseases* **10**, 223-227 (1988).
- 8 Collier, L. H. The development of a stable smallpox vaccine. *The Journal of hygiene* **53**, 76-101 (1955).
- 9 Weller, T. H., Robbins, F. C. & Enders, J. F. Cultivation of poliomyelitis virus in cultures of human foreskin and embryonic tissues. *Proceedings of the Society for Experimental Biology and Medicine. Society for Experimental Biology and Medicine* **72**, 153-155 (1949).
- 10 McAleer, W. J. *et al.* Human hepatitis B vaccine from recombinant yeast. *Nature* **307**, 178-180 (1984).
- 11 Koff, W. C. *et al.* Accelerating next-generation vaccine development for global disease prevention. *Science* **340**, 1232910, doi:10.1126/science.1232910 (2013).
- 12 Bland, J. & Clements, J. Protecting the world's children: the story of WHO's immunization programme. *World health forum* **19**, 162-173 (1998).
- 13 Keja, K., Chan, C., Hayden, G. & Henderson, R. H. Expanded programme on immunization. *World health statistics quarterly. Rapport trimestriel de statistiques sanitaires mondiales* **41**, 59-63 (1988).

- 14 Plotkin, S. History of vaccination. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 12283-12287, doi:10.1073/pnas.1400472111 (2014).
- 15 Plotkin, S. A. Correlates of protection induced by vaccination. *Clinical and vaccine immunology : CVI* **17**, 1055-1065, doi:10.1128/CVI.00131-10 (2010).
- 16 Shim, E. & Galvani, A. P. Distinguishing vaccine efficacy and effectiveness. *Vaccine* **30**, 6700-6705, doi:10.1016/j.vaccine.2012.08.045 (2012).
- 17 Mentzer, A. J., O'Connor, D., Pollard, A. J. & Hill, A. V. Searching for the human genetic factors standing in the way of universally effective vaccines. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **370**, doi:10.1098/rstb.2014.0341 (2015).
- 18 Minor, P. D. Live attenuated vaccines: Historical successes and current challenges. *Virology* **479-480**, 379-392, doi:10.1016/j.virol.2015.03.032 (2015).
- 19 Belongia, E. A. & Naleway, A. L. Smallpox vaccine: the good, the bad, and the ugly. *Clinical medicine & research* **1**, 87-92 (2003).
- 20 Lalvani, A. & Sridhar, S. BCG vaccination: 90 years on and still so much to learn. *Thorax* **65**, 1036-1038, doi:10.1136/thx.2010.140996 (2010).
- 21 Koprowski, H., Jervis, G. A. & Norton, T. W. Immune responses in human volunteers upon oral administration of a rodent-adapted strain of poliomyelitis virus. *American journal of hygiene* **55**, 108-124 (1952).
- 22 Sabin, A. B., Hennessen, W. A. & Winsser, J. Studies on variants of poliomyelitis virus. I. Experimental segregation and properties of avirulent variants of three immunologic types. *The Journal of experimental medicine* **99**, 551-576 (1954).
- 23 Enders, J. F. & Peebles, T. C. Propagation in tissue cultures of cytopathogenic agents from patients with measles. *Proceedings of the Society for Experimental Biology and Medicine. Society for Experimental Biology and Medicine* **86**, 277-286 (1954).
- 24 Katz, S. L. John F. Enders and measles virus vaccine--a reminiscence. *Current topics in microbiology and immunology* **329**, 3-11 (2009).
- 25 Sabin, A. B. Vaccination against poliomyelitis in economically underdeveloped countries. *Bulletin of the World Health Organization* **58**, 141-157 (1980).
- 26 Hinman, A. R. *et al.* The case for global eradication of poliomyelitis. *Bulletin of the World Health Organization* **65**, 835-840 (1987).
- 27 Uzicanin, A. & Zimmerman, L. Field effectiveness of live attenuated measles-containing vaccines: a review of published literature. *The Journal*

- of infectious diseases* **204 Suppl 1**, S133-148, doi:10.1093/infdis/jir102 (2011).
- 28 Roy, A. *et al.* Effect of BCG vaccination against Mycobacterium tuberculosis infection in children: systematic review and meta-analysis. *Bmj* **349**, g4643, doi:10.1136/bmj.g4643 (2014).
 - 29 Wilkie, M. E. & McShane, H. TB vaccine development: where are we and why is it so difficult? *Thorax* **70**, 299-301, doi:10.1136/thoraxjnl-2014-205202 (2015).
 - 30 Kimmey, J. M. *et al.* Unique role for ATG5 in neutrophil-mediated immunopathology during M. tuberculosis infection. *Nature* **528**, 565-569, doi:10.1038/nature16451 (2015).
 - 31 Nunes-Alves, C. *et al.* In search of a new paradigm for protective immunity to TB. *Nature reviews. Microbiology* **12**, 289-299, doi:10.1038/nrmicro3230 (2014).
 - 32 Madsen, T. Vaccination against whooping cough. *Jama* **101**, 187-188 (1933).
 - 33 Shapiro-Shapin, C. G. Pearl Kendrick, Grace Eldering, and the pertussis vaccine. *Emerging infectious diseases* **16**, 1273-1278, doi:10.3201/eid1608.100288 (2010).
 - 34 Marks, H. M. The Kendrick-Eldering-(Frost) pertussis vaccine field trial. *Journal of the Royal Society of Medicine* **100**, 242-247, doi:10.1177/014107680710000516 (2007).
 - 35 PREVENTION of whooping-cough by vaccination; a Medical Research Council investigation. *British medical journal* **1**, 1463-1471 (1951).
 - 36 Juskewitch, J. E., Tapia, C. J. & Windebank, A. J. Lessons from the Salk polio vaccine: methods for and risks of rapid translation. *Clinical and translational science* **3**, 182-185, doi:10.1111/j.1752-8062.2010.00205.x (2010).
 - 37 Francis, T., Jr. & Korns, R. F. Evaluation of 1954 field trial of poliomyelitis vaccine: synopsis of summary report. *The American journal of the medical sciences* **229**, 603-612 (1955).
 - 38 Meldrum, M. "A calculated risk": the Salk polio vaccine field trials of 1954. *Bmj* **317**, 1233-1236 (1998).
 - 39 A.T., G. & B.E., H. Diphtheria toxoid as an immunising agent. *Br J Exp Pathol* **4**, 283-288 (1923).
 - 40 Moloney, P. J. The Preparation and Testing of Diphtheria Toxoid (Anatoxine-Ramon). *American journal of public health* **16**, 1208-1210 (1926).
 - 41 CDC. (ed Centres for disease control and prevention) (2015).

- 42 Nanthavong, N. *et al.* Diphtheria in Lao PDR: Insufficient Coverage or Ineffective Vaccine? *PloS one* **10**, e0121749, doi:10.1371/journal.pone.0121749 (2015).
- 43 Bisgard, K. M. *et al.* Diphtheria toxoid vaccine effectiveness: A case-control study in Russia. *The Journal of infectious diseases* **181 Suppl 1**, S184-187, doi:10.1086/315562 (2000).
- 44 Kendrick, P. L. Use of Alum-Treated Pertussis Vaccine, and of Alum-Precipitated Combined Pertussis Vaccine and Diphtheria Toxoid, for Active Immunization. *American journal of public health and the nation's health* **32**, 615-626 (1942).
- 45 Editorial: Vaccination against whooping cough. *British medical journal* **3**, 539-540 (1974).
- 46 Zhang, L., Prietsch, S. O., Axelsson, I. & Halperin, S. A. Acellular vaccines for preventing whooping cough in children. *The Cochrane database of systematic reviews*, CD001478, doi:10.1002/14651858.CD001478.pub6 (2014).
- 47 Plotkin, S. A. The pertussis problem. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* **58**, 830-833, doi:10.1093/cid/cit934 (2014).
- 48 Pinto, M. V. & Merkel, T. J. Pertussis disease and transmission and host responses: insights from the baboon model of pertussis. *The Journal of infection* **74 Suppl 1**, S114-S119, doi:10.1016/S0163-4453(17)30201-3 (2017).
- 49 Bancroft, T. *et al.* Th1 versus Th2 T cell polarization by whole-cell and acellular childhood pertussis vaccines persists upon re-immunization in adolescence and adulthood. *Cellular immunology* **304-305**, 35-43, doi:10.1016/j.cellimm.2016.05.002 (2016).
- 50 Plotkin, S. A. Complex correlates of protection after vaccination. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* **56**, 1458-1465, doi:10.1093/cid/cit048 (2013).
- 51 Public Health England. Laboratory confirmed pertussis in England: data to end-February 2017. (2017).
- 52 Macleod, C. M., Hodges, R. G., Heidelberger, M. & Bernhard, W. G. Prevention of Pneumococcal Pneumonia by Immunization with Specific Capsular Polysaccharides. *The Journal of experimental medicine* **82**, 445-465 (1945).
- 53 Austrian, R. Pneumococcal polysaccharide vaccines. *Reviews of infectious diseases* **11 Suppl 3**, S598-602 (1989).
- 54 Anderson, P., Peter, G., Johnston, R. B., Jr., Wetterlow, L. H. & Smith, D. H. Immunization of humans with polyribophosphate, the capsular antigen

- of *Haemophilus influenzae*, type b. *The Journal of clinical investigation* **51**, 39-44, doi:10.1172/JCI106794 (1972).
- 55 Vipond, C., Care, R. & Feavers, I. M. History of meningococcal vaccines and their serological correlates of protection. *Vaccine* **30 Suppl 2**, B10-17, doi:10.1016/j.vaccine.2011.12.060 (2012).
 - 56 Kelly, D. F., Moxon, E. R. & Pollard, A. J. *Haemophilus influenzae* type b conjugate vaccines. *Immunology* **113**, 163-174, doi:10.1111/j.1365-2567.2004.01971.x (2004).
 - 57 Avery, O. T. & Goebel, W. F. Chemo-Immunological Studies on Conjugated Carbohydrate-Proteins : Ii. Immunological Specificity of Synthetic Sugar-Protein Antigens. *The Journal of experimental medicine* **50**, 533-550 (1929).
 - 58 Martin, N. G., Sadarangani, M., Pollard, A. J. & Goldacre, M. J. Hospital admission rates for meningitis and septicaemia caused by *Haemophilus influenzae*, *Neisseria meningitidis*, and *Streptococcus pneumoniae* in children in England over five decades: a population-based observational study. *The Lancet. Infectious diseases* **14**, 397-405, doi:10.1016/S1473-3099(14)70027-1 (2014).
 - 59 Elliman, D. & Bedford, H. Safety and efficacy of combination vaccines. *Bmj* **326**, 995-996, doi:10.1136/bmj.326.7397.995 (2003).
 - 60 Hepatitis B vaccines. *British medical journal* **280**, 203-204 (1980).
 - 61 Szmunes, W. *et al.* Hepatitis B vaccine: demonstration of efficacy in a controlled clinical trial in a high-risk population in the United States. *The New England journal of medicine* **303**, 833-841, doi:10.1056/NEJM198010093031501 (1980).
 - 62 Valenzuela, P., Medina, A., Rutter, W. J., Ammerer, G. & Hall, B. D. Synthesis and assembly of hepatitis B virus surface antigen particles in yeast. *Nature* **298**, 347-350 (1982).
 - 63 Jilg, W. *et al.* Clinical evaluation of a recombinant hepatitis B vaccine. *Lancet* **2**, 1174-1175 (1984).
 - 64 Schoub, B. D. *et al.* Integration of hepatitis B vaccination into rural African primary health care programmes. *Bmj* **302**, 313-316 (1991).
 - 65 Chang, M. H. *et al.* Universal hepatitis B vaccination in Taiwan and the incidence of hepatocellular carcinoma in children. Taiwan Childhood Hepatoma Study Group. *The New England journal of medicine* **336**, 1855-1859, doi:10.1056/NEJM199706263362602 (1997).
 - 66 Hsu, H. M. *et al.* Efficacy of a mass hepatitis B vaccination program in Taiwan. Studies on 3464 infants of hepatitis B surface antigen-carrier mothers. *Jama* **260**, 2231-2235 (1988).

- 67 Hepatitis B vaccine in the expanded programme of immunisation: The Gambian experience. The Gambia Hepatitis Study Group. *Lancet* **1**, 1057-1059 (1989).
- 68 Mendy, M. *et al.* Observational study of vaccine efficacy 24 years after the start of hepatitis B vaccination in two Gambian villages: no need for a booster dose. *PloS one* **8**, e58029, doi:10.1371/journal.pone.0058029 (2013).
- 69 Peto, T. J. *et al.* Efficacy and effectiveness of infant vaccination against chronic hepatitis B in the Gambia Hepatitis Intervention Study (1986-90) and in the nationwide immunisation program. *BMC infectious diseases* **14**, 7, doi:10.1186/1471-2334-14-7 (2014).
- 70 Pollard, A. J. Hepatitis B vaccination. *Bmj* **335**, 950, doi:10.1136/bmj.39315.677396.BE (2007).
- 71 Torjesen, I. UK adds hepatitis B to infant vaccination schedule. *Bmj* **358**, j3357, doi:10.1136/bmj.j3357 (2017).
- 72 Pollard, A. J. *et al.* Human microbial challenge: the ultimate animal model. *The Lancet. Infectious diseases* **12**, 903-905, doi:10.1016/S1473-3099(12)70292-X (2012).
- 73 Janeway, C. *Immunobiology : the immune system in health and disease*. 6th ed. edn, (Garland Science, 2005).
- 74 Saadi, S., Wrenshall, L. E. & Platt, J. L. Regional manifestations and control of the immune system. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology* **16**, 849-856, doi:10.1096/fj.01-0690hyp (2002).
- 75 Akira, S. & Takeda, K. Toll-like receptor signalling. *Nature reviews. Immunology* **4**, 499-511, doi:10.1038/nri1391 (2004).
- 76 Netea, M. G. *et al.* Trained immunity: A program of innate immune memory in health and disease. *Science* **352**, aaf1098, doi:10.1126/science.aaf1098 (2016).
- 77 Levy, O. Innate immunity of the newborn: basic mechanisms and clinical correlates. *Nature reviews. Immunology* **7**, 379-390, doi:10.1038/nri2075 (2007).
- 78 Lacy, P. & Stow, J. L. Cytokine release from innate immune cells: association with diverse membrane trafficking pathways. *Blood* **118**, 9-18, doi:10.1182/blood-2010-08-265892 (2011).
- 79 Worbs, T., Hammerschmidt, S. I. & Forster, R. Dendritic cell migration in health and disease. *Nature reviews. Immunology* **17**, 30-48, doi:10.1038/nri.2016.116 (2017).

- 80 Market, E. & Papavasiliou, F. N. V(D)J recombination and the evolution of the adaptive immune system. *PLoS biology* **1**, E16, doi:10.1371/journal.pbio.0000016 (2003).
- 81 Klein, L., Kyewski, B., Allen, P. M. & Hogquist, K. A. Positive and negative selection of the T cell repertoire: what thymocytes see (and don't see). *Nature reviews. Immunology* **14**, 377-391, doi:10.1038/nri3667 (2014).
- 82 Davey, M. S. *et al.* Clonal selection in the human Vdelta1 T cell repertoire indicates gammadelta TCR-dependent adaptive immune surveillance. *Nature communications* **8**, 14760, doi:10.1038/ncomms14760 (2017).
- 83 Wakim, L. M. & Bevan, M. J. Cross-dressed dendritic cells drive memory CD8+ T-cell activation after viral infection. *Nature* **471**, 629-632, doi:10.1038/nature09863 (2011).
- 84 Kakaradov, B. *et al.* Early transcriptional and epigenetic regulation of CD8+ T cell differentiation revealed by single-cell RNA sequencing. *Nature immunology*, doi:10.1038/ni.3688 (2017).
- 85 Farber, D. L., Yudanin, N. A. & Restifo, N. P. Human memory T cells: generation, compartmentalization and homeostasis. *Nature reviews. Immunology* **14**, 24-35, doi:10.1038/nri3567 (2014).
- 86 Kaech, S. M. & Cui, W. Transcriptional control of effector and memory CD8+ T cell differentiation. *Nature reviews. Immunology* **12**, 749-761, doi:10.1038/nri3307 (2012).
- 87 Mueller, S. N. & Mackay, L. K. Tissue-resident memory T cells: local specialists in immune defence. *Nature reviews. Immunology* **16**, 79-89, doi:10.1038/nri.2015.3 (2016).
- 88 Gattinoni, L. *et al.* A human memory T cell subset with stem cell-like properties. *Nature medicine* **17**, 1290-1297, doi:10.1038/nm.2446 (2011).
- 89 Salehi, Z. *et al.* Differential Frequency of CD8+ T Cell Subsets in Multiple Sclerosis Patients with Various Clinical Patterns. *PloS one* **11**, e0159565, doi:10.1371/journal.pone.0159565 (2016).
- 90 Kapsenberg, M. L. Dendritic-cell control of pathogen-driven T-cell polarization. *Nature reviews. Immunology* **3**, 984-993, doi:10.1038/nri1246 (2003).
- 91 DuPage, M. & Bluestone, J. A. Harnessing the plasticity of CD4(+) T cells to treat immune-mediated disease. *Nature reviews. Immunology* **16**, 149-163, doi:10.1038/nri.2015.18 (2016).
- 92 Sehra, S. *et al.* TH9 cells are required for tissue mast cell accumulation during allergic inflammation. *The Journal of allergy and clinical immunology* **136**, 433-440 e431, doi:10.1016/j.jaci.2015.01.021 (2015).

- 93 Tesmer, L. A., Lundy, S. K., Sarkar, S. & Fox, D. A. Th17 cells in human disease. *Immunological reviews* **223**, 87-113, doi:10.1111/j.1600-065X.2008.00628.x (2008).
- 94 Josefowicz, S. Z., Lu, L. F. & Rudensky, A. Y. Regulatory T cells: mechanisms of differentiation and function. *Annual review of immunology* **30**, 531-564, doi:10.1146/annurev.immunol.25.022106.141623 (2012).
- 95 Li, M. O. & Rudensky, A. Y. T cell receptor signalling in the control of regulatory T cell differentiation and function. *Nature reviews. Immunology* **16**, 220-233, doi:10.1038/nri.2016.26 (2016).
- 96 Maloy, K. J. & Powrie, F. Regulatory T cells in the control of immune pathology. *Nature immunology* **2**, 816-822, doi:10.1038/ni0901-816 (2001).
- 97 Greiff, V. *et al.* A bioinformatic framework for immune repertoire diversity profiling enables detection of immunological status. *Genome medicine* **7**, 49, doi:10.1186/s13073-015-0169-8 (2015).
- 98 Nutt, S. L., Hodgkin, P. D., Tarlinton, D. M. & Corcoran, L. M. The generation of antibody-secreting plasma cells. *Nature reviews. Immunology* **15**, 160-171, doi:10.1038/nri3795 (2015).
- 99 Cyster, J. G. B cell follicles and antigen encounters of the third kind. *Nature immunology* **11**, 989-996, doi:10.1038/ni.1946 (2010).
- 100 Heesters, B. A., Myers, R. C. & Carroll, M. C. Follicular dendritic cells: dynamic antigen libraries. *Nature reviews. Immunology* **14**, 495-504, doi:10.1038/nri3689 (2014).
- 101 Schroeder, H. W., Jr. & Cavacini, L. Structure and function of immunoglobulins. *The Journal of allergy and clinical immunology* **125**, S41-52, doi:10.1016/j.jaci.2009.09.046 (2010).
- 102 Pillai, S. & Cariappa, A. The follicular versus marginal zone B lymphocyte cell fate decision. *Nature reviews. Immunology* **9**, 767-777, doi:10.1038/nri2656 (2009).
- 103 Cerutti, A., Cols, M. & Puga, I. Marginal zone B cells: virtues of innate-like antibody-producing lymphocytes. *Nature reviews. Immunology* **13**, 118-132, doi:10.1038/nri3383 (2013).
- 104 Kurosaki, T., Kometani, K. & Ise, W. Memory B cells. *Nature reviews. Immunology* **15**, 149-159, doi:10.1038/nri3802 (2015).
- 105 Crotty, S. T follicular helper cell differentiation, function, and roles in disease. *Immunity* **41**, 529-542, doi:10.1016/j.immuni.2014.10.004 (2014).
- 106 Crotty, S. A brief history of T cell help to B cells. *Nature reviews. Immunology* **15**, 185-189, doi:10.1038/nri3803 (2015).

- 107 Crotty, S. Follicular helper CD4 T cells (TFH). *Annual review of immunology* **29**, 621-663, doi:10.1146/annurev-immunol-031210-101400 (2011).
- 108 Croft, M. Control of immunity by the TNFR-related molecule OX40 (CD134). *Annual review of immunology* **28**, 57-78, doi:10.1146/annurev-immunol-030409-101243 (2010).
- 109 Dan, J. M. *et al.* A Cytokine-Independent Approach To Identify Antigen-Specific Human Germinal Center T Follicular Helper Cells and Rare Antigen-Specific CD4+ T Cells in Blood. *Journal of immunology* **197**, 983-993, doi:10.4049/jimmunol.1600318 (2016).
- 110 Havenar-Daughton, C. *et al.* Cytokine-Independent Detection of Antigen-Specific Germinal Center T Follicular Helper Cells in Immunized Nonhuman Primates Using a Live Cell Activation-Induced Marker Technique. *Journal of immunology* **197**, 994-1002, doi:10.4049/jimmunol.1600320 (2016).
- 111 Farooq, F. *et al.* Circulating follicular T helper cells and cytokine profile in humans following vaccination with the rVSV-ZEBOV Ebola vaccine. *Scientific reports* **6**, 27944, doi:10.1038/srep27944 (2016).
- 112 Plotkin, S. *Vaccines*. 6 edn, (Elsevier, 2013).
- 113 Hill, A. V. Pre-erythrocytic malaria vaccines: towards greater efficacy. *Nature reviews. Immunology* **6**, 21-32, doi:10.1038/nri1746 (2006).
- 114 Faldella, G. *et al.* The preterm infant's antibody response to a combined diphtheria, tetanus, acellular pertussis and hepatitis B vaccine. *Vaccine* **16**, 1646-1649 (1998).
- 115 Poland, G. A., Ovsyannikova, I. G. & Jacobson, R. M. Vaccine immunogenetics: bedside to bench to population. *Vaccine* **26**, 6183-6188, doi:10.1016/j.vaccine.2008.06.057 (2008).
- 116 Ekstrom, N. *et al.* Kinetics and avidity of antibodies evoked by heptavalent pneumococcal conjugate vaccines PncCRM and PncOMPC in the Finnish Otitis Media Vaccine Trial. *Infection and immunity* **73**, 369-377, doi:10.1128/IAI.73.1.369-377.2005 (2005).
- 117 Aaby, P. *et al.* Is early measles vaccination better than later measles vaccination? *Transactions of the Royal Society of Tropical Medicine and Hygiene* **109**, 16-28, doi:10.1093/trstmh/tru174 (2015).
- 118 Siegrist, C. A. & Aspinall, R. B-cell responses to vaccination at the extremes of age. *Nature reviews. Immunology* **9**, 185-194, doi:10.1038/nri2508 (2009).
- 119 Klein, S. L., Marriott, I. & Fish, E. N. Sex-based differences in immune function and responses to vaccination. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **109**, 9-15, doi:10.1093/trstmh/tru167 (2015).

- 120 Voysey, M., Pollard, A. J., Perera, R. & Fanshawe, T. R. Assessing sex-differences and the effect of timing of vaccination on immunogenicity, reactogenicity and efficacy of vaccines in young children: study protocol for an individual participant data meta-analysis of randomised controlled trials. *BMJ open* **6**, e011680, doi:10.1136/bmjopen-2016-011680 (2016).
- 121 Hsu, L. C. *et al.* Ethnic differences in immune responses to hepatitis B vaccine. *American journal of epidemiology* **143**, 718-724 (1996).
- 122 Mendel, G. L. Versuche uber panzenhybriden. *Verhandlungen des Naturforschenden Veriens* **4**, 3-47 (1865).
- 123 Garrod, A. E. About Alkaptonuria. *Medico-chirurgical transactions* **85**, 69-78 (1902).
- 124 Watson, J. D. & Crick, F. H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**, 737-738 (1953).
- 125 Mullis, K. *et al.* Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harbor symposia on quantitative biology* **51 Pt 1**, 263-273 (1986).
- 126 Botstein, D., White, R. L., Skolnick, M. & Davis, R. W. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American journal of human genetics* **32**, 314-331 (1980).
- 127 Weber, J. L. & May, P. E. Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *American journal of human genetics* **44**, 388-396 (1989).
- 128 Abecasis, G. R., Cherny, S. S., Cookson, W. O. & Cardon, L. R. Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. *Nature genetics* **30**, 97-101, doi:10.1038/ng786 (2002).
- 129 Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516-1517 (1996).
- 130 Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921, doi:10.1038/35057062 (2001).
- 131 Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304-1351, doi:10.1126/science.1058040 (2001).
- 132 McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nature genetics* **48**, 1279-1283, doi:10.1038/ng.3643 (2016).
- 133 Nielsen, R., Paul, J. S., Albrechtsen, A. & Song, Y. S. Genotype and SNP calling from next-generation sequencing data. *Nature reviews. Genetics* **12**, 443-451, doi:10.1038/nrg2986 (2011).
- 134 Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-291, doi:10.1038/nature19057 (2016).

- 135 Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J. & Lander, E. S. High-resolution haplotype structure in the human genome. *Nature genetics* **29**, 229-232, doi:10.1038/ng1001-229 (2001).
- 136 Gabriel, S. B. *et al.* The structure of haplotype blocks in the human genome. *Science* **296**, 2225-2229, doi:10.1126/science.1069424 (2002).
- 137 Reich, D. E. *et al.* Human genome sequence variation and the influence of gene history, mutation and recombination. *Nature genetics* **32**, 135-142, doi:10.1038/ng947 (2002).
- 138 Slatkin, M. Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. *Nature reviews. Genetics* **9**, 477-485, doi:10.1038/nrg2361 (2008).
- 139 Sachidanandam, R. *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928-933, doi:10.1038/35057149 (2001).
- 140 Genomes Project, C. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65, doi:10.1038/nature11632 (2012).
- 141 Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74, doi:10.1038/nature15393 (2015).
- 142 Gurdasani, D. *et al.* The African Genome Variation Project shapes medical genetics in Africa. *Nature*, doi:10.1038/nature13997 (2014).
- 143 Teo, Y. Y., Small, K. S. & Kwiatkowski, D. P. Methodological challenges of genome-wide association analysis in Africa. *Nature reviews. Genetics* **11**, 149-160, doi:10.1038/nrg2731 (2010).
- 144 de Lange, K. M. *et al.* Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nature genetics* **49**, 256-261, doi:10.1038/ng.3760 (2017).
- 145 Ellinghaus, D. *et al.* Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci. *Nature genetics* **48**, 510-518, doi:10.1038/ng.3528 (2016).
- 146 Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376-381, doi:10.1038/nature12873 (2014).
- 147 Fuchsberger, C. *et al.* The genetic architecture of type 2 diabetes. *Nature* **536**, 41-47, doi:10.1038/nature18642 (2016).
- 148 Marouli, E. *et al.* Rare and low-frequency coding variants alter human adult height. *Nature* **542**, 186-190, doi:10.1038/nature21039 (2017).
- 149 Astle, W. J. *et al.* The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell* **167**, 1415-1429 e1419, doi:10.1016/j.cell.2016.10.042 (2016).

- 150 Sham, P. C. & Purcell, S. M. Statistical power and significance testing in large-scale genetic studies. *Nature reviews. Genetics* **15**, 335-346, doi:10.1038/nrg3706 (2014).
- 151 Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature genetics* **39**, 906-913, doi:10.1038/ng2088 (2007).
- 152 Altshuler, D., Daly, M. J. & Lander, E. S. Genetic mapping in human disease. *Science* **322**, 881-888, doi:10.1126/science.1156409 (2008).
- 153 Emilsson, V. *et al.* Genetics of gene expression and its effect on disease. *Nature* **452**, 423-428, doi:10.1038/nature06758 (2008).
- 154 Grant, S. F. *et al.* Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. *Nature genetics* **38**, 320-323, doi:10.1038/ng1732 (2006).
- 155 Genovese, G. *et al.* Association of trypanolytic ApoL1 variants with kidney disease in African Americans. *Science* **329**, 841-845, doi:10.1126/science.1193032 (2010).
- 156 Chu, A. Y. *et al.* Multiethnic genome-wide meta-analysis of ectopic fat depots identifies loci associated with adipocyte development and differentiation. *Nature genetics* **49**, 125-130, doi:10.1038/ng.3738 (2017).
- 157 Trowsdale, J. & Knight, J. C. Major histocompatibility complex genomics and human disease. *Annual review of genomics and human genetics* **14**, 301-323, doi:10.1146/annurev-genom-091212-153455 (2013).
- 158 Horton, R. *et al.* Gene map of the extended human MHC. *Nature reviews. Genetics* **5**, 889-899, doi:10.1038/nrg1489 (2004).
- 159 Singal, D. P. & Blajchman, M. A. Histocompatibility (HL-A) antigens, lymphocytotoxic antibodies and tissue antibodies in patients with diabetes mellitus. *Diabetes* **22**, 429-432 (1973).
- 160 Sollid, L. M. *et al.* Evidence for a primary association of celiac disease to a particular HLA-DQ alpha/beta heterodimer. *The Journal of experimental medicine* **169**, 345-350 (1989).
- 161 Stastny, P. Association of the B-cell alloantigen DRw4 with rheumatoid arthritis. *The New England journal of medicine* **298**, 869-871, doi:10.1056/NEJM197804202981602 (1978).
- 162 Parham, P. & Ohta, T. Population biology of antigen presentation by MHC class I molecules. *Science* **272**, 67-74 (1996).
- 163 Gragert, L., Madbouly, A., Freeman, J. & Maiers, M. Six-locus high resolution HLA haplotype frequencies derived from mixed-resolution DNA typing for the entire US donor registry. *Human immunology* **74**, 1313-1320, doi:10.1016/j.humimm.2013.06.025 (2013).

- 164 International, H. I. V. C. S. *et al.* The major genetic determinants of HIV-1 control affect HLA class I peptide presentation. *Science* **330**, 1551-1557, doi:10.1126/science.1195271 (2010).
- 165 Goyette, P. *et al.* High-density mapping of the MHC identifies a shared role for HLA-DRB1*01:03 in inflammatory bowel diseases and heterozygous advantage in ulcerative colitis. *Nature genetics* **47**, 172-179, doi:10.1038/ng.3176 (2015).
- 166 the International Multiple Sclerosis Genetics, C. Class II HLA interactions modulate genetic risk for multiple sclerosis. *Nature genetics* **47**, 1107-1113, doi:10.1038/ng.3395
<http://www.nature.com/ng/journal/v47/n10/abs/ng.3395.html#supplementary-information> (2015).
- 167 Sveinbjornsson, G. *et al.* HLA class II sequence variants influence tuberculosis risk in populations of European ancestry. *Nature genetics* **48**, 318-322, doi:10.1038/ng.3498 (2016).
- 168 Dunstan, S. J. *et al.* Variation at HLA-DRB1 is associated with resistance to enteric fever. *Nature genetics* **46**, 1333-1336, doi:10.1038/ng.3143 (2014).
- 169 Patsopoulos, N. A. *et al.* Fine-mapping the genetic association of the major histocompatibility complex in multiple sclerosis: HLA and non-HLA effects. *PLoS genetics* **9**, e1003926, doi:10.1371/journal.pgen.1003926 (2013).
- 170 Raychaudhuri, S. *et al.* Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. *Nature genetics* **44**, 291-296, doi:10.1038/ng.1076 (2012).
- 171 Marsh, S. G. *et al.* Nomenclature for factors of the HLA system, 2010. *Tissue antigens* **75**, 291-455, doi:10.1111/j.1399-0039.2010.01466.x (2010).
- 172 Dilthey, A. *et al.* Multi-population classical HLA type imputation. *PLoS computational biology* **9**, e1002877, doi:10.1371/journal.pcbi.1002877 (2013).
- 173 Scally, S. W. *et al.* A molecular basis for the association of the HLA-DRB1 locus, citrullination, and rheumatoid arthritis. *The Journal of experimental medicine* **210**, 2569-2582, doi:10.1084/jem.20131241 (2013).
- 174 van Heemst, J. *et al.* Crossreactivity to vinculin and microbes provides a molecular basis for HLA-based protection against rheumatoid arthritis. *Nature communications* **6**, 6681, doi:10.1038/ncomms7681 (2015).
- 175 Miyadera, H., Ohashi, J., Lernmark, A., Kitamura, T. & Tokunaga, K. Cell-surface MHC density profiling reveals instability of autoimmunity-

- associated HLA. *The Journal of clinical investigation* **125**, 275-291, doi:10.1172/JCI74961 (2015).
- 176 Kulkarni, S. *et al.* Differential microRNA regulation of HLA-C expression and its association with HIV control. *Nature* **472**, 495-498, doi:10.1038/nature09914 (2011).
 - 177 Craven, D. E. *et al.* Nonresponsiveness to hepatitis B vaccine in health care workers. Results of revaccination and genetic typings. *Annals of internal medicine* **105**, 356-360 (1986).
 - 178 Hennig, B. J. & Hall, A. J. Host genetic factors in hepatitis B infection, liver cancer and vaccination response: a review with a focus on Africa. *The Science of the total environment* **423**, 202-209, doi:10.1016/j.scitotenv.2010.09.036 (2012).
 - 179 Alper, C. A. *et al.* Genetic prediction of nonresponse to hepatitis B vaccine. *The New England journal of medicine* **321**, 708-712, doi:10.1056/NEJM198909143211103 (1989).
 - 180 Davila, S. *et al.* New genetic associations detected in a host response study to hepatitis B vaccine. *Genes and immunity* **11**, 232-238, doi:10.1038/gene.2010.1 (2010).
 - 181 De Silvestri, A. *et al.* Family study of non-responsiveness to hepatitis B vaccine confirms the importance of HLA class III C4A locus. *Genes and immunity* **2**, 367-372, doi:10.1038/sj.gene.6363792 (2001).
 - 182 Desombere, I., Willems, A. & Leroux-Roels, G. Response to hepatitis B vaccine: multiple HLA genes are involved. *Tissue antigens* **51**, 593-604 (1998).
 - 183 Hohler, T. *et al.* The influence of major histocompatibility complex class II genes and T-cell Vbeta repertoire on response to immunization with HBsAg. *Human immunology* **59**, 212-218 (1998).
 - 184 McDermott, A. B., Zuckerman, J. N., Sabin, C. A., Marsh, S. G. & Madrigal, J. A. Contribution of human leukocyte antigens to the antibody response to hepatitis B vaccination. *Tissue antigens* **50**, 8-14 (1997).
 - 185 Wu, T. W. *et al.* Responses to booster hepatitis B vaccination are significantly correlated with genotypes of human leukocyte antigen (HLA)-DPB1 in neonatally vaccinated adolescents. *Human genetics* **132**, 1131-1139, doi:10.1007/s00439-013-1320-5 (2013).
 - 186 Sekar, A. *et al.* Schizophrenia risk from complex variation of complement component 4. *Nature* **530**, 177-183, doi:10.1038/nature16549 (2016).
 - 187 Desombere, I., Cao, T., Gijbels, Y. & Leroux-Roels, G. Non-responsiveness to hepatitis B surface antigen vaccines is not caused by defective antigen presentation or a lack of B7 co-stimulation. *Clinical and experimental immunology* **140**, 126-137, doi:10.1111/j.1365-2249.2004.02749.x (2005).

- 188 Kruger, A. *et al.* Hepatitis B surface antigen presentation and HLA-DRB1*- lessons from twins and peptide binding studies. *Clinical and experimental immunology* **140**, 325-332, doi:10.1111/j.1365-2249.2005.02765.x (2005).
- 189 Salazar, M. *et al.* Normal HBsAg presentation and T-cell defect in the immune response of nonresponders. *Immunogenetics* **41**, 366-374 (1995).
- 190 Milich, D. R. & Leroux-Roels, G. G. Immunogenetics of the response to HBsAg vaccination. *Autoimmunity reviews* **2**, 248-257 (2003).
- 191 Xu, B. *et al.* Minimal association of alleles of human leukocyte antigen class II gene and long-term antibody response to hepatitis B vaccine vaccinated during infancy. *Vaccine*, doi:10.1016/j.vaccine.2017.03.021 (2017).
- 192 Marchant, A. *et al.* Predominant influence of environmental determinants on the persistence and avidity maturation of antibody responses to vaccines in infants. *The Journal of infectious diseases* **193**, 1598-1605, doi:10.1086/503775 (2006).
- 193 Png, E. *et al.* A genome-wide association study of hepatitis B vaccine response in an Indonesian population reveals multiple independent risk variants in the HLA region. *Human molecular genetics* **20**, 3893-3898, doi:10.1093/hmg/ddr302 (2011).
- 194 Pan, L. *et al.* A genome-wide association study identifies polymorphisms in the HLA-DR region associated with non-response to hepatitis B vaccination in Chinese Han populations. *Human molecular genetics* **23**, 2210-2219, doi:10.1093/hmg/ddt586 (2014).
- 195 Newport, M. J. *et al.* Genetic regulation of immune responses to vaccines in early life. *Genes and immunity* **5**, 122-129, doi:10.1038/sj.gene.6364051 (2004).
- 196 Hohler, T. *et al.* Differential genetic determination of immune responsiveness to hepatitis B surface antigen and to hepatitis A virus: a vaccination study in twins. *Lancet* **360**, 991-995, doi:10.1016/S0140-6736(02)11083-X (2002).
- 197 Hennig, B. J. *et al.* Host genetic factors and vaccine-induced immunity to hepatitis B virus infection. *PloS one* **3**, e1898, doi:10.1371/journal.pone.0001898 (2008).
- 198 Haralambieva, I. H. *et al.* Associations between single nucleotide polymorphisms and haplotypes in cytokine and cytokine receptor genes and immunity to measles vaccination. *Vaccine* **29**, 7883-7895, doi:10.1016/j.vaccine.2011.08.083 (2011).
- 199 Haralambieva, I. H. *et al.* Genetic polymorphisms in host antiviral genes: associations with humoral and cellular immunity to measles vaccine. *Vaccine* **29**, 8988-8997, doi:10.1016/j.vaccine.2011.09.043 (2011).

- 200 Ovsyannikova, I. G. *et al.* Effects of vitamin A and D receptor gene polymorphisms/haplotypes on immune responses to measles vaccine. *Pharmacogenetics and genomics* **22**, 20-31, doi:10.1097/FPC.0b013e32834df186 (2012).
- 201 Ovsyannikova, I. G. *et al.* The role of polymorphisms in Toll-like receptors and their associated intracellular signaling genes in measles vaccine immunity. *Human genetics* **130**, 547-561, doi:10.1007/s00439-011-0977-x (2011).
- 202 Ovsyannikova, I. G., Pankratz, V. S., Vierkant, R. A., Jacobson, R. M. & Poland, G. A. Human leukocyte antigen haplotypes in the genetic control of immune response to measles-mumps-rubella vaccine. *The Journal of infectious diseases* **193**, 655-663, doi:10.1086/500144 (2006).
- 203 Jacobson, R. M., Ovsyannikova, I. G., Vierkant, R. A., Pankratz, V. S. & Poland, G. A. Human leukocyte antigen associations with humoral and cellular immunity following a second dose of measles-containing vaccine: persistence, dampening, and extinction of associations found after a first dose. *Vaccine* **29**, 7982-7991, doi:10.1016/j.vaccine.2011.08.060 (2011).
- 204 Haralambieva, I. H. *et al.* Genome-wide associations of CD46 and IFI44L genetic variants with neutralizing antibody response to measles vaccine. *Human genetics* **136**, 421-435, doi:10.1007/s00439-017-1768-9 (2017).
- 205 Oldstone, M. B. *et al.* Measles virus infection in a transgenic model: virus-induced immunosuppression and central nervous system disease. *Cell* **98**, 629-640 (1999).
- 206 Zilliox, M. J., Parmigiani, G. & Griffin, D. E. Gene expression patterns in dendritic cells infected with measles virus compared with other pathogens. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 3363-3368, doi:10.1073/pnas.0511345103 (2006).
- 207 Feenstra, B. *et al.* Common variants associated with general and MMR vaccine-related febrile seizures. *Nature genetics* **46**, 1274-1282, doi:10.1038/ng.3129 (2014).
- 208 Clifford, H. D. *et al.* Polymorphisms in key innate immune genes and their effects on measles vaccine responses and vaccine failure in children from Mozambique. *Vaccine* **30**, 6180-6185, doi:10.1016/j.vaccine.2012.07.063 (2012).
- 209 Ovsyannikova, I. G., Larrabee, B. R., Schaid, D. J. & Poland, G. A. Immunoglobulin GM and KM genes and measles vaccine-induced humoral immunity. *Vaccine*, doi:10.1016/j.vaccine.2017.02.046 (2017).
- 210 Ovsyannikova, I. G. *et al.* A large population-based association study between HLA and KIR genotypes and measles vaccine antibody responses. *PloS one* **12**, e0171261, doi:10.1371/journal.pone.0171261 (2017).

- 211 Ladhani, S. N. *et al.* Association between single-nucleotide polymorphisms in Mal/TIRAP and interleukin-10 genes and susceptibility to invasive haemophilus influenzae serotype b infection in immunized children. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* **51**, 761-767, doi:10.1086/656236 (2010).
- 212 McMahon, G., Ring, S. M., Davey-Smith, G. & Timpson, N. J. Genome-wide association study identifies SNPs in the MHC class II loci that are associated with self-reported history of whooping cough. *Human molecular genetics* **24**, 5930-5939, doi:10.1093/hmg/ddv293 (2015).
- 213 Webb, E. L. *et al.* Effect of single-dose anthelmintic treatment during pregnancy on an infant's response to immunisation and on susceptibility to infectious diseases in infancy: a randomised, double-blind, placebo-controlled trial. *Lancet* **377**, 52-62, doi:10.1016/S0140-6736(10)61457-2 (2011).
- 214 Nash, S. *et al.* The impact of prenatal exposure to parasitic infections and to anthelmintic treatment on antibody responses to routine immunisations given in infancy: Secondary analysis of a randomised controlled trial. *PLoS neglected tropical diseases* **11**, e0005213, doi:10.1371/journal.pntd.0005213 (2017).
- 215 Nunes, M. C. *et al.* Duration of Infant Protection Against Influenza Illness Conferred by Maternal Immunization: Secondary Analysis of a Randomized Clinical Trial. *JAMA pediatrics* **170**, 840-847, doi:10.1001/jamapediatrics.2016.0921 (2016).
- 216 Bliss, C. M. *et al.* Viral Vector Malaria Vaccines Induce High-Level T Cell and Antibody Responses in West African Children and Infants. *Molecular therapy : the journal of the American Society of Gene Therapy* **25**, 547-559, doi:10.1016/j.ymthe.2016.11.003 (2017).
- 217 Smits, G. P., van Gageldonk, P. G., Schouls, L. M., van der Klis, F. R. & Berbers, G. A. Development of a bead-based multiplex immunoassay for simultaneous quantitative detection of IgG serum antibodies against measles, mumps, rubella, and varicella-zoster virus. *Clinical and vaccine immunology : CVI* **19**, 396-400, doi:10.1128/CI.05537-11 (2012).
- 218 van Gageldonk, P. G., van Schaijk, F. G., van der Klis, F. R. & Berbers, G. A. Development and validation of a multiplex immunoassay for the simultaneous determination of serum antibodies to Bordetella pertussis, diphtheria and tetanus. *Journal of immunological methods* **335**, 79-89, doi:10.1016/j.jim.2008.02.018 (2008).
- 219 Hijnen, M., van Gageldonk, P. G., Berbers, G. A., van Woerkom, T. & Mooi, F. R. The Bordetella pertussis virulence factor P.69 pertactin retains its immunological properties after overproduction in Escherichia coli. *Protein expression and purification* **41**, 106-112, doi:10.1016/j.pep.2005.01.014 (2005).

- 220 Waterboer, T. *et al.* Multiplex human papillomavirus serology based on in situ-purified glutathione s-transferase fusion proteins. *Clinical chemistry* **51**, 1845-1853, doi:10.1373/clinchem.2005.052381 (2005).
- 221 Chen, R. *et al.* Presence of DNA of human papillomavirus 16 but no other types in tumor-free tonsillar tissue. *Journal of clinical microbiology* **43**, 1408-1410, doi:10.1128/JCM.43.3.1408-1410.2005 (2005).
- 222 Illumina. Genome-Wide DNA Analysis BeadChips. (2013).
- 223 Purcell, S., Cherny, S. S. & Sham, P. C. Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics* **19**, 149-150 (2003).
- 224 Anderson, C. A. *et al.* Data quality control in genetic case-control association studies. *Nature protocols* **5**, 1564-1573, doi:10.1038/nprot.2010.116 (2010).
- 225 Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics* **81**, 559-575, doi:10.1086/519795 (2007).
- 226 Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS genetics* **2**, e190, doi:10.1371/journal.pgen.0020190 (2006).
- 227 Rayner, N. W. & McCarthy, M. I. in *American Society of Human Genetics* (Montreal, Canada, 2011).
- 228 Weir, B. S., Anderson, A. D. & Hepler, A. B. Genetic relatedness analysis: modern data and new challenges. *Nature reviews. Genetics* **7**, 771-780, doi:10.1038/nrg1960 (2006).
- 229 Price, A. L. *et al.* Long-range LD can confound genome scans in admixed populations. *American journal of human genetics* **83**, 132-135; author reply 135-139, doi:10.1016/j.ajhg.2008.06.005 (2008).
- 230 Hardy, G. H. Mendelian Proportions in a Mixed Population. *Science* **28**, 49-50, doi:10.1126/science.28.706.49 (1908).
- 231 International Multiple Sclerosis Genetics, C. *et al.* Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature* **476**, 214-219, doi:10.1038/nature10251 (2011).
- 232 Price, A. L., Zaitlen, N. A., Reich, D. & Patterson, N. New approaches to population stratification in genome-wide association studies. *Nature reviews. Genetics* **11**, 459-463, doi:10.1038/nrg2813 (2010).
- 233 Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature genetics* **44**, 955-959, doi:10.1038/ng.2354 (2012).

- 234 Delaneau, O., Marchini, J. & Zagury, J. F. A linear complexity phasing method for thousands of genomes. *Nature methods* **9**, 179-181, doi:10.1038/nmeth.1785 (2011).
- 235 Loh, P. R., Palamara, P. F. & Price, A. L. Fast and accurate long-range phasing in a UK Biobank cohort. *Nature genetics* **48**, 811-816, doi:10.1038/ng.3571 (2016).
- 236 Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nature reviews. Genetics* **11**, 499-511, doi:10.1038/nrg2796 (2010).
- 237 Howie, B., Marchini, J. & Stephens, M. Genotype imputation with thousands of genomes. *G3* **1**, 457-470, doi:10.1534/g3.111.001198 (2011).
- 238 Pirinen, M., Donnelly, P. & Spencer, C. C. Including known covariates can reduce power to detect genetic effects in case-control studies. *Nature genetics* **44**, 848-851, doi:10.1038/ng.2346 (2012).
- 239 Balding, D. J. A tutorial on statistical methods for population association studies. *Nature reviews. Genetics* **7**, 781-791, doi:10.1038/nrg1916 (2006).
- 240 Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M. & Price, A. L. Advantages and pitfalls in the application of mixed-model association methods. *Nature genetics* **46**, 100-106, doi:10.1038/ng.2876 (2014).
- 241 Pirinen, M., Donnelly, P. & Spencer, C. C. A. Efficient Computation with a Linear Mixed Model on Large-Scale Data Sets with Applications to Genetic Studies. *Ann Appl Stat* **7**, 369-390, doi:10.1214/12-AOAS586 (2013).
- 242 Kang, H. M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nature genetics* **42**, 348-U110, doi:10.1038/ng.548 (2010).
- 243 Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nature genetics* **44**, 821-824, doi:10.1038/ng.2310 (2012).
- 244 Zhou, X. & Stephens, M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature methods* **11**, 407-409, doi:10.1038/nmeth.2848 (2014).
- 245 Clayton, D. G. *et al.* Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nature genetics* **37**, 1243-1246, doi:10.1038/ng1653 (2005).
- 246 Yin, T., Cook, D. & Lawrence, M. ggbio: an R package for extending the grammar of graphics for genomic data. *Genome biology* **13**, R77, doi:10.1186/gb-2012-13-8-r77 (2012).

- 247 Evangelou, E. & Ioannidis, J. P. Meta-analysis methods for genome-wide association studies and beyond. *Nature reviews. Genetics* **14**, 379-389, doi:10.1038/nrg3472 (2013).
- 248 Han, B. & Eskin, E. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *American journal of human genetics* **88**, 586-598, doi:10.1016/j.ajhg.2011.04.014 (2011).
- 249 Cochran, W. G. Some Methods for Strengthening the Common X2 Tests. *Biometrics* **10**, 417-451, doi:Doi 10.2307/3001616 (1954).
- 250 Higgins, J. P. & Thompson, S. G. Quantifying heterogeneity in a meta-analysis. *Statistics in medicine* **21**, 1539-1558, doi:10.1002/sim.1186 (2002).
- 251 Morris, A. P. Transethnic meta-analysis of genomewide association studies. *Genetic epidemiology* **35**, 809-822, doi:10.1002/gepi.20630 (2011).
- 252 Replication, D. I. G. *et al.* Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nature genetics* **46**, 234-244, doi:10.1038/ng.2897 (2014).
- 253 Hammer, C. *et al.* Amino Acid Variation in HLA Class II Proteins Is a Major Determinant of Humoral Response to Common Viruses. *American journal of human genetics* **97**, 738-743, doi:10.1016/j.ajhg.2015.09.008 (2015).
- 254 Wood, A. R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature genetics* **46**, 1173-1186, doi:10.1038/ng.3097 (2014).
- 255 Band, G. *et al.* Imputation-based meta-analysis of severe malaria in three African populations. *PLoS genetics* **9**, e1003509, doi:10.1371/journal.pgen.1003509 (2013).
- 256 Griffin, S. M. *et al.* Application of salivary antibody immunoassays for the detection of incident infections with Norwalk virus in a group of volunteers. *Journal of immunological methods* **424**, 53-63, doi:10.1016/j.jim.2015.05.001 (2015).
- 257 Jones, C. E. *et al.* Maternal HIV infection and antibody responses against vaccine-preventable diseases in uninfected infants. *Jama* **305**, 576-584, doi:10.1001/jama.2011.100 (2011).
- 258 Ovsyannikova, I. G., Pankratz, V. S., Vierkant, R. A., Jacobson, R. M. & Poland, G. A. Consistency of HLA associations between two independent measles vaccine cohorts: a replication study. *Vaccine* **30**, 2146-2152, doi:10.1016/j.vaccine.2012.01.038 (2012).

- 259 Tian, Y., Schreiber, R. & Kunzelmann, K. Anoctamins are a family of Ca²⁺-activated Cl⁻ channels. *Journal of cell science* **125**, 4991-4998, doi:10.1242/jcs.109553 (2012).
- 260 Hammer, C. *et al.* A Coding Variant of ANO10, Affecting Volume Regulation of Macrophages, Is Associated with Borrelia Seropositivity. *Molecular medicine* **21**, 26-37, doi:10.2119/molmed.2014.00219 (2015).
- 261 Schwab, A., Fabian, A., Hanley, P. J. & Stock, C. Role of ion channels and transporters in cell migration. *Physiological reviews* **92**, 1865-1913, doi:10.1152/physrev.00018.2011 (2012).
- 262 Maniero, C. *et al.* Role of ANO4 in regulation of aldosterone secretion in the zona glomerulosa of the human adrenal gland. *Lancet* **385 Suppl 1**, S62, doi:10.1016/S0140-6736(15)60377-4 (2015).
- 263 Engler, R. J. *et al.* Half- vs full-dose trivalent inactivated influenza vaccine (2004-2005): age, dose, and sex effects on immune responses. *Archives of internal medicine* **168**, 2405-2414, doi:10.1001/archinternmed.2008.513 (2008).
- 264 Fang, J. W., Lai, C. L., Chung, H. T., Wu, P. C. & Lau, J. Y. Female children respond to recombinant hepatitis B vaccine with a higher titre than male. *Journal of tropical pediatrics* **40**, 104-107 (1994).
- 265 Hess, G., Hingst, V., Cseke, J., Bock, H. L. & Clemens, R. Influence of vaccination schedules and host factors on antibody response following hepatitis B vaccination. *European journal of clinical microbiology & infectious diseases : official publication of the European Society of Clinical Microbiology* **11**, 334-340 (1992).
- 266 Klein, S. L., Jedlicka, A. & Pekosz, A. The Xs and Y of immune responses to viral vaccines. *The Lancet. Infectious diseases* **10**, 338-349, doi:10.1016/S1473-3099(10)70049-9 (2010).
- 267 Tian, C. *et al.* Genome-wide association and HLA region fine-mapping studies identify susceptibility loci for multiple common infections. *bioRxiv*, doi:10.1101/073056 (2016).
- 268 Gonzalez-Galarza, F. F. *et al.* Allele frequency net 2015 update: new features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations. *Nucleic acids research* **43**, D784-788, doi:10.1093/nar/gku1166 (2015).
- 269 Mayor, N. P. *et al.* HLA Typing for the Next Generation. *PloS one* **10**, e0127153, doi:10.1371/journal.pone.0127153 (2015).
- 270 Schofl, G. *et al.* 2.7 million samples genotyped for HLA by next generation sequencing: lessons learned. *BMC genomics* **18**, 161, doi:10.1186/s12864-017-3575-z (2017).

- 271 Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nature genetics* **44**, 369-375, S361-363, doi:10.1038/ng.2213 (2012).
- 272 Mack, S. J. *et al.* Common and well-documented HLA alleles: 2012 update to the CWD catalogue. *Tissue antigens* **81**, 194-203, doi:10.1111/tan.12093 (2013).
- 273 Cereb, N., Kim, H. R., Ryu, J. & Yang, S. Y. Advances in DNA sequencing technologies for high resolution HLA typing. *Human immunology* **76**, 923-927, doi:10.1016/j.humimm.2015.09.015 (2015).
- 274 Marsh, S. G. *Nomenclature for Factors of the HLA System*, <<http://hla.alleles.org/nomenclature/naming.html>> (2017).
- 275 Robinson, J. *et al.* The IPD and IMGT/HLA database: allele variant databases. *Nucleic acids research* **43**, D423-431, doi:10.1093/nar/gku1161 (2015).
- 276 Holsinger, K. E. & Weir, B. S. Genetics in geographically structured populations: defining, estimating and interpreting F(ST). *Nature reviews. Genetics* **10**, 639-650, doi:10.1038/nrg2611 (2009).
- 277 Wright, S. The genetical structure of populations. *Annals of eugenics* **15**, 323-354 (1951).
- 278 Weir, B. S. & Cockerham, C. C. Estimating F-Statistics for the Analysis of Population Structure. *Evolution; international journal of organic evolution* **38**, 1358-1370, doi:10.1111/j.1558-5646.1984.tb05657.x (1984).
- 279 Nei, M. Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences of the United States of America* **70**, 3321-3323 (1973).
- 280 Jost, L. G(ST) and its relatives do not measure differentiation. *Molecular ecology* **17**, 4015-4026 (2008).
- 281 Gerlach, G., Jueterbock, A., Kraemer, P., Deppermann, J. & Harmand, P. Calculations of population differentiation based on GST and D: forget GST but not all of statistics! *Molecular ecology* **19**, 3845-3852, doi:10.1111/j.1365-294X.2010.04784.x (2010).
- 282 Keenan, K., McGinnity, P., Cross, T. F., Crozier, W. W. & Prodohl, P. A. diveRsity: An R package for the estimation and exploration of population genetics parameters and their associated errors. *Methods Ecol Evol* **4**, 782-788, doi:10.1111/2041-210X.12067 (2013).
- 283 Henn, B. M. *et al.* Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 5154-5162, doi:10.1073/pnas.1017511108 (2011).

- 284 Szpiech, Z. A., Jakobsson, M. & Rosenberg, N. A. ADZE: a rarefaction approach for counting alleles private to combinations of populations. *Bioinformatics* **24**, 2498-2504, doi:10.1093/bioinformatics/btn478 (2008).
- 285 Jia, X. *et al.* Imputing amino acid polymorphisms in human leukocyte antigens. *PloS one* **8**, e64683, doi:10.1371/journal.pone.0064683 (2013).
- 286 Zheng, X. *et al.* HIBAG--HLA genotype imputation with attribute bagging. *The pharmacogenomics journal* **14**, 192-200, doi:10.1038/tpj.2013.18 (2014).
- 287 Sutton, V. R., Kienzle, B. K. & Knowles, R. W. An altered splice site is found in the DRB4 gene that is not expressed in HLA-DR7,Dw11 individuals. *Immunogenetics* **29**, 317-322 (1989).
- 288 Yates, A. *et al.* Ensembl 2016. *Nucleic acids research* **44**, D710-716, doi:10.1093/nar/gkv1157 (2016).
- 289 Zhou, X. *et al.* Epigenomic annotation of genetic variants using the Roadmap Epigenome Browser. *Nature biotechnology* **33**, 345-346, doi:10.1038/nbt.3158 (2015).
- 290 Hill, A. V. *et al.* Common west African HLA antigens are associated with protection from severe malaria. *Nature* **352**, 595-600, doi:10.1038/352595a0 (1991).
- 291 McLaren, P. J. *et al.* Association study of common genetic variants and HIV-1 acquisition in 6,300 infected cases and 7,200 controls. *PLoS pathogens* **9**, e1003515, doi:10.1371/journal.ppat.1003515 (2013).
- 292 Tshabalala, M., Mellet, J. & Pepper, M. S. Human Leukocyte Antigen Diversity: A Southern African Perspective. *Journal of immunology research* **2015**, 746151, doi:10.1155/2015/746151 (2015).
- 293 Gourraud, P. A. *et al.* HLA diversity in the 1000 genomes dataset. *PloS one* **9**, e97282, doi:10.1371/journal.pone.0097282 (2014).
- 294 Zhou, F. *et al.* Deep sequencing of the MHC region in the Chinese population contributes to studies of complex disease. *Nature genetics* **48**, 740-746, doi:10.1038/ng.3576 (2016).
- 295 Vince, N. *et al.* HLA-C Level Is Regulated by a Polymorphic Oct1 Binding Site in the HLA-C Promoter Region. *American journal of human genetics* **99**, 1353-1358, doi:10.1016/j.ajhg.2016.09.023 (2016).
- 296 Jain, M. *et al.* Nanopore sequencing and assembly of a human genome with ultra-long reads. *bioRxiv*, doi:10.1101/128835 (2017).
- 297 Greenberg, D. A., Hodge, S. E. & Rotter, J. I. Evidence for recessive and against dominant inheritance at the HLA-"linked" locus in coeliac disease. *American journal of human genetics* **34**, 263-277 (1982).

- 298 MacDonald, M. J., Gottschall, J., Hunter, J. B. & Winter, K. L. HLA-DR4 in insulin-dependent diabetic parents and their diabetic offspring: a clue to dominant inheritance. *Proceedings of the National Academy of Sciences of the United States of America* **83**, 7049-7053 (1986).
- 299 Neefjes, J., Jongsma, M. L., Paul, P. & Bakke, O. Towards a systems understanding of MHC class I and MHC class II antigen presentation. *Nature reviews. Immunology* **11**, 823-836, doi:10.1038/nri3084 (2011).
- 300 Diethelm-Okita, B. M., Okita, D. K., Banaszak, L. & Conti-Fine, B. M. Universal epitopes for human CD4+ cells on tetanus and diphtheria toxins. *The Journal of infectious diseases* **181**, 1001-1009, doi:10.1086/315324 (2000).
- 301 Ergonul, O. *et al.* An unexpected tetanus case. *The Lancet. Infectious diseases* **16**, 746-752, doi:10.1016/S1473-3099(16)00075-X (2016).
- 302 Miyadera, H. & Tokunaga, K. Associations of human leukocyte antigens with autoimmune diseases: challenges in identifying the mechanism. *Journal of human genetics* **60**, 697-702, doi:10.1038/jhg.2015.100 (2015).
- 303 McMichael, A. J. & Jones, E. Y. Genetics. First-class control of HIV-1. *Science* **330**, 1488-1490, doi:10.1126/science.1200035 (2010).
- 304 Stern, L. J. & Calvo-Calle, J. M. HLA-DR: molecular insights and vaccine design. *Current pharmaceutical design* **15**, 3249-3261 (2009).
- 305 Liu, X. S. & Mardis, E. R. Applications of Immunogenomics to Cancer. *Cell* **168**, 600-612, doi:10.1016/j.cell.2017.01.014 (2017).
- 306 Amirthalingam, G., Gupta, S. & Campbell, H. Pertussis immunisation and control in England and Wales, 1957 to 2012: a historical review. *Eurosurveillance* **18**, 19-27 (2013).
- 307 Lindestam Arlehamn, C. S. *et al.* Memory T cells in latent Mycobacterium tuberculosis infection are directed against three antigenic islands and largely contained in a CXCR3+CCR6+ Th1 subset. *PLoS pathogens* **9**, e1003130, doi:10.1371/journal.ppat.1003130 (2013).
- 308 Weiskopf, D. *et al.* Comprehensive analysis of dengue virus-specific responses supports an HLA-linked protective role for CD8+ T cells. *Proceedings of the National Academy of Sciences of the United States of America* **110**, E2046-2053, doi:10.1073/pnas.1305227110 (2013).
- 309 Frazier, A. *et al.* Allergy-associated T cell epitope repertoires are surprisingly diverse and include non-IgE reactive antigens. *The World Allergy Organization journal* **7**, 26, doi:10.1186/1939-4551-7-26 (2014).
- 310 Vita, R. *et al.* The immune epitope database (IEDB) 3.0. *Nucleic acids research* **43**, D405-412, doi:10.1093/nar/gku938 (2015).
- 311 Karosiene, E. *et al.* NetMHCIIpan-3.0, a common pan-specific MHC class II prediction method including all three human MHC class II isotypes,

- HLA-DR, HLA-DP and HLA-DQ. *Immunogenetics* **65**, 711-724, doi:10.1007/s00251-013-0720-y (2013).
- 312 Wang, P. *et al.* Peptide binding predictions for HLA DR, DP and DQ molecules. *BMC bioinformatics* **11**, 568, doi:10.1186/1471-2105-11-568 (2010).
- 313 Sidney, J. *et al.* Measurement of MHC/peptide interactions by gel filtration or monoclonal antibody capture. *Current protocols in immunology* **Chapter 18**, Unit 18 13, doi:10.1002/0471142735.im1803s100 (2013).
- 314 Trolle, T. *et al.* The Length Distribution of Class I-Restricted T Cell Epitopes Is Determined by Both Peptide Supply and MHC Allele-Specific Binding Preference. *Journal of immunology* **196**, 1480-1487, doi:10.4049/jimmunol.1501721 (2016).
- 315 Sharon, E. *et al.* Genetic variation in MHC proteins is associated with T cell receptor expression biases. *Nature genetics* **48**, 995-1002, doi:10.1038/ng.3625 (2016).
- 316 Battle, A. *et al.* Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome research* **24**, 14-24, doi:10.1101/gr.155192.113 (2014).
- 317 Kirkwood, B. R., Sterne, J. A. C. & Kirkwood, B. R. E. o. m. s. *Essential medical statistics*. 2nd ed. / Betty R. Kirkwood, Jonathan A.C. Sterne. edn, (Blackwell Science, 2003).
- 318 Li, Y. R. *et al.* Meta-analysis of shared genetic architecture across ten pediatric autoimmune diseases. *Nature medicine* **21**, 1018-1027, doi:10.1038/nm.3933 (2015).
- 319 International Stroke Genetics, C. *et al.* Genome-wide association study identifies a variant in HDAC9 associated with large vessel ischemic stroke. *Nature genetics* **44**, 328-333, doi:10.1038/ng.1081 (2012).
- 320 Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nature genetics* **47**, 1236-1241, doi:10.1038/ng.3406 (2015).
- 321 Sulzer, D. *et al.* T cells from patients with Parkinson's disease recognize alpha-synuclein peptides. *Nature* **546**, 656-661, doi:10.1038/nature22815 (2017).
- 322 Stalika, E. *et al.* High-Throughput Profiling of the T-Cell Receptor Gene Repertoire Supports Antigen Drive in the Pathogenesis of Chronic Idiopathic Neutropenia. *Blood* **124** (2014).
- 323 Yin, G. H. *et al.* Screening of differentially expressed genes and predominant expression of beta variable region of T cell receptor in peripheral T cells of psoriatic patients. *Eur J Dermatol* **21**, 938-944, doi:10.1684/ejd.2011.1551 (2011).

- 324 Kaur, G. *et al.* Structural and regulatory diversity shape HLA-C protein expression levels. *Nature communications* **8**, 15924, doi:10.1038/ncomms15924 (2017).
- 325 Horowitz, A. *et al.* Class I HLA haplotypes form two schools that educate NK cells in different ways. *Science immunology* **1**, doi:10.1126/sciimmunol.aag1672 (2016).
- 326 Handunnetthi, L., Ramagopalan, S. V., Ebers, G. C. & Knight, J. C. Regulation of major histocompatibility complex class II gene expression, genetic variation and disease. *Genes and immunity* **11**, 99-112, doi:10.1038/gene.2009.83 (2010).
- 327 Vincent, R. *et al.* Quantitative analysis of the expression of the HLA-DRB genes at the transcriptional level by competitive polymerase chain reaction. *Journal of immunology* **156**, 603-610 (1996).

Supplementary Tables

Table S1: Non-genetic factors explaining the greatest proportion of variance for each vaccine response trait in each population. Step-wise forward and reverse model testing was applied to identify the combination of variables that explained the greatest proportion of variance in antibody responses within each population. Akaike's information criterion (AIC) was used to test between models. Not all variables were available for all populations.

Vaccine trait		Entebbe[†]	P-value (r²)	Soweto[‡]	P-value (r²)	Banfora[§]	P-value (r²)
PT	1) Time since vaccination		1.9x10 ⁻³ (0.006)	1) Time since vaccination 2) Breast feeding	4.6x10 ⁻⁵ (0.03)	1) Time since vaccination 2) Asymptomatic malaria parasitaemia	3.26x10 ⁻⁶ (0.08)
FHA	1) Time since last vaccination		1.0x10 ⁻¹³ (0.04)	1) Time since vaccination 2) Breast feeding	2.1x10 ⁻³ (0.02)	1) Time since vaccination	3.1x10 ⁻³ (0.04)
PRN	1) HIV status at 1 year		1.9 x 10 ⁻⁵ (0.01)	NA	NA	1) Time since vaccination	1.6x10 ⁻⁸ (0.10)
DT	1) Time since vaccination 2) HIV status at 1 year		2.7x10 ⁻¹⁵ (0.05)	1) Time since vaccination 2) Breast feeding 3) Sex	1.9x10 ⁻⁸ (0.05)	1) Time since vaccination	5.9x10 ⁻¹⁰ (0.15)
TT	1) Time since vaccination 2) HIV status at 1 year 3) Household socioeconomic group 4) Ethnicity		7.9x10 ⁻¹⁰ (0.05)	1) Time since vaccination 2) Breast feeding	4.2x10 ⁻³ (0.01)	1) Time since vaccination	1.8x10 ⁻⁵ (0.06)

Table S1: continued

Hib	1) Time since vaccination	1) Time since vaccination	1) Time since vaccination
	2) HIV status at 1 year	2) Breast feeding	
	3) Sex		
	0.03 (0.005)	2.7x10 ⁻⁴ (0.02)	3.9x10 ⁻⁴ (0.06)
HBsAg	1) Time since vaccination	1) Sex	1) Time since vaccination
	2) HIV status at 1 year		
	3) Sex		
	3.1x10 ⁻⁴ (0.02)	0.05 (0.004)	0.03 0.02
Measles	1) Time since measles vaccination	NA	1) Time since vaccination
	2) HIV status at 1 year		
	3) Sex		
	4) Maternal measles concentration		
	5) Asymptomatic malaria parasitaemia		
	6) Number of LRTI events in infancy		
	1.7x10 ⁻⁷ (0.07)	NA	0.02 (0.04)

†: Variables tested included sex (binary), ethnicity (categorical), HIV diagnosis at 1 year (binary), asymptomatic parasitaemia detected at 1 year of age (binary), weight-for-length Z score (continuous), total number of malaria episodes within first year of life (categorical), total number of diarrhoeal episodes within first year of life (categorical), total number of lower respiratory tract infection episodes within first year of life (categorical), maternal socio-economic group (categorical), number of weeks between last vaccination (either EPI or measles) and sampling (continuous)

‡: Variables tested included sex (binary), ethnicity (categorical), HIV diagnosis at 6 months (binary), weight-for-length Z score (continuous), infant breast fed (binary), number of weeks between last vaccination (EPI) and sampling (continuous)

§: Variables tested included sex (binary), ethnicity (categorical), asymptomatic parasitaemia detected at 1 year of age (binary), weight-for-length Z score (continuous), number of weeks between last vaccination (either EPI or measles) and sampling (continuous)

Table S2: List of class II alleles included in amended MiSeq 6-digit ‘G’ groups

Gene	6-Digit G[†]	Included Alleles
<i>DPA1</i>	01:03:01G ²	01:03:01:01/01:03:01:02/01:03:01:03/01:03:01:04/01:03:01:05
<i>DPB1</i>	01:01:01G ²	01:01:01/417:01
<i>DPB1</i>	01:01:01G ²³	01:01:01/417:01
<i>DPB1</i>	02:01:02G ²	02:01:02/141:01/352:01/414:01/416:01
<i>DPB1</i>	02:01:02G ²³	02:01:02/414:01
<i>DPB1</i>	03:01:01G ²³	03:01:01/104:01
<i>DPB1</i>	03:01:01G ²³	03:01:01/104:01/124:01
<i>DPB1</i>	04:01:01G ²³	04:01:01:01/04:01:01:02
<i>DPB1</i>	04:01:01G ²³	04:01:01:01/04:01:01:02/350:01
<i>DPB1</i>	04:01:01G ²	04:01:01:01/04:01:01:02/126:01/350:01/415:01
<i>DPB1</i>	04:02:01G ²	04:02:01:01/04:02:01:02/105:01
<i>DPB1</i>	13:01:01G ²³	13:01:01/107:01
<i>DPB1</i>	13:01:01G ²³	13:01:01/107:01/133:01
<i>DPB1</i>	17:01:01G ²³	17:01/131:01

[†]: Group call of alleles with a superscript if the exon sequenced included to make the group calls (exons 2 with or without 3)

Table S2: continued

<i>DPB1</i>	17:01:01G ²	17:01/131:01/168:01
<i>DQA1</i>	01:03:01G ²	01:03:01:01/01:03:01:02
<i>DQA1</i>	01:02:01G ²	01:02:01:01/01:02:01:02/01:02:01:03/01:02:01:04/01:02:02:01:02:03/ 01:02:04/01:08/01:09/01:11
<i>DQA1</i>	01:01:01G ²	01:01:01/01:01:02/01:04:01:01/01:04:01:02/01:04:02/01:05:01/01:05:02/01:12
<i>DQA1</i>	03:01:01G ²	03:01:01/03:02/03:03:01/03:03:02
<i>DQA1</i>	04:01:01G ²	04:01:01/04:01:02:01/04:01:02:02/04:02/04:04
<i>DQA1</i>	05:01:01G ²	05:01:01:01/05:01:01:02/05:03/05:05:01:01/05:05:01:02/05:05:01:03/05:06/ 05:07/05:08/05:09/05:11
<i>DQB1</i>	02:01:01G ²	02:01:01/02:01:08/02:02:01/02:02:02/02:04/02:06/02:09/02:10/02:48
<i>DQB1</i>	03:01:01G ²³	03:01:01:01/03:01:01:02/03:01:01:03
<i>DQB1</i>	03:03:02G ²³	03:03:02:01/03:03:02:02/03:03:02:03/03:03:02:04
<i>DQB1</i>	03:01:01G ²	03:01:01:01/03:01:01:02/03:01:01:03/03:01:04/03:01:05/03:01:09/03:01:10/03:01:11/ 03:01:12/03:01:20/03:01:26/03:09/03:19/03:21/03:22/03:24/03:29/03:35/03:42/03:49/ 03:50/03:51/03:84N/03:94/03:115/03:116/03:164/03:165/03:169
<i>DQB1</i>	04:02:01G ²	04:02:01/04:13/04:23

[†]: Group call of alleles with a superscript of the exons sequenced to make the group calls (exons 2 with or without 3)

Table S2: continued

<i>DQB1</i>	05:01:01G ²³	05:01:01:01/05:01:01:02/05:01:01:03
<i>DQB1</i>	05:03:01G ²³	05:03:01:01/05:03:01:02
<i>DQB1</i>	05:01:01G ²	05:01:01:01/05:01:01:02/05:01:01:03/05:18/05:27/05:31/05:32/05:45/05:84
<i>DQB1</i>	06:02:01G ²	06:02:01/06:02:12/06:02:23/06:47/06:84/06:109/06:111/06:116/06:117/06:127
<i>DQB1</i>	06:02:01G ²³	06:02:01/06:84
<i>DQB1</i>	06:04:01G ²³	06:04:01/06:39
<i>DQB1</i>	06:04:01G ²	06:04:01/06:34/06:36/06:38/06:39/06:52/06:86
<i>DQB1</i>	06:09:01G ²	06:09:01/06:88
<i>DQB1</i>	06:09:01G ²³	06:09:01/06:88
<i>DQB1</i>	06:03:01G ²	06:03:01/06:41/06:44/06:110
<i>DRB1</i>	03:01:01G ²³	03:01:01:01/03:01:01:02
<i>DRB1</i>	07:01:01G ²³	07:01:01:01/07:01:01:02
<i>DRB1</i>	07:01:01G ²	07:01:01:01/07:01:01:02/07:34
<i>DRB1</i>	09:01:02G ²³	09:01:02/09:21
<i>DRB1</i>	12:01:01G ²	12:01:01/12:06/12:10/12:17
<i>DRB1</i>	12:01:01G ²³	12:01:01/12:10

[†]: Group call of alleles with a superscript of the exons sequenced to make the group calls (exons 2 with or without 3)

Table S2: continued

<i>DRB1</i>	13:01:01G ²	13:01:01/13:117/13:190
<i>DRB1</i>	15:01:01G ²³	15:01:01:01/15:01:01:02/15:01:01:03/15:01:01:04
<i>DRB1</i>	15:03:01G ²³	15:03:01:01/15:03:01:02
<i>DRB3</i>	01:01:02G ²	01:01:02:01/01:01:02:02/01:16
<i>DRB3</i>	02:02:01G ²	02:02:01:01/02:02:01:02/02:28/02:29N
<i>DRB3</i>	03:01:01G ²	03:01:01/03:01:03
<i>DRB4</i>	01:01:01G ²	01:01:01:01/01:03:01:01/01:03:01:03/01:03:02/01:06
<i>DRB5</i>	02:02:01G ²	02:02/02:06

†: Group call of alleles with a superscript of the exons sequenced to make the group calls (exons 2 with or without 3)

“Our greatest glory is not in never falling, but in rising every time we fall”

- Confucious