

Parasite genetic effects and host interactions in *Plasmodium falciparum* malaria



Annie J. Forster
St Peter's College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Michaelmas 2024

For my dear Granda Jim
(1932-2022)

Acknowledgements

First and foremost, I extend my deepest gratitude to my primary supervisors, Dr. Gavin Band and Dr. Alex Mentzer. Their knowledge and skills have helped shape me as a scientist and I am certain that I would not have been able to complete this doctorate without them.

I would like to thank Prof. Dominic Kwiatkowski, Prof. Brian Greenwood, Prof. Adrian Hill, and their teams for the collection, exploration, and storage of the GAMCC samples. I also extend my gratitude to the study participants and their families who contributed to the GAMCC study. Dr. Alfred Amambua-Ngwa was instrumental in updating the ethics framework for the GAMCC samples and overseeing project design. I am thankful for his thoughtful insight throughout my project.

I would also like to further acknowledge Prof. Adrian Hill and Prof. Dominic Kwiatkowski for their long-standing and inspirational dedication to malaria research.

Further thanks go to the Malaria Genomic Epidemiology Network for providing a wealth of excellent data which were heavily relied upon in this work.

I would also like to extend my thanks to the members of the Centre for Human Genetics that have helped along the way, in particular members of Prof. Julian Knight's and Dr. Mentzer's groups, who offered invaluable practical support. Dr. Amanda Chong in particular ensured I was able to access the necessary resources to complete my work.

I would like to express my appreciation to my family. My parents have been my foundation of support and provided cake and tea in times of need. My brothers, along with their wonderful partners, have been a source of great friendship: Omin and Emji helped me settle in to Oxford and made lockdown bearable, if not enjoyable, while Budin and Cell-dog let me escape to Amsterdam and are great fun to be around. I am also deeply thankful to Grandma Jean and Granda Jim, who are very dear to me and instilled in me the value of learning, for which I am very grateful.

I also would like to thank my “chosen family”: my soul sister Irene, who kept me sane from across the globe, my other smol ones Bianca and Serena, and my tol one Adelaide who all provided much needed escape and laughter. Adelaide, in particular, offered me rent-free accommodation in London whenever I needed it, and is a wonderful host. I would also like to acknowledge my newer friends, in particular my basketball team, “zesty” college pals, and amazing housemates who made tough times at Oxford bearable. I also extend special thanks to George who supported me through to the finish line.

My DPhil was supported by the Wellcome Trust.

Chapter specific contributions

Chapter 3

- Oxford Genomics Centre sequenced all GAMCC study samples.
- Doris Quinkert (Dept. of Biochemistry) contributed *Plasmodium falciparum* control DNA for use in sequencing trials.
- Katie Healey contributed PCR runs prior to Sanger sequencing for *HBB* genotypes in the main study samples.
- Zhipeng Zhang contributed imputed genotype calls from low-coverage whole genome sequence data generated for a subset of GAMCC samples.

Chapter 5

- Gavin Band collaborated with me on a pipeline to produce a phased, polarised variant data for samples in MalariaGEN Pf7 and compute extended haplotype homozygosity metrics.
- William Hamilton (Wellcome Sanger Institute) and Lucas Amenga-Etego (University of Ghana) provided data from mild infections from Ghana.

Chapter 6

- Gavin Band developed the coverage caller used to type structural variation at a locus in the *P. falciparum* genome.

- Richard Pearson (Wellcome Sanger Institute) provided access to Malaria-GEN Pf6 alignment files.
- Thomas N. Williams, Silvia Kariuki, Alex Macharia and Patrick Ombati (KEMRI-Wellcome Trust Research Programme) provided *P. falciparum* DNA samples from Kenyan isolates for long-read sequencing.
- BEI resources (<https://www.beiresources.org/>) provided *P. falciparum* DNA samples from Kenyan isolates for long-read sequencing.
- James Docker (Technology Platforms Group) processed and sequenced additional *P. falciparum* DNA samples on the long-read PacBio platform.
- Gavin Band implemented a genome assembly and annotation pipeline for long-read sequenced samples and contributed the relevant Methods sections.
- Jason Hendry, Anna Jeffreys and Mariateresa de Cesare contributed amplicon-based Nanopore sequencing of 3D7 and FUP-H genomes.

Abstract

Malaria remains a substantial public health threat. Over recent decades improvements in rates of mortality from *Plasmodium falciparum* malaria have been observed in Africa where most disease burden persists. This success however, is threatened by parasite counter evolution against anti-malarial drugs and recently approved vaccines for *falciparum* malaria. Human genetics has been extensively studied in terms of malarial disease. Comparatively little is understood about parasite genetic effects on disease outcomes, and their response to natural or vaccine induced host immunity. There is therefore a need to investigate parasite genetics and how they interact with the host's in order to inform intervention strategies, control disease, and reach the eventual goal of elimination. This requires the generation of large datasets of host and parasite genetic information, alongside accurate phenotypic data, to infer how these factors influence disease outcomes.

In this thesis, I develop a new resource to investigate parasite and human genetic variation in malarial disease, using severe and mild cases collected in Gambian infants from 1988-90 (the 'GAMCC' resource). I integrate this with public datasets to perform powered, cross population analyses including an unbiased, genome-wide scan for parasite genetic effects on disease severity. This indicates a potential association on parasite chromosome 1 with evidence of replication but reveals minimal evidence of strong parasite genetic effects overall. I further investigate the recent finding that *P. falciparum* is evolving at loci known as *Pfsa* to counteract HbS-mediated resistance, identifying a novel putative region of interaction on parasite chromosome 2 (named *Pfsa2B*). I also uncover additional potential population-specific effects, highlighting the intricacy of HbS-*Pfsa* relationships. Through novel methods development I characterise complex structural variation at *Pfsa3*, which extends the potential that this might underlie the association of this region with HbS. My analyses provide important insights into host-pathogen interactions, opening possibilities for several future functional studies.

Contents

1	Introduction	1
1.1	Introduction	1
1.1.1	<i>Plasmodium falciparum</i> malaria	2
1.1.2	The complex life cycle of <i>P. falciparum</i>	3
1.1.3	The current interventions used to tackle malaria	4
1.1.4	The outlook for malaria control	7
1.2	An overview of human genetic analysis	7
1.2.1	Genetic variation	7
1.2.2	Candidate loci analysis	8
1.2.3	Genome-wide analyses	9
1.2.3.1	Array-based genotyping	9
1.2.3.2	Whole genome sequencing	9
1.2.3.3	Short read sequencing	10
1.2.3.4	Long read sequencing	11
1.3	Human genetics and malaria disease outcomes	12
1.3.1	Blood-mediated resistance	13
1.3.1.1	Polymorphisms in the <i>HBB</i> gene	13
1.3.1.2	Alterations to the production of alpha- or beta-globin	14
1.3.1.3	Erythrocyte cell surface molecules	14
1.3.1.4	G6PD deficiency	16
1.3.1.5	Alterations of calcium flux	17
1.3.2	Immune-mediated resistance	17
1.4	Genomics methods for malaria parasites	20
1.4.1	Short read sequencing	20

1.4.2	Large scale short read sequencing of <i>P. falciparum</i> genomes	21
1.4.3	Long read sequencing	22
1.5	Parasite genetics and malaria disease outcomes	23
1.5.1	The population genetics of <i>P. falciparum</i>	23
1.5.2	Evidence of evolution in the <i>P. falciparum</i> genome	25
1.5.2.1	Anti-malarial drug resistance	25
1.5.2.2	Parasite-origin erythrocyte surface molecules	27
1.6	Emerging evidence of host-parasite coevolution	29
1.7	Summary	31
1.7.1	Structure of this thesis	31
2	Materials and Methods	33
2.1	Description of data used within this thesis	33
2.1.1	The Genetic Analysis of Malaria Cases and Controls cohort	33
2.1.1.1	Ethics statement	34
2.1.2	The MalariaGEN Consortial Project 1 (CP1) cohort	34
2.1.3	The MalariaGEN Pf7 cohort	34
2.1.4	Long-read assemblies	35
2.2	Preparation of samples for <i>Pf</i> DNA sequencing	35
2.2.1	Initial preparation of samples in the GAMCC cohort	35
2.2.2	Depletion of human DNA	36
2.2.2.1	McrBC	36
2.2.2.2	Bead enrichment	37
2.2.3	<i>Pf</i> Whole Genome Amplification	37
2.2.3.1	Selective Whole Genome Amplification (sWGA)	37
2.2.3.2	Random Whole Genome Amplification (rWGA)	39
2.3	Description of DNA sequencing methods for <i>Pf</i> parasites	40
2.3.1	Short read sequencing	40
2.3.1.1	NEBNext Ultra DNA Library Prep kit	40
2.3.1.2	NEBNext Ultra II FS kit	40
2.3.1.3	NEBNext Ultra II (PCR-free) kit	40
2.3.2	Long read sequencing	41
2.3.2.1	PacBio sequencing	41

2.3.2.2	Nanopore amplicon sequencing	42
2.4	Targetted human sequencing	42
2.4.1	HLA typing	42
2.4.2	HBB typing	43
2.4.2.1	Low-coverage sequencing and array-based typing of GAMCC samples	44
2.4.2.2	The Sanger sequencing method	44
2.4.3	Previously performed typing	47
2.5	Genetic data alignment and processing	47
2.5.1	Reference assemblies used and <i>Pf</i> genome information . . .	47
2.5.2	Alignment of short read sequence data	48
2.5.3	Quality control assessments for short read alignments	49
2.5.3.1	Assessment of human contamination	49
2.5.3.2	Assessment of breadth of sequencing coverage . . .	49
2.5.3.3	Assessment of duplication levels	50
2.5.4	Processing of long read sequence data	50
2.5.4.1	Generation of assemblies from PacBio data	50
2.5.4.2	Analysis of amplicons generated using Nanopore sequencing	52
2.6	Variant call sets developed for this thesis	52
2.6.1	The production of a combined GAMCC/CP1 call set	52
2.6.2	Generating an intersected variant call set from the GAMCC, CP1 and Pf7 cohorts	53
2.6.3	Quality control and filtering of variant call sets	54
2.6.3.1	Sample filtering	54
2.6.3.2	Variant filtering	56
2.6.3.3	Principal component analysis and other quality con- trol	57
2.6.4	Generating phased polarised haplotypes across African sam- ples in the Pf7 resource	57
2.7	Genome wide association analyses	58
2.7.1	Phenotyping infections as mild or severe	58
2.7.1.1	GAMCC samples	58

2.7.1.2	CP1 and Pf7 samples	59
2.7.2	Basic association model using SNPTEST	59
2.7.3	Modelling true discovery of genetic associations and performing power analyses	61
2.7.4	Meta-analysing genome-wide associations between populations using BINGWA	63
2.7.5	Modelling host-parasite genetic associations using HPTEST	65
2.8	Other statistical analyses	66
2.8.1	Estimation of population relative risks using malaria cases and controls	66
2.8.2	Assessing the accuracy of the HLA-B*53 typing assay	68
2.9	Population genetic analyses	69
2.9.1	Measuring complexity of infection in <i>Pf</i> WGS data	69
2.9.2	Modelling allele frequency shift using the X^tX statistic	70
2.9.3	Computing between chromosome linkage disequilibrium	74
2.9.4	Determining ancestral allele frequencies in the <i>Pf</i> genome	75
2.9.5	Extended haplotype homozygosity (EHH) and integrated haplotype score (iHS) analyses	76
2.10	Structural variant calling	78
2.10.1	Approximating structural variant haplotypes at the CLAG3.1 and CLAG3.2 locus	78
2.10.2	Analysis of structural variation at the <i>Pf</i> sa3 locus	80
2.10.2.1	Sample selection for these analyses	81
2.10.2.2	Coverage call	81
2.10.2.3	Insert size call	83
2.10.2.4	Kmer call	84
3	Generation of the ‘Genomic Analysis of Malaria Cases and Controls’ (GAMCC) resource	87
3.1	Introduction	87
3.2	Results	89
3.2.1	Enriching <i>Pf</i> DNA in GAMCC samples prior to sequencing	89

3.2.2	Establishing the optimal library preparation method for the GAMCC set	91
3.2.3	Drop out of regions within sequencing data	93
3.2.4	Optimisation of sWGA parameters	97
3.2.5	Completing <i>Pf</i> WGS for the full GAMCC set	100
3.2.6	Comparison of the GAMCC to a public dataset	101
3.2.7	Performing a surrogate “sex check” using contaminant hu- man reads	102
3.2.8	Variant Calling in a large, combined <i>Pf</i> WGS set	104
3.2.9	Targeted human typing in the GAMCC set	104
3.2.9.1	HLA typing	104
3.2.9.2	Haemoglobin S/C typing	105
3.3	Discussion	107
4	Does parasite genetic variation influence the severity of malaria infection?	110
4.1	Introduction	110
4.2	Results	112
4.2.1	Does parasite genetic diversity influence risk of severe malaria disease?	112
4.2.1.1	Assessing parasite genetic effects in the GAMCC dataset	112
4.2.1.2	Replication analysis using the MalariaGEN Pf7 and CP1 resources	115
4.2.1.3	Inspecting a putative severity locus	124
4.2.2	Does parasite genetics influence severe disease subtype? . . .	131
4.3	Discussion	135
4.A	Supplementary Figures and Tables	140
4.A.1	Supplementary Figures	140
4.A.2	Supplementary Tables	146

5	Investigation of co-evolutionary interactions between <i>Plasmodium falciparum</i> parasites and the human HbS and <i>HLA-B</i> loci	148
5.1	Introduction	148
5.2	Results	151
5.2.1	Assessing HbS-associated <i>Pfsa</i> loci in the <i>Pf</i> genome using the GAMCC resource and other datasets	151
5.2.1.1	<i>Pfsa1</i> and <i>Pfsa3</i> replicate within the GAMCC dataset	151
5.2.1.2	A putative <i>Pfsa</i> locus discovered in the GAMCC dataset	156
5.2.2	Joint analysis of HbS associations using cross-population analysis	157
5.2.2.1	Assessment of the <i>Pfsa2B</i> association across datasets	158
5.2.2.2	Further population specific <i>Pfsa</i> relationships are found in public datasets	161
5.2.2.3	A re-evaluation of how HbS protection varies with parasite genotype	173
5.2.3	A review of the association of HLA-B*53 and protection against severe malaria	178
5.2.3.1	Power analysis of the original study	179
5.2.3.2	Inclusion of ethnicity to account for population structure	181
5.2.3.3	An evaluation of the accuracy of the HLA-B*53 typing assay	183
5.2.3.4	An assessment of how accuracy of the HLA-B*53 typing assay may have affected results of the original study	186
5.2.3.5	A genome-wide scan for associations in the <i>Pf</i> genome to HLA-B*53	186
5.2.4	Population genetics at the <i>Pfsa</i> loci	189
5.2.4.1	Unusual signals of LD in the GAMCC set	190
5.2.4.2	The allele frequencies of the <i>Pfsa</i> loci have increased in The Gambia	194

5.2.4.3	Haplotypes at the <i>Pfsa</i> loci are shared across populations	196
5.2.4.4	There is minimal evidence of recent positive selection at the <i>Pfsa</i> loci in global populations	199
5.3	Discussion	202
5.A	Supplementary Figures and Tables	207
5.A.1	Supplementary Figures	207
5.A.2	Supplementary Tables	219
6	Structural variation at a sickle-associated locus of the <i>Plasmodium falciparum</i> genome	225
6.1	Introduction	225
6.2	Results	227
6.2.1	Using coverage and split reads to infer structural variant genotypes from short-read <i>Pf</i> genome sequencing	227
6.2.1.1	The coverage caller	228
6.2.1.2	The insert size caller	229
6.2.1.3	Inspecting discordant samples and discovering new SVs	229
6.2.1.4	Generating a large set of consensus SV calls	230
6.2.2	A catalogue of <i>Pfsa3</i> structural variant types segregating in global populations	231
6.2.3	Confirmation of SV haplotype structures using long-read sequencing	234
6.2.3.1	Development and testing of a long-read amplicon based assay for <i>Pfsa</i> regions	236
6.2.4	The DUP-B and DUP-F haplotypes are strongly linked to <i>Pfsa3</i> + SNPs	236
6.2.5	Paralogous sequence variation within the <i>Pfsa3</i> SVs	240
6.2.6	A third method to call <i>Pfsa3</i> SVs using PSVs	241
6.2.6.1	Testing the kmer calling method	242
6.2.7	Genomic prediction of effect of <i>Pfsa3</i> SVs on gene transcription	243

6.2.7.1	Data to assess gene expression in parasites carrying different SV haplotypes	243
6.2.7.2	Gene expression within a DUP-F carrying parasite compared to 3D7	243
6.2.7.3	The duplication of SNRPF gives rise to additional RNA transcripts	244
6.2.7.4	Assessments of RNA expression data from clinical infections	247
6.3	Discussion	249
6.A	Supplementary Figures and Tables	253
6.A.1	Supplementary Figures	253
6.A.2	Supplementary Tables	257
7	Concluding Remarks	260
7.0.1	Limitations and Future Directions	262
	Bibliography	264

List of Figures

1.1	The life cycle of <i>Plasmodium</i> parasites	3
1.2	The process of Illumina-based cluster formation for whole genome short-read sequencing.	10
1.3	The method of long-read sequencing by PacBio	11
1.4	Challenges in resolving repetitive sequence and structural variation using short read genomic sequencing data.	22
2.1	Example sanger-sequencing traces used to genotype the HbS locus .	46
2.2	Prior on effect for a meta-analysis of malaria severity subtype in BINGWA	64
2.3	X^tX statistics from an empirical model of allele frequencies in The Gambia for different allele frequency bins.	73
2.4	Calling the presence of the CLAG3h hybrid in four datasets	79
2.5	Normalised coverage across the <i>Pfsa3</i> locus used to identify struc- tural variation	82
3.1	Proportions of <i>Pf</i> DNA in GAMCC samples prepared using different methods.	90
3.2	Average breadth of coverage of the <i>Pf</i> 3D7 genome between different sample preparations	92
3.3	Coverage across the <i>Pfsa</i> 1-3 regions in the <i>Pf</i> 3D7 genome between different sample preparations	95
3.4	Coverage across the LSA-1 and LSA-3 regions in the <i>Pf</i> 3D7 genome between different sample preparations	96
3.5	Comparison of sWGA parameters on GAMCC samples.	98

3.6	Comparison of sequence data quality across the pilot sequencing run, the full GAMCC set, the filtered GAMCC analysis set, and 100 random alignments from the MalariaGEN CP1 dataset	100
3.7	A “sex check” of the GAMCC <i>Pf</i> WGS data.	103
4.1	Power and true discovery rate analysis for an analysis of <i>Pf</i> parasite genetic effects on severity using the GAMCC dataset	113
4.2	Strength of evidence for the association of variants across the <i>Pf</i> genome with disease severity in the GAMCC dataset	114
4.3	PCA in the Gambian population.	118
4.4	PCA in the Kenyan population.	119
4.5	Changes in allele frequency over time at the chloroquine resistance locus CRT in Gambia and Kenya	121
4.6	Evidence of association of variants across the <i>Pf</i> genome with disease severity in the Kenyan population across different GWAS models	122
4.7	Final strength of evidence for the association of variants across the <i>Pf</i> genome with disease severity in Kenya using the CP1 and Pf7 datasets	124
4.8	Estimated effect sizes compared between different GWAS scans at top hits from GWAS of parasite genetic effects on malaria severity .	125
4.9	PCA in the GAMCC dataset.	127
4.10	Comparison of strength of association at Pf3D7_01.v3:557240:T>A correcting for possible confounders	129
4.11	Estimated effect sizes for Fws and COI on malaria severity	131
4.12	Strength of evidence within a meta-analysis of association across variants in the <i>Pf</i> genome with severe malaria subtype	132
4.13	Estimated effect sizes compared between different GWAS scans at top hits from the meta-analysis of parasite genetic effects on severe malaria subtype	134
4.14	PCA results without highly related samples removed	141
4.15	Strength of evidence for the association of variants across the <i>Pf</i> genome with disease severity in the Gambia and Kenya using the CP1 and Pf7 datasets	141

4.16	Sample sizes of <i>Pf</i> genomes in the Gambian and Kenyan populations across time	142
4.17	Evidence of association of variants across the <i>Pf</i> genome with disease severity in the Gambian population across different GWAS models	142
4.18	Strength of evidence for the association of variants across the <i>Pf</i> genome with disease severity in Kenya using the CP1 and Pf7 datasets without using drug resistance loci as covariates.	143
4.19	Differential missingness and heterozygosity between variants in the CP1 and Pf7 datasets	143
4.20	Depth and mapping quality for Pf3D7_14_v3:2,092,254:A>G with the GAMCC/CP1 callset.	144
4.21	Depth and mapping quality for Pf3D7_01_v3:557240:T>A with the GAMCC/CP1 callset.	144
4.22	PCA results without highly related samples removed in the GAMCC	145
4.23	Comparison of association results in the GAMCC with and without highly related samples removed	145
4.24	Comparison of Fws statistics and COI calculated using coiaf software in <i>Pf</i> WGS data	146
5.1	Evidence for association between <i>Pf</i> variation and HbS across four datasets	153
5.2	Estimated effect sizes of associations between <i>Pf</i> sa1-4 and HbS across four datasets	155
5.3	Comparison of evidence for the association of <i>Pf</i> sa2B to HbS	158
5.4	Allele frequencies of <i>Pf</i> sa variants within populations in the MalariaGEN Pf7 resource	160
5.5	Comparison of evidence for the association of Pf3D7_07_v3:614420:C>T to HbS	162
5.6	Estimated effect sizes of associations between Pf3D7_07_v3:614420:C>T and HbS within CP1 Gambian samples conditioned on the DR locus in the CRT gene	163

5.7	Comparison of evidence for the association of Pf3D7_03_v3:140060:T>A to HbS	165
5.8	An schematic of association analyses conducted between CLAG3 genes and HbS	166
5.9	Associations between Pf3D7_03_v3:140060:T>A and HbS across four datasets	167
5.10	Correlation between potential <i>Pfsa</i> loci and <i>Pfsa3</i> within populations in the MalariaGEN Pf7 resource	170
5.11	Comparison of evidence for the association of Pf3D7_09_v3:1428875:A>T to HbS	172
5.12	Comparison of evidence for the association of Pf3D7_03_v3:79845:G>A to HbS	173
5.13	Estimated relative risk of severe malaria for HbS carriers in infections of different <i>Pfsa</i> genotypes	177
5.14	Power and true discovery rate analysis for the analysis of an association of HLA-B*53 to protection against severe malaria	180
5.15	A comparison of results for the association of HLA-B*53 and protection against severe malaria discovered in the GAMCC dataset	182
5.16	Differences between HLA-B alleles and oligonucleotides used to type HLA-B*53	185
5.17	Models of interactions between HLA-B*53 status, parasite genotype and malaria disease severity.	187
5.18	Evidence for association between <i>Pf</i> variation and HLA-B*53 in the GAMCC dataset	189
5.19	Genome-wide between chromosome LD in the GAMCC dataset	191
5.20	Allele frequencies of <i>Pfsa</i> variants within populations in the MalariaGEN Pf7 resource	193
5.21	Increase in <i>Pfsa</i> allele frequencies over time in The Gambia	195
5.22	Haplotypes at the <i>Plasmodium falciparum</i> sickle associated (<i>Pfsa</i>) loci in samples from the MalariaGEN Pf7 resource	198
5.23	Haplotype at the <i>Pfsa2B</i> locus in samples from the MalariaGEN Pf7 resource	199

5.24	iHS results for the <i>Pfsa</i> loci in samples from the MalariaGEN Pf7 resource	201
5.25	Estimated effect sizes of associations between <i>Pfsa</i> 2B and HbS across four datasets conditioned of genotype at <i>Pfsa</i> 1	208
5.26	Assessment of quality at the <i>Pfsa</i> 2B (Pf3D7_02_v3:825357:T>A) variant	209
5.27	Assessment of quality of Pf3D7_07_v3:614420:T>A variant	210
5.28	Visualisation of PSV sites between the CLAG3.1 and CLAG3.2 genes in the <i>Pf</i> genome	211
5.29	Assessment of quality of Pf3D7_03_v3:140060:T>A variant	212
5.30	Kmer sharing across the CLAG3.2 and CLAG3.1 locus in a number of whole genome assemblies	213
5.31	Assessment of quality of Pf3D7_09_v3:1428875:A>T variant	214
5.32	Allele frequencies of potential <i>Pfsa</i> variants within populations in the MalariaGEN Pf7 resource	215
5.33	Assessment of quality of Pf3D7_03_v3:79845:G>A variant	216
5.34	Estimated relative risk of severe malaria for HbS carriers in infections of different <i>Pfsa</i> genotypes in two datasets from Gambia	217
5.35	iHS results for known drug resistance loci in samples from the MalariaGEN Pf7 resource	218
6.1	Catalogue of structural variant haplotypes at the <i>Pfsa</i> 3 locus	232
6.2	Allele frequencies of structural variants at the <i>Pfsa</i> 3 locus in global populations	233
6.3	<i>Pfsa</i> 3 haplotypes for a subset of long-read genome assemblies	235
6.4	Allele frequencies at SNPs across the <i>Pfsa</i> 3 region in samples with different SV haplotypes	240
6.5	Differential gene expression between the FUP-H and 3D7 <i>Pf</i> isolates cultured in HbS-trait and non-trait blood culture	244
6.6	Proportions of PSV alleles between the FUP-H and 3D7 <i>Pf</i> isolates cultured in HbS-trait and non-trait blood culture	246
6.7	Gene expression within clinical <i>Pf</i> isolates collected in Mali	248

6.8	Proportions of PSV alleles within clinical <i>Pf</i> infections collected in Mali	249
6.9	Nucleotide sequence (kmer) sharing between <i>Pf</i> assemblies at the <i>Pfsa3</i> locus	254
6.10	Primers designed to amplify the <i>Pfsa3</i> locus for amplicon sequencing on the ONT MinION	255
6.11	Targetted amplicon sequencing of the <i>Pfsa3</i> locus within the FUP-H on ONT MinION	256
6.12	RNA-seq read pairs from FUP-H align to PF3D7_1127000 and the partial duplication of SNRPF	256

List of Tables

1.1	Research questions and datasets used in each chapter	32
2.1	The 10A sWGA primer set	38
2.2	The 6A sWGA primer set	38
2.3	PCR amplification steps for the <i>HBB</i> gene	45
3.1	sWGA parameters tested on the GAMCC set	99
3.2	Summary of sample numbers in the GAMCC set.	101
3.3	Comparison of original and retyped <i>HBB</i> genotypes in non-reference samples in the GAMCC cohort	106
3.4	Results of <i>HBB</i> retyping in the GAMCC cohort	107
4.1	Sample sizes for analysis of malaria disease severity	116
4.2	Drug resistance loci included as covariates in GWAS models.	147
5.1	Summary of datasets used in Chapter 5	151
5.2	<i>Pf</i> genotype counts at <i>Pfsa</i> loci in HbA vs. HbS individuals in West-African populations	175
5.3	<i>Pf</i> genotype counts at <i>Pfsa</i> loci in HbA vs. HbS individuals in Kenya (an East-African population)	176
5.4	Results of HLA-B*53 typing compared to original typed calls	183
5.5	<i>Pfsa</i> -HbS association statistics across datasets	221
5.6	Regions with highest values of between chromosome linkage dise- quilibrium in the GAMCC dataset	222
5.7	iHS results for <i>Pfsa</i> and drug resistance loci in samples from the MalariaGEN Pf7 resource	224

6.1	Linkage between Pf3D7_11_v3:1058035:T>A and SV haplotypes at the <i>Pfsa3</i> locus	238
6.2	Linkage between Pf3D7_11_v3:1057437:T>C and SV haplotypes at the <i>Pfsa3</i> locus	239
6.3	Segments across the <i>Pfsa3</i> involved in structural variation	257
6.4	Structural variant frequencies at the <i>Pfsa3</i> locus across global populations in the consensus call set	258
6.5	Primers for targetted amplicon sequencing of the <i>Pfsa3</i> locus within the FUP-H on ONT MinION	259
6.6	Structural variant calls at the <i>Pfsa3</i> locus made using RNA-seq data from clinical infections collected in Mali using the kmer caller . . .	259

Abbreviations

AAT1 amino acid transporter 1

AC allele count

ACS acyl-CoA synthetase

AF allele frequency

AH Adrian Hill

AMA1 apical membrane antigen 1

ART artemisinin

AT adenosine-thymine

ATP adenosine triphosphate

ATP2B4 ATPase plasma membrane Ca^{2+} transporting gene 4

BAM Binary Alignment Map

BF Bayes Factor

BoCM ‘breadth of coverage metric’

bp base pair

BSA bovine serum albumin

BSG basigin

CHG Centre for Human Genetics

CI confidence interval

CLAG cytoadherence linked asexual gene

CM cerebral malaria

COI complexity of infection

CP1 Consortial Project 1

CQ chloroquine

CRT chloroquine resistance transporter

CSP circumsporozoite protein

DARC Duffy antigen receptor for chemokines

DBL duffy binding-like

DBS dried blood spot

ddNTP dideoxynucleotide

DHPS dihydropteroate synthetase

DHFR dihydrofolate reductase

DK Dominic Kwiatkowski

DNA deoxyribonucleic acid

DR drug resistance

DRC Democratic Republic of Congo

dsDNA double stranded DNA

DV digestive vacuole

EBA erythrocyte binding antigen

EBL erythrocyte binding-like

EHH extended haplotype homozygosity

FIKK serine/threonine protein kinase

FUP-H Ugandan Palo-Alto

GAMCC Genetic Analysis of Malaria Cases and Controls

GC guanine-cytosine

gDNA genomic DNA

GWAS genome wide association study

GYP glycophorin

G6PD glucose-6-phosphate dehydrogenase

Hb haemoglobin

HBA alpha-globin

HBB beta-globin

HIV-1 human immunodeficiency virus 1

HLA human leukocyte antigen

hyp10 hypothetical gene family 10

IBD identical by descent

IGV Integrative Genomics Viewer

iHH integrated haplotype homozygosity

iHS integrated haplotype score

INDEL short insertion or deletion

Kb kilobase

KENI Kenyan isolate

lcWGS low coverage WGS

LD linkage disequilibrium

LSA liver stage antigen

MalariaGEN Malaria Genomic Epidemiology Network

MAPQ mapping quality score

Mb megabase

MDR1 multidrug resistance 1

MHC major histocompatibility complex

mM millimole

MRC Medical Research Council

MRScyt methionine-tRNA ligase

MSP merozoite surface protein

ng nanogram

NGS Next Generation Sequencing

nM nanomole

ONT Oxford Nanopore Technologies

OR odds ratio

P P-value

PacBio Pacific Biosciences

PCA principal component analysis

PCR polymerase chain reaction

PEXEL *Plasmodium* export element

Pf *Plasmodium falciparum*

PfEMP1 *Plasmodium falciparum* erythrocyte membrane protein 1

Pfsa *Plasmodium falciparum* sickle associated

PMF probability mass function

PPPK hydroxymethyldihydropterin pyrophosphokinase

PSV paralogous sequence variant

PVL Panton-Valentine leucocidin

QC quality control

RBC red blood cell

RBL reticulocyte binding-like

REX ring exported protein

RH5 reticulocyte binding protein homologue 5

RIFIN repetitive interspersed family

RNA ribonucleic acid

RNA-seq RNA sequencing

RR relative risk

RTS,S RTS,S/AS01

rWGA random whole genome amplification

R21 R21/Matrix-M

SEA South-East Asia

SMA severe malarial anaemia

SNRPF small nuclear ribonucleoprotein polypeptide F

SNP single nucleotide polymorphism

SP sulfadoxine and pyrimethamine

SSA Sub-Saharan Africa

ssDNA single stranded DNA

STEVR subtelomeric variable open reading frame

SV structural variant

sWGA selective whole genome amplification

TDR true discovery rate

TNF tumour necrosis factor

TPM transcripts per million

UTR untranslated region

VCF variant call format

VQSLOD Variant Quality Score log odds

VQSR Variant Quality Score Recalibration

WGS whole genome sequencing

WHO World Health Organisation

5mC 5-methylcytosine

μL microlitre

μM micromolar

Chapter 1

Introduction

1.1 Introduction

Malaria has been an object of scientific fascination since the birth of science itself. Historical reports of high fever come from as early as the Ancient Egyptians (~1300 BC) (1) who are known to have been heavily afflicted (2). More suggestive descriptions of cyclical fevers and an association with water appeared later in texts by the Ancient Greek philosopher Hippocrates (460 BC) (3). The disease later earned its modern name “bad air” (mal’aria) in medieval Italy, christened to describe the putrid miasma which emanated from marshlands and was reputed to be causative of the disease (4). As the field of medicine has progressed, so has our understanding of the factors which cause this disease and effect its outcomes, from the discovery of the malarial parasite in 1880, earning Charles L. A. Laveran the Nobel Prize for Medicine (5), and culminating in a catalogue of loci throughout the human genome which are now known to influence disease risk (6). However, for as much as is uncovered, more mysteries materialise. This is partly due to the complex nature of the genomes of humans, malaria parasites, and their mosquito vectors, which technological advances are gradually helping to unravel. For example, only recently a rare group of *Plasmodium falciparum* (*Pf*) parasites, which have been dubbed a “cryptotype”, were identified. This group appear to be genetically related but are distributed widely across Africa, suggesting they may have been maintained by a yet unidentified evolutionary force (7). The purpose of this thesis is to further leverage genomic sequencing to continue to elucidate

the influence of genetics on malarial disease, with a particular focus on parasite genetic effects and how these interact with human genetics to impact outcomes.

1.1.1 *Plasmodium falciparum* malaria

Malaria is a pervasive disease which afflicts many vertebrates, from our closest relative the Chimpanzee, to more distantly related avian cousins (8). Five species of malaria parasite are known to cause disease in humans: *Pf*, *P. vivax*, *P. ovale*, *P. malariae* and *P. knowlesi* (9). Malaria is estimated to have caused more than 600,000 deaths in 2022 (95% confidence interval (CI) 566,000-738,000), the vast majority of which were in Africa (>95%) (9). *P. vivax*, the second most prevalent and most widely distributed malaria species, poses a significant global health challenge due to its substantial contribution to morbidity (10). However, *Pf* accounts for over 97% of global malaria deaths, highlighting the critical need for targeted research and control measures to combat this more lethal form.

In endemic regions, including much of Sub-Saharan Africa (SSA), many will become infected with *Pf* parasites at a young age and may experience disease symptoms (11). This includes a high fever which has classically been described as occurring periodically (“tertian fever”) due to the cyclical nature of the human stages of *Pf* infection (described below) (12). These bouts of “uncomplicated” malaria have low fatality rates. A very small proportion however (~1%) will proceed to develop the much more fatal syndrome “severe malaria”, a heterogeneous presentation which can include many different symptoms; most predominantly cerebral malaria (CM) (involving coma and convulsions (12)), severe malarial anaemia (SMA) (a severe reduction in haemoglobin (Hb) levels) and respiratory distress (11). CM is considered the more severe of these and is associated with greater fatality rates (13). Consequently, most malaria deaths (more than three quarters in 2022) occur within infants under the age of five (9). Over successive infections individuals will eventually acquire natural immunity to the disease and will rarely experience overt symptoms (14). Understanding this process, and how the development of severe malaria is thus prevented, has been fundamental to the search for an effective malaria vaccine which has been ongoing for the past four decades (15; 11).

1.1.2 The complex life cycle of *P. falciparum*

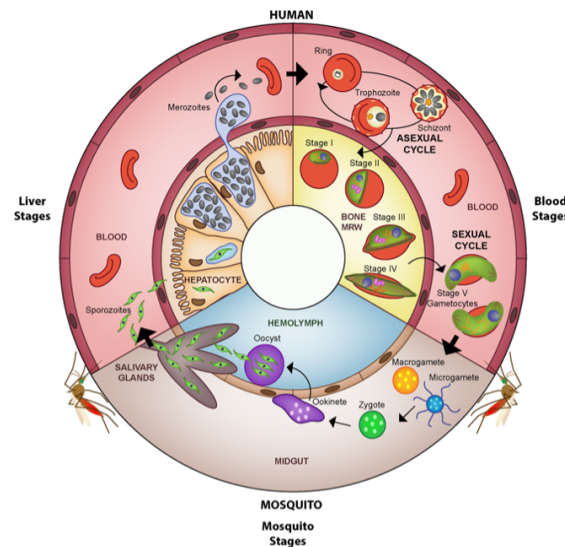


Figure 1.1: **The life cycle of *Plasmodium* parasites.** Reproduced from (16) (available under the Creative Commons Attribution 4.0 International (CC BY) licence (<https://creativecommons.org/licenses/by/4.0/>)).

To grasp the complexities of this disease, it is important to understand the *Pf* parasite's complicated life cycle. The malaria parasite is a master of shapeshift, completely altering its biology to become minutely suited to infecting its mosquito or human hosts (depicted in Fig. 1.1). A malaria infection begins with the bite of a female mosquito in the *Anopheles* genus. While the female feeds, small and motile parasite cells called sporozoites enter the blood and quickly migrate to the liver where they invade hepatic parenchymal cells. Here, pre-erythrocytic schizogony begins, producing tens of thousands of merozoites. Within *P. vivax* and *P. ovale* infections a proportion of liver cells remain infected by dormant hypnozoites which can cause later relapses of disease (17; 12). The infected liver cells eventually burst allowing the merozoites into the blood stream where they invade erythrocytes (also referred to here as red blood cells (RBCs)) (12). *Pf* invasion occurs via the binding of multiple parasite antigens to erythrocyte surface proteins (11; 18) (which will be discussed further below). Inside the RBCs, over a period of 48 hours, merozoites transform from smaller rings stages to large trophozoites. Successive rounds of mitotic division lead to the final schizont form which eventually rupture to release

further merozoites into the blood stream. Within minutes these infect new RBCs repeating the blood stages in further rounds of asexual division and over time generating 10^{13} - 10^{14} circulating parasites (12; 19). This period of infection, known as the asexual cycle, is responsible for malaria associated disease symptoms (16).

A small proportion of parasites within this stage, instead of becoming further merozoites, commit to becoming gametocytes. This is required to enable transmission of the disease to new hosts. The proportion of parasites which commit to the sexual cycle, as well as the sex ratio of produced gametes, are adaptively controlled by parasites to maximise transmission success (20). Once within a mosquito, following three rounds of mitosis of the microgametes, a single micro- and macrogamete fuse to form a zygote within the insect's midgut (12). These penetrate the gut wall forming an ookinete and eventually an oocyst. Sporozoites are produced by cell division within the oocyst and are transferred via the insect haemolymph to the salivary glands in preparation to infect a new host upon the next blood feed (12).

1.1.3 The current interventions used to tackle malaria

Given the significant global impact of *Pf* malaria, numerous intervention strategies have been developed to mitigate its effects. Intervention and control strategies generally aim for at least one of two goals: prevention and case-management. Preventative measures include non-clinical interventions such as the use of insecticide treated bed-nets which thwart mosquito bites (21). Bed-nets remain the front-line tool used to tackle malaria contributing the most to the aversion of malaria cases achieved since the year 2000 (22). They are predicted to reduce malarial disease by up to around 50% (23; 24; 21). However, the use of bed nets and insecticides are vulnerable as they impose strong selective pressures on mosquito populations. Genetic resistance to many insecticides is widespread across Africa (9) and suggestions of behavioural changes which cause mosquitoes to favour day-biting when bed-nets are ineffective have been observed (25). Further advanced methods such as the development of genetically modified gene-drive mosquitoes which will flexibly suppress vector populations are also undergoing development, although the consequences of their use has been questioned by some (26).

Clinical case-management includes the treatment of malaria using anti-malarial drugs, the use of which is under severe threat due to the spread of resistance in malaria parasites (9). The impact that this has had on the *Pf* genome will be discussed in detail later. By far the most keenly sought after intervention is the development of an effective malaria vaccine, which the World Health Organisation (WHO) has proclaimed a core ingredient to an effective malaria control strategy (27). Vaccines act to reduce the likelihood or severity of infection by presenting pathogen-derived antigens to the recipient's immune system. This primes the host to more effectively combat the pathogen upon true infection (28). The sophisticated life cycle of the malaria parasite however complicates vaccine design, forcing a single life cycle stage to be targeted (17). The most progress has been gained for vaccines which target the pre-erythrocytic stages of infection (spanning from the entry of sporozoites during mosquito bite to the egress of merozoites from liver cells (Fig. 1.1 (pg.3))). Two vaccines, named RTS,S/AS01 (RTS,S) and R21/Matrix-M (R21), have both been designed to target a protein which coats the surface of sporozoites (circumsporozoite protein (CSP)), and are to date the only vaccines recommended for use by the WHO (17). During the phase III trial of R21, formulated in the adjuvant Matrix-M, it reached 68% efficacy against clinical episodes over the first 12 months at sites of perennial transmission and 75% at sites of seasonal malaria transmission (29). Very recently (July 2024) this vaccine has started to be deployed owing to the efficient large-scale production by the Serum Institute of India who have pledged to produce 100 million affordable doses per year (at less than 4US\$/dose) (30). Questions remain as to the longer-term efficacy of R21; protection is incomplete and is observed to wane (29). It is also yet unknown the response vaccine pressure may incur within parasite populations. RTS,S has been shown to have differential efficacy based on infecting parasite CSP genotype raising the possibility of resistance evolution (31). Genomic surveillance will be vital as these historic vaccines continue to be instituted.

Since most symptoms occur during the erythrocyte stages of parasite infection, this is an attractive stage to target the parasite (17). Success here has historically been more limited, with candidates targeting parasite merozoite surface protein (MSP)1, MSP3, apical membrane antigen 1 (AMA1) and erythrocyte binding antigen (EBA)-175 all showing minimal efficacy at trial (19). These ligands

are known to be involved in invasion pathways. MSP1 has been shown to interact with erythrocyte membrane protein Band 3, allowing attachment of merozoites to the targeted erythrocyte cell (32). Less is understood about the role of MSP3, but it appears to form a complex with MSP1 and other parasite surface proteins (33). AMA1 is not responsible for initial attachment but rather appears to mediate the orientation of parasites in preparation for invasion (18; 34). Actual invasion is mediated by two major families in *Plasmodium spp.*; the erythrocyte binding-like (EBL) family (also known as duffy binding-like (DBL)) and reticulocyte binding-like (RBL) family (35; 18). EBA-175 is an EBL type ligand which binds to the major erythrocyte surface molecule glycophorin (GYP)A (36). The disappointing potency of vaccines targeting these proteins may be due to their highly polymorphic nature and the redundancy of many of these invasion pathways (17). EBA-175, for example, is a dimorphic gene with both F- and C-types present at over 10% frequency within every available population in a large global dataset of *Pf* genomes (37). MSP1 exhibits diversity to such an extreme that it is routinely used to assess for levels of complexity of infection (COI) within samples (38; 39).

A promising erythrocyte stage vaccine target which appears to overcome these barriers is reticulocyte binding protein homologue 5 (RH5), a member of the RBL family. RH5 binds to the human protein basigin (BSG) with potent specie-specific affinity (40), and unlike EBA-175 is essential for invasion (41). RH5 is also highly conserved with only 5 common (at >10% allele frequency) non-synonymous single nucleotide polymorphism (SNP)s within its gene sequence (42). RH5.1/Matrix-M is currently undergoing a phase II trial in West Africa following the assertion that it is safe and well-tolerated, and the tantalising indication that parasite growth was inhibited during phase Ib (43). Recently published interim results (December 2024) showed an encouraging 55% efficacy under a delayed third-dose regimen, marking a significant breakthrough in the development of blood-stage vaccines (44).

A third stage of the parasite life cycle that might be beneficially impeded is the sexual phase (Fig. 1.1 (pg.3)). So called “transmission blocking” vaccines would procure the transferral of antibodies into the mosquito blood meal which target sporogonic-stage parasites preventing onward transmission by the vector (15; 17). Although these would not directly prevent disease symptoms, this vaccine

type may be required to eventually eradicate malaria. Although malaria vaccine development tends to favour infants who are the most affected demographically, asymptomatic infections in adults can act as reservoirs for transmission allowing infection to persist within the population (45; 17). Targeting ookinete surface proteins (e.g. Pfs25) and gametocyte antigens (e.g. Pfs230, Pfs48/45) has been attempted (15) and Pfs230 domain 1 has proceeded to phase II trial (17) having shown functional immunogenic activity at phase Ib (46).

1.1.4 The outlook for malaria control

Recent successes in the field of malaria vaccines have cultivated a bright outlook for malaria prevention and control. However, it is clear that global elimination is not yet within reach. The WHO expresses concern over biological threats such as antimalarial and insecticide resistance, as well as the potential for increased transmission periods and the expansion of climates suitable for competent vectors due to climate change (9). Funding for research and development within malaria has successively decreased since 2018, despite the call for novel and improved interventions (47; 9).

The study of the genomics of malaria, that is of humans, parasites, and mosquitoes, can be harnessed to bolster control efforts. It can aid in the assessment of current interventions, empowering their improvement and implementation, and importantly it may reveal novel targets with which to tackle malaria (48). In particular, efforts have been invested into discovering the human genetic factors which affect disease susceptibility and severity which have yielded vital insights.

1.2 An overview of human genetic analysis

1.2.1 Genetic variation

All genomes within a species are broadly similar. They contain the genes required for the organism's survival and function. However, there are differences between individuals, broadly termed genetic variation, which interact with other factors to contribute to differences in observable characteristics (phenotypes). Genetic

variation can take many forms, however the four main classes are: nucleotide substitutions (referred to as a SNP), short insertion or deletions (INDELs), structural variants (SVs) and larger chromosomal abnormalities (49). Assessing for these differences is fundamental to genetic analysis.

1.2.2 Candidate loci analysis

Much was discovered about the genomes of humans, mosquitoes, and malaria parasites through the use of ‘pre-genomic’ approaches (48). The basis of many of these methods was the polymerase chain reaction (PCR) which utilises naturally occurring enzymes to copy deoxyribonucleic acid (DNA), allowing the generation of large amounts of alike DNA molecules (50). Genetic variation can then be assessed, for example, by differences in amplicon size visualized on a gel, by designing a PCR assay that only amplifies DNA containing a specific variant, through the use of restriction enzymes that cut DNA depending on the presence or absence of variation, or by employing oligonucleotides that hybridize to amplified fragments containing a particular variant. These methods have all been used to determine the genotype of a sample at a specific locus (e.g., (51; 52; 39)).

To ascertain whether a human genetic locus impacts malaria risk, targeted genotyping followed by testing for differences in allele frequency between malaria cases and carefully matched controls has been commonly implemented, and many loci have been hypothesised to influence malaria risk and severity in this manner (11). However, because these assays are low throughput, such studies frequently suffered from small sample sizes and were vulnerable to confounding due to inadequate control over population structure. Population structure refers to systematic genetic differences between groups of individuals due to non-random mating throughout the population, potentially leading to false positive results (53). Efforts to overcome this were later attempted by collecting larger sample sets from geographically dispersed populations and comparing them (54). Only a proportion of loci originally hypothesised to associate with malaria were systematically replicated in this manner (6; 54).

1.2.3 Genome-wide analyses

A more systematic control of population structure can be achieved using genome wide association studies (GWAS). Rather than a focussed approach, genetic variation is assessed across the entire genome of an individual, allowing genetic relatedness within the set to be directly assessed and controlled for within statistical analyses. Differences in allele frequency between case and control groups, or across a phenotype range, are assessed at every typed locus, generating an impartial assessment of the genetic basis of a phenotype.

1.2.3.1 Array-based genotyping

Genome wide human genotyping is commonly performed using arrays. This in essence performs targeted genotyping but for many loci in parallel, revealing genotypes at a subset of genetic positions. Variation at untyped sites is then imputed using reference panels of complete human genomes, which is possible due to extensive linkage disequilibrium (i.e. shared inheritance) between nearby sites. However, this approach presented considerable challenges in malaria research. Within African populations who carry the brunt of malaria burden, blocks of linkage disequilibrium are generally shorter due to greater genetic diversity compared to European populations, and African genomes are poorly represented within available reference panels (55). By enriching reference panels with African genomes, the largest GWAS of malaria susceptibility in multiple populations to date was released in 2019 (6). This revealed an even smaller subset of loci to associate with malaria disease than previously suspected. Interestingly however, these loci explained only a small fraction of variation in malaria incidence indicating much is yet to be understood regarding the architecture of malaria susceptibility (6; 56).

1.2.3.2 Whole genome sequencing

The need for imputation is negated through sequencing most nucleotides within the genome (whole genome sequencing (WGS)). The first sequenced genomes were generated using shotgun sequencing. The basis of this method is that of Sanger sequencing, the early form of DNA sequence determination. Using this method, multiple copies of a single fragments are amplified by the PCR but

DNA synthesis is randomly terminated by the inclusion of fluorescently labelled dideoxynucleotide (ddNTP)s which produce a mixture of labelled DNA fragments of every intermediary input length. Fragments are then separated by length using electrophoresis and the identity of the final incorporated ddNTP determined by laser excitation, allowing the identity of the nucleotide present at each position to be determined. This method can be used to produce 1 kilobase (Kb) fragments of extremely high accuracy (>99.999%) (50). However, this is laborious, expensive and slow; famously the first human genome took 13 years and \$3 billion to generate (57). Modern methods have led to the dramatically cheaper, more feasible and faster generation of many WGS genomes which will be described next, although it should be noted that array typing is still heavily used within human genetic research due to its lower cost and often sufficient efficacy (58).

1.2.3.3 Short read sequencing

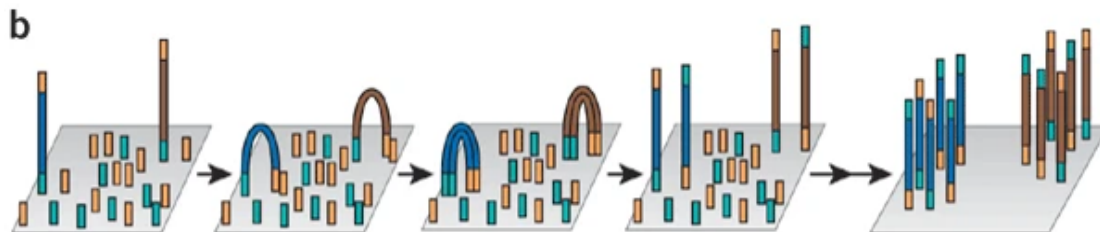


Figure 1.2: The process of Illumina-based cluster formation for whole genome short-read sequencing. In brief, small genomic DNA fragments are attached to the flow cell by an adapter sequence (green). The distal adapter (orange) anneals nearby to form a bridge. Anchored primers attach and the fragment is amplified via PCR. Complementary strands dissociate. The resultant 2 single strands enter further rounds of amplification to form clonal clusters. Each cluster is then sequenced simultaneously in the forward and then the reverse direction. Reproduced from (50) (licence number: 5856540201050)

A revolution in genomics research was initiated with the advent of Next Generation Sequencing (NGS) (48). This wave was led by the Illumina platform which is still heavily relied upon today. This method is able to perform massively parallel sequencing of short DNA fragments (50-300 base pairs (bps)) covering the entire genome producing high levels of genome coverage for many samples at once (at a lower accuracy rate compared to sanger sequencing) (59). The method begins with the preparation of sample genomic (g)DNA into fragments. Unique index

fragments which allow for the identification of sample of origin within resultant data and adapter sequences are appended to each fragment using PCR. The addition of adapters allows each fragment to anneal to fixed complementary adapter oligos anchored to the sequencing flow-cell to allow cluster formation (Fig. 1.2 on the previous page). Attached single stranded DNA (ssDNA) folds to attach its distal adapter to a nearby fixed oligo forming a bridge. The fragment is then amplified via PCR to produce double stranded DNA (dsDNA). Complementary strands are dissociated to reform ssDNA which enter new rounds of amplification forming local clusters of identical fragments. Following this step the nucleotide sequence of the forward strands of all fragments in each cluster is determined via the addition of fluorescently labelled ddNTPs. These are imaged at each incorporation (50). Once imaged, fluorescent labels are cleaved, and the termination of sequencing reversed allowing the addition of a further ddNTP in the next cycle. Each clonal fragment within a cluster is thus sequenced simultaneously. In paired-end sequencing, both the forward and reverse strands of each DNA fragment are sequenced. After the forward strand is sequenced, the newly synthesized strand is washed away. Then, similar to the process of cluster formation, the reverse strand is synthesized through bridge amplification and sequenced in the same way as the forward strand.

1.2.3.4 Long read sequencing

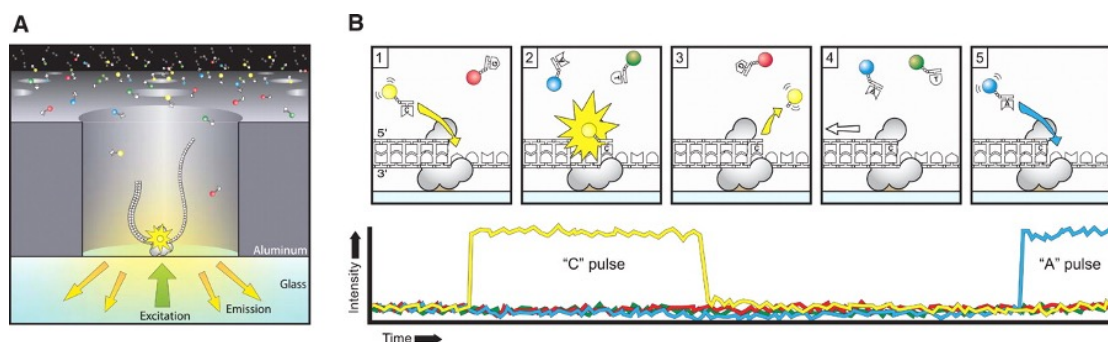


Figure 1.3: The method of long-read sequencing by PacBio **A.** A DNA fragment is sequenced by an anchored polymerase within a zero-mode wave guide which is excited by light below. **B.** fluorescently labelled ddNTPs are sequentially incorporated, and their identity determined by excitation. Borrowed from (60). Reprinted with permission from AAAS.

More recently, efforts have been made to sequence longer stretches of DNA. One such platform by Pacific Biosciences (PacBio) can sequence >10Kb fragments. This method, like the Illumina method described above, is based on sequencing by synthesis but differs in that each read reflects the sequencing of a single strand of DNA without the need for terminating nucleotides. The identity of incorporated bases are determined as they are sequentially added by a fixed polymerase molecule (Fig. 1.3 on the preceding page) (60). This allows fragments to be longer in comparison to the Illumina method, where incomplete reversal of termination or failed cleavage of fluorescent labels causes clusters of sequenced fragments to eventually fall out of phase, rendering fluorescent signals illegible (50).

A second method of long read sequencing, by Oxford Nanopore Technologies (ONT), is based on a distinct method of sequencing. Nucleotide sequence is determined via ionic fluctuations which occur as a ssDNA molecule passes through a microscopic pore. The pore is charged such that DNA (which is negatively charged) is drawn through it allowing sequencing to take place (61).

Compared to short read sequencing, long-read is less widely used due largely to prohibitive costs, low throughput ability and traditionally lower accuracy (62). However, accuracy has recently been dramatically improved for PacBio sequencing through HiFi technology which performs repeated sequencing of circularised fragments (63). ONT has also continually reported improved accuracy rates (61). Furthermore, the recently released PacBio Revio machine is able to perform long read sequencing in massive parallel which may reduce costs moving forward (64). A further discussion of the relative strengths of these technologies within malaria research will be made later in this chapter.

1.3 Human genetics and malaria disease outcomes

Malaria causes high rates of mortality in children, prior to reproductive age. This disease therefore imparts a strong selection pressure to evolve molecular mechanisms that protect against infection or disease, especially severe malaria and death (48). Malaria has consequently been considered the strongest evolutionary force acting upon the human genome (11). J.B.S Haldane first proposed that a fitness

advantage against malaria might underlie the distribution of certain haematological disorders, and the “malaria hypothesis” was shortly thereafter posited to explain high frequencies of HbS and other haemoglobinopathies in endemic regions (56). Since then, many human loci have been proposed to influence disease outcomes. These broadly categorise as affecting the erythrocyte, rendering them impedimentary to infection, or impacting the activity of the mounted immune response against the *Plasmodium* parasite (reviewed in (11)).

1.3.1 Blood-mediated resistance

1.3.1.1 Polymorphisms in the *HBB* gene

Many loci now widely accepted through GWAS analysis to impact disease risk influence erythrocyte biology (56). The primary role of erythrocytes is the transport of oxygen, facilitated by the haemoglobin protein. Haemoglobin is composed of two alpha-globin (HBA) and two beta-globin (HBB) chains, encoded by the *HBA1* or *HBA2* and *HBB* genes (65). Three non-synonymous SNPs in *HBB* are thought to confer a fitness advantage to malaria infected individuals: HbS, HbC and HbE (11). HbS in the heterozygous form (HbAS) is observed to induce a drastic reduction in disease odds in severe cases compared to non-infected controls (by 80-90%) (54). Homozygotes (HbSS) however are afflicted with the severe and often fatal disorder sickle-cell anaemia (11); the consequent heterozygote advantage at this locus causes HbS frequency to undergo forces of balancing selection, maintaining moderate frequencies (roughly 2-18%) across SSA (66). Despite its well-established connection to malaria, the major mechanism by which HbS protects carriers, if there is one, is not clear. Increased clearance of infected erythrocytes, a differential immune response, impaired *Pf* protein trafficking, and inhibition of parasite growth have all been suggested experimentally (56).

The connection between HbE with malaria is less robustly established. HbE displays a broadly opposing distribution to HbS; it is common in South-East Asia (SEA) but rare in SSA (11). Populations carrying the allele however are relatively understudied within malaria research limiting the drawing of firm conclusions (65). HbC on the other hand is common in central west-African countries including Ghana and Mali (67), and locally HbS is often comparatively rare in these groups

(11). A reduction in severe malaria risk of almost 30% in a large multi-population meta-analysis was observed within HbCC individuals, who may also suffer from at most mild anaemia (54).

1.3.1.2 Alterations to the production of alpha- or beta-globin

Thalassaemias, a group of related disorders, are caused by the defective production of alpha- or beta-globin chains, rather than changes to their amino acid composition (11). Disruption of one of the two homologous *HBA1* and *HBA2* genes is referred to as alpha+ thalassaemia, and this is thought to be protective against severe malaria while only inducing mild anaemia (11). However, because the molecular basis of this disorder can be structural, and because the underlying variation falls on numerous haplotype backgrounds, genotypes were inaccessible to the largest GWAS performed for severe malaria (6). Alpha+ thalassaemia has thus not been tested for effect on malaria risk on a wide scale (6). A large study of severe cases (N=2244) and controls (N=3949) from Kenya did indicate a protective impact of around 20% (68). Alpha-thalassaemia and HbS are both common with overlapping distributions (69), and there is evidence these genotypes interact to influence disease, adding further complexity to the dynamics of malaria risk in Africa (70; 71). Although beta-thalassaemia was originally proposed by Haldane to influence malaria susceptibility, in contradiction of the malaria hypothesis, beta-thalassaemia is relatively rare in African populations and thus evidence for its association is scant. *In vitro* experiments suggest beta-thalassaemia does not impede *Pf* invasion (72).

1.3.1.3 Erythrocyte cell surface molecules

The recognition, attachment, and entry of merozoites into erythrocytes via surface molecules has driven the evolution of high variation at these loci. These molecules serve as the interface between humans and microorganisms and are under strong evolutionary pressure to diversify (73).

Duffy antigen receptor for chemokines As a noteworthy example of how malaria imposed evolution can affect these ligands, most populations within Africa

are resistant to *P. vivax* infection due to the absence of a single surface chemokine receptor (known as the ‘duffy-negative’ blood group) (74). This is caused by a SNP in a transcription factor binding site for the gene which encodes the FY glycoprotein. This prevents the presentation of the Duffy antigen receptor for chemokines (DARC) protein on the erythrocyte surface that *P. vivax* necessarily utilises for entry (11). Fst values, a genetic measure of differentiation between populations, have been found to be at their highest value genome wide at the duffy-negative FY allele (74). No such polymorphism which confers near-complete resistance to *Pf* infection is known to exist (11).

The ABO blood group The RBC surface antigens denoted ABO also vary in their worldwide distribution to a significant degree (18). The variation underlying these blood groups in the ABO gene appears to be under long term balancing selection as it is found within multiple primate species, many with several independently evolved silencing O alleles (75). The A, B, and O blood groups are determined by the presence or absence of specific glycosyltransferases. These enzymes modify the H-antigen on RBCs: A-glycosyltransferase adds a sugar molecule to create the A-antigen, B-glycosyltransferase adds a different sugar to produce the B-antigen, and the absence of these enzymes in type O blood leaves the H-antigen unmodified. The single nucleotide deletion underlying the O group in humans has been found to decrease malaria disease risk by 26%, and the nearby non-synonymous variant underlying the expression of B-glycotransferase (and so yielding the B blood group) increased risk by 25% (54). However, disentangling the individual effects of these linked loci has traditionally been challenging. Modern statistical evidence suggests the effect is driven by the O group (6). The O-group may be protective due to a decreased propensity for rosetting, a phenomenon whereby infected erythrocytes bind to other uninfected RBCs obstructing the micro-vascular network and causing severe disease (76). Other mechanisms have been suggested, such as changes to cell structure, an impact on cytoadherence events or direct influence on invasion pathways (18).

Glycophorins Variation in genes encoding members of another group of erythrocyte surface molecule, the glycophorins, also show evidence of long-term bal-

ancing selection based on comparisons of human and chimpanzee genomes (77; 78). The distinct role of GYPA, B and C antigens in merozoite entry through junction formation with parasite DBL family EBA proteins suggests the prospect of resistance evolution at these loci (11).

GYPA and *GYPB* cluster alongside *GYPE* in a 400Kb region of chromosome 4. Variation across this region is challenging to resolve due to significant sequence paralogy and the presence of structural variation (78). Changes in coverage in high depth sequencing data were used to catalogue structural variants in the region, allowing the identification of a particular rearrangement (called DUP4), present only in East Africa, which included two copies of a GYPA/B hybrid gene. This was shown to be the molecular basis of the Dantu blood group phenotype, which was associated with reducing disease risk to one eighth of that of non-carriers in Kenya (78). Remarkably, despite the anticipation that inhibited junction formation might underly this resistance, increased tension of the erythrocyte membrane was identified as the primary factor limiting merozoite entry into Dantu blood cells (79).

1.3.1.4 G6PD deficiency

Another mechanism of malaria resistance is predicted to occur through the deficiency of the glucose-6-phosphate dehydrogenase (G6PD) enzyme, which plays a role in response to oxidative stress. The G6PD gene falls on the X chromosome, resulting in a single copy in genetic males. A SNP known to underly the majority of G6PD deficiency in Africa was found to be protective against CM but was a risk variant for severe malarial anaemia in male hemizygotes and female homozygotes (54). This is perhaps biologically plausible as G6PD deficiency itself causes haemolytic anaemia, potentially exacerbating this symptom (11). CM is a particularly damaging symptom with potentially long-lasting effects (80), perhaps allowing the evolution of a mechanism with opposing effects by subtype.

However, it was later suggested that the observed protective effect against CM might be due to statistical confounding through collider bias (81). Furthermore, in the largest GWAS to date there was only marginal evidence of an effect on the likelihood of SMA at the G6PD locus, bringing its connection to malaria further

into question (6). On the other hand, signals of positive selection at this locus have been detected, which might be suggestive of a role in malaria resistance (82).

1.3.1.5 Alterations of calcium flux

A surprising discovery emerged just over ten years ago. A population specific GWAS in Ghanaian samples indicated that a SNP proximal to the *ATPase plasma membrane Ca^{2+} transporting gene 4* (*ATP2B4*) was associated with protection against severe malaria (83), and an Africa-wide meta-analysis later showed a protective effect of 23% in homozygote carriers (54). Although initially hypothesised to effect *ATP2B4* expression through the disruption of an enhancer, it was later predicted that the associated variant instead prevents the expression of an erythroid specific *ATP2B4* transcript (84). The variant disrupts a GATA transcription factor binding motif at the transcription start site of this alternative transcript. This association was replicated in the large GWAS across multiple populations securing its connection to disease (6).

While the specific mechanism by which this influences malaria risk remains unidentified, it is hypothesized to relate to intracellular calcium concentrations because *ATP2B4* encodes the primary calcium channel in erythrocytes. It is noteworthy that *Pf* invasion induces a spike in intra-erythrocytic calcium in a process linked to the interaction between RH5 and BSG. This binding occurs with greater affinity when BSG is complexed with ATP2B4 hinting at a potential role for this protein in malarial disease (85).

1.3.2 Immune-mediated resistance

The impact malaria has imparted on variation of erythrocyte biology is evident. The development of acquired immunity observed in adults in endemic regions indicates the immune system also certainly plays a role in this disease. Bewilderingly, few examples of variation in known immune system genes have been confidently associated with malaria protection. For many resistance loci which relate to erythrocyte biology described above, the precise mechanisms of protection are not yet clear and may intersect with the immune system, but no clear relationship has yet been defined.

Many immune system genes, like those which produce other cell-surface receptors, are under impetus to evolve high levels of variation due to their interaction with infectious pathogens. This is exemplified by the major histocompatibility complex (MHC) locus on chromosome 6 which contains a high density of genes known to play a crucial role in immune system function (86). Although not expressed by erythrocytes, the human leukocyte antigen (HLA) genes found within the MHC locus are thus potential candidates for a relationship with malaria. Classical class I and class II HLA genes, *HLA-A*, *-B* and *-C*, and *HLA-DR*, *-DQ* and *-DP* respectively, display the most diversity with the MHC locus. Their role is to present antigens to T cells in order to facilitate an appropriate immune response (86; 87).

An example of an association of the HLA and malaria risk was the identification that HLA-B*53, a common West African allele of *HLA-B*, held significantly lower allele frequencies in severe malaria groups compared to non-malaria groups within Gambian infants (52). This was of significant excitement because it indicated that HLA-B*53 restricted antigens might be inductive of a protective immune response, which has implications for vaccine development. Further work went on to identify that peptides within liver stage antigen (LSA)-1 and -3 proteins expressed during the infection of hepatic cells, which express HLA-B, are recognised by HLA-B*53 restricted cytotoxic T-lymphocytes (88; 89). However, the original association has not been widely replicated since its discovery (90; 91).

This was similarly observed for many other immune loci thought to associate with malaria protection which were discovered based on targeted sequencing experiments in geographically isolated samples. This includes members of the interleukin and interferon families, and tumour necrosis factor (TNF) (11). An upstream variant of the CD40 ligand gene was found to significantly reduce malaria risk by almost a half in large multicentre meta-analysis, the only association of many immune genes tested (54). In the largest GWAS of malaria susceptibility to date however, no strong associations were discovered in the MHC or any other immune genes (6).

Within other major infectious diseases, such as human immunodeficiency virus 1 (HIV-1), Hepatitis B and C, and Tuberculosis, associations have been observed between immune loci and various disease traits (87). It seems improbable therefore

that there would be no similar genetic effects involved in malaria outcomes; this lack might explain some of the observed missing heritability in severe malaria susceptibility. Methodological artefacts might underly a lack of immune GWAS hits. Firstly, HLA genes and other immune loci, such as the immunoglobulins, T-cell receptors, and Killer immunoglobulin, are challenging to genotype using array-based typing or short read whole genome sequencing, especially within African populations for which appropriate reference panels used for imputation are often not available (55). The largest malaria GWAS however did use a dedicated HLA reference panel enriched for African genomes and so the HLA locus is considered to be accurately called to at least 2-digit resolution (6). Secondly, the tested phenotype used within GWAS has important implications for results. For example, in a GWAS analysis of HIV-1, the genetic correlates of viral load were tested within HIV-1 infected individuals (92). Within the malaria GWAS however, because all are likely to be exposed within endemic regions, cases and population controls were compared (6). It is perhaps the case that immune genetic factors play a finer role in affecting outcomes within infections than this analysis could detect. Malarial parasitaemia, unlike viral load in HIV-1 infections, does not correlate well with disease severity. This is perhaps because parasitaemia cyclically varies throughout the asexual stages of infection and the final measure depends on when the blood sample is taken (6; 92). Selecting medically relevant phenotypes to evaluate is therefore challenging. Thirdly, sample selection could underlie the lack of immune hits. The large malaria GWAS was conducted in infants but perhaps immunity may be more appropriately assessed in adults who have had sufficient time to develop acquired immunity.

A potential major contributor to the lack of immune genetic GWAS associations, and a small number of GWAS hits overall, is confounding by parasite genetic factors. The *Pf* genome, compared to for example viruses, is large and shows vast genetic diversity, and thus the impact of pathogen genetic factors on GWAS may be pronounced potentially hiding true signals completely. The complexities of interactions of highly diverse human and highly diverse parasite loci might be inaccessible if their genetics is considered in isolation. Further to this, it has become increasingly apparent that immunity to malaria is highly complex, involving intricate humoral and cellular responses (93). No singular immunological response

has reliably been associated to malaria protection (11). An integrated view which considers both human and parasite genetics on a large scale might reveal more about the acquired immune response to malaria allowing it to be more effectively harnessed via vaccination. To facilitate combined analyses, precise sequencing of malaria parasites is essential, ensuring robust and well-powered findings can be drawn.

1.4 Genomics methods for malaria parasites

1.4.1 Short read sequencing

The first complete *Pf* genome sequence was released in 2002, consisting of 14 chromosomes and an apicoplast and mitochondrion genome, generated via shotgun sequencing (94). Because the *Pf* genome is comparatively small in relation to the human genome (roughly 23 megabases (Mbs) compared to 3 gigabases), WGS by the Illumina method can be performed relatively cheaply. However, the reliance of this method on PCR poses a problem for *Pf* genome sequencing. It is well established that a strong correlation between genome coverage and guanine-cytosine (GC)-content exists. This is because PCR is inherently GC biased. It is not well understood what underlies this, but it is proposed to be due to the differential melting temperatures of DNA dependent on GC content. Three hydrogen bonds bind GC nucleotides in comparison to two between adenosine-thymine (AT), causing AT-rich DNA to be less thermostable (94). The *Pf* genome is particularly AT-rich with a high distribution of GC content within genes which vastly compounds this bias, leading to uneven genome coverage (94). Due to this, PCR-free methods of library preparation have been successfully developed for *Pf* genomes (95). Since then further reformulations of reagents and the minimisation of PCR cycles have been implemented (96) such that PCR based library preparation kits have now been used to sequence *Pf* genomes successfully (e.g. (97; 98)).

1.4.2 Large scale short read sequencing of *P. falciparum* genomes

Once the ability to generate parasite genome sequences was established, a further challenge was to isolate parasite DNA (48). Although erythrocytes contain no human DNA, within an infected blood sample there is nevertheless typically 1000-fold more human DNA due to the presence of immune cells which contain the comparatively much larger human genome (48). Sequencing venous blood derived DNA for the goal of *Pf* genome analysis is therefore limited to high parasitaemia infections, and even in this case a vast majority of reads will be human origin which, if not desired, is wasteful. Other methods have included culturing parasites in erythrocytes, allowing pure *Pf* DNA to be obtained. This is however not feasible on a large scale and poses the risk of genetic changes due to culture adaption (48). A common method therefore has been to selectively remove human immune cells from blood samples using columns (“leucocyte depletion”) prior to sequencing (99). Although effective, this requires equipment which may not be available in remote regions of the world where malaria is endemic (99).

These issues were solved with the advent of a new technique which is able to selectively amplify parasite DNA from mixed sources (selective whole genome amplification (sWGA)) (99). This technique even allows dried blood spots (DBSs) to be utilised as input, negating the need for blood draws entirely (99). This method relies on the extraordinarily AT-rich genome of the *Pf* parasite. Primers which solely contain AT nucleotides will preferentially anneal to parasite DNA within a mixed source, allowing the PCR amplification of *Pf* derived DNA (99). This technique has allowed the generation of large global *Pf* WGS datasets such as the Malaria Genomic Epidemiology Network (MalariaGEN) Pf7 resource which contains over 20,000 samples, the majority of which were generated from DBS using sWGA (37). This method however has a downside: sWGA has been found to produce uneven genome coverage, exacerbating already present GC-biasing in sequence data (48). This leads to the loss of certain genomic regions which are of potential interest.

Short read data is typically examined by aligning sequenced fragments onto a reference genome assembly. The low complexity and highly repetitive nature of

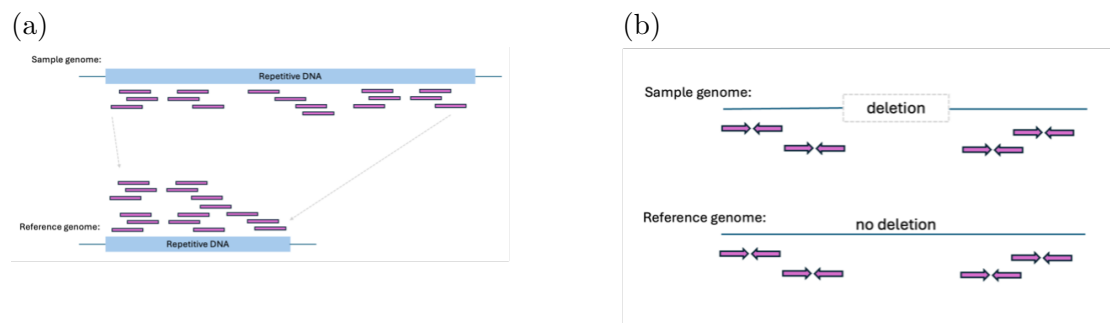


Figure 1.4: Challenges in resolving repetitive sequence and structural variation using short read genomic sequencing data. For example, **a.** reads (purple) sequenced from multiple locations in a repetitive region (blue) of a sample genome align to single locations in the reference. **b.** a deletion in a sample genome does not produce sequence reads (purple). If sequencing depth is not sufficient, discriminating this from low coverage is difficult.

the *Pf* genome introduces challenges to this method. For example, inference of the exact order or length of highly repetitive sequences is not simple to achieve (Fig. 1.4a), nor is it easily possible to deduce the presence of structural variation (Fig. 1.4b) (100; 101). The identification of variation within *Pf* genomes using short read data is often limited to what are called the ‘core’ regions for this reason, which comprise only about 90% of the total genome (102).

1.4.3 Long read sequencing

Many of the limitations of short-read WGS can be overcome using long read sequencing. Using these methods, there is no need for a reference sequence as overlapping fragments can be pieced together to generate full length chromosomes (“*de novo* assembly”). This has been used to generate multiple new *Pf* reference assemblies from a diverse range of clinical isolates (103). Within this data telomeric and repetitive sequence can be analysed, and structural variation observed directly.

Nanopore technology has also been employed on malaria parasites. The major upside of this technology for *Pf* sequencing is the portability of the sequencing device (the “MinION”) which allows it to be utilised in low resource settings such as parts of SSA (104). Comparisons of these platforms in whole genome sequencing a rodent *Plasmodium* strain indicates that Pacific Biosciences (PacBio)

HiFi assemblies are currently more accurate and cover more of the genome (63). It is yet unclear how well these methods will be able to produce whole genome sequence data following sWGA DNA preparation which may interfere with fragment integrity which is vital for the success of long read quality. Should this be investigated, and the recent development of high-throughput potential of PacBio sequencing capitalised upon, the production of large numbers of sequences which span the entire parasite genome might be realised.

1.5 Parasite genetics and malaria disease outcomes

Using genome sequencing to uncover how malaria parasites overcome selective forces, as has been achieved in the analysis of human genomes described above, has already revealed important insights which can inform malaria control.

1.5.1 The population genetics of *P. falciparum*

Understanding the population genetic features of *Pf* is vital to disentangling what parasite loci may be clinically important. The unusual life cycle of *Pf* parasites makes them particularly well suited to counter-evolve against measures used to control them. The vast amounts of mitotic division during the asexual phases of infection, coupled with a moderate mutation frequency of an estimated 1×10^{-9} per base per division, allows the generation and rapid selection of fit mutations (105). Only a small number of parasites within an infection go on to develop into gametocytes, and only a proportion of these will be ingested by a competent vector. Even fewer may successfully develop into oocysts in the mosquito midgut wall and go on to produce sporozoites which will infect a new human host. These sequential bottlenecks mean that the effects of stochastic fluctuation in allele frequency (genetic drift) are very strong, allowing the rapid fixation (or loss) of alleles (106). Therefore, despite a large *Pf* population size in endemic regions, *Pf* evolutionary dynamics mirrors some features of much smaller populations. In addition, the fact that *Pf* parasites do obligately sexually reproduce may add

an evolutionary advantage. In endemic regions where opportunities for outbreeding are common, beneficial mutations are able to recombine onto other haplotype backgrounds, which holds evolutionary implications (20; 107). This has been exemplified by the spread of anti-malaria drug resistance (DR) which is discussed further below.

Transmission intensity also affects the evolutionary dynamics of malaria parasites. In regions of seasonal transmission large population bottlenecks periodically occur, pronouncing the effects of genetic drift (7). Consequently, *Pf* populations in regions of low or seasonal transmission are generally more inbred compared to endemic year-round transmission areas. Low transmission will also decrease levels of acquired immunity and, if longstanding, alter the genetic features underlying susceptibility of the host population. These in turn affect the regional reliance on therapeutics, altering selective pressures exerted on parasites (106; 105). These factors are thought to contribute to why anti-malarial DR generally appears in SEA where transmission rates are low. Relatively unfit mutations like those relating to DR, which generally hold a fitness cost, may also be able to spread within these regions as within-host competition between parasites is less common (105).

High endemicity on the other hand means that infections often contain multiple genetically distinct parasites and so within-host competition is pronounced. Parasite fitness is composed of within-host survival and between host transmission, and thus the relative investment into asexual and sexual life cycle stages is adaptively controlled (20). Parasites attempt to balance the trade-off between these facets by the sensing of environmental stress and genetic diversity within the infection in order to maximise fitness.

Based on principal component analysis, a method used to assess structure, African *Pf* populations exist as a group of sub-populations. Locally however structure is often not pronounced; this appears to depend on transmission intensity and seasonality as well as the co-transmission of other malaria species (108; 7). Gambia for example has experienced a significant reduction in transmission intensity over the past two decades changing from year-round to seasonal transmission; *Pf* populations here have accordingly become more inbred and structured (7). Ethiopian parasites are also divergent from other African populations based on genetic clus-

tering analysis, and this may be to do with the co-transmission of *P. vivax* in this region (108).

High transmission areas hold additional features which complicate the efficacy of intervention strategies. The effective pseudo-polyploidy of complex infections means they are robust to environmental changes such as the use of antimalarial drugs (105). In regions with large numbers of individuals with acquired immunity, who are prone to carrying asymptomatic infections, these act as a reservoir for diversity, allowing the persistence of alleles (106). It is evident that the evolutionary dynamics of malaria parasites are complex, with biological and population factors playing a role.

1.5.2 Evidence of evolution in the *P. falciparum* genome

1.5.2.1 Anti-malarial drug resistance

One area which has displayed the adaptive capacity of *Pf* parasites is the development of anti-malarial DR. This has been revealed to us using large-scale parasite genome sequencing. Chloroquine (CQ), discovered in the mid 1930s, soon became the frontline antimalarial drug because of its low cost, safety and high efficacy (109; 105). Unfortunately, resistance rapidly emerged; genetic crosses between resistant and sensitive isolates indicated a gene on parasite chromosome 7, thereafter named chloroquine resistance transporter (CRT), as underlying resistance. CRT produces a transporter protein in the membrane of the digestive vacuole (DV) in infected erythrocytes. The DV is the active site of CQ which inhibits the detoxification of heme, a byproduct of infection, which occurs there. Amino acid substitutions in the CRT protein enable the more effective expulsion of CQ from the DV preventing its activity (105). Multiple SNP mutations within CRT gene have been shown to associate with CQ resistance, but the major molecular marker which occurs in almost all resistant field isolates is a nonsynonymous change procuring K76T (110; 111). This mutation, originating at least 4 times, spread rapidly across Africa from its origins in SEA (106).

Temporal *Pf* WGS in the Gambia recently revealed the coincidental allele frequency increase of a SNP in an amino acid transporter 1 (AAT1) gene on parasite chromosome 6, which showed high linkage disequilibrium (LD) with the K76T

SNP and was a quantitative trait locus for CQ resistance in a cross experiment (112). AAT1 also sits in the DV membrane and is likely to also mediate the efflux of CQ. The possession of the S258L variant in AAT1 and K76T in CRT was found to be relatively unfit compared to the CQ sensitive alleles in the absence of drug pressure (112). In accordance with this, Malawi has shown a rapid decrease in the K76T mutation following the cessation of CQ usage (113). It is yet unclear whether S258L follows the same pattern. The purging of these alleles from populations will depend on their individual fitness costs, time since CQ usage was stopped, and population transmission factors outlined above.

Two further drugs, sulfadoxine and pyrimethamine (SP), target the dihydropteroate synthetase (DHPS) and dihydrofolate reductase (DHFR) enzymes involved in folate synthesis (114). Clinical resistance through multiple non-synonymous SNPs within the hydroxymethyldihydropterin pyrophosphokinase (PPPK)-DHPS and DHFR genes emerged within a year of roll out (105). Amplification, rather than amino acid substitution, has also been shown to underly drug resistance. For example the duplication of multidrug resistance 1 (MDR1) on parasite chromosome 12, an adenosine triphosphate (ATP) binding cassette which also sits on the DV membrane (115), has been shown to associate with mefloquine, lumafentrine and artemisinin (ART) resistance (116). SNPs in MDR1 have also been connected to CQ resistance (105).

Resistance to ART is of major current concern because it is the core component of current front line antimalarials. ART was first used as a Chinese medical remedy as early as 340AD (via the plant *Artemisia annua*), but despite this its mode of activity has not yet been defined (105). The appearance of resistance mediated by the kelch protein K13 gene which appear to lengthen the ring stage and shorten the trophozoite stage of infection indicates ART's action may be stage specific (117).

The positive selection of DR has had a profound impact on *Pf* genomes across SSA. Genomic regions that are identical by descent (IBD) at the CRT, AAT1, DHFR, PPPK-DHPS and MDR1 loci can be observed between geographically distant and otherwise unrelated populations (108). Surrounding these causal SNPs are stretches of shared neutral variation which spread through genetic hitchhiking causing long flanking regions of elevated LD (106). This highlights the importance

of genomic surveillance in intervention implementation and in the progress to the goal of eventual malaria elimination (118). Concerningly, ART resistance mutations appear to have spread from their origin in SEA to regions of Africa which requires close monitoring (119; 120). There is hope that lessons learned from the spread of other DR alleles will inform the limitation of this precarious situation (121).

1.5.2.2 Parasite-origin erythrocyte surface molecules

It is not solely man-made pressures that have shaped the *Pf* genome. Like humans, parasites have evolved mechanisms which benefit their fitness in the face of human resistance. One element of this which holds clinical significance is the mechanisms of immune evasion parasites employ. Mature infected erythrocytes are “sticky”; they’re able to cytoadhere to host vasculature preventing their clearance by the spleen (122). Within the spleen erythrocytes are passed through inter-endothelial slits of 4 micrometer diameter. Because *Pf* parasites extensively physically remodel erythrocytes in order to facilitate infection, infected RBC will not pass through, leading to their envelopment and destruction by immune cells (123). Three major classes of parasite surface antigen facilitate cytoadherence: subtelomeric variable open reading frame (STEVAR), repetitive interspersed family (RIFIN), and *Plasmodium falciparum* erythrocyte membrane protein 1 (*PfEMP1*) proteins. These are trafficked out of the compartment within the RBCs which contains the infecting parasites (the parasitophorous vacuole), to the surface of the erythrocyte during the asexual stages of infection.

The best studied of these are *PfEMP1*s, encoded by the VAR gene family. Each VAR gene consists of two exons: the first encoding the extracellular portion and the second the transmembrane and intercellular domains (122). Because of its exposure to the host immune system, the first exon is highly diverse, containing between 2 and 10 DBL domains interspersed with cysteine rich interdomain regions. The majority of VAR sequences are found in the subtelomeric genome regions. These are delineated into group A and group B VAR genes according to the direction of transcription. A smaller group of C VAR genes exist in the internal chromosomes. *PfEMP1*s can bind to many different host adhesins in various tissues, which is thought to contribute to different manifestations of disease. For

example, binding to the endothelial receptor intracellular adhesion molecule-1 has been connected to sequestration in cerebral vasculature, which is a feature of CM (124). Binding of *Pf*EMP1 to complement receptor 1 is thought to contribute to rosetting whereby infected and non-infected erythrocytes clump together, also thought to lead to severe symptoms (125).

Additional complexity underlies VAR gene arrangement and expression. This feature exemplifies an additional premier trait which allows *Pf* to evade host-mediated destruction: antigen switching. Roughly 60 complete VAR gene sequences capable of producing a *Pf*EMP1 protein exist within the *Pf* genome. Despite an extensive repertoire, only one VAR gene is expressed by a parasite. This allelic exclusion prevents unneeded exposure to the host immune system (126). Expression is switched to another VAR gene at a very low rate, such that if the host develops an effective immune response to the major *Pf*EMP1 allele, a resistant population can expand. Suppression of all but one VAR gene is achieved epigenetically, and these controls are inherited by daughter merozoites ensuring singular expression is maintained (126).

Despite a clear relation to virulence, discovering replicable links between specific *Pf* surface antigens and clinical outcomes has proven challenging. The expression of A and B group VAR genes, assessed via quantifying ribonucleic acid (RNA) transcripts via PCR, has been associated with severe malaria (127), but this focussed view is potentially prone to false positive associations. Genetic studies have suggested there is preferential binding of *Pf*EMP1s based on variation in host ligands, and it is probable that a combination of parasite and host genetics thus act to influence outcomes (128). Further complexity is added when STEVOR and RIFIN are considered, for which there are about 30 and 150 genes respectively (129). The effects of the arms race between parasites and humans are evident in both genomes and incorporating sequencing information for both species may be vital to understanding these interactions. Unpacking this requires precise sequencing methodology and careful statistical analysis.

1.6 Emerging evidence of host-parasite coevolution

Sequencing studies which take whole human and parasite genomes into account are rare but yield important results. Recently, four loci within the malaria parasite genome have been identified that exhibit strong associations with the HbS allele (130; 131). As discussed in an earlier section, this polymorphism causes sickle cell anaemia when homozygous but provides substantial protection against malaria in its heterozygous form. The first three loci to be discovered, named *Plasmodium falciparum* sickle associated (*Pfsa*)1-3, were revealed in two sets of severe cases collected in Gambia and Kenya for which whole genome *Pf* and human sequencing was performed (130). Distributions of alleles for all *Pf* SNPs were assessed across a set of candidate human loci which had previously been associated with malaria to simplify analysis and limit computational load. Recently, *Pfsa4* was discovered in a set of mild cases from Ghana based on targeted HbS typing for a set of *Pf* WGS data (131). This analysis also replicated the associations at *Pfsa1* and *Pfsa3*. *Pfsa2* is restricted to East African parasite populations so is not present in Ghana. *Pfsa4* however did not replicate in the original discovery sample sets but did in an independent sample set of Mali which sequenced RNA and DNA of natural infections in a small number of HbAS and HbAA individuals. This suggests there is spatial variation in these *Pfsa* signals.

Based on the discovery study, carrying the first three *Pfsa* alleles completely obviated the protective effect of carrying an HbS allele, suggesting parasites have evolved at these loci to overcome HbS-mediated resistance. Discovering by what mechanisms parasites achieve this may reveal a vulnerability in parasite biology that might be therapeutically targeted, as well as elucidate what underlies HbS-mediated resistance, which is broadly unclear (56). The strongest associations at all four *Pfsa* loci were at non-synonymous variants in the genes: acyl-CoA synthetase (ACS)8 (on parasite chromosome 2), PF3D7_0220300 (on chromosome 2), PF3D7_1127000 (on chromosome 11) and serine/threonine protein kinase (FIKK)4.2 (on chromosome 4). ACS and FIKK are both expanded gene families in the *Laverania* subgenus (which contains *Pf*) compared to other *Plasmodium spp.* (132). The genes at *Pfsa2* and *Pfsa3*, like many genes in the *Pf*

genome, have no known function, but similarly to ACS8 and FIKK4.2, do contain *Plasmodium* export element (PEXEL)-like motifs. This suggests they produce proteins that are exported out of the parasitophorous vacuole and into the host cell cytosol during the asexual stages of infection.

Parasites carrying the *Pfsa3* SNP were also found to carry structural variation within the region in several long-read sequence assemblies. A laboratory parasite isolate called Ugandan Palo-Alto (FUP-H), which carries *Pfsa1-3* alleles, showed significantly increased expression of PF3D7_1127000 (the gene at the *Pfsa3* locus) compared to the 3D7 isolate which does not carry the *Pfsa+* alleles or structural variation at the *Pfsa3* locus. Not much more is known about these proteins, and experiments to elucidate their cellular function have thus been set in motion.

A further fascinating feature of these loci is the distinct levels of LD observed between them, even though they predominantly fall on separate chromosomes. This is surprising because recombination occurring during meiosis within the mosquito stages of infection acts to randomly assort chromosomes, such that if a micro- and macrogamete derived from parasite lineages which differ at these loci fuse (out-crossing), between chromosome LD should be broken down. If this feature was caused by elevated levels of inbreeding, we might expect between chromosome LD to appear at other loci in the genome. However, Pearson's R, a measure of correlation, between *Pfsa1* and *Pfsa2* (on parasite chromosome 2) with *Pfsa3* (on parasite chromosome 11) was 0.81 and 0.67 respectively in Kenya, falling at the extreme of the distribution of all between chromosome LD. Elevated LD between these loci is also observed in all other so far assessed African populations.

A pair of loci which hold similar levels of between chromosome LD in the Gambia are the antimalarial DR loci CRT and AAT1, which confer resistance to CQ. Drug pressure is known to give a significant selective advantage to parasites carrying these alleles, removing diversity at these loci and generating LD. A strong possibility is that the presence of the HbS allele applies a similar selective pressure onto *Pfsa* loci, however it seems unlikely that this should be so strong as to generate the observed LD given HbS is rare. The unusual evolutionary dynamics of *Pf* parasites described above, such as extreme population bottlenecks throughout the life cycle, as well as reservoirs in asymptomatic infections which are common in HbAS individuals (133), could contribute to this outcome although this

is as yet unclear. Another possibility is that the *Pfsa* loci are epistatically linked, such that parasites of intermediary *Pfsa* genotypes are unfit and are purged from populations. Whether epistasis is required to produce this pattern is unknown and should be experimentally investigated.

1.7 Summary

It is clear that both human and parasite genetic factors affect disease and it is likely that they interact in ways we have not yet ascertained. If we can understand the molecular mechanisms underlying these interactions, we might tackle malaria more effectively. Advances in sequencing technologies, particularly in methods to sequence large numbers of malaria parasites, means we are in a greater position to achieve this than ever before. This endeavour requires the selection of appropriate sample sets with relevant disease phenotype information, the election of appropriate methods, the optimisation of these for accurate parasite and human sequencing, and the performance of well-controlled statistical comparisons in order to glean clinical insight. Foremost, further investigation of the potentially clinically important *Pfsa* loci should be undertaken to assess the possibility of exploiting this for human advantage in the battle against the disease malaria. As this field advances, how the genetics of the immune system impacts malaria and how natural immunity thus develops may finally be exposed. This could have important implications for future vaccine and anti-malarial drug design.

1.7.1 Structure of this thesis

I will therefore investigate parasite genetics throughout this thesis across four results chapters. In the first results chapter (Chapter 3 (pg.87)), I will describe the generation of a new sequencing resource from The Gambia called the Genetic Analysis of Malaria Cases and Controls (GAMCC) set. The GAMCC was collected from severe and mild cases in infants allowing an analysis of the impact of parasite genetics on disease severity, and consists of *Pf* WGS and targeted human genotyping at the *HBB* and HLA loci allowing an assessment for host-pathogen genetic interactions. Hence, as indicated in Table 1.1 on the following page, I will use

the GAMCC set in conjunction with other publicly available datasets to answer specific questions in the two following results chapters. These datasets include a set of severe cases collected in Gambia and Kenya referred to as MalariaGEN Consortial Project 1 (CP1) (130), and the MalariaGEN Pf6 and Pf7 resources of globally sampled community *Pf* infections (37). I will also present a fourth results chapter containing an investigation of structural variation at the *Pfsa3* locus performed using public datasets alone.

Chapter	Question	Datasets used
Chapter 3 (pg.87)	How can we optimise <i>Pf</i> whole genome sequencing for the historical GAMCC set?	GAMCC, CP1
Chapter 4 (pg.110)	Does parasite genetic variation influence the severity of malaria infection?	GAMCC, CP1, Pf7
Chapter 5 (pg.148)	Does the HbS- <i>Pfsa</i> interaction replicate in The Gambia at an earlier timepoint?	GAMCC, CP1, Pf7
Chapter 5	Did parasite variation confound the observed strong protective effect of the HLA-B*53 allele against severe malaria?	GAMCC
Chapter 5	What does population genetic analyses tell us about the <i>Pfsa</i> loci?	GAMCC, CP1, Pf7
Chapter 6 (pg.225)	Is the <i>Pfsa3+</i> allele really linked to a structural variant? Can we map the haplotype structure and predict function?	CP1, Pf6

Table 1.1: **Research questions and datasets used in each chapter.**

Chapter 2

Materials and Methods

2.1 Description of data used within this thesis

2.1.1 The Genetic Analysis of Malaria Cases and Controls cohort

These samples, which for the purposes of this work we have named the Genetic Analysis of Malaria Cases and Controls (GAMCC) cohort, were originally collected with a view to comparing human leukocyte antigen (HLA) class I and class II allele frequencies in severe *Plasmodium falciparum* (*Pf*) malaria cases, non-malaria severe and mild infections, and healthy adults (52). The mild and severe malaria cases, which are focussed on in this work, were collected in children up to the age of 10 from Royal Victoria Hospital (Banjul, The Gambia) and the Medical Research Council (MRC) Hospital (Fajara, The Gambia) during months of high malaria transmission (August to November) in 1988-90 (52) (further details of phenotyping are given later in Section 2.7.1.1 (pg.58)). Within Chapter 3 (pg.87) of this thesis I describe the work I carried out to produce genetic data for these samples. Briefly, *Pf* deoxyribonucleic acid (DNA) within these samples underwent selective whole genome amplification (sWGA) (see Section 2.2.1 (pg.35)) followed by whole genome sequencing (WGS) (see Section 2.3.1 (pg.40)). In addition, targeting human genotyping was performed (Section 2.4 (pg.42)).

2.1.1.1 Ethics statement

The genetic sequencing of human and parasite derived DNA preserved in these samples was approved by the London School of Hygiene and Tropical Medicine observational/intervention research ethics committee in May 2022 (Reference: 26354) and the Oxford Tropical Research Ethics Committee.

2.1.2 The MalariaGEN Consortial Project 1 (CP1) cohort

The Malaria Genomic Epidemiology Network (MalariaGEN) Consortial Project 1 (CP1) samples were collected for the purpose of assessing human genetic effects on severe malaria disease susceptibility in a large, geographically widespread sample (6; 54). These were gathered from multiple partner sites across Africa and Asia from infants with severe malaria infections (54) and analysed for human genetic variation. As part of a later project, malaria parasite genomes from the Kenyan and Gambian contingents of this cohort were sequenced (6). These were either sequenced following sWGA (like those in the GAMCC) or using high coverage WGS on whole blood DNA extractions (referred to as the WGS CP1 sample type) generating large amounts of human and parasite reads.

2.1.3 The MalariaGEN Pf7 cohort

MalariaGEN Pf7 is a very large (>20,000) dataset of *Pf* genomes which, similarly to the CP1 cohort, was gathered from multiple partner sites across Africa and Asia. This cohort was composed in order to empower *Pf* genomics research and give an improved understanding of *Pf* population structure and diversity (37). Therefore, this and a previous, smaller release of it (called Pf6) have been utilised to assess multiple facets of *Pf* genomics. Earlier releases were generally produced from blood samples which had been leukocyte depleted, but Pf7 contains a large proportion (>60%) of genomes generated using sWGA from dried blood spots (DBSs).

2.1.4 Long-read assemblies

Publicly available long-read assemblies used in this thesis were produced from laboratory and clinical isolates by Otto et al. (103). A number of additional assemblies were also newly generated during this research; used in this thesis is a clinical isolate from Kenya referred to as Kenyan isolate (KENI)1, and an assembly generated from a laboratory isolate Ugandan Palo-Alto (FUP-H). Cultured DNA for these samples was obtained from the KEMRI institute (Kilifi, Kenya) and BEI Resources respectively. Details of the sequencing and assembly of these new references are detailed in Sections 2.3.2.1 (pg.41) and 2.5.4.1 (pg.50).

2.2 Preparation of samples for *Pf* DNA sequencing

2.2.1 Initial preparation of samples in the GAMCC cohort

Two sets of aliquots of this cohort were available to the present study, referred to as the Dominic Kwiatkowski (DK) and Adrian Hill (AH) sets (reflecting the principal investigators of the laboratories in which the samples were stored). The DK set is believed to have been aliquoted from the original AH set and kept at -80°C away from general laboratory space. The AH set has been kept at -20°C since its collection and is known to have been used in multiple studies. The DK set was therefore prioritised for this work due to the decreased likelihood of contamination and sample degradation.

Aliquots of the whole blood phenol-chloroform extracted DNA were collected from the freezer and organised into a 96-well plate plan in numerical order. Depleted or evaporated samples were rehydrated using 12.5 microlitres (μLs) of nuclease-free H_2O and kept at room temperature overnight to recover any DNA which may still be available within tubes. Photographs of tubes organised into plans were taken in order to be able to verify sample IDs, should the occurrence of a mix up be indicated during the study. Available database files indicated the existence of 984 malaria case samples in the GAMCC cohort. Of this number, 906 were found in the DK set and a further 14 were found in the AH set. 64 were not found in either set.

DNA concentration was assessed using a Quant-iT PicoGreen dsDNA Assay (Invitrogen, P7589) following the manufacturer's instructions. 100 nanograms (ngs) of DNA was then normalised into a volume of 20 μ Ls using nuclease free water. If the PicoGreen assay indicated that 100ng was not available within tubes, then 2.5 μ L was taken forward for human DNA depletion.

2.2.2 Depletion of human DNA

Samples within the GAMCC collection were generally very highly concentrated. Because these individuals were unwell, potentially leading to an elevated white blood cell count, and because a significant proportion of cases in the collection were mild malaria cases (and therefore may have had lower parasitaemia measurements (52)), it is likely that human DNA makes up a large proportion of the mass within each sample. The DNA of *Pf* parasites, like all eukaryotic genomes, can be methylated at cytosine bases, but this occurs much less so than within the human genome (134). Because of this, human DNA can be specifically removed or degraded in samples to enrich *Pf* genomic DNA by depleting 5-methylcytosine (5mC)-methylated DNA. Two methods of this were trialled as part of this thesis: using the enzyme McrBC and via a bead-based method, both described below. Following this trial (described in Chapter 3 (pg.87)), McrBC digestion was performed for all samples in the GAMCC cohort. Where multiple sWGA conditions were tested on the same sample as part of the trialling process, depletion was performed as a pool and divided between different conditions.

2.2.2.1 McrBC

5mC-methylated DNA was specifically digested using the endonuclease McrBC (NEB, M0272) as previously described (98). 100ng of DNA, 10units of McrBC, 1 $\times$ NEBuffer2, 0.5 μ L of 100 $\times$ bovine serum albumin (BSA), and 0.5 μ L of 100 $\times$ guanosine triphosphate (which are included with in the NEB kit) were combined in a 30 μ L reaction and heated at 37°C for 120 minutes, followed by 20 minutes of enzyme inactivation at 65°C.

2.2.2.2 Bead enrichment

A method of enrichment for *Pf* DNA was trialled for a small subset of samples using the NEBNext Microbiome DNA Enrichment Kit (NEB, E2612S) following the manufacturer’s instructions. 400ng was enriched as a pool using this method, so as per the protocol, 64 μ L of Protein A magnetic beads was utilised per sample. Beads were mixed with the sample and incubated at room temperature for 10 minutes on a rotating mixer to allow DNA to bind. Beads were washed three times using 1 x Bind/wash buffer at room temperature for 3 minutes, and the buffer removed through the use of a magnetic rack. Following the third wash the supernatant, which contains unmethylated DNA in the sample, was retained. This was cleaned using the DNA Clean & Concentrator-10 kit (Zymo Research, D4011) following the manufacturer’s instructions prior to input into WGA.

2.2.3 *Pf* Whole Genome Amplification

Once human DNA had been partially removed using the above methods, WGA was performed using parasite genome selective primers (sWGA) or using random hexamer primers (random whole genome amplification (rWGA)). sWGA was eventually performed on the entire GAMCC set.

2.2.3.1 Selective Whole Genome Amplification (sWGA)

sWGA reactions were performed as previously described (99). Half ($\sim$ 50ng) of the resultant McrBC digest or bead enriched supernatant was normalised using nuclease free water to 30 μ L. One of two previously published primer sets were used for sWGA: the 10A (Table 2.1 on the next page) set alone, or a combination of the sets 10A and 6A (Table 2.2 on the following page) (99; 135).

An sWGA master mix was added to the sample to a final volume of 50 μ L with the concentrations of 1 $\times$ NEB Phi29 buffer, 0.1milligram/millilitre BSA or recombinant albumin (which are included in the kit), 2.5 micromolar (μ M) of each primer pool, 1mM of dideoxynucleotide (ddNTP) mixture, and 10 units of Phi29 enzyme (NEB, M0269L). ddNTP ratios were adjusted to 70% adenosine-thymine (AT) bases as trialled in (98). Following this, produced mass was tested using a PicoGreen assay. Samples that did not produce an expected yield were

re-processed. 13 samples produced no yield after two rounds of sWGA and were thus excluded. It is worth noting that negative controls are prone to produce high mass due to the amplification of primers within the reaction, however this tends to be much less than a successful sWGA amplification. sWGA reactions were stored at -20°C in wait for DNA sequencing.

Name	Sequence
Pf1	ATATATATAT*A
Pf2	TATATATATAT*T
Pf3	TATATATATA*A
Pf4	TAATATATA*T
Pf5	TATATATATT*T
Pf6	ATTATTATTA*T
Pf7	TAATAATAAT*A
Pf8	AAAAAAAAAAAA*A
Pf9	AATAATAATA*A
Pf10	TATTATATA*T

Table 2.1: **The 10A sWGA primer set.**

Name	Sequence
6A_0	TAAATAAAAA*A*A
6A_1	CATAAAAAA*A*A
6A_2	TAAATAATAA*T*A
6A_3	ATCATAATA*A*T
6A_4	TAACAAAAAA*A*A
6A_5	TAATAAATAA*A*A
6A_6	TAACATAGG*T*C
6A_7	TAGTAGTAG*T*A
6A_8	ATAATAAATA*A*T
6A_9	CATAATAATA*A*T

Table 2.2: **The 6A sWGA primer set.** The asterisk * represents a phosphonothioate bond.

2.2.3.2 Random Whole Genome Amplification (rWGA)

rWGA was performed using the QIAseq FX Single Cell DNA Library Kit (Qiagen, 181703). This kit is originally designed to amplify genomes from single cells (i.e. with minimal DNA mass input) but follows the same principles of other WGA methods. This was performed under a Class II Microbiological Safety Cabinet which had been cleaned thoroughly using a 1% bleach solution and DNAZap solution (Invitrogen, AM9890). rWGA was performed following the manufacturer's instructions but using 50% volumes of the master mix buffers. To a final volume of 20 μ L, a master mix of 14.5 μ L REPLI-g buffer, 1.25 μ L DTT (diluted to a ratio of 1:10), 1 μ L REPLI-g sc DNA polymerase, and 3.25 μ L of H₂O were added to a 1 μ L volume of McrBC digested or bead enriched genomic DNA (gDNA). The protocol recommends 1-10ng mass input; the approximate concentration of McrBC digests was 0.33ng/ μ L. This mixture was incubated at 30°C for 2.5hours before polymerase inactivation at 65°C for 3 minutes and the results stored at -20°C prior to sequencing.

Non-amplified control DNA was extracted from a sample of 3D7 erythrocyte culture using the Quick-DNA Miniprep Plus (Zymo Research, D3024) and DNA Clean & Concentrator-10 Kits. Further gDNA for 3D7 (the 3D7A isolate which was used as the amplified control input) and other parasite laboratory isolates including FUP-H were sourced from BEI resources. If gDNA was input for sWGA, human DNA was added to parasite DNA to mimic the standard conditions for this prep.

During the developmental stages of this work, the outcome of *Pf* DNA bead enrichment and WGA was tested using a nested polymerase chain reaction (PCR) for the *Pf* gene merozoite surface protein (MSP)2 (described in (136)). The presence of strong bands on a gel was taken to indicate the successful amplification of *Pf* DNA.

2.3 Description of DNA sequencing methods for *Pf* parasites

2.3.1 Short read sequencing

Sequencing and library preparation was performed by Oxford Genomics Centre (Centre for Human Genetics (CHG), University of Oxford) (this group has now been discontinued, with key staff subsumed into the CHG Technology Platforms Group). Three library preparation methods were tested (see Chapter 3 (pg.87)). The NEBNext Ultra II FS kit was used on the entirety of the GAMCC set.

2.3.1.1 NEBNext Ultra DNA Library Prep kit

gDNA was normalised to 500ng and fragmented by mechanical shearing to 350 base pairs (bps) using a MultiFunctional Bioprocessor. Library preparation was performed using standard Illumina multiplexing adapters with minor modifications to the manufacturer's protocol. Libraries were PCR-amplified (10 cycles) using in-house unique dual indexing primers (137). Post-PCR purification was performed using SPRIselect beads (Beckman, B23318).

2.3.1.2 NEBNext Ultra II FS kit

Samples were purified prior to normalisation using Beckman SPRIselect beads with a ratio of 1.8 $\times$. 100ng of gDNA was then used to prepare libraries with minor modifications. For fragmentation, the enzyme was diluted 1:2 and incubated with the sample for 8 minutes to achieve an average fragment size of 450bp. Size-selection was adjusted to recover $\sim$ 550bp and eluted in 15 μ L. The sample was then split and 7.5 μ L was taken forward into a full volume PCR performed at 6 cycles also using in-house unique dual indexing primers (137). Post-PCR purification was performed using Beckman SPRIselect.

2.3.1.3 NEBNext Ultra II (PCR-free) kit

500ng of gDNA was fragmented using the Covaris ME220 for inserts of $\sim$ 330bp and prepared using the standard protocol. Samples were indexed using the NEBNext Multiplex Oligos for Illumina (96 Unique Dual Index Primer Pairs). Libraries

were size selected to ~ 550 bp. Library sizes were quality checked by PCR amplifying $1\mu\text{L}$ of sample for 12 cycles and run on the Agilent 4200 TapeStation. Samples were pooled to equimolar ratios and quality-checked using quantitative PCR.

All libraries were then normalised using a PicoGreen or Quantifluor assay and pooled. The size profile of the pooled library was analysed on a TapeStation. The pooled library was quantified using Qubit (Invitrogen) and diluted to ~ 10 nanomoles (nMs) for storage. The 10nM library was denatured and further diluted prior to loading on the sequencer. Paired-end sequencing was performed using a NovaSeq6000 platform (Illumina, NovaSeq 6000 SP Reagent Kit v1.5 (300 cycles)).

2.3.2 Long read sequencing

Long read sequencing was performed for a small selection of *Pf* samples as part of a trial of DNA fragmentation methods. Whole-genome assemblies were generated (as described in Section 2.5.4.1 (pg.50)), which were used throughout this work. In addition, the Oxford Nanopore Technologies (ONT) MinION platform was also used to perform amplicon sequencing to investigate structural variation at a parasite locus (see Chapter 6 (pg.225) for full details).

2.3.2.1 PacBio sequencing

Library preparation for PacBio sequencing was performed by Mr. James Docker (Technology Platforms Unit, CHG). Sample (sWGA amplified or non-amplified gDNA) was first normalised to 300ng to a final volume of $50\mu\text{L}$ in nuclease free water. DNA was then fragmented to 10 kilobases (Kbs) at 1500 rotations per minute for 5 minutes using the 1600 MiniG Homogenizer. Samples that had been identified as containing 1.6Kb fragments (or smaller) were input to a size selection step using Ampure PB beads (diluted to 35% volume to volume) at a ratio of $3.1\times$ to remove fragments $< 5\text{Kb}$. Fragment size distribution was then checked using the Agilent Femto Pulse.

Fragmented samples were prepared using the PacBio SMRTBell Prep. Kit 3.0 and indexed with the SMRTbell Barcoded Adapter Plate 3.0. Libraries were then pooled and prepared for sequencing using the Sequel II binding kit 3.2. Libraries were loaded at a concentration of 105picomoles with adaptive loading and

sequenced using the PacBio Sequel IIe platform with a 30-hour movie time (time of sequencing reaction).

2.3.2.2 Nanopore amplicon sequencing

Multiplex primers were designed using a pre-release version of the software Pf-multiply, developed by Dr. Jason Hendry and others (104). Pf-multiply implements a pipeline using the primer design software primer3 and several downstream *in silico* analyses to generate PCR primers for multiple genome regions simultaneously (a multiplex PCR panel), while minimising potential primer cross- or self-binding. In brief, this pipeline was used (working jointly with Dr. Gavin Band and Dr. Jason Hendry) to generate three multiplex primer designs for the *Plasmodium falciparum* sickle associated (*Pfsa*) and amino acid transporter 1 (AAT1) genomic regions. Due to structural variation at the *Pfsa3* locus (Chapter 6 (pg.225)), and to balance amplicon length across the panel, the software was adjusted to cover the *Pfsa3* region with two $\sim$ 3Kb amplicons, while the other regions were covered by one amplicon each. PCR amplification, optimisation, and Nanopore sequencing was then performed by Ms. Anna Jeffreys and Dr. Mariateresa De Cesare at the CHG, using DNA from the 3D7 and FUP-H laboratory isolates obtained from BEI resources. Both the multiplex assay and single primer pair amplifications were tested. This was used to PCR *Pf* DNA generated from an sWGA reaction. The quality of the multiplex PCR was then visually assessed using a gel, and the set of amplicons with the most even amplification was then sequenced using an ONT MinION.

2.4 Targetted human sequencing

2.4.1 HLA typing

Samples within the DK collection were considered to be more likely to be high quality due to their storage at -80°C , however their volumes were minimal. Samples within the AH collection (see Section 2.2.1 (pg.35)) of the GAMCC cohort were chosen for HLA-typing because the large mass of DNA required for this was available from these aliquots. However, samples within the AH collection were

heterogenous and highly glutinous in consistency. This caused substantial inaccuracies in pipetting and quantification during initial attempts to aliquot these samples. This was identified by repeat quantification of the samples showing inconsistent results. Several steps were therefore taken to encourage fluidity: first, samples were defrosted at room temperature overnight and when ready to aliquot were gently heated to 60 °C, vortexed by pulsing, and quickly spun several times. This made it more possible to more reliably transfer DNA from these samples, although the mass was often still difficult to predict.

High resolution typing of HLA alleles was performed by the company Histogenetics (Ossining, New York, USA) (using the manufacturer’s “1×” version). This included antigen recognition site typing across exons 2 and 3 of the HLA Class I genes and across exon 2 of HLA Class II genes. This was performed using a combination of short and long read sequencing platforms including the Illumina MiSeq and NovaSeq, as well as the PacBio Sequel machine.

2.4.2 HBB typing

Two polymorphisms of the *beta-globin* (*HBB*) gene were called as part of this work: haemoglobin (Hb)S (encoded by rs334; chr11:5,227,002 T>A on GRCh38) and HbC (encoded by rs33930165; chr11:5,227,003 C>T). HbS typing for many samples in the GAMCC cohort was performed at the time of their collection (referred to as the AH-calls). However, 125 of the samples for which *Pf* WGS was available did not have an HbS genotype within the inherited metadata. Furthermore, an initial set of *HBB* retyping revealed substantial inconsistencies within the AH-calls (described in detail in Chapter 3 (pg.87)), which prompted the seeking of renewed genotyping information. Some genotypes were available through human genotyping data available for a subset of GAMCC samples. Most HbS/HbC retyping was performed using a Sanger-based approach described below (a complete display of sample numbers stratified by method is given in 3.2 (pg.101) in Chapter 3).

2.4.2.1 Low-coverage sequencing and array-based typing of GAMCC samples

HbS genotypes in a subset of the GAMCC cohort were recently retyped through a separate project to compare array-based genotyping and low coverage WGS (lcWGS) of the human genome in this dataset. Specifically, genome-wide genotypes for 190 samples were assayed using the ThermoFisher Axiom Precision Medicine Array (HbS is directly assayed on this array). In addition, low-coverage WGS of a superset of 210 samples was performed, followed by genome-wide imputation using the software QUILT (138). Of these, $\frac{1}{3}$ rd of the samples were non-malaria controls. In total 139 of the lcWGS-imputed samples and 122 chip-typed samples were among those with *Pf* WGS data in the GAMCC dataset.

2.4.2.2 The Sanger sequencing method

This method has been previously described (131; 54). Ms. Katie Healey (Oxford Genomics Core, CHG) assisted with the PCR step of this method. For each sample, using the AH aliquots (see Section 2.2.1 (pg.35)), the region was first amplified via PCR using the primers acgttggatgGTCTCCTTAAACCTGTCTTG and acgttggatgTCAAACAGACACCATGGTGC. To a final volume of 25 μ L, a master mix of 1.55 $\times$ BioTaq buffer, 1.24millimoles (mMs) ddNTPs, 3.1mM MgCl₂, 0.3 μ M of each primer, and 0.4U/sample of BioTaq enzyme (Bioline, BIO-21040) was added to a 2 μ L aliquot of sample (which had been pre-diluted with nuclease free water to ratio of 1:15). Two other PCR recipes were tested during this work: one using KAPA HiFi (Roche, KR0370) and another using the REDTaq ReadyMix PCR Reaction Mix (Sigma-Aldrich, R2523), however neither produced a visible amplicon on a gel. The PCR programme is shown in Table 2.3 on the next page. The original protocol was adapted to perform more amplification cycles (35 instead of 29) in order to generate more product.

Products were initially checked on a gel for an 162bp band. As the project progressed and the assay became reliable, a TapeStation was used to check a random subset of 10 samples per plate to check the majority had produced an amplicon. Plates were sent to the company Azenta Life Sciences (Takeley, United Kingdom) for sequencing. Each amplicon was sequenced in one direction using

Step	Temperature	Time
1	96°C	1 min
2	94°C	45 secs
3	56°C	45 secs
4	72°C	30 secs
5	Go to Step 2	x5
6	94°C	45 secs
7	65°C	45 secs
8	72°C	30 secs
9	Go to Step 6	X35*
10	72°C	10 mins
11	15°C	15 mins

Table 2.3: **PCR amplification steps for the *HBB* gene.** * was increased from 29 in the original protocol.

the primer acgttgatgGTCTCCTTAAACCTGTCTTG. On each 96 well plate a positive HbAA control (human DNA from a Centre d'Etude du Polymorphisme Humain (CEPH) individual) and a negative control (nuclease free water) was included. Sequencing was repeated once if the first attempt failed to produce usable data.

Sanger traces were inspected using the software 4Peaks (version 1.8), and sample genotypes were determined visually. An example of an HbAS trace is shown in Fig. 2.1b on the following page, and an example of a reference genotype (HbAA) is shown in Fig. 2.1a. These traces sometimes showed noise at parts of the amplicon which would obscure the sequence. Data would at times come with QC warnings provided by the sequencing company. This noise may be due to the increased number of PCR cycles used during this assay which can lead to poorer quality sequence.

Because it is guanine-cytosine (GC)-rich, the *HBB* gene can be challenging to amplify (139), exemplified by failed attempts at amplification using adjusted master-mix recipes including different commercially available polymerases alluded to above. As long as the sequence locally at the HbS and HbC positions was clear, genotypes were considered to be able to be judged sufficiently.

There were two instances of positive controls not producing any sequence, and

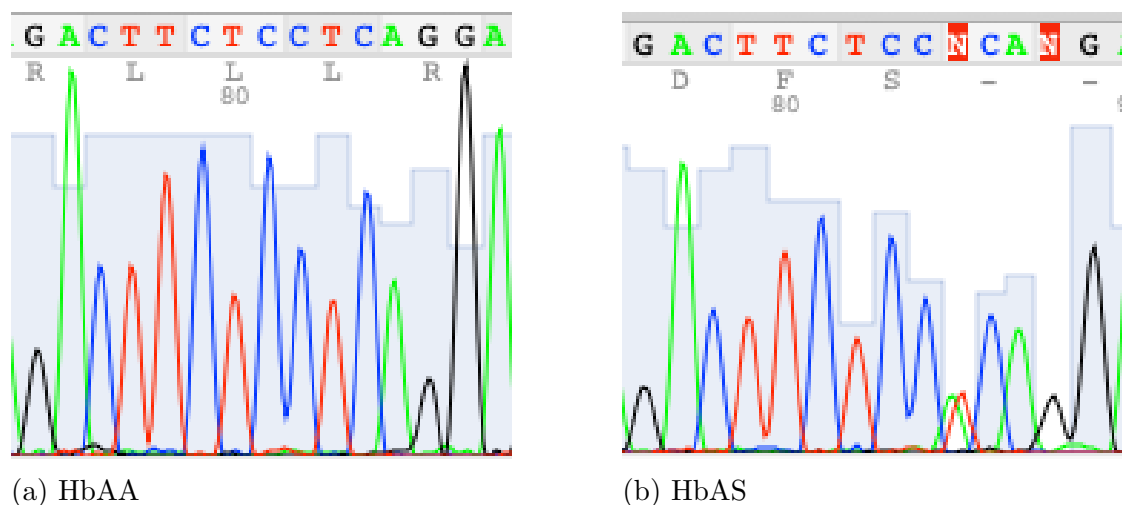


Figure 2.1: **Example sanger-sequencing traces used to genotype the HbS locus.** **A.** shows an HbAA sample and **B.** shows an HbAS sample.

the negative control showing an amplicon. This amplicon appears identical to that of positive controls from other sequencing runs, and can therefore likely be attributed to a sample mix up. Controls were added to plates after the test samples had been carefully plated, so there can be reasonable security that this issue is not indicative of broader errors in the process. Low levels of poor-quality sequence signal was also occasionally seen within negative controls which may be suggestive of minor contamination. The low signal observed in the Sanger run indicates minimal material was present in the well, indicating that contamination would likely have occurred after the PCR step. While this contamination is not ideal, it is unlikely to have caused false positive calls that could significantly impact the analysis. Notably, no samples identified as carrying one or two HbS alleles were in adjacent wells, suggesting that cross-contamination between wells causing false positive signals did not occur.

Plates aliquoted from the DK set used to generate *Pf* WGS would have been the preferable source of input for this typing, as there would be a reduced risk of sample mix up leading to falsely paired HbS calls and *Pf* genotypes, however due to the high occurrence of low quality control (QC) samples within these plates, and due to time constraints, targeted sequencing samples was undertaken. There is of course a concern that samples in the DK and AH collections may be mixed in

relation to each other causing parasite WGS and HbS types to mismatch. Samples for which aliquots from both the AH and DK collections underwent *Pf* WGS showed matching variation (as assessed using BamMatcher (for further detail see Section 2.5.2 on the next page)), and a subset of DK aliquots also underwent *HBB* genotyping (N=10) to give matching results (7/7 successful pairs of sequencing matched).

2.4.3 Previously performed typing

No further renewed human typing was performed as part of the work presented in this thesis, although numerous other loci in these samples have been genotyped (e.g. (140; 141; 142)). ABO blood type information was used as part of Chapter 4 (pg.110). The method of ABO typing performed on these samples is not clear from metadata, but was likely to be serologically based, and therefore types are likely to refer to a phenotype: the presence of A or B antigens on the surface of red cells, a combination, or neither of these antigens, rather than a genotype.

2.5 Genetic data alignment and processing

2.5.1 Reference assemblies used and *Pf* genome information

Pf short read WGS data was processed by aligning to a concatenated assembly including the human genome with decoy HLA (GRCh38 version GCA_000001405.15), the *Pf* 3D7 reference (Pf3D7 version 3) and PhiX-174 genome (accession NC_001422). The human assembly was included to allow the determination of levels of human-origin DNA present in sequenced libraries, and so that these reads could be removed later when data is made publicly available in order to comply with ethics approval for the use of these samples. The bacteriophage PhiX genome was included because its DNA had been spiked into libraries in order to add complexity (PhiX has a balanced nucleotide composition) and enhance sequencing chemistry. This tended to have virtually non-existent coverage in our data.

The 3D7 reference is ~23 megabases (Mbs) in length and contains 14 chromosomes as well as a mitochondrion and apicoplast genome (94). In this assembly,

chromosome contigs are designated as “Pf3D7_01_v3” for chromosome 1, continuing sequentially up to “Pf3D7_14_v3” for chromosome 14. AT content is just over 80% in the assembly, rising to around 90% in non-coding regions. Regions of the genome are divided into core, sub-telomeric hypervariable, sub-telomeric repeat and centromeric regions. ‘Core’ regions of the *Pf* genome, which make up about 90% of the genome total (20.8Mb), are defined as sections of the 3D7 assembly to which short read illumina data can be accurately aligned to, as previously annotated (102). These core regions will be referred to throughout this thesis.

2.5.2 Alignment of short read sequence data

To process the large amounts of generated short read sequencing data in a consistent fashion pipelines were generated using the workflow tool ‘snakemake’ (version 7.18.1) (143). For the analysis presented in Chapter 3 (pg.87), data produced using the PCR-free library preparation method were first subsampled to 16 million reads using ‘seqtk’ (version 1.3-r117) to allow efficient processing (144). This is because a small number of samples were sequenced as part of a single sequencing run, generating superfluous read depth. All other genetic data produced in this thesis was more highly multiplexed and so were processed entirely.

Firstly, duplicate reads were first removed from raw sequencing reads in fastq format using ‘FastUniq’ (version 1.1) (145) and the adapters removed using ‘Trimomatic’ (version 0.39) (146). ‘FastQC’ (version 0.11.9) and ‘MultiQC’ (version 1.9) (147) was used prior and post-duplicate and adapter removal to allow manual inspection of the efficacy of these steps. The cleaned reads were then aligned to the concatenated assembly (see the previous section) using the bwa-mem algorithm (version 0.7.12) (148) and the resulting files converted to a sorted Binary Alignment Map (BAM) format using ‘samtools’ (version 1.8) (149).

Certain libraries were sequenced in multiple rounds, and so data produced from multiple runs was merged following alignment. First, it was checked that each BAM file produced for the same sample over separate runs came from the same source (i.e. sample IDs had not become mismatched between runs) using the tool ‘BamMatcher’ (version 2.7) (150). This calls variation within alignments (by aligning to the reference sequence and calling variation using GATK) and

outputs a contingency table of variation compared at each site to give the fraction of variation in common. Only samples which were determined to match (i.e. with “BAM files are very likely from the same source” within the output report) were then merged. As well as matching, some pairs were reported to also show “the possibility of allele specific genotype”. The flagging of an allele-specific genotype may be due to difficulties aligning to non-core and low complexity regions of the *Pf* genome. All alignments produced from separate runs of a single sample were found to match. Alignments were merged using the samtools ‘merge’ command.

2.5.3 Quality control assessments for short read alignments

Following alignment a number of quality control measures were performed (as described in Chapter 3 (pg.87)). These analyses were limited to “core” regions of the genome.

2.5.3.1 Assessment of human contamination

Levels of human derived gDNA present within sequenced libraries were determined by measuring the proportion of reads mapping to contigs in the human GRCh38 reference compared to those aligning to the parasite *Pf*3D7 reference. This was calculated using primary alignments only using the tool samtools ‘idxstats’.

2.5.3.2 Assessment of breadth of sequencing coverage

To allow direct comparison of breadth of coverage between sequenced genomes, alignments were downsampled using seqtk. Alignments were limited to primary reads and converted to the fastq format (using the samtools command ‘fastq’). 3 million reads were subsampled, and original alignments filtered to contain only those reads using the Picard (version 2.21.1) tool ‘FilterSamReads’. ‘Bedtools genomecov’ (version 2.27) was used to compute number of base positions in the genome at each depth. Cumulative coverage distributions were compared to a Poisson distribution where μ was the average genome coverage (total reads/size of the *Pf* core genome in base pairs).

2.5.3.3 Assessment of duplication levels

Numbers of duplicate reads within alignments were assessed using the Picard tool ‘MarkDuplicates’.

2.5.4 Processing of long read sequence data

2.5.4.1 Generation of assemblies from PacBio data

New whole-genome assemblies were assembled by Dr. Gavin Band using the following process. The PacBio platform was used to perform long-read WGS on eight parasite samples. This included an isolate called FUP-H (gDNA obtained from BEI Resources (MRA-915G)); four clinical isolates from Kenya (KENI1-KENI4), two isolates from the GAMCC resource (described in Chapter 3 (pg.87)), and the widely-used laboratory reference clone 3D7. First, separately for each sample, hifiasm (version 0.19.5-r587) was used to create a graph assembly (.gfa file) based on the PacBio reads using default parameters. Hifiasm outputs a set of ‘primary’ contigs and separate files containing ‘alternate’ contigs; the latter may reflect small assembly fragments or other low-quality contaminants. Assembly graphs were inspected using Bandage-NG (version 2022.8-dev). We noted that for most samples processed, nuclear chromosomes were typically assembled in a single contig or (occasionally) in two contigs with little branching. Further refining these contigs to produce complete assemblies was not required for the purposes of this thesis and represents future work. The primary contigs were converted to FASTA format using the ‘gfatools gfa2fa’ tool; this outputs the graph vertex sequences as separate FASTA records.

In addition to contigs compatible with *Pf* nuclear chromosomes ranging in length from ~640Kb–3.3Mb, the apicoplast (~29-35Kb), and multiple shorter contigs putatively representing the mitochondrion (~5-6kb), inspection of KENI assemblies revealed the presence of a prominent ~720Kb contig. Alignment of this to available sequences using BLAST (version 2.7.1) suggested that this was *Mycoplasma arginini*, a common bacterial contaminant of *Pf* culture medium.

Next, minimap2 was used to align contigs to a combined reference of the *Pf* Pf3D7, human GRCh38, PhiX and *M. arginini* reference genomes using the -x asm5 option which is suitable for assembled genomes of limited divergence. Then,

a custom program was used to identify the primary alignment of each contig to assign it to a nuclear chromosome. If necessary, the reverse complement sequence was produced to match the aligned reference contig orientation. This output a new FASTA file with re-named and re-oriented sequences. The oriented version of the assembly was used for all subsequent analyses.

Gene locations were annotated onto newly generated assemblies at the *Pfsa3* locus (used in Chapter 6 (pg.225)) as follows. Seperate structural variant (SV) regions of the *Pfsa3* SVs were extracted from the 3D7 reference assembly (based on segments R1-R10 from Table 6.3 (pg.257)). These regions were then aligned to each assembly using minimap2 using the options '-xasm5' (appropriate for assemblies), '-eqx' (to output 'X' instead of 'M' in mismatching bases), and '-cs' (to output the CS alignment field). The gene information for each region was subsequently extracted from the 3D7 reference annotation (.GFF) file using a custom script, and lifted over to the coordinates of each new assembly. Specifically, for each GFF record intersecting one of the segments, and for each alignment of the segment, the alignment location was computed by parsing the "cs field" output by minimap2 (this was necessary since the alignment might have a different length compared to the original segment if it contained insertions or deletions). This produced a new GFF file including the duplication of the SNRPF segments upstream of PF3D7_1127000. This file was manually edited to group neighbouring exons into genes.

Generating a multiple sequence alignment of whole genome assemblies

To compare assemblies across samples, we generated a multiple sequence alignment of the whole genome assemblies as follows. We included a set of publicly available assemblies previously published by Otto et al. (103), and the newly generated assemblies described above. To produce this, for each region of interest we first extracted the 101bp segment from the 3D7 reference genome centred on the SNP of interest. We used minimap2 to align this segment to all the assemblies. We then extended the alignments outwards by 10kb on either side to generate an approximately 20kb region of each genome centred on the variant of interest.

Using these 20kb segments, we then used MAFFT (version 7.490) with default options to compute a multiple sequence alignment of the genomes. The output of

MAFFT is a FASTA file with sequences of identical length; the sequences contain a ‘-’ character to indicate alignment locations that are not present in a given genome.

To visualise the alignments, we used a custom software tool called zoomsa (unpublished work by Dr. Band) which reflects the alignment onto a webpage for dynamic viewing.

To generate certain images of these assemblies presented in this thesis (e.g. Fig. 5.26b (pg.209)), the alignments of the 101bp segment were extracted and aligned using MAFFT, and the small multiple sequence alignment visualised using pyMSAviz to generate an image (version 0.5.0).

2.5.4.2 Analysis of amplicons generated using Nanopore sequencing

As described in Section 2.3.2.2 (pg.42), and in Chapter 6 (pg.225), an assay for targeted amplicon sequencing of the *Pfsa1-3* loci on the ONT MinION was developed in collaboration with Dr. Band and the Hendry group. As shown in Chapter 6 Supplementary Figure 6.10 (pg.255), three amplicons were predicted to be produced from parasites carrying the DUP-F haplotype. To analyse data produced from FUP-H, raw sequenced reads were aligned to a version of the 3D7 reference assembly which had been rearranged based on the structure of DUP-F shown in Figure 6.1 (pg.232) (using regions in Supp. Table 6.3 (pg.257)). This appeared to show the presence of three amplicons (Fig. 6.11 (pg.256)). Since this was conducted, a whole genome assembly for FUP-H has been generated which would be better suited to this purpose moving forward.

2.6 Variant call sets developed for this thesis

2.6.1 The production of a combined GAMCC/CP1 call set

Variants within the GAMCC and MalariaGEN CP1 sample sets were called together as part of a large set of 5,077 samples (referred to in this thesis as the GAMCC/CP1 call set). Although a *Pf* variant call set was already available for the MalariaGEN CP1 cohort, recalling was performed due to the large numbers of comparisons to perform between this set and the GAMCC, and the fact that both sets contain Gambian samples. Sequence data in the form of fastq files

were downloaded for all samples in the published “analysis” set of CP1 (described in Section 2.6.3.1 on the next page) and processed via the alignment pipeline described in Section 2.5.2 (pg.48).

Variants were called in *Pf* data using a workflow inspired by that published for the Pf7 dataset (37). First the GATK (version 4.1.4) base quality score recalibration tool was applied over each chromosome within each sequencing run for each sample. This step, which uses observed differences from the reference genome to empirically recalibrate sequencing base qualities, determines systematic sequencing errors present in the set. GATK’s haplotype caller was then utilised to call genotypes per sample and samples were then grouped using the DBImport functionality. GenotypeGVCF was then applied to call variation jointly across the dataset. Variants were annotated by their region type (core, subtelomeric hyper-variable, subtelomeric repeat or centromeric regions (102)), and their functional impact annotated using SNPEFF (151). This was performed across 200Kb segments of genome to accelerate processing. Resultant variant call format (VCF) files were then merged into chromosomes using bcftools.

GATK’s Variant Quality Score Recalibration (VQSR) was then applied to all calls. This tool makes assessments of the quality of variants based on multiple metrics such as quality normalised by depth, strand odds ratio, and fisher strand measures to then use machine learning to annotate potentially erroneous variants (152). The tool requires a training set of true variation to determine annotations. For this I used high confidence calls published from crosses of laboratory clones (3D7 with HB3, 7G8 with GB4 and HB3 with Dd2) (102). VQSR was performed for single nucleotide polymorphisms (SNPs) and short insertion or deletions (INDELs) separately for each chromosome in the *Pf* genome.

2.6.2 Generating an intersected variant call set from the GAMCC, CP1 and Pf7 cohorts

Raw sequencing data in the form of fastq files were not available for the Pf7 release (only alignment files in BAM format), so it was not possible to apply the described alignment processing steps to samples in Pf7. A large variant call set, called across the entire cohort of around 20,000 samples, which was generated

using a process very similar to that described in the previous section, is publicly accessible. In order to save time computationally, variation in Pf7 was not re-called alongside GAMCC/CP1, but instead the already available call set was intersected with the GAMCC/CP1 call set described in the previous section. This produced the ‘intersected call set’ analysed in this thesis. Because mainly the Kenyan and Gambian populations were assessed in this work, intersection was limited to these populations in Pf7 (sample filtering is described in full in Section 2.6.3.1).

Due to the large numbers of samples in both the CP1 and Pf7 datasets, many rare, AT-rich and overlapping INDELs were present in the variant call sets, which may have occurred due to poor resolution in the alignment of low complexity reads (153). The overlapping nature of these alleles means that Pf7 allele encoding will often differ to those in the GAMCC/CP1 call set despite referring to the same variant. To attempt to overcome this issue, the `vcflib` (version 1.0.3) software ‘`vt normalise`’ tool was used to alter the encoding of variants into a universal format (154). Alleles in the Pf7 and GAMCC/CP1 sets with counts of less than 5 were then removed entirely using the `vcflib` ‘`vcffilter`’ command. This retained truly biallelic sites which were called as multiallelic due to rare alleles, which would otherwise prevent the intersection at that site (155). Core variants within each independent dataset were combined using ‘`bcftools`’ to generate an intersected set of 168,412 variants (including 81,160 SNPs, 10,332 of which were multiallelic, and 29,646 INDELs). An equivalent process without the ‘`vt normalise`’ step produced only 110,270 variants, indicating that variant normalisation is useful in this context. Finally, genotypes for the Gambian and Kenyan severe cases from CP1, the severe and mild cases from the GAMCC, and the Gambian and Kenyan Pf7 samples were merged across these variants. The combined callset is referred to in Chapter 4 (pg.110) as the ‘intersected call set’.

2.6.3 Quality control and filtering of variant call sets

2.6.3.1 Sample filtering

CP1 For samples in the MalariaGEN CP1 set, the previously published “analysis set” was relied upon for sample filtering (6). Samples were then excluded under the following criteria:

- (i) Samples where the GATK v.3.8.0 CallableLoci metric (defined as the proportion of genomic bases in the core regions of the *Pf* genome with at least $5\times$ fold coverage and where at least 90% of covering reads have a mapping quality score ≥ 10) was less than 50% were removed.
- (ii) Samples which had $>5\%$ genotype missingness were excluded.

In total, 2235 samples were included from Gambia and 1936 from Kenya in this set (Table 3.2 (pg.101)).

GAMCC The GAMCC set was sequenced as part of this thesis and thus required sample filtering. Once samples were sequenced, they were filtered by criteria (i) and (ii) in the previous section with one alteration: for the GAMCC samples which had genotype missingness $>10\%$ were excluded. This more lenient threshold was adopted to retain as many samples as possible given the relatively small size of the GAMCC set. Furthermore, the potential DNA degradation due to the age of the samples may have caused dropout in certain regions without necessarily compromising accuracy at other loci. Moreover, the CP1 samples were sequenced to very high coverage (with many exceeding $100\times$ fold coverage, compared to on average $55\pm 18.4\times$ fold coverage for the GAMCC samples), which may lead to a higher proportion of non-missing calls. In total, of 905 sequenced samples, 831 were retained in the GAMCC analysis set (Table 3.2 (pg.101)).

Pf7 The total available samples from The Gambia and Kenya from Pf7 were included in the intersected call set: 1247 samples from Gambia and 726 from Kenya (Table 3.2 (pg.101)). The previously published “QC pass” analysis set of Pf7 was also relied upon for the filtering of these samples. QC pass samples were judged by the original authors to be those which satisfied criteria (i) from Section 2.6.3.1 on the preceding page, and samples which did not have a high proportion of singleton SNP calls (37).

For the analysis of parasite genetic effects on disease severity presented in Chapter 4 (pg.110), samples from Kenya were further filtered to those collected in the Kilifi district (“Admin level 1” = “Kilifi” within released metadata) to match the general location of those collected in the CP1 set. This led to a total sample

size of 602 mild (Pf7) samples and 1,936 severe (CP1) samples for this analysis (Table 3.2 (pg.101)).

For Gambia, CP1 samples were collected from the MRC laboratories in Banjul and were recruited from hospitals near this region (54). Pf7 samples were recorded as having been collected in three districts: North Bank, Western and Upper River. To maximise potential sample size, and because the country Gambia covers a small region of space, all of these samples were retained. This generated a sample size of 818 mild (Pf7) samples and 2235 severe (CP1) samples from Gambia (Table 3.2 (pg.101)).

2.6.3.2 Variant filtering

To ensure the accuracy and reliability of variant calls tested within this thesis, several stringent filtering steps were applied. Firstly, I observed that analyses run while retaining deletion variants, denoted by an asterisk in VCF files, led to a high rate of false positive associations at these sites which did not appear credible. Analysis was therefore restricted to SNPs to ensure tested variation did not reflect sequencing or calling errors.

In addition, in GWAS analyses (Section 2.7.2 (pg.59)), only variants which had a minimum predictor count (that is the minimum number of observations of the allele in homozygous or heterozygous form) of at least 5 were included to further ensure the allele existed and was not an artefact of sequencing error. Moreover, because of the parameters required by certain software (e.g. SNPTEST), and because multiallelic SNPs may be due to sequencing error, only biallelic SNPs were considered in analyses. Further refinement involved selecting variants with a Variant Quality Score log odds (VQSLOD) > 2 , a missingness proportion $< 10\%$, and those within core genomic regions.

Although *Pf* parasites exist as haploid during the human portion of infection, variants in the GAMCC/CP1 and Pf7 call sets were called (following standard practice (5)) as diploid, which provides a basic way to account for levels of mixture within malaria parasite infections. This produces heterozygous calls for some samples at certain loci, which do not reflect individual parasite genomes. These genotype calls were therefore excluded from GWAS. Unless stated otherwise, only variants for which at least 90% of samples had non-missing genotype calls following

this removal were included within the presented results to ensure expected power was roughly retained.

2.6.3.3 Principal component analysis and other quality control

Principal component analysis (PCA) was performed using the programme QCTOOL (v.2.2) (6) using biallelic SNPs with a minor allele frequency of at least 2.5% that have been thinned such that no variant was closer than 500bp. Other metrics, such as variant or sample missingness and relatedness between samples, were also determined using QCTOOL.

2.6.4 Generating phased polarised haplotypes across African samples in the Pf7 resource

To create a curated African-specific set of genotypes for population genetic analysis, VCF files (representing genetic variant sites and sample genotypes) were downloaded from the Pf7 resource page (<https://www.malariagen.net/resource/34/>) and post-processed as follows. First, samples were restricted to those from African study sites which were marked as passing Pf7 QC processes (N=8492) (sample QC is described in 2.6.3.1 (pg.54)). Genotypes in these VCF files are called as diploid with heterozygous calls indicating potentially mixed infections. Samples with non-trivial rates of mixed infection (identified using the Fws metric as $Fws < 0.9$) were excluded retaining 4788 samples. We also restricted attention to variants marked as 'PASS' (i.e. passing the genotype calling pipeline's quality filters; this also excludes variants outside the core genome).

The VCFs above have a mixture of biallelic and multiallelic sites; however, many of the multiallelics are effectively biallelic with a single rare allele. To handle this we use vcflib 'vcffilter' tool to exclude alleles with fewer than 20 copies across the included 4,788 samples, and included the remaining alleles as biallelic variants in the callset (the vcffilter tool sets genotypes for removed alleles to missing and excludes variant records where the alternative allele is "*", indicating that the site overlaps a deletion defined on a different row of the VCF file). A total of 2,959,655 variants with reference and alternate allele assignments were retained by this method.

Data was further processed to produce a set of fully haploid, non-missing genotype calls with alleles polarised to ancestral/derived state. First, remaining heterozygous calls were replaced with missing genotypes and then re-imputed in haploid form using beagle (version 5.4, beagle.01Mar24.d36.jar). 24 phasing iterations were specified to ensure convergence. A very small number of genotype calls (5486 across the whole callset) in the beagle output were still heterozygous and were replaced with reference homozygous genotypes. Finally, we used QCTOOL with the inferred ancestral/derived allele designation (Section 2.9.4 (pg.75)) to “polarise” the genotypes (i.e. so that the ancestral allele is the first allele and the derived allele is the second allele).

2.7 Genome wide association analyses

2.7.1 Phenotyping infections as mild or severe

2.7.1.1 GAMCC samples

Of the 831 samples in the GAMCC analysis set (for details of sample filtering see 2.6.3.1 (pg.55)), 502 were categorised as severe cases and 329 were categorised as mild upon their collection (Table 3.2 (pg.101)). Severe cases were carefully classified as those involving:

- (i) cerebral malaria (CM) (with an unrousable coma score < 3 (by standard criteria) or repeated prolonged seizures lasting $> 30\text{min}$).
- (ii) severe malarial anaemia (SMA) (Hb level < 5 grams/decilitre).
- (iii) Hypoglycaemia (blood glucose level $< 2.2\text{mM}$).

Mild cases were considered those with a parasitaemia $> 2500/\mu\text{L}$ (a level described to be associated with illness in this setting) but did not otherwise involve any of the three attributes defining severity (i – iii) (52). For the GWAS comparison all severe cases were compared to all mild cases.

2.7.1.2 CP1 and Pf7 samples

Samples within CP1 were classified as severe if they involved CM, SMA, or a combination of these traits using identical thresholds as the GAMCC set. In contrast to the GAMCC set, infections without these two traits (or severe hypoglycaemia) but with other uncharacterised severe pathologies were included as severe within a category denoted as “other” (Table 3.2 (pg.101)).

Generating the largest sample possible was the primary goal of the Pf7 collection. Contributors were asked to donate samples they had available, and because severe malaria is comparatively rare Pf7 is thought to contain mainly non-severe but clinical infections sampled within clinics and hospitals. Disease status was therefore not controlled and phenotypes are likely to be highly heterogenous. Although top-level descriptions of sample collection objectives are given within the Pf7 release, specific phenotypes are not linked to individual samples. The majority of samples available in the Pf7 resource from Kilifi, Kenya are reported to have come from both asymptomatic and symptomatic patients. Those from Gambia, which were sourced from six studies, include clinical cases (personal communication with Prof. D. Conway and Dr. A. Amambua-Ngwa) but also some from asymptomatic infections. Overall, samples within the Pf7 set are therefore likely to represent more mild infections than the CP1 resource, and so for the purposes of replication analyses presented in Chapter 4 (pg.110) are considered as “mild”.

2.7.2 Basic association model using SNPTEST

The programme SNPTEST (version 2.5.6) was used to perform genome wide association studies (GWAS) using a logistic regression-based model (6). As described in Chapter 4 (pg.110), this tested for associations between genome-wide genetic variants and the binary outcome of whether an infection was reported to be severe or mild. Because the intersected call set contains a reduced number of variants due to the intersection process (see Section 2.6.2 (pg.53)), the GAMCC/CP1 call set was used for any GWAS comparison that did not involve Pf7 samples.

In summary, using the following notation:

- Y_i : The binary outcome for individual i (e.g., whether the infection is severe ($Y_i=1$) or mild ($Y_i=0$)).

- X_{ij} : The genotype of individual i at SNP j (encoded as 0 or 2 corresponding to the number of minor alleles (heterozygote genotypes were excluded from the analysis)).
- β_0 : The intercept of the regression.
- β_j : The effect size (log-odds) associated with SNP j .

Then:

$$\text{logit}(P(Y_i = 1 \mid X_{ij})) = \beta_0 + \beta_j X_{ij}$$

If covariates were included in the model, these were added as additional terms:

- Z_{ik} : The value of covariate k for individual i (e.g. HbS status).
- γ_k : The effect size associated with covariate k .

$$\text{logit}(P(Y_i = 1 \mid X_{ij})) = \beta_0 + \beta_j X_{ij} + \sum_k \gamma_k Z_{ik}$$

β_j therefore represents the modelled change in the log-odds of the sample being a severe (vs. mild) infection when the infecting parasite carries SNP j . The resulting P-value tests the null hypothesis that $\beta_j = 0$, and thus small P-values indicate strong evidence of a genetic effect on the severity of infection at SNP j .

An important feature of genome-wide data is that it permits empirical assessment of potential confounding factors (which affect loci across the genome). The genomic inflation factor Lambda (λ) was used to assess this. λ compares the observed distribution of P-values with the expected distribution under the null hypothesis of no association between genetic variation across the parasite genome and malaria disease severity. Values significantly greater than 1 (i.e. many more significant P-values than expected by chance) indicate the occurrence of confounding through sources of bias. Regardless of the λ value, confounding at specific variants may still be present due to factors such as genotyping artifacts or environmental influences. Therefore, it is important to investigate potential confounders beyond λ to ensure robust results.

2.7.3 Modelling true discovery of genetic associations and performing power analyses

To aid in the interpretation of P-values for GWAS conducted in Chapter 4 (pg.110), and to re-assess the previously reported association between HLA-B*53 and severe malaria protection in Chapter 5 (pg.148), power analyses were conducted using simulations. For each, 2000 sets of genotypes were estimated using a binomial distribution. Input population allele frequencies (f) and tested effect sizes (odds ratio (OR)) were used to generate the allele frequency in cases (P_{cases}) using the formula:

$$P_{cases} = \frac{OR \times \frac{f}{1-f}}{1 + OR \times \frac{f}{1-f}}$$

This formula reflects the calculation of relative risk (RR), and how we use the OR to estimate this. Here RR is:

$$RR = \frac{P(\text{severe} \mid G = 1, S)}{P(\text{severe} \mid G = 0, S)}$$

In the analysis presented in Chapter 4, the RR reflects the probability of severe disease (as opposed to mild malaria) (denoted as ‘severe’ in the formula above) in infections where parasites carry a certain genotype ($G = 1$), as opposed to infections that do not ($G = 0$), in the study population S . In the analysis presented in Chapter 5, this RR reflects the probability of severe disease as opposed to uninfected individuals who carry HLA-B*53 ($G = 1$) and those that do not ($G = 0$). Because severe malaria is rare within the population, even within the population of malaria infections, the OR calculated within regression analyses can be used to estimate RR for both versions. By expanding the formula above using Bayes Theorem therefore:

$$RR = \frac{P(G = 1 \mid \text{severe}, S)}{P(G = 0 \mid \text{severe}, S)} \times \frac{P(G = 0 \mid S)}{P(G = 1 \mid S)} = OR$$

To convert this from odds that $G = 1$ to probability, we can use $x \rightarrow \frac{x}{1+x}$. This gives the formula for P_{cases} above. For cases, genotypes (G) were then drawn from the distribution:

$$G \sim \text{Binomial}(1, P_{cases})$$

Whereas for controls, genotypes were drawn from:

$$G \sim \text{Binomial}(1, f)$$

The number of cases and controls drawn were chosen to reflect the sample sizes of the assessed association.

Using the generated sets of genotype calls, a logistic regression was then performed. For a P-value threshold denoted T , given that an association is true (represented as A), the power of a test of association can be expressed as $P(p < T \mid A)$. For a range of P-value thresholds, the proportion of logistic regression association results where $p < T$ represents the power. From this, true discovery ‘rate’ (that is, the probability that an association is true given the P-value of the association is below a threshold) of the association analysis was assessed.

By inputting this into Bayes Theorem, the probability of true discovery (given the P-value is smaller than T) can be simplified to give:

$$P(A \mid p < T) = \frac{\text{power} \times P(A)}{\text{power} \times P(A) + T \times (1 - P(A))}$$

Here $P(A)$ therefore refers to the prior belief in A , which can have a large impact on the estimated true discovery. Ideally, this prior would be learned from other studies, however this is not possible for the presented discovery analyses. Therefore, these priors need to be estimated. In the analysis presented in Chapter 4 (pg.110) for example, a prior belief that 1 in an estimated 20,000 non-recombining blocks in the *Pf* genome is truly associated with disease severity was adopted. In Chapter 5 (pg.148), two priors were assessed. The first was a more relaxed prior that one of the 58 assessed HLA class I and class II loci in the original study

was associated with severe malaria protection. The second was a more stringent prior, to reflect a more commonly used GWAS prior, that one locus in 1 million recombining segments in the human genome might truly be associated with severe malaria protection.

2.7.4 Meta-analysing genome-wide associations between populations using BINGWA

In Chapter 4 (pg.110), a Bayesian meta-analysis of parasite genetic effects on severe malaria subtype was conducted on CP1 samples in the GAMCC/CP1 call set (Section 2.6.1 (pg.52)). Severity subtype categorisation as CM and/or SMA for the CP1 resource was attained from MalariaGEN. To perform the meta-analysis, filtering was performed as outlined in Section 2.6.3.2 (pg.56) but including an additional step to filter variants with less than 2.5% allele frequency in the Kenyan or Gambian populations, leaving a total of 20,079 variants across which to perform GWAS.

Genetic differences between parasite genomes sampled from CM, SMA and other severe infections was first assessed using a multinomial logistic regression GWAS separately for the Kenyan and Gambian samples using the programme SNPTTEST. Infections containing both CM and SMA were excluded for simplicity.

Then, the Bayesian meta-analysis was performed using the software BINGWA (version 2.1.2) (77). This programme takes SNPTTEST output files as inputs and performs a frequentist fixed effect meta-analysis and a Bayesian meta-analysis based on the approximate Bayes Factor (BF) approach. This allows a flexible set of models to be tested.

Specifically, it was considered whether an effect in this analysis might be driven by CM (the ‘CM only’ model), by SMA (‘SMA only’), whether it might be a general severity effect independently of whether a sample carried CM or SMA (the ‘independent’ model) or whether the effect was the same between CM and SMA (the ‘fixed effect’ model). The value of the resultant BF depends on the data as well as the prior distribution assumed to represent true association effect sizes. For this analysis, a prior was used which allowed for large effects while still concentrating

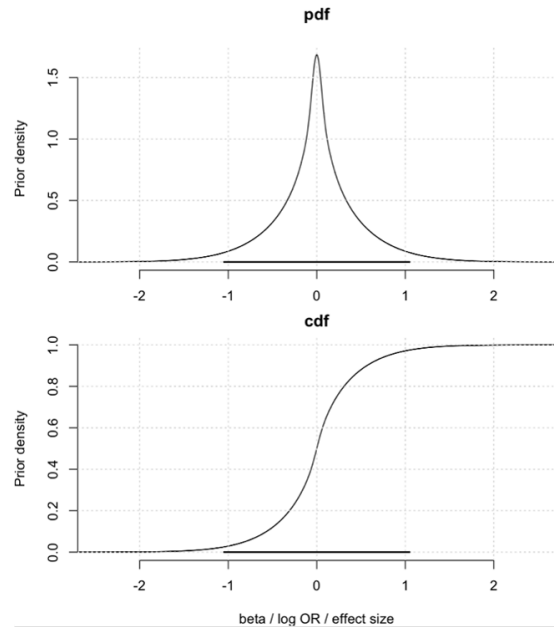


Figure 2.2: Prior on effect for a meta-analysis of malaria severity subtype in BINGWA The probability density function (pdf) and cumulative density function (cdf) of the prior on effect used in the Bayesian meta-analysis of the effect of parasite genetic effects on severe malaria subtype. This was generated using mixed gaussians with standard deviations ranging from 0.05 – 1 in intervals of 0.05.

probability mass within a close range of 0 (i.e. no effect), constructed as a mixture of gaussians (Fig. 2.2 on the preceding page).

2.7.5 Modelling host-parasite genetic associations using HPTEST

A genome-wide scan for associations between parasite biallelic SNP variation and human loci (HbS genotype or HLA-B*53 status) was conducted in Chapter 5 (pg.148). The GAMCC/CP1 call set (see Section 2.6.1 (pg.52)) was used to perform these analyses. *Pf* genotype calls and HbS calls from Ghana (made for a subset of the MalariaGEN Pf7 and presented by Hamilton et al. (131)) were acquired from the original authors. Briefly, samples denoted as "QC-pass" in the Pf7 resource which had been successfully typed for HbS status were included. Biallelic QC-pass variants from the Pf7 release were made available, and further variant filtering was performed as described in Section 2.6.3.2 (pg.56).

This analysis utilized the software HPTEST (version 2.1.9) (130), which performs logistic regressions within a Bayesian framework, with human genotype as the predictor and parasite genotype as the outcome variable. No covariates were included in these computations. The default prior of $LogF(2, 2)$ was applied to the effect of the HbS genotype on parasite genotype as described in the original study. This prior is weakly informative, providing mild regularization of effects, which is particularly important when either the host or parasite allele frequency is low, such as with the HbS allele. Despite imposing more stringent regularisation than, for example, a $LogF(1, 1)$ prior, $LogF(2, 2)$ still accommodates relatively large effect sizes whilst reducing the risk of overfitting and overestimating effects.

HPTEST outputs a BF as well as a P-value for each variant. The BF represents the ratio of the probability of the observed data given the association is true (A) relative to the probability given the association is false ($\neg A$) (i.e. $\frac{P(data|A)}{P(data|\neg A)}$). The posterior probability ($P(A | data)$) can be expressed using the BF and the prior probability $P(A)$ as follows:

$$P(A | data) = \frac{BF \times P(A)}{BF \times P(A) + (1 - P(A))}$$

Or, on an odds scale:

$$\text{posterior odds}(A \mid \text{data}) = BF \times \text{prior odds}(A)$$

Posterior odds can be understood as how much more likely the hypothesis is to be true than false, and depends both on the data and on the assumed prior odds of association.

2.8 Other statistical analyses

2.8.1 Estimation of population relative risks using malaria cases and controls

By comparing frequencies of HbS in malaria cases and non-malaria population controls it is possible to estimate the protective effect of HbS on disease. This is commonly quantified as a relative risk (RR). In the analyses presented in Chapter 5 (pg.148), HbAS and SS genotypes were grouped such that focus was placed upon the ratio of the probability of developing malaria in individuals carrying the HbS allele to the probability in those without:

$$RR = \frac{P(\text{disease} \mid E = 1)}{P(\text{disease} \mid E = 0)}$$

Here $E = 1$ stands for ‘individual has HbAS or SS genotype’ and $E = 0$ stands for ‘individual has HbAA genotype’. The RR is measured with regards to a baseline of disease in non-HbS carriers ($E = 0$). An RR less than 1 indicates individuals with the HbS allele are less likely to develop malaria compared to those without it.

In Chapter 5 (pg.148) an estimation of how the infecting parasite genotype alters this protection is presented. This can be determined by making similar comparisons of HbS frequency to population controls but stratified by possible parasite genotype combinations. The RR of disease observed with a particular parasite genotype (g) in population S is:

$$RR_{E=1}(g) = \frac{P(\text{disease}_g \mid E = 1, S)}{P(\text{disease}_g \mid E = 0, S)}$$

Using Bayes Theorem this can be rearranged into an odds ratio:

$$RR_{E=1}(g) = \frac{P(E = 1 \mid \text{disease}_g)}{P(E = 0 \mid \text{disease}_g)} \div \frac{P(E = 1 \mid S)}{P(E = 0 \mid S)} = OR$$

This expression implies that $RR_{E=1}(g)$ can be estimated by comparing the frequency of HbS genotypes in a sample of cases involving parasites carrying genotype g (first term) to the frequency in a control sample from the general population (second term). In practice $RR_{E=1}(g)$ can be estimated across parasite genotypes g using multinomial logistic regression. To implement this, custom python code was created to perform multinomial logistic regression, using E as the predictor variable, and parasite genotypes at specified *Pfsa* loci as the outcome levels.

Population controls for this analysis were sourced from several datasets: the CP1 resource included non-malaria cases from Kenya ($N_{controls}^{Kenya}=4,046$) and Gambia ($N_{controls}^{Gambia}=2,784$) sampled from nearby the collection at a similar time point (130). For these samples, HbS was directly genotyped using the Sequenom MassArray platform. Controls for the analysis of Ghana were taken from the MalariaGEN release of human severe cases specifically from Navrongo in Northern Ghana (ethnicity labelled “Noguchi”) where the malaria cases within the Ghana dataset were also collected ($N_{controls}^{Ghana}=918$) (6). The GAMCC resource lacks population controls with recent HbS genotyping, therefore 1,000 controls from the CP1 Gambia dataset were used for the analysis, while an independent set of 1,000 was used for CP1 Gambia. The CP1 Gambia samples were sampled in a similar region of the country (Banjul, Gambia) but are separated by between 6 and 21 years (Fig. 4.16 (pg.142)). However, HbS frequencies are thought to remain comparable over these time scales within modern human populations (156), so we find it likely that these controls are appropriate.

It is important to note that, because malaria cases in individuals who carry HbS is rare, estimating RR for certain parasite genotypes using the framework described is impeded by small or non-existent sample sizes. To endeavour to make

these estimates for rare parasite genotypes with minimal data, and to prevent overfitting of the multinomial model, a Bayesian regularisation method was used (as described in (6)). This was implemented using Stan with a Gaussian prior on the effect with a mean of 0 and a standard deviation of 2, and the correlation between parameters across the model set to 0.5. For genotype combinations with low sample counts, the resulting parameter values can be interpreted as plausible estimates given the current data, but larger samples would be needed to fully specify these effects.

2.8.2 Assessing the accuracy of the HLA-B*53 typing assay

The presence of HLA-B*53, previously called for samples in the GAMCC, were determined using a PCR and oligo-hybridization based assay (described further in Chapter 5 (pg.148)). These were used to report an association between HLA-B*53 and protection from severe malaria (52). HLA sequencing was recently performed on a subset of the GAMCC resource which can be used to validate the original calls (Section 2.4.1 (pg.42)). The results of this are shown in Chapter 5 Table 5.4 (pg.183). The probability of a sample genotype (G) being HLA-B*53 ($G = 1$) or not ($G = 0$) given the observed genotype (O), determined by the PCR assay, was directly estimated using counts of true positive (TP), false positive (FP), true negative (TN) and false negative (FN) calls within the retyped set.

$$P(G = 1 \mid O = 1) = \frac{TP}{TP + FP}$$

$$P(G = 1 \mid O = 0) = \frac{FN}{FN + TN}$$

$P(G = 1 \mid O = 1)$ is the probability that a sample is truly HLA-B*53 positive given that the PCR assay called it as such (74%). $P(G = 1 \mid O = 0)$ is the probability that a sample is truly HLA-B*53 positive given that the PCR assay called it as negative (6%).

In order to test how this level of accuracy affected the reported association of HLA-B*53 to severe malaria protection, two approaches were taken: using simulations, and using genotype probabilities within GWAS software. For the first method, simulations were generated using the above probabilities as described in Section 2.7.3 (pg.61). For each simulation a logistic regression of disease status against HLA-B*53 status was made and the empirical distribution of maximum likelihood estimates of effect sizes displayed in Chapter 5 (Fig. 5.15 (pg.182)) by its average (mean) and 95% confidence interval (CI)s. For the second, regressions were performed using the programme SNPTEST using genotype probabilities (GP field) instead of genotype calls (GT field) as predictor variables. The SNPTEST model incorporates genotype uncertainty internally, summing over the unobserved possible true genotypes and making estimates accordingly.

2.9 Population genetic analyses

2.9.1 Measuring complexity of infection in *Pf* WGS data

The complexity of infection (COI), that is, the number of genetically distinct parasites within a host, can be indirectly assessed using the Fws statistic, which compares within-host heterozygosity (denoted H_w) to within-population heterozygosity (denoted H_s). Heterozygosity measures how diverse a population or infection is. It's calculated using the Hardy-Weinberg principle:

$$H_w = 2p_wq_w$$

$$H_s = 2p_sq_s$$

where p and q represent allele counts at a biallelic SNP within the sample (p_w , q_w) and within the population (p_s , q_s). We expect each sample to be made up of a sub-sample of the entire population, and therefore $H_w < H_s$. Fws is calculated using the following formula:

$$Fws = 1 - \frac{H_w}{H_s}$$

To calculate Fws genome-wide the gradient of the linear relationship between H_w and H_s for each sample is used (157). The Fws value falls between 0 and 1, with values near 1 indicating an almost clonal infection (a commonly used threshold is >0.95), and values close to 0 representing an abundance of unrelated isolates present within an infection (158). Fws values within this thesis were calculated jointly across all variants within the intersected call set (see Section 2.6.2 (pg.53)).

The software coiaf was also used to calculate complexity of infection (COI) directly from WGS data and was specifically designed for malaria parasite analysis (159). The programme was used across 6,477 loci within the intersected call set using biallelic SNPs which were at least 1000bp apart to ensure independency of sites. Continuous and discrete COI values were output by the programme, and the continuous measures were used within Chapter 4 (pg.110) as they are reported to be more reflective of the varying levels of parasite relatedness across the genome which occur over successive generations of coinfection (159).

2.9.2 Modelling allele frequency shift using the X^tX statistic

Genome wide allele frequency differentiation in *Pf* parasites from Gambia presented in Chapter 5 (pg.148) was determined using a model similar to Bayenv (160) adopting simplifications first presented in Band et al. (6).

We can imagine that a sample population have evolved from an ancestral population (time point t_0). A very simple approximation of genetic drift since this time point within a single population can be made using the normal distribution. Let θ_i be the population allele frequency of a derived allele at variant i at a certain timepoint (t_x). Therefore, in a sampled population at t_x :

$$\theta_{it_x} \sim N(\theta_{i0}, \sigma_{t_0}(\theta_{i0}(1 - \theta_{i0})))$$

The parameter δ represents drift between time point t_0 and t_x . δ can therefore be estimated as variance. Bayenv expanded this model to encompass multiple populations by estimating this drift a covariance matrix (denoted W) using genome wide data which reflects the effects of shared ancestry between those populations.

The original implementation of Bayenv employed a Monte Carlo Markov Chain scheme to compute their estimations. The simplified version used within this thesis was as follows.

Let F denote the vector of true allele frequencies in the population at the sampled time points. In the formulation above, F can be modelled as a multivariate normal depending on the ancestral allele frequency:

$$F \sim MVN(f_0, f_0(1 - f_0)W)$$

Here f_0 represents the ancestral allele frequency (i.e. the frequency at timepoint t_0) within Gambia, and the matrix W represents the covariance effects of genetic drift acting between f_0 and populations at later timepoints. This formula can then be re-written in terms of transformed frequencies denoted F' :

$$F' = \frac{F - f_0}{\sqrt{f_0(1 - f_0)}} \sim MVN(0, W)$$

An important point is that W is assumed to be the same for all variants. In principle, if the ancestral frequencies f_0 of a set of variants are known, the matrix W can therefore be estimated as the covariance of the transformed frequencies on the left-hand side.

In the approach taken here, f_0 was empirically estimated from a set of samples collected in 1984 (t_0) in the Pf7 resource (N=140), which were the earliest available samples from The Gambia. To estimate the matrix W , a random selection of biallelic, common variants across the *Pf* genome was used (these are treated as representing neutrally evolving variation). Allele frequencies for each variant were estimated within the GAMCC resource (“old” timepoint t_1 (1988-1989), N=814) and within a more recent set of samples in the Pf7 resource (“contemporary” timepoint t_2 (2013-17), N=604). These were then transformed using the formula $F' = \frac{(F-f_0)}{\sqrt{f_0(1-f_0)}}$. Finally, an empirical estimate Ω of W was computed as the covariance of these scaled frequencies:

$$\Omega = cov(F')$$

The model estimated above provides an empirical ‘null’ model of genome-wide allele frequencies across time points, taking genetic drift into account. Using this, it is possible to test specific variants against the model to determine whether unusual levels of allele frequency shift have occurred. This is achieved by computing an X^tX statistic and P-value, as follows. For a given variant, the empirical model states that:

$$F' \sim MVN(0, \Omega)$$

To compute a P-value, the values in F' can be transformed into independent Gaussian variables using the inverse Cholesky decomposition of the estimated matrix ω' and the scaled allele frequencies F' :

$$G = (g_i) = L \cdot F \text{ where } LL^t = \Omega^{-1}$$

The entries of G now follow independent univariate Gaussian distribution; if we define the X^tX statistic as $z = \sum_i g_i^2$, then z follows a χ^2 distribution (with 2 degrees of freedom since two time points are tested in our implementation). A P-value can now be computed as the mass under the tail of this chi-squared distribution.

For each variant, the X^tX P-value indicates how closely the variant aligns with neutral frequency variation. The approximation outlined above is expected to hold when ancestral allele frequencies are not too close to zero or one. To assess this, X^tX statistics were plotted separately for the allele frequency (AF) bins $0 < AF \leq 0.05$, $0.05 < AF \leq 0.95$, and $0.95 < AF < 1$. This simplified model appeared to describe data well for the 5-95% allele frequency bin (Fig. 2.3 on the next page) and therefore, the primary analysis was restricted to these variants only.

This method models the old and contemporary time points as having diverged from f_0 as independent lineages which is not a true reflection of time. This model could be improved by jointly modelling differentiation across connected time points ($f_0 \rightarrow \text{old} \rightarrow \text{contemporary}$). This was not attempted here to avoid complexity, but the results of this analysis should therefore be appraised with this limitation in

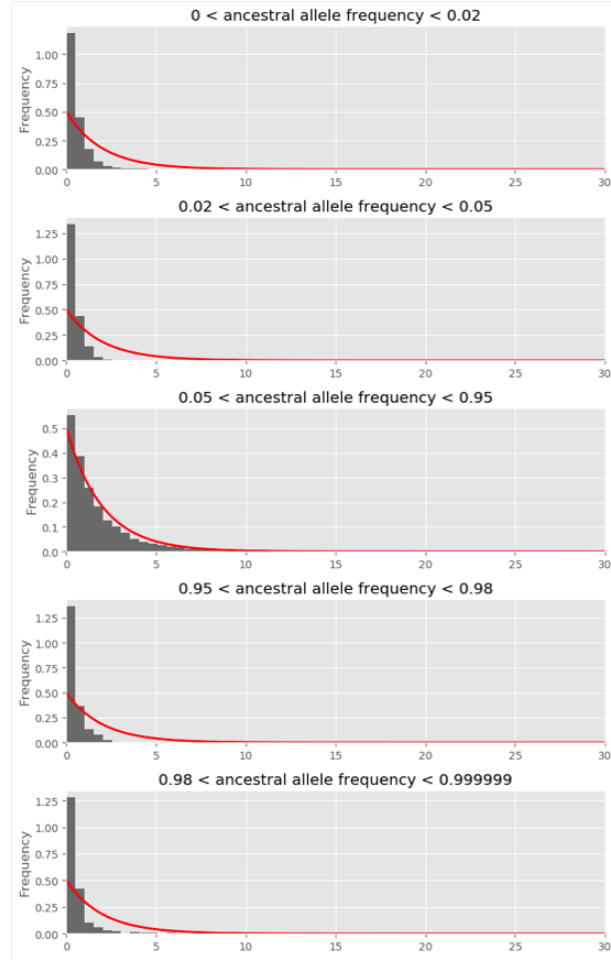


Figure 2.3: $X^t X$ statistics from an empirical model of allele frequencies in The Gambia for different allele frequency bins. Histograms of the $X^t X$ statistic for variants in the 0-0.02, 0.02-0.05, 0.05-0.95, 0.95-0.98 and 0.98-1 allele frequency bins. Red lines show the χ^2 distribution where degrees of freedom = 2. This theoretical distribution matches the observed distribution well for the 0.05-0.95 allele frequency bin.

mind. It is also worth noting that samples in the Pf7 resource are heterogenous and may come from different types of infection (e.g. clinical illness or asymptomatic) and from different times of year which vary in transmission intensity (37). These factors may lead to allele frequency differences that impact the present results. Despite these caveats, should variants present with unusual X^tX values, it is likely this represents a potential deviation from genome-wide drift which warrants further inspection.

2.9.3 Computing between chromosome linkage disequilibrium

Calculating between chromosome linkage disequilibrium (LD) iteratively between every variant pair in the *Pf* genome (a total which scales quadratically with number of tested variants) would be time consuming and computationally laborious. To overcome this, a series of matrix operations were designed.

To illustrate, let A and B denote matrices of genotypes (as 0s, 1s and missing values) for the first and second chromosomes in a comparison pair. Matrix A has the dimensions $m \times n$, where m represents the number of variant sites on the first chromosome and n represents the number of samples. Matrix B has dimensions $p \times n$, where p is number of variants on the second chromosome. Correlation is calculated using the formula:

$$Cor(A, B) = \frac{Cov(A, B)}{sd(A)sd(B)}$$

Where $Cov(A, B)$ is the covariance of the two matrices and $sd(A)$ and $sd(B)$ are the standard deviations of the two matrices. The covariance matrix can be calculated as:

$$Cov(A, B) = \frac{\sum(A - \bar{A})(B - \bar{B})}{N}$$

By this implementation, each variant (row) within A and B were first mean centred and divided by their standard deviation to form matrices X and Y respectively. Then an unnormalised correlation was calculated by:

$$uCor = X \cdot Y^t$$

This was calculated in a single computation using the matrix operations package ‘numpy’ within the programming language python. The genetic data used in these calculations often contains missing genotypes. This can prevent computations being made, so to get around this masked numpy arrays were used, which prevent missing values affecting the calculations of covariance or standard deviation. Resulting values must then be normalised by sample size (N) to produce a Pearson’s correlation r value. In order to get a count of samples where both genotypes were not missing, the following was used:

$$count = A \cdot B^t$$

The count matrix of N values has the same dimensions as $uCor$ ($m \times p$). Finally, a matrix of Pearson’s correlation r values for each pair was computed by: $Cor = \frac{uCor}{count}$.

The method described above was used to calculate LD between chromosomes for all common (>5%) biallelic SNPs with low levels of missingness or heterozygosity (<20%). To gather LD signals into independent ‘regions’ of LD, SNPs were iteratively grouped as follows. First, pairwise LD values were filtered to those where $|R| > 0.35$. Then, for example, if site i on chromosome a (SNP _{ai}) was correlated with site i on chromosome b (SNP _{bi}), SNP _{ai} would be grouped with a neighbouring site j (SNP _{aj}) if SNP _{aj} was also in LD ($|R| > 0.35$) with SNP _{bi} or a neighbour of SNP _{bi} . A neighbour was considered SNPs within 10Kbs of the focal SNP.

2.9.4 Determining ancestral allele frequencies in the *Pf* genome

Ancestral alleles were determined using the software *est-sfs* (version 2.04) (161). Firstly, a multiple sequence alignment of the *Pf*3D7 reference sequence, an assembly for the related Gorilla malaria parasite *P. praefalciparum* (PfG01), and an

isolate of the chimpanzee infecting *P. reichenowi* (PrG01) was computed using the software mafft for each chromosome (using the command line argument — auto and using four threads). The nucleotide base at positions of VQSR pass variants in the MalariaGEN Pf6 resource were input into est-sfs, formatted such that *P. praefalciparum* was considered the first outgroup (the closer relative to *Pf*) and *P.reichenowi* a more distantly related outgroup (132). Based on this, ancestral state probabilities were determined using the maximum likelihood method and the Kimura 2-parameter model. This models the rate of transitions (substitutions between pyrimidines or between purines) and the rate of transversions (purine to pyrimidine and *vice versa*) across the genome in order to estimate genetic distances. The nucleotide type with the highest probability was determined to be the ancestral state allele.

2.9.5 Extended haplotype homozygosity (EHH) and integrated haplotype score (iHS) analyses

Under a basic model of evolution, mutations arise once on a single haplotype. If beneficial, this mutation and thus variants in its surrounding haplotype will undergo positive selection and increase in frequency. This is called genetic hitch-hiking (82). Extended lengths of shared haplotypes that are identical by descent (IBD) within a population can therefore be indicative of recent positive selection in that genomic region, a phenomenon measured by extended haplotype homozygosity (EHH). Over time, haplotypes are broken down by recombination events, such that the length of sharing reflects the evolutionary history of the allele.

Let n_a represent the total sequences carrying allele a at position s . Considering the genomic region between position s and position t , let K denote all distinct haplotypes in the sample, and K_a the proportion of these which carry allele a . EHH is the probability that two randomly chosen haplotypes both carrying allele a at locus s are IBD across the genomic region between s and t (as inferred by homozygosity at all intermediary SNPs) (82). EHH can therefore be defined across the region as:

$$EHH = \frac{1}{n_a(n_a - 1)} \sum_{k=1}^{K_a} n_k(n_k - 1)$$

Where n_k represents the number of sequences sharing haplotype k . The denominator of this equation represents the number of ways to select two sequences carrying allele a , whilst the numerator represents the number of ways to choose matching haplotypes that are IBD (the 2 within both elements of the equation have cancelled each other out). EHH starts at 1 near the focal allele, and decays as you move away from a focal locus due to recombination events.

EHH can be computed at any distance out from a focal allele. In order to compare a single-number summary of EHH, the integrated haplotype homozygosity (iHH) is typically computed, that is, the area under the EHH curve for each allele. In practice, iHH for the ancestral allele is usually compared to iHH for the derived allele at any locus, controlling for local variation in recombination or mutation rates, giving the metric:

$$uIHS = \log \frac{iHH_{derived}}{iHH_{ancestral}}$$

This is referred to as the un-normalised integrated haplotype score (iHS), and this indicates which allele may be under selection.

Under pure genetic drift, the value of uIHS will depend on the allele frequency (since, typically, rarer alleles tend to be younger and therefore have larger IHH values). Therefore, to detect selected alleles it is usual to work with standardised uIHS values within allele frequency bins (162).

In this implementation, EHH, iHH and iHS was calculated across the *Pf* genome over all biallelic SNPs with a minimum minor allele frequency of 5%. *Pf* genetic data is called as diploid due to high levels of mixed infections which contain multiple parasite genetic versions (163). Although methods exist to divide these into separate haplotypes (e.g dEplod (164)), these are statistically complex and challenging to validate. For simplicity in this analysis, a filtered variant call set developed from the Pf7 resource was used which contained entirely homozygous non-missing haplotype calls, which had been polarised to represent derived and ancestral alleles (described in Section 2.6.4 (pg.57)). Diploid genomes were then converted to haploid genomes and re-formatted into “.hap” and “.map” file formats. EHH and iHS was then calculated for every variant across each chromosome using the R package rehh within 2.5% allele frequency bins (165). Data was

treated as “phased” (as each *Pf* whole genome sequence had been simplified to a haploid form).

2.10 Structural variant calling

2.10.1 Approximating structural variant haplotypes at the CLAG3.1 and CLAG3.2 locus

Within Chapter 6 (pg.225), a potential association between variation in the paralogous genes CLAG3.1 and CLAG3.2 and HbS is described. A deletion between the CLAG3.1 and CLAG3.2 genes on Pf3D7_03_v3 forming a CLAG3-hybrid (termed CLAG3h) has been previously identified (166). To assess this in available long-read assemblies, a region surrounding the annotated CLAG3.1 and 3.2 genes in the 3D7 reference assembly was extracted alongside the aligning regions of other assemblies within a multiple sequence alignment (described in Section 2.5.4.1 (pg.50)). Based on this, kmer sharing across the region was computed, which confirmed the presence of high paralogy between sequence within CLAG3.1 and 3.2 (shown in Chapter 5, Supp. Fig. 5.30 (pg.213)). This also confirmed the presence of CLAG3h; for example, the laboratory isolate GB4 and the clinical isolate KE01 (from Kenya) carry the deletion (Supp. Fig. 5.30) (103).

Inspecting these plots, this deletion was observed to remove two intermediary genes Pf3D7_0302300 and Pf3D7_0302400 within these assemblies (Supp. Fig. 5.30 (pg.213)). To assess this within short read data, distributions of the proportion of missing variants across these genes were plotted (from the start of deleted gene Pf3D7_0302300 to the end of Pf3D7_0302400) within call sets of each assessed cohort. Missingness across this region followed a bimodal distribution with samples generally having either a very high proportion of missing calls at variant sites (as expected if a deletion across this region was present) or a very low proportion of missing variants (expected if this sequence was not deleted in the sample) (Fig. 2.4 on the next page). Based on these distributions, samples with a proportion of missing variants >0.7 across this region were determined to carry the deletion, and samples where this proportion was <0.2 were not, as a means to generate a crude

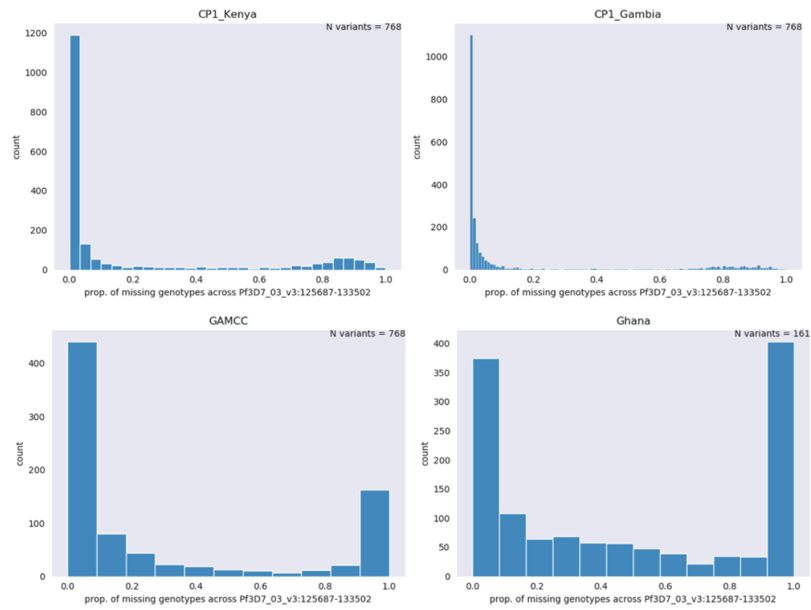


Figure 2.4: **Calling the presence of the CLAG3h hybrid in four datasets.** Distributions of proportions of missing variants in the GAMCC/CP1 callset for analysis samples in the CP1 Kenyan, CP1 Gambian and GAMCC cohorts and the MalariaGEN Pf7 callset of Ghanaian samples. The number of variants across which this proportion is calculated is in the top right corner.

method by which to call CLAG3h in these samples. Samples with proportions between these values were excluded.

This led to a significant reduction in samples for the GAMCC and Ghanaian cohorts which contained many samples with intermediary values (Fig. 2.4), perhaps due to levels of mixed infection. It should also be noted that the Ghanaian call set contained fewer variant sites than the GAMCC/CP1 call set, and so calling within this data was across a smaller number of sites. Consequently, calls for this dataset may be less accurate. This method serves as an initial approach to investigate this region, and future studies will employ more advanced techniques to resolve this region with greater confidence.

2.10.2 Analysis of structural variation at the *Pfsa3* locus

The most HbS-associated variant in the *Pfsa3* region was Pf3D7_11_v3:1058035:T>A. A region of roughly 6Kb flanking this SNP (Pf3D7_11_v3:1053750-1060000 in co-ordinates of the 3D7 reference assembly) has been indicated to carry structural variation in a set of 15 published long read assemblies. The two SV haplotypes present in these assemblies, for the purposes of this work, have been named DUP-B1 and DUP-B2. As described in Chapter 6, one objective was to assess structural variation at the *Pfsa3* locus throughout *Pf* parasite populations. To do so, a number of statistical methods were generated to call SVs from short-read sequencing data. For this analysis, MalariaGEN CP1 BAM files were downloaded from the original release and a subset of alignment files from the MalariaGEN Pf6 dataset were recieved directly from MalariaGEN which are gratefully acknowledged.

Overall, three calling methods were generated. The intersection of calls made using the coverage and insert size callers (described below), as described in Chapter 6, were used to build a ‘consensus call set’ in the MalariaGEN Pf6 and CP1 datasets. During the latter stages of this work, a third method to call structural variation was developed called the kmer call which was not used to develop the consensus set.

2.10.2.1 Sample selection for these analyses

sWGA has recently become a commonly used method of *Pf* DNA sample preparation for WGS (see Section 2.2.3.1 (pg.37) for a further description of the method). The sWGA data type now makes up the bulk of publicly available *Pf* WGS data (making up >60% of the MalariaGEN Pf7 resource (37)) but a set of non-sWGA data can be found in the MalariaGEN Pf6 release (N=7113) and a subset of MalariaGEN CP1 samples (N=1181). Non-sWGA data was used for this work as coverage is thought to be more even in non-sWGA data (as depicted in Chapter 3, Fig. 3.2 (pg.92)), facilitating the development of the structural variant callers.

The previously published Pf6 “QC pass” and CP1 “analysis set” (which were not prepared using sWGA) were used for sample selection (details of these filtering processes are in Section 2.6.3.1 (pg.54). Pf6 used the same sample filters as Pf7 (167)). In addition, both sets were further filtered to have a percent callable metric (described in Section 2.6.3.1 (pg.54)) $\geq 80\%$ ($N^{Pf6}=5380$, $N^{CP1}=518$). This was to ensure there was likely to be sufficient coverage across the *Pf*sa3 region to allow the a call to be made.

2.10.2.2 Coverage call

This model was implemented by Dr. Band and will be summarised here. In brief, the caller works based on a set of read alignment files (BAM files) from a population of samples. First, per-base read coverage values were computed for each sample generating an $L \times N$ coverage matrix X (where L is the number of bases in the region, and N is the number of samples). Given a particular set of structural variant genotypes, which are input into the caller, the per-base coverage values are then modelled as a gaussian distribution with mean that depends on the total coverage for the sample and on the local copy number of the sample.

In practice, a great deal of variation in coverage is observed in *Pf* data due to effects of AT rich genome and amplification artifacts. To deal with this an iterative approach was used as follows. In the first iteration it was assumed that all samples carry the reference-like (3D7) haplotype structure. To serve as a proxy for per-base variation in coverage due to amplification and sequencing artifacts,

we computed the per-base coverage averaged over samples (after normalising each sample by the overall sample coverage, reflecting sequencing depth). All coverage calculations were then performed after dividing each sample's coverage by this per-base mean coverage. This substantially homogenised coverage between bases as shown in Figure 2.5. For any given copy number profile we then modelled this normalised coverage using a Gaussian distribution with mean and variance parameterised by the sample's mean coverage and the copy number at each site, leading to a likelihood for each SV in each sample. This generates a new posterior estimate made of the sample genotypes. This process was then repeated, with the per-base mean coverage dependent on the updated genotype distributions. After 5 iterations, the posterior genotype probabilities for each sample were determined.

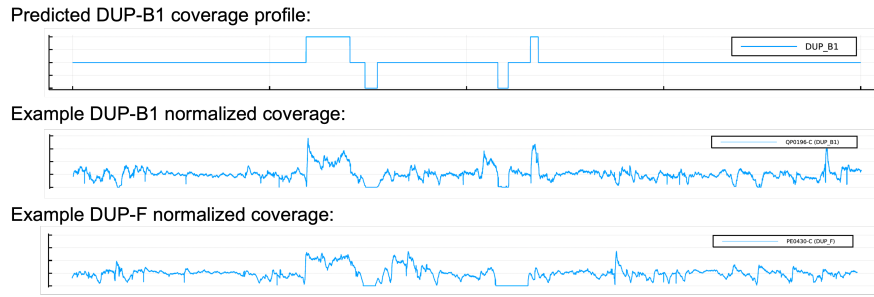


Figure 2.5: **Normalised coverage across the *Pfsa3* locus used to identify structural variation.** These plots are generated as part of the coverage caller. The top panel shows the coverage profile predicted for the DUP-B1 haplotype. Below shows coverage for two real samples which were identified as carrying the DUP-B1 and DUP-F haplotypes.

Following this process, the scaled coverage values for all samples were plotted and visually compared to predicted coverage profiles. This revealed the presence of visually discordant samples (for an example of how a DUP-F haplotype differs from DUP-B, see Fig. 2.5). These samples were examined for the presence of split read pairs in Integrative Genomics Viewer (IGV). Using this method, samples with additional structural variant types were identified, including deletion types that were not present in the current genome assemblies. These types were added to the catalogue and the calling process repeated until no obvious discordant samples were found. The set of structural haplotypes present is shown in Chapter 6 Figure 6.1 (pg.232).

2.10.2.3 Insert size call

In addition to the coverage caller described in the previous section, a further method was developed using an independent information source. This caller models the distance between read pairs (the insert size). First, for each SV haplotype, a dummy reference sequence was constructed, representing the structure of this haplotype. For each sample, read pairs which aligned to breakpoints within the region were extracted from the BAM file and realigned to each false reference using bwa-mem and the “-a” flag to allow for more than one alignment location. Insert sizes (i.e. the distance between the aligned location of the start of read 1 and the end of read 2 in each pair) were modelled as a Gaussian distribution:

$$(s_1 - s_2) \sim N(\mu, \sigma^2)$$

Here, s_1 and s_2 are the first positions of the alignment of each read pair on the false reference. The mean (μ) and standard deviation (σ^2) insert size for each sample was learned from the alignment of all reads to the Pf3D7 reference (which generally contain little structural variation). A likelihood for read pair r (r_1, r_2) on each false reference was generated by summing over all possible alignment locations:

$$P(r|\text{genome}) = \sum_{s_1, s_2} P(r_1|s_1)P(r_2|s_2) \cdot P(s_1 - s_2|\mu, \sigma^2)$$

If read pairs are being aligned to a reference which mirrors the sample they originate from, they should align within the standard fragment length distribution. The probability formulated above will therefore be large for at least one possible alignment location of that read pair. Conversely, if reads are aligned to a reference which does not emulate the sample they were sourced from, reads will not fall within the learned distribution (for instance, they may align too close to each other, or too far apart). This will cause the probability above to be small.

For each read pair, $P(r_1 | s_1)$ was estimated by the alignment score output by bwa-mem. Read pairs were treated independently; a likelihood for each dummy reference was calculated through summing likelihoods for all read pairs on each

reference. The dummy reference with the maximum likelihood was then determined to be the genotype of the sample. Thus, in contrast to the coverage caller, which uses an average across the entire sample to perform its call, the insert size caller processes one sample individually.

Implementation of mixture handling for the insert size caller Due to the possibility of an individual being infected with multiple different genetically distinct parasites carrying different SV haplotypes, a simple implementation of this model with mixture handling was also developed. One can imagine that genome G is a mixture of two genomes $G = (G_1, G_2)$ with mixture proportions α_1 and $\alpha_2 = 1 - \alpha_1$. Therefore:

$$P(r|G) = \sum_{i=1,2} P(r|G_i)P(G_i|G) = \sum_{i=1,2} P(r|G_i) \cdot \alpha_i$$

However, calls using this calculation were impossible to replicate during the initial stages of development of Chapter 6 (pg.225), as there was no similar determination by the coverage caller. Therefore, although this has not been widely tested, it will form the basis of future refinements of these methods.

2.10.2.4 Kmer call

This method is based upon the linkage between certain SNP sites to structural variant haplotypes (as displayed in Chapter 6 Fig. 6.4 (pg.240)). This method cannot differentiate all haplotypes within the presented catalogue of structural variation (Fig. 6.1 (pg.232)), but can be used to distinguish between 3D7-like, DUP-B-like, and DUP-F-like samples.

First, lists of structural variant regional kmers (short sequences of nucleotides), which differentiated each haplotype, were generated. To do so, common (AC>30), high quality (VQSR ‘pass’) SNPs were extracted from the SV region (Pf3D7_11_v3:1053960-1059852) in the MalariaGEN Pf6 variant call set filtered for African samples. Then, allele frequencies for samples in the consensus call set determined as 3D7, DUP-B, or DUP-F type were computed. SNPs which differed in allele frequency between these groups by more than 25% were retained as potentially usable

to differentiate between haplotypes. Kmers of 31 bases in length (31mers) centred on these SNPs for both the reference and alternate alleles were then generated for each potential site.

31mer pairs were counted within *Pf* long-read assemblies representative of each SV haplotype (Pf3D7 as an example of no structural variation, a newly generated assembly from Kenya (KENI1) as an example of DUP-B1, FUP-H as an example of DUP-F and CD01 (from (103)) as an example of DUP-B2) using the software jellyfish2 (Fig. 6.3 (pg.235)). These counts were tested to ensure the expected ratio of kmers was returned. For example, a DUP-B type parasite possesses a paralogous sequence variant at Pf3D7_11_v3:1054021:T>C, and therefore the expected ratio of each allele in the kmer pair is 50:50. Pf3D7 however has no duplication and therefore no paralogous sequence variant (PSV) at this site, so we would expect the ratio of reference to alternate centred 31mers to be 1:0. Because the DUP-B, DUP-F, and 3D7 haplotypes each have a distinct pattern of expected kmer count ratios, the likelihood of different mixture proportions between haplotypes can be tested by this caller to give an estimate of levels of mixture within the infection.

To implement this caller, genome wide 31mer counts were computed directly from fastq files using jellyfish2. From the predicted ratios of kmers in assemblies (e.g. 50:50 or 1:0), we can predict kmer ratio proportions in infections of different mixtures of each SV haplotype (for example, at PSV Pf3D7_11_v3:1054021:T>C, an equal mixture of 3D7-like parasites and DUP-like parasites in a single infection, would have an expected ratio at this site of 75:25). For each mixture a log likelihood was then computed as follows.

Let A be the alternate count, T be the total count of alternate and reference 31mers, and E be the expected proportions given the mixture. The log likelihood L for each element in these arrays is then calculated using the probability mass function (PMF) for a binomial distribution:

$$L = \log(PMF(A, T, E))$$

Where PMF is the binomial probability mass function:

$$PMF(k, n, p) = C(n, k) \times p^k \times (1 - p)^{(n-k)}$$

$$C(n, k) = \frac{n!}{(n - k)!}$$

Here, k is the number of alternate kmer counts, n is the total number of alternate and reference kmers, and p is the expected proportion. $C(n, k)$ is the binomial coefficient, which calculates the number of possible combinations of n items taken k at a time. Log likelihoods are then summed across sites, and the mixture with the maximum likelihood determined to reflect that of the sample. Within the mixture, the majority haplotype was taken as the primary call. Samples with a haplotype holding a mixture proportion >95% were deemed clonal whereas those without were deemed mixed.

Currently, within this method only a single site (Pf3D7_11_v3:1054447:A>G) distinguishes DUP-B and DUP-F haplotypes, which limits its ability to determine mixtures between these haplotypes. The kmer caller would therefore benefit from further optimisation and development, such as the addition of missing sites in the R8 segment (Chapter 6 Fig. 6.1 (pg.232)) to differentiate DUP-B and DUP-F with greater confidence.

Chapter 3

Generation of the ‘Genomic Analysis of Malaria Cases and Controls’ (GAMCC) resource

3.1 Introduction

The World Health Organisation has set out in its strategic framework that basic research is fundamental to the intended elimination of malaria. This is because understanding *Plasmodium falciparum* (*Pf*) parasite biology might lead to improvements upon established interventions or the development of entirely novel approaches to tackle disease (47). A foundation of this strategy is to build large datasets of human, parasite and mosquito genetic data. Collecting the samples required for this is both costly and time-consuming. Therefore exploiting historical sample sets, collected prior to the advent of Next Generation Sequencing (NGS), proves an attractive method to generate large genetic data resources whilst also providing a retrospective assessment of genomic diversity capturing the inherent evolution in genomes over time. The Genetic Analysis of Malaria Cases and Controls (GAMCC) is one such historical sample set. It was collected from Gambian infants between 1988-90 and first described by Hill et al. (52) who retained them at the University of Oxford for the 30 years since their collection. This sample set was used to define the widely cited association between a human leukocyte antigen (HLA)-B allele, HLA-B*53 which is common in West Africa, and protection against severe malaria. It was therefore collected with the intent of human

genetic analysis and possesses information for this specific human locus, among other biologically relevant loci such as the haemoglobin (Hb)S sickle allele. Since the samples were collected at the point of malaria diagnosis, the samples also likely contain parasite genetic material. Given the uniqueness of the GAMCC dataset and the opportunities it therefore offered for understanding both parasite and human genetic variation simultaneously, I planned to perform detailed sequencing of the *Pf* parasites within this dataset. This chapter describes the steps taken to generate this data. The first step was to ensure we had appropriate permissions to undertake this analysis. Once I had received renewed ethics for both human and parasite sequencing on the GAMCC set from the Medical Research Council unit in The Gambia, I then proceeded to trial and develop modern sequencing methodologies for these samples, aiming to establish a robust resource for comprehensive analysis.

Generating high quality whole genome sequencing (WGS) data for *Pf* parasites however comes with several challenges. Increasing the proportion of *Pf* DNA content within samples which primarily contain human DNA, such as those in the GAMCC collection, can be achieved either by specifically degrading or removing human DNA, by amplifying the DNA of parasite origin, or by a combination of the two. Examples of human DNA removal methods include the digestion of methylated DNA that is present at low levels within parasite genomes via the MspJI enzyme family (98; 168). The most widely used method for the enrichment of *Pf* DNA is selective whole genome amplification (sWGA) (99). The suitability of sWGA, which was developed and optimised for dried blood spots (DBSs), was uncertain for whole blood phenol-chloroform DNA extractions like those in the GAMCC set, and therefore application of such an approach required optimisation. In the following sections I describe my approach to optimise sWGA specifically for this context, with the aim of generating high-quality malaria parasite WGS data for samples in the GAMCC set (Table 3.2 (pg.101)).

3.2 Results

3.2.1 Enriching *Pf* DNA in GAMCC samples prior to sequencing

The first trial of the sWGA approach involved sequencing four samples from the GAMCC dataset (prepared using the NEBNext Ultra DNA Library Prep Kit (UltraI)) under three conditions which have been reported in the literature as beneficial for parasite sequencing (98). These conditions included using sWGA directly; using sWGA with prior digestion of human DNA using the endonuclease McrBC (a member of the MspIJ enzyme family); and using a non-specific whole genome amplification approach based on random primer sets (random whole genome amplification (rWGA)) following McrBC digestion. As shown in Figure 3.1 on the following page, under the condition of McrBC digestion followed by rWGA, only $10.10 \pm 12.4\%$ of the sequenced read pairs aligned to the *Pf* genome on average. Using sWGA a larger proportion of reads mapped to the malaria parasite reference genome (Pf3D7) ($70.36 \pm 16.1\%$). Digestion using McrBC beforehand showed the greatest proportion of reads mapping to Pf3D7 for all four samples ($83.83 \pm 7.6\%$), although the level of improvement observed following McrBC digestion varied between samples. This indicated that McrBC-mediated human DNA degradation, followed by sWGA-mediated enrichment of *Pf* DNA content produced the highest specificity for *Pf* genome sequencing.

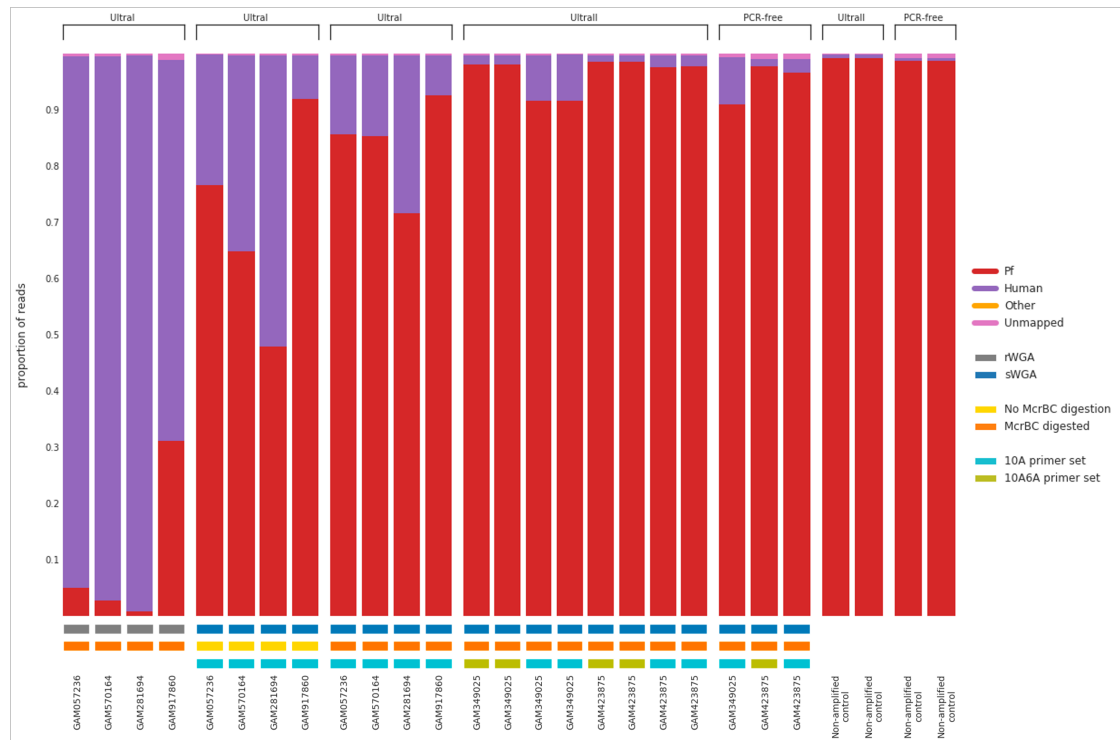


Figure 3.1: Proportions of *Pf* DNA in GAMCC samples prepared using different methods. Proportion (between 0 and 1) of reads aligning to the *Pf* 3D7 reference genome ('Pf', red), the Hg38 human genome (purple), other aligned-to genomes (PhiX-174) (orange) and unmapped reads (pink) for each GAMCC sample (identifiers starting with 'GAM') and non-amplified control DNA extracted from culture of the *Pf* clone 3D7. Library preparation for each sample was performed using the NEBNext Ultra DNA Library Prep Kit (UltraI), the NEBNext Ultra II FS Library Prep Kit (UltraII) or using the NEBNext Ultra II PCR-Free kit (PCR-free), as denoted by brackets at the top of the plot. Conditions for the whole genome amplification (WGA) of the GAMCC samples are shown as coloured rectangles at the bottom of each proportion bar. WGA was performed for the GAMCC samples using primers which are selective for the *Pf* genome (sWGA) (denoted in blue) or random primers (rWGA) (denoted in grey). Prior to WGA, enrichment for *Pf* DNA was performed via digestion of methylated-DNA (which is present at low levels in the *Pf* genome) using the endonuclease McrBC (orange) or was not performed (yellow). For sWGA, the primer sets 10A (light blue) or 10A6A were used (green).

3.2.2 Establishing the optimal library preparation method for the GAMCC set

A known issue with sWGA is that genome coverage is typically highly uneven (99). I therefore assessed the breadth and depth of coverage of our samples compared to four randomly selected Malaria Genomic Epidemiology Network (MalariaGEN) Consortium Project 1 (CP1) samples (which were sequenced at the Wellcome Sanger Institute (Cambridge, UK)). Coverage was assessed across previously defined ‘core’ regions of the *Pf* genome which short-reads are known to robustly align to (see Materials and Methods section 2.5.1 (pg.47) for further detail) (102). In comparison to alignments of CP1 samples (down-sampled to allow direct comparison with presented GAMCC data), I observed relatively poor breadth of coverage within GAMCC samples; for example, on average 12% of base positions in the core regions of the Pf3D7 were uncovered ($11.9 \pm 0.09\%$) and $>12\%$ showed greatly expanded coverage of $100\times$ or more ($12.4 \pm 0.21\%$), much higher than the genome-wide sub-sampled average of $39\times$ fold coverage (Fig. 3.2 on the following page).

Available literature suggested that polymerase chain reaction (PCR) steps which occurred during sequencing library preparation may have biased library content to regions of higher guanine-cytosine (GC) composition (95). Malaria parasites have low genomic GC content of around 19% (48). GC bases are unevenly distributed throughout the *Pf* genome and are concentrated in genic regions (51). In accordance with this we found coverage within the GAMCC (UltraI) samples to be concentrated within genic regions; for example, across Pf3D7_01.v3 exons had a 28.4% greater fold coverage than the total chromosome (with an average of $42.61 \pm 8.17\times$ fold coverage across the chromosome and $56.73 \pm 9.75\times$ fold across regions annotated as exons). The association between gene regions of higher GC content and increased coverage can be visualised in Figs. 3.3 (pg.95) & 3.4 (pg.96).

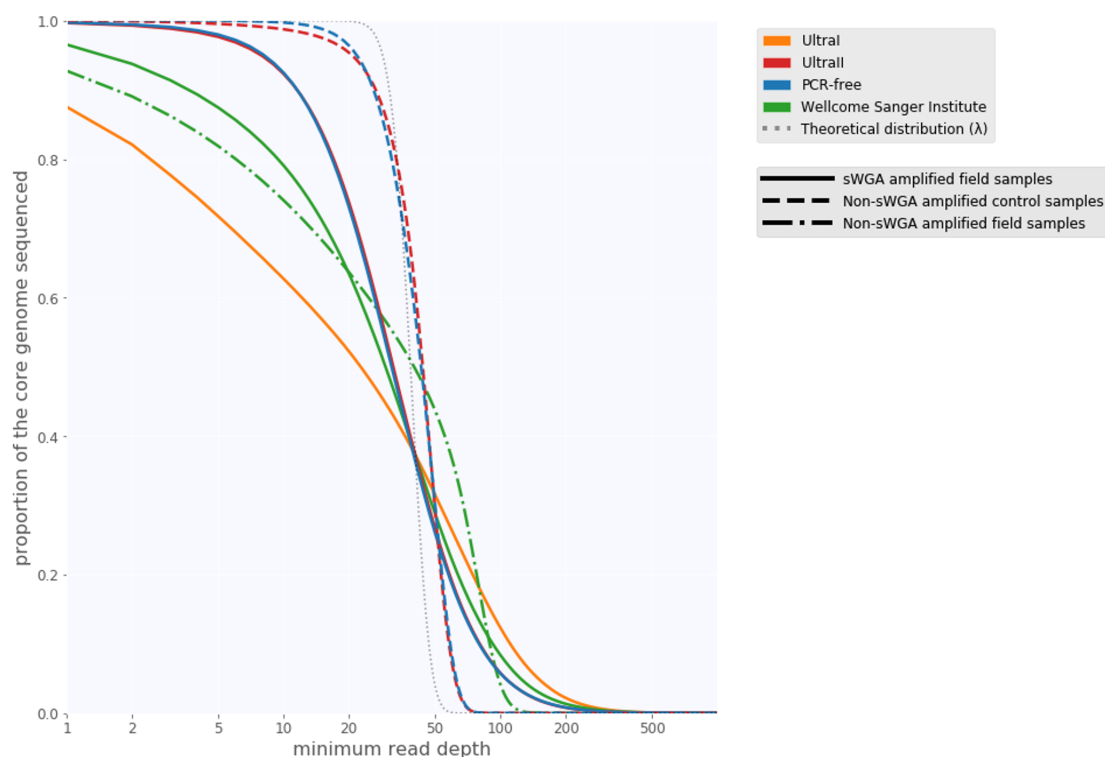


Figure 3.2: Average breadth of coverage of the *Pf*3D7 genome between different sample preparations. Average breadth of coverage of the core regions of the *Pf* 3D7 reference genome for alignments produced from samples sequenced for this chapter using the NEBNext Ultra DNA Library Prep Kit (‘UltraI’, orange), using the NEBNext Ultra II FS Library Prep Kit (‘UltraII’, red) or a NEBNext Ultra II PCR-free kit (‘PCR-free’, blue) or samples sequenced at the Wellcome Sanger Institute (Cambridge, UK) (green). GAMCC samples which have undergone sWGA-amplification are shown as solid lines (UltraI $N = 4$, UltraII $N = 4$, PCR-free $N = 3$), other non-GAMCC field samples which have undergone sWGA are shown as a dotted and dashed line ($N = 4$) and non-amplified control DNA extracted from culture of the *Pf* clone 3D7 are shown as dashed lines (UltraII $N = 1$, PCR-free $N = 1$). A theoretical best-case Poisson distribution, based on uniform sampling of reads across the genome, is also shown as a dotted grey line. The y-axis shows proportion of the genome covered to at least the given minimum read depth shown on the x-axis. All samples were down sampled to a genome-wide average of $39\times$ coverage (6 million reads) to allow for direct comparison.

We therefore evaluated whether alternate library preparation kits improved breadth of coverage. To do this, we repeated sequencing on the same samples from the first round (prepared using UltraI), trialling two additional DNA sequencing library preparation approaches: the NEBNext Ultra II FS Library Prep Kit (UltraII) and a the NEBNext Ultra II PCR-free kit (PCR-free). UltraII utilises

PCR steps but is reported to be less biased by adenosine-thymine (AT) content. Both the UltraII and PCR-free approaches produced alignments with greatly improved breadth of coverage (Fig. 3.2 on the previous page). GAMCC samples sequenced via these methods showed almost identical depth profiles (on average only 0.5% ($\pm 0.19\%$) and 0.4% ($\pm 0.19\%$) of core genome bases were uncovered and 6.0% ($\pm 1.31\%$) and 5.9% ($\pm 1.71\%$) had more than $100\times$ coverage respectively) (Fig. 3.2). Both methods performed better on non-amplified DNA compared to amplified samples, indicating that sWGA itself does introduce bias into the process (Fig. 3.2).

Contamination with human DNA in samples sequenced using the UltraII and PCR-free methods was also less than those sequenced via UltraI; between $3.54 \pm 2.8\%$ and $2.67 \pm 2.6\%$ of reads mapped to the human reference genome respectively (compared to $>14\%$ for comparable UltraI samples) (Fig. 3.1 (pg.90)). This may be due to reduced GC biasing, as the human genome has a much greater GC content than *Pf* (169). The levels of human contamination within the UltraII and PCR-free samples however varied between samples (Fig. 3.1 (pg.90)). This may be related to variation in parasitaemia, direct measurements of which are unavailable for this dataset. Proportions of unmapped reads were also very low across all library preparation methods indicating that other exogenous contamination levels were minimal (Fig. 3.1 (pg.90)). Sequencing via the PCR-free method costed an average of £91.44/sample whereas the UltraII method costed £34.64/sample, prompting the UltraII kit to be the choice method for *Pf* sequencing the entire GAMCC set.

3.2.3 Drop out of regions within sequencing data

Despite the improvements in sequence coverage outlined above, I noted that some regions of the genome were still not captured within this sequence data. To understand these differences in more detail I examined performance in specific regions of interest, including the *Plasmodium falciparum* sickle associated (*Pfsa*)1-3 regions (which have been shown to associate with the HbS-trait in humans (130)) and the liver stage antigen (LSA)-1 and LSA-3 genes (which produce proteins that have been shown to be HLA-B*53 restricted (170)). I observed that GAMCC

samples sequenced via the Ultra II and PCR-free methods had more even coverage across those regions than samples sequenced via UltraI and had more similar profiles to non-amplified controls (Figs. 3.3 on the next page & 3.4 (pg.96)). Unusually however, the read depth across the LSA-1 locus and extending towards the 3' direction was very low for all samples which had undergone sWGA in comparison to non-amplified controls and four randomly selected CP1 samples. Average depth calculated in 100bp windows within UltraII GAMCC samples was only $2\times$ across LSA-1 (Fig. 3.4 (pg.96)). This is despite the moderate GC content within this region. Coverage was also low across the LSA-3 gene region with an average of $9\times$ fold coverage. The *Pfsa1-3* loci showed more than adequate coverage for the UltraII and PCR-free sequenced GAMCC samples with more than an average of $33\times$ across the three gene regions (Fig. 3.3 on the following page).

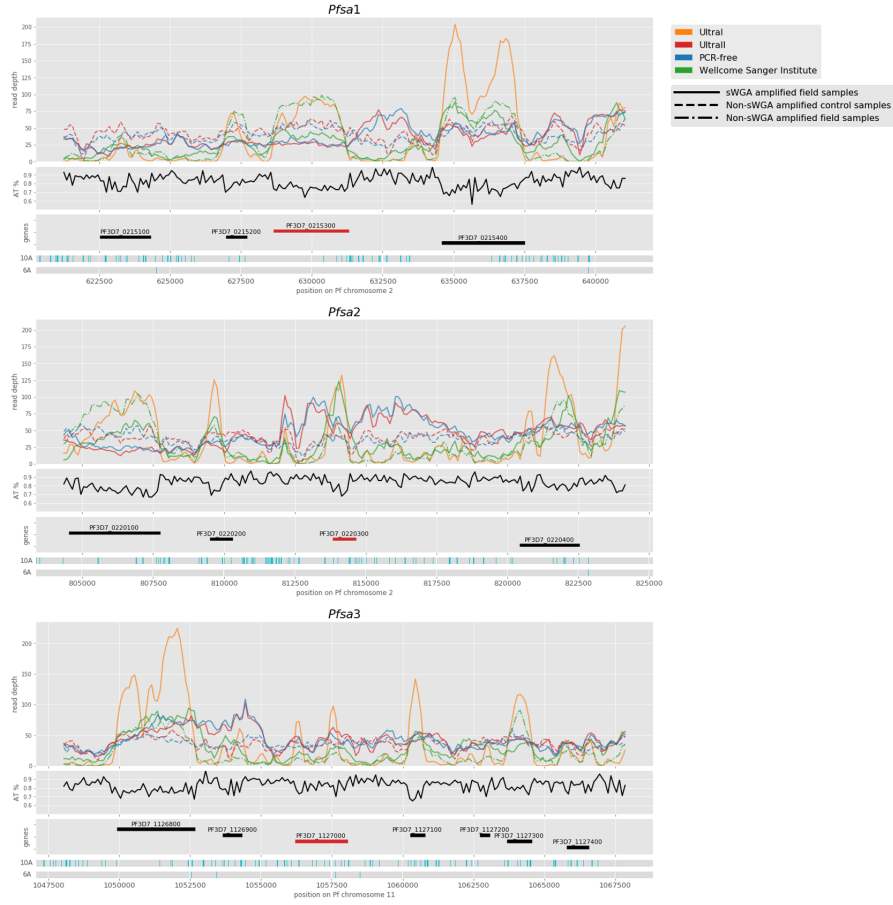


Figure 3.3: Coverage across the *Pfsa* 1-3 regions in the *Pf*3D7 genome between different sample preparations. Read depth across the three *Pf* sickle associated (*Pfsa*) regions and the flanking regions for alignments produced from samples sequenced by Oxford Genomics Core using the NEBNext Ultra DNA Library Prep Kit ('UltraI', orange), using the NEBNext Ultra II FS Library Prep Kit ('UltraII', red) or a NEBNext Ultra II PCR-free kit ('PCR-free', blue) or samples sequenced at the Wellcome Sanger Institute (Cambridge, UK) (green). GAMCC samples which have undergone sWGA-amplification are shown as solid lines (UltraI N=4, UltraII N=4, PCR-free N=3), other non-GAMCC field samples which have undergone sWGA are shown as a dotted and dashed line (N=4) and non-amplified control DNA extracted from culture of the *Pf* clone 3D7 are shown as a dashed line (UltraII N=1, PCR-free N=1). All samples were down sampled to a genome-wide average of 39 $\times$ coverage (6 million reads) to allow for direct comparison. In the second plot from the top AT nucleotide percentage for the genomic region is shown (calculated in 100bp windows). In the third plot from the top regional genes are annotated, with the gene of interest (which contains the top HbS-associated variants in each region) indicated in red. At the bottom, positions of primer binding sites for the 10A primer set (used in all sWGA reactions) and 6A primer set (used in some sWGA reactions sequenced by OGC using UltraII and PCR-free) are shown in blue. Genomic position on the indicated chromosome in the *Pf* genome is the x-axis for all four plots.

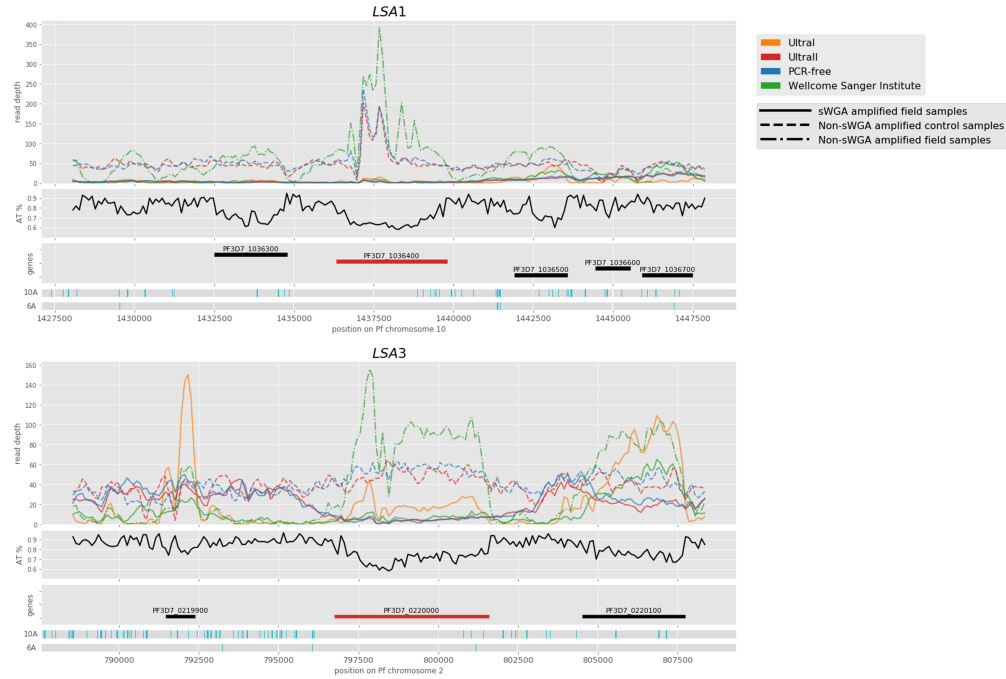


Figure 3.4: Coverage across the LSA-1 and LSA-3 regions in the *Pf*3D7 genome between different sample preparations. Read depth across the LSA-1 and LSA-3 *Pf* genes and the flanking regions for alignments produced from samples sequenced by Oxford Genomics Core using the NEBNext Ultra DNA Library Prep Kit ('UltraI', orange), using the NEBNext Ultra II FS Library Prep Kit ('UltraII', red) or a NEBNext Ultra II PCR-free kit ('PCR-free', blue) or samples sequenced at the Wellcome Sanger Institute (Cambridge, UK) (green). GAMCC samples which have undergone sWGA-amplification are shown as solid lines (UltraI N=4, UltraII N=4, PCR-free N=3), other non-GAMCC field samples which have undergone sWGA are shown as a dotted and dashed line (N=4) and non-amplified control DNA extracted from culture of the *Pf* clone 3D7 are shown as a dashed line (UltraII N=1, PCR-free N=1). All samples were down sampled to a genome-wide average of 39 $\times$ coverage (6 million reads) to allow for direct comparison. In the second plot from the top AT nucleotide percentage for the genomic region is shown (calculated in 100bp windows). In the third plot from the top regional genes are annotated, with the gene of interest (the name of which is written at the top of the plot) indicated in red. At the bottom, positions of primer binding sites for the 10A primer set (used in all sWGA reactions) and 6A primer set (used in some sWGA reactions sequenced by OGC using UltraII and PCR-free) are shown in blue. Genomic position on the indicated chromosome in the *Pf* genome is the x-axis for all four plots.

3.2.4 Optimisation of sWGA parameters

By improving sequencing efficiency, that is widening the breadth of coverage and enlarging the proportion of non-duplicate sequenced reads of parasite origin, it is possible to increase the throughput of the sequencing run and reduce experimental costs. Duplicate reads occur when the input sequencing library has low complexity, likely due to an enrichment bias during the PCR steps of library preparation, and therefore indicate a “waste” of sequencing during the run (171). To optimise sWGA with these factors in mind, I tested seven sets of parameter conditions for sWGA, including varying input DNA mass, testing two different primer sets and trialing two extension times (as displayed in Table 3.1 (pg.99)). All seven sWGA conditions were tested on the same 8 samples to attempt to control for inter-sample variation. To compare the conditions, I assessed three quality control (QC) metrics; a measurement of the breadth of coverage (the ‘breadth of coverage metric’ (BoCM)), levels of human DNA contamination and the proportion of duplicate reads.

The BoCM indicates the proportion of core *Pf* genome bases with a coverage value within 3 standard deviations of the mean genome wide coverage (as calculated in alignments subsampled to 3 million core primary alignments to make them directly comparable, where the standard deviation is computed by assuming coverage follows a Poisson distribution). Samples with a greater proportion of the genome with coverage within this range have less variable coverage and therefore a more even spread of coverage across the genome. Samples with a high BoCM are therefore more likely to produce non-missing well-informed variant calls across the core genome.

As above, human contamination was assessed by measuring the proportion of reads within the sequence data for that sample which aligned to the Hg38 human reference. Levels of duplication are calculated as the proportion of reads which were marked as duplicate by the software Picard (further details of these methods can be found in Materials and Methods section 2.5.3 (pg.49)).

Changing sWGA conditions was found to make little difference to these metrics, indicating that improvements to the sWGA assay in the context of these samples cannot be easily made (Fig. 3.5 on the following page). The 8 samples had

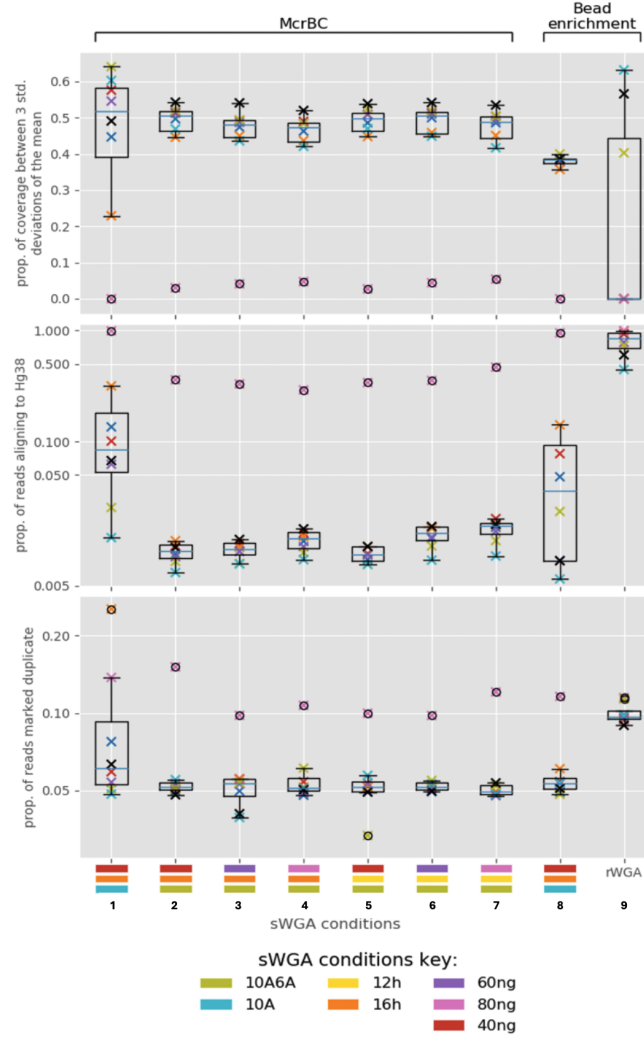


Figure 3.5: Comparison of sWGA parameters on GAMCC samples. Boxplots indicating the quality of sequencing data following nine different malaria parasite DNA amplification protocols. The metrics shown are the BoCM (which measures evenness of coverage compared to a theoretically uniform distribution) (top), proportion of aligned reads within the sequencing data which aligned to the human genome (middle) and proportion of aligned reads which have been marked as duplicate (bottom). Values range from 0-1; higher BoCM, lower human contamination and lower levels of duplicate reads are considered beneficial. These metrics are shown for 9 sets of conditions (labelled 1-9); sequencing data was generated following either selective whole genome amplification (sWGA) using AT enriched primers (Cond1-Cond8) or following random whole genome amplification (rWGA) using a set of random primers (Cond9). Details of the tested conditions are indicated by coloured bars, the key for which is beneath the plot (these are also detailed in Table 3.1 on the next page). Human DNA was degraded within samples using the restriction endonuclease McrBC (Cond1-Cond8) or using a method of bead-enrichment specific to di-methylated DNA (Cond8 and Cond9). Eight identical samples underwent each condition as indicated by the colour of the 'x' symbol denoting the quality control metric for that sample.

Condition	Human DNA depletion	Input (ng)	Primer sets	Extension time (hours)	rWGA or sWGA?
Cond1	McrBC	40	10A	16	sWGA
Cond2	McrBC	40	10A, 6A	12	sWGA
Cond3	McrBC	40	10A, 6A	16	sWGA
Cond4	McrBC	60	10A, 6A	16	sWGA
Cond5	McrBC	80	10A, 6A	16	sWGA
Cond6	McrBC	60	10A, 6A	12	sWGA
Cond7	McrBC	80	10A, 6A	12	sWGA
Cond8	Bead enrichment	NA	10A, 6A	16	sWGA
Cond9	Bead enrichment	NA	random	4	rWGA

Table 3.1: **sWGA parameters tested on the GAMCC set.**

consistent measures of breadth of coverage, human contaminations and duplicate read level across the conditions, apart from for Cond1 (Fig. 3.5 on the previous page), the single condition testing the ‘10A’ primer set in isolation, in which certain samples showed unusually low or unusually high BoCM compared to Cond2-Cond7. Cond1, although producing a high breadth of coverage on average (0.442 ± 0.21 BoCM), also showed much higher levels of human contamination (with an average of $21.4 \pm 30.7\%$ of reads aligning to Hg38 compared to $5.36 \pm 11.6\%$ for Cond2) and duplicate reads (with an average of $9.30 \pm 6.66\%$ of reads marked as duplicate in contrast to $6.38 \pm 3.31\%$ for Cond2). The results here suggested that the addition of another primer set known as ‘6A’ to produce ‘10A6A’ (Cond2–Cond7) generated higher quality data overall for the samples in the GAMCC set than 10A alone. Cond2 produced the second highest breadth of coverage with a 0.442 ± 0.16 BoCM (Fig. 3.5 on the preceding page) and therefore I opted to use this set of conditions to sequence the entire set.

In order to further validate my approach, as part of this pilot I tested an alternative method of human DNA depletion using beads which selectively bind to CpG-methylated DNA (Fig. 3.5 on the previous page, Cond8 and Cond9). I found this method followed by sWGA or rWGA produced poorer quality data than comparable samples following McrBC digestion and sWGA. On average $15.82 \pm 0.3\%$ and $80.17 \pm 18.7\%$ of reads aligned to the Hg38 reference genome respectively (compared to, for example, $5.35 \pm 11.6\%$ for Cond2 (Fig. 3.5 on the preceding page)).

3.2.5 Completing *Pf* WGS for the full GAMCC set

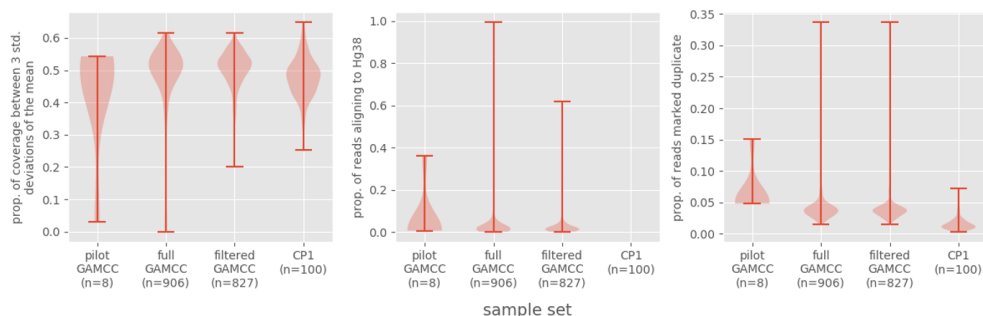


Figure 3.6: **Comparison of sequence data quality across the pilot sequencing run, the full GAMCC set, the filtered GAMCC analysis set, and 100 alignments randomly selected from the MalariaGEN CP1 dataset.** Violin plots showing BoCM (left), proportion of aligned reads within the sequencing data which aligned to the human genome (Hg38) (middle) and proportion of aligned reads which have been marked as duplicate (right).

918 samples were processed using sWGA under Cond2 (Table. 3.1 on the previous page & Table 3.2 on the following page). Of these, only 903 produced measurable DNA mass post sWGA and were therefore submitted for WGS. There was on average greater quality of sequencing for the entire set than the pilot; with 0.472 ± 0.12 BoCM across the 903 samples compared to 0.442 ± 0.16 BoCM in the pilot test (Fig. 3.6). There was however a significant drop out rate. For example, 7.31% of samples had a high proportion of human contamination ($> 20\%$ of reads mapping to Hg38) or poor breadth of coverage (with a BoCM < 0.2) or large proportion of PCR duplicate reads ($> 10\%$). This could be due to the build-up of contaminants or the occurrence of DNA degradation over the considerable time scale of 30 years since the GAMCC samples were collected.

To produce an analysis set, samples were filtered based on two metrics; (i) the GATK CallableLoci metric (the proportion of genomic bases in the core regions of the *Pf* genome with at least $5\times$ coverage and where at least 90% of covering reads have a mapping quality score ≥ 10), and (ii) genotype missingness (for further details see Materials and Methods section 2.6.3.1 (pg.54)). 831 samples for which $>50\%$ of bases were recognised as callable and $< 10\%$ of genotypes were missing were retained.

The coverage distribution for the entire set was comparable to that of the pilot measured genome wide (Fig. 3.6 on the previous page) and at relevant candidate loci. For example, the gene ACS8 (*Pfsa1* locus) had an average of $34.3 \pm 16.5\times$ fold coverage across its sequence, PF3D7_0220300 (*Pfsa2* locus) had $34.0 \pm 22.0\times$, PF3D7_1127000 (*Pfsa3* locus) had $32.2 \pm 17.7\times$ (130) and serine/threonine protein kinase (FIKK)4.2 (*Pfsa4* locus) had $27.0 \pm 15.5\times$ average fold coverage. Coverage across LSA-1 and LSA-3 however did not improve in the fully sequenced set with $4.32 \pm 5.15\times$ and $4.19 \pm 2.78\times$ average fold coverage respectively.

	Number in metadata	Number in original study	Number located in collection	Number with parasite WGS	Number in analysis GAMCC set	Number with HLA typing	Number with HLA-B53 call	Number with HbS re-typing
Severe malaria cases	620	619	571	558	502	89	582	487
Mild malaria cases	364	354	347	345	329	91	361	319
Mildly ill, non-malaria infected controls	547	510	-	-	-	33	520	2
Severely ill, non-malaria infected controls	347	332	-	-	-	37	339	1
Total malaria cases	984	972	918	903	831	180	943	806

Table 3.2: **Summary of sample numbers in the GAMCC set.** Malaria cases were the focus of this study and so counts of non-missing non-malaria cases were not made.

3.2.6 Comparison of the GAMCC to a public dataset

To ensure the quality of the GAMCC dataset was equivalent to other publicly available datasets, I compared quality control metrics to a random selection of 100 samples from the MalariaGEN CP1 dataset which were also prepared using the sWGA method (130). I found that both the filtered and unfiltered GAMCC sets had very similar breadth of coverage to CP1 (0.472 ± 0.12 BoCM in the unfiltered GAMCC compared to 0.479 ± 0.064 BoCM in the subset of CP1) (Fig. 3.6 on the preceding page). I was unable to compare levels of human contamination as genetic content of human origin was not included in the CP1 release (130). Duplicate reads did however appear to make up a lower proportion of reads within CP1 samples (only $1.6 \pm 1.0\%$ of reads were marked duplicate compared to $4.2 \pm 2.6\%$ in the GAMCC).

Following filtering, breadth of coverage in GAMCC surpassed that of CP1 (0.504 ± 0.058 BoCM) however duplicate read proportion remained higher in the filtered analysis GAMCC set ($3.8 \pm 1.7\%$). CP1 is made up of only severe cases

which may naturally have higher parasitaemia and therefore higher initial *Pf* DNA content within DNA extractions. This could improve the complexity of libraries following sWGA compared to the GAMCC which contains a mixture of severe and mild cases. On the other hand, the connection between parasitaemia and severity is not clear cut (172), and it is perhaps more likely that methodological differences between research centres underlies this disparity.

3.2.7 Performing a surrogate “sex check” using contaminant human reads

To ensure there was minimal sample mix-up over the longstanding storage of the GAMCC samples, I performed a *de facto* sex check on the data. Although referred to in the previous sections of this chapter as “contamination”, the human reads that were sequenced alongside those of *Pf* origin during sWGA offered useful insights into the individual sample. To attempt to infer the biological sex of the sample, I plotted proportion of reads mapping to the X chromosome (out of all human reads) against proportion mapping to the Y chromosome and found samples to visually cluster into two groups (Fig. 3.7 on the next page). These groups appeared to correlate to the reported sex of the sample (as indicated by marker colour in Fig. 3.7). Using a density-based clustering method, 337 samples were clustered into the predominantly female ‘cluster 1’ (represented by triangles symbols in Fig. 3.7) and 375 samples into the predominantly male ‘cluster 2’ (represented as ‘x’ symbols (Fig. 3.7)) (102 samples were not clustered, indicated by dots (Fig. 3.7)). Six samples reported as female were placed into cluster 2 (1.75% of all clustered females in the set) and 8 males into cluster 1 (2.08% of all clustered males). However, some of these samples fell between the two clusters such that only 4 (1.16%) and 6 (1.56%) respectively visually appear to have the sex which they were not assigned in the metadata (Fig. 3.7 on the following page).

The discrepancy between apparent biological and recorded sex could be due to, for example, erroneous record making at the time of their collection, which is unlikely to affect analysis if all other records were taken correctly, or subsequent sample mix up or pipetting mistakes during the present study. Complete sample mix ups can be investigated due to thorough plate records, and having investigated

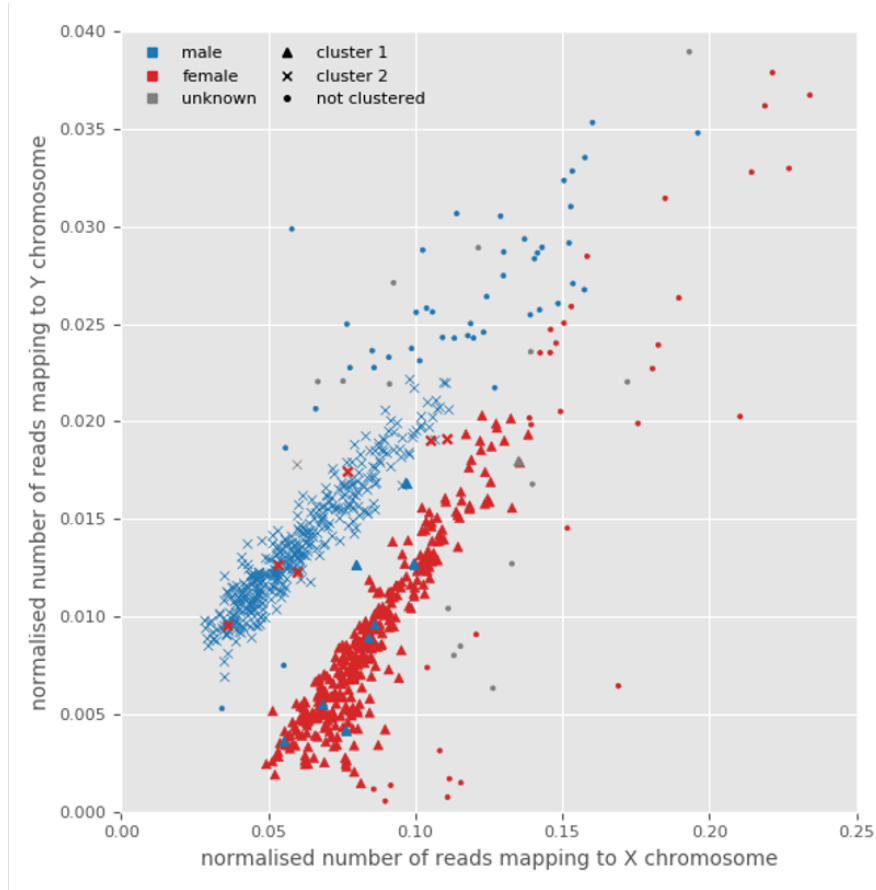


Figure 3.7: A “sex check” of the GAMCC *Pf* WGS data. The proportion of all human reads in the alignment of sequences generated from malaria parasite DNA amplifications which align to the human X chromosome against the proportion of all human reads aligning to the human Y chromosome. All generated reads were aligned to a concatenated Pf3D7 and human Hg38 reference genome. All alignments in the filtered dataset are shown ($N = 829$). Points are coloured red to indicate that the sample was recorded as female, blue to indicate male and grey if that data was missing. These points were standardised and clustered into two groups using DBSCAN indicated as triangles or ‘x’. Noise samples are shown as points.

these samples I believe it is unlikely that this occurred. Discriminating between pipetting and record errors is challenging, but because there is a large concordance between reported and inferred sex this gives confidence in the broad accuracy of the metadata we have for this set.

On average only $22,700 \pm 476$ reads aligned to the Hg38 genome in the GAMCC set, resulting in an average fold coverage of less than $0.01\times$ across the human genome. Consequently, the scope for deriving further insights from generated human data was severely restricted.

3.2.8 Variant Calling in a large, combined *Pf* WGS set

To streamline downstream analyses, I generated a combined variant call set for GAMCC and CP1 (from samples indicated as the ‘analysis set’ in the original CP1 study in both Gambia and Kenya (130)). CP1 sequence data was downloaded and processed in the same manner as GAMCC and on the resultant alignments I performed variant calling for the combined set of 5074 samples (GAMCC N=903, CP1 N=4171), calling a total of 3,097,708 single nucleotide polymorphisms (SNPs) and 1,243,684 short insertion or deletions (INDELs) across the 14 main chromosomes in the malaria parasite genome. Further details of this process can be found in Materials and Methods section 2.5.2 (pg.48) and 2.6.1 (pg.52).

3.2.9 Targeted human typing in the GAMCC set

3.2.9.1 HLA typing

Certain HLA information was already available for these samples. Shortly following their collection (1988-90), an assay to type the presence of the HLA-B*53 allele was performed (52). Recently, 70 severe and 70 mild cases (as well as 70 non-malaria infected controls, totalling 230 samples) from the GAMCC set also underwent array-typing using the Affymetrix array and low-coverage Illumina WGS (for more detail of this experiment see Materials and Methods section 2.4.2.1 (pg.44)). High-resolution HLA-typing was also performed on the same 230 samples by the company HistoGenetics (using a combination of short and long read sequencing).

In addition to this, I commissioned HLA sequencing for a further 40 samples generating a total of 180 high accuracy HLA calls for malaria cases in the GAMCC with sequenced *Pf* genomes (21.7% of the analysis set) (Table 3.2 (pg.101)). These additional 40 were selected so that half had been previously called as carrying the class I HLA-B*53 allele and half had not, with a view to improving power to estimate the accuracy of the HLA-B*53 typing assay developed by Hill et al. (52) (as described later in Chapter 5 (pg.148)). Three novel HLA allele subtypes were discovered as part of this endeavour at the class I HLA-A and HLA-C loci and at the class II DQA1 locus.

3.2.9.2 Haemoglobin S/C typing

Two beta-globin hematopathologies known to confer resistance to *Pf* infection (6) were assessed within the GAMCC. These included the more frequent HbS that occurs within the 6th codon of the *beta-globin* (*HBB*) gene, and the less frequent HbC variant that occurs in the preceding base position in the same codon (17, 18). HbS and HbC-typing calls existed in the historical metadata for the GAMCC sample set (referred to as the AH calls). However, it was noted that there were inconsistencies between sample records from different sources. Therefore I performed initial validity HbS/C-typing on a subset of samples recorded as carrying non-reference alleles (that is carrying one or two copies of an HbS or HbC allele) using a Sanger-sequencing based method. This showed a significant degree of non-concordance; of 11 samples which had been called as HbS-carriers (HbAS), 4 were in fact homozygous reference samples (HbAA) and 1 was a homozygous HbS individual (HbSS). In addition, a sample called as HbSS was in fact HbAA upon re-sequencing (Table 3.3 on the following page).

Because HbS is very rare within the population, false positive calls would significantly compromise downstream analysis, particularly the genome wide association scan between parasite variation and HbS-type (as conducted in Chapter 5 (pg.148)). I therefore decided to obtain renewed typing information for the entirety of the GAMCC set.

The affymetrix array used to perform whole-genome genotyping on a subset of samples in the GAMCC (described above) directly typed the presence of HbS polymorphism. HbS and HbC genotypes have also been imputed for these samples

		Sanger sequencing call			
		AA	AS	SS	AC
AH calls	AS	4	6	1	0
	SS	1	0	0	0
	AC	0	0	0	5

Table 3.3: **Comparison of original and retyped *HBB* genotypes in non-reference samples in the GAMCC cohort.** The table shows samples which were called as carrying a non-reference alleles at the HbS and HbC polymorphisms in the original metadata (AH calls) and the genotype they were found to upon re-sequencing these samples using Sanger sequencing. This indicated significant inconsistencies.

using lcWGS data, giving genotypes for 133 samples in the set. For the remaining samples in the set I performed HbS/C typing on the remaining samples using the same Sanger-sequencing based method as above. The generated calls, and method used to make the call, are shown in Table 3.4 on the following page.

The *HBB* gene is challenging to amplify and sequence and the approach adopted here has shown high dropout rates in other studies (e.g. (131)). Consequently, despite submitting all initial failures for re-sequencing, 2.6% of severe cases and 2.4% of mild cases in the GAMCC set produced no readable data. The age and state of the GAMCC samples may have contributed to the difficulty to produce clear sequence data at the *HBB* locus. Calls were made based on careful manual inspection, therefore an alternative approach, such as WGS, could be adopted to secure the further reliability of this aspect of the GAMCC resource (a full discussion of this is made in Materials and Methods section 2.4.2.2 (pg.44)).

	PCR	PCR,array	PCR,array,lcWGS	PCR,lcWGS	array	array,lcWGS	lcWGS	not assessed
HbS calls								
T/T	672	1	13	3	12	92	7	0
T/A	8	1	1	1	0	1	1	0
A/A	0	0	1	0	0	0	0	0
Missing	0	0	0	0	0	0	0	89
HbC calls								
C/C	676	0	0	18	0	0	102	0
C/T	7	0	0	0	0	0	0	0
T/T	0	0	0	0	0	0	0	0
Missing	0	0	0	0	0	0	0	100

Table 3.4: **Results of *HBB* retyping in the GAMCC cohort.** results are stratified by genotyping method: PCR followed by Sanger sequencing (PCR), using the Affymetrix array, or following low coverage (lc)WGS followed by imputation. Results at rs334 (HbS, chr11:5,227,002 T>A) and rs33930165 (HbC, chr11:5,227,003 C>T) are shown.

3.3 Discussion

Generating whole genome *Pf* data for the GAMCC set posed a substantial challenge due to the relatively low parasite DNA content within the naturally infected samples and the low complexity high AT content of the malaria parasite genome. To overcome the former, I employed a widely used approach utilising AT rich primer sequences to perform sWGA of the parasite genome (Figs. 3.1 (pg.90) & 3.2 (pg.92)). I observed that incubating samples with the methylated DNA cleaving restriction enzyme McrBC prior to sWGA aided this process by further reducing levels of human DNA present in final sequenced libraries (Fig. 3.1), mirroring the findings of previous optimisation studies (98). To counter the latter, working with the Oxford Genomics Centre (which has now been replaced by the Technology Platforms Group), three library preparation methods were trialled to assess which mitigated GC-biases introduced during the PCR-steps of Illumina sequencing to the greatest degree (Fig. 3.1). A major downside of the sWGA method is the uneven coverage in Illumina data that it produces. Over two pilot rounds of sequencing, breadth of coverage was improved for GAMCC samples prepared using sWGA to be on a par with the publicly available high-quality data set MalariaGEN CP1 using the UltraII and PCR-free library preparation methods (Fig. 3.2).

Although we achieved major improvements, certain regions of the *Pf* genome

appeared unrecoverable by sWGA. For example, both the LSA-1 and LSA-3 loci were poorly covered in GAMCC samples using any library preparation method following sWGA, but not those sequenced from unamplified DNA, indicating that this might be a perplexing artefact of amplification (Fig. 3.4 (pg.96)). Visually, primer binding sites appeared to be sparse within these regions, however primer binding sites have not been found to correlate with coverage following sWGA (99). Targeted long-read amplicon based sequencing, methods for which are currently being developed (104), may be required to access these and other potentially complex regions of the *Pf* genome. In contrast the *Pfsa* regions, which will be investigated later in this thesis, were well covered in data produced by this method (Fig. 3.3 (pg.95)).

Despite our efforts, significant improvements in breadth of coverage through adjustment to sWGA conditions were not achieved in this study, however the addition of the 6A primer set appeared to marginally improve quality in terms of the breadth of coverage and the recovery of specifically *Pf* DNA within libraries, reproducing the findings of others on DBS samples (98) (Fig. 3.5 (pg.98)). Further optimisation in this regard, if possible, may necessitate substantial adjustments to the current methodology and therefore could require significant resources to develop.

Following the determination of the most optimal sWGA parameter set within the constraints of the current workflow (Cond2, Fig. 3.5), data for the entire GAMCC set was generated. Overall, from the set of 972 samples, 918 of which were physically located and 831 passed quality control filtering (Table 3.2 (pg.101)). A 90% retention of available samples was therefore achieved indicating that historical sample sets can be exploited with minimal loss. Key to this is the faithful retention of metadata as time progresses, which appears to have been broadly achieved here due to the concordance of the recorded sample biological sex to the levels of sex chromosome aligning reads in sequenced libraries (Fig. 3.7 (pg.103)).

The quality of the GAMCC set compared to another publicly available dataset of similarly prepared samples showed they were closely comparable in terms of quality (Fig. 3.6 (pg.100)), affirming the calibre of the GAMCC derived data. The quality and accuracy of this *Pf* WGS dataset should therefore facilitate well-evidenced inferences, building on our understanding of *Pf* populations in The Gam-

bia. The vintage of this sample set makes it particularly impactful because it affords an opportunity to investigate a currently unassessed time-point. To my knowledge, no public Gambian *Pf* WGS data includes samples from 1988-90, and in fact the vast majority ($> 98\%$) of Gambian samples in the recently released MalariaGEN Pf7 resource (37) were collected beyond the year 2000.

However, a significant limitation of the GAMCC resource is the absence of human WGS, which would allow a complete genome-to-genome investigation into host-parasite interactions. This was not attempted due to the prohibitive costs of sequencing the relatively much larger human genome. Instead, targeted sequencing of specific loci, including the *HBB* and HLA loci, was successfully conducted. The HbS/C variants at the human *HBB* locus were typed using a Sanger-sequencing based method and high resolution class I and II HLA calls (6-digit resolution) were generated for a subset of 180 samples by the external company Histogenetics (Table 3.2).

Within the Centre for Human Genetics (University of Oxford, UK), efforts have been made to generate genome-wide data for human samples in the GAMCC using an ultra-low coverage sequencing approach. This method presents challenges; for example, because the generally high levels of genetic diversity and lower levels of linkage disequilibrium in African populations, imputation panels which this method relies upon may perform poorly on underrepresented populations such as those from Gambia (173). These challenges are actively being addressed, and with further optimization, the envisioned expansion into a complete host and parasite whole genome resource may eventually be realised. Irrespective of the inclusion of human WGS, the targeted sequencing available in the GAMCC promises to yield valuable insights into genetic interactions between hosts and parasites, thus serving as a foundational pilot study for these types of analyses.

Chapter 4

Does parasite genetic variation influence the severity of malaria infection?

4.1 Introduction

While malaria remains a persistent global health issue, it is important we build our understanding of factors which contribute to higher mortality and morbidity of infection. Acute episodes of malaria, broadly termed “severe”, make up a small fraction ($\sim 1\%$) of malaria cases (55) but constitute the most clinically significant group. Severe malaria remains the leading cause of preventable childhood deaths in endemic regions (80). The human genetic correlates of protection against severe malaria have been studied in depth, but only a handful of these have been found to robustly replicate across multiple populations (6; 174; 54). Furthermore, the key susceptibility loci which have been identified using genome wide association studies (GWAS) account for only a small fraction of the variation in observed disease severity (6; 55).

It is therefore recognised that parasite factors might contribute to disease severity (55; 130). This has been observed for other infectious diseases for which GWAS have provided important insights. For example, GWAS performed using bacterial genomes established that Pyomyositis, a severe *Staphylococcus aureus* infection of skeletal muscle, strongly depended on the genetic ability of bacteria to produce a single toxin called Panton-Valentine leucocidin (PVL), offering valuable guidance

for potential vaccine design (175). To our knowledge no similar efforts to assess for the correlates of severity throughout the malaria parasite genome have been made. This is partly due to significant challenges imposed by such an analysis. Severe malaria infections present clinically in diverse ways since malaria can affect multiple body systems (80). This makes it difficult to define clinical outcomes for analyses that can in turn make GWAS results difficult to interpret.

Therefore, an analysis of this type requires a large set of samples which have been collected following consistent and precise definitions of severity phenotypes, ideally over a short time frame within a narrow geographic region to mitigate the risk of confounding by temporal and geographical population structure. These results could then be meta-analysed with additional populations to identify both population-specific and fixed effects. Although the Genetic Analysis of Malaria Cases and Controls (GAMCC) dataset, described in Chapter 3 (pg.87), involves a relatively small sample size, the samples were collected based on defined criteria of what constituted a severe or mild case. As a result, this set provides a unique opportunity to conduct a pilot investigation into the genome-wide parasite genetic correlates of malaria infection severity.

Levels of malaria disease severity exist in a continuum ranging from asymptomatic infections to fatal aetiologies. Parasite genetic severity effects might enact their effect at any point in this scale. Expanding on the example above, within Pyomyositis infections the presence of the PVL encoding gene differentiated severe disease from asymptomatic carriage (175). In malaria infections however parasite genetic effects might plausibly increase the likelihood of certain symptoms such as cerebral malaria (CM) or severe malarial anaemia (SMA) within severe infections, or may be more generalised, acting to increase severity at any level. Identifying effects and understanding their impact by leveraging more precise sub-type phenotypic information might help uncover the biological mechanisms which underlie severity and eventually lead to the design of more effective treatment approaches.

In this chapter I therefore set out to explore the genetic correlates of malaria severity using *Plasmodium falciparum* (*Pf*) whole genome sequencing (WGS) data in the GAMCC cohort using a GWAS approach and attempt to replicate this analysis using public *Pf* WGS resources released by the Malaria Genomic Epidemiology Network (MalariaGEN). I also aim to identify candidate loci within the *Pf*

genome that correlate with the severe malaria subtypes CM and SMA using the MalariaGEN Consortial Project 1 (CP1) dataset.

4.2 Results

4.2.1 Does parasite genetic diversity influence risk of severe malaria disease?

4.2.1.1 Assessing parasite genetic effects in the GAMCC dataset

The GAMCC analysis set, collected from Gambian infants (1988-90), contains substantial numbers of both severe (N=502) and mild (N=329) cases which were classified at the time of collection. Severe infections were defined as those involving CM, SMA, or hypoglycaemia, while mild infections were characterised by moderate parasitaemia (*Pf* parasitised blood cells) exceeding a specific threshold of 2,500 infected cells/ μ L. These cases displayed clinical illness typical of the setting but without the critical symptoms observed in severe cases (for further details of phenotyping, see Materials and Methods section 2.7.1.1 (pg.58)). By comparing the newly sequenced parasite genomes between infections from each of these categories using GWAS, it may be possible to identify parasite loci that predispose infections to increased severity.

First, to aid in the interpretation of the results of GWAS using this data, and to help determine an appropriate P-value threshold to dictate whether an association might be of interest, I conducted a power analysis. Specifically, I simulated sets of genotypes under the assumption of a true association, using the sample sizes available in the GAMCC. The architecture of parasite effects on severity is not known, so these simulations included two effect sizes (odds ratio (OR) = 2 and OR = 4) and varying baseline population allele frequencies (5%, 10%, 25%, 50%) to test plausible scenarios. By performing association tests using these simulations, I computed the statistical power (the probability that a P-value would fall below a given threshold when an association truly exists). By using Bayes Theorem, under a relatively modest prior assumption that a single 1Kb region of the *Pf* genome was associated with severe disease, I then used the calculated power to estimate the true discovery rate (TDR) (the probability that an observed association is true

when the P-value is below the chosen threshold). Further details of this statistical process can be found in Materials and Methods section 2.7.3 (pg.61).

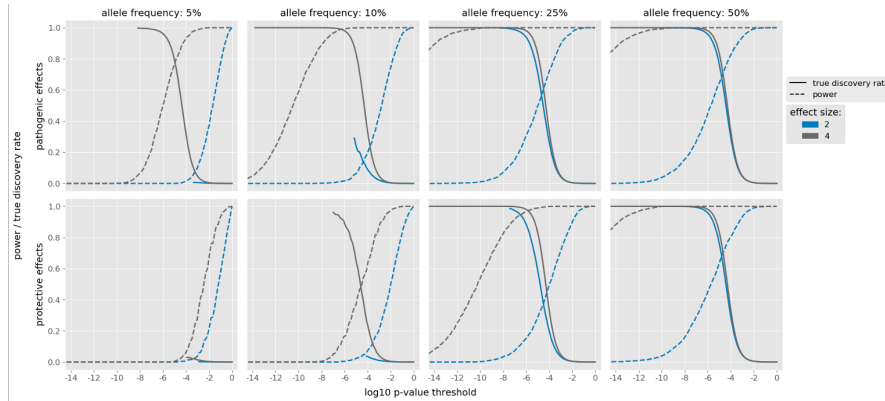


Figure 4.1: Power and true discovery rate analysis for an analysis of *Pf* parasite genetic effects on severity using the GAMCC dataset. Model computed power (dashed lines) and true discovery rates (solid lines) for different P-value thresholds ($\log_{10}$) on the x axis. These values were calculated based on 2000 simulations using a prior assumption that 1 in the approximately 20,000 regions in the *Pf* genome is associated with severe malaria. Five baseline allele frequencies (that is, the frequency in mild malaria cases) were assessed: 5%, 10%, 25% and 50% (from left to right). Two effect size magnitudes were used. Pathogenic effects (top panels) indicate the variant increases the chance of severe malaria (i.e. OR = 2 (blue) and OR = 4 (grey)). Protective variants indicate the variant reduces the risk of severe malaria (i.e. OR = $\frac{1}{2}$ (blue) and OR = $\frac{1}{4}$ (grey)).

This found that a GWAS testing for severity using this data would be powered to find large severity increasing effects at low to moderate frequencies, (>5-10% allele frequency) but would be less well-powered to resolve severity reducing effects at these frequencies (Fig. 4.1). This imbalance in power between effect type is due to the uneven sample sizes between phenotype categories (with $N_{\text{mild}}=329$ and $N_{\text{severe}}=502$). At large allele frequencies however (25-50%) we would be well powered to infer both effect types. To produce a TDR of 80% across most allele frequencies for strong severity increasing effects (OR=4) a P-value threshold at roughly the Bonferroni correction threshold ($1/\text{total number of variants} = 2.14 \times 10^{-5}$ in the present study) appeared a suitable threshold with which to interpret this analysis (Fig. 4.1).

Next, I performed GWAS using a variant call set generated from the GAMCC samples. This dataset was jointly called with samples from the MalariaGEN CP1 resource to enable multiple comparisons between the GAMCC and CP1 samples

throughout this thesis (referred to as the GAMCC/CP1 call set) (for further detail, see section 2.6.1 (pg.52)). Association tests were performed across a set of high confidence biallelic SNP variants (Variant Quality Score log odds (VQSLOD) > 2 , sample missingness $< 10\%$, allele count (AC) > 5) (N=46,757) filtered from the analysable regions of the “core” *Pf* genome. Descriptions of sample and variant filtering, and details of the GWAS method employed, are described in Materials and Methods section 2.6 (pg.52) and 2.7.2 (pg.59).

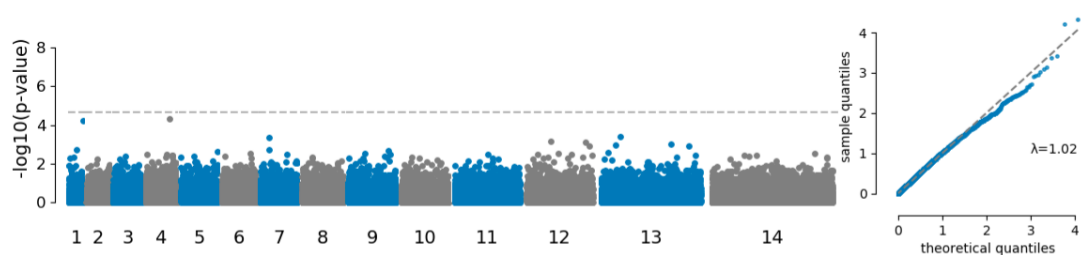


Figure 4.2: Strength of evidence for the association of variants across the *Pf* genome with disease severity in the GAMCC dataset. Strength of evidence ($-\log_{10}$ P-values) for the association of variants (y-axis) across the 14 chromosomes in the *Pf* genome (x-axis) with disease severity (left) and quantile-quantile plot displaying observed (y-axis) versus expected (x-axis) $-\log_{10}$ P-values from the GWAS. Genomic inflation factor lambda is also indicated.

The results of this GWAS analysis are shown in Figure 4.2. No variants reached the Bonferroni level of statistical significance indicated as potentially compelling by the presented power analysis (Fig. 4.1 on the previous page). However, two variants reached levels of near significance. The first was at position 557,240 on chromosome 1 (Pf3D7_01.v3 in the Pf3D7 reference assembly). This variant showed evidence of a protective effect (OR=0.355 (0.214-0.589), allele frequency (AF)=9%, P-value (P)= 6.09×10^{-5}). Based on our model of TDR, I estimated that there would be modest evidence of this variant being truly associated with severity at this P-value threshold (under these parameters power=75.6% and TDR=67.9%), though this relies on approximations of prior quantities that are not known in practice.

The second variant fell on chromosome 4 (Pf3D7_04.v3) at position 898,419 (P = 4.82×10^{-5}). This allele was rare within the GAMCC set with an allele frequency of only 5%. At allele frequencies $< 10\%$, I estimated that the power of this analysis to detect true effects would be significantly compromised (Fig. 4.1

on the preceding page), and therefore it was less clear how much meaning can be derived from this observation from the GAMCC dataset alone.

4.2.1.2 Replication analysis using the MalariaGEN Pf7 and CP1 resources

Since no associations of clear statistical significance were observed, I next attempted to independently replicate this analysis. There are very few other parasite genomic datasets available that categorise infections as severe or mild. However, two public resources of parasite WGS data curated by MalariaGEN, specifically the CP1 dataset of severe cases from Gambia and Kenya (first described in (130)), and the Pf7 resource of community infections (37), offered a potential opportunity to undertake replication.

Creating a merged, stratified case-control dataset In order to compare variation from multiple resources, the Pf7 and GAMCC/CP1 call sets were intersected (the “intersected call set”) (see Materials and Methods section 2.6.2 (pg.53)). Using these datasets for GWAS comparison posed notable challenges. Foremost, the Pf7 resource was assembled from multiple sample collections with differing phenotyping conditions, including community and clinical cases, and does not reflect a specific severity phenotype. However, these samples likely do not include severe cases. A complete discussion of the sample phenotyping and a description of filtering are available in the Materials and Methods section 2.7.1.2 (pg.59) and 2.6.3.2 (pg.56). In summary, and for the purposes of this analysis, Pf7 samples were all considered as “mild” and CP1 as “severe” (Table 4.1 on the following page). In addition, since the Pf7 samples were collected at different times to the CP1 set, this means that the severe and mild cases were not only stratified by disease status but also by time.

Dataset	Country	Total samples (no filtering)	Considered as	Attributes											Total (filtered)
				CM	SMA	CM & SMA	Hyp	CM+Hyp	SMA+Hyp	CM, SMA & Hyp	Other severe pathologies	parasitaemia > 2500/ μ L	Unresolved	Unknown	
GAMCC	Gambia	905	Severe	205	109	37	24	108	8	10				1	502
			Mild									329			329
Pf7	Gambia	1247	Mild										818		818
	Kenya	726	Mild										602		602
CP1	Gambia	2235	Severe	673	384	117									2235
	Kenya	1936	Severe	786	247	218							685		1936

Table 4.1: **Sample sizes for analysis of malaria disease severity.** Abbreviations: CM = cerebral malaria, SMA = severe malarial anaemia, Hyp = hypoglycaemia. Unresolved refers to samples which are likely to be clinical but not severe malaria, although this cannot be verified due to the nature of the collection.

Therefore, there is the potential for confounding of GWAS comparisons using these data due to genetic drift or other modes of allele frequency shift within parasite populations between time points. *Pf* populations tend to undergo more pronounced genetic drift (176) and evolve in response to selection pressures, such as anti-malarial drug use, very rapidly (112; 177; 178; 179). This can be seen by dramatic changes in the frequency of chloroquine (CQ) resistance mutations in both the Gambian and Kenyan populations (as shown in Fig. 4.5 (pg.121)) (112). This may alter parasite genetics on a genome wide scale and make it difficult to distinguish between differences due to this shift (i.e. batch effects) and true genetic differences between phenotypic groups (in this case between the severe infections within the CP1 resource or the mild infections in Pf7). Therefore, I attempted to find a GWAS model which limited this confounding to the greatest degree.

Generating an association model for Pf7 and CP1 First, to assess for the magnitude of temporal and spatial structure within Kenya and Gambia I performed a principal component analysis (PCA) across samples in the CP1 and Pf7 sets individually for each population. There were several highly related samples within Gambia (706/3053 (23.1%)) and Kenya (202/2538 (8.0%)) which initially strongly defined the PCs (Supp. Fig. 4.14 (pg.141)). Following their removal, samples within each country clustered together with samples derived from the Pf7 and CP1 sets overlapping (Figs. 4.3a on the next page & 4.4a (pg.119)), which was promising because it suggested there was not major differences between the two data sets. However, within the single cluster, samples appeared to group by year and sample set (Pf7 or CP1).

Assessing the loadings of these PCs across the genome within Gambia showed a high level of PCA loading within regions of known anti-malarial drug resistance (DR) loci within the genes dihydrofolate reductase (DHFR) (on Pf3D7_04.v3), multidrug resistance 1 (MDR1) (on Pf3D7_05.v3), amino acid transporter 1 (AAT1) (Pf3D7_06.v3), a peak proximal to the chloroquine resistance transporter (CRT) gene (on Pf3D7_07.v3) and at hydroxymethyldihydropterin pyrophosphokinase (PPPK)-dihydropteroate synthetase (DHPS) (Pf3D7_08.v3). Interestingly, the *Plasmodium falciparum* sickle associated (*Pfsa*)3 locus (PF3D7_1127000 on Pf3D7_11.v3)

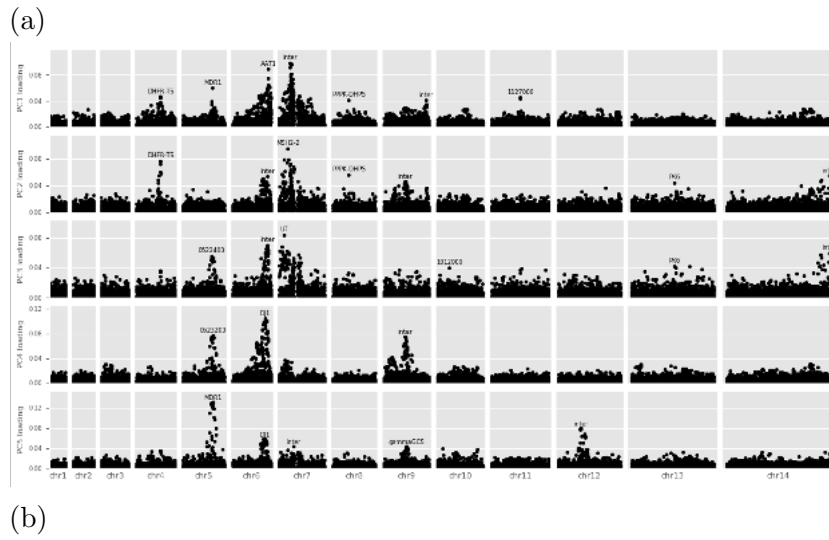
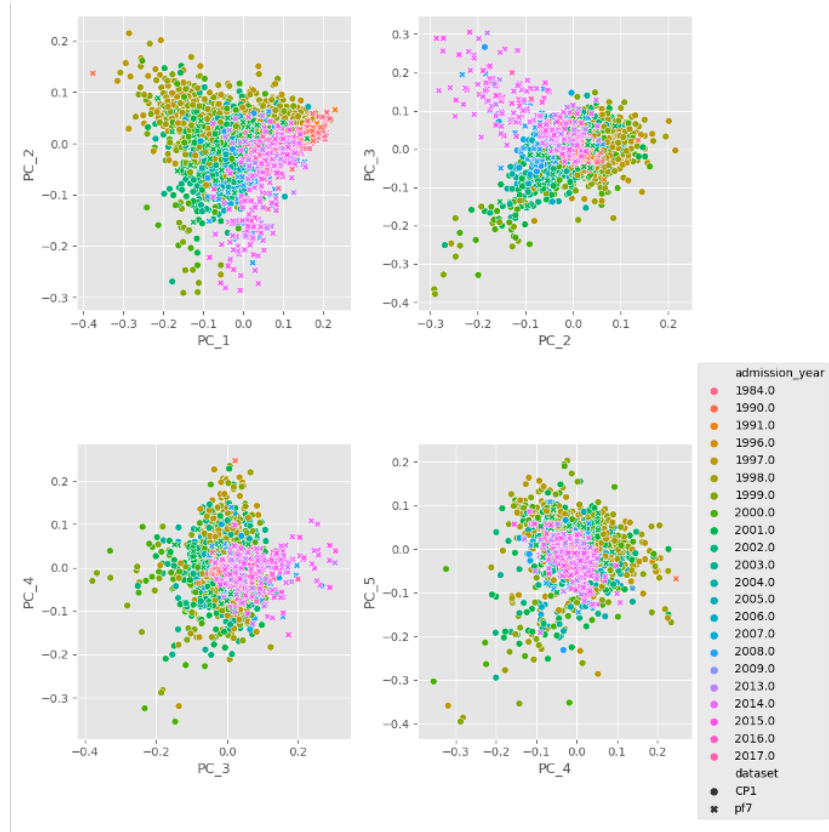
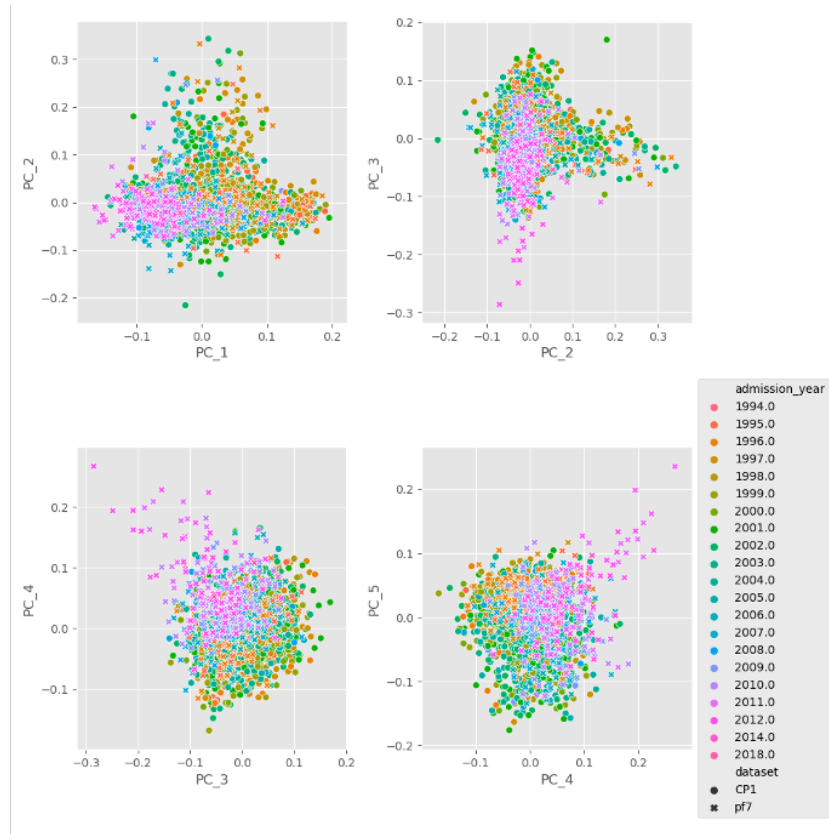
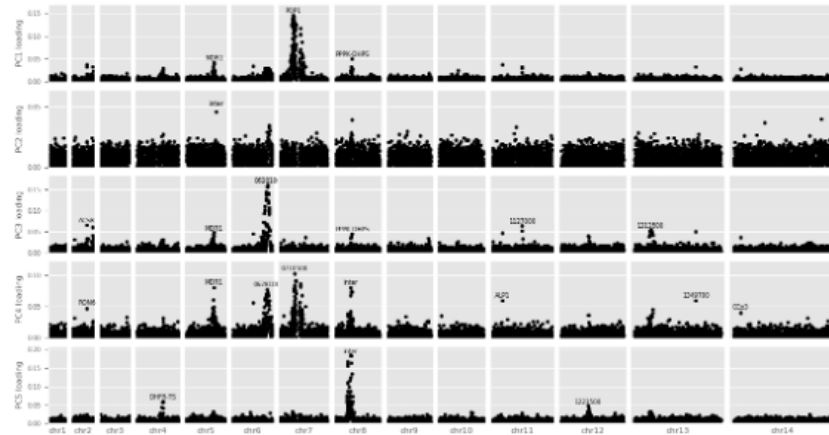


Figure 4.3: PCA in the Gambian population. A. PCA with highly related samples (kinship>0.5) removed. Markers are coloured by sample collection year. Marker shape indicates the source of the sample: the GAMCC set (circles), MalariaGEN CP1 set (plus symbols) and MalariaGEN Pf7 set (triangles). **B.** Loadings of each PC (indicated on the y-axis of each panel) across the 14 chromosomes in the malaria parasite genome.



(a)



(b)

Figure 4.4: PCA in the Kenyan population. **A.** PCA with highly related samples (kinship > 0.5) removed. Markers are coloured by sample collection year. Marker shape indicates the source of the sample: the GAMCC set (circles), MalariaGEN CP1 set (plus symbols) and MalariaGEN Pf7 set (triangles). **B.** Loadings of each PC (indicated on the y-axis of each panel) across the 14 chromosomes in the malaria parasite genome.

which has recently been shown to associate with the human resistance polymorphism within the HBB gene called HbS (130), and an intergenic region on parasite chromosome 9 (Pf3D7_09_v3) (most proximal to the gene gametocyte development protein 1), also strongly dictated these PCs in Gambian samples (Fig. 4.3a (pg.118)).

The influence of drug resistance on population structure was also clearly present within the Kenyan population. PC1 and PC3-5 were strongly driven by MDR1, CRT, AAT1, PPPK-DHPS and DHFR in various combinations (Fig. 4.4b on the preceding page). The gene at a further HbS associated locus *Pfsa1* (acyl-CoA synthetase (ACS)8, Pf3D7_02_v3) and *Pfsa3* also appeared to influence PCs 3 and 1 and 3 respectively. PC2 within Kenya appeared to show loadings distributed across the entire genome indicating it may have encapsulated genetic drift.

In order to assess the effects this might impart upon association results, I compared GWAS using different sets of covariates. To avoid potential confounding caused by genotype errors, I focussed on a set of high-confidence biallelic SNP sites (VQSLOD > 2, sample missingness < 10%, AC > 5) within the core regions of the genome. After filtering, 38,889 sites for the Kenyan population and 43,342 sites for the Gambian population were retained for use in these analyses (for details, see Materials and Methods section 2.6.3.2 (pg.56)).

GWAS comparing the mild Pf7 samples and severe CP1 samples for each population including no covariates produced strong signals of association (7.8% of tested variants with P-values < 0.05) in Gambia at the MDR1, AAT1 and CRT loci (Supp. Fig. 4.15a (pg.141)) and less extreme associations at the CRT and MDR1 loci in Kenya (Supp. Fig. 4.15b (pg.141)). A metric commonly used to assess genome-wide confounding in GWAS is the lambda inflation factor. Lambda values greater than 1 indicate an inflated GWAS model which likely occurs through the effects of systematic biasing (for a more detailed description Materials and Methods section 2.7.2 (pg.59)). As expected, P-values were more greatly inflated in the Gambian analysis than in Kenya ($\lambda_{Nocovariates}^{Gambia} = 3.108$ compared to $\lambda_{Nocovariates}^{Kenya} = 1.256$).

The discrepancy in levels of inflation between the two countries might be explained by the temporal distribution of the sets. The Gambian dataset covers a wider range of years (1984-2017 as opposed to 1994-2014 in Kenya) which fall in

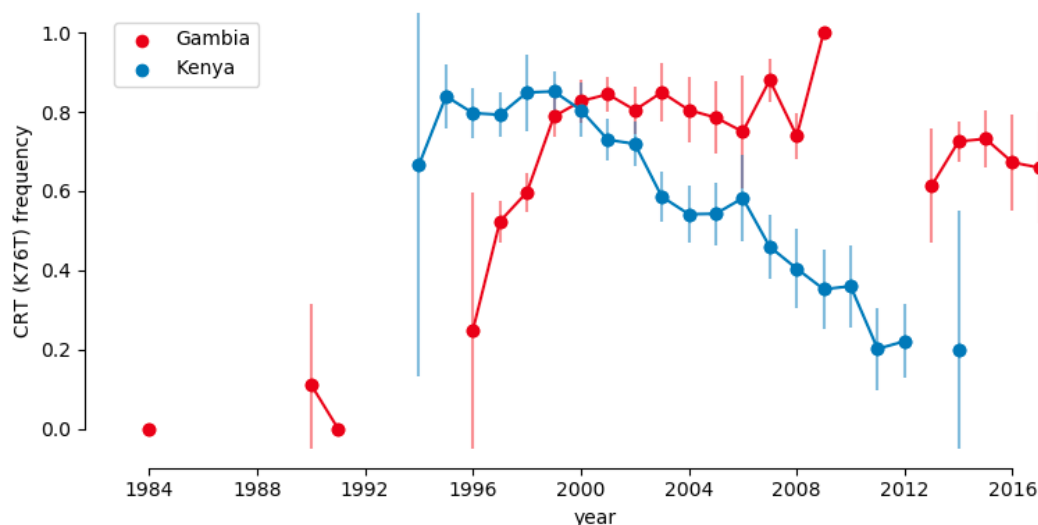


Figure 4.5: **Changes in allele frequency overtime at the chloroquine resistance locus CRT in Gambia and Kenya.** Changes in allele frequency at the chloroquine drug resistance locus (amino acid substitution K76T) in the CRT gene in samples collected from Gambia (red) and Kenya (blue). Samples come from different data resources: the GAMCC set, the MalariaGEN CP1 and MalariaGEN Pf7.

disconnected time points with minimal overlap between the CP1 and Pf7 resources (Supp. Fig. 4.16 (pg.142)). Another contributing factor might be the spread of the CQ resistance mutation within each country. Allele frequency of the CQ resistance mutation in the CRT gene was more stable across time in the Kenyan dataset, indicating the sampling initiated following the occurrence of the selective sweep of this allele (Fig. 4.5). In contrast, within Gambia the CRT resistance mutation underwent a complete selective sweep within the sampling period (changing in frequency from 0% in 1984 to 100% in 2009). This alteration in parasite genomes, particularly when considering the long-range linkage disequilibrium known to be associated with the selective sweeps of drug resistance (179) might have lead to the grossly inflated results within the Gambian data which may not be possible to control. Furthermore, malaria transmission in Gambia has been much reduced in recent years leading to potentially complex patterns of selection which may affect allele frequencies (180).

I therefore opted to focus further on the Kenyan population. In an attempt to diminish the occurrence of spurious inflated signals, genotype calls at a set of

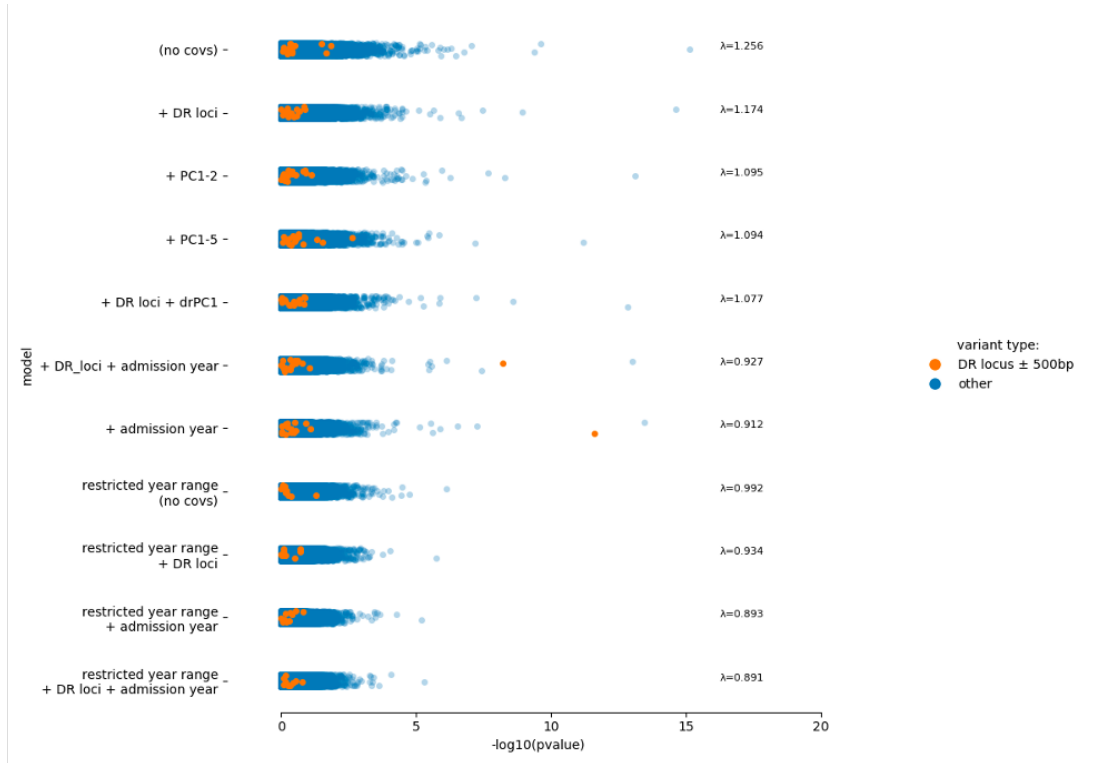


Figure 4.6: Evidence of association of variants across the *Pf* genome with disease severity in the Kenyan population across different GWAS models. Evidence of association ($-\log_{10}$ P-values) for all tested variants in a GWAS scan comparing severe cases from the MalariaGEN CP1 dataset and mild cases from the MalariaGEN Pf7 dataset in Kenya. Different sets of included covariates are on the y-axis (abbrev. Principle components (PC), drug resistance (DR)). Restricted year range indicates that samples have been reduced to only those collected in 1995-1998 and 2005-2009. Genomic inflation factor lambda is also indicated. Variants are coloured orange if they fall within 500 base pairs of a known drug resistance locus (listed in Supp. Table 4.2 (pg.147)).

known causal anti-malarial DR loci were included in the GWAS model as covariates (Supp. Table 4.2 (pg.147)). This visually appeared to remove a large proportion of strong signals and less inflated results of association ($\lambda_{DRloci}^{Kenya} = 1.174$). Multiple further association analyses using different covariates (PC1-2, PC1-5, admission year, admission year + DR loci, PC1 calculated after removal or DR loci + DR loci) were also compared (Fig. 4.6 on the previous page). The most well-controlled model in Kenya (with a Lambda value closest to 1) was the use of DR loci and admission year which produced a deflated association statistic ($\lambda_{DRloci+admissionyear}^{Kenya} = 0.927$). Despite this, several strongly associated variants remained in the results.

To assess whether the temporal separation in this sample could be contributing to these remaining strong associations in the Kenyan population, I tested a GWAS limited to samples collected in a subset of years where there was a considerable number of both Pf7 and CP1 samples (1995-8, 2005-9). This significantly reduced inflation of the association results even without the use of other covariates ($\lambda_{Nocovariates}^{Kenya(restricted\ years)} = 0.992$). In this version, an association close to the CQ resistance locus (Supp. Fig. 4.18 (pg.143)) still appeared within the results. If real, confidently untangling an association at this locus from confounding due to the spread of anti-malarial DR would be challenging. To adopt a more cautious approach, an additional model incorporating DR loci genotypes was applied to the restricted years set, aiming to eliminate potentially spurious signals, producing a slightly deflated model ($\lambda_{DRloci}^{Kenya(restrictedyears)} = 0.934$). Overall, this model was determined to provide the most suitable comparison to results of the above scan in the GAMCC. For the Gambian population, similar attempts were not found to reduce inflation to a value near 1 and therefore only the Kenyan population was considered for replication (Supp. Fig. 4.17 (pg.142)).

Final severity replication association analysis The results of the final replication model GWAS are shown in Figure 4.7 on the following page (a version of the Manhattan where DR loci were not included as covariates can be seen in Supp. Fig. 4.18 (pg.143)). In this analysis, one rare variant, Pf3D7_14_v3:2092254:A>G, demonstrated significant evidence of association with severity with a very strong effect (OR=0.03 (0.007-0.127), $P=1.96 \times 10^{-6}$, AF=3%). To ensure this is not

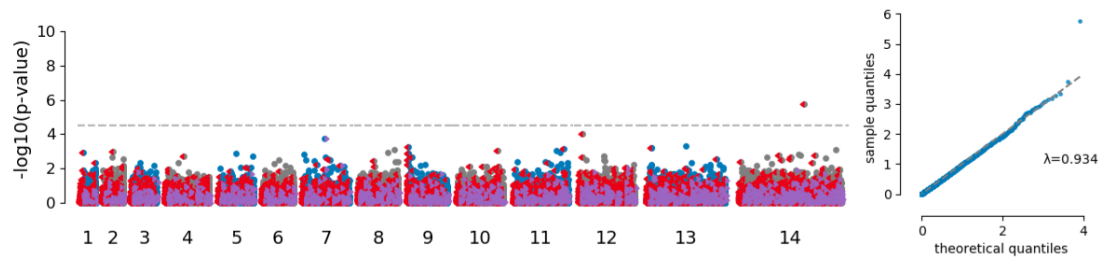


Figure 4.7: Final strength of evidence for the association of variants across the *Pf* genome with disease severity in Kenya using the CP1 and Pf7 datasets. Strength of evidence ($-\log_{10}$ P-values) for the association of variants (y-axis) across the 14 chromosomes in the *Pf* genome (x-axis) with disease severity (left) and quantile-quantile plot displaying observed (y-axis) versus expected (x-axis) $-\log_{10}$ P-values from the GWAS. Genomic inflation factor lambda is also indicated. Variants are coloured purple if the sample sets differed by more than 20% in scaled levels of heterozygosity, and red if they differed by more than 20% (scaled) in sample missingness.

an artefact of independently performed variant calling in the CP1 and Pf7 call sets, variants on Figure 4.7 have been marked to indicate whether the variant had unusually high levels of differential heterozygosity ($>20\%$ scaled difference) or differential missingness ($>20\%$ scaled difference) between the CP1 and Pf7 cohorts (as characterised in Supp. Fig. 4.19 (pg.143)). High levels of differential missingness or heterozygosity might be reflective of technical biases or batch effects in variant calling which might confound the association. Variant Pf3D7_14.v3:2,092,254:A>G did have marginally high differential missingness by this threshold (actual missingness in CP1 was 1.2% compared to no missingness in Pf7), although this does not seem to clearly indicate that this variant was inaccurately called. The position was also aligned to with a high mapping quality in the GAMCC/CP1 call set (Supp. Fig. 4.20 (pg.144)).

4.2.1.3 Inspecting a putative severity locus

Comparing the results of both GWAS scans, the variant Pf3D7_01.v3:557240:T>A discovered in the GAMCC set also showed evidence of association with severity within the CP1/Pf7 GWAS (Fig. 4.8 on the following page). This consistent pattern across datasets suggested a potentially genuine link between this variant and malaria severity. Furthermore, reads in the GAMCC cohort mapped to Pf3D7_01.v3:557240:T>A with high mapping quality scores, indicating this variant was

confidently called and was unlikely to be an artefact of sequencing error (Supp. Fig. 4.21 (pg.144)).

Other variants indicated in these scans as associated with severity, Pf3D7_04_v3:898419:C>T and Pf3D7_14_v3:2092254:A>G, were too rare in the CP1/Pf7 and GAMCC comparisons respectively to be able to compare results (Fig. 4.8).

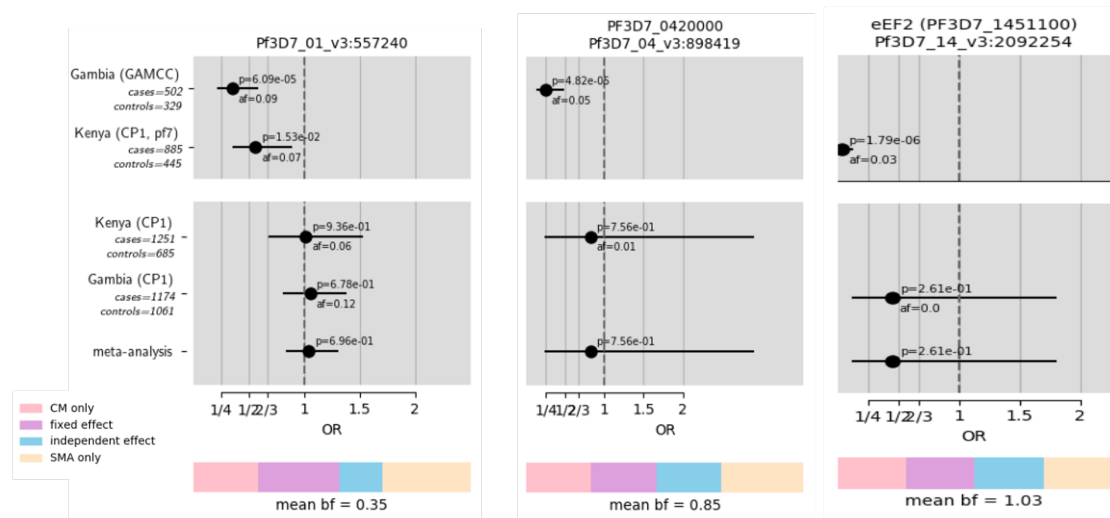


Figure 4.8: Estimated effect sizes compared between different GWAS scans at top hits from GWAS of parasite genetic effects on malaria severity. Estimated effects (odds ratio (OR)) of *Pf* variants associated with malaria disease severity across a GWAS scan between severe and mild cases in the GAMCC cohort (top row) or a GWAS scan between severe cases in the MalariaGEN CP1 dataset and mild cases within the MalariaGEN Pf7 dataset in samples from Kenya in a restricted set of sample collection years (second row). Results of a multinomial Bayesian meta-analysis of the correlates of severe malaria subtype (CM or SMA) in MalariaGEN CP1 samples from Kenya and Gambia. 95% confidence intervals (CIs) are indicated as error bars.

A putative severity locus The putative severity variant, Pf3D7_01_v3:557240:T>A, is at the distal end of chromosome 1 (Pf3D7_01_v3), falling 18.6Kb within the designated “core” region (102). It is a non-coding single nucleotide polymorphism (SNP) 459bp downstream of the protein coding gene PF3D7_0114500 of unknown function which is listed as a *Plasmodium* exported protein within hypothetical gene family 10 (hyp10) (181). The variant is also 1802bp upstream of an apparent VAR pseudogene (PF3D7_0114400, *Plasmodium falciparum* erythrocyte membrane protein 1 (*PfEMP1*) exon 2). Protein-coding VAR genes encode

*Pf*EMP1 proteins which are believed to be significant virulence factors because of their role in cytoadherence and immune avoidance, however the role of VAR pseudogenes is less well characterised (126).

Evaluating possible confounders The conclusions we could draw from the above analyses were however limited. Firstly, the discovery analysis in the GAMCC set had a small sample size and was underpowered limiting its ability to resolve true associations (Fig. 4.1 (pg.113)). Secondly, although the replication analysis had a more appropriate sample size, the results were arguably more difficult to trust owing to the challenges of intersecting and reliably comparing datasets from separate collections which lack complete sample provenance.

Overall, the discovery GWAS analysis did not produce inflated P-values ($\lambda_{Nocovariates}^{GAMCC} = 1.02$) indicating minimal confounding for this analysis overall. Nevertheless sources of bias may have influenced the association at this locus specifically which would not have influenced lambda. Therefore, I undertook further targeted investigation of possible confounding at the Pf3D7_01_v3:557240:T>A locus itself in order to build confidence in the result.

Relatedness and population structure Population structure within a sample is well known to impact GWAS results. A PCA ($N_{variants}=8845$) made clear that there were highly related samples in the GAMCC set (highlighted in black in Supp. Fig. 4.22 (pg.145)) which strongly defined the PCs. To determine the nature of the structure of this sample more clearly, a second PCA was therefore performed with these highly related individuals (with a kinship value > 0.5 , $N = 74$) excluded (Fig. 4.9a on the next page) and the resulting PCA was broadly shapeless with loadings across PC1-5 spread moderately evenly across the *Pf* genome (Fig. 4.9b on the following page). This indicated that there was minimal population structure in the set. Outliers within this PCA appeared to be caused by moderate levels of relatedness at the margin of the relatedness filtering threshold (Fig. 4.9a on the next page).

High levels of observed relatedness within *Pf* WGS sets are not unusual (e.g. (51)) and could be caused by a number of factors. The GAMCC samples were collected from individuals living in a relatively small area and within a short time

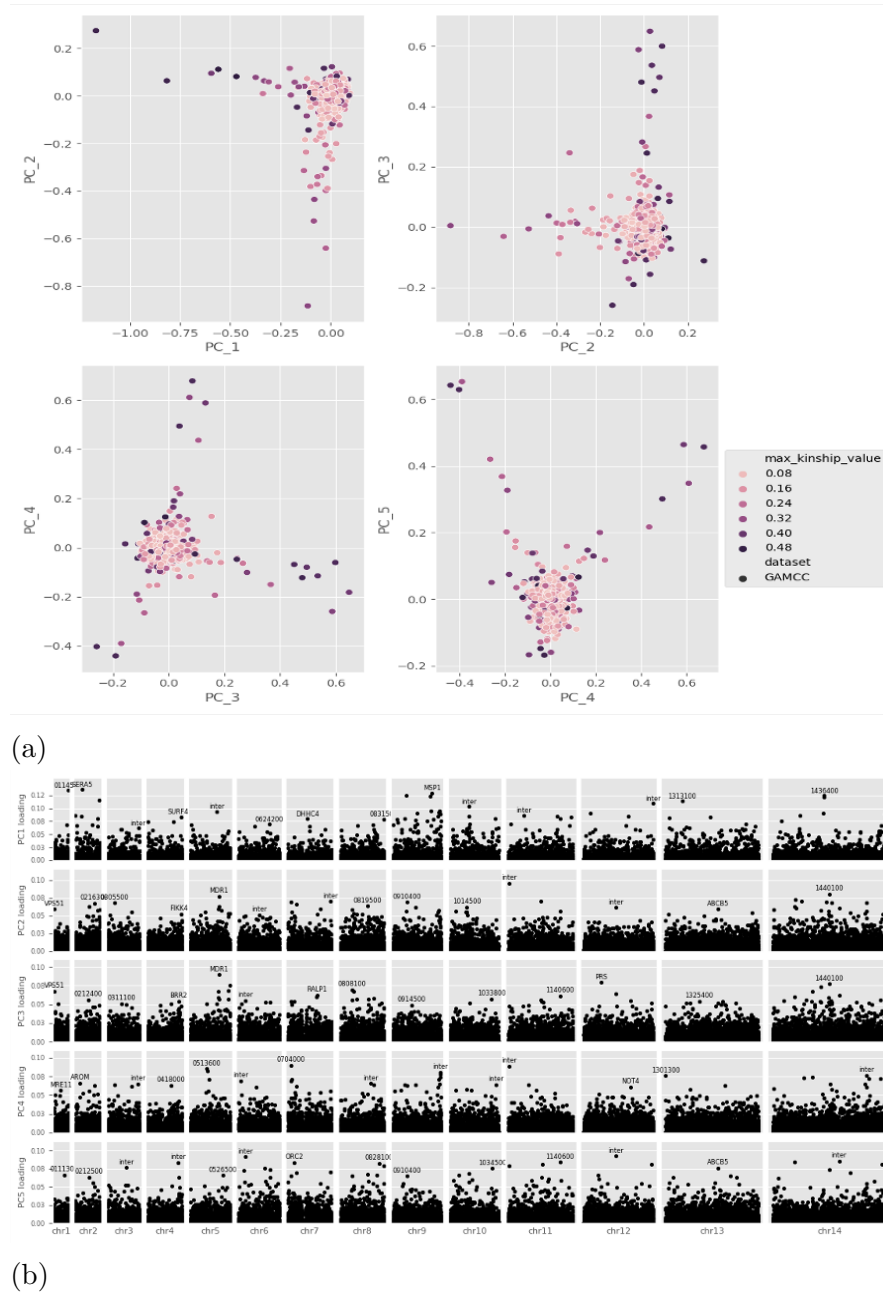


Figure 4.9: **PCA in the GAMCC dataset.** **A.** PCA with highly related samples (kinship > 0.5) removed. Markers are coloured by sample collection year. Marker shape indicates the source of the sample: the GAMCC set (circles), MalariaGEN CP1 set (plus symbols) and MalariaGEN Pf7 set (triangles). **B.** Loadings of each PC (indicated on the y-axis of each panel) across the 14 chromosomes in the malaria parasite genome.

frame, therefore it is not unlikely that individuals may have become infected with closely related parasites. Infected individuals might therefore have shared their environments which might separately influence disease severity and could thus have introduced confounding into the model. It is unfortunately impossible to determine whether individuals shared households this long after their collection in the late 1980s. However, although challenging to confirm, we find it likely that there were attempts to avoid this confounding in the original study this sample set was collected for (52). To reinforce that this notion was valid, the GWAS scan itself in the GAMCC cohort was reperformed following the removal of highly related parasite samples (kinship value>0.5) and including PCs 1-3 (to account for any population structure and marginal relatedness following this exclusion). This made little difference to the calculated P-values across the GWAS (Pearson's $R=0.87$, Supp. Fig. 4.23 (pg.145)).

Human genetics It is known that human factors have strong impacts on the severity of malaria (described in detail in Introduction section 1.3 (pg.12)). Human genetics is also known to impact parasite genetics as has been recently shown in the case of the human resistance allele HbS and the *Pf*sa loci (130). There is therefore the potential that human genetic factors were impacting the results of these GWAS. Unfortunately, complete human WGS in the GAMCC set was not available, but genotyping was present for two clinically important malaria resistance loci (6); the HbS allele and the ABO blood group (the human leukocyte antigen (HLA)-B locus was also typed, investigations into which are made in Chapter 4 (pg.110)). To help ensure that the association of Pf3D7_01.v3:557240:T>A was not confounded by these, the GWAS was rerun: (i) excluding samples who carry one or two HbS alleles and (ii) using ABO blood group as a covariate. The HbS typing of the GAMCC set was reperformed as part of the present research (see Chapter 3 (pg.87)), but ABO blood groups were ascertained at the time of sample collection and have not since been validated. Neither had major impacts on the confidence of the association of Pf3D7_01.v3:557240:T>A with disease severity (Fig. 4.10a on the following page).

Complexity of infection In areas where malaria has been historically highly endemic, such as The Gambia and Kenya, poly-clonal or “complex” infections

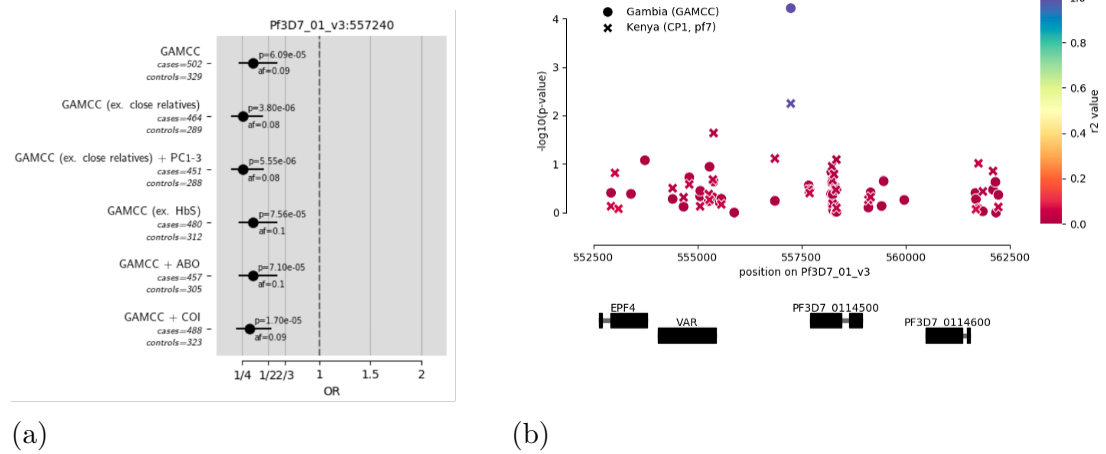


Figure 4.10: Comparison of strength of association at Pf3D7_01_v3:557240:T>A correcting for possible confounders. A. Estimated OR and 95% CIs for the association of variant Pf3D7_01_v3:557240:T>A with disease severity from a GWAS scan performed between severe and mild cases in the GAMCC set using different covariates indicated on the y-axis. Allele frequencies and P-values are indicated on the OR values, and sample sizes in cases (severe cases) and controls (mild cases) are indicated in the y-axis. For results of the Bayesian meta-analysis, ORs for a comparison between CM and SMA cases against other severe cases are shown. **B.** a detailed view of the association at Pf3D7_01_v3:557240 within the two scans of severe against mild cases included in the chapter; GAMCC (circles) and severe CP1 cases against mild Pf7 cases from a restricted time period in Kenya (crosses). Pearson's R-squared values between the central variant and other local variants are indicated by the colour of the marker, as described by the colour bar (right). At the bottom the positions of genes for the region of the genome are shown. Exons within genes are indicated as thicker black bars.

(single infections carrying multiple genetically distinct parasite isolates of the same species) are common (163; 159). Whether complexity of infection (COI) has an effect on disease severity is not fully resolved with apparent positive (182) and negative relationships between COI and severity having been observed previously (38). If COI does impact severity, it is possible that this could have confounded the results of these analyses. COI can be measured in WGS data by identifying variants with multiple alleles present, however this depends on sequencing depth. The GAMCC dataset, because it was entirely sequenced via the same method to a similar depth and at the same sequencing centre, offers a rare opportunity to compare COI between severe and mild cases with minimal risk of confounding due to batch effects.

Two metrics were used to assess COI within the GAMCC cohort; the Fws statistic and using ‘coiaf’, a statistical software which was designed to infer COI directly from *Pf* WGS data (methods are described in full in Materials and Methods section 2.9.1 (pg.69)). These measures were found to strongly correlate ($R=-0.96$) but coiaf’s ‘continuous COI’ metric appeared to provide enhanced resolution at higher levels of mixture (as shown by the non-linear relationship in Supp. Fig. 4.24 (pg.146)). Fws (transformed into its z-scores using the inverse normal cumulative distribution) was not found to strongly associate with disease severity in the GAMCC set but there was some evidence that continuous COI was associated with severity in this set ($P=2.43 \times 10^{-2}$) (Fig. 4.11 on the next page). Severe cases appeared to be more clonal (reduced COI) in comparison to the mild cases within the GAMCC set.

Similar assessments of differences in COI between severe and mild cases in the Kenyan population had very small P-values, despite the inclusion of various covariates such as admission year and sample preparation type (Fig. 4.11 on the following page). Other features such as sequencing depth or sequencing platform warrants further exploration as these, or other confounding factors may be impacting these results.

Due to the suggestion that differences in COI existed between severe and mild cases in the well-controlled GAMCC comparison, a GWAS scan including COI as a covariate was conducted. The inclusion of this covariate also did not markedly reduce the association of *Pf*3D7_01_v3:557240:T>A with severity (Fig. 4.10a

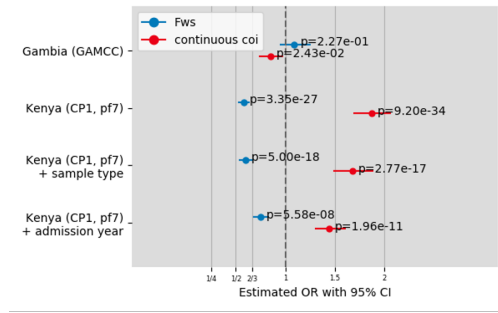


Figure 4.11: **Estimated effect sizes for Fws and COI on malaria severity in different dataset comparisons.** Estimated effects (odds ratio (OR)) of continuous COI (as calculated by the programme coiaf)(red)and the Fws statistic (blue) on whether an infection is severe or mild computed using severe and mild cases in the GAMCC cohort (top row) and a GWAS scan between severe cases in the MalariaGEN CP1 dataset and mild cases within the MalariaGEN Pf7 dataset in samples from Kenya in a restricted set of sample collection years (second row), and within Kenya including the covariate of sample type (that is, whether the sample was prepared using sselective whole genome amplification (sWGA) or not) or admission year.

(pg.129)). Overall, this suggests the association seen at Pf3D7_01.v3:557240:T>A to disease severity in the GAMCC set is not clearly the product of confounding, furthering the notion that it may be truly related to severity.

4.2.2 Does parasite genetics influence severe disease sub-type?

Although only severe cases are present in the CP1 dataset, this set provides further granularity by including sub-type information such as whether the infection included the symptoms CM or SMA. In contrast to the above analysis across the Pf7 and CP1 datasets, a within-CP1 comparison is expected to exhibit minimal noise and a lower risk of confounding. Given the clinical significance of these manifestations, a deeper understanding of the parasite genetics contributing to each symptom could offer critical insights into the mechanisms underlying disease severity, potentially guiding the development of targeted interventions.

To do so *Pf* genomes in the CP1 set associated with the clinical presentations of CM and SMA were compared to other severe infections using multinomial logistic regressions. This method allowed for the evaluation of differential genetic effects across multiple subcategories of a single phenotype. Following this, effects

were meta-analysed using a Bayesian framework across the Gambian and Kenyan populations. Although subtype classification information was available for the GAMCC set making it a potential resource for this analysis, this cohort was not included because it lacked a comparable baseline category required to perform a comparable investigation (other severe malaria cases without CM/SMA). The much larger sample size, and the fact that it encompassed two populations, made the CP1 resource a more powerful tool to assess for CM/SMA effects.

To help distinguish between the influence of any suggested effect, four types of genetic effect were specifically tested for in the meta-analysis: a CM only effect, a SMA only effect, a fixed effect, and an independent effect. The CM and SMA only models tested the assumption that certain genotypes predispose severe infections to each of the respective traits. The fixed effect model tested whether genetic variation contributes uniformly to each severe disease subtype and the independent model allows for the possibility that variation might independently impact each condition (for further details of this method, see Materials and Methods section 2.7.4 (pg.63)).

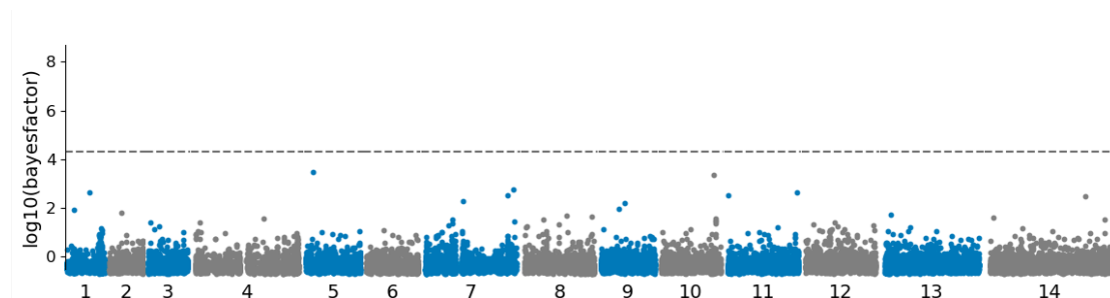


Figure 4.12: Strength of evidence within a meta-analysis of association across variants in the *Pf* genome with severe malaria subtype. Strength of evidence ($-\log_{10}$ P-values) for the association of variants (y-axis) across the 14 chromosomes in the *Pf* genome (x-axis) with disease severity (left). Results reflect a meta-analysis performed across the Kenyan ($N_{variants}^{Kenya} = 38,889$) and Gambian populations ($N_{variants}^{Gambia} = 43,342$) in the MalariaGEN CP1 resource of multinomial logistic regressions testing for genetic effects on the outcomes of CM and SMA against a baseline of other severe infections which do not involve CM or SMA. $\log_{10}$ Bayes Factor (BF) represents the average BF under four tested models of effects (given equal weights): a CM only model, a SMA only model, a fixed effect model and an independent effect model. The dashed line indicates a prior belief that 1 in the approximately 20,000 unlinked regions in the *Pf* genome is associated with severe malaria subtype.

The Bayes Factor (BF) averaged across the four tested models are shown in

Figure 4.12 on the previous page. A reasonable prior belief might be that there are 20,000 independent (unlinked) regions of the *Pf* genome, and one of these might be associated with severity subtype. However, no variants reached this nominal level of significance in this analysis (indicated by the dashed line) (Fig. 4.12 on the preceding page) (further descriptions of the interpretation of BFs can be found in Materials and Methods section 2.7.5 (pg.65)). The most significantly associated hit within the meta-analysis fell on Pf3D7_05_v3:196418:T>G (meta-analysis $OR^{CM+SMAv.other}=0.572$ (0.461-0.709), $AF_{CP1}^{Kenya}=3\%$, $AF_{CP1}^{Gambia}=15\%$, mean BF = 2981.21). This variant was not compared between severe and mild cases in the GWAS performed on the GAMCC set because it failed to reach the minimum AC required but was strongly associated with reduced likelihood of CM and SMA within severe infections from the Gambia (driven by a fixed effect) (Fig. 4.13 on the following page). Pf3D7_05_v3:196418:T>G falls in the protein coding gene PF3D7_0504800 which has no known function. It is a missense mutation (Lys2772Thr) annotated as having a moderate effect by the programme snpeff (151).

Another common variant, Pf3D7_10_v3:1383191:T>A, was an upstream gene modifier to the protein coding gene methionine-tRNA ligase (MRScyt). It was found to be associated with CM and SMA (mainly by a fixed effect) in both Gambia and Kenya (meta-analysis $OR^{CM+SMAv.other} = 1.406$ (1.224-1.614), $AF_{CP1}^{Kenya}=47\%$, $AF_{CP1}^{Gambia}=47\%$, mean BF = 2320.35) (Fig. 4.13 on the next page). No strong evidence of effect of this variant was observed in the GAMCC set between severe and mild cases despite a high allele frequency in this set.

The third most associated variant, Pf3D7_07_v3:1339175:T>A, fell within an upstream gene modifier to the non-coding RNA gene PF3D7_0731000, a downstream gene modifier of the protein coding gene PF3D7_0731200 (an exported protein of unknown function) and an intronic variant to the protein coding gene EMP1-trafficking protein (PTP2, PF3D7_0731100) (Fig. 4.13 on the following page). Evidence that this variant is associated with CM/SMA was however very minimal (mean BF = 570.9).

Assessing the putative severity variant Pf3D7_01_v3:557240:T>A identified as potentially associated with disease severity within the earlier analyses (Section 4.2.1.3 (pg.124)) in this meta-analysis, indicated there was no strong evidence of

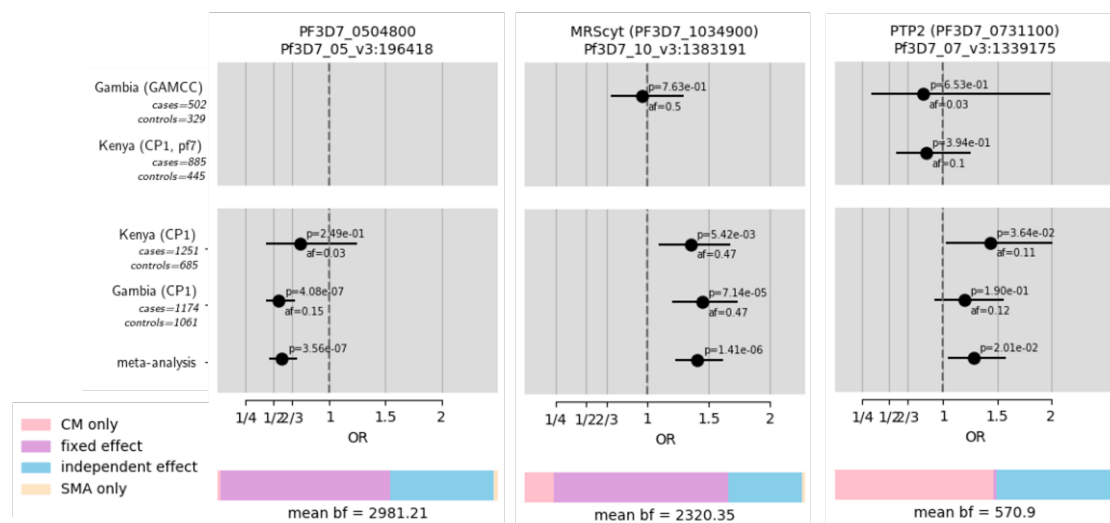


Figure 4.13: **Estimated effect sizes compared between different GWAS scans at top hits from the meta-analysis of parasite genetic effects on severe malaria subtype.** Estimated effects (odds ratio (OR)) of *Pf* variants associated with malaria disease severity across a GWAS scan between severe and mild cases in the GAMCC cohort (top row) or a GWAS scan between severe cases in the MalariaGEN CP1 dataset and mild cases within the MalariaGEN Pf7 dataset in samples from Kenya in a restricted set of sample collection years (second row). Results of a multinomial Bayesian meta-analysis of the correlates of severe malaria subtype (CM or SMA) in MalariaGEN CP1 samples from Kenya and Gambia. 95% CIs are indicated as error bars. Below shows colour bars indicating the proportion of the mean BF in the severe subtype meta-analysis that is contributed by each of the four tested models: CM only (pink), SMA only (yellow), fixed effect (purple) and independent effect (blue). Gene name (where applicable), variant chromosome and position are indicated at the top of each panel.

effect on severe disease subtype (Fig. 4.8 (pg.125)). This could indicate that Pf3D7_01_v3:557240:T>A, if truly associated with severity, acts between mild and severe cases rather than within severe infections.

4.3 Discussion

In this chapter, I presented investigations into the impact of parasite loci on disease severity. My analysis encompassed two levels of severity, initially comparing severe versus mild cases using the rigorously phenotyped GAMCC resource, alongside a comparable replication contrasting severe cases from CP1 with community infections from Kenya in the Pf7 resource. Further, I presented evidence of the influence of variants on severity subtypes within severe cases, utilizing CP1 data from both Kenya and Gambia.

A key result from this chapter was the discovery that the Pf3D7_01_v3:557240:T>A variant appeared to associate with severity in the GAMCC sample set and this association was replicated within the Kenyan population (Fig. 4.8 (pg.125)). The association of this variant was not found to be explained by population structure, relatedness in the sample, host HbS or ABO status or levels of COI in the GAMCC (Fig. 4.10a (pg.129)). Under an idealised model of TDR, there was just under a 70% probability that the observed association represented a real signal in the GAMCC cohort. This calculation did not account for sample missingness, which, given the already small sample size, would likely reduce this value. Although further investigations are required to validate whether Pf3D7_01_v3:557240:T>A effects disease severity, the replication evidence is suggestive of a real effect.

It is however puzzling that this variant did not associate with severe disease subtypes in a meta-analysis of CM and SMA cases from Gambia and Kenya (Fig. 4.8 (pg.125)). The association of Pf3D7_01_v3:557240:T>A with severity was identified in the GAMCC by comparing severe cases (CM, SMA, or hypoglycaemia) with mild cases (children with parasitaemia >2,500 parasites/ μ L). In contrast, the sub-type meta-analysis compared severe cases with CM or SMA to other severe cases with different features, such as hypoglycaemia or respiratory distress, without a parasitaemia threshold. It is possible that there may be some phenotypic overlap between the mild cases in the GAMCC and “other” severe cases in CP1,

so we might therefore expect that there should be an observable replicative signal at the Pf3D7_01_v3:557240:T>A locus. However, given our limited understanding of the genetic architecture of severity, it is difficult to predict whether a replicative signal should be expected. It may be that Pf3D7_01_v3:557240:T>A plays a broader role in the progression from mild to severe disease, rather than influencing specific severe subtypes. Further research is required to clarify whether this variant modulates overall severity or is context-dependent within specific clinical presentations.

Pf3D7_01_v3:557240:T>A was not in linkage disequilibrium (LD) with any nearby SNPs assessed as part of this study (Fig. 4.10b (pg.129)). Minimal local LD is not unusual in the *Pf* genome within which LD tends to decay within 500bp (167). This does not exclude the possibility that Pf3D7_01_v3:557240:T>A tags an unseen causal feature or otherwise affects the expression of local genes. The closest gene is called PF3D7_0114500, which appears within a list of candidate virulence genes within the *Pf* genome based on clustering with other human pathogenic malaria species (183), but otherwise very little is known about the function of the encoded protein.

The proximity of this variant to an apparent partial VAR gene (PF3D7_0114400, located 1.8Kb away) is intriguing. The VAR gene family is thought to play a significant role in malaria severity due to their production of *Pf*EMP1 proteins, which mediate the cytoadherence of parasitized red blood cells to host tissues and enable immune evasion through antigenic variation (126). The *Pf* genome contains an extensive repertoire of approximately 60 full-length VAR gene copies, although only one is expressed at a time. In addition, there are partial copies, such as PF3D7_0114400, which encode only one of the two exons which encode *Pf*EMP1s. PF3D7_0114400 encodes the second exon, which is smaller and generally more conserved than the first, and has received comparatively less scientific attention (184).

Specific investigations have been made with a focused view into the impact of certain VAR genes on disease using expression-based quantification studies. For instance, some subgroups of VAR genes have been shown to have differential expression in severe malaria cases, while others have not demonstrated significant expression differences between severe and uncomplicated cases (185; 127). There

have also been suggestions that partial VAR genes may be expressed during malaria infection, although the functional significance of this, if any, is largely unclear (186). Given the current observations, it is not yet possible to definitively link the association of Pf3D7_01.v3:557240:T>A with the VAR gene family or disease severity. Further *in silico* studies are necessary to investigate the potential role of Pf3D7_01.v3:557240:T>A in the pathogenesis of severe malaria.

Performing the replication analysis within the Kenyan population presented in this chapter was challenging. Methodologically, the intersection of the public Pf7 call set with the GAMCC/CP1 call set generated in this work led to the retention of only a small number of variants available for analysis, and temporal separation of samples between the datasets meant any true relationships were difficult to distinguish from the effects of confounding through genetic drift. Through the inclusion of different covariates and the sub-setting of samples to those with more temporal overlap, a model with low levels of genome-wide P-value inflation (a lambda close to 1) was generated for the Kenyan population (Fig. 4.6 (pg.122)), but not The Gambia (Supp. Fig. 4.17 (pg.142)). The large shift in AF of the CQ resistance mutation in the CRT gene (Fig. 4.5 (pg.121)) and of other DR loci in Gambia may have contributed to this. PCA analyses of the Gambian population further support this, as these regions exert a strong influence on the results (Fig. 4.3a (pg.118)). Furthermore, changes in transmission intensity in this region over the time frame of these samples to a more seasonal as opposed to endemic transmission type have influenced genomes within the Gambian population (7). Seasonal transmission leads to pronounced population bottlenecks which act to amplify the effects of genetic drift (106). Evidence from a further independent set where these factors are not so influential would therefore be beneficial to this work to further judge the observed signals. The MalariaGEN CP1 cohort consists of severe infections sampled from other populations which are also represented in Pf7, such as Mali and Cameroon, and so efforts to sequence these severe parasite genomes could be undertaken to this end.

While an appropriate lambda value was achieved for the Kenyan population, this does not rule out the potential for confounding at individual variants. Unlike the GAMCC set (Fig. 4.10a (pg.129)), further exploration of this issue would be challenging due to the limited metadata available for the CP1 and Pf7 samples,

restricting the use of this dataset for discovery analysis. For example, Pf3D7_14_v3:2,092,254:A>G was strongly associated with severity in this set because it was only seen in Pf7 Kenyan samples and not at all in CP1 (Fig. 4.8 (pg.125)). This is interesting as, given the size of the CP1 set, we might expect even rare alleles should be represented. However, further investigation is needed to determine whether this may be an artefact of confounding.

A significant caveat of this analysis is the uncertain phenotyping categorisation of the Pf7 samples as “mild”. We find it likely that the Pf7 samples contain few, if any, severe cases and range from moderately ill individuals collected in clinics to asymptomatic infections from community screening (see Materials and Methods section 2.7.1.2 (pg.59)). Heterogeneity within this group could act to reduce the likelihood of finding effects which act within a certain category of infection and make it challenging to understand the impact of any assessed effects. Cataloguing more specific disease phenotypes for these samples may be possible with collaboration with partner institutes, but this was determined to be beyond the scope of this thesis.

An analysis of severe disease subtypes in the large MalariaGEN CP1 cohort showed minimal evidence of parasite genetic influence on the development of CM or SMA, despite more statistical power than the GAMCC GWAS. The strongest association detected at variant Pf3D7_05_v3:196418:T>G, exhibited a significant effect on predisposing to CM/SMA in the Gambian but not the Kenyan population, suggesting the possibility of population-specific effects (Fig. 4.13 (pg.134)). This variant is located within the gene PF3D7_0504800, which currently lacks a known function. Notably, while this variant has a moderate AF in the Gambian CP1 dataset ($AF_{CP1}^{Gambia}=15\%$), it is much rarer in the Kenyan population ($AF_{CP1}^{Kenya}=3\%$) and was absent in the earlier GAMCC cohort (collected in 1988-9 compared to 1996-2009), indicating a relatively rapid increase in frequency. The recent spread of DR might have caused this through genetic hitchhiking, and this may not have adequately been controlled for as no covariates were included in these scans. However, this seems unlikely given that the nearest known DR locus in the MDR1 gene (Supp. Table 4.2 (pg.147)) is located more than 700Kb away from this variant.

A second variant, Pf3D7_10_v3:1,383,191:T>A in the MRScyt (PF3D7_1034900) showed a consistent effect in Gambia and Kenya (Fig 4.13 (pg.134)). This gene encodes enzymes involved in protein synthesis within the parasite cytoplasm and has been identified as potential drug target (187). However, given the limited evidence for association at these variants (both showing low BF_s), it remains difficult to draw definitive conclusions without additional data.

The relatively modest genetic effects observed in parasite loci on severity in all the analyses presented in this chapter raises important questions about the factors driving severe malaria. Rather than large effects that can be observed in these samples, it may be that parasite genetic effects are small, requiring very large sample sizes to detect. Adjusting sample selection might also aid explorations if this is the case. The mild samples in the GAMCC set, the central pilot performed here, were in fact clinically ill rather than representing asymptomatic infections. Mirroring the approach taken by investigations into the effect of human genetics on severe malarial disease (6; 11; 54), it may be more revealing to test severe malaria cases against asymptomatic cases in the first instance because effects may be more pronounced.

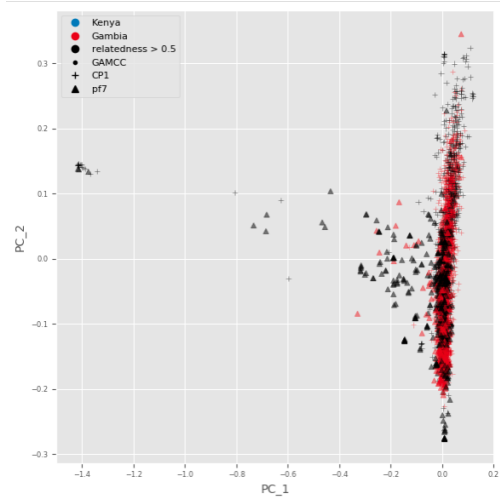
Another possibility is that these assessments, which focused purely on the core regions of the *Pf* genome, were not able to capture the relevant genetics driving severity. Gene families involved in cytoadherence, such as VAR, subtelomeric variable open reading frame (STEVAR), and repetitive interspersed family (RIFIN), are predominantly located in the sub-telomeric regions, which are known for their repetitive and highly variable nature (123). This variability complicates the analysis of these genes. Even though specialised methodologies have been developed to recover VAR gene sequences from short-read data (184), accessing these sequences remains challenging with this data type. The high variability within these regions poses difficulties for alignment software, which rely on a reference genome that may significantly differ from the sample (102). The long-read sequencing platform is much more able to resolve these regions because sample genomes can be assembled *de novo*, allowing variable regions to be pieced together (103). Currently only few datasets of long read *Pf* data exist (184), but as producing long read sequencing data becomes more cost-effective, the ability to test associations across the entire parasite genome might be realised.

Finally, a further possibility is that malaria severity may be more closely related to host-pathogen interactions, rather than being driven by parasite genetics. For example, the cytoadherence mediated by infected red blood cell surface proteins might also be mediated by host variation at tissue binding sites. Only recently it was shown that certain parasite loci known as Pfsa associated with the human resistance polymorphism HbS (130), although the functional mechanism this refers to is not yet understood. This suggests a need for further research into the interface between human and parasite genetic factors, which will be explored in the next chapter.

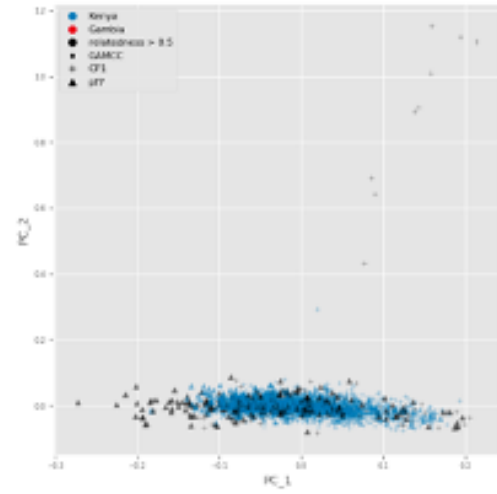
Importantly, this analysis provides the first genome-wide assessment of the parasite genetic effects on severity and once published, the summary statistics will provide a resource for researchers to better understand the potential effects of parasite loci their work uncovers. Because there appears to be no major genetic differences between severe and mild cases in the GAMCC resource revealed here, it is possible to consider these samples as one group for further analyses presented in this thesis.

4.A Supplementary Figures and Tables

4.A.1 Supplementary Figures

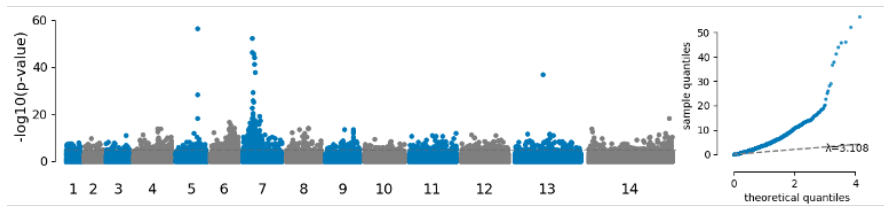


(a) Gambia

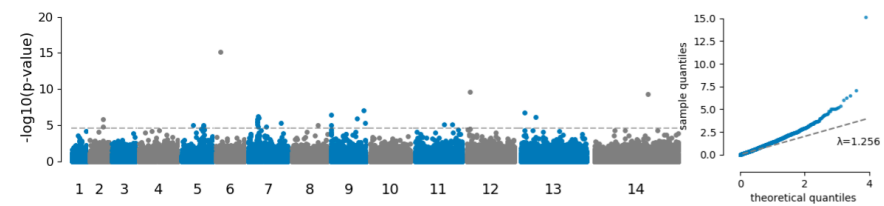


(b) Kenya

Figure 4.14: PCA results without highly related samples removed. Principal component analysis (PCA) results within the Gambian and Kenyan populations. PC1 v. PC2 are shown. Markers coloured in black indicate those where kinship values (as calculated by QCTOOL) are greater than 0.5. The marker shape indicates the source of the sample: the GAMCC set (circles), MalariaGEN CP1 set (plus symbols) and MalariaGEN Pf7 set (triangles).

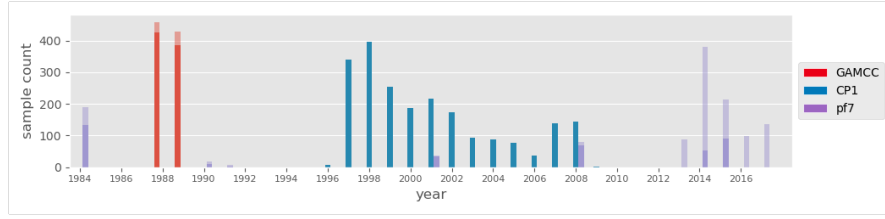


(a) Gambia

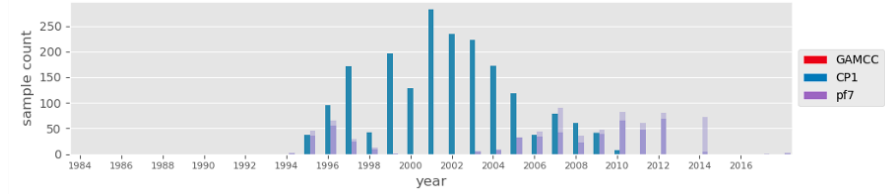


(b) Kenya

Figure 4.15: Strength of evidence for the association of variants across the *Pf* genome with disease severity in the Gambia and Kenya using the CP1 and Pf7 datasets. Strength of evidence ($-\log_{10}$ P-values) for the association of variants (y-axis) across the 14 chromosomes in the *Pf* genome (x-axis) with disease severity (left) and quantile-quantile plot displaying observed (y-axis) versus expected (x-axis) $-\log_{10}$ P-values from the GWAS. Genomic inflation factor lambda is also indicated.



(a) Gambia



(b) Kenya

Figure 4.16: Sample sizes of *Pf* genomes in the Gambian and Kenyan populations across time. Sample sizes (y-axis) by sample collection year (x-axis) for samples collected in A. Gambia and B. Kenya. The colour of the bar indicates the source dataset: GAMCC (red), MalariaGEN CP1 (blue) and MalariaGEN Pf7 (purple).

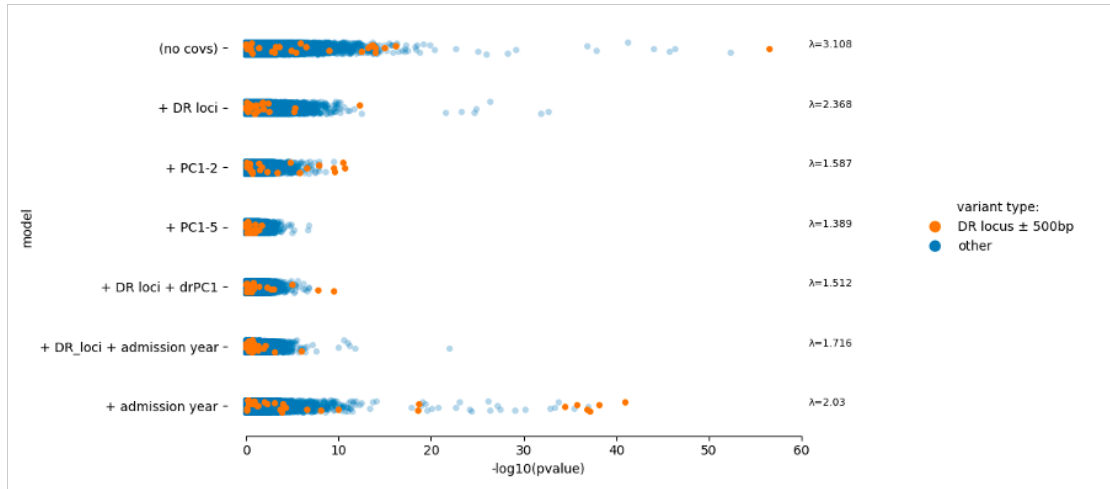


Figure 4.17: Evidence of association of variants across the *Pf* genome with disease severity in the Gambian population across different GWAS models. Evidence of association ($-\log_{10}$ P-values) for all tested variants in a GWAS scan comparing severe cases from the MalariaGEN CP1 dataset and mild cases from the MalariaGEN Pf7 dataset in Gambia. Different sets of included covariates are on the y-axis (abbrev. Principle components (PC), drug resistance (DR)). Restricted year range indicates that samples have been reduced to only those collected in 1995-1998 and 2005-2009. Genomic inflation factor lambda is also indicated. Variants are coloured orange if they fall within 500 base pairs of a known drug resistance locus (listed in Supp. Table 4.2 (pg.147)).

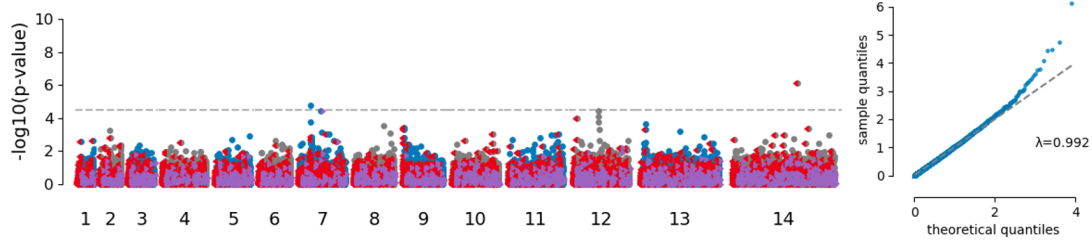
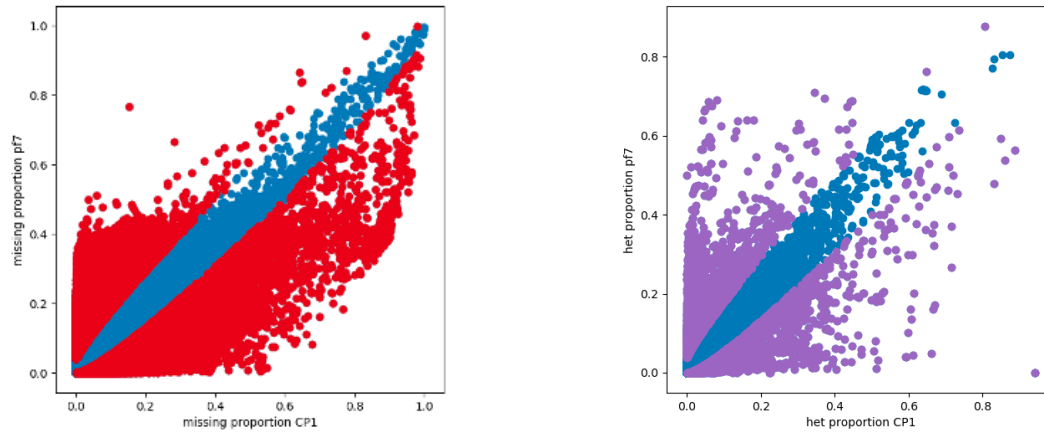


Figure 4.18: **Strength of evidence for the association of variants across the *Pf* genome with disease severity in Kenya using the CP1 and Pf7 datasets without using drug resistance loci as covariates.** Strength of evidence ($-\log_{10}$ P-values) for the association of variants (y-axis) across the 14 chromosomes in the *Pf* genome (x-axis) with disease severity (left) and quantile-quantile plot displaying observed (y-axis) versus expected (x-axis) $-\log_{10}$ P-values from the GWAS. Genomic inflation factor lambda is also indicated.



(a) Missingness

(b) Heterozygosity

Figure 4.19: **Differential missingness and heterozygosity between variants in the CP1 and Pf7 datasets.** Proportions of heterozygous calls (A) and missing calls (B) for variants in the *Pf* genome is samples from Kenya from the MalariaGEN CP1 resource (x-axis) and MalariaGEN (Pf7) resource. Variants with a scaled heterozygosity difference and scaled missingness difference $>20\%$ have been coloured (purple or red respectively). Scaled differences were calculated using the formula $scaled difference = \frac{|p_2 - p_1|}{\sqrt{p_2(1-p_2)}}$ where p_1 and p_2 are the proportions in the CP1 and Pf7 dataset respectively.

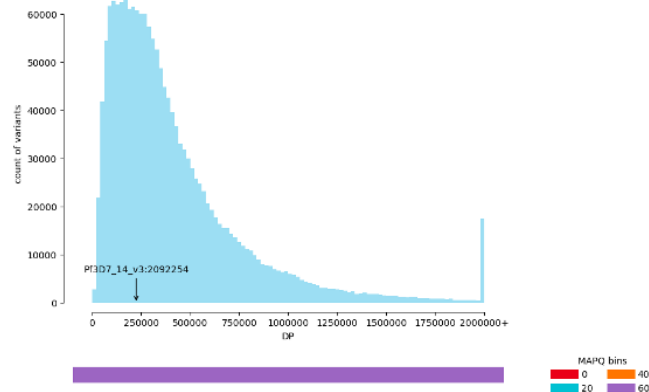


Figure 4.20: **Depth (DP) and mapping quality scores (MAPQ) for Pf3D7_14_v3:2,092,254:A>G with the GAMCC/CP1 callset.** In the top panel the distribution of depth (DP) values for all VQSR pass variants in the combined GAMCC/CP1 call set generated in this thesis are shown. The DP value for the tested variant is indicated by an arrow. Variants with too high DP values indicate they reside in a repetitive region and too low indicates there may be insufficient data to type the variant. The bottom colour bar indicates the proportion of reads in the GAMCC set which have a mapping quality (MAPQ) score of 0-20 (red), 20-40 (blue), 40-60 (orange) or 60+ (purple). MAPQ score indicates the probability that a read aligns elsewhere within the reference genome.

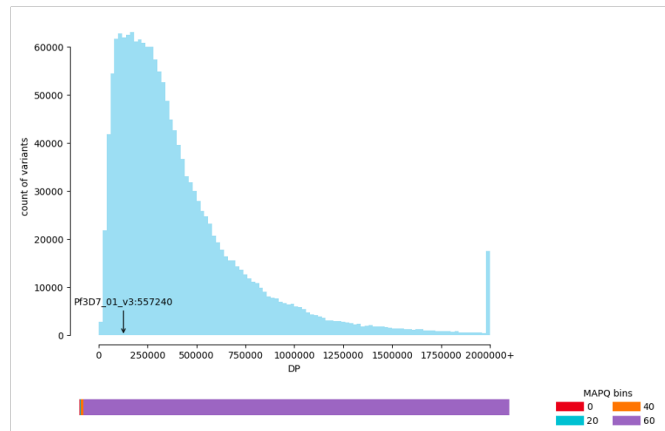


Figure 4.21: **Depth (DP) and mapping quality scores (MAPQ) for Pf3D7_01_v3:557240:T>A with the GAMCC/CP1 callset.** In the top panel the distribution of depth (DP) values for all VQSR pass variants in the combined GAMCC/CP1 call set generated in this thesis are shown. The DP value for the tested variant is indicated by an arrow. Variants with too high DP values indicate they reside in a repetitive region and too low indicates there may be insufficient data to type the variant. The bottom colour bar indicates the proportion of reads in the GAMCC set which have a mapping quality (MAPQ) score of 0-20 (red), 20-40 (blue), 40-60 (orange) or 60+ (purple). MAPQ score indicates the probability that a read aligns elsewhere within the reference genome.

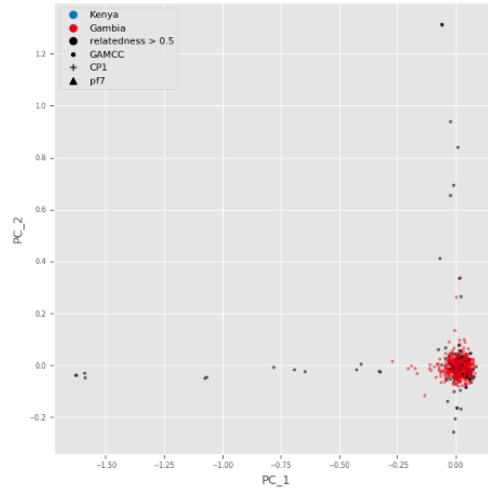


Figure 4.22: **PCA results without highly related samples removed in the GAMCC.** Principal component analysis (PCA) results within the GAMCC samples. PC1 v. PC2 are shown. Markers coloured in black indicate those where kinship values (as calculated by QCTOOL) are greater than 0.5. The marker shape indicates the source of the sample: the GAMCC set (circles), MalariaGEN CP1 set (plus symbols) and MalariaGEN Pf7 set (triangles).

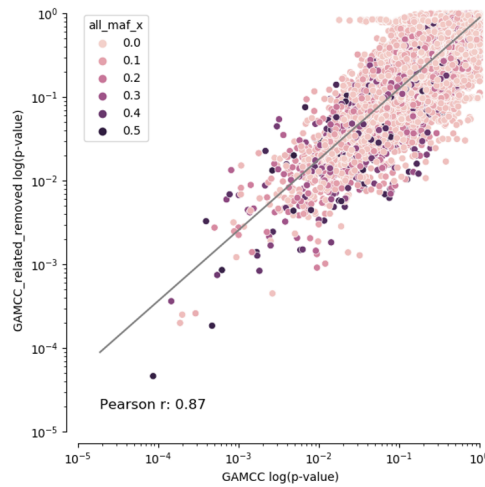


Figure 4.23: **Comparison of association results in the GAMCC with and without highly related samples removed.** Comparisons of strength of association ($-\log_{10}(\text{P-value})$) between GWAS scans performed between severe and mild cases in the GAMCC set and the same scan performed in the samples restricted to unrelated samples. The Pearson's correlation R value is shown. Colours of points on both plots indicate the minor allele frequency of the variant in the set as indicated in the legend.

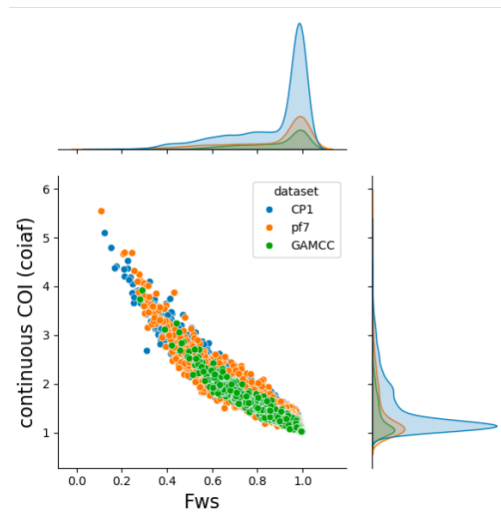


Figure 4.24: **Comparison of Fws statistics and COI calculated using coiaf software in *Pf* WGS data.** Correlation between continuous COI values (y-axis) and the Fws statistics (x-axis). The colour of the point indicates the sample source: GAMCC (green), MalariaGEN CP1 (blue) and MalariaGEN Pf7 (CP1). Distributions for these statistics are also shown opposite each axis.

4.A.2 Supplementary Tables

Locus	Chromosome	Position	Amino acid substitution
CRT	Pf3D7_07_v3	403625	K76T
AAT1	Pf3D7_06_v3	1215233	S258L
DHFR	Pf3D7_04_v3	748410	S108N
MDR1 (A)	Pf3D7_05_v3	958145	N86Y
MDR1 (B)	Pf3D7_05_v3	958440	Y184F
MDR1 (C)	Pf3D7_05_v3	961625	D1246Y
DHPS-PPPK	Pf3D7_08_v3	549685	A437G

Table 4.2: **Drug resistance loci included as covariates in GWAS models.**
Information on drug resistance loci in the malaria parasite genome used as covariates in GWAS scans in this chapter.

Chapter 5

Investigation of co-evolutionary interactions between *Plasmodium falciparum* parasites and the human HbS and *HLA-B* loci

5.1 Introduction

It is well known that humans have evolved genetic resistance mutations to counter malaria infection (as discussed in the Introduction section 1.3 (pg.12)). What is less well-characterised is whether parasites have co-evolved in response. Investigating this is challenging due to a sparsity of datasets which include both human and parasite genetic data. Within one of the few datasets of this type, Band et al. (130) revealed an association between genetic variants in three regions of the *Plasmodium falciparum* (*Pf*) genome and the human haemoglobin (Hb) S polymorphism in severe malaria cases from The Gambia and Kenya (referred to as the MalariaGEN Consortial Project 1 (CP1) dataset) (130). HbS, a point mutation in the beta-globin gene on human chromosome 11, confers strong protection against severe malaria infection (188; 54), and is found to be the region most associated with protection against severe malaria genome wide compared to other human loci (6). Despite this, the mechanisms by which HbS protects carriers from malaria infection remain poorly resolved (71). Band et al. (130) revealed that the degree of protection conferred by HbS depends on parasite genotype and suggested that

malaria parasites have co-evolved to counteract HbS-mediated resistance (130). The parasite genomic regions involved have been termed *Plasmodium falciparum* sickle associated (*Pfsa*)1-3.

Although the study identified three *Pfsa* variants, a further interesting observation was that one of the alleles (*Pfsa*2+) was absent from West African populations, whilst being common in East African populations (130). Moreover, a more recent study of mild malaria cases in Ghana revealed a fourth *Pfsa* locus (termed *Pfsa*4) (131). Although this association was replicated in an independent study in Mali, it was not observed to associate to HbS in Gambian samples in the original study and the *Pfsa*4+ allele was found to be absent in Kenya. This highlights the potential for significant spatial variation in the HbS-*Pfsa* interaction signals across Africa which merits exploration.

Both of the so far published studies contain only a small number of malaria cases in HbS individuals (those who are heterozygous (HbAS) or homozygous (HbSS)) due to the rare occurrence of this state (60 within a sample size of 4054 in (130) and 35 in a sample of 1274 in (131)). The Genetic Analysis of Malaria Cases and Controls (GAMCC) resource contains 13 HbS individuals (total N=805). Although a modest amount, this adds additional power with which to assess *Pfsa* relationships, aiding our comprehension of the complex selective forces acting at these loci.

Clues to the roles of the four *Pfsa* regions in malarial disease may be established through the investigation into their evolutionary history. A natural hypothesis is that HbS-associated (*Pfsa*+) haplotypes have been positively selected by the presence of HbS across Africa. *Pfsa*+ variants are not found in *Pf* parasites where HbS frequency is low such as within Asia (130). In addition, there appears to be a positive correlation between HbS and *Pfsa*+ allele frequency. However, *Pfsa*+ alleles exist at higher frequencies than HbS across some parts of Africa (reaching 60% in certain populations) and parasites which carry these variants have been found to infect individuals who do not carry HbS in high numbers (130). The possible selective forces maintaining the substantial but intermediate allele frequencies of the *Pfsa*+ variants are therefore difficult to explain. Disentangling these forces might offer clues as to how *Pfsa*+ haplotypes contribute to disease.

The GAMCC set holds an additional benefit of allowing us to reassess a widely-cited association that was reported using the GAMCC samples between the class I human leukocyte antigen (HLA)-B*53 and malaria susceptibility (52). It was determined that HLA-B*53-restricted cytotoxic T-lymphocytes recognise a conserved peptide of the liver stage antigen (LSA)-1 protein (88) and later peptides in LSA-3 were also found to be HLA-B*53-restricted (89). This led to the inclusion of these peptides in a malaria vaccine candidate ME-TRAP (189; 190; 191; 192). However, other similar targeted genotyping studies, for example from Kenya (90) and Malawi (91), did not replicate the association between malaria susceptibility and HLA-B*53, and more modern and highly powered genome wide association studies (GWAS) have not found any HLA alleles to be associated with protection against malaria infection (6). ME-TRAP has since been deprioritised due to poor efficacy after testing in a Phase IIB randomised clinical trial (193).

There are multiple possible explanations for this discrepancy in association results. Using recent human typing performed on the GAMCC dataset it is possible to test: (i) whether the original signal was a false positive, (ii) the possibility that the association was caused by genotyping errors in the original study and (iii) the effects of population structure upon the original results. Moreover, motivated by the discovery of the *Pfsa* loci outlined above, it has become clear that variability in parasite genomes might contribute to disparity seen in the human genetic correlates of malaria susceptibility. The generation of *Pf* WGS data available for the GAMCC, described in Chapter 3 (pg.87), allows us to further address the hypothesis that parasite genetic variation might have impacted the observed HLA-B*53 association.

In this chapter I conduct an investigation into potential interactions between parasite and human genetics at the *beta-globin* (*HBB*) and HLA loci. The analysis is split across three sections. First, I report results of genome-wide association analyses between *Pf* variation and the HbS allele in the GAMCC dataset, and compare these results to previously published data, with a view to detecting population specific *Pfsa* interactions. Secondly, I re-evaluate the discovery that HLA-B*53 protects carriers from severe malaria disease and test for regions of parasite genetic interaction with HLA-B*53 (52). Finally, I describe population genetic

analyses of the *Pfsa* loci to initiate exploration of the complex evolutionary forces which may be acting on them.

5.2 Results

5.2.1 Assessing HbS-associated *Pfsa* loci in the *Pf* genome using the GAMCC resource and other datasets

5.2.1.1 *Pfsa1* and *Pfsa3* replicate within the GAMCC dataset

Collection, generation, and quality control (QC) of *Pf* WGS data and production of HbS calls for samples in the GAMCC resource are described in Chapter 3 (pg.87). Of the 831 samples in the GAMCC analysis set, 805 also have HbS genotypes available (Table 5.1). Using these samples I performed a genome-wide scan for associations between parasite biallelic single nucleotide polymorphism (SNP)s and HbS genotype using the program HPTEST. HPTEST fits logistic regression models under a Bayesian framework using HbS genotype as the predictor and parasite genotype as the outcome variable, outputting effect size estimates (odds ratio (OR)s), standard errors, and both frequentist (P-value) and Bayesian measures of evidence for association (Bayes Factor (BF)).

Dataset	Study	Country	Case types	N samples	N samples	N not called	HbAA	HbAS	HbSS	HbAC	HbCC	HbSC	N variants (filtered)
GAMCC		Gambia	Severe and mild	905	831	26	792	12	1	6	0	1	16477
CP1	Band <i>et al.</i> (2022)	Gambia	Severe	2235	2235	20	2194	18	3	Unknown	Unknown	Unknown	23485
CP1	Band <i>et al.</i> (2022)	Kenya	Severe	1936	1936	8	1887	32	9	Unknown	Unknown	Unknown	17619
PF7	Hamilton <i>et al.</i> (2023)	Ghana	Mild	1555	1309	NA	962	29	1	301	11	5	10627

Table 5.1: **Summary of datasets used in this Chapter.**

The BF, which was used to assess strength of evidence, compares the likelihood of the observations under association compared to no association. How to directly interpret this depends on a prior; for example, we might assume that the core regions of the *Pf* genome, which is just under 20 megabases in length, contains $20,000 \times 1$ kilobase (Kb) non recombining blocks. Under a prior that two of these regions associate with HbS in the Gambian population, as has been previously shown through the presence of the *Pfsa1* and *Pfsa3* loci within this population

(130), a $\text{BF}=10,000$ ($\log_{10}\text{BF}=4$) indicates an equal likelihood of association and no association at the locus. It is important to note that this calculation is only illustrative, since the true number of associated loci and their effect sizes are not known. For the purposes of analysing the GAMCC cohort, I determined that $\text{BF}>4000$ ($\log_{10}\text{BF}=3.6$) was suggestive of an association to HbS within this population ($>15\%$ posterior probability using the calculation outlined above). Further associations reported later in this chapter show $\text{BF}>100000$ ($\log_{10}\text{BF}=5$) which is indicative of strong evidence of association ($>80\%$ posterior probability of association). For this analysis I used the HPTEST default prior on the effect of HbS genotype on parasite genotype, which (as described in the original study (130)) allows for relatively large effects while providing a weakly informative, mildly regularising impact on predictions. This prevents model fitting errors when host or parasite allele frequency is small, as is the case for the HbS genotype within malaria infections.

In the GAMCC cohort, variants within three regions across the genome showed $\text{BF}>4000$, including those within the *Pf*sa1 and *Pf*sa3 loci (Fig. 5.1 on the next page). At the *Pf*sa1 locus, the most-associated variant was Pf3D7_02_v3:630,545:C>T ($\log_{10}\text{BF}=3.62$, $\text{OR}=11.28$ (3.74-33.99), $P=1.68\times 10^{-5}$, $N=746$, allele frequency (AF)=12.2%). This was -645base pair (bp)s from the previously identified lead *Pf*sa+ SNP (Pf3D7_02_v3:631190:T>A) which was also strongly associated ($\log_{10}\text{BF}=2.82$, $\text{OR}=9.08$ (2.93-28.2), $P=1.35\times 10^{-4}$, $N=755$, AF=12.2%). At the *Pf*sa3 locus the most-associated variant was Pf3D7_11_v3:1054282:T>C ($\log_{10}\text{BF}=3.88$, $\text{OR}=10.04$ (3.40-29.68), $P=3.03\times 10^{-5}$, $N=724$, AF=19.0%), which fell -3753bp from the previously identified lead *Pf*sa3+ SNP (Pf3D7_11_v3:1058035:T>A, $\log_{10}\text{BF}=2.44$, $\text{OR}=7.44$ (2.43-22.75), $P=4.31\times 10^{-4}$, $N=744$, AF=14.0%). This corroborates previous results that these regions associate with HbS in the Gambia at the earlier time point of the GAMCC samples.

The *Pf*sa4 locus however, mirroring previously reported results of tests for associations in the CP1 Gambian dataset (131), did not show strong evidence of association in the GAMCC dataset (at Pf3D7_04_v3:1121472:T>A $\log_{10}\text{BF}=0.63$, $\text{OR}=0.38$ (0.12-1.22), $P=1.04\times 10^{-1}$, $N=726$, AF=86.5%) (Fig. 5.2 (pg.155)). The lead HbS-associated SNPs at the *Pf*sa2 locus (Pf3D7_02_v3:814288:C>T, referred

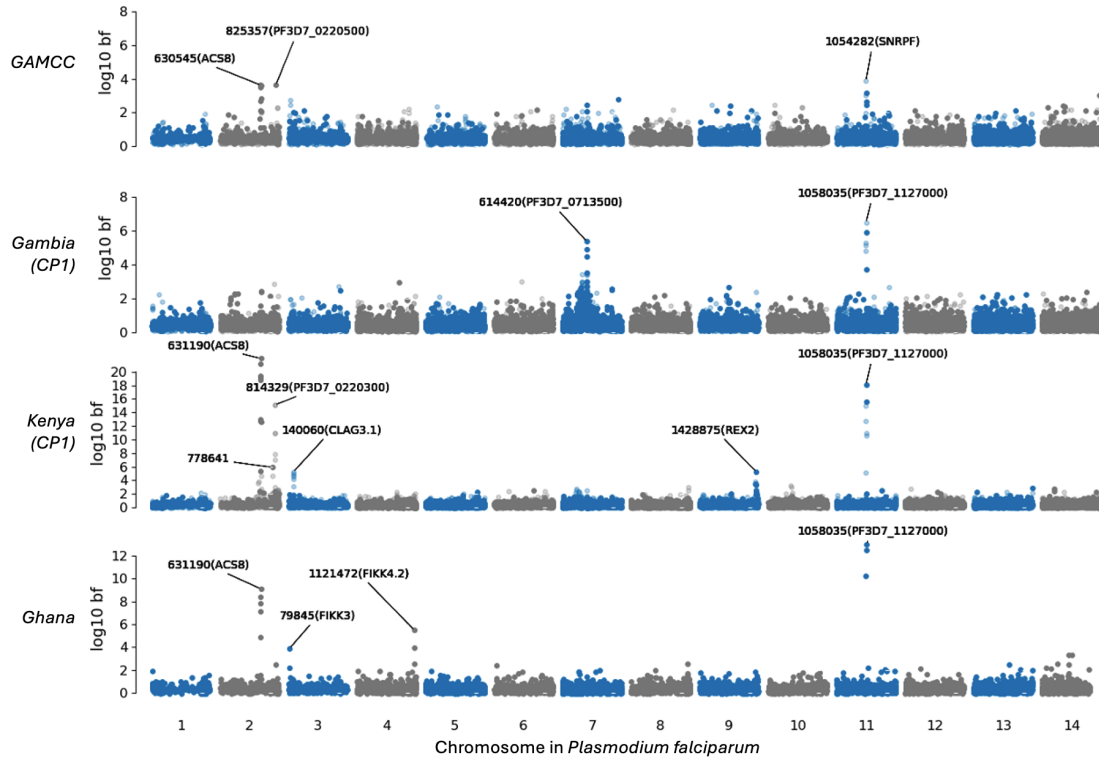


Figure 5.1: **Strength of evidence for associations between *Pf* variation and HbS within the GAMCC, CP1 Gambia, CP1 Kenya and Ghanaian datasets.** Strength of evidence ($\log_{10}$ Bayes Factor, y axis) for variants across the 14 chromosomes in the *Pf* genome (x axis) with the human HbS polymorphism. Solid points indicate that for that variant less than 10% of samples had missing or heterozygous calls (i.e. the sample size was at least $0.9 \times N$ samples). Transparent points indicate variants with high levels of heterozygous calls. Peak associations are labelled with their position on the chromosome and the gene name within the Pf3D7 reference assembly. Panels reflect: Gambian cases and controls (GAMCC) cohort generated as part of this thesis (top) (N samples=805, N variants=16,477), MalariaGEN CP1 samples of severe cases in infants in The Gambia (second) (N samples=2215, N variants=23,485), MalariaGEN CP1 samples of severe cases in infants from Kenya (third) (N samples=1928, N variants=17619) and mild cases collected in Ghana, from the MalariaGEN Pf7 resource (bottom) (N samples=1309, N variants=10627).

to as *Pfsa2+*), previously discovered to be associated to HbS in Kenya but shown to be absent in West Africa (130; 131), was also absent in the GAMCC (AF=0%).

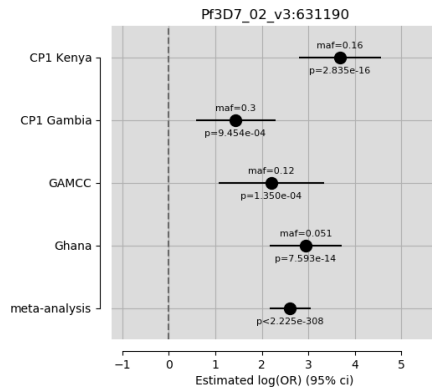
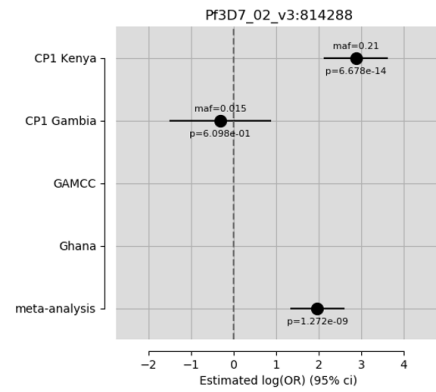
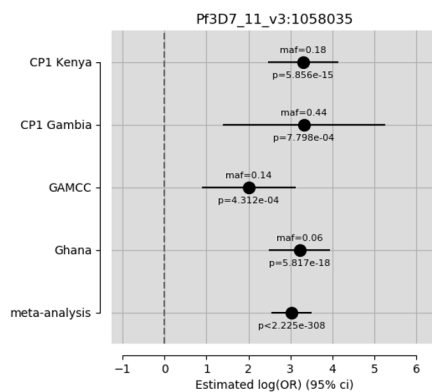
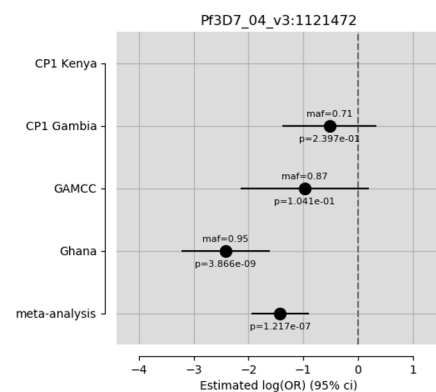
(a) *Pfsa1*(b) *Pfsa2*(c) *Pfsa3*(d) *Pfsa4*

Figure 5.2: **Estimated effect sizes of associations between *Pfsa1-4* and HbS within the GAMCC, CP1 Gambia, CP1 Kenya and Ghanaian datasets.** Replication of evidence across the four studies (indicated on the y-axis), indicating the effect size as a log(odds ratio (OR)) and showing 95% confidence interval (CI) as lines. Two-sided P-values for the logistic regression and the allele frequency of the tested allele (“maf”) are shown with each point. The fixed-effect meta-analysis effect size is indicated as the bottom row. If a dataset did not contain the variant due to low allele frequency, the effect size is not shown.

5.2.1.2 A putative *Pfsa* locus discovered in the GAMCC dataset

In addition to the above, inspection of the results indicated a further association with $\text{BF} > 4000$ at a locus which had not been identified in previous literature (Pf3D7_02_v3:825357:T>C, $\log_{10}\text{BF}=3.63$, $\text{OR}=10.96$ (3.71-31.92), $P=1.12 \times 10^{-5}$, $N=747$, $\text{AF}=11.2\%$) (Fig. 5.1 (pg.153)). This locus was $\sim 11\text{Kb}$ from the *Pfsa2*+ polymorphism (Pf3D7_02_v3:2:814288:C>T) discovered in the original study to be associated with HbS in Kenya. This is surprising because it indicated that a further genomic region, which I will hereafter refer to as *Pfsa2B*, may play a role in the ability of parasites to infect HbS-carrying individuals. I therefore set out to probe this association further.

It is possible that the newly identified *Pfsa2B* locus is physically linked to *Pfsa1*+, also on Pf3D7_02_v3, which may drive the observed association at *Pfsa2B*. However, although correlation between *Pfsa1* and *Pfsa2B* was elevated ($D'=0.467$ (0.344-0.605)), correlation between *Pfsa1* and an intermediary variant (Pf3D7_02_v3:749523:A>C, $\text{AF}=5.7\%$) was reduced ($D'=0.134$ (0.017-0.271)) suggesting that the observed linkage disequilibrium (LD) is not due to a lack of recombination between the *Pfsa1* and *Pfsa2B* regions (D' was used to report these correlations because it depends on the presence of specific combinations of alleles, as opposed to Pearson's R which calculates the linear correlation between genotypes. D' is therefore more indicative of the frequency of recombination (194; 195)). Furthermore, it has been shown that there is a tendency for LD to decay in *Pf* genomes beyond $\sim 500\text{bp}$ physical distance (167), so it is unlikely that the association of this locus is driven by proximity to *Pfsa1*.

To test this further, I reperformed the association conditioned on the genotype at *Pfsa1*, and found that evidence of effect at *Pfsa2B* on HbS, although diminished, remained ($\log_{10}\text{BF}=1.46$, $\text{OR}=4.64$ (1.41-15.24), $P=1.16 \times 10^{-2}$, $N=747$), further suggesting this effect is somewhat independent of the previously discovered *Pfsa1* locus (Supp. Fig. 5.25 (pg.208)).

Given that the variant calls are based on sequence data, I also considered the possibility that the presence of *Pfsa2B* in the dataset may be due to a genotype calling artifact. I therefore inspected this locus in available whole genome assemblies (from (103)) which have undergone multiple sequence alignment (the gener-

ation of which is described in Materials and Methods section 2.5.4.1 (pg.51)). I found this variant was present in a clinical isolate named SN01 (from Senegal) and the laboratory isolate HB3 (103) (Supp. Fig. 5.26a (pg.209)). Furthermore, short-read data in the GAMCC set aligned across the *Pfsa2B* site with high mapping quality scores (Supp. Fig. 5.26b). This suggested this variant was unlikely to be an artefact of inaccurate mapping and was likely to represent a real polymorphism within *Pf* populations.

5.2.2 Joint analysis of HbS associations using cross-population analysis

The most effective way to elucidate *Pfsa*-HbS interactions, and validate the observed effect at the *Pfsa2B* locus, will be to use all available data. Furthermore, the discovery of the *Pfsa1-3* loci revealed that the *Pfsa2+* alleles were only present in East Africa, and a fourth associated locus (*Pfsa4*) discovered in Ghana shows the converse pattern and is absent in East Africa but present in West Africa (131). Taken together with the new indication of an association at *Pfsa2B* described above, this suggests that *Pfsa* relationships may extend to population or region-specific patterns.

Motivated by this I re-analysed other available datasets, investigating *Pfsa* signals within and across populations (Fig. 5.1 (pg.153)). Specifically, I performed association analyses using HPTTEST on three previously published datasets; the MalariaGEN CP1 samples from Gambia and Kenya collected from severe cases in infants in which the original discovery of *Pfsa1-3* was made (130), and a set of mild malaria infections collected in Ghana in which the association at *Pfsa4* was revealed (131). To perform this scan in CP1 samples, I used variant calls within a combined call set generated for the GAMCC and CP1 sets (referred to as the GAMCC/CP1 call set, described in full in Chapter 4 (pg.110)). To analyse the Ghanaian population, I made use of a pre-filtered dataset made available by the original authors (described in Materials and Methods section 2.7.5 (pg.65)). To compare signals across populations I performed a fixed effect meta-analysis, yielding a P-value.

5.2.2.1 Assessment of the *Pfsa2B* association across datasets

To further scrutinise the association at the *Pfsa2B* locus, I sought to discover whether the signal was replicated within the additional datasets. To display evidence at the *Pfsa2B* locus, as well as other loci described later in this chapter, I have developed a visualisation to compare regional HbS-association signals (Fig. 5.3b) and the estimated effects (Fig. 5.3a) across datasets. Figure 5.3a indicates that an effect at this locus is supported by the Ghanaian dataset ($\log_{10}\text{BF}=2.49$, $\text{OR}=6.68$ (2.61-17.17), $P=7.52\times 10^{-5}$, $N=1254$, $\text{AF}=4.3\%$) and Gambian samples in the CP1 dataset ($\log_{10}\text{BF}=1.08$, $\text{OR}=2.86$ (1.16-7.04), $P=2.20\times 10^{-2}$, $N=1972$, $\text{AF}=25.6\%$). Effect direction and magnitude appears to be consistent between these populations (Fig 5.3a). However, evidence of an effect at the *Pfsa2B* locus was not observed in Kenya, the only available East African population, despite being present within this population ($\log_{10}\text{BF}=0.23$, $\text{OR}=1.42$ (0.50-3.98), $P=0.51$, $N=1799$, $\text{AF}=8.4\%$).

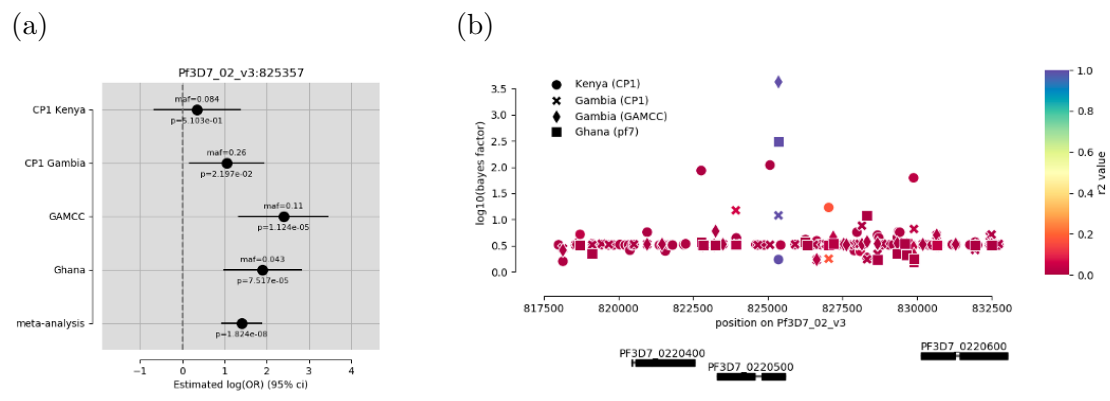


Figure 5.3: Comparison of evidence for the association of *Pfsa2B* (Pf3D7_02_v3:825357:T>C) to HbS. (a) Estimated $\log_{10}(\text{OR})$ and 95% confidence intervals (CIs) for the association of *Pfsa2B* with HbS within the GAMCC, CP1 Gambia, CP1 Kenya and Ghanaian datasets (y-axis). Allele frequencies and P-values are indicated on the OR values. (b) A detailed view of the association at *Pfsa2B* within the two scans for association with HbS; GAMCC (diamonds), severe CP1 cases from Gambia (crosses) and Kenya (circles) and mild cases from Ghana (squares). Pearson's R-squared values between the central variant and other local variants are indicated by the colour of the marker, as described by the colour bar (right). At the bottom the positions of genes for the region of the genome are shown. Exons within genes are indicated as thicker black bars.

Given that other *Pfsa* alleles have been observed to vary dramatically across global populations, I further analysed allele frequencies of the *Pfsa2B*+ polymorphism in the MalariaGEN Pf7 resource of *Pf* WGS data sampled from global populations (37). To avoid excessive sampling noise, I restricted the analysis to samples which are denoted as QC pass within the release, and to countries where at least 50 genomes were available. The frequencies of *Pfsa2B*+ vary drastically across Africa (Fig. 5.4e on the next page), varying in a similar way to *Pfsa1*+ (Fig. 5.4a) and *Pfsa3*+ alleles (Fig. 5.4c). However, unlike the *Pfsa1*+ and *Pfsa3*+ alleles, there is nontrivial frequency of this polymorphism in regions of Asia, reaching 5.2% in samples from Eastern regions of India in this dataset (N=484).

Collectively, the evidence strongly supports an association between the novel *Pfsa2B* locus and the HbS allele (meta-analysis P-value^{*allpopulations*} $< 2 \times 10^{-8}$). While the previously discovered *Pfsa2* allele is absent in West African populations, the nearby *Pfsa2B* signal (~ 11 Kb away) has been replicated in other West African populations but not in Kenya, an East African population (Fig. 5.3 on the preceding page).

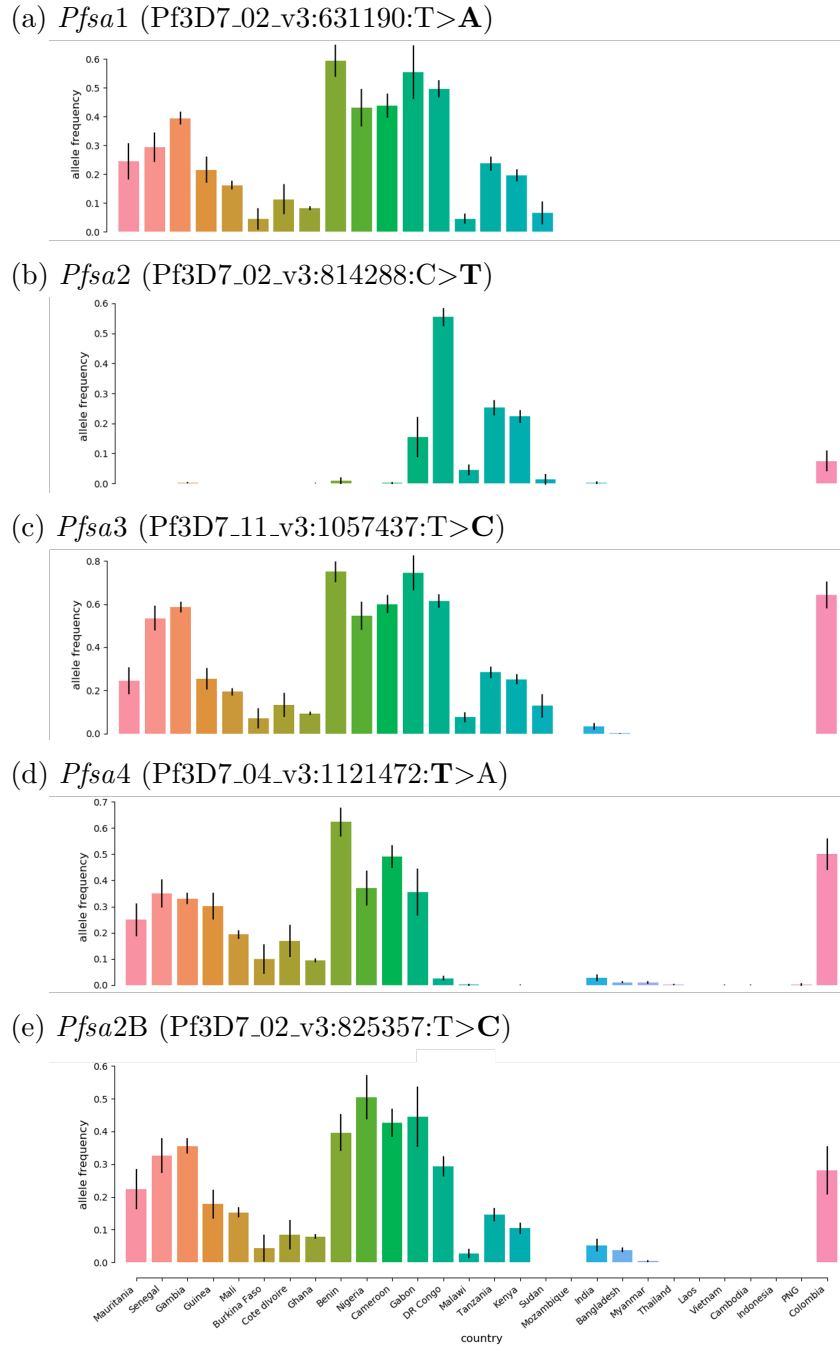


Figure 5.4: **Allele frequencies of *Pfsa* variants within populations in the MalariaGEN Pf7 resource.** Allele frequency of the effect alleles at the *Pfsa1-4* and putative *Pfsa2B* loci within different populations in the MalariaGEN Pf7 resource. Only populations with more than 50 samples are shown. Countries are shown in geographical order roughly from West to East. 95% CI error bars are show. Pf3D7_11_v3:1057437:T>C was used instead of the identified most-associated variant at *Pfsa3* (Pf3D7_11_v3:1058035:T>A) as the latter was not present in the Pf7 dataset. Within the caption for each figure, the effect allele is highlighted in bold.

5.2.2.2 Further population specific *Pfsa* relationships are found in public datasets

The original discovery of the *Pfsa*1-3 loci used a meta-analysis approach across the Gambian and Kenyan populations within the CP1 dataset but did not investigate population specific signals. The individual analysis of each population performed within this chapter implicated three new loci with strong evidence ($BF > 100000$) and one with marginal evidence ($BF \sim 8000$) of association to HbS outlined in Supp. Table 5.5 (pg.221) and described further below.

A Gambia-specific association on Pf3D7_07_v3 in CP1 data A strong association was observed in the CP1 Gambian dataset for a locus on parasite chromosome 7 (Pf3D7_07_v3:614420:C>T, $\log_{10}BF=5.36$, $OR=10.50$ (3.6-30.61), $P=1.63 \times 10^{-5}$, $N=2058$, $AF=29.4\%$) (Fig. 5.1 (pg.153)). However, no strong evidence of association was apparent in the Kenyan ($\log_{10}BF=0.44$, $OR=1.82$, $P=1.19 \times 10^{-5}$, $N=1648$, $AF=18.6\%$) or Ghanaian ($\log_{10}BF=0.37$, $OR=1.35$ (0.35-5.27), $P=6.64 \times 10^{-1}$, $N=1239$, $AF=2.0\%$) populations, nor in samples from the GAMCC resource ($\log_{10}BF=0.53$, $OR=0.54$ (0.13-2.23), $P=3.96 \times 10^{-1}$, $N=780$, $AF=5.3\%$), with inconsistent directions of effect estimated between these datasets (Fig. 5.5 on the next page).

An assessment of alignments at this locus suggests that this variant is well called with almost all reads mapping with a mapping quality score (MAPQ)>60 (i.e. the estimated probability that the read is misaligned at < 1 in 1,000,000) within alignments of GAMCC data (Supp. Fig. 5.27a (pg.210)). The variant also existed within available long read assemblies (this variant was present within the laboratory clones IT, 7G8 and Dd2, as well as a clinical isolate called KE01 from Kenya (103)) and did not appear within a highly repetitive region (Supp. Fig. 5.27b (pg.210)). Corroborating this, in the GAMCC/CP1 variant call set, the coverage of this variant appeared within the range of depth observed genome-wide (Supp. Fig. 5.27a (pg.210)). Together, this suggested this variant was well genotyped and unlikely to be caused by a sequencing artefact. It is worth noting that upstream of this variant appears to be dense with variation (Fig. 5.5b on the next page), and this region contains a pseudogene of the VAR family, and a

member of the RIFIN family (which falls just outside of Fig. 5.5b), which are known to be highly polymorphic (129).

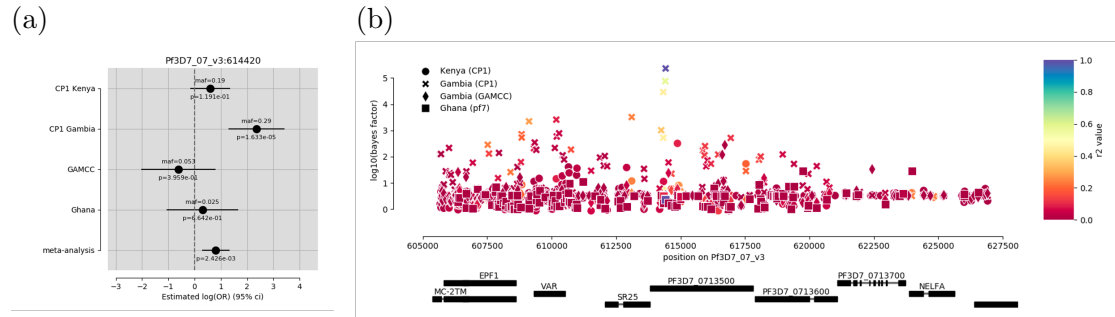


Figure 5.5: Comparison of evidence for the association of *Pf3D7_07_v3:614420:C>T* to HbS. (a) Estimated log₁₀(OR) and 95% CIs for the association of *Pfsa2B* with HbS within the GAMCC, CP1 Gambia, CP1 Kenya and Ghanaian datasets (y-axis). Allele frequencies and P-values are indicated on the OR values. (b). a detailed view of the association at *Pf3D7_07_v3:614420:C>T* within the two scans for association with HbS; GAMCC (diamonds), severe CP1 cases from Gambia (crosses) and Kenya (circles) and mild cases from Ghana (squares). Pearson's R-squared values between the central variant and other local variants are indicated by the colour of the marker, as described by the colour bar (right). At the bottom the positions of genes for the region of the genome are shown. Exons within genes are indicated as thicker black bars.

The location of this association on parasite chromosome *Pf3D7_07_v3* raises the possibility that it has been influenced by chloroquine (CQ) resistance within the chloroquine resistance transporter (CRT) gene. CRT mutations are known to have undergone a strong selective sweep (179), and I observed a rapid increase in AF of a variant identified to be linked to a CQ-resistant haplotype (*Pf3D7_07_v3:403625:A>C*) within the timeframe of these samples (Fig. 4.5 (pg.121) in Chapter 4). Although the discussed HbS-associated locus occurs >200Kb away from the resistance mutation, selective sweeps can generate long-range LD, potentially driving the observed association. Further contextual evidence suggests this may have been the case. The AF of *Pf3D7_07_v3:614420:C>T* within the GAMCC samples (collected 1988-90, AF=5.3%), which were sampled prior to the spread of CQ resistance in Gambia, was much less than in the Gambian samples in the CP1 dataset (collected 1996-2009) (AF=29%), hinting that this variant and CQ resistance within CRT may have become correlated. Moreover, the CRT locus itself

was found to be marginally associated with HbS within the Gambian CP1 samples (at Pf3D7_07_v3:403625:A>C $\log_{10}\text{BF}=2.25$, $\text{OR}=8.58$ (1.41-52.19), $P=1.96 \times 10^{-2}$, $N=2104$, $\text{AF}=72.8\%$).

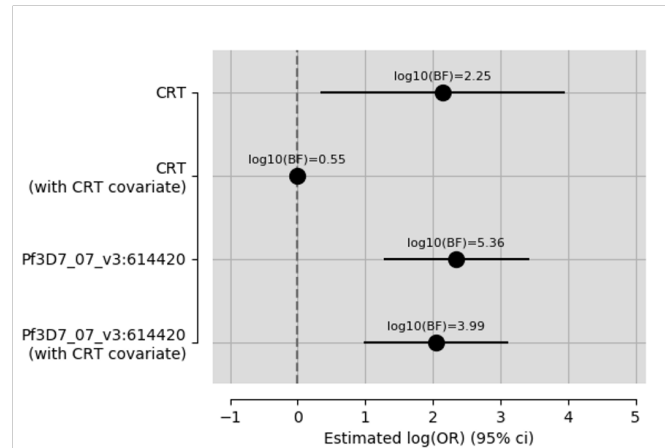


Figure 5.6: **Estimated effect sizes of associations between Pf3D7_07_v3:614420:C>T and HbS within samples from Gambia in the CP1 dataset conditioned on genotype at the CQ resistance locus CRT (Pf3D7_07_v3:403625:A>C).** Effect sizes are shown as a $\log(\text{odds ratio (OR)})$ with 95% confidence interval (CI)s. Strength of evidence ($\log_{10}\text{BF}$) are shown at each point.

To investigate whether the spread of CQ resistance might underlie the observed association of Pf3D7_07_v3:614420:C>T to HbS, I examined LD between these loci and found that moderate LD existed ($R=0.295$ (0.256-0.332)). To test whether this affected the reported results, I reperformed the GWAS scan within the CP1 Gambian cohort conditioning on the parasite genotype at the CRT locus. Contrary to our predictions, this did not remove the association at Pf3D7_07_v3:614420:C>T but did remove it at the CRT locus (Fig. 5.6), suggesting that the sweep of CQ resistance does not explain the association. Until further replication evidence is obtained, the validity of this association remains uncertain. Nevertheless, the potential linkage of this locus to HbS has not been conclusively ruled out and warrants further investigation.

A potential cross-population association at the cytoadherence linked asexual gene (CLAG)3.1-3.2 genes As described within the Introduction Chapter (Section 1.6 (pg.29)), elevated LD between *Pfsa* loci is a feature of

HbS-*Pfsa* interactions (this will be revisited later within this chapter) (130; 131). Within the previously published discovery analysis of *Pfsa*1-3, elevated levels of LD between the *Pfsa* loci and variation within each of the CLAG3.1 and CLAG3.2 genes, which sit ~ 9.5 Kb apart on chromosome Pf3D7_03.v3, was reported within the Kenyan population, but no association at this region to HbS was described (130). In my analysis of the same data, I found evidence of an association within the CLAG3.1 gene at Pf3D7_03.v3:140060:T>A within Kenya ($\log_{10}\text{BF}=5.22$, $\text{OR}=0.14$ (0.07-0.30), $P=3.85 \times 10^{-7}$, $N=1368$, $\text{AF}=75.8\%$) (Fig. 5.1 (pg.153)). This may not have been assessed within the discovery study because this site and other nearby HbS-associated variants had high levels of heterozygous calls. As common practice, variation within the *Pf* genome is called as diploid to account for levels of mixture present within natural infections (37). Consequently, loci of mixed allele calls are called as heterozygous. Because this is not representative of individual parasite genomes these sites are usually excluded from GWAS analysis of parasite genomes, and are not included within computations performed by HPTEST. Variants with low levels of missing and heterozygous calls ($<10\%$ within each dataset) are indicated on Figure 5.1 as solid points, whereas those with high heterozygosity are transparent.

Inspecting this association across populations, estimated effects of Pf3D7_03.v3:140060:T>A occurred in the same direction to within Kenya in the CP1 Gambian ($\log_{10}\text{BF}=0.83$, $\text{OR}=0.38$ (0.14-1.01), $P=5.29 \times 10^{-2}$, $N=1656$, $\text{AF}=66.9\%$) and GAMCC datasets ($\log_{10}\text{BF}=0.84$, $\text{OR}=0.31$ (0.09-1.08), $P=6.62 \times 10^{-2}$, $N=482$, $\text{AF}=72.3\%$) (Fig. 5.9a (pg.167)). This variant was not present within the Ghanaian dataset so was not initially tested for replication. These results are compatible with a genuine association between variation at this locus and HbS, although strength of evidence differs between these populations (meta-analysis $P\text{-value}^{\text{GAMCC,CP1Kenya,CP1Gambia}} < 5 \times 10^{-8}$). I therefore set out to determine what might underlie the high levels of heterozygosity at this site with a view to performing a more rigorous analysis of the association of this locus to the HbS polymorphism.

Variants across the CLAG3.1 gene region showed high levels of heterozygosity across the four tested datasets (e.g. 14% of samples in GAMCC, 22% and 23% in CP1 Gambia and Kenya were heterozygous at the Pf3D7_03.v3:140060:T>A

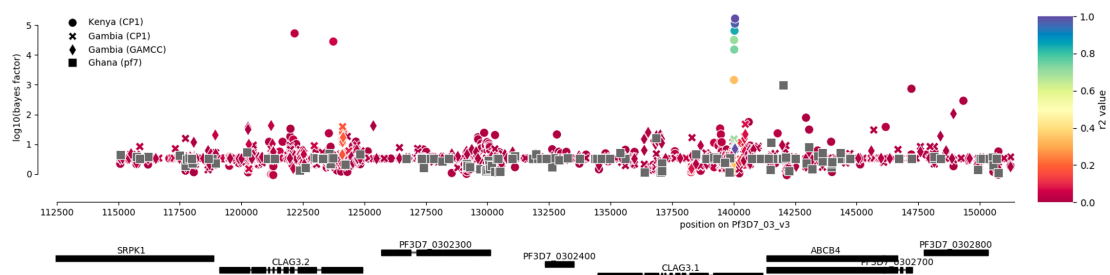


Figure 5.7: Comparison of evidence for the association of Pf3D7_03.v3:140060:T>A to HbS. A detailed view of the association at Pf3D7_03.v3:140060:T>A within the two scans for association with HbS; GAMCC (diamonds), severe CP1 cases from Gambia (crosses) and Kenya (circles) and mild cases from Ghana (squares). Pearson's R-squared values between the central variant and other local variants are indicated by the colour of the marker, as described by the colour bar (right). At the bottom the positions of genes for the region of the genome are shown. Exons within genes are indicated as thicker black bars.

variant). This may be because CLAG3.1 shares much of its sequence identity with CLAG3.2 (196). If nucleotides differ between paralogous positions in CLAG3.1 and CLAG3.2, known as a paralogous sequence variants (PSVs), sequenced reads containing both alleles may align to the same copy of the gene within the reference assembly causing a heterozygous call. To investigate this, I located the position of the paralogous site of Pf3D7_03.v3:140060:T>A variant (within CLAG3.1) within the CLAG3.2 gene (using the software BLAST) in a multiple sequence alignment of whole genome assemblies. Inspecting both of these sites (Pf3D7_03.v3:140060:T>A and Pf3D7_03.v3:124135:T>A in Pf3D7 reference co-ordinates), indicated that these positions, and many other nearby sites, appeared to be PSVs. For example, within the genome assemblies for the laboratory isolate HB3, and the clinical isolates GN01 and GA01, the Pf3D7_03.v3:124135 site carries the A allele and Pf3D7_03.v3:140060 carries T (Supp. Fig. 5.28 (pg.211)).

MAPQ scores across this region were poor, as would be expected when reads are able to align to each copy of CLAG3 (Supp. Fig. 5.29a (pg.212)). In fact, within the Pf7 call set this variant was not annotated as of high-quality based on annotations by GATK's VQSR tool and was excluded from the Ghanaian analysis. Although VQSR was also performed on the GAMCC/CP1 call set, this variant was marked as pass and was retained.

A hybrid of CLAG3.1/2, called CLAG3h, has also been previously described and we observed this within long-read genome assemblies (166) (Supp. Fig. 5.30 (pg.213)). Because computations by HPTEST were limited to homozygous variation, the sample across which the association analysis was performed was potentially biased towards samples carrying the CLAG3h structure (who should not contain PSV sites). It is possible therefore that structural variation might underlie the observed association of this locus to HbS.

In order to assess this locus, I performed a number of additional analyses (a schematic representation of these is provided in Figure 5.8). First, I made a basic call of structure using high levels of missingness across the deletion as an indication that the hybrid CLAG3h form was present in the sample (described in Materials and Methods section 2.10.1 (pg.78)). This deletion varied in frequency between datasets ($AF_{CP1}^{Kenya}=18.4\%$ (N=1770), $AF_{CP1}^{Gambia}=14.4\%$ (N=2143), $AF^{GAMCC}=27.1\%$ (N=726), $AF^{Ghana}=48.9\%$ (N=990)). Based on this preliminary investigation, carrying CLAG3h did not appear to associate strongly to HbS in the datasets (Fig. 5.9b on the following page). This suggested that SNP variation instead might drive the association of this region to HbS.

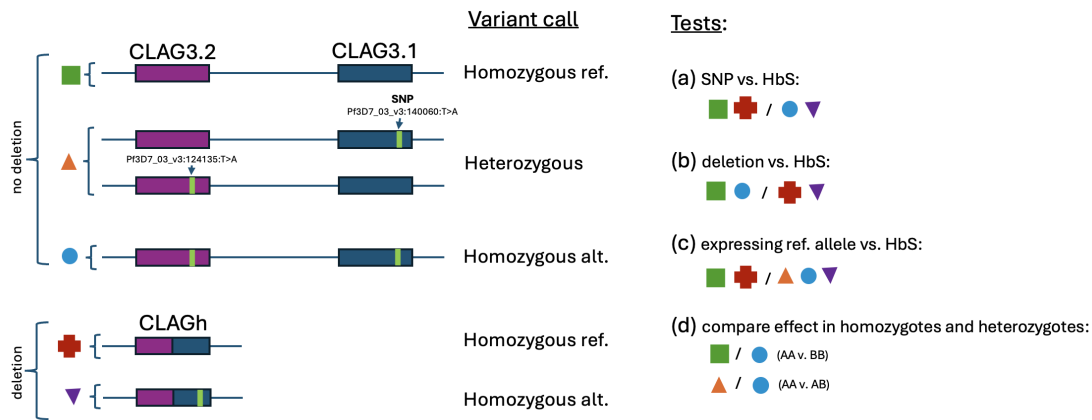


Figure 5.8: **An explanation of conducted analyses into associations between CLAG3 and HbS.** An infographic explaining analyses which produced results in Figure 5.9.

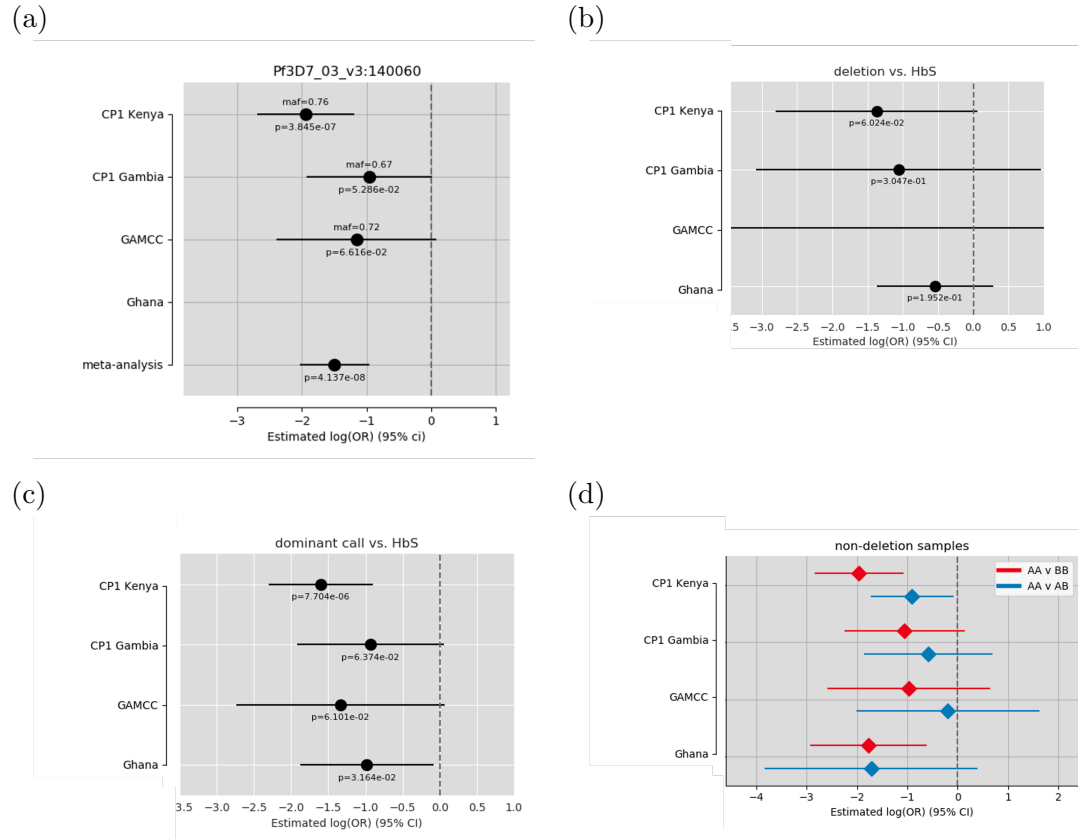


Figure 5.9: Associations between Pf3D7_03_v3:140060:T>A and HbS within the GAMCC, CP1 Gambia, CP1 Kenya and Ghanaian datasets with and without allowing for structural variation. (a) Replication of evidence across the four studies (indicated on the y-axis), indicating the effect size as a log(odds ratio (OR)) and showing 95% confidence interval (CI)s as lines. Two-sided P-values for the logistic regression and the allele frequency of the tested allele (maf) are shown with each point. The fixed-effect meta-analysis effect size is indicated as the bottom row. (b)-(d) Results of logistic regressions against HbS status allowing for CLAG3 deletion polymorphism as follows: (b) association between HbS and CLAG3h hybrid gene status, (c) association between Pf3D7_03_v3:140060 and HbS, after treating homozygous alternate and heterozygous calls as one group (in comparison to (a) where heterozygous calls are excluded) and (d) in samples called as having both copies of CLAG3.1 and CLAG3.2, association between heterozygous (blue; 'AA v AB') or homozygous (red; 'AA v BB') alternate allele genotypes and HbS, compared to homozygous reference genotypes. The x-axis shows the estimated log odds ratio and 95% confidence interval error bars.

It has been observed that CLAG3.1 and CLAG3.2 are generally expressed in mutually exclusive manner where one copy is epigenetically silenced (196). To attempt to recover statistical power in the assessment of SNP variation, I per-

formed a supplementary regression comparing samples carrying homozygous reference alleles (which was the identified effect allele), who must express this allele, to parasites carrying heterozygous and homozygous alternate alleles, of which an unknown proportion of heterozygotes will express this allele (test (c) in Fig. 5.8 (pg.166)). With these adjustments, evidence of association persisted across the studied populations (Fig. 5.9c on the preceding page).

If SNP variation underlies the observed association, we might hypothesise that within samples that carry both copies of *CLAG3*, the association signal would be weaker in heterozygous samples (test (d) in Fig. 5.8 (pg.166)). This is because only a proportion of these will be expressing the gene carrying the effect allele due to the mutually exclusive expression between *CLAG3.1* and *3.2*. This appears to be the case in samples we predict to carry both copies (i.e. those that do not carry the deletion) where a comparison of heterozygous-to-homozygous reference samples produced a weaker effect compared to homozygous reference-to-homozygous alternate samples in all tested populations apart from within Ghana (Fig. 5.9d on the preceding page). It is worth noting that sample sizes were significantly reduced by stratifying and performing the deletion call, and so results should be interpreted with caution. Future analysis of this locus poses potential challenges due to difficulties mapping to a structurally variable region and the presence of significant sequence homology. Resolving associations here may therefore require more precise custom methods which require development.

Compelling evidence supports the connection of this locus to HbS. As reported by the discovery study, the LD between *Pfsa1* with *CLAG3.1* and *CLAG3.2* ranked 4th and 5th strongest across the Kenyan genome in the CP1 dataset. LD between *Pfsa3* with *CLAG3.1* and *CLAG3.2* ranked 6th and 10th, and LD between *Pfsa2* and *CLAG3.1* ranked 12th. LD was also observed to a lesser extent in Gambia, where LD between *Pfsa3* and *CLAG3.2* ranked 58th genome-wide (130). In accordance with this, the lead variant here shows elevated levels of LD to *Pfsa3* in many countries within the Pf7 dataset including Mauritania ($R=0.472$ (0.317-0.664)), Guinea ($R=0.341$ (0.204-0.493)), Mali ($R=0.292$ (0.232-0.354)) and Cote D'Ivoire ($R=0.303$ (0.177-0.585)) (Fig. 5.10b (pg.170)). The preliminary analysis presented here provides suggestive evidence for the association, but a comprehensive consolidation of nucleotide and structural variation at this locus will be

essential to solidify confidence in the association and fully understand its potential connection to HbS.

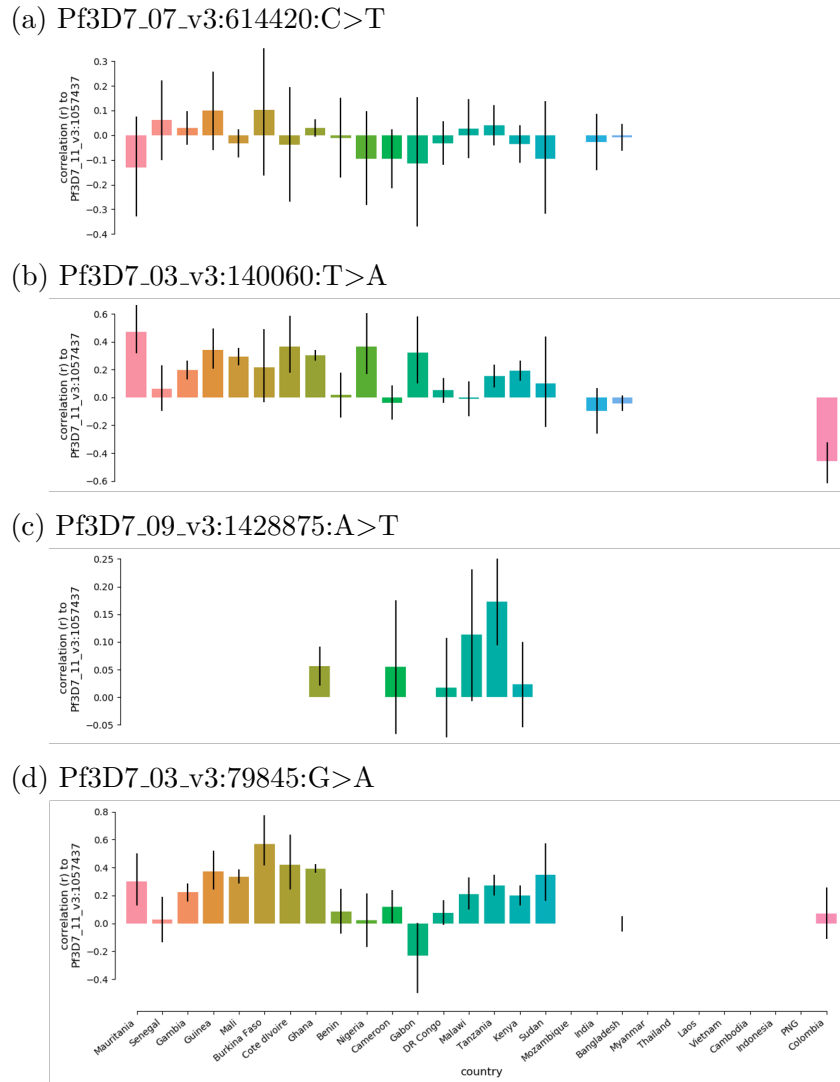


Figure 5.10: **Correlation between potential *Pfsa* loci and *Pfsa3* within populations in the MalariaGEN Pf7 resource.** Only populations with more than 50 samples are shown. Countries are shown in geographical order roughly from West to East. Pearson's R value between the tested variant and a strongly HbS-associated variant at the *Pfsa3* locus (Pf3D7_11_v3:1057437:T>C) for populations in the MalariaGEN pf7 resource. This *Pfsa3* variant was chosen because it appears associated to HbS in all the so far tested populations and is present across all the tested call sets. 95% CIs calculated using a Fisher transformation are shown.

A Kenya-specific association at ring exported protein (REX)2 My analysis revealed a second previously undescribed association in the CP1 Kenyan dataset, at Pf3D7_09_v3:1428875:A>T ($\log_{10}\text{BF}=5.25$, $\text{OR}=10.09$ (4.66-21.84), $P=4.51\times 10^{-9}$, $N=1828$, $\text{AF}=6.1\%$) (Fig. 5.1 (pg.153)). This variant fell within the 5' untranslated region (UTR) of the gene REX2 (PF3D7_0936000) and downstream of transcribed membrane protein 9 (ETRAPM9, PF3D7_0936100) (Fig. 5.11b on the next page). This variant was confidently called in the present dataset (with high MAPQ scores and a reasonable depth profile, Supp. Fig. 5.31 (pg.214)) but it was not found in West African regions (Supp. Fig. 5.32c (pg.215)) and therefore no comparison of evidence for this association could be provided by Gambian or Ghanaian genomes (Fig. 5.11b on the following page). Where it was present there was limited evidence of LD of this region to known *Pfsa* loci, for example in Tanzania correlation of this locus to *Pfsa3* was only $R=0.173$ (0.094-0.251). REX1, the adjacent gene to REX2 just 1311bp downstream (Fig. 5.11b on the next page), is found to be in prominent between chromosome LD with *Pfsa1*, *Pfsa2* and *Pfsa3* in the Kenyan samples in the CP1 dataset (ranking 8th, 13th and 7th of all genome wide trans-chromosomal LD values in Kenya respectively) (130).

REX1 and REX2 are PEXEL-negative exported proteins which, along with other proteins of this type, are highly expressed shortly after cell invasion, and so are thought to play a role in host cell remodelling (197). More data sets from East Africa are required to validate this locus as a potential *Pfsa* region.

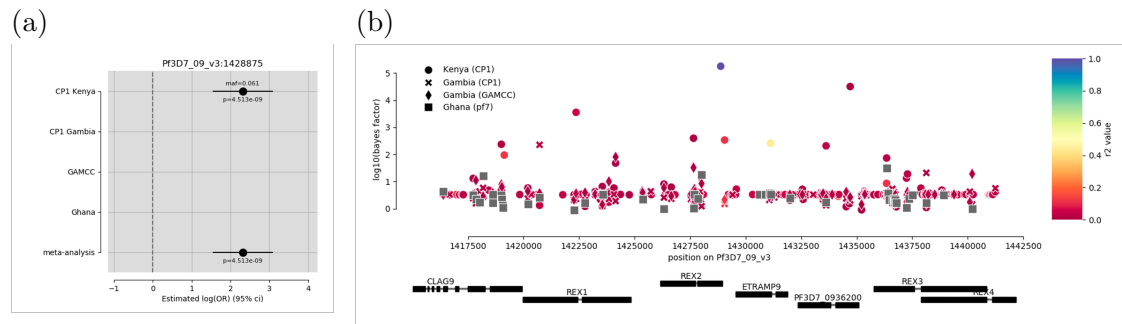


Figure 5.11: **Comparison of evidence for an association of Pf3D7_09_v3:1428875:A>T to HbS.** (a) Estimated $\log_{10}(\text{OR})$ and 95% CIs for the association of *Pfsa2B* with HbS within the GAMCC, CP1 Gambia, CP1 Kenya and Ghanaian datasets (y-axis). Allele frequencies and P-values are indicated on the OR values. (b) A detailed view of the association at Pf3D7_09_v3:1428875:A>T within the two scans for association with HbS; GAMCC (diamonds), severe CP1 cases from Gambia (crosses) and Kenya (circles) and mild cases from Ghana (squares). Pearson's R-squared values between the central variant and other local variants are indicated by the colour of the marker, as described by the colour bar (right). At the bottom the positions of genes for the region of the genome are shown. Exons within genes are indicated as thicker black bars.

A marginal association in Ghana at serine/threonine protein kinase (FIKK)3 In the Ghanaian dataset, a marginal association was observed on chromosome 3 (Pf3D7_03_v3:79845:G>A), ($\log_{10}\text{BF}=3.90$, $\text{OR}=0.09$ (0.04-0.23), $P=4.7 \times 10^{-7}$, $N=1263$, $\text{AF}=96.8\%$) (Fig. 5.1 (pg.153)) which also showed evidence of association to HbS in the GAMCC ($\log_{10}\text{BF}=1.04$, $\text{OR}=0.23$ (0.07-0.81), $P=2.19 \times 10^{-2}$, $N=756$, $\text{AF}=92.1\%$) and Kenyan ($\log_{10}\text{BF}=1.95$, $\text{OR}=0.17$ (0.06-0.46), $P=4.52 \times 10^{-4}$, $N=1838$, $\text{AF}=95.2\%$) datasets, but not Gambian genomes in the CP1 resource ($\log_{10}\text{BF}=0.22$, $\text{OR}=1.09$ (0.35-3.41), $P=8.83 \times 10^{-1}$, $N=2096$, $\text{AF}=86.3\%$) (Fig. 5.12a on the next page). This suggested this locus may represent a further cross-population *Pfsa* locus.

This variant fell in the gene *FIKK3*, another member of the same serine/threonine kinase gene family as *FIKK4.2*, the gene present at the *Pfsa4* locus (which was discovered using the same cohort) (Fig. 5.12b on the following page). The effect of *Pfsa4* was also not found to replicate in Gambian CP1 data (131). The frequency of this variant appears to covary across Africa in a manner similar to *Pfsa1* and *Pfsa3* (Fig. 5.4 (pg.160) & Supp. Fig. 5.32 (pg.215)) and showed levels of

elevated LD in some populations in the Pf7 resource, particularly in West Africa (e.g. in Burkina Faso correlation to *Pfsa3* was $R=0.569$ (0.416-0.784), Fig. 5.10d (pg.170)). The variant was likely to be real as it existed in long read assemblies available for *Pf* (Supp. Fig. 5.33b (pg.216)) and reads have mapped to this locus in the GAMCC with high MAPQ scores, indicating it was likely to be confidently called within datasets (Supp. Fig. 5.33a (pg.216)). The association of FIKK4.2 and the marginal association presented here implicate this gene family as possibly playing a role in infecting to HbS individuals, but further evidence is required to understand their possible function.

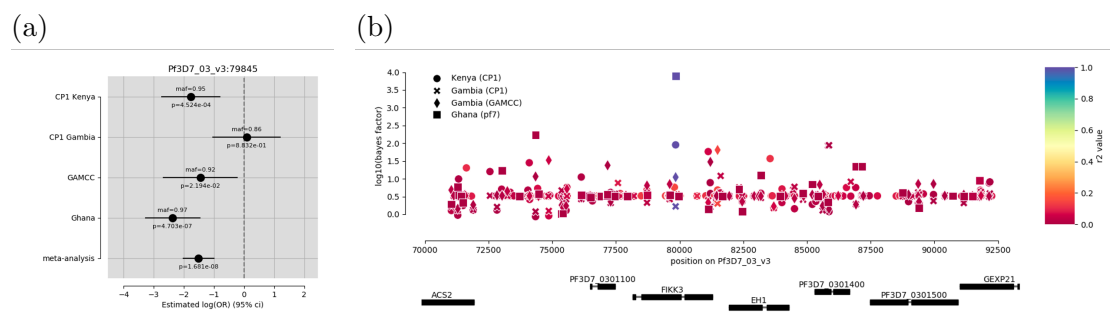


Figure 5.12: Comparison of evidence for the association of Pf3D7_03_v3:79845:G>A to HbS. (a) Estimated log₁₀(OR) and 95% CIs for the association of *Pfsa2B* with HbS within the GAMCC, CP1 Gambia, CP1 Kenya and Ghanaian datasets (y-axis). Allele frequencies and P-values are indicated on the OR values. (b). a detailed view of the association at Pf3D7_03_v3:79845:G>A within the two scans for association with HbS; GAMCC (diamonds), severe CP1 cases from Gambia (crosses) and Kenya (circles) and mild cases from Ghana (squares). Pearson's R-squared values between the central variant and other local variants are indicated by the colour of the marker, as described by the colour bar (right). At the bottom the positions of genes for the region of the genome are shown. Exons within genes are indicated as thicker black bars.

5.2.2.3 A re-evaluation of how HbS protection varies with parasite genotype

Within the publication describing the discovery of the *Pfsa1-3* loci, Band et al. (130) estimated that HbS genotypes may confer much less protection against infections caused by malaria parasites carrying all three *Pfsa1-3+* alleles. This was based on 49 severe malaria cases in individuals carrying at least one HbS allele,

after excluding samples with mixed genotypes at the three *Pfsa* loci. Intermediate levels of HbS-protection were observed where parasites carried a subset of the *Pfsa* alleles, however these were based on very small sample sizes, causing considerable uncertainty around these estimates.

The additional 35 cases procured during the *Pfsa4* discovery study (131), as well as a further 13 cases within the GAMCC resource, means these estimates can be improved. Specifically, in light of the evidence of spatial variation in HbS-*Pfsa* interactions outlined above, and because a large amount of data for West African regions is now available ($N_{HbAS, HbSS}^{W-AFR}=64/4329$ total samples (Table 5.2 on the following page)), the protective effect of HbS dependent on parasite genotype can also begin to be assessed between geographical areas. To estimate the protective effect of HbS considering *Pfsa*, I compared the frequency of HbS genotypes in malaria cases and non-malaria-infected population controls, stratified by parasite genotype in the West and East African populations separately. Estimates were made using multinomial logistic regression as described previously (130). Because some parasite genotypes have small counts in these data, I used a Bayesian formulation with a mild regularising prior, which produces similar estimates to non-Bayesian regression for common genotypes, and plausible estimates for parasite genotypes with small sample sizes (for further detail of this method, and information about population controls, see Materials and Methods section 2.8.1 (pg.66)).

Genotype	Gambia (GAMCC)		Gambia (CP1)		Ghana (pf7)		W-AFR Total	
	AA	AS/ SS	AA	AS/ SS	AA	AS/ SS	AA	AS/ SS
Control	835	152	844	139	850	67	2529	358
- - - -	520	2	610	0	1060	15	2190	17
- - - +	22	1	120	0	17	1	159	2
- - + -	14	0	181	4	8	1	203	5
- - + +	5	0	81	0	2	1	88	1
- + - -	11	0	91	0	23	0	125	0
- + - +	2	0	26	0	0	0	28	0
- + + -	5	0	52	2	0	0	57	2
- + + +	4	0	25	0	1	0	30	0
+ - - -	9	1	56	0	1	0	66	1
+ - - +	0	0	16	0	2	0	18	0
+ - + -	13	0	107	2	9	3	129	5
+ - + +	8	0	82	1	4	2	94	3
+ + - -	4	1	32	0	2	0	38	1
+ + - +	2	0	14	0	0	0	16	0
+ + + -	6	3	89	1	2	3	97	7
+ + + +	8	1	66	2	6	3	80	6

Table 5.2: *Pf* genotype counts at *Pf**sa* loci in HbA vs. HbS individuals in West-African populations. Genotypes at the *Pf**sa*1, *Pf**sa*2B, *Pf**sa*3 and *Pf**sa*4 loci were included in this table. Infections involving the non-reference allele at each site are indicated by a ‘+’ (i.e. “++++” indicates the infecting parasite carried all four *Pf**sa*+ alleles.)

Genotype	Kenya (CP1)	
	AA	AS/ SS
Control	3389	636
- - -	1276	4
- - +	35	1
- + -	55	0
- + +	12	0
+ - -	4	0
+ - +	19	2
+ + -	14	1
+ + +	134	24

Table 5.3: ***Pf* genotype counts at *Pfsa* loci in HbA vs. HbS individuals in Kenya (an East-African population).** Genotypes at the *Pfsa1*, *Pfsa2* and *Pfsa3* loci were included in this table. Infections involving the non-reference allele at each site are indicated by a ‘+’ (i.e. “+++” indicates the infecting parasite carried all three *Pfsa*+ alleles.)

Results of this analysis are shown in Figure 5.13 on the following page. Two differences between the West-African and Kenyan estimates were of note. Firstly, corroborating Band et al. (130), HbS carrying individuals were strongly protected from malaria infections caused by *Pfsa*- parasites in West Africa (only genotypes at the *Pfsa1*, *Pfsa3*, *Pfsa4* and the putative *Pfsa2B* locus described within this chapter were included as *Pfsa2* is absent from Western populations) (RR=0.06 (0.035-0.092)) (Fig. 5.13a). However, protection against infection by *Pfsa*- parasites was stronger in Kenya (here the *Pfsa1*-3 sites were included) (RR=0.02 (0.007-0.043)). Secondly, although the level of protection conferred by HbS was dramatically decreased in the context of infections with parasites carrying the *Pfsa1*+, *Pfsa2B*+, *Pfsa3*+ and *Pfsa4*+ alleles in West Africa, it did not disappear (RR=0.43 (0.166-1.0)). This contrasts with Kenya where the level of protection afforded by HbS when infected by *Pfsa*+ parasites was negligible (RR=0.97 (0.660-1.553)) (Fig. 5.13b).

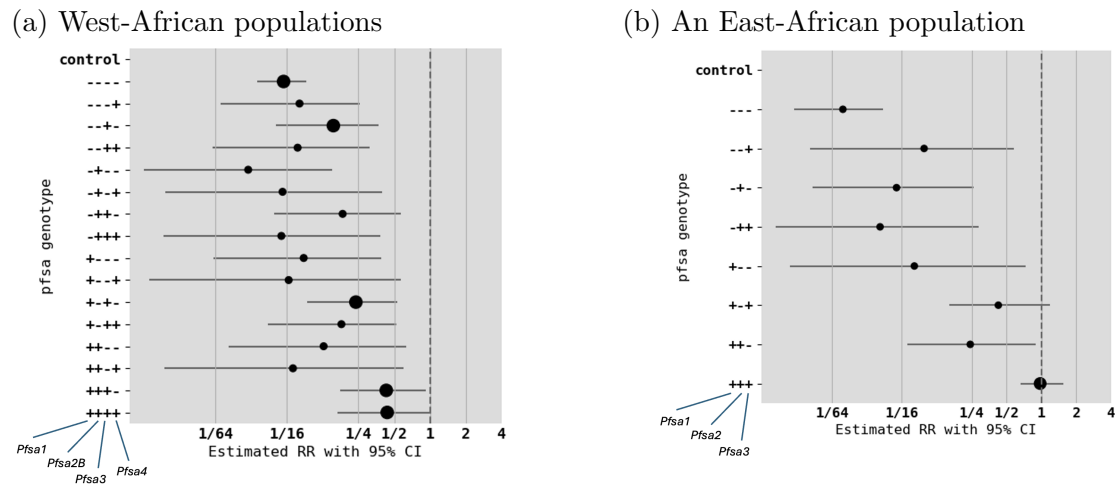


Figure 5.13: **Estimated relative risk of severe malaria for HbS carriers in infections of different *Pf*sa genotypes.** relative risk (RR) is stratified by infecting malaria parasite genotype at the *Pf*sa loci. **(a)** Shows this analysis for populations in Western Africa (GAMCC, Gambian CP1 samples and Ghana) for the *Pf*sa1, *Pf*sa2B, *Pf*sa3 and *Pf*sa4 alleles, and **(b)** shows results in Kenya (the only East African population for which this data exists) for the *Pf*sa1-3 variants. RRs were calculated using a multinomial logistic regression using non-malaria controls as a baseline. To avoid over-fitting of points with small sample sizes, bayesian regularisation was performed using a weakly informative Gaussian prior with mean 0 and standard deviation of 2 for each parameter, and between-parameter correlation set to 0.5. A ‘+’ for the *Pf*sa genotype indicates possessing the effect allele.

Reasons behind this are not immediately obvious. Based on counts of cases of each *Pfsa* genotype stratified by human HbS genotype (shown in Tables 5.2 (pg.175) & 5.3 on the preceding page), the Ghanaian population appeared unusual. In the Ghanaian population the protective effect of HbS against malaria infection was smaller than that of the other populations due to its decreased AF in non-malaria controls ($AF_{HbS}^{Ghana}=7.3\%$, compared to $AF_{HbS}^{Gambia}=14.7\%$ and $AF_{HbS}^{Kenya}=15.8\%$). Furthermore, the Ghanaian dataset contained a relatively high number of HbAS/SS individuals who were infected with parasites that did not carry any of the assessed *Pfsa* alleles (N=15 out of 29 infections of HbS individuals were with *Pfsa*- parasites (5.2)). This contrasted to Kenya and Gambia, where very few infections of HbS carriers do not carry any *Pfsa* alleles ($N_{CP1}^{Kenya}=4/32$, $N_{CP1}^{Gambia}=0/12$, $N_{GAMCC}=2/9$).

Removing Ghana from the analysis of West-Africa and considering the Gam-

bian data in isolation however retained some of the observed differences between the West African populations and Kenya (Supp. Fig. 5.34 (pg.217)). The RR of disease in infections caused by *Pfsa*- parasites became closer to estimates made in Kenya (RR=0.02 (0.004-0.033)), however RR was still not close to 1 in the context of infections caused by *Pfsa*+ parasites (RR=0.23 (0.044-0.84)). This suggested that Ghana is not the sole driver of these observed differences between west and East African populations. It should be noted that these analyses are based on small samples of HbS individuals with malaria so it is possible that these estimates may converge given more data. If true however, this might suggest that the level of protection conferred by HbS may vary between populations in a way that is not clearly explained by variation in parasite genetics at these loci.

This analysis was based on the previously established *Pfsa*1-4 loci, and the putative association at *Pfsa*2B. However, as depicted in Figure 5.1 (pg.153), the research presented within this chapter suggests a more complex picture where multiple alleles potentially contribute in ways that vary across populations. Efforts to increase power of this type of analysis, by determining parasite genotypes in more infected HbAS/SS individuals, will be invaluable to determining the protective effect of HbS in the context of further *Pf* genotypes, and accurately assessing differences between populations. Additionally, methods to effectively analyse multiple signals will need to be developed.

5.2.3 A review of the association of HLA-B*53 and protection against severe malaria

A goal of this work has been to reassess the widely cited association between HLA-B*53 and protection from severe malaria discovered using GAMCC (52). The original publication reported a strong effect of HLA-B*53 on malaria severity, which replicated within Gambia (52). However, additional replication of this signal was not widely observed (90), and no association of the HLA-B*53 allele was observed with malaria protection within the largest GWAS analysis to date (6). The disparity in these results may have various sources which I will investigate in the following sections.

5.2.3.1 Power analysis of the original study

One potential criticism of the original analysis is that the study may have been underpowered due to the small sample size by modern standards ($N_{cases}=613$ and $N_{controls}=510$) leading to an overestimated effect. To investigate this, I simulated genotypes under the assumption of a true association, using the sample sizes available within the original study and considering two effect sizes ($OR=0.5$, $OR=0.75$). I also performed these simulations at varying allele frequencies of HLA-B*53 within the population (5%, 25% and 50%). Through association testing on these simulated datasets, I determined the probability of obtaining a P-value below a given threshold when an association is genuinely present (the power). Using Bayes Theorem, it is possible to use these power calculations, in conjunction with a prior assumption on the probability of a true association, to estimate the probability that an association is true given it is below a certain P-value threshold (sometimes called the true discovery rate) (details of this are displayed in Materials and Methods section 2.7.3 (pg.61)).

Within this calculation, I assessed two different priors; the first was a stringent, GWAS-type prior that of 1 million recombining segments in the human genome, one might truly be associated with severe malaria protection. Figure 5.14 on the following page makes clear that under this prior the probability of true discovery was low under all sets of parameters such that only a very small P-value threshold ($P < 1 \times 10^{-7}$), a level lower than observed within the original study, would be indicative of robust effect (Fig. 5.14). However, the original study focussed on only class I and II HLA loci given their importance within infectious disease (190). Because of this, I also tested a second was a more relaxed prior that 1 of the 58 HLA loci investigated within the original study was associated with malaria severity. Under this prior, there was a modest level of observed TDR (Fig. 5.14); at 25% allele frequency (like the observed frequency) with an effect size of $OR=0.5$, $TDR=0.597$ and when $OR=0.75$, $TDR=0.307$. These values do not meet standard thresholds for power utilised during the preparation of modern studies. Furthermore, there was a non-trivial probability of a false positive association. However there is a possibility that the association of HLA-B*53 and protection against severe malaria could be true under this model.

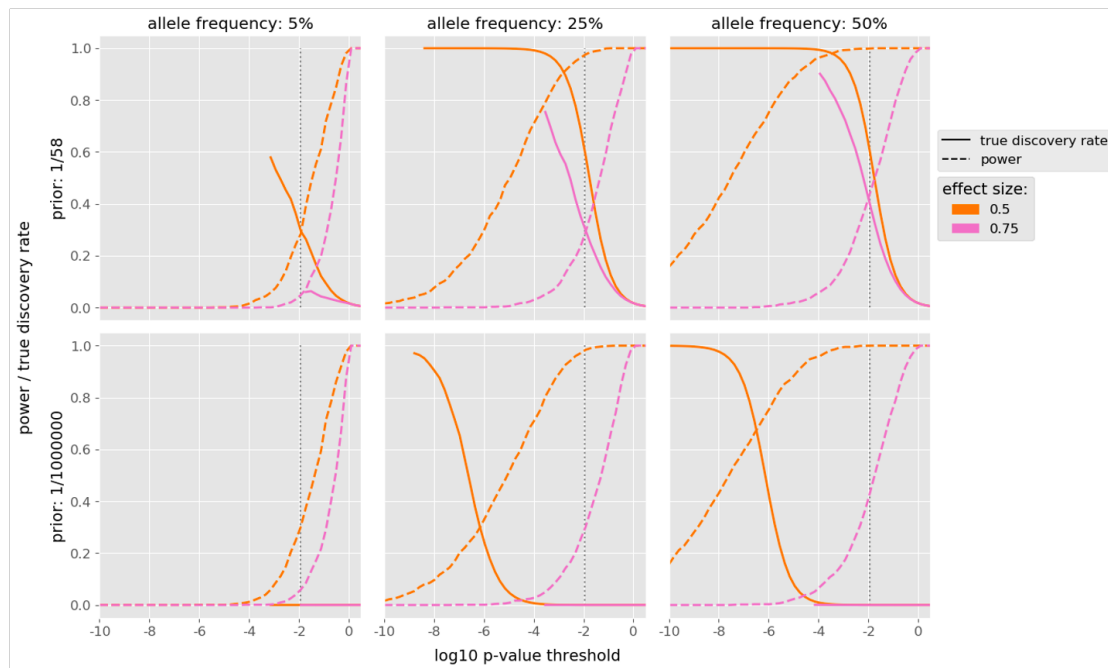


Figure 5.14: **Power and true discovery rate analysis for the analysis of an association of HLA-B*53 to protection against severe malaria.** Model computed power (dashed lines) and true discovery rates (solid lines) for p-value thresholds ($\log_{10}$) on the x axis. These values were calculated based on 1000 simulations using two prior probabilities of association, 1/58 (top) and 1 in 1000000 (bottom), two effect sizes (odds ratio (OR)=0.5 in orange and OR=0.75 in pink) and across three baseline allele frequencies; 5% (left), 25% (middle) and 50% (right). Sample sizes matching those used in the discovery study were used.

5.2.3.2 Inclusion of ethnicity to account for population structure

Another factor which has been known to significantly hinder association analyses in human genetic data is population structure (198). In the original study, this was controlled for by including ethnic group within the association computation using a Mantel-Haenszel test which, similarly to a standard chi-squared test, compares observed to expected counts in categorical data stratified by a single covariate (52). Since then, the favoured methodology for this type of analysis has become regression because it allows the inclusion of multiple covariates, such as principal components which are typically used to control for population structure (199). Unfortunately, human whole genome sequencing (WGS) data is not available for the full GAMCC dataset, and as such a comprehensive test for population structure, such as principal component analysis, cannot be attempted. However, to confirm the original method produced comparable results to a regression-based method, I reperformed the analysis using logistic regression, with and without including ethnic group as a covariate. These produced results very similar to those reported, with ethnic group having little influence on estimated effect size (Fig. 5.15 on the next page).

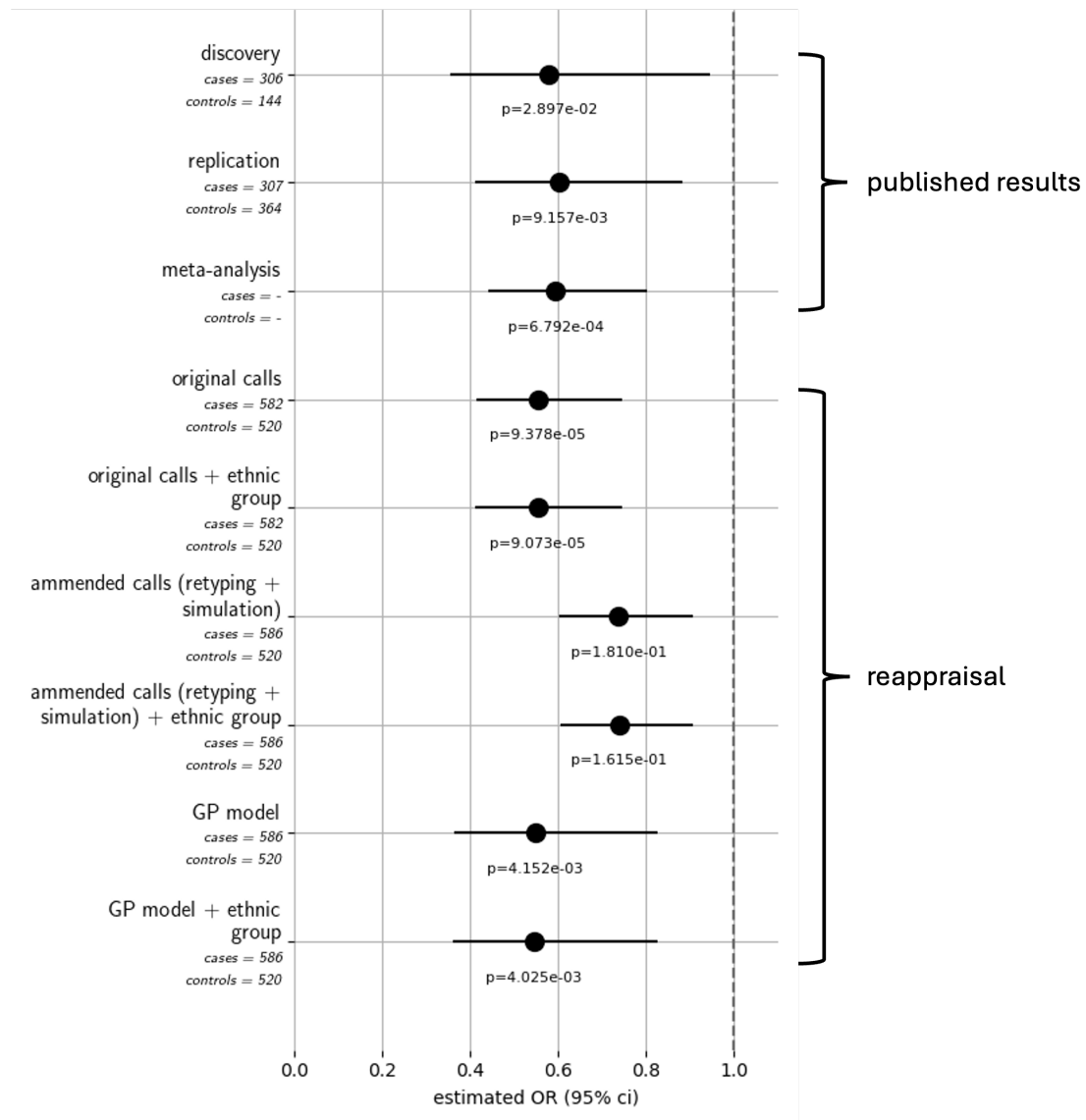


Figure 5.15: A comparison of results for the association of HLA-B*53 and protection against severe malaria discovered in the GAMCC dataset. The upper panel shows results of a logistic regression performed using counts reported in the original study (published results). The bottom panel shows results of additional regressions performed using data inherited for the set (original calls), using ethnic group as a covariate in the regression (+ ethnic group), using HLA-B*53 genotyping calls across a set of 1000 simulations using assay accuracy with corrections on any sample for which we have high resolution HLA typing (retyping + simulation), the same simulations using ethnic group as a covariate, and using genotype probabilities as calculated using the accuracy of the HLA-B*53 assay determined using high resolution HLA typing calls (GP model) with and without ethnic group. Sample sizes for each group are indicated with each y-axis label. 95% confidence interval error bars are shown.

5.2.3.3 An evaluation of the accuracy of the HLA-B*53 typing assay

A potential possibility is that the assay used to type HLA-B*53 in the original study was inaccurate, leading to an erroneous estimate of the frequency of HLA-B*53. To make the initial discovery that HLA-B*53 appeared associated with protection against severe malaria, serological typing was utilised (52). In this process, sample lymphocytes are tested against a panel of well-characterised antisera containing HLA-specific antibodies. Cell lysis in response to a specific antiserum indicates the HLA type positivity of the cells (200). This was performed on a subset of 450 samples to gain near complete HLA genotypes. Following the initial discovery that HLA-B*53 frequencies varied significantly between severe malaria and non-malaria control groups, a polymerase chain reaction (PCR) and oligo hybridization assay was designed to type the presence of HLA-B*53 and this was performed on 671 samples within a replication set (52).

	True	False
Observed True	47	16
Observed False	11	171

Table 5.4: **Results of HLA-B*53 typing compared to original typed calls.** Counts of the number of samples in the retyped set (N=250) that genuinely carried HLA-B*53 (True) or not (False) stratified by the assay used to type HLA-B*53 in the original study by Hill et al. (52) (observed true and false).

To assess the accuracy of this assay in the GAMCC data resource, we obtained high resolution HLA typing for 250 samples in the set (180 malaria cases and 70 non-malaria controls, see Chapter 3 (pg.87) for details). As shown in Table 5.4, of the 250 accurately retyped samples, 218 were correctly called by the PCR/oligo-based assay; 16 of 63 were wrongly assayed as HLA-B*53 positive by the original assay, and 11 of 182 were wrongly assayed as HLA-B*53 negative, leading to an overall accuracy of 89%.

At the time of the study, complete reference sequences for HLA alleles were not available, nor had a thorough catalogue of existing *HLA-B* alleles been developed. To understand the above results, I compared sequence homology at the binding sites of the PCR primers and probe within alleles present in the retyped

set (Fig. 5.16 on the following page). Multiple alleles appeared with false positive calls. Of note, one allele (HLA-B*51:01:01) appeared within 4 of 16 false positive samples. This allele also had the fewest cumulative number of differences between its sequence and the designed assay oligos. This suggested that mispriming to the HLA-B*51 allele may have contributed to the inaccuracy of the assay. This allele was reported to have moderate frequency within The Gambia at the time (AF=5.6%), although was less frequent within our retyped sample (AF=2.1%). Although infrequent, at scale this may have contributed to the degree of false positives seen within data.

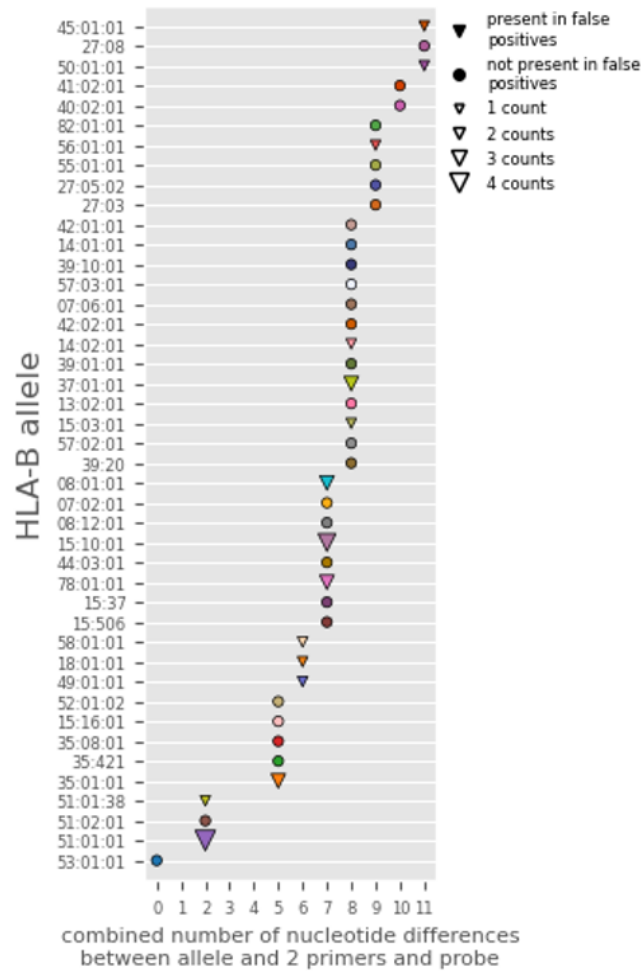


Figure 5.16: **Differences between HLA-B alleles and oligonucleotides used to type HLA-B*53.** Counts of cumulative nucleotide differences (x-axis) between the HLA-B*53 PCR primers and oligo probes used to type HLA-B*53 in malaria cases and non-malaria controls by Hill et al. (52) and *HLA-B* alleles (y-axis) found to be present in the set based on high resolution retyping of N=250 samples in the set. Triangle markers indicate that the allele was observed within a false positive HLA-B*53 call by the assay. Circles were not observed. The size of the triangle indicates the number of instances that allele was observed as a false positive.

5.2.3.4 An assessment of how accuracy of the HLA-B*53 typing assay may have affected results of the original study

It is possible that the level of inaccuracy of the HLA-B*53 assay may have confounded the reported results, leading to a false positive association between this allele and severe malaria protection. The retyped calls allowed us to estimate the probability of a sample truly carrying HLA-B*53 if the assay did and did not report a positive result (which was computed as 74% and 6% respectively). I then performed simulations of true genotypes based on this inaccuracy (for samples with renewed HLA-typing their true call was retained) (see Materials and Methods section 2.8.2 (pg.68) for details). Within these simulations the average observed effect was reduced compared to the original association (OR=0.740 (0.601-0.906), Fig. 5.15 (pg.182)).

To trial another method which incorporated the uncertainty in calls made by the assay, I performed the regression using the software SNPTEST based on genotype probabilities. This method averages the likelihood function over possible genotypes using a ‘GP’ field, as opposed to averaging estimates over different possible genotypes (as in the simulation approach). This however did not significantly alter results in relation to those originally reported (OR=0.548 (0.363-0.827)) but did widen the errors and reduce the significance of the association (Fig. 5.15 (pg.182)). The reasons behind the discrepancy in results from each of these methods are not obvious but may relate to differences in the handling of uncertainty. Overall, this suggested that the marginal inaccuracy of the assay did not strongly affect association results.

5.2.3.5 A genome-wide scan for associations in the *Pf* genome to HLA-B*53

The discovery of the *Pfsa* loci however hints that the unassessed variability in parasite genomes might underlie the temporal and spatial disparity seen in the association between *Pf* variation and HLA-B*53. The GAMCC resource includes *Pf* WGS data for these samples which allows the opportunity to explore the HLA-B*53 association considering parasite variation.

To better understand how parasite variation might influence the observed association between HLA-B*53 and protection from severe malaria, I will first briefly

outline models of potential interactions that could have impacted the ascertained HLA-B*53 association. Specifically, I will consider two contrasting hypotheses: one where the association between HLA-B*53 and protection from severe malaria is genuine, and one where it is spurious. By examining how parasite genetic diversity might confound these scenarios, we can gain insights into whether the observed relationship is robust or influenced by unaccounted parasite variability.

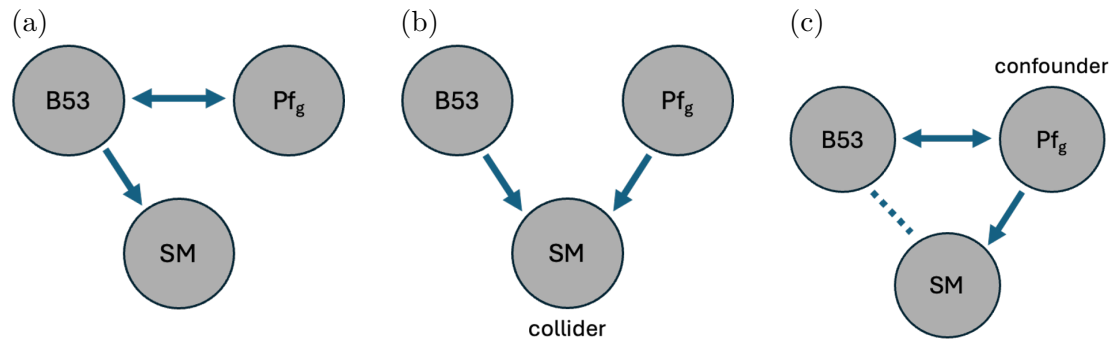


Figure 5.17: **Models of interactions between HLA-B*53 status, parasite genotype (Pf_g) and malaria disease severity.** (a) A genetic interaction between HLA-B*53 and a parasite genotype exists. This affects the observed correlation between HLA-B*53 and disease severity. (b) HLA-B*53 status and parasite genotype independently affect disease severity (which becomes a collider). Parasite genetics confounds the observed association between HLA-B*53 and severe malaria. (c) A genetic interaction between parasite genotype and HLA-B*53 and disease severity exists. Parasite genotype acts as a confounder to produce an false positive association between HLA-B*53 and disease severity.

The association of HLA-B*53 is true Evidence presented within the original study suggests a true association between HLA-B*53 and disease severity. If this association was real and causal, this could have led to the emergence of parasites with an ability to overcome HLA-B*53 mediated protection, causing this association to disappear, explaining a lack of subsequent replication of this signal in more recent datasets. Alternative to a selective-sweep type event is that of a longer-term host-parasite coevolutionary interaction between HLA alleles and parasite genotypes. It is known that genetic interactions like this exist between the *Pf* and human genomes like between the HbS and *Pf*sa loci (130; 131). The absence of replication of this association in other tested populations might therefore be due to

the variable presence of HLA-B*53 resistant parasites (Fig. 5.17a on the preceding page). In both cases we would expect that infections of HLA-B*53 carriers would involve parasites who carry more HLA-B*53 resistant genotypes compared to non-HLA-B*53 carriers, as these parasites are better adapted to infect this group. This can be interrogated by assessing for associations between HLA-B*53 and parasite loci.

The association of HLA-B*53 is false Alternatively, parasite genetic variation might have acted to cause a false-positive association between HLA-B*53 and disease severity. If parasite genetics impacts disease severity, and this locus also interacts with HLA-B*53, this would lead to a perceived correlation between HLA-B*53 and severity through a confounding effect (Fig. 5.17c on the previous page). Although in this thesis I have reported no evidence of strong parasite genetic effects on malaria severity (Chapter 4 (pg.110)), this was based on a small sample and restricted to the core genome so did not investigate variation in highly complex, repetitive or diverse genome regions. It therefore may still be the case that parasite genetic factors influence severity. Indeed, under this scenario the HLA-B*53 association may still exist but is hidden by the effect of *Pf* genetics on severity (Fig. 5.17b on the previous page).

A scan for associations with HLA-B*53 In order to evaluate the scenarios outlined above, I performed a genome wide scan for associations between parasite genetic variation and host HLA-B*53 status. For the 831 samples in the GAMCC analysis set, logistic regressions were performed using the programme HPTTEST (130) using HLA-B*53 carrier status as the predictor and parasite genotype at common biallelic SNPs as the outcome.

No variants showed strong evidence of association with HLA-B*53 (Fig. 5.18 on the following page). The most associated variant was Pf3D7_01_v3:192084:G>A, a missense variant in the gene ubiquitin carboxyl-terminal hydrolase 1 (UBP1, PF3D7_0104300); however this only showed modest strength of evidence of association ($\log_{10}BF=3.19$, $OR=3.36$ (1.88-5.99), $P=4.16 \times 10^{-5}$, $N=762$, $AF=8.7\%$). These results suggest little evidence of a parasite genetic interaction with this

HLA allele, rendering it unlikely that this has acted to confound assessments of resistance effects of the HLA-B*53 allele.

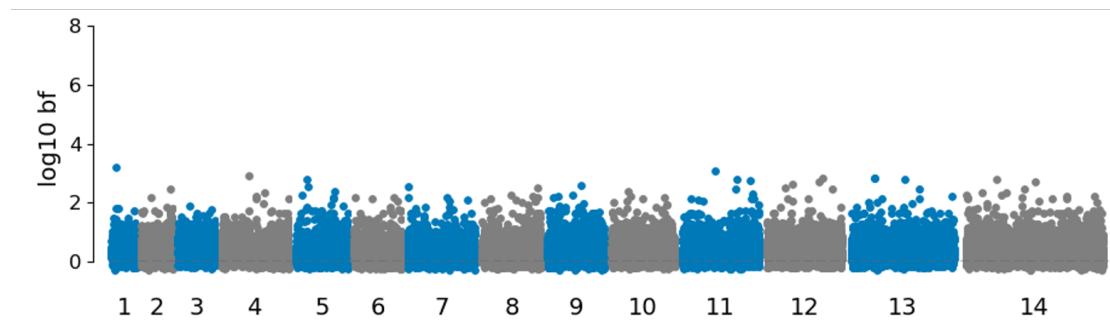


Figure 5.18: **Evidence for association between *Pf* variation and HLA-B*53 in the GAMCC dataset.** Strength of evidence ($\log_{10}(\text{Bayes Factor (BF)})$) for the association of variants across the 14 chromosomes in the *Pf* genome with the human class I HLA allele HLA-B*53 in the GAMCC cohort (N=807). HLA-B*53 calls were partially generated from high resolution HLA typing (N=171) and partially from historical data inherited for the cohort (N=636). Only variants with less than 10% of samples missing or heterozygous are shown (N variants = 48969).

5.2.4 Population genetics at the *Pf*sa loci

The presence of host-parasite interactions within malaria revealed at the *Pf*sa loci raises important questions as to their evolution. In particular, the spatial variation in HbS-*Pf*sa interactions described in this chapter indicates that these forces are not uniform and are thus potentially influenced by several factors. Certain contextual features offer clues in this regard. For one, *Pf*sa is associated with HbS which is maintained through forces of balancing selection due to the resistance advantage conferred to heterozygote individuals, in contrast to the significant disadvantage of sickle-cell disease in the homozygous form (201). The *Pf*sa+ alleles, like HbS, show intermediate allele frequencies across Africa which appear correlated to the frequency of HbS, but are not fixed in any population (130). This might suggest correspondingly complex modes of selection in the *Pf* genome. For the purposes of this thesis, I will describe some observations and initial analyses made into the population genetics of the *Pf*sa loci using the GAMCC and other public datasets in order to work towards carefully unpicking this interaction.

5.2.4.1 Unusual signals of LD in the GAMCC set

As alluded to earlier in this chapter, a distinctive yet unexplained feature of the previously discovered *Pfsa* loci is that of strong LD observed between them across populations. This is particularly striking because *Pfsa* loci often fall on different chromosomes in the *Pf* genome and thus are generally expected to randomly segregate (130; 131). In order to explore this within the newly generated GAMCC dataset, I developed a custom programme to calculate Pearson's correlation coefficient R between all common ($AF > 5\%$) biallelic pairs of SNPs with low missingness/heterozygosity ($< 20\%$) within the GAMCC call set ($N^{variants} = 51,429$), leveraging matrix operations for enhanced computational efficiency (Fig. 5.19 on the next page). Pairs of alleles with high measurements of LD ($|R| > 0.35$) were then grouped into regions if nearby SNPs also showed elevated LD to the same region (Materials and Methods section 2.9.3 (pg.74)). This produced seven regions pairs within the *Pf* genome which showed high levels of between-chromosome correlation in the GAMCC set (appearing at the upper end of the distribution of genome wide values as displayed in Fig. 5.19). Interestingly, six of the identified region pairs identified involved *Pfsa* loci (Supp. Table 5.6 (pg.222)), the strongest of which was between variation within the *Pfsa1* and *Pfsa3* regions (peaking at $R = 0.59$ (0.547-0.636) between Pf3D7_02_v3:631092:T>C ($AF = 9\%$) and Pf3D7_11_v3:1057437:T>C ($AF = 14.4\%$)).

Among these signals also appeared the newly described *Pfsa2B* locus (Pf3D7_02_v3:825357:T>C ($AF = 7.7\%$)) which showed strong levels of correlation to the *Pfsa3* locus (with Pf3D7_11_v3:1057437:T>C $R = 0.43$ (0.374-0.492)) and, notably, to variation at the *Pfsa4* locus (with Pf3D7_04_v3:1122147:A>C ($AF = 83.6\%$) $R = -0.36$ (-0.419 - -0.292)). This is surprising given the absence of strong evidence of association of the *Pfsa4* region to HbS in the GAMCC and other Gambian datasets (Fig. 5.2 (pg.155)). The presence of elevated LD between the putative *Pfsa2B* locus and the *Pfsa3* and *Pfsa4* loci adds weight to the results of the genome-wide scan presented above which suggested the *Pfsa2B* locus is associated with HbS. Assessing this in further populations, levels of LD between the *Pfsa2B* locus and *Pfsa3* (which was the most highly correlated pair for the *Pfsa2B* locus

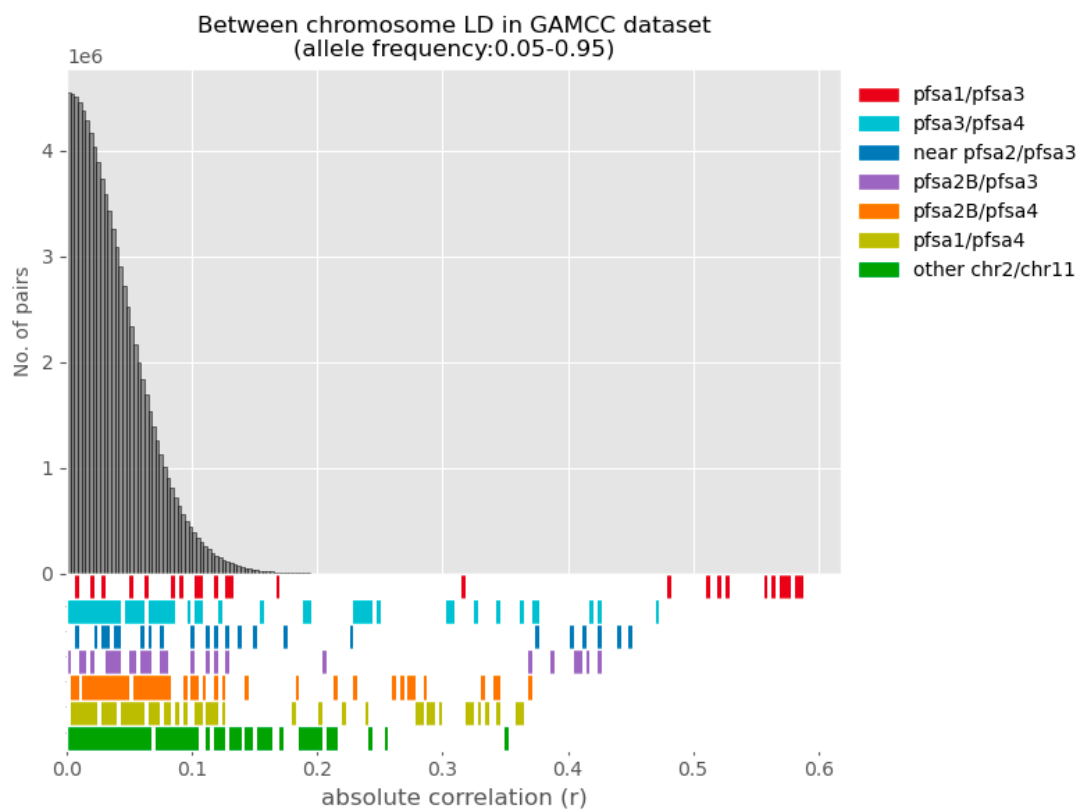


Figure 5.19: Genome-wide between chromosome LD in the GAMCC dataset. Distributions of absolute Pearson's correlation (r) values between all pairs of variants on separate chromosomes in the malaria parasite genome in the GAMCC cohort. Only biallelic single nucleotide polymorphism variants with an allele frequency between 5-95% were included. Below the distribution indicates correlation values for pairs of variants within the 10Kb region containing the highest ranking LD values genome-wide as indicated by the colour key (these values are tabulated in Table 5.6 (pg.222)).

in the GAMCC dataset) vary extensively between populations in the Malaria-GEN Pf7 resource, showing moderately elevated levels in Gambia ($R=0.27$ (0.201-0.33), $AF_{Gambia}^{Pfsa2B}=35.58$, $AF_{Gambia}^{Pfsa3}=58.62$, $N=1650$), Guinea ($R=0.34$ (0.189-0.478), $AF_{Guinea}^{Pfsa2B}=17.81$, $AF_{Guinea}^{Pfsa2B}=25.33$, $N=292$) and Mali ($R=0.47$ (0.424-0.517), $AF_{Mali}^{Pfsa2B}=15.27$, $AF_{Mali}^{Pfsa3}=19.36$, $N=2168$) in Western regions, but also in Malawi ($R=0.44$ (0.335-0.53), $AF_{Malawi}^{Pfsa2B}=2.64$, $AF_{Malawi}^{Pfsa3}=7.55$, $N=530$), Tanzania ($R=0.32$ (0.246-0.394), $AF_{Tanzania}^{Pfsa2B}=14.6$, $AF_{Tanzania}^{Pfsa3}=28.41$, $N=1144$) and Kenya ($R=0.24$ (0.16-0.312), $AF_{Kenya}^{Pfsa2B}=10.42$, $AF_{Kenya}^{Pfsa3}=25.15$, $N=1200$) in the East African populations, although this may be confounded by the presence of the nearby *Pfsa2+* allele within these populations (Fig. 5.20d on the following page). The observed levels of LD were not as extreme as of that between *Pfsa1* and *Pfsa3* across Africa which reach $R>0.75$ in many populations (Fig. 5.20a on the next page), nor were they like that of *Pfsa4* where LD appears restricted to Western populations (Fig. 5.20c on the following page).

An additional signal of LD was observed in the GAMCC set between a variant adjacent to the original *Pfsa2* locus (Pf3D7_02.v3:814297:C>T (+9bp from the lead *Pfsa+* variant in the discovery analysis) ($AF=8.8\%$)) and variation at the *Pfsa3* locus (with Pf3D7_11.v3:1054282:T>C ($AF=14.1\%$), $R=0.45$ (0.394-0.511)). This is more surprising still given the lack of *Pfsa2+* alleles in West Africa (Fig. 5.4b (pg.160)). This SNP was recorded as a multiallelic site within our call sets; however, this site was in fact biallelic within Gambia and Kenya, but via a different non-reference nucleotide. Specifically, Pf3D7_02.v3:814297:C>T occurred at an 8.8% and 18.3% AF in the GAMCC and CP1 Gambian cohorts (compared to $AF^{Kenya}=0\%$), whereas Pf3D7_02.v3:814297:C>A occurred at a 3.2% AF within Kenyan samples ($AF_{CP1}^{Gambia}=0\%$, $AF^{GAMCC}=0\%$). Both alleles cause a non-synonymous amino acid substitution: the first causes Ala57Thr and the second Ala57Ser. Because this SNP was not recorded as biallelic within our call sets, it was excluded from the association analysis between *Pf* variation and HbS. Efforts to recover this within this analysis should be undertaken in future to understand this LD signal.

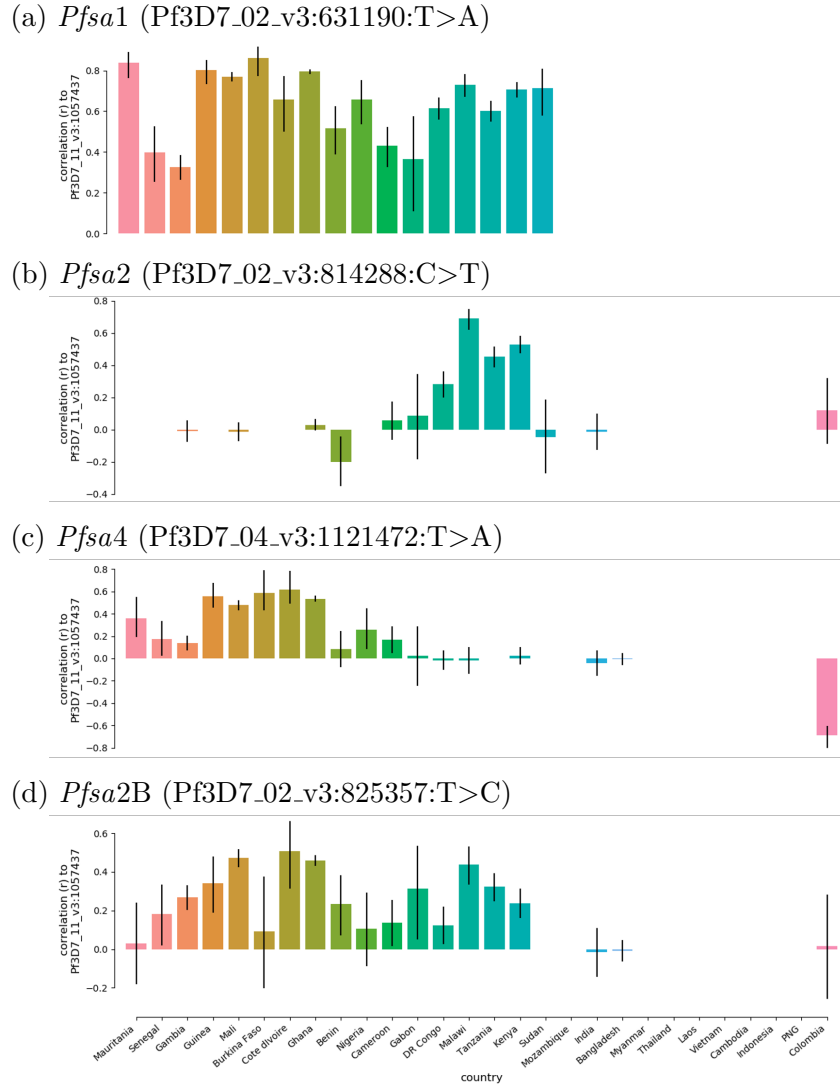


Figure 5.20: **Correlation between *Pfsa* variants and the *Pfsa3* locus within populations in the MalariaGEN Pf7 resource.** Pearson's R value between each variant and a strongly HbS-associated variant at the *Pfsa3* locus (Pf3D7_11_v3:1057437:T>C) for populations in the MalariaGEN Pf7 resource. This *Pfsa3* variant was chosen because it appears associated to HbS in all the so far tested populations and is present across all the tested call sets. 95% CIs calculated using a Fisher transformation are shown.

5.2.4.2 The allele frequencies of the *Pf*sa loci have increased in The Gambia

During the analyses presented in this chapter it was noted that allele frequencies of the *Pf*sa loci are much higher in samples collected in Gambia in the CP1 resource (1996-2009) than within the GAMCC resource (1988-90). Using the ‘intersected’ call set (described in Chapter 4 (pg.110)) which comprises all samples from Gambia from the CP1, GAMCC and Pf7 resources, we assessed AFs of the *Pf*sa loci over time and noticed a distinctive trend of AF increase for *Pf*sa1+, *Pf*sa2B+ and *Pf*sa3+ alleles in the region (Fig. 5.21a on the next page). This was most stark for *Pf*sa3 which had an AF of 15.0% (12.6-17.4 95% CI) in 1988 (N=420) and peaked with an allele frequency of 72.8% (67.6-78.1 95% CI) in 2015 (N=138) (Fig. 5.21a).

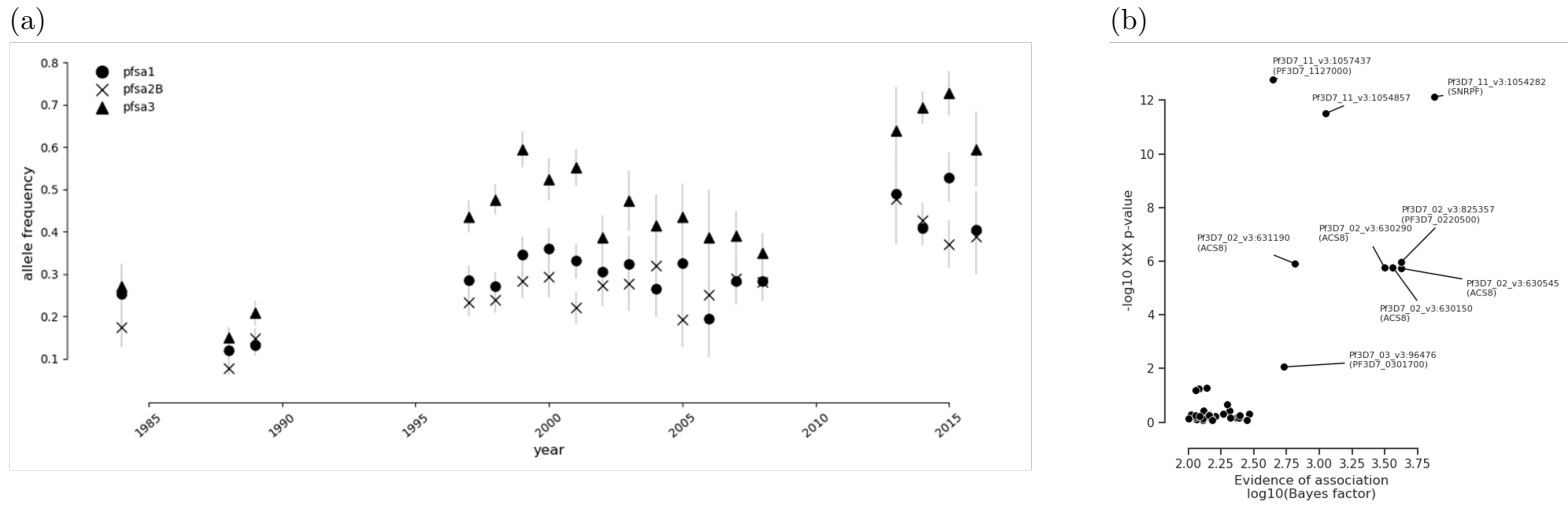


Figure 5.21: **Increase in *Pfsa* allele frequencies overtime in The Gambia.** (a) The allele frequency of a lead variant at the *Pfsa1* as circles, *Pfsa3* (Pf3D7_11_v3:1057437:T>C) as triangles and the putative *Pfsa2B* locus as crosses. 95% confidence interval error bars are shown. Pf3D7_11_v3:1057437:T>C as opposed to Pf3D7_11_v3:1058035:T>A was chosen for *Pfsa3* due to the consistency of allele denotation of this variant across separate datasets. (b) $-\log_{10}(P_{X^tX})$ values for the most HbS-associated variants in the GAMCC cohort. P_{X^tX} values indicate the evidence of differentiation for the variant within an empirical model of allele frequency shift computed between Gambian samples from 1988-89 to samples from 2013-17, using samples from 1984 as an ancestral allele frequency. Variants which are associated with HbS ($\log_{10}(\text{Bayes Factor}) > 2.5$) and which show evidence of differentiation ($-\log_{10}(P_{X^tX})$ value > 2) are labelled with their chromosome, position and gene ID.

To determine whether this change in AF over time was beyond that expected due to genetic drift, these loci were compared to a model of genome-wide AF variation generated for two time points; “past” samples from the GAMCC resource (1988-90) and “contemporary” samples within the Pf7 dataset (from 2013-17) compared to an ancestral frequency (determined from a sample collected in 1984 available with Pf7 dataset). Briefly, AFs from the past and contemporary samples were transformed and compared through a covariance matrix, which allowed the identification of variants with unusual shifts from the expected neutral variation, indicating the potential occurrence of evolutionary pressures (described in full in Materials and Methods section 2.9.2 (pg.70)). Figure 5.21b on the previous page shows that many variants in the Gambia shown to be associated with HbS, including the putative *Pfsa2B* variant, showed evidence of extreme allele frequency shift compared to genome-wide frequency variation. In particular, very strong signals were observed for the *Pfsa3+* alleles (with very small P_{X^tX} values $< 1 \times 10^{-12}$ indicating a major deviation from the genome-wide expectation). This contrasts with current evidence from Kenya where allele frequencies at the *Pfsa1-3* loci have remained constant over time (130). This suggests that these variants may have undergone positive selection within this population.

5.2.4.3 Haplotypes at the *Pfsa* loci are shared across populations

To investigate the potential for positive selection at the *Pfsa* loci further, I investigated haplotype sharing at the *Pfsa* loci across global populations. To do so, in collaboration with Dr. Gavin Band, we generated a filtered version of the MalariaGEN Pf7 call set, focussing only on African samples with little evidence of mixed infection ($Fws > 0.9$). In this sample, biallelic SNP sites were polarized to represent ancestral and derived alleles. At missing or heterozygous sites, variant calls were imputed using the Beagle software to produce a dataset of complete fully homozygous haplotypes (see Materials and Methods section 2.6.4 (pg.57) for further detail).

In principle, there is the possibility that the *Pfsa+* mutations arose independently in separate locations, or that they have a shared evolutionary origin and have spread between populations. To assess this, I have created images of polarised

haplotypes which have been clustered using a hierarchical method where samples are grouped together based on pairwise allele matching.

At the *Pfsa1*, *Pfsa2*, *Pfsa3* and *Pfsa4* loci, haplotypes carrying the *Pfsa+* allele appeared distinct from those carrying the *Pfsa-* allele (highlighted in neon green), causing haplotypes to be clustered by *Pfsa+* genotype rather than by population (Fig. 5.22 on the next page). Furthermore, at all the loci several other derived mutations appeared to be essentially perfectly correlated to the *Pfsa+* site (e.g. at *Pfsa1* Pf3D7_02.v3:630150 and 630290, at *Pfsa2* Pf3D7_02.v3:813895-814329, at *Pfsa3*, Pf3D7_11.v3:1054282 and 1054587 and at *Pfsa4* Pf3D7_04.v3:1121457 and 1119762). At the same time, at each locus there were also derived mutations which appeared on a subset of *Pfsa+* haplotypes (but were still not seen on *Pfsa-* haplotypes). For example, at the *Pfsa1* region some *Pfsa1+* parasites carried the derived allele at the Pf3D7_02.v3:630420 locus whereas others did not.

These features strongly indicate that haplotypes at each locus have a single mutational origin as it is highly unlikely that further mutations would coincide with the central *Pfsa+* allele independently within multiple evolutionary lineages. Following the *Pfsa+* mutation, it appears that other mutations have occurred within these lineages leading to modern *Pfsa+* haplotypes. Taken together, this suggests that these haplotypes are shared across Africa and thus have a potentially ancient evolutionary origin.

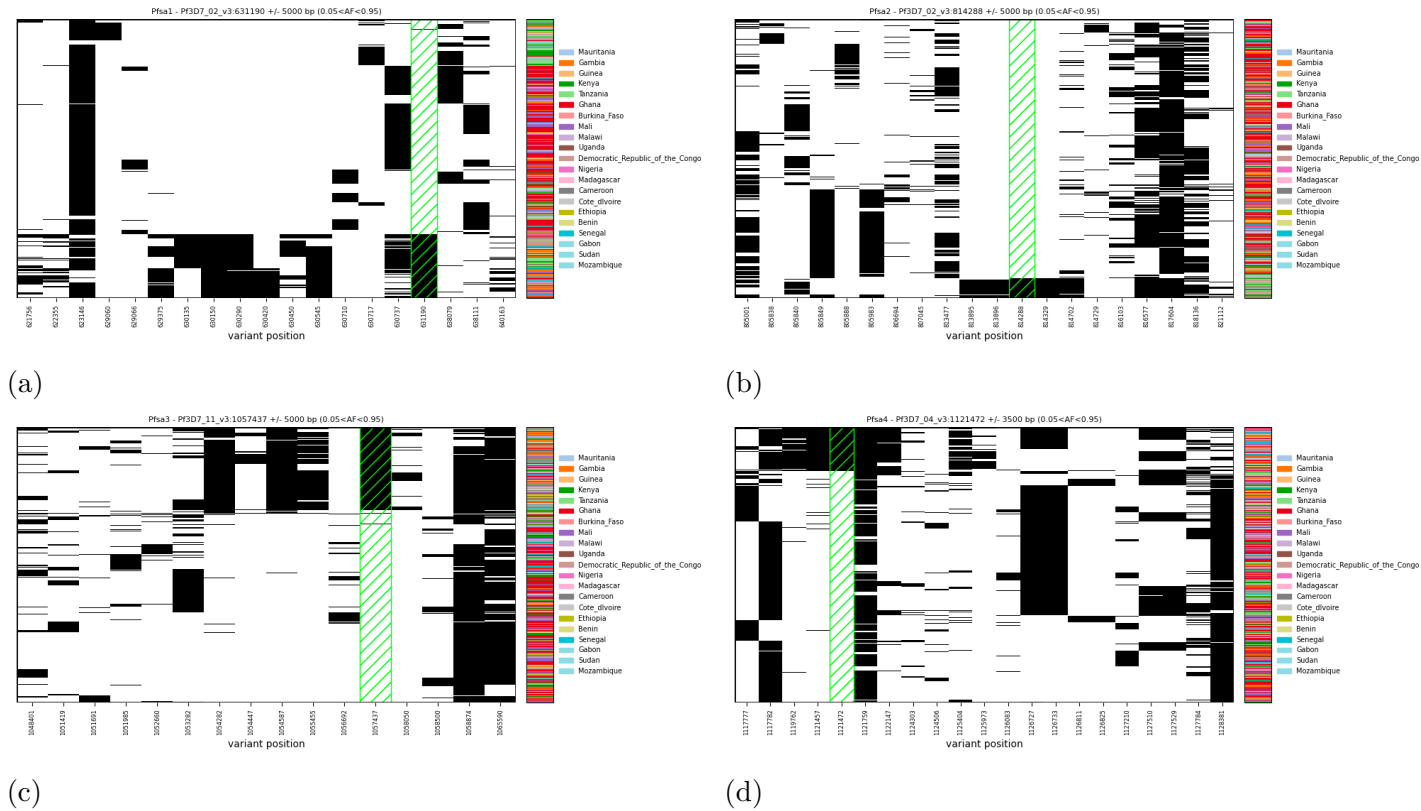


Figure 5.22: Haplotypes at the *Pfsa* loci in samples from the MalariaGEN Pf7 resource. Haplotypes at the previously discovered *Pfsa*1-4 loci are shown. SNPs with an allele frequency (AF) < 0.05 or AF > 0.95 were filtered prior to clustering. A colour bar indicating the country of origin of the sample is shown on the right. The top HbS-associated variant at each locus is highlighted in hashed neon green.

Inspecting haplotypes at the *Pfsa2B* locus however showed that, although parasites carrying this allele generally clustered together, there was not a clear indication of a common haplotype at this locus across populations (Fig. 5.23). It should be noted that this is a simplified view of genetic variation at the locus, and other features such as INDELs, or multiallelic variation, might be needed to more accurately infer relationships.

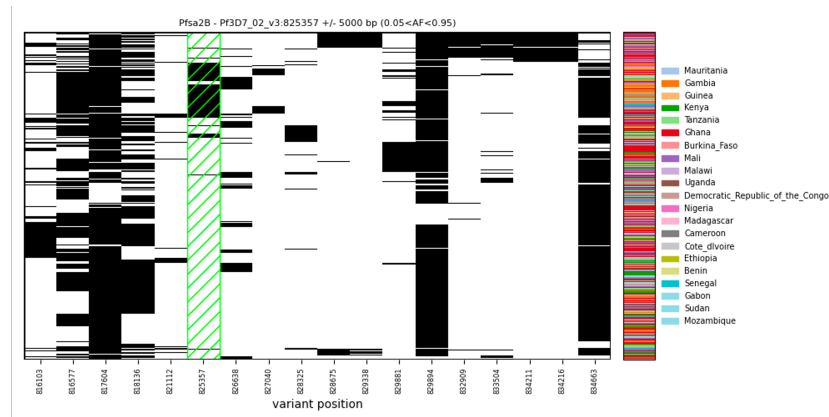


Figure 5.23: **Haplotype at the *Pfsa2B* locus in samples from the MalariaGEN Pf7 resource.** Haplotypes at the putative *Pfsa2B* locus described within this chapter. SNPs with an allele frequency (AF)<0.05 or AF>0.95 were filtered prior to clustering. A colour bar indicating the country of origin of the sample is shown on the right. The top HbS-associated variant is highlighted in neon.

5.2.4.4 There is minimal evidence of recent positive selection at the *Pfsa* loci in global populations

One way to empirically assess the occurrence of positive selection is using extended haplotype homozygosity (EHH) which measures the degree to which a haplotype, extending out from a focal allele, is shared within a dataset. Differences in EHH between ancestral and derived alleles can be indicative of positive selection, because a positively selected allele tends to be flanked by other neutral mutations leading to high EHH (also known as genetic hitchhiking) (82). Over time however, these haplotypes are broken down by recombination events; EHH is therefore only likely to detect recent positive selection. In practice, EHH is compared across the genome using the integrated haplotype score (iHS) statistic

which compares EHH for the ancestral and derived alleles at a locus standardised within allele frequency bins.

To assess for evidence of recent positive selection, I therefore computed iHS across the polarised haplotypes generated from the Pf7 call set within the Gambian, Ghanaian and Kenyan populations and across all African samples combined (including Gambia, Ghana and Kenya). In my calculations, large negative iHS values indicated an extended haplotype carrying the derived allele, suggesting the occurrence of positive selection of the derived allele, whereas an unusually positive value indicated a short haplotype (a full description of the method is within Materials and Methods section 2.9.5 (pg.76)).

Standardised iHS values follow an approximately normal distribution; these distributions, and the iHS value for the indicated *Pfsa* locus within each tested population, are shown in Figure 5.24 on the following page. We found that iHS values did not deviate significantly from other genome-wide variants at the *Pfsa1*, *Pfsa2*, *Pfsa3*, *Pfsa4* or *Pfsa2B* loci within Gambia, Kenya, Ghana, or across all African populations. The largest iHS value was observed in Kenya at the *Pfsa1* locus but the evidence of positive selection was nevertheless weak ($iHS_{Pfsa1}^{Kenya} = -1.111$, P-value=0.267). This contrasts with iHS values calculated at known drug resistance loci within the same data; for example, at the CQ resistance locus within the CRT gene in Gambia, $iHS_{CRT}^{Gambia} = -3.04$ showing more striking evidence of positive selection (P=0.002) (Supp. Fig. 5.35 (pg.218)). Overall, these results suggest that the *Pfsa* regions are unlikely to have undergone a recent selective sweep in any of the studied populations.

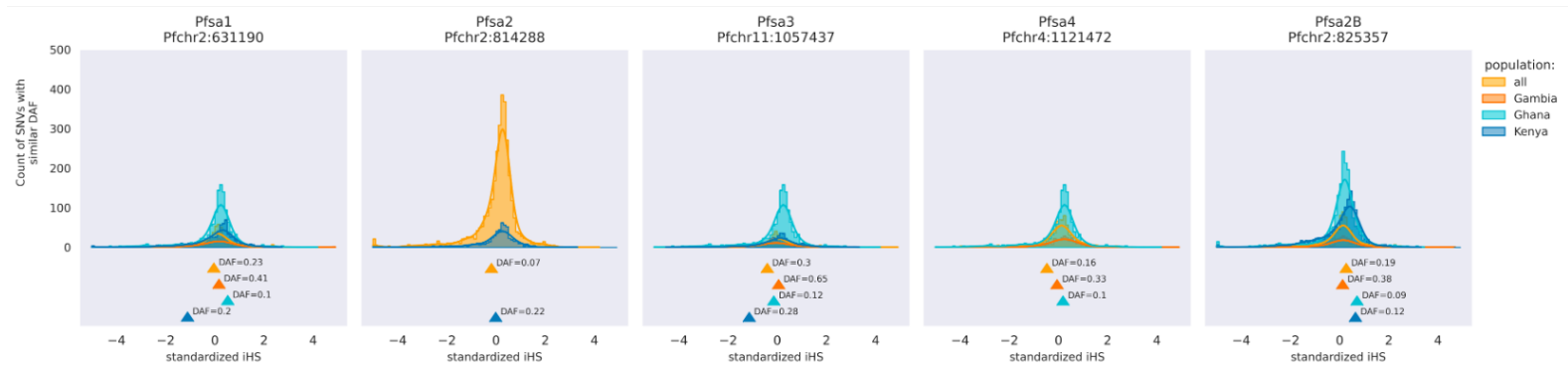


Figure 5.24: **iHS results for the *Pfsa* loci in samples from the MalariaGEN Pf7 resource.** Results for the previously discovered *Pfsa*1-4 loci, and the putative *Pfsa*2B locus discovered in the GAMCC dataset are shown. integrated haplotype score (iHS) values across the entire filtered dataset (“all”), within Gambia, Ghana and Kenya are shown. Histograms show the distributions of iHS values for all variants within the same AF bin (2.5% width), and arrows indicate the iHS value for each target SNP.

5.3 Discussion

In this chapter, I have presented multiple avenues of investigations into genetic interactions between humans and *Pf* parasites. First, I described a genome-wide analysis of HbS associations in the newly generated GAMCC dataset. This revealed a putative *Pfsa* locus which I have named *Pfsa2B*. This locus was with ~10Kb of the previously discovered *Pfsa2* locus which is restricted to East African *Pf* populations (130). The association at *Pfsa2B* replicated in a sample of mild cases collected in Ghana, and severe cases collected in Gambia at a later time point (CP1 Gambia) (Fig. 5.3a (pg.158)). The *Pfsa2B* locus also showed elevated levels of LD to other known *Pfsa* loci, including *Pfsa1* (Pf3D7_02.v3), *Pfsa3* (Pf3D7_11.v3) and *Pfsa4* (Pf3D7_04.v3), which has been identified previously as a feature of *Pfsa*-HbS interactions (130; 131). LD between *Pfsa2B* and *Pfsa3* ranked 4th of all genome-wide LD between biallelic variants on separate chromosomes (Fig. 5.19 (pg.191)). Taken together this suggests that the *Pfsa2B* locus is associated with HbS and therefore might play a role in an ability of parasites to overcome HbS resistance (the HbS-escape phenotype).

The association of *Pfsa2B* (Pf3D7_02.v3:825357:T>C) occurs in the 5' UTR of the gene PF3D7_0220500. This gene is listed as a member of the small hypothetical gene family 2 (hyp2) (181) although it has no listed cellular function. It is also listed among 493 candidate genes associated with virulence based on clustering analysis of *Pf* and other human pathogenic malaria species (183), and was found to be differentially expressed in severe compared to uncomplicated malaria cases in an albeit modest sample of N=23 and N=21 respectively (202). Like the previously identified *Pfsa1-3* loci (130), PF3D7_0220500 has been noted to contain a *Plasmodium* export element (PEXEL) amino acid motif ('RTLNY') and to be a member of the essential exportome of *Pf* parasites, although the precise localisation of this particular protein within the host red cell has not been tested to my knowledge (197). This further suggests that the site of interaction between HbS-*Pfsa* may be within the host cytosol (130). Further datasets will be necessary to validate this signal and functional experiments needed to characterise its role in the interplay of *Pf* parasites and erythrocytes.

The work presented in this chapter further suggested that *Pfsa* associations vary by population, with new potential associations identified in East and West African datasets (Supp. Table 5.5 (pg.221)). One potential association within the gene CLAG3.1 (Pf3D7_03_v3:140090:T>A) was identified in the CP1 Kenyan population. The CLAG gene family, of which five are present in *Pf* parasites, play a wide spread role in infection biology, having been shown to mediate cytoadherence, invasion and nutrient uptake (203). Intriguingly, only one of CLAG3.1 and CLAG3.2 is expressed at a time. Despite this suggesting an overlapping function, analyses of their sequence has previously shown that CLAG3.1 and 3.2 segregate into separate clades suggesting distinct evolutionary forces are acting at these genes (204). Although suggestions of an association at this locus were observed across populations (Fig. 5.9a (pg.167)), variant calls from short-read data utilised in this chapter appear overall to be insufficient to resolve this locus due to high levels of homology between CLAG3.1 and CLAG3.2 and the complexity introduced by the presence of structural variation (166). More targetted methods such as graph-based calling or kmer-based approaches might enable a clearer elucidation of variation across the CLAG3.1 and CLAG3.2 copies, allowing a more robust assessment of the relationship of this locus to HbS.

A variant within the FIKK3 gene (Pf3D7_03_v3:79845:G>A) showed modest evidence for association with HbS in Ghana, within the same dataset that the *Pfsa4* locus within the FIKK4.2 gene was identified (131). This also showed evidence of replication in the GAMCC samples from Gambia and in the Kenyan population (Fig. 5.12a (pg.173)). The implication of a further FIKK gene in HbS-*Pfsa* interactions is compelling; there are 18 members in this expanded gene family all of which are exported into the erythrocyte to play a role in host cell remodelling (205). The FIKK gene family appears to have expanded within the *Laverania* subgenus (which contains *Pf* but not other human-infecting *Plasmodium spp.*) comprising 20 members compared to one in other Apicomplexa (206; 135). This and the location of many in the highly mutable subtelomeric chromosomal regions means they are thought to be involved specifically in human-parasite cellular interactions (207). As alluded to previously, a shared attribute of the previously identified *Pfsa* loci and many putative *Pfsa* regions described here is that many

have been identified as exported proteins, further suggesting that the interaction during the blood stages of infection is crucial to the development of *Pfsa*.

I also identified additional population-specific loci, including an association at Pf3D7_07_v3:614420:C>T in the CP1 samples from Gambia and an association at REX-2 in Kenya. However, the association in Gambia did not replicate in other tested populations. As a result, the conclusions we can draw remain limited (Fig. 5.5a (pg.162)). Further investigation is required to determine whether this represents an emerging *Pfsa* association specific to Gambia or is an artifact of confounding. The association in the REX2 gene in Kenya involved a relatively rare allele, Pf3D7_09_v3:428875:A>T ($AF^{Kenya}=6.1\%$), which was absent in all other populations analysed in this study. Additional datasets from East African populations where this allele is present are eagerly awaited, as they may allow for replication of these findings and provide more robust insights into this relationship.

Previous estimates suggested that HbS protection depends strongly on parasite genotype at the *Pfsa* loci (130). These predictions were made based on a small number of severe malaria cases in HbS carrying individuals. Three independent datasets sampled from West-African populations were recently made available for this analysis (two from Gambia from different time points and one from Ghana), meaning it was possible to compare these effects between Western populations and a single East African population (Kenya). Reviewing relative risks, stratified by parasite genotypes at the *Pfsa* loci, confirmed a strong impact of *Pfsa* on the HbS-protective effect (Fig. 5.13 (pg.177)). However, carrying *Pfsa*+ variation appeared to diminish the protective effect of HbS to a larger degree in Kenya compared to the West African populations for reasons that are yet unclear (5.13). Furthermore, the protective effect of HbS appeared much greater in Kenya than in the West African populations. This might suggest that the protective effect of HbS may vary independently of parasite genotype.

One possible factor underlying the geographical differences is the *Pfsa2* allele, which is only present in East African populations according to current population genetic data (Fig. 5.4b (pg.160)). A further factor which might contribute to these differences is the distribution of HbC in West African populations, a neighbouring polymorphism in the *HBB* gene known to offer weaker malaria protection that is at high frequency in Ghana (54). HbS frequency is known to vary considerably

between populations (156). Many factors have been posited to influence the frequency of HbS in populations including the presence of HbC (208). Individuals carrying HbC alleles were retained in the presented analysis, however, the presence of HbC did not appear to explain the presence of many *Pf*sa- infections in HbS carrying individuals in Ghana as the majority of these did not carry any HbC alleles (3/15 carried the HbSC genotype). The interplay between HbS, HbC and *Pf*sa are potentially complex and may require more careful modelling and improved power to unpick. Whether HbC or another factor accounts for these variations remains to be clarified.

An opportunity afforded by the generation of sequence data within the GAMCC set was the reappraisal of a widely-cited but rarely replicated association between the HLA-B*53 allele and protection against severe malaria (52). Given the growing evidence for parasite genetic interactions with human resistance factors, it is plausible that parasite variation may play a role in the association between HLA-B*53 and severe malaria. To investigate this, I presented a genome wide scan for associations between HLA-B*53 and parasite variation, which did not produce strong evidence for interactions between this HLA locus and the *Pf* genome (Fig. 5.18 (pg.189)). This suggested that parasite variation was unlikely to have confounded the association between HLA-B*53 and severe malaria protection. However, it should be noted that this scan was limited to only a portion of the core regions of the *Pf* genome and was limited in sample size. As a result, the analysis may have been underpowered or unable to detect true loci of interaction. Suggestions of associations between HLA alleles and *Pf* variation have been previously observed (130) and as larger datasets become available, robust evidence of interactions between HLA loci and the *Pf* genome may yet be uncovered. Overall, the evidence presented here suggests that the HLA-B*53 association is more likely to be a false positive association due to statistical underpower, or another unknown source of confounding.

Since their discovery, the distinctive population genetic characteristics of the *Pf*sa loci have become more evident. Assessing long-range LD in the GAMCC revealed that prominent correlations existed between the *Pf*sa1, *Pf*sa3 and the newly identified *Pf*sa2B loci. Intriguingly however, strong levels of LD were also observed between these sites and the *Pf*sa2 and *Pf*sa4 loci, despite current evidence

indicating that these loci are not associated with HbS in The Gambia (Fig. 5.2b (pg.155) & 5.2d (pg.155)). The lack of replication of *Pfsa4* in Gambia, which was discovered in Ghana and replicated in Mali (131), is yet to be explained. The LD signal observed in the GAMCC suggests that *Pfsa4* may be involved in interactions with HbS in this population, the detection of which may have been otherwise masked. In addition, the LD between *Pfsa3* and a locus just 9bp from the *Pfsa2* locus (Pf3D7_02_v3:814288) was striking. The *Pfsa2* locus was shown to be associated with HbS in Kenya but is absent from populations in West Africa like the Gambia (130). This result is challenging to interpret given current available data, but this might suggest that the *Pfsa2* locus is playing a role within western populations which has not been detected due to the complexity of this mutation.

Inspecting haplotypes at the *Pfsa1-4* regions indicated they have a shared evolutionary origin (Fig. 5.22 (pg.198)). In our study we observed a significant increase in the frequency of alleles at the *Pfsa1*, *Pfsa3*, and *Pfsa2B* loci between the GAMCC dataset, collected in 1988-90, and more recent samples from the Pf7 dataset (2013-17) which might be suggestive of a selective sweep in The Gambia (Fig. 5.21 (pg.195)). This trend contrasts with the Kenyan population, where *Pfsa* allele frequencies appear to have remained stable (130). However, my analysis of haplotype homozygosity at these loci using the iHS metric, which allows the inference of the occurrence of a recent selective sweep, indicated that this has not occurred in The Gambia, nor in Kenya or Ghana (Fig. 5.24 (pg.201)). An alternative possibility therefore is that these loci are under forces of balancing selection which cause them to be maintained within populations at stable intermediate frequencies.

One possible explanation for the allele frequency shift in The Gambia, consistent with this model, is that a change in selection occurred, acting on pre-existing standing variation at the *Pfsa+* mutations. If multiple related *Pfsa+* haplotypes were already present within Gambia with a limited extent of sharing due to ancestral recombination events, as could be the case if these loci are under forces of balancing selection, all may have increased in frequency simultaneously. This would likely not result in a strong iHS signal (209). Further population genetic analyses would be needed to investigate this and other possibilities.

This raises questions as to what might have caused a change of frequencies in The Gambia. The malaria landscape in this region has undergone substantial changes since the collection of the GAMCC samples including the spread of CQ resistance, which has led to marked allele frequency shifts surrounding the CRT (on Pf3D7_07_v3) and amino acid transporter 1 (on Pf3D7_06_v3) genes (Fig. 4.5 (pg.121) in Chapter 4). Notably in The Gambia, CQ resistance alleles in the CRT gene exhibit moderate LD with *Pf*sa3+ alleles (ranking 9th of all genome wide between chromosome LD signals in the Gambia published for the CP1 dataset (130)), suggesting that *Pf*sa alleles may have, in some unexplained manner, become associated with this locus, leading to concurrent changes in their frequencies. Additionally, the intensity of malaria transmission has significantly declined in The Gambia, leading to altered parasite genetics in this region (108; 180; 9). Reduced transmission intensity might diminish the strength of competition between parasites, increase levels of inbreeding within populations and reduce effective population size (106; 105), which in turn may have altered the evolutionary dynamics acting on these loci.

The results presented in this chapter highlight the significant knowledge gaps regarding human-malaria parasite interactions. Resolving this can only be achieved if large, cross population datasets are generated which enable investigations into this area. A better understanding of the host-parasite interplay will offer valuable insights into the evolution of malaria parasites and may aid in identifying strategies to mitigate their harmful effects.

5.A Supplementary Figures and Tables

5.A.1 Supplementary Figures

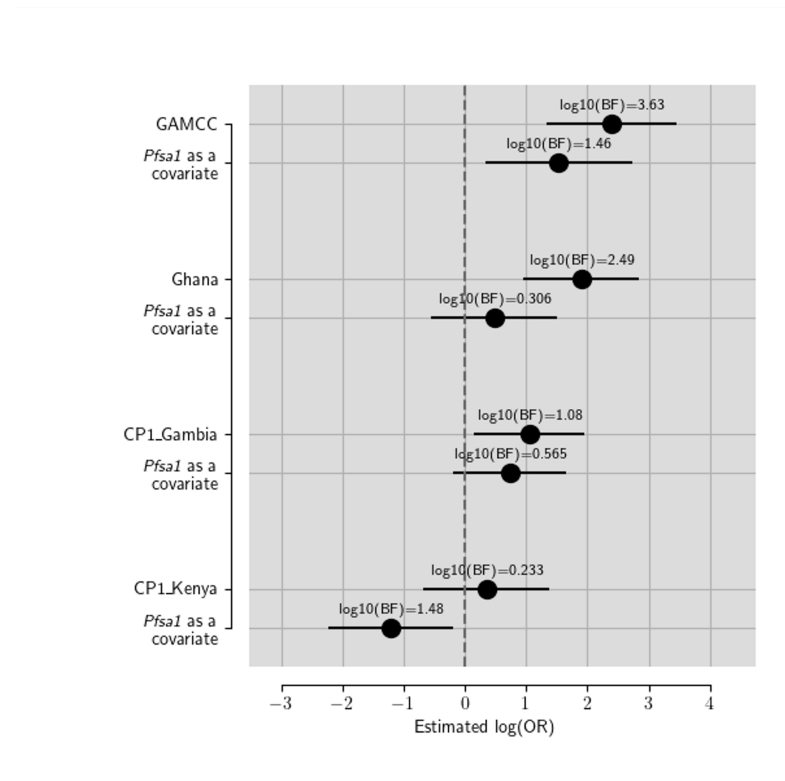


Figure 5.25: **Estimated effect sizes of associations between *Pfsa2B* and HbS across four datasets conditioned of genotype at *Pfsa1*.** Replication of evidence across the four studies with and without including *Pfsa1* genotype as a covariate in the analysis (indicated on the y-axis), indicating the effect size as a log(odds ratio (OR)) and showing 95% confidence interval (CI)s as lines. Strength of evidence (log₁₀Bayes Factor (BF)) are shown at each point.

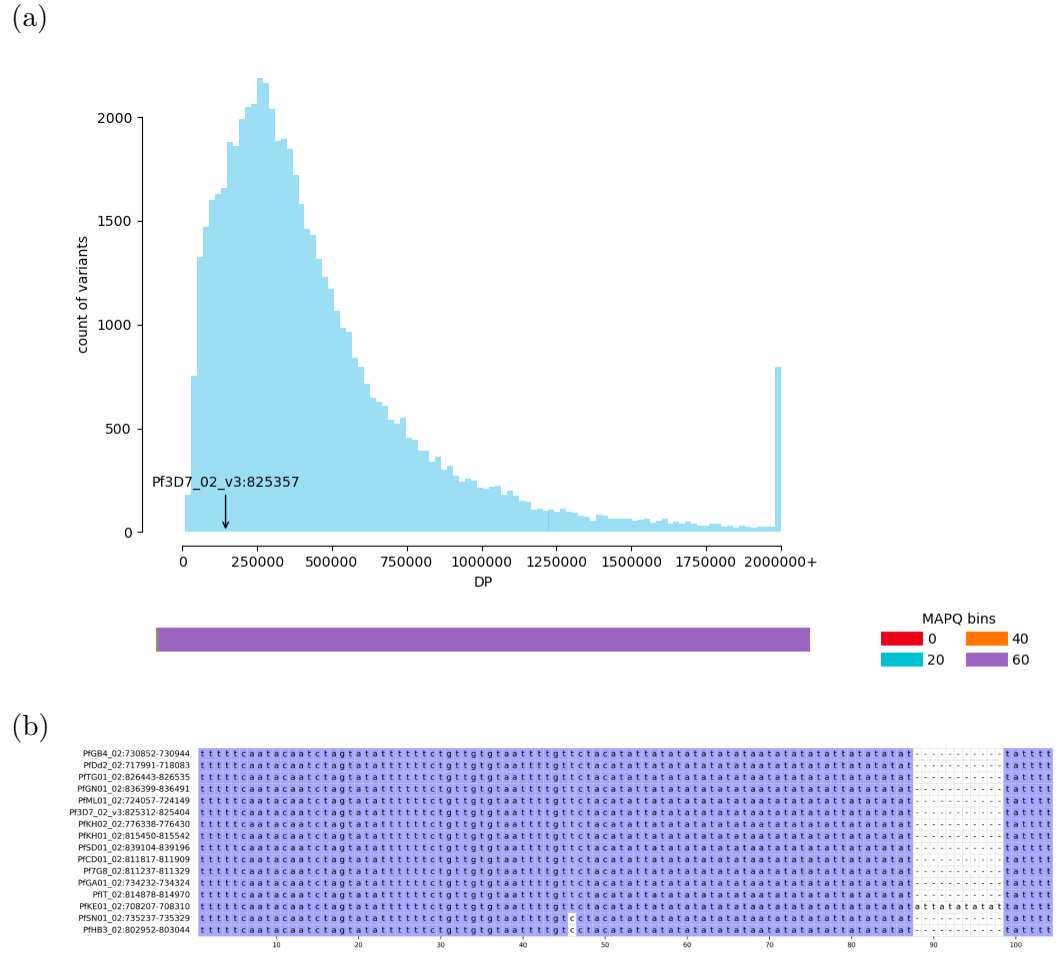


Figure 5.26: **Assessment of quality at the *Pfsa2B* (Pf3D7_02_v3:825357:T>A) variant.** (a) shows the distribution of depth (DP) values for all VQSR pass variants in the combined GAMCC/CP1 call set generated in this thesis. The DP value for the tested variant is indicated by an arrow. Variants with too high DP values indicate they reside in a repetitive region and too low indicates there may be insufficient data to type the variant. The bottom colour bar indicates the proportion of reads in the GAMCC set which have a mapping quality (MAPQ) score of 0-20 (red), 20-40 (blue), 40-60 (orange) or 60+ (purple). MAPQ score indicates the probability that a read aligns elsewhere within the reference genome. (b) Shows alignments for a subset of long read *Pf* PacBio assemblies generated by Otto et al. (103). A 101bp region surrounding the lead variant was blasted within assemblies and extracted from a multiple sequence alignment.



Figure 5.28: Visualisation of PSV sites between the CLAG3.1 and CLAG3.2 genes in the *Pf* genome using the visualisation software zoomsa. Pf3D7_03_v3:140060:T>A (in CLAG3.1) and its paralogous site in CLAG3.2 (Pf3D7_03_v3:124135:T>A) in whole genome assemblies. The HB3 clone, for example, carries an A at Pf3D7_03_v3:124135:T>A and a T at Pf3D7_03_v3:140060:T>A and is therefore a paralogous sequence variant (PSV).

(a)

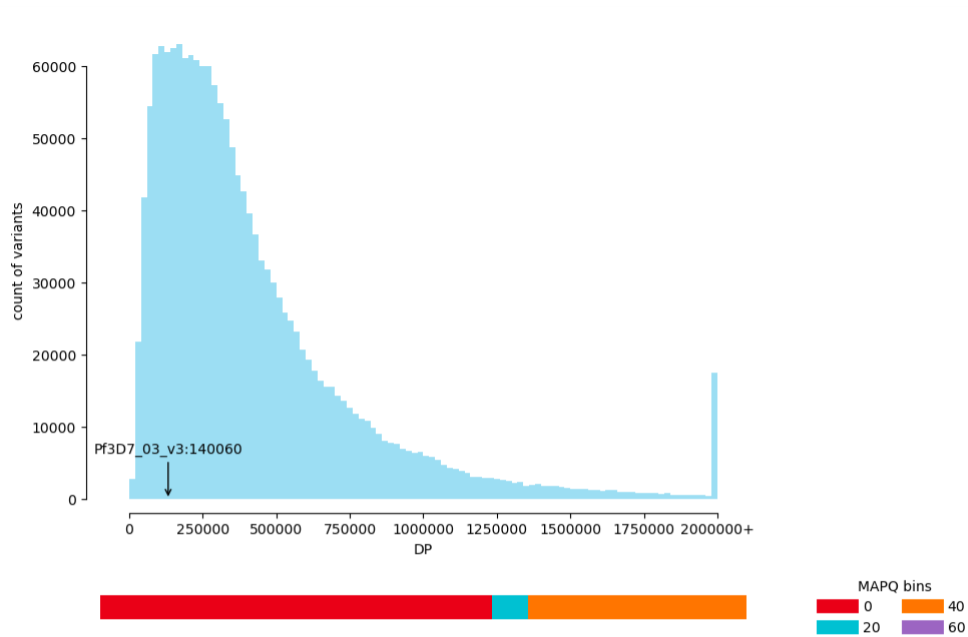


Figure 5.29: Assessment of quality of Pf3D7_03_v3:140060:T>A variant. (a) shows the distribution of depth (DP) values for all VQSR pass variants in the combined GAMCC/CP1 call set generated in this thesis. The DP value for the tested variant is indicated by an arrow. Variants with too high DP values indicate they reside in a repetitive region and too low indicates there may be insufficient data to type the variant. The bottom colour bar indicates the proportion of reads in the GAMCC set which have a mapping quality (MAPQ) score of 0-20 (red), 20-40 (blue), 40-60 (orange) or 60+ (purple). MAPQ score indicates the probability that a read aligns elsewhere within the reference genome.

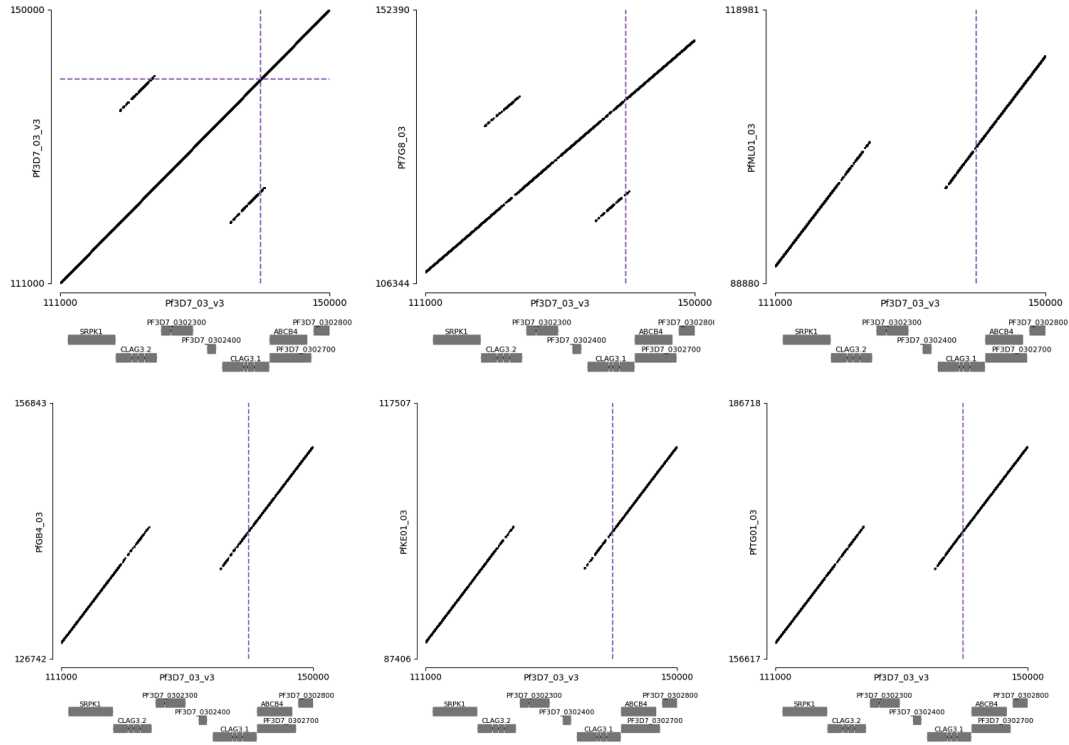


Figure 5.30: **Kmer sharing across the CLAG3.2 and CLAG3.1 locus in a number of whole genome assemblies.** An assembly (on the y-axis) compared to the Pf3D7 reference assembly (x-axis) at the CLAG3 locus. 51bp were extracted across the Pf3D7_03.v3:111000-150000 region. If this kmer is present in the partner assembly then a point is marked. Homology between CLAG3.1 and CLAG3.2 is shown by multiple points at that position (e.g. Pf3D7 and Pf7G8). Several samples have a single hybrid copy of CLAG3.1 and 3.2, leading to the deletion of the intermediary genes (e.g. ML01, GB4, KE01, TG01). The position of Pf3D7_03.v3:140060:T>A is shown as a dashed purple line.

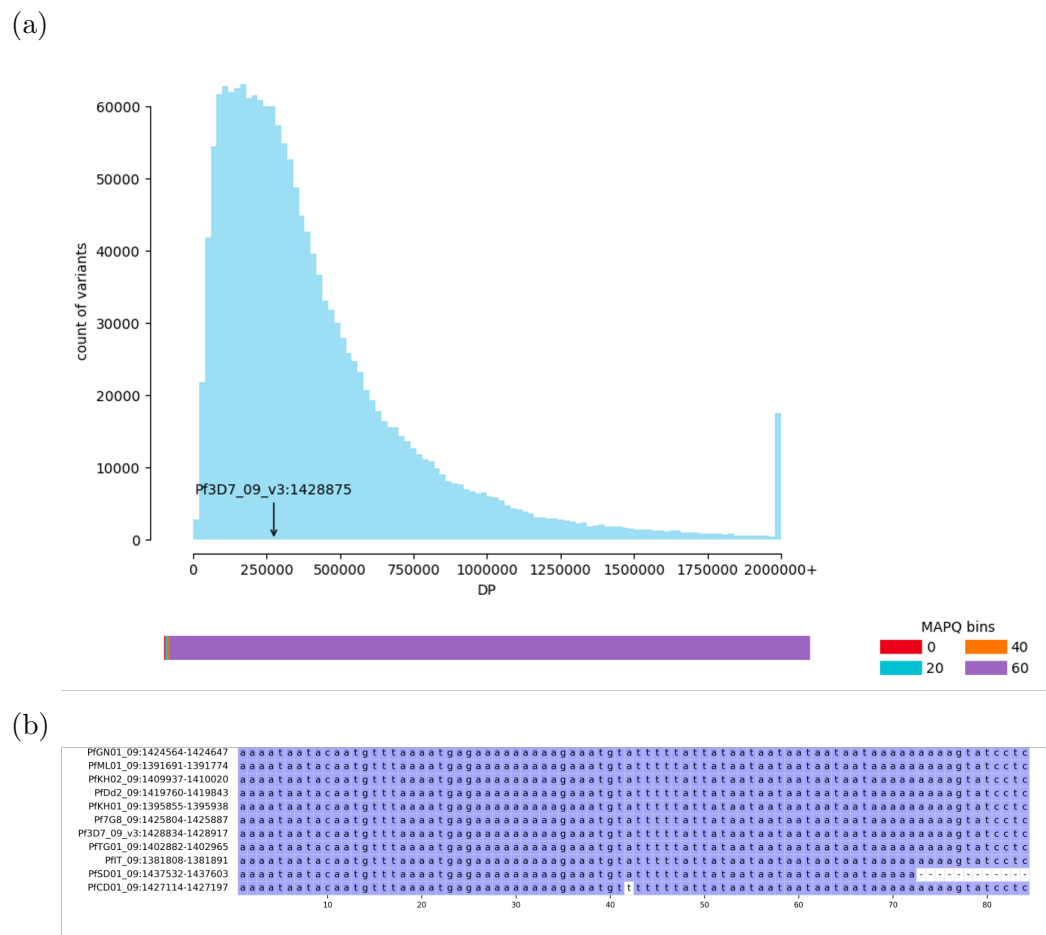
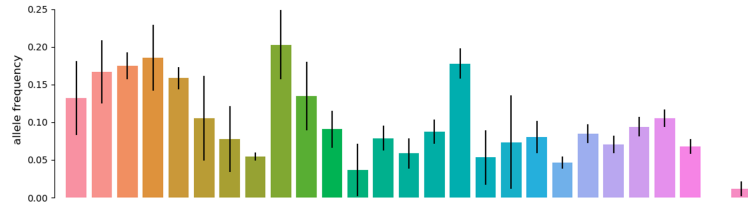
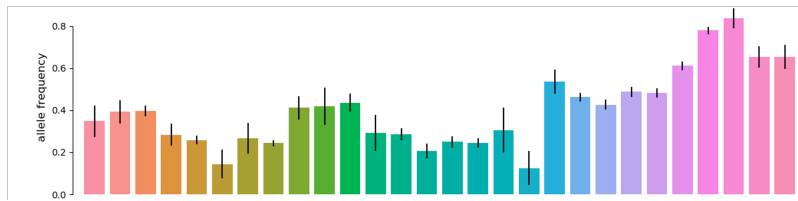


Figure 5.31: Assessment of quality of Pf3D7_09_v3:1428875:A>T variant. (a) shows the distribution of depth (DP) values for all VQSR pass variants in the combined GAMCC/CP1 call set generated in this thesis. The DP value for the tested variant is indicated by an arrow. Variants with too high DP values indicate they reside in a repetitive region and too low indicates there may be insufficient data to type the variant. The bottom colour bar indicates the proportion of reads in the GAMCC set which have a mapping quality (MAPQ) score of 0-20 (red), 20-40 (blue), 40-60 (orange) or 60+ (purple). MAPQ score indicates the probability that a read aligns elsewhere within the reference genome. (b) Shows alignments for a subset of long read *Pf* PacBio assemblies generated by Otto et al. (103). A 101bp region surrounding the lead variant was blasted within assemblies and extracted from a multiple sequence alignment.

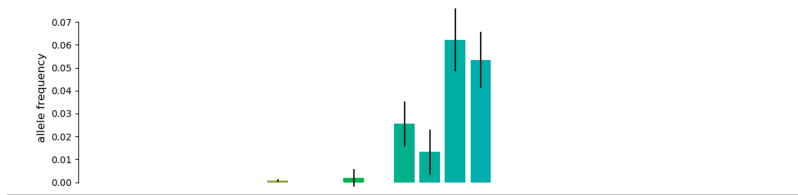
(a) Pf3D7_07_v3:614420:C>**T**



(b) Pf3D7_03_v3:140060:T>**A**



(c) Pf3D7_09_v3:1428875:A>**T**



(d) Pf3D7_03_v3:79845:G>**A**

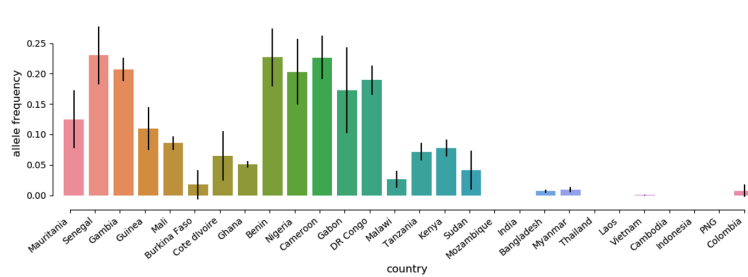
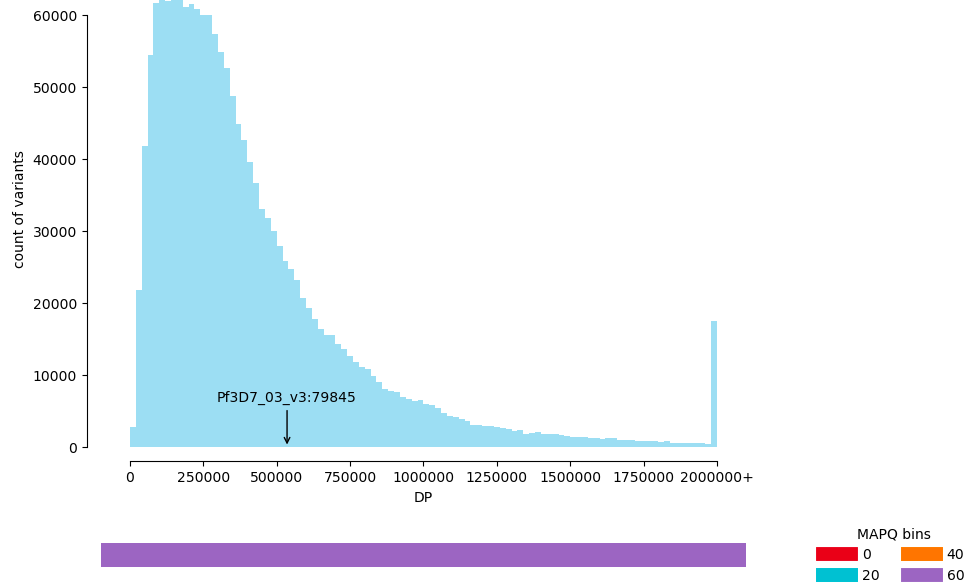


Figure 5.32: Allele frequencies of potential *Pfsa* variants within populations in the MalariaGEN Pf7 resource. Allele frequency of the effect alleles at potential *Pfsa* loci identified using a cross population analysis within different populations in the MalariaGEN Pf7 resource. Only populations with more than 50 samples are shown. Countries are shown in geographical order roughly from West to East. 95% CI error bars are show. Within the caption for each figure, the effect allele is highlighted in bold.

(a)



(b)

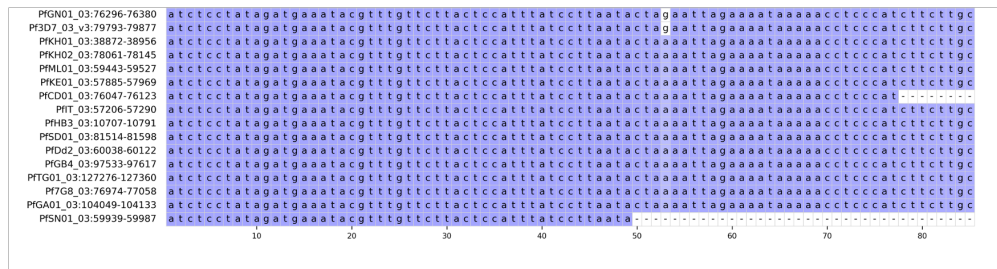


Figure 5.33: **Assessment of quality of Pf3D7_03_v3:79845:G>A variant.**

(a) shows the distribution of depth (DP) values for all VQSR pass variants in the combined GAMCC/CP1 call set generated in this thesis. The DP value for the tested variant is indicated by an arrow. Variants with too high DP values indicate they reside in a repetitive region and too low indicates there may be insufficient data to type the variant. The bottom colour bar indicates the proportion of reads in the GAMCC set which have a mapping quality (MAPQ) score of 0-20 (red), 20-40 (blue), 40-60 (orange) or 60+ (purple). MAPQ score indicates the probability that a read aligns elsewhere within the reference genome. (b) Shows alignments for a subset of long read *Pf* PacBio assemblies generated by Otto et al. (103). A 101bp region surrounding the lead variant was blasted within assemblies and extracted from a multiple sequence alignment.

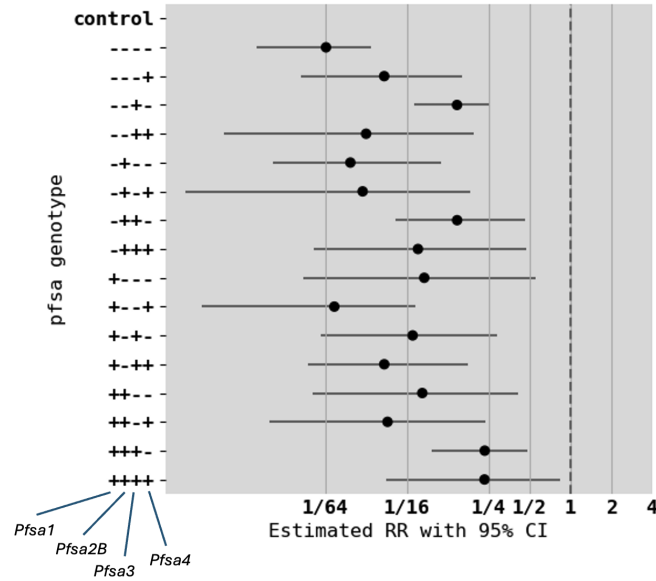


Figure 5.34: **Estimated relative risk of severe malaria for HbS carriers in infections of different *Pfsa* genotypes in two datasets from Gambia.** This is a variation of Figure 5.13 (pg.177) but Ghana has been excluded. RR is stratified by infecting malaria parasite genotype at the *Pfsa* loci. This version uses the GAMCC and Gambian CP1 samples for the *Pfsa*1, *Pfsa*2B, *Pfsa*3 and *Pfsa*4 alleles. RRs were calculated using a multinomial logistic regression using non-malaria controls as a baseline. To avoid over-fitting of points with small sample sizes, Bayesian regularisation was performed using a weakly informative Gaussian prior with mean 0 and standard deviation of 2 for each parameter, and between-parameter correlation set to 0.5. A ‘+’ for the *Pfsa* genotype indicates possessing the effect allele.

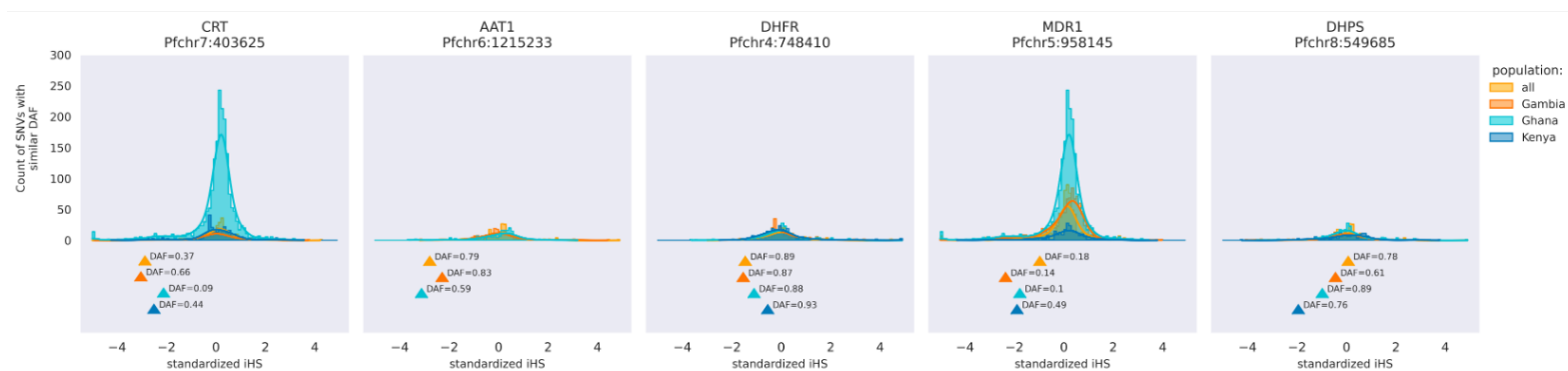


Figure 5.35: **iHS results for for known drug resistance loci in samples from the MalariaGEN Pf7 resource.** Results for drug resistance loci in the CRT, amino acid transporter 1 (AAT1), dihydrofolate reductase (DHFR), multidrug resistance 1 (MDR1) and dihydropteroate synthetase (DHPS) genes across the entire filtered dataset (“all”), within Gambia, Ghana and Kenya are shown. Histograms show the disributions of iHS values for all variants within the same AF bin (2.5% width), and arrows indicate the iHS value for each target SNP (positions are listed in Supp. Table 4.2 (pg.147)).

5.A.2 Supplementary Tables

Population	Dataset	Odds Ratio	Standard Error	Lower Bound	Upper Bound	Bayes Factor	Sample Size
Previously identified <i>Pfsa</i> loci							
Pfsa1 (ACS8, PF3D7_0215300) : Pf3D7_02_v3:631190:T>A							
Gambia	GAMCC	9.08	1.78	2.93	28.2	653.39	755
Gambia	CP1	4.2	1.54	1.79	9.84	240.50	2058
Kenya	CP1	39.44	1.57	16.35	95.13	8.78E+21	1774
Ghana	Pf7	18.93	1.48	8.76	40.92	1.25E+09	1282
Pfsa2 (PF3D7_0220300) : Pf3D7_02_v3:814288:C>T							
Gambia	GAMCC						
Gambia	CP1	0.73	1.84	0.22	2.42	3.24	2198
Kenya	CP1	17.67	1.47	8.34	37.46	1.12E+15	1682
Ghana	Pf7						
Pfsa3 (PF3D7_1127000) : Pf3D7_11_v3:1058035:T>A							
Gambia	GAMCC	7.44	1.77	2.43	22.75	275.87	744
Gambia	CP1	27.61	2.68	3.99	191.3	3.07E+06	1984
Kenya	CP1	27.25	1.53	11.89	62.48	1.27E+18	1752
Ghana	Pf7	25.08	1.45	12.07	52.1	9.86E+12	1277
Pfsa4 (FIKK4.2, PF3D7_0424700) : Pf3D7_04_v3:1122147:A>C							
Gambia	GAMCC	0.58	1.88	0.17	1.98	2.37	739
Gambia	CP1	0.62	1.58	0.25	1.51	1.95	2036
Kenya	CP1	0.66	1.45	0.32	1.36	1.69	1622
Ghana	Pf7	0.11	1.55	0.05	0.26	8530.02	1255
New associations							
Putative locus <i>Pfsa2B</i> (PF3D7_0220500) : Pf3D7_02_v3:825357:T>C							
Gambia	GAMCC	10.96	1.72	3.77	31.92	4239.26	747
Gambia	CP1	2.86	1.58	1.16	7.04	11.93	1972
Kenya	CP1	1.42	1.69	0.5	3.98	1.71	1799

Population	Dataset	Odds Ratio	Standard Error	Lower Bound	Upper Bound	Bayes Factor	Sample Size
Ghana	Pf7	6.68	1.62	2.61	17.12	310.94	1254
Gambia specific association within CP1 (PF3D7_0713500) : Pf3D7_07_v3:614420:C>T							
Gambia	GAMCC	0.54	2.06	0.13	2.23	3.40	780
Gambia	CP1	10.5	1.73	3.6	30.61	229134.24	2058
Kenya	CP1	1.82	1.47	0.86	3.85	2.77	1648
Ghana	Pf7	1.35	2	0.35	5.27	2.34	1239
Potential cross population association at CLAG3.1 (PF3D7_0302500) : Pf3D7_03_v3:140060:T>A							
Gambia	GAMCC	0.31	1.88	0.09	1.08	6.91	482
Gambia	CP1	0.38	1.64	0.14	1.01	6.80	1656
Kenya	CP1	0.14	1.47	0.07	0.3	167795.38	1368
Ghana	Pf7						
Kenya specific association at REX2 (PF3D7_0936000) : Pf3D7_09_v3:1428875:A>T							
Gambia	GAMCC						
Gambia	CP1						
Kenya	CP1	10.09	1.48	4.66	21.84	177533.37	1828
Ghana	Pf7						
Marginal association in Ghana at FIKK3 (PF3D7_0301200) : Pf3D7_03_v3:79845:G>A							
Gambia	GAMCC	0.23	1.89	0.07	0.81	11.08	756
Gambia	CP1	1.09	1.79	0.35	3.41	1.66	2096
Kenya	CP1	0.17	1.65	0.06	0.46	89.91	1838
Ghana	Pf7	0.09	1.6	0.04	0.23	7951.15	1263

Table 5.5: *Pfsa*-HbS association statistics across tested datasets at implicated loci.

Chromosome A					Chromosome B					Correlation		
Chr	Region	Lead Position	AF	Genes in Range	Chr	Region	Lead Position	AF	Genes in Range	Average R Across Range	Lead R	Lead N
chr2	629375-631190	631092	0.096	ACS8 (PF3D7_0215300)	chr11	1054282-1065275	1057437	0.155	SNRPF (PF3D7_1127000), PF3D7_1127000	0.536	0.589	747
chr4	1119762-1122147	1122147	0.908	FIKK4.2 (PF3D7_0424700)	chr11	1054282-1065275	1057437	0.155	SNRPF (PF3D7_1127000), PF3D7_1127000	-0.389	-0.472	724
chr2	814297-814297	814297	0.094	PF3D7_0220300	chr11	1054282-1065275	1054282	0.157	SNRPF (PF3D7_1127000), PF3D7_1127000	0.419	0.449	709
chr2	825357-825357	825357	0.083	PF3D7_0220500	chr11	1054282-1065275	1057437	0.155	SNRPF (PF3D7_1127000), PF3D7_1127000	0.403	0.426	730
chr2	825357-825357	825357	0.083	PF3D7_0220500	chr4	1122147-1122147	1122147	0.908	FIKK4.2 (PF3D7_0424700)	-0.370	-0.370	719
chr2	631092-631190	631092	0.096	ACS8 (PF3D7_0215300)	chr4	1119762-1122147	1119762	0.916	FIKK4.2 (PF3D7_0424700)	-0.364	-0.368	730
chr2	845721-845721	845721	0.053	PF3D7_0220900	chr11	191902-191902	191902	0.050	PF3D7_1104300	0.351	0.351	747

Table 5.6: Regions with highest values of between chromosome linkage disequilibrium in the GAMCC dataset.

Country	N	Freq	iHS	p-value
Pfsa1 (Pf3D7_02_v3:631190)				
all	4788	0.23	-0.044	0.965
Gambia	574	0.413	0.167	0.867
Ghana	1582	0.101	0.525	0.600
Kenya	401	0.197	-1.111	0.267
Pfsa2 (Pf3D7_02_v3:814288)				
all	4788	0.072	-0.187	0.852
Gambia	574	0.0	NA	NA
Ghana	1582	0.0	NA	NA
Kenya	401	0.224	-0.016	0.987
Pfsa3 (Pf3D7_11_v3:1057437)				
all	4788	0.305	-0.406	0.685
Gambia	574	0.650	0.055	0.956
Ghana	1582	0.116	-0.140	0.888
Kenya	401	0.277	-1.143	0.253
Pfsa4 (Pf3D7_04_v3:1121472)				
all	4788	0.158	-0.469	0.639
Gambia	574	0.326	-0.057	0.955
Ghana	1582	0.101	0.184	0.854
Kenya	401	0.0	NA	NA
Pfsa2B (Pf3D7_02_v3:825357)				
all	4788	0.188	0.268	0.789
Gambia	574	0.382	0.120	0.904
Ghana	1582	0.088	0.707	0.479
Kenya	401	0.125	0.642	0.521
CRT (Pf3D7_07_v3:403625)				
all	4788	0.367	-2.874	0.004
Gambia	574	0.657	-3.037	0.002
Ghana	1582	0.087	-2.118	0.034
Kenya	401	0.439	-2.506	0.012
AAT1 (Pf3D7_06_v3:1215233)				
all	4788	0.785	-2.786	0.005
Gambia	574	0.829	-2.285	0.022
Ghana	1582	0.594	-3.118	0.002
Kenya	401	0.0	NA	NA
DHFR (Pf3D7_04_v3:748410)				
all	4788	0.885	-1.455	0.146
Gambia	574	0.868	-1.539	0.124
Ghana	1582	0.884	-1.101	0.271

Country	N	Freq	iHS	p-value
Kenya	401	0.928	-0.549	0.583
MDR1 (Pf3D7_05_v3:958145)				
all	4788	0.178	-0.971	0.331
Gambia	574	0.141	-2.358	0.018
Ghana	1582	0.098	-1.781	0.075
Kenya	401	0.486	-1.892	0.059
DHPS (Pf3D7_08_v3:549685)				
all	4788	0.776	0.071	0.943
Gambia	574	0.611	-0.445	0.656
Ghana	1582	0.892	-0.980	0.327
Kenya	401	0.763	-1.950	0.051

Table 5.7: **integrated haplotype scores (iHS) for *Pf*sa and drug resistance loci in samples from the MalariaGEN Pf7 resource.** Results across the dataset (all), within Gambia, Ghana and Kenya are shown.

Chapter 6

Structural variation at a sickle-associated locus of the *Plasmodium falciparum* genome

6.1 Introduction

Carriers of the haemoglobin (Hb)S allele, a point mutation in the *beta-globin* (*HBB*) gene, show increased resistance to *Plasmodium falciparum* (*Pf*) malaria infection (71). The recent discovery of the association between regions of the *Pf* genome and the HbS polymorphism suggests that parasite genetic variation contributes to a yet-uncharacterised ability of parasites to overcome this resistance (an ‘HbS-escape’ phenotype) (130). Understanding this phenotype and determining whether it can be pharmacologically exploited, has therefore become a research priority within the field. Determining the biological function of these regions, termed *Plasmodium falciparum* sickle associated (*Pfsa*), will require functional experiments, the design of which requires a detailed understanding of the genetic landscape of the associated regions.

In this chapter, I focus on one of the *Pfsa* regions, the *Pfsa3* locus on chromosome 11 (Pf3D7_11.v3). The most significantly HbS-associated (*Pfsa*+) single nucleotide polymorphism (SNP) at the *Pfsa3* locus (Pf3D7_11.v3:1058035:T>A) falls within an uncharacterised gene PF3D7_1127000 which has been predicted *in silico* to be a putative phosphatase (197). Evidence suggesting two possible functional mechanisms underlying the association between the *Pfsa3* region and

HbS were outlined in the original study. Firstly, it was noted that one of the two lead *Pfsa*+ SNPs at the *Pfsa3* locus (Pf3D7_11.v3:1057437:T>C) lies immediately downstream of *Plasmodium* export element (PEXEL)-like motif, causing a non-synonymous substitution in the immediately neighbouring amino acid in the expressed protein (130). PEXEL motifs play an important role in malaria infection mediating the transport of parasite expressed proteins out of the parasitophorous vacuole and into the erythrocyte cytosol (197). The base-pair change at this position could plausibly alter the transport of the protein expressed by PF3D7_1127000, affecting cellular function in a way which contributes to the HbS-escape phenotype. However, the export of Pf3D7_1127000 protein has not, to our knowledge, been observed experimentally, so the true impact of this mutation is not straightforward to predict. Secondly, the PF3D7_1127000 gene appeared to be differentially over-expressed within the trophozoite-stage of malaria for the Ugandan Palo-Alto (FUP-H) isolate which carries *Pfsa*+ variants compared to a *Pfsa*- isolate (130). It might therefore be the case that increased expression, rather than an alteration to protein export, may be responsible for the HbS-escape phenotype.

In addition to these possible mechanisms, it was noted that structural variation exists at the *Pfsa3* locus in 15 parasites for which genome assemblies have been generated using the Pacific Biosciences (PacBio) long-read approach (130). In this small sample, *Pfsa*+ SNP genotypes were perfectly associated with the observed structural variation. Although this sample is insufficient to be certain, this raises the possibility that *Pfsa*+ mutations are linked to structural variant haplotypes, which might suggest that this structural variation contributes to the HbS-escape phenotype. This would offer insight into the functional basis of the *Pfsa3* association, which could eventually aid in uncovering the underlying mechanism.

Understanding the relative importance of structural variation and SNP variation to the association of *Pfsa3* locus to HbS is therefore of considerable interest. Motivated by this, I set out to investigate genome structure and genetic variation at the locus in depth using both newly generated and publicly available DNA and RNA sequencing datasets. Specifically, I determined to develop methods to identify structural variation at the *Pfsa3* locus within whole genome sequencing (WGS) data, examine the distribution of these haplotypes across global populations, ascertain whether the *Pfsa*+ variants at the *Pfsa3* locus are genuinely linked to the

described structural variant (SV) haplotypes and assess the impact of the SVs on the genetic landscape in the region.

6.2 Results

6.2.1 Using coverage and split reads to infer structural variant genotypes from short-read *Pf* genome sequencing

Firstly, we sought to generate a more thorough catalogue of structural variation at the *Pfsa3* locus present within global parasite populations. A wealth of publicly available *Pf* short-read Illumina sequencing data provides an ideal resource for this analysis. Unlike from long-read data, alterations in the structure of the genome are more technically challenging to infer from short-reads. WGS long-reads can often be assembled *de novo*, allowing the structure of the genome to be determined. In contrast, the process of short read sequencing generates large volumes of data, often containing errors, which are aligned to a reference assembly. If the structure of the reference differs from the sample, several artifacts may occur within alignments. These include read pairs aligning unusually far apart or too close together or in the incorrect orientation relative to one another. Reads may also align with low confidence, be partially aligned (resulting in “clipped” reads), or fail to map altogether (101). Although indicative of structural variation, these can be challenging to differentiate from sequencing error or noise.

Methods for SV calling in short-read data will often utilise changes in the fold coverage across duplicated or deleted regions of the reference genome (210). A prerequisite of these methodologies is that coverage should be broadly uniform across the region of interest so that changes can be differentiated. This is not the case for short-read data produced from malaria parasites which is often highly variable in terms of coverage (95). In light of this, we developed custom methods to call SVs at the *Pfsa3* locus from a set of haplotypes based on coverage (the coverage caller), and a second distinct source of information; insert size distribution (insert-size caller). These methods will be detailed below. These callers were incorporated into an iterative process, during which I examined read data from all samples with

inconsistent or uncertain calls using Integrative Genomics Viewer (IGV). This analysis was then used to refine and expand the initial set of SV haplotypes for further calling.

To develop this process, I focussed on *Pf* WGS data sequenced without the use of the amplification step known as selective whole genome amplification (sWGA). As explored in Chapter 3 (pg.87), samples sequenced following sWGA are more affected by guanine-cytosine (GC)-biasing than non-amplified controls leading to more uneven genome coverage, which might impede method development (Fig. 3.2 (pg.92)). For the initial stages of development of this caller, we further restricted samples to those within the Malaria Genomic Epidemiology Network (MalariaGEN) Pf6 dataset from the Democratic Republic of Congo (DRC) (N=344). DRC was selected because it has a high frequency of the *Pf*sa+ variant (allele frequency (AF)>60% at *Pf*sa3) and so was potentially more likely to contain higher frequencies of SVs (130). Following this, we expanded the analysis to include additional populations and datasets, as will be detailed later.

6.2.1.1 The coverage caller

The coverage caller, which was developed by Dr. Gavin Band, iterates across sites to compare an individual sample's coverage to a set of possible predicted coverage profiles. For these profiles, we initially input the two SV haplotypes observed in PacBio assemblies within the original study (here termed 'DUP_B1' and 'DUP_B2' (Fig. 6.1 (pg.232))) (130). To overcome residual levels of uneven genome coverage present within non-sWGA data, the coverage caller employed the use of a per-site normalisation factor, calculated iteratively across the assessed sample set (for details of this caller see Materials and Methods section 2.10.2.2 (pg.81)). This approach generated a normalized coverage value for each sample at each site, accounting for coverage variation. Subsequently, a likelihood for each possible coverage profile was computed using a Gaussian model that described the effect of copy number on coverage, resulting in a posterior estimate of the SV present in the sample.

6.2.1.2 The insert size caller

To increase reliability in our calls, I implemented a second distinct caller termed the ‘insert size caller’. Based on a secondary, independent information source, this method modelled the base pair distance between two reads within a single read pair (known as the insert size). Read pairs sequenced from structurally variant samples align to the reference genome at the variable region with longer or shorter insert sizes than at non-variable sites due to the presence of deletions and duplications. In order to exploit this facet of the data, I realigned reads from the *Pfsa3* region to “dummy” reference genomes which had been rearranged to mimic the SV haplotypes discovered using the coverage caller. The likelihood for the set of reads on each dummy reference was then calculated, and the dummy reference with the maximum likelihood was determined to most closely reflect the structure of the sample. For a complete description of the mathematical basis of this method see Materials and Methods section 2.10.2.3 (pg.83).

6.2.1.3 Inspecting discordant samples and discovering new SVs

By plotting the normalized coverage generated by the coverage caller across the *Pfsa3* region, we visually inspected samples with low-confidence calls. This analysis revealed that the coverage profiles of certain samples did not align well with their predicted call profiles. Consequently, we manually examined the read alignments for these samples using IGV software (211), which uncovered additional SV haplotypes within the dataset (see Materials and Methods section 2.10.2.2 (pg.81) for further detail). These newly identified SVs were then incorporated into the coverage and insert size callers, and the calling process was repeated. This integration would result in higher confidence statistical determinations from both callers for these samples, building confidence in these new structures. Eventually, 10 break-points were found to be implicated in different SVs, fragmenting the 3D7 reference assembly within the region into 11 sections (Supp. Table. 6.3 (pg.257)). These have been named ‘R1’ to ‘R11’, which are differentially arranged within each SV haplotype.

6.2.1.4 Generating a large set of consensus SV calls

In order to generate a large set of structural variant calls for further analysis, we expanded the callers to the wider MalariaGEN Pf6 resource of infections from multiple countries (N=5380) (167) and a set of non-sWGA samples in the Malaria-GEN Consortial Project 1 (CP1) dataset of severe cases from Gambia and Kenya (N=518) (for details of sample filtering see Materials and Methods section 2.10.2.1 (pg.81)). Overall, we observed a high concordance between our callers in these samples. 84% of SV haplotype calls generated using the coverage and insert-size call methods matched within the CP1 samples (based on all haplotypes shown in Fig. 6.1 (pg.232)). DUP-B2 was only distinguishable from DUP-B1 by a small 737bp duplication (R9 and R10 (Fig. 6.1 (pg.232))) which causes an increase in the factor of relative fold coverage over this length from 2 to 3 times. This was likely to be difficult for the coverage caller to identify over noise within the sequencing data. Therefore, although we acknowledge the presence of both the DUP-B1 and DUP-B2 haplotypes circulating within populations, within this call set we opted to combine these haplotypes into a single DUP-B category to account for the limitations of the calling methods. In doing so, the concordance between callers increased to 94.6% (Supp. Table. 6.4 (pg.258)). Bolstered by this, using the intersection of calls from the coverage and insert size callers, we generated a robust consensus set of calls for a combined dataset of 5898 samples (henceforth referred to as the ‘consensus call set’).

A factor which may contribute to the remaining level of discordance between the two calling methods is the presence of complex infections. Complexity of infection (COI) occurs when a host is infected by multiple genetically distinct parasite clones (163). In cases where a single infection contained parasites with differing structural variant haplotypes, the callers may have been unable to distinguish between them, leading to the mismatch. However, the high level of concordance observed between the two independent callers strongly suggested that structural variation had been reliably identified within the short-read data, as it was unlikely that both callers would produce the same erroneous result.

6.2.2 A catalogue of *Pfsa3* structural variant types segregating in global populations

The catalogue of the extensive structural variation, discovered at the *Pfsa3* locus using the coverage and insert size callers, can be seen in Figure 6.1 on the following page. The previously identified DUP_B1 and DUP_B2 haplotypes were observed in the consensus call set at moderate frequencies (AF=11.8%). However, in addition to DUP_B1 and DUP_B2, we observed a rarer novel duplication haplotype which we termed DUP_F (AF=1.2%). DUP_F appears to be ancestrally related to the DUP_B haplotypes because it shares the duplication of a specific segment (segment R2 in Fig. 6.1 on the next page). Inspecting these haplotypes on a population-by-population basis, frequency of these haplotypes varied considerably, and within many African populations structural variant haplotypes were common (Fig. 6.2 (pg.233)). Interestingly, the DUP-B SV haplotypes were distributed across the continent whereas DUP_F was restricted to countries in East Africa. Three genes exist locally to *Pfsa3*: small nuclear ribonucleoprotein polypeptide F (SNRPF), PF3D7_1127000 (which contains the most associated *Pfsa3*+ SNPs) and dUTPase. The duplication of the R2 segment shared by the DUP haplotypes appears to bisect the gene SNRPF, but the PF3D7_1127000 gene remains intact (Fig. 6.1 on the next page). Structural variation was found to be absent in *Pf* genomes sampled within Asia, but DUP_B type structural variants were found outside of the African region within Colombia, South America. Sample sizes within this region were small (N=15), but recent studies of *Pf* genetic data acquired from ancient human remains have suggested that African *Pf* lineages were likely introduced to the Americas from Africa during the trans-Atlantic slave trade (212).

Alongside the DUP haplotypes, we discovered 5 previously unobserved haplotypes containing deleted segments (DEL_1-5). In the consensus set, DEL-1 and DEL-3 were each observed 10 (AF=0.17%) and 3 (AF=0.05%) times respectively which is strongly suggestive of their circulation within populations. DEL-3 notably deletes the entirety of the gene PF3D7_1127000, suggesting this gene is not essential to establish infections *in vivo*. DEL-5 was observed once whereas DEL-2 and DEL-4 were not called with a consensus between the coverage and insert size callers. Although we inferred their existence based on IGV inspection, it's

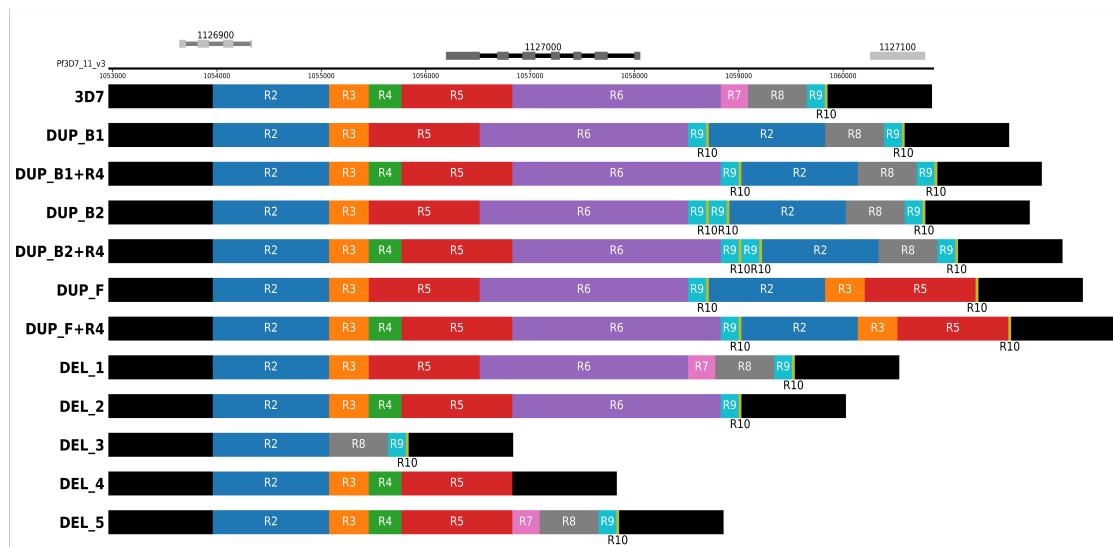


Figure 6.1: **Catalogue of structural variant haplotypes at the *Pfsa3* locus.** The *Pf* 3D7 reference genome has been sectioned into segments R1-11 based on break points involved in the structural variants. The locations of genes within the 3D7 reference are displayed across the top of the plot. Exons within these genes are shown as thicker light grey rectangles. The gene containing the top HbS-associated single nucleotide variants (PF3D7.1127000) is shown in darker grey. Haplotypes involving duplications of segments have been named “DUP” and haplotypes containing only deleted segments have been named “DEL”.

possible that other aspects of these structures were not evident during the inspection and confirmation of their presence in populations may require larger datasets or direct typing. In addition, we observed that segment ‘R4’ (Fig. 6.1 on the preceding page) appears to have been lost on several haplotype backgrounds and may be a recurrent deletion. However, this segment is short (~ 300 bp) and may be challenging to differentiate from a mixed infection, so we opted not to include this distinction within our consensus set until direct typing has corroborated these haplotypes.

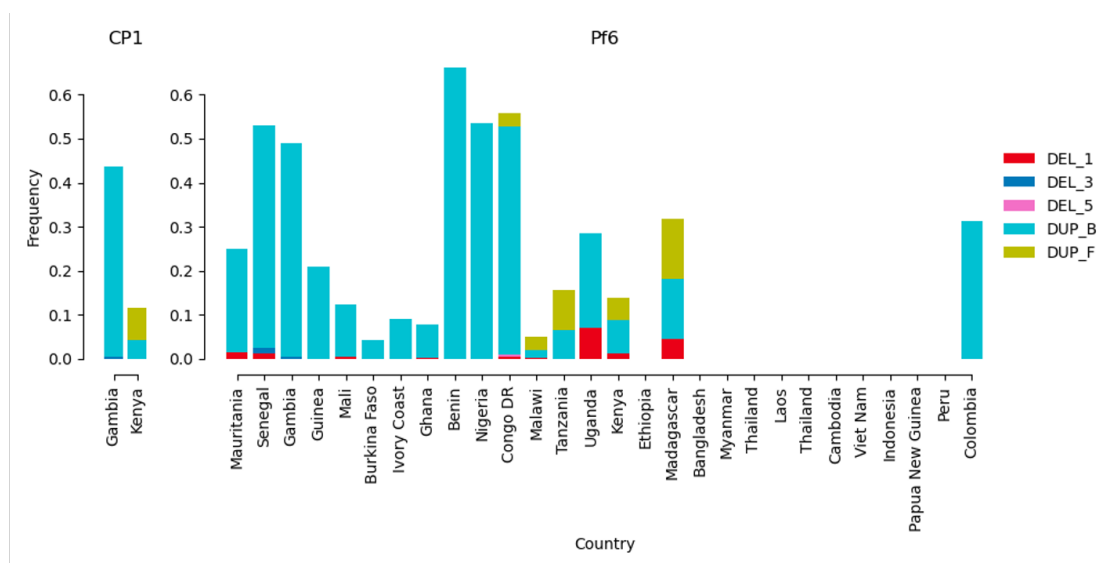


Figure 6.2: Allele frequencies of structural variants at the *Pfsa3* locus in global populations. Allele frequency of each structural variant haplotype in a set of CP1 severe cases from The Gambia and Kenya (left) and from the Pf6 resource of community sampled malaria infections from various countries. Country of origin is shown along the x-axis and frequency as a proportion between 0-1 on the y-axis. Countries on the x-axis are roughly ordered based on their geographical position east to west. Each SV haplotype is represented by a different colour as stipulated in the figure legend. The remainder of frequency (summing to 1 in all regions) is made up of non-structurally varying “3D7-like” haplotypes.

The degree of variation observed at this locus hints that complex forces of selection may be acting. Repeated, independent duplication events have been observed to occur in species under significant selective pressure (such as anti-insecticide resistance in *Anopheles* mosquitoes (213)). Extensive variation can however also be consistent with non-homologous recombination events between SV haplotypes, as is suggested to have occurred for the glycophorin locus in the human genome

(214). Preliminary observations indicate that the structural variants identified in this study share breakpoints and are likely to be ancestrally related, which is corroborated by haplotype analysis presented in Chapter 5 (Fig. 5.22 (pg.198)); however, further research is needed to elucidate this.

6.2.3 Confirmation of SV haplotype structures using long-read sequencing

In order to explore the structure of predicted SV haplotypes and attempt to validate those which had been predicted *in silico*, I made use of a set of new long-read assemblies developed alongside this work (for details of their sequencing and assembly see Materials and Methods section 2.3.2.1 (pg.41) and 2.5.4.1 (pg.50)). The PacBio platform was used to perform long read WGS for 8 parasite samples: an isolate called FUP-H (which to our knowledge is the only laboratory adapted clone available from public repositories which carries *Pfsa*+ alleles), four clinical isolates from Kenya (KENI1-KENI4) (selected to include two *Pfsa*+ and two *Pfsa*- parasites), two isolates from the Genetic Analysis of Malaria Cases and Controls (GAMCC) resource (the generation of which were described in Chapter 3 (pg.87) and which were called as *Pfsa*+) and the widely-used laboratory reference clone 3D7 (which does not carry *Pfsa* alleles).

It was not known whether the GAMCC samples, which had undergone sWGA, would produce high quality assemblies because it was suspected that the amplification process may introduce branched chimeric DNA, potentially inhibiting the generation of long-reads (215). However, for most samples, including those from the GAMCC set, long reads assembled into a single contig for each chromosome and the mitochondrial and apicoplast genomes to produce a high-quality assembly (see Materials and Methods section 2.5.4.1 (pg.50)).

Through manual inspection and an assessment of sequence sharing between the FUP-H and reference 3D7 assemblies (performed using short nucleotide sequences called kmers), we found that the FUP-H isolate carried a DUP-F haplotype (Supp. Fig. 6.9 (pg.254)). This was significant because it confirmed the existence of this haplotype which previously only been predicted using short read data and statistical inference. FUP-H was collected from a severe infection in California

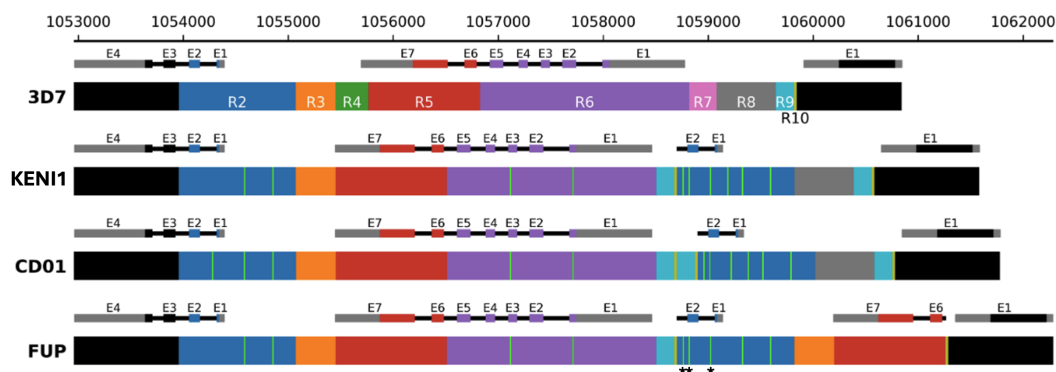


Figure 6.3: *Pfsa3* haplotypes for a subset of long-read genome assemblies The structure of the SV haplotypes at the *Pfsa3* locus within three long-read *Pf* genome assemblies compared to the reference assembly 3D7. The locations of SNPs on these assemblies are indicated by neon green lines. The assemblies FUP-H and Kenyan Isolate 1 (KENI1) were generated alongside this study and CD01 was released by Otto et al. (103). The locations of genic regions based on assembly annotations are displayed across the top of the structure plot. Exons within these genes are shown as thicker rectangles, coloured by the segment within which they are found, and untranslated regions are coloured grey. Sites on the FUP-H assembly marked with an * were analysed in RNA-seq data.

from a returning traveller from Uganda, an East African population, in 1966 (216). This corresponds to the observed restriction of this haplotype to East African populations (Fig. 6.2 (pg.233)).

One Kenyan isolate which carried *Pfsa*+ alleles, referred to as KENI1, was found to carry the DUP-B1 haplotype (Supp. Fig. 6.9 (pg.254)). Although the DUP-B1 haplotype had been previously observed in a public set of long read assemblies, this haplotype was present in a sample (called ML01) which had multiple contigs present for parasite chromosome 11 where *Pfsa3* sits (130). This was likely due to mixture in the sample which was collected from a clinical infection (103). This reduced confidence in this assembly because there was a possibility that variation could not be phased onto contigs to reflect the distinct haplotypes. With the addition of KENI1 we therefore generated a more reliable reference for the DUP-B1 haplotype. Of the remaining three Kenyan isolates (KENI2-4), the other carrying *Pfsa*+ alleles carried the DUP-B haplotype and those which were *Pfsa*- did not carry structural variation.

6.2.3.1 Development and testing of a long-read amplicon based assay for *Pfsa* regions

As part of further work investigating these haplotypes, and in order to trial a newly developed workflow by de Cesare et al. (104), we developed a polymerase chain reaction (PCR) amplicon panel targeting the *Pfsa* regions. With the help of members of the Hendry group (Centre for Human Genetics, University of Oxford), amplicons were generated by PCR from the FUP-H and 3D7 isolates and sequenced on the Oxford Nanopore Technologies (ONT) MinION. Primers for the *Pfsa3* region were designed such that different SV haplotypes would produce different sizes or different numbers of amplicons (Supp. Table. 6.5 (pg.259) and Supp. Fig. 6.10 (pg.255)). By aligning the sequenced reads to a reference constructed to emulate the structure of DUP-F, and using manual inspection of the data, it appeared that three amplicons were produced for the FUP-H line compared to two produced for 3D7, the sizes of which roughly corresponded to *in silico* predictions (Supp. Fig. 6.11 (pg.256)). This again corroborated our finding that the FUP-H line carried the DUP_F haplotype and illustrated the potential for the assessment of these haplotypes using this workflow which might be more appropriate for use in low resource settings (104).

6.2.4 The DUP-B and DUP-F haplotypes are strongly linked to *Pfsa3*+ SNPs

Having generated a consensus call set of structural variation present at the *Pfsa3* locus across global populations, I set out to establish whether the *Pfsa*+ variants are linked to SV haplotypes in this much larger sample. Using the consensus call set we determined that the *Pfsa*+ SNPs were closely associated with the duplication SV haplotypes. Specifically, alternate alleles for the two lead *Pfsa*+ SNPs in the *Pfsa3* region (Pf3D7_11.v3:1058035:T>A and Pf3D7_11.v3:1057437:T>C) were present on the vast majority of DUP_B and DUP_F haplotypes. The Pf3D7_11.v3:1058035:A allele was present within 95% and 98% of samples called as carrying DUP_B and DUP_F haplotypes respectively (Table 6.1 (pg.238)). Stronger associations were observed for the Pf3D7_11.v3:1057437:C allele which was carried by 98% of DUP_B and 100% of DUP_F samples (Table 6.2 (pg.239)).

Conversely, samples called as lacking structural variation almost never carried these variants (0% and 0.1% of samples called as carrying the 3D7 haplotype carried Pf3D7_11.v3:1058035:A and Pf3D7_11.v3:1057437:C respectively) (Tables 6.1 on the following page & 6.2 (pg.239)).

The primary contribution to the imperfect association of Pf3D7_11.v3:1058035:T>A and Pf3D7_11.v3:1057437:T>C to structural variation was the presence of mixed (both reference and alternate) alleles which were likely to be caused by mixed infections carrying both DUP and 3D7-like haplotypes. However, the DUP_B structure carried the reference Pf3D7_11.v3:1058035:T allele without mixture in 2% of samples in our set which might be due to incomplete linkage of these SNP alleles to structural variation. By examining coverage in these samples I found that each had only two read counts across this variant, suggesting that a low-quality variant call was likely the source of this discrepancy. This evidence suggests that the *Pfsa+* alleles are very strongly associated with the DUP-B and DUP-F structural variant haplotypes, extending the possibility that these structures underlie the association of this region to HbS.

Structural variant	Reference allele	%	Mixture of reference and Pfsa alleles	%	Pfsa allele	%	Missing	%	Not Called
3D7	4635	97.6	116	2.4	0	0.0	0	0.0	47
DEL_1	8	80.0	2	20.0	0	0.0	0	0.0	0
DEL_3	0	0.0	0	0.0	0	0.0	2	100.0	1
DEL_5	0	0.0	0	0.0	0	0.0	1	100.0	0
DUP_B	16	2.2	20	2.7	692	95.1	0	0.0	7
DUP_F	0	0.0	1	1.5	64	98.5	0	0.0	6
No Consensus	81	29.5	130	47.3	52	18.9	12	4.4	5

Table 6.1: **Linkage between Pf3D7_11_v3:1058035:T>A and SV haplotypes at the *Pfsa3* locus.** Each row represents a structural variant haplotype at the *Pfsa3* locus. Columns show counts and percentages of *Pf* samples in the consensus call set, called by the coverage call and the insert size call as carrying each SV haplotype, which also carry homozygous reference (T/T), heterozygous (T/A) or homozygous alternate (A/A) alleles or have missing genotype calls within the publicly available Pf7 or CP1 variant call set.

Structural variant	Reference allele	%	Mixture of reference and Pfsa alleles	%	Pfsa allele	%	Missing	%	Not Called
3D7	4619	97.2	126	2.7	6	0.1	0	0.0	47
DEL_1	3	30.0	0	0.0	7	70.0	0	0.0	0
DEL_3	0	0.0	0	0.0	0	0.0	2	100.0	1
DEL_5	1	100.0	0	0.0	0	0.0	0	0.0	0
DUP_B	1	0.1	15	2.1	711	97.7	1	0.1	7
DUP_F	0	0.0	0	0.0	65	100.0	0	0.0	6
No Consensus	75	27.3	114	41.5	77	28.0	9	3.3	5

Table 6.2: **Linkage between Pf3D7_11_v3:1057437:T>C and SV haplotypes at the *Pfsa3* locus.** Each row represents a structural variant haplotype at the *Pfsa3* locus. Columns show counts and percentages of *Pf* samples in the consensus call set, called by the coverage call and the insert size call as carrying each SV haplotype, which also carry homozygous reference (T/T), heterozygous (T/C) or homozygous alternate (C/C) alleles or have missing genotype calls within the publicly available Pf7 or CP1 variant call set.

6.2.5 Paralogous sequence variation within the *Pfsa3* SVs

SNPs in the surrounding *Pfsa3* region also show levels of association with HbS (130). By assessing allele frequencies at common SNPs (AC>30 in African samples) across the *Pfsa3* locus within each of the different haplotypes within the consensus call set, I saw that surrounding variation was also linked to certain SV haplotypes (Fig. 6.4). In particular, I noted several apparent SNPs within the duplicated R2 segment of the DUP-B/F haplotypes which appeared to be predominantly called as heterozygous (positions Pf3D7_11_v3:1054021:T>C, Pf3D7_11_v3:1054078:C>T (within DUP-B and DUP-F) and Pf3D7_11_v3:1054447: A>G (within DUP-B)) (Fig. 6.4). A heterozygous site would usually represent a single infection containing distinct parasites carrying disparate alleles at that site. However, because the R2 segment of the genome appears twice in parasites carrying the DUP haplotypes, heterozygosity might instead represent variation between paralogous copies (referred to as paralogous sequence variants (PSVs)).

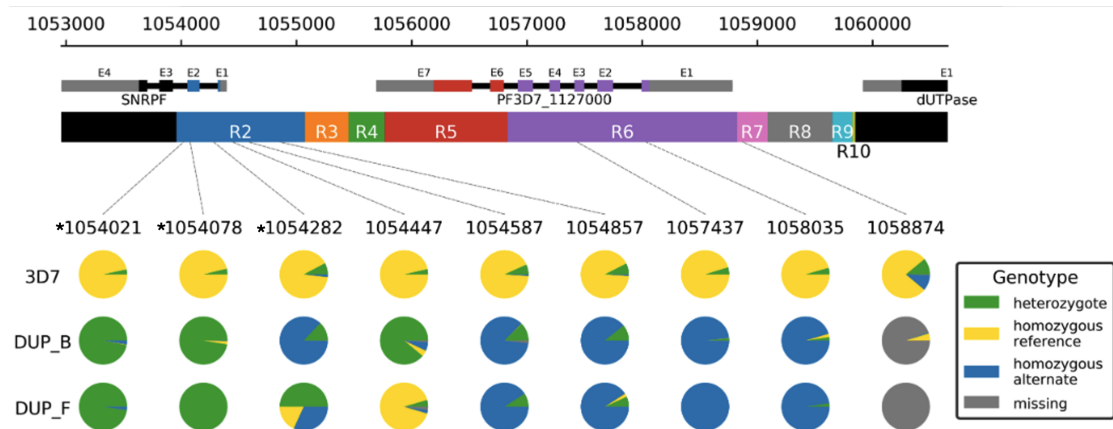


Figure 6.4: **Allele frequencies at SNPs across the *Pfsa3* region in samples with different SV haplotypes.** Pie charts indicate the proportions of homozygous and heterozygous genotypes at SNP locations in samples within the consensus call set which have been called as carrying DUP-B (DUP-B1, DUP-B2, DUP-B1+R4, DUP-B2+R4), DUP-F (DUP-F, DUP-F+R4) or 3D7 *Pfsa3* SV haplotypes by both the insert size and coverage callers. Sites marked by an asterisk were analysed in RNA-seq data.

Using the newly available FUP-H and KENI1 assemblies (which are examples of a DUP-F and DUP-B1 SV haplotype respectively), as well as the publicly available CD01 assembly (which carries a DUP-B2 haplotype (Supp. Fig. 6.9

(pg.254))), we were able to map polymorphisms to each copy of R2. This identified that the second copy of the R2 segment carried up to 6 non-reference alleles (Fig. 6.3 (pg.235)). Based on our assemblies, I observed that 4 of these can be PSVs. The sites Pf3D7_11_v3:1054021:T>C and Pf3D7_11_v3:1054078:C>T were PSVs in the CD01, KENI1 and FUP-H assemblies. A further PSV Pf3D7_11_v3:1054447:A>G appeared within CD01 and KENI1 but not FUP-H, and Pf3D7_11_v3:1054282:T>C was a PSV in FUP-H (Fig. 6.3 (pg.235)).

The observation that Pf3D7_11_v3:1054021:T>C and Pf3D7_11_v3:1054078:C>T were heterozygous in the majority of samples identified as carrying DUP structural variants within the consensus call set, along with Pf3D7_11_v3:1054447:A>G in samples carrying DUP-B, strongly suggests that these PSVs are linked to these SV haplotypes in global populations (Fig. 6.4 on the preceding page). The PSV Pf3D7_11_v3:1054282:T>C, despite being a PSV on the FUP-H assembly, was only sometimes called as heterozygous within both the DUP-B and DUP-F haplotypes in the consensus call set suggesting it is not strongly linked to either SV haplotype (Fig. 6.4).

6.2.6 A third method to call *Pfsa3* SVs using PSVs

The tight linkage of certain SNPs and PSVs to the DUP-B and DUP-F structural variant haplotypes raised the possibility of calling structural variant types using this nucleotide variation. To develop a systematic approach to utilise sequence variation to call structural variation, I designed a method which involved counting the appearance of short nucleotide sequences (referred to as kmers) which flanked this nucleotide variation in sequence data. An advantage of using a kmer based approach over the insert-size and coverage callers described above is that there is no need to align reads to a reference genome. Kmers can be counted within raw sequence data in a “fastq” format, making it potentially more straightforward to apply to additional datasets.

Pairs of non-overlapping alternate and reference 31mers (kmers of 31bp length) surrounding common SNPs (AC>30), now understood to also represent PSVs, were generated across the *Pfsa3* region and systematically filtered to reflect variation that was fixed on the 3D7, DUP-B and DUP-F backgrounds (with over 95%

of samples within those groups in the consensus call set having the same genotype call). The 3D7, KENI1 and FUP-H assemblies were used to verify that kmer pairs produced the predicted ratio of alternate to reference kmers and were not likely to be found elsewhere in the *Pf* genome. Five pairs of fixed, site specific 31mers (flanking variation at sites Pf3D7_11.v3:1054021:T>C, Pf3D7_11.v3:1054078:C>T, Pf3D7_11.v3:1054447:A>G, Pf3D7_11.v3:1057437:T>C and Pf3D7_11.v3:1058035:T>A) were then counted within samples using the programme jellyfish2. A likelihood was then calculated for the 3D7, DUP-B and DUP-F haplotypes based on the observed kmer counts compared to expected proportions derived from predicted mixtures of the haplotypes, utilizing a binomial distribution. The haplotype with the highest likelihood is therefore likely to represent the haplotype of that sample. This method is referred to hereafter as the ‘kmer call’ (for a full description of the statistical methods of this caller see Materials and Methods section 2.10.2.4 (pg.84)).

6.2.6.1 Testing the kmer calling method

The kmer call was tested in a subset of the consensus set (N=100), producing an 100% match (N=70 as 30 samples had fewer than 2 sites with any kmer counts). Samples within the CP1 resource were prepared using two different methods: some were sequenced from whole blood extractions to a high coverage (producing many human reads), and some following the amplification method sWGA (130). The coverage and insert size callers were not developed for use on sWGA data because of the uneven genome coverage this preparation produces (see Chapter 3 (pg.87)), so it is not known whether they can produce reliable calls for this datatype. As the kmer caller does not require alignment to an assembly, it is less likely to be influenced by uneven coverage. I tested the kmer caller on a subset of samples from CP1 in the consensus call set which also have sWGA data available for the sample. Using the kmer caller on the sWGA data produced an 100% match to the consensus call (N=48), indicating the kmer caller may be just as reliable on the sWGA data type. However, the kmer caller required that the region was well covered, which may not be the case for all sWGA samples. sWGA data now makes up the majority of WGS sequences available for *Pf* (37). Once this method has

undergone further development and testing, it therefore has the potential to be widely applied.

6.2.7 Genomic prediction of effect of *Pfsa3* SVs on gene transcription

6.2.7.1 Data to assess gene expression in parasites carrying different SV haplotypes

The changes in the genomic structure at the *Pfsa3* locus in parasites carrying DUP haplotypes raised questions as to whether this influences local gene expression. To investigate this I used two RNA sequencing (RNA-seq) datasets from Saelens et al. (217). In the first dataset, 3D7 and FUP-H parasites were synchronised, and messenger (m)RNA extracted every three hours post-invasion over a time course of 48 hours (an infection cycle). This was performed within blood from non-HbS carrier (HbAA) and carrier (HbAS) donors. Parasite stage was determined based on time point. This data therefore provides an opportunity to assess differences in gene expression between a parasite carrying a DUP-F (FUP-H) and 3D7 haplotype throughout infection.

The second dataset performed RNA sequencing on natural infections in children from Mali (N=32), half of which occurred in HbAS patients and half in HbAA patients. Within this data, parasite stage was inferred based on gene expression compared to the time series data generated from FUP-H and 3D7. Many of these infections carry *Pfsa3*+ alleles and are thus likely to carry structural variation, allowing for assessments of expression in samples carrying these haplotypes within natural infections.

6.2.7.2 Gene expression within a DUP-F carrying parasite compared to 3D7

It has been previously noted using this data that there was an increase of expression of the PF3D7_1127000 gene (which contains the lead *Pfsa3*+ variants) in the FUP-H isolate compared to 3D7 during the middle stages of infection (Fig. 6.5 on the next page). This hints at a possible basis for the association of *Pfsa3*

to HbS. Using the same data, I noted that dUTPase, the gene immediately downstream of PF3D7_1127000, was also differentially expressed between these isolates during the earlier stages of infection in this data (Fig. 6.5). The expression of SNRPF appeared to remain stable throughout *Pf* infection for the FUP-H and 3D7 isolates, regardless of whether the parasite was cultured in red blood cells donated from an HbAS or HbAA individual (Fig. 6.5).

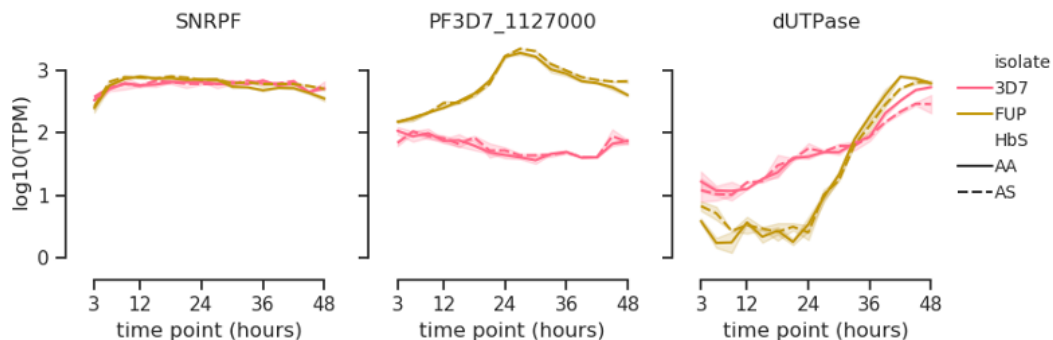


Figure 6.5: Differential gene expression between the FUP-H and 3D7 *Pf* isolates cultured in HbS-trait (AS) and non-trait (AA) blood culture
 RNA-seq data collected from two laboratory *Pf* isolates; 3D7 (pink) and FUP-H (gold). transcripts per million (TPM) values as calculated by Saelens et al. (217) for three genes local to the *Pfsa3* locus over 48 hours of infection culture are shown. Isolates were cultured in two different blood types; heterozygous HbS carriers (AS, dashed lines) and homozygous non-HbS individuals (AA, solid lines). This data was collected and averaged across two replicates.

Our SV analysis however, suggested that the duplication of the R2 segment within the DUP SV haplotypes would lead to a partial duplication of the first two exons of the SNRPF gene, which are placed just upstream of Pf3D7_1127000 (Fig. 6.1 (pg.232)) (see Materials and Methods section 2.5.4.1 (pg.50) for details of this annotation in long-read assemblies). I therefore examined the impact of this within expression data.

6.2.7.3 The duplication of SNRPF gives rise to additional RNA transcripts

As shown in Figure 6.3 (pg.235), the second copy of the duplicated R2 segment contained 3 PSVs which fell within the partial copy of SNRPF in the FUP-H parasite (indicated with an * on Figs. 6.4 (pg.240) & 6.3 (pg.235)). One of these

(Pf3D7_11_v3:1054078:C>T) fell within an exon. This meant this region should be well represented in the RNA-seq dataset, which sequenced primarily exonic portions of mRNA transcripts. By searching for the specific 31mer (sequence of 31 nucleotides) centred on the SV containing the alternate PSV allele (which occurs only within the second copy of R2 in FUP-H), I found it to be present in the RNA-seq data produced from this isolate. This indicated that the partial duplicate of SNRPF is producing RNA transcripts. Importantly, no corresponding kmer counts were found in expression data for 3D7, confirming that the second copy of R2 was responsible for the transcripts, ruling out false-positive results from another genome location.

To endeavour to assess the levels of expression of this transcript throughout infection, I counted the same kmer across the time series of expression data available within the dataset. This was compared to the total kmer count (made up of kmers carrying both the alternate and reference alleles for each PSV). I used this as a proxy to determine the relative expression levels of the partial SNRPF duplicate compared to expression of the complete SNRPF gene. As shown in Figure 6.6 on the next page, the expression pattern of the partial duplication of SNRPF appeared to inversely mirror the expression of the PF3D7_1127000 gene seen in FUP-H (Fig. 6.5 on the preceding page). Proportions of the paralogous copy of the Pf3D7_11_v3:1054078:C>T PSV kmer (containing the T allele) were initially high (41% when cultured in HbAA blood and 43% in HbAS culture on average across 2 replicates) but decreased during the middle stages of infection (reaching a minimum level of ~9% at 24 hours post-infection). This reduction was not attributable to increased expression of SNRPF, which remained constant throughout this infection (Fig. 6.5 on the previous page).

I also performed the same analysis for two other PSVs within the partial SNRPF duplication carried by FUP-H: Pf3D7_11_v3:1054021:T>C and Pf3D7_11_v3:1054282:T>C. Because both sites fell within introns, they were less well represented in the RNA-seq data because introns are not found in mature mRNA. Rather than representing differences in expression, their presence might therefore have indicated changes in splicing of introns from RNA transcripts.

31mers originating from Pf3D7_11_v3:1054021:T>C were almost entirely absent in 3D7, indicating this intron was generally efficiently spliced from the SNRPF

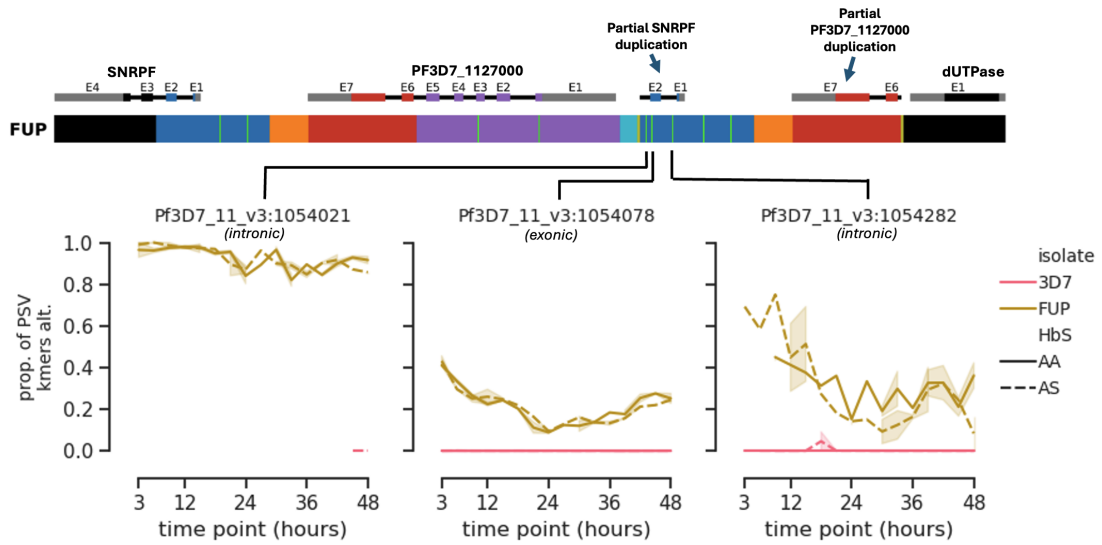


Figure 6.6: Proportions of PSV alleles between the FUP-H and 3D7 *Pf* isolates cultured in HbS-trait (AS) and non-trait (AA) blood culture
 Counts of 31mers (short nucleotides sequences of 31bp in length) originating from the second copy of the R2 segment within RNA-seq data collected from two laboratory *Pf* isolates; 3D7 (pink) and FUP-H (gold). Counts are divided by total count from both the first and second copy of R2 (called the proportion of PSV alternate kmers or ‘prop. of PSV kmers alt’). Data was collected by Saelens et al. (217) over 48 hours of infection culture. Isolates were cultured in two different blood types; heterozygous HbS carriers (AS, dashed lines) and homozygous non-HbS individuals (AA, solid lines). This data was collected and averaged across two replicates.

transcript. However, within the RNA-seq data from FUP-H this 31mer was present and almost entirely originated from the second copy of R2 (containing the C allele) (Fig. 6.6 on the preceding page). As SNRPF expression appeared to be similar between FUP-H and 3D7 isolates (Fig. 6.5 (pg.244)), this indicated that the intron may not be as effectively spliced from the extra transcript associated with FUP-H. In contrast, the second intronic PSV Pf3D7_11.v3:1054282:T>C appeared to follow an expression pattern similar to Pf3D7_11.v3:1054078:C>T (Fig. 6.6 on the preceding page).

6.2.7.4 Assessments of RNA expression data from clinical infections

To assess whether this pattern extended to natural parasites and was not in some way an artefact of the long-term culture of the laboratory-adapted 3D7 and FUP-H isolates, I extended the above analysis to the RNA-seq data generated from *Pf* parasites in 32 clinical samples collected in Mali (from carriers and non-carriers of the HbS allele) as part of the same expression study (217).

The kmer caller described above predominantly uses sites within genes to produce a call. This enabled the use of this caller to estimate structural variant haplotypes in the clinical samples from the available RNA-seq data. Because the only site which differentiated DUP-B and DUP-F haplotypes (Pf3D7_11.v3:1054447:A>G) occurred within an untranslated region and was poorly covered within this data, I opted to solely classify infections as DUP or 3D7 haplotypes (based on population frequencies in Mali shown in Fig. 6.2 (pg.233), it was likely these samples contained DUP-B haplotypes). The results of these calls are shown in Supp. Table. 6.6 (pg.259). The majority of samples collected from HbAA individuals contained 3D7-like haplotypes, but a single infection was predicted to carry a DUP haplotype. Conversely, most infections sampled in HbAS individuals were associated with DUP haplotypes and a single infection carried a 3D7-like haplotype.

Based on these calls, the clinical samples which carried DUP haplotypes showed a similar pattern of gene expression to the DUP-F carrying FUP-H isolate for the three genes at the *Pfsa3* locus, as assayed by published transcript quantification data. Clinical infections carrying 3D7-like haplotypes showed similar patterns of expression to the 3D7 laboratory isolate (as indicated by comparing Fig. 6.7 &

Fig. 6.5 (pg.244)). In particular, PF3D7_1127000 was more greatly expressed by naturally infecting parasites carrying the DUP haplotype during the middle trophozoite stages of infection (Fig. 6.7). This confirmed that the expression pattern observed within FUP-H extended to clinical infections, as has been previously shown (130), and that this strongly corresponded to structural variant haplotype.

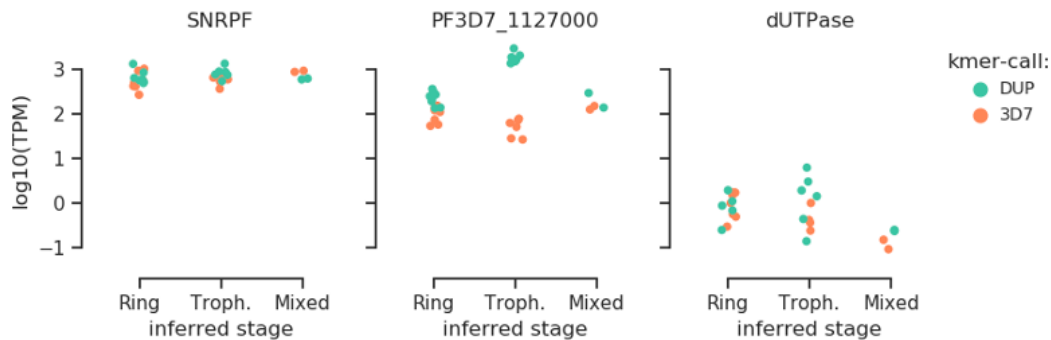


Figure 6.7: Gene expression within clinical *Pf* isolates collected in Mali
 RNA sequence data collected from two laboratory clinical infections in Mali. Infecting *Pf* parasites have been determined to carry “DUP” SV (green) or “3D7” SV (orange) haplotypes using the kmer caller described in the main text. transcripts per million (TPM) values were calculated by Saelens et al. (217) for three genes local to the *Pf*sa3 locus. Infections were categorised into stages using time series data in the laboratory isolates FUP-H and 3D7.

Patterns of expression from the partial duplication of SNRPF in the clinical infections also appeared to mirror that observed in FUP-H and 3D7 analysis (shown by comparing Fig. 6.8 on the following page & Fig. 6.6 (pg.246)). The relative levels of RNA from the second copy of R2 (assessed using counts of kmers surrounding the Pf3D7_11_v3:1054078:C>T PSV as described above), also reached its lowest point during the middle trophozoite stage within the clinical infection data (Fig. 6.8). Furthermore, sequence originating from the intronic site Pf3D7_11_v3:1054021:T>C was absent in clinical infections carrying 3D7 haplotypes as was observed for the 3D7 laboratory isolate. The paralogous C allele of this PSV was however present in expression data from infections carrying DUP-like haplotypes (Fig. 6.8). It should be noted that this clinical RNA-seq data was of lower coverage than the dataset for 3D7/FUP-H (displayed in Fig. 6.6) and kmer counts were in general very low for all non-exonic sites, limiting the power with which to accurately assess these proportions. However, this suggested that

the inferred levels of expression of the paralogous partial copy of SNRPF, and the inclusion of intronic sequence in this RNA, replicated in natural infections.

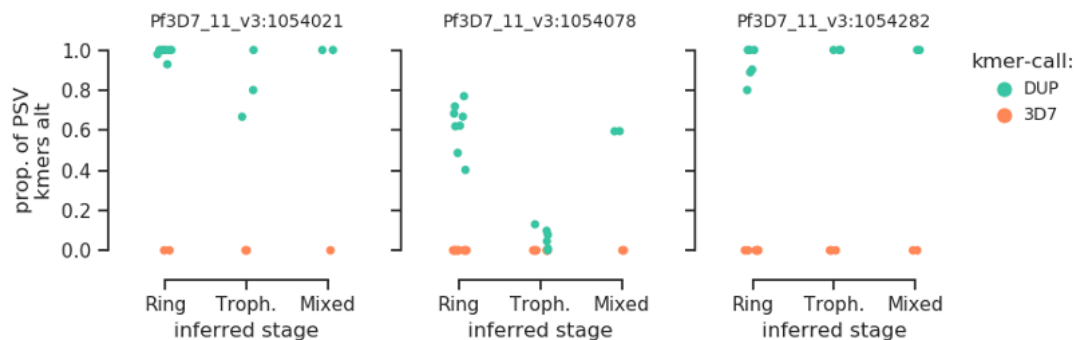


Figure 6.8: Proportions of PSV alleles within clinical *Pf* infections collected in Mali Counts of 31mers (short nucleotides sequences of 31base pairs (bps) in length) originating from the second copy of the R2 segment within RNA sequence data collected from clinical infections in Mali. Infecting *Pf* parasites have been determined to carry “DUP” SV (green) or “3D7” SV(orange) haplotypes using the kmer caller described in the main text. Counts are divided by total count from both the first and second copy of R2 (called the proportion of PSV alternate kmers or ‘prop. of PSV kmers alt.’). Data was collected by Saelens et al. (217) over 48 hours of infection culture.

Whether this additional transcript is functionally important is however not clear. Comparing the expression of PF3D7_1127000 and the partial duplication of SNRPF in parasites carrying DUP haplotypes, it was not immediately obvious whether they interact. Overall, the analyses of these two datasets has not provided clear evidence for a role of the SNRPF partial duplication in driving the increased expression of PF3D7_1127000 but given the regulation of gene expression is complex, these data may be insufficient to draw definitive conclusions regarding this relationship. Further assessments of other data types, such as of regulatory elements or epigenetic modifications, may be required to understand the genomic landscape at this locus and how it is impacted by structural variation.

6.3 Discussion

It has previously been shown that regions of the *Pf* genome, termed *Pf*sa, associate with the human resistance genotype HbS and therefore may mediate the escape of this resistance by parasites (via the yet uncharacterised ‘HbS-escape

phenotype'). We have discovered high amounts of previously undescribed structural diversity at one locus called *Pfsa3* within two large publicly available datasets (Fig. 6.1 (pg.232)). This included the discovery of a new SV haplotype, here called DUP-F, which appears ancestrally related to the previously described DUP-B1 and DUP-B2 haplotypes (referred to collectively as DUP-B). These DUP haplotypes all share the duplication of a segment of the genome referred to as R2 (Fig. 6.3 (pg.235)). The SV haplotypes at *Pfsa3* were distributed across Africa yet, like alleles at another identified locus *Pfsa2*, DUP-F appeared restricted to East Africa (Fig. 6.2 (pg.233)). This could indicate that there is spatial variation in the evolutionary forces with respect to these regions across Africa or that their evolutionary history differs.

We have shown that the lead *Pfsa+* variants at this region are strongly linked to the DUP-B and DUP-F haplotypes (Tables 6.1 (pg.238) & 6.2 (pg.239)), extending the possibility that structural, rather than nucleotide, variation underlies the association of the *Pfsa3* region to HbS. The HbS-escape phenotype itself is not understood, but it has been observed that PF3D7_1127000, the gene containing the lead *Pfsa+* variants in the region, showed a marked increase in gene expression within *Pf* isolate FUP-H (a *Pfsa+* parasite) compared to 3D7 (a *Pfsa-* parasite). The function of PF3D7_1127000 itself is unknown but the presence within the population of the DEL-3 haplotype, which entirely deletes the gene, indicates that the gene is not essential to the survival of parasites. The PF3D7_1127000 gene is not directly disrupted by the structural variation within the DUP-B and DUP-F haplotypes, but a duplicated region in these haplotypes (segment R2 in Fig. 6.1 (pg.232)) bisects the adjacent gene SNRPF and transplants the first two exons upstream of PD3D7_1127000 (Fig. 6.3 (pg.235)).

The existence of this duplication on multiple haplotype backgrounds hints at its importance. It's possible that the duplication might underlie the alteration of expression of PF3D7_1127000. One possibility is that the initiation machinery of SNRPF, which may respond to different regulatory factors, has been transplanted upstream of the Pf3D7_1127000 coding sequence leading to the measured differential expression. Although the assessment for open chromatin and histone marks which indicate the presence of such machinery has been performed on *Pf* parasites

(reviewed in (218)), to our knowledge this has not been performed on a *Pf*sa+ isolate which limits our ability to understand the regulatory landscape in the context of this structural variation.

There is also a possibility that the duplication of the R2 segment leads to the occurrence of an extended transcript combining the paralogous section of SNRPF with all or part of PF3D7_1127000. Due to the presence of RNA reads aligning across the discovered PSVs we can verify that a transcript is produced from the paralogous R2 segment (Fig. 6.6 (pg.246)), but whether this commonly extends into the PF3D7_1127000 sequence is not clear. We found some evidence for an extended transcript in RNA-seq data produced from the FUP-H isolate (Supp. Fig. 6.12 (pg.256)) however it appeared to exist in small numbers relative to other transcripts produced from the paralogous segment. It should be noted however that the short-read nature of this data somewhat limits the ability to identify new transcripts.

This data was generated using a poly-A capture method, which enriches RNA-seq libraries for mRNA (217), so the presence of this additional transcript in the sequencing library suggests the possibility that this duplicate transcript is a poly-adenylated. However, internal poly-A sequences, which are common throughout the *Pf* genome, have been shown to bind to poly-T oligos during RNA-seq library preparation leading to their capture and truncation (219). This suggests the possibility that non-terminally poly-adenylated transcripts were captured during this process. A poly-T sequence of 21bp in length is found upstream of the paralogous SNRPF exon 2 in the FUP-H assembly (SNRPF is on the reverse strand).

Interestingly, it appears that expression of the paralogous SNRPF segment inversely mirrors the expression of PF3D7_1127000, that is, when the expression of PF3D7_1127000 increases during the trophozoite infection stage in FUP-H (Fig. 6.5 (pg.244)) and in clinical Malian samples (Fig. 6.7 (pg.248)), the proportion of kmers from the second copy of the duplicated R2 segment decreases within the same expression data to a similar degree (Figs. 6.6 (pg.246) & 6.8 (pg.249)). This could suggest that the two transcripts, the partial SNRPF paralogue and PF3D7_1127000, in some way negatively influence one another, however further experimentation is required for this position to become anything but speculation.

Overall, the evidence presented here suggests that a SNP focused approach to functional assessment of the *Pfsa3* locus may not be sufficient to uncover the HbS-escape phenotype. In particular, gene-editing the lead *Pfsa3*+ variants revealed to be associated to HbS by Band et al. (130) onto non-structurally varied backgrounds, as has been performed effectively for other loci (e.g. (112)), may not appropriately target the desired hypotheses.

The evidence presented here however does not exclude the possibility that nucleotide variation contributes to the HbS-escape phenotype. One already posited possibility is that Pf3D7_11.v3:1057437:T>C alters a PEXEL motif. Here, we show that Pf3D7_11.v3:1057437:T>C is tightly linked to the structural variant haplotypes, even more so than the most sickle-associated variant at the *Pfsa3* locus (Pf3D7_11.v3:1058035:T>A) (Table 6.2 (pg.239)). In addition, in this chapter I have found multiple other highly linked SNVs and PSVs which may not have been directly assessed due to presenting as highly heterozygous in variant call data. It is certainly striking that up to 6 alternate alleles are found on the paralogous R2 segment. There is a suggestion that intronic sequence retention is not the same between SNRPF paralogues due to absence of the non-paralogous kmer at position Pf3D7_11.v3:1054021:T>C in sequencing data generated from FUP-H and 3D7 parasites and the presence of the paralogous kmer in FUP-H (Fig. 6.6 (pg.246)). It may be the case that variation within this paralogue somehow alters the regulation of splicing within the region, however more data is required to conclude this with certainty. Functional work, such as long-read transcriptomic studies or assessments of Pf3D7_1127000 protein transport in *Pfsa3*+/- parasites, will prove essential to the determination of whether structural variation, SNP genotypes (such as the PEXEL-motif adjacent *Pfsa3*+ variant at Pf3D7_11.v3:1057437:T>C) or a combination of the two functionally underly the association of *Pfsa3* to HbS and thus the HbS-escape phenotype.

We initially developed two methods by which to call structural variation at the *Pfsa3* locus: the coverage caller and the insert-size caller. These showed high concordance indicating they were accurately identifying structural variation at the *Pfsa3* loci (Supp. Table 6.4 (pg.258)). Our results suggested that a potential problem for these callers was the presence of mixture in parasite infections as samples which were not called with a consensus using the coverage and insert size

callers typically carried mixtures of alleles at *Pfsa*+ SNPs (Tables 6.1 (pg.238) & 6.2 (pg.239)). Furthermore, the coverage and insert size callers were developed for use on non-sWGA amplified samples. The sWGA data type is becoming commonplace within available malarial parasite genome datasets because this method allows the ability to enrich small amounts of parasite DNA to a high mass. A third caller called the kmer caller, which appeared to work well in sWGA data, was developed. However, this caller is not able to determine new haplotypes that may be present at rare frequencies and may be detectable as datasets continue to expand. Therefore, further refinement of the coverage and insert size callers to handle mixture (as is discussed for the insert size caller in Materials and Methods section 2.10.2.3 (pg.83)), and to accommodate the sWGA data-type would be beneficial. This would enable further characterisation of structural variation at this locus, allowing well-powered investigations into both the relationship of this locus to HbS, and the evolutionary forces acting upon it. Ultimately, these efforts will contribute to unravelling the *Pfsa* associations, which is crucial for advancing our understanding of malaria parasites and developing effective strategies to target them.

6.A Supplementary Figures and Tables

6.A.1 Supplementary Figures

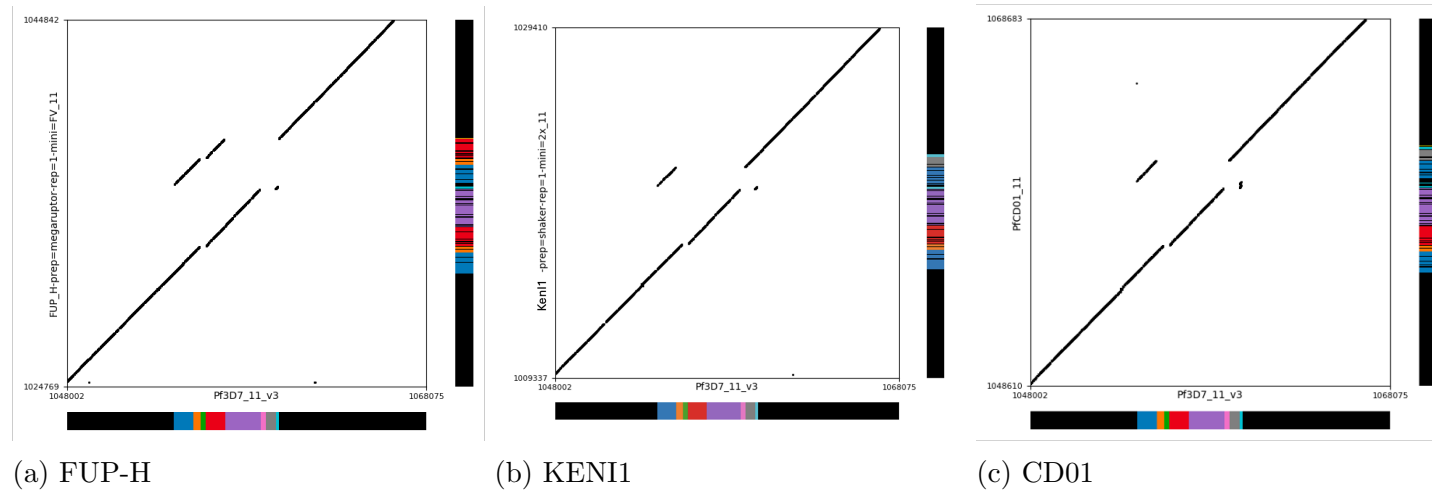


Figure 6.9: **Nucleotide sequence (kmer) sharing between *Pf* assemblies at the *Pfsa3* locus.** Kmer sharing plots comparing *Pf* genome assemblies of FUP-H, KENI1 and CD01 to the 3D7 reference assembly. Points represent the sharing of a 51 base pair kmer in both assemblies, and the position on the x and y axes indicate the position of that kmer within the reference and other assembly respectively. Colour bars along the bottom and right hand sides of the plots indicate the segment from which each kmer in the respective assembly belongs as displayed in Figure 6.1 (pg.232). Black sections within the colour bars for the y axis assembly indicate that there is no match to a kmer within the reference at that position due to variation within that parasite with respect to the 3D7 reference.

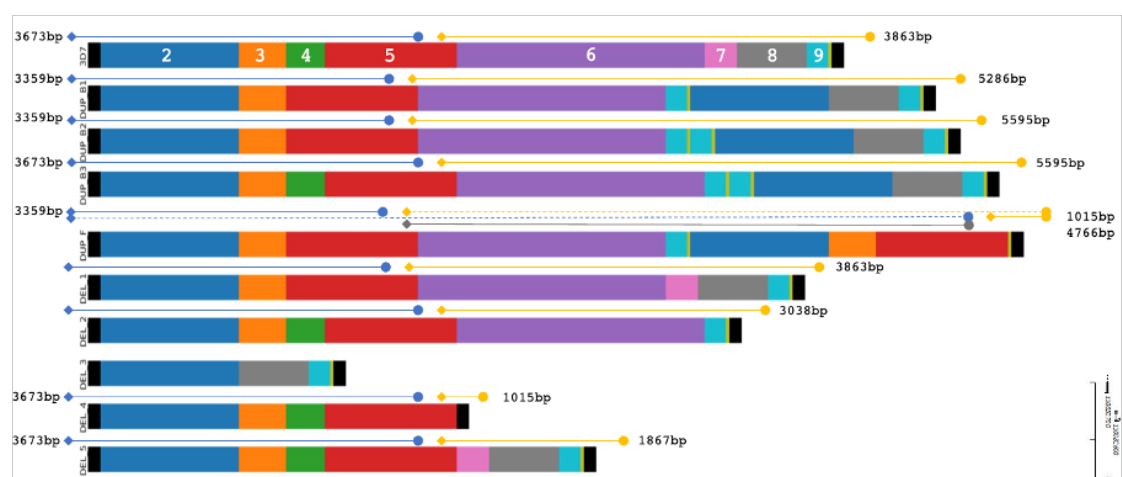


Figure 6.10: **Primers designed to amplify the *Pfsa3* locus for amplicon sequencing on the ONT MinION.** Amplicon sizes for the ONT MinION assay generated using a workflow developed by de Cesare et al. (104). Amplicon sizes vary depending on the structural variant haplotype carried by a tested sample.



Figure 6.11: **Targetted amplicon sequencing of the *Pfsa3* locus within the FUP-H on ONT MinION.** Alignments of reads generated from sequencing on the ONT MinION of amplicons produced from the FUP-H *Pf* isolate. These reads were aligned to a 3D7 assembly which had been rearranged to emulate the predicted structural variant haplotype DUP-F. Three reads appear to have been generated, as predicted for DUP-F haplotypes.

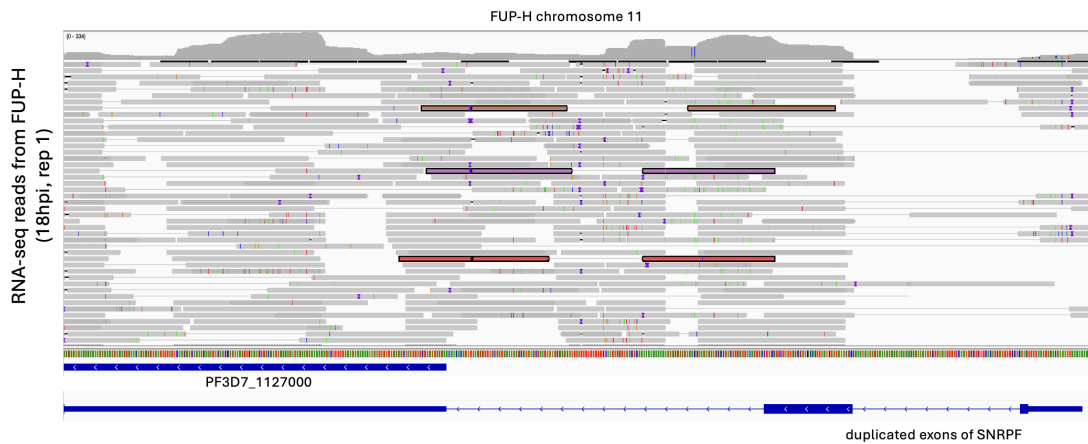


Figure 6.12: **RNA-seq read pairs from FUP-H align to PF3D7_1127000 and the partial duplication of SNRPF.** Highlighted read pairs align to the gene PF3D7_1127000 and the partial duplication of SNRPF across PSV sites with a MAPQ > 20. This image was made using the Integrative Genomics Viewer (IGV) (211).

6.A.2 Supplementary Tables

Chromosome	Start	End	Name
Pf3D7_11_v3	1038038	1053959	R1
Pf3D7_11_v3	1053960	1055073	R2
Pf3D7_11_v3	1055074	1055454	R3
Pf3D7_11_v3	1055455	1055768	R4
Pf3D7_11_v3	1055769	1056830	R5
Pf3D7_11_v3	1056831	1058826	R6
Pf3D7_11_v3	1058827	1059087	R7
Pf3D7_11_v3	1059088	1059652	R8
Pf3D7_11_v3	1059653	1059824	R9
Pf3D7_11_v3	1059825	1059852	R10
Pf3D7_11_v3	1059853	1078034	R11

Table 6.3: Segments across the *Pfsa3* involved in structural variation

Country	3D7	DUP_B	DUP_F	DEL_1	DEL_3	DEL_5	No Consensus	Total
<i>African Populations</i>								
Ghana	734	62	0	1	0	0	29	826
Gambia	190	189	0	0	2	0	46	427
Kenya	283	14	25	1	0	0	42	365
Mali	301	43	0	2	0	0	19	365
Malawi	309	4	10	1	0	0	14	338
Tanzania	227	19	26	0	0	0	15	287
Congo DR	74	114	7	1	0	1	34	231
Cameroon	84	110	0	0	0	0	35	229
Guinea	97	27	0	0	0	0	9	133
Senegal	36	39	0	1	1	0	4	81
Ivory Coast	58	6	0	0	0	0	3	67
Mauritania	45	15	0	1	0	0	3	64
Benin	17	35	0	0	0	0	1	53
Nigeria	12	15	0	0	0	0	1	28
Burkina Faso	19	1	0	0	0	0	3	23
Ethiopia	23	0	0	0	0	0	0	23
Madagascar	14	3	3	1	0	0	1	22
Uganda	9	0	0	1	0	0	4	14
<i>Other</i>								
Colombia	4	5	0	0	0	0	7	16
Non-African Populations	2254	0	0	0	0	0	54	2308
Lab Isolates	12	0	0	0	0	0	2	14
Total	4798	696	71	10	3	1	319	5898
Percentage	81.35	11.8	1.2	0.17	0.05	0.02	5.41	100.0

Table 6.4: Structural variant frequencies at the *Pfsa3* locus across global populations in the consensus call set

Target	Direction	Sequence	Position	Product (bp)
AAT1	F	AGAGCGTTGACGAAATGAATGG	1210980	6558
AAT1	R	TGGCCTTGTTTCATATAACCTGT	1217537	6558
Pfsa1	F	TTTTGGTTTCAGGGGGTTGTGA	627771	5233
Pfsa1	R	TTCTCGTTTGCTGACCCTT	633003	5233
Pfsa2	F	AGTTCACGAAATTCAAGTTACA	812125	4079
Pfsa2	R	TGGGTATACACATCACGATGGT	816203	4079
Pfsa3A	F	GCTTAAAGCGTTCGTGGTT	1052706	3674
Pfsa3A	R	ACGGGGAATTTTGATGAGTGT	1056379	3674
Pfsa3B	F	CCGTGTTCCAACCTGACTGTC	1056475	3837
Pfsa3B	R	ACTTCCAGAGGAAGGAGGT	1060311	3837

Table 6.5: **Primers to amplify the *Pfsa3* locus for sequencing on the ONT MinION.** Primer information for the ONT MinION assay generated using a workflow developed by de Cesare et al. (104). Amplicon sizes vary depending on the structural variant haplotype carried by a tested sample.

	3D7	DUP
AA	15	1
AS	1	15

Table 6.6: **Structural variant calls at the *Pfsa3* locus made using RNA-seq data from clinical infections collected in Mali using the kmer caller.**

Chapter 7

Concluding Remarks

Pf malaria remains a major scourge faced by society. Finding methods to tackle this disease is a global aim. Scientific enquiry has so far struggled to develop strategies with established long-term efficacy; the use of anti-malarial drugs and vector-controlling insecticides has become increasingly limited due to the evolution of resistance within the parasite and vector (48). Although major strides in vaccine development have been made over the previous four decades, protection conferred by the leading vaccine candidate Matrix-M/R21 is incomplete and diminishes with time (29). It is therefore unlikely that these tools alone will enable the eventual elimination of the disease (17).

Human evolution has developed by far the most effective mechanisms for protection against malaria. As described in the Introduction section 1.3 (pg.12), several human resistance polymorphisms strongly impact disease risk, with the most striking example being the HbS polymorphism in the *HBB* gene (54). Independent of HbS, many adults in malaria endemic countries are completely protected from experiencing malaria symptoms through naturally acquired immunity (14). Understanding this process is crucial for vaccine development, but is for the most part mysterious (11). It is my belief that integrating parasite genetics with human genetics will enhance our understanding of human resistance and immunity to malaria, and provide greater insight into how parasites evade these defences. This will facilitate the development of improved interventions that may ultimately lead to eventual elimination.

The goal of this thesis was to generate a new data resource and use it to perform

analyses which illuminate parasite genetic effects on disease, advance our understanding of the emerging field of human-parasite genetic interactions in malaria and facilitate future research of parasite genomics. To achieve these goals, I presented four results chapters.

Within the first results chapter (Chapter 3 (pg.87)), I described the generation of the new GAMCC resource of *Pf* WGS data and targeted human genotyping at the *HBB* and HLA loci. The availability of large genetic datasets is crucial to perform highly powered statistical analyses of evolution (48). The GAMCC collection, which was sampled in the late 1980s, offers opportunities to perform temporal as well as spatially based analyses. Like other datasets which have empowered malaria genomics research (e.g. (37; 130; 217)), my intention is to publish this data as an openly accessible resource as soon as possible following the completion of this thesis.

Severe malaria is clinically very significant, causing a high number of deaths. Therefore, unravelling the factors which contribute to this presentation is keenly sought after (11). As I described within the Introduction section 1.3 (pg.12), many initial investigations into the human genetic factors which impact disease severity led to associations which failed to replicate in larger studies. This suggests that some were potentially false positive associations, perhaps due to a focussed approach where confounders were not adequately controlled for. Since then, GWAS scans have more robustly revealed the human genetic loci which impact severity across populations (6). However, no such assessment has been made for the parasite genome. Therefore, within the second results chapter (Chapter 4 (pg.110)) I performed the first unbiased whole genome assessment of parasite genetic effects on disease severity using severe and mild malaria cases within the GAMCC. This effort faced a few limitations, including a relatively small sample size and challenges in obtaining appropriate datasets for replication. A potential severity locus on parasite chromosome 1 showed replication in Kenyan severe and non-severe cases sourced from two separate public datasets. However, overall, I found minimal evidence of strong parasite genetic effects on disease severity, highlighting the need for further equivalent sample sets to improve power. It is my hope that this research sets a precedent for future analyses of this kind and encourages the sharing of disease phenotype information to support such efforts.

A significant recent discovery is that *Pf* parasites are evolving to counteract the HbS resistance polymorphism, with the relevant regions of the *Pf* genome referred to as *Pfsa* loci (130; 131). Understanding the impacts of these loci has rapidly become a research aim within the malaria genomics community. In the third results chapter of this thesis (Chapter 5 (pg.148)), I revealed a novel potential *Pfsa* locus (referred to as *Pfsa2B*) using GAMCC, which is one of the few datasets which contain *Pf* WGS data and HbS typing. By performing a joint analysis across all four available datasets, I revealed four further potential *Pfsa* loci which warrant interrogation by the research community, drawing attention to the complexity of HbS-*Pfsa* interactions.

In addition, I developed custom statistical methods to reveal large amounts of previously undescribed structural variation at the *Pfsa3* locus, presented in the final results chapter (Chapter 6 (pg.225)). This revealed that these structures, depicted in Figure 6.1 (pg.232), could underlie the association of this region to HbS. By uncovering the mechanisms parasites use to overcome HbS-mediated resistance, we may identify a critical vulnerability that can be targeted pharmacologically to inhibit infection. Investigations into the transcriptomics and proteomics of parasites carrying these genome structures will be key to this endeavour. Furthermore, although HbS has long been understood to confer protection against malaria, the mechanisms by which this occurs are not resolved (56). By uncovering the interactions between parasites differentially overcoming HbS effects, we may learn how this resistance is achieved, potentially leading to ways to exploit these effects across individuals and reduce morbidity and mortality attributable to malaria.

I have thus conducted a comprehensive set of investigations into the genetic effects of the parasite on disease severity using a novel dataset, the sharing of which will facilitate future studies on parasite evolution. Moreover, I have added to our understanding of the *Pfsa* loci, which represent a key area of malaria research.

7.0.1 Limitations and Future Directions

The work presented in this thesis faced a set of consistent constraints which I will briefly outline. Firstly, Chapters 4 and 5 presented multiple analyses using

the GAMCC resource. This data was carefully constructed and was of high quality (as described in Chapter 3) but had a relatively small sample size, meaning comparisons within this data suffered from limited statistical power.

This limitation underscored the need for replication of signals. However, this was challenging due to a scarcity of comparable datasets. For example, other than the GAMCC resource, to my knowledge no other single WGS dataset contains both severe and mild cases, so there was limited choice for replication of the GWAS comparing severe and mild cases (Chapter 4). This is in part due to challenges in classifying the severity of malaria cases, which is crucial to the ability to robustly compare analysis across resources and interpret results. Furthermore, analysis of HbS associations is constrained by the rarity of cases in HbS carriers (Chapter 5). Data from three countries (Gambia, Kenya and Ghana) has been assessed for HbS-*Pfsa* interactions so far, and further datasets will be necessary to fully understand these relationships.

The analyses of the *Pf* genome within this thesis were limited to the core regions which can be analysed using short read data (102). However, this excludes the subtelomeric and other hypervariable regions which contain many genes which likely impact disease such as the VAR, RIFIN and STEVOR families (129). Additionally, the exploration of the human genome was limited to a single polymorphism in the *HBB* gene and the presence of a single HLA allele. Even with complete sequencing, which is often prohibitively expensive, certain regions of the human genome also remain challenging to access. This is particularly the case within African genomes, which are crucial to the study of *Pf* malaria, because they are underrepresented in resources used to resolve these regions (55).

This highlights the need for large public *Pf* WGS datasets with detailed phenotyping from diverse populations, allowing well-powered precise cross-population comparisons. This should be coupled with human genetic data to facilitate assessments of host-parasite interactions, ideally through a genome-to-genome approach. Accessing complex gene regions within the host and parasite requires advanced sequencing technologies, alongside computational methods for effective statistical analyses. Importantly, efforts must be made to ensure these technologies are accessible within Africa, with careful consideration given to developing methods that are suitable for widespread use.

Bibliography

- [1] A. Nerlich. *Paleopathology and Paleomicrobiology of Malaria*. Paleomicrobiology of Humans. 2017. ISBN 978-1-55581-916-3. doi: 10.1128/microbiolspec.PoH-0006-2015.
- [2] E. Massa, N. Cerutti, and D. Savoia. Malaria in Ancient Egypt: Paleomimmunological investigation on predynastic mummified remains. *Chungará (Arica)*, 32, 2000. doi: 10.4067/S0717-73562000000100003.
- [3] C. B. Cunha and B. A. Cunha. Brief history of the clinical diagnosis of malaria: from hippocrates to osler. *J Vector Borne Dis*, 45(3):194–9, 2008.
- [4] M. Pombi, D. Modiano, and G. Corbellini. *Malaria eradication in Italy: the story of a first success*. Oxford University Press, 2018. ISBN 9780198789833. doi: 10.1093/oso/9780198789833.003.0013.
- [5] F. Nosten, D. Richard-Lenoble, and M. Danis. A brief history of malaria. *La Presse Médicale*, 51(3):104130, 2022. doi: <https://doi.org/10.1016/j.lpm.2022.104130>.
- [6] G. Band, Q. S. Le, G. M. Clarke, K. Kivinen, C. Hubbard, A. E. Jeffreys, K. Rowlands, E. M. Leffler, M. Jallow, D. J. Conway, F. Sisay-Joof, G. Sirugo, U. d’Alessandro, O. B. Toure, M. A. Thera, S. Konate, S. Sissoko, V. D. Mangano, E. C. Bougouma, S. B. Sirima, L. N. Amenga-Etego, A. K. Ghansah, A. V. O. Hodgson, M. D. Wilson, A. Enimil, D. Ansong, J. Evans, S. A. Ademola, T. O. Apinjoh, C. M. Ndila, A. Manjurano, C. Drakeley, H. Reyburn, N. H. Phu, N. T. N. Quyen, C. Q. Thai, T. T. Hien, Y. Y. Teo, L. Manning, M. Laman, P. Michon, H. Karunajeewa, P. Siba, S. Allen,

- A. Allen, M. Bahlo, T. M. E. Davis, V. Simpson, J. Shelton, C. C. A. Spencer, G. B. J. Busby, A. Kerasidou, E. Drury, J. Stalker, A. Diltthey, A. J. Mentzer, G. McVean, K. A. Bojang, O. Doumbo, D. Modiano, K. A. Koram, T. Agbenyega, O. K. Amodu, E. Achidi, T. N. Williams, K. Marsh, E. M. Riley, M. Molyneux, T. Taylor, S. J. Dunstan, J. Farrar, I. Mueller, K. A. Rockett, D. P. Kwiatkowski, and Network of Malaria Genomic Epidemiology. Insights into malaria susceptibility using genome-wide data on 17,000 individuals from Africa, Asia and Oceania. *Nature Communications*, 10, 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-13480-z.
- [7] O. Miotto, A. Amambua-Ngwa, L. Amenga-Etego, M. M. A. Hamid, I. Adam, E. Aninagyei, T. Apinjoh, G. A. Awandare, P. Bejon, G. I. Bertin, M. Bouyou-Akotet, A. Claessens, D. J. Conway, U. D'Alessandro, M. Diakite, A. Djimdé, A. M. Dondorp, P. Duffy, R. M. Fairhurst, C. I. Fanello, A. Ghansah, D. Ishengoma, M. Lawniczak, O. Maïga-Ascofaré, S. Auburn, A. Rosanas-Urgell, V. Wasakul, N. F. D. White, J. Almagro-Garcia, R. D. Pearson, S. Goncalves, C. Ariani, Z. Bozdech, W. Hamilton, V. Simpson, and D. P. Kwiatkowski. A complex *Plasmodium falciparum* cryptotype circulating at low frequency across the African continent. *bioRxiv*, 2024. doi: 10.1101/2024.01.20.576496.
- [8] S. C. Galen, J. Borner, E. S. Martinsen, J. Schaer, C. C. Austin, C. J. West, and S. L. Perkins. The polyphyly of *Plasmodium*: comprehensive phylogenetic analyses of the malaria parasites (order Haemosporida) reveal widespread taxonomic conflict. *Royal Society Open Science*, 5, 2018. doi: 10.1098/rsos.171780.
- [9] *World Malaria Report 2023*. World Health Organization, Geneva, 2023. Licence: CC BY-NC-SA 3.0 IGO.
- [10] R. N. Price, R. J. Commons, K. E. Battle, K. Thriemer, and K. Mendis. *Plasmodium vivax* in the era of the shrinking *P. falciparum* map. *Trends in Parasitology*, 36:560–570, 2020. doi: 10.1016/j.pt.2020.03.009.

- [11] D. P. Kwiatkowski. How malaria has affected the human genome and what human genetics can teach us about malaria. *American Journal of Human Genetics*, 77(2):171–192, 2005. ISSN 0002-9297. doi: 10.1086/432519.
- [12] N. J. White and A. M. Dondorp. *Malaria*. Oxford University Press, 2020. ISBN 9780198746690. doi: 10.1093/med/9780198746690.003.0160.
- [13] L. von Seidlein, R. Olaosebikan, I. C. E. Hendriksen, S. J. Lee, O. T. Adedoyin, T. Agbenyega, S. B. Nguah, K. Bojang, J. L. Deen, J. Evans, C. I. Fanello, E. Gomes, A. J. Pedro, C. Kahabuka, C. Karema, E. Kivaya, K. Maitland, O. A. Mokuolu, G. Mtove, J. Mwanga-Amumpaire, B. Nadjm, M. Nansumba, W. P. Ngum, M. A. Onyamboko, H. Reyburn, T. Sakulthaew, K. Silamut, A. K. Tshefu, N. Umulisa, S. Gesase, N. P. J. Day, N. J. White, and A. M. Dondorp. Predicting the clinical outcome of severe *falciparum* malaria in African children: Findings from a large randomized trial. *Clinical Infectious Diseases*, 54:1080–1090, 2012. doi: 10.1093/cid/cis034.
- [14] D. L. Doolan, C. Dobano, and J. K. Baird. Acquired immunity to malaria. *Clinical Microbiology Reviews*, 22:13–36, 2009. doi: 10.1128/CMR.00025-08.
- [15] S. J. Draper, B. K. Sack, C. R. King, C. M. Nielsen, J. C. Rayner, M. K. Higgins, C. A. Long, and R. A. Seder. Malaria vaccines: Recent advances and new horizons. *Cell Host & Microbe*, 24(1):43–56, 2018. ISSN 1931-3128. doi: 10.1016/j.chom.2018.06.008.
- [16] SK Nilsson, LM Childs, C Buckee, and M Marti. Targeting human transmission biology for malaria elimination. *PLOS Pathogens*, 11, 2015. doi: 10.1371/journal.ppat.1004871.
- [17] P. E. Duffy, J. P. Gorres, S. A. Healy, and M. Fried. Malaria vaccines: a new era of prevention and control. *Nature Reviews Microbiology*, 2024. ISSN 1740-1534. doi: 10.1038/s41579-024-01065-7.
- [18] M. P. Loscertales, S. Owens, J. O’Donnell, J. Bunn, X. Bosch-Capblanch, and B. J. Brabin. ABO blood group phenotypes and *Plasmodium falciparum* malaria: Unlocking a pivotal mechanism. In *Advances in Par-*

- asitology*, volume 65, pages 1–50. 2007. ISBN 978-0-12-374166-0. doi: 10.1016/S0065-308X(07)65001-5.
- [19] P. E. Duffy and J. P. Gorres. Malaria vaccines since 2000: progress, priorities, products. *NPJ Vaccines*, 5, 2020. doi: 10.1038/s41541-020-0196-3.
 - [20] P. Schneider and S. E. Reece. The private life of malaria parasites: Strategies for sexual reproduction. *Molecular and Biochemical Parasitology*, 244, 2021. doi: 10.1016/j.molbiopara.2021.111375.
 - [21] J. Pryce, M. Richardson, and C. Lengeler. Insecticide-treated nets for preventing malaria. *Cochran Database of Systematic Reviews*, 2018. doi: 10.1002/14651858.CD000363.pub3.
 - [22] S. Bhatt, D. J. Weiss, E. Cameron, D. Bisanzio, B. Mappin, U. Dalrymple, K. E. Battle, C. L. Moyes, A. Henry, P. A. Eckhoff, E. A. Wenger, O. Briët, M. A. Penny, T. A. Smith, A. Bennett, J. Yukich, T. P. Eisele, J. T. Griffin, C. A. Fergus, M. Lynch, F. Lindgren, J. M. Cohen, C. L. J. Murray, D. L. Smith, S. I. Hay, R. E. Cibulskis, and P. W. Gething. The effect of malaria control on *Plasmodium falciparum* in Africa between 2000 and 2015. *Nature*, 526:207, 2015. doi: 10.1038/nature15535.
 - [23] H. W. Choi, J. G. Breman, S. M. Teutsch, S. M. Liu, A. W. Hightower, and J. D. Sexton. The effectiveness of insecticide-impregnated bed nets in reducing cases of malaria infection - a meta-analysis of published results. *American Journal of Tropical Medicine and Hygiene*, 52:377–382, 1995.
 - [24] G. G. Yang, D. Kim, A. Pham, and C. J. Paul. A meta-regression analysis of the effectiveness of mosquito nets for malaria control: The value of long-lasting insecticide nets. *International Journal of Environmental Research and Public Health*, 15, 2018. doi: 10.3390/ijerph15030546.
 - [25] N. J. Govella, P. P. Chaki, and G. F. Killeen. Entomological surveillance of behavioural resilience and resistance in residual malaria vector populations. *Malaria Journal*, 12, 2013. doi: 10.1186/1475-2875-12-124.

- [26] K. Kyrou, A. M. Hammond, R. Galizi, N. Kranjc, A. Burt, A. K. Beaghton, T. Nolan, and A. Crisanti. A crispr-cas9 gene drive targeting *double-sex* causes complete population suppression in caged *Anopheles gambiae* mosquitoes. *Nature Biotechnology*, 36:1062, 2018. doi: 10.1038/nbt.4245.
- [27] *Malaria vaccine: WHO position paper*, volume 99. World Health Organization, Geneva, 2024.
- [28] A. J. Pollard and E. M. Bijker. A guide to vaccinology: from basic principles to new developments. *Nature Reviews Immunology*, 21:83–100, 2021. doi: 10.1038/s41577-020-00479-7.
- [29] M. S. Dattoo, A. Dicko, H. Tinto, J. B. Ouédraogo, M. Hamaluba, A. Olotu, E. Beaumont, F. R. Lopez, H. M. Natama, S. Weston, M. Chemba, Y. D. Compaore, D. Issiaka, D. Salou, A. M. Some, S. Omenda, A. Lawrie, P. Bejon, H. Rao, D. Chandramohan, R. Roberts, S. Bharati, L. Stockdale, S. Gairola, B. M. Greenwood, K. J. Ewer, J. Bradley, P. S. Kulkarni, U. Shaligram, A. V. S. Hill, and R. 21 Matrix-M Phase 3 Trial Grp. Safety and efficacy of malaria vaccine candidate R21/Matrix-M in African children: a multicentre, doubleblind, randomised, phase 3 trial. *Lancet*, 403:533–544, 2024. doi: 10.1016/S0140-6736(23)02511-4.
- [30] Côte d’Ivoire makes history as first nation to deploy R21/Matrix-M malaria vaccine, 2024. URL <https://www.ox.ac.uk/news/2024-07-15-c-te-d-ivoire-makes-history-first-nation-deploy-r21matrix-m-malaria-vaccine>. Accessed: 2024-07-04.
- [31] D. E. Neafsey, M. Juraska, T. Bedford, D. C. Benkeser, C. Valim, A. D. Griggs, M. Lievens, T. Agbenyega, S. Anderson, M. Bruls, K. M. Connolly, U. D’Alessandro, C. Dobano, J. Grimsby, M. J. Hamel, A. Leach, N. J. Lennon, E. L. Moss, D. J. Park, K. G. Pelle, D. Robbins, C. Russ, E. M. Ryan, B. Sogoloff, S. K. Volkman, Q. Yu, D. Lapierre, B. W. Birren, P. B. Gilbert, and D. F. Wirth. Parasite genetic diversity and protective efficacy in a phase 3 trial of the RTS,S/AS01 malaria vaccine. *American Journal of Tropical Medicine and Hygiene*, 93(4):112–112, 2015. ISSN 0002-9637.

- [32] V. K. Goel, X. R. Li, H. Q. Chen, S. C. Liu, A. H. Chishti, and S. S. Oh. Band 3 is a host receptor binding merozoite surface protein 1 during the *Plasmodium falciparum* invasion of erythrocytes. *Proceedings of the National Academy of Sciences of the United States of America*, 100:5164–5169, 2003. doi: 10.1073/pnas.0834959100.
- [33] A. Deshmukh, B. K. Chourasia, S. Mehrotra, I. H. Kana, G. Paul, A. Panda, I. Kaur, S. K. Singh, S. Rathore, A. Das, P. Gupta, M. Kalamuddin, S. K. Gakhar, A. Mohammed, M. Theisen, and P. Malhotra. *Plasmodium falciparum* msp3 exists in a complex on the merozoite surface and generates antibody response during natural infection. *Infection and Immunity*, 86, 2018. doi: 10.1128/IAI.00067-18.
- [34] G. H. Mitchell, A. W. Thomas, G. Margos, A. R. Dlugewski, and L. H. Bannister. Apical membrane antigen 1, a major malaria vaccine candidate, mediates the close attachment of invasive merozoites to host red blood cells. *Infection and Immunity*, 72:154–158, 2004. doi: 10.1128/IAI.72.1.154-158.2004.
- [35] S. Dankwa, M. Chaand, U. Kanjee, R. H. Y. Jiang, L. V. Nobre, J. M. Goldberg, A. K. Bei, M. A. Moechtar, C. Grüning, A. D. Ahouidi, D. Ndiaye, T. N. Dieye, S. Mboup, M. P. Weekes, and M. T. Duraisingh. Genetic evidence for erythrocyte receptor glycophorin b expression levels defining a dominant *Plasmodium falciparum* invasion pathway into human erythrocytes. *Infection and Immunity*, 85, 2017. doi: 10.1128/IAI.00074-17.
- [36] P. A. Orlandi, F. W. Klotz, and J. D. Haynes. A malaria invasion receptor, the 175-kilodalton erythrocyte binding antigen of *Plasmodium falciparum* recognizes the terminal Neu5Ac(α -2-3)gal- sequences of Glycophorin-A. *Journal of Cell Biology*, 116:901–909, 1992.
- [37] M. M. Abdel Hamid, M. H. Abdelraheem, D. O. Acheampong, A. Ahouidi, M. Ali, J. Almagro-Garcia, A. Amambua-Ngwa, C. Amaratunga, L. Amenga-Etego, B. Andagalu, T. Anderson, V. Andrianaranjaka, I. Aniebo, E. Aninagyei, F. Ansah, P. O. Ansah, T. Apinjoh, P. Ar-

naldo, E. Ashley, S. Auburn, G. A. Awandare, H. Ba, V. Baraka, A. Barry, P. Bejon, G. I. Bertin, M. F. Boni, S. Borrmann, T. Bousema, M. Bouyou-Akotet, O. Branch, P. C. Bull, H. Cheah, K. Chindavongsa, T. Chookajorn, K. Chotivanich, A. Claessens, D. J. Conway, V. Corredor, E. Courtier, A. Craig, U. D'Alessandro, S. Dama, N. P. J. Day, B. Denis, M. Dhorda, M. Diakite, A. Djimde, C. Dolecek, A. Dondorp, S. Doumbia, C. Drakeley, E. Drury, P. Duffy, D. F. Echeverry, T. G. Egwang, S. M. M. Enosse, B. Erko, R. M. Fairhurst, A. Faiz, C. A. Fanello, M. Fleharty, M. Forbes, M. Fukuda, D. Gamboa, A. Ghansah, L. Golassa, S. Goncalves, G. L. A. Harrison, S. A. Healy, J. A. Hendry, A. Hernandez-Koutoucheva, T. T. Hien, C. A. Hill, F. Hombhanje, A. Hott, Y. Htut, M. Hussein, M. Imwong, D. Ishengoma, S. A. Jackson, C. G. Jacob, J. Jeans, K. J. Johnson, C. Kamaliddin, E. Kamau, J. Keatley, T. Kochakarn, D. S. Konate, A. Konaté, A. Kone, D. P. Kwiatkowski, M. P. Kyaw, D. Kyle, M. Lawniczak, S. K. Lee, M. Lemnge, P. Lim, C. Lon, K. M. Loua, C. I. Mandara, J. Marfurt, K. Marsh, R. J. Maude, M. Mayxay, O. Maïga-Ascofaré, O. Miotto, T. Mita, V. Mobegi, A. O. Mohamed, O. A. Mokuolu, J. Montgomery, C. M. Morang'a, I. Mueller, K. Murie, P. N. Newton, T. Ngo Duc, T. Nguyen, T. N. Nguyen, T. Nguyen Thi Kim, H. Nguyen Van, H. Noedl, F. Nosten, R. Noviyanti, V. N. N. Ntui, A. Nzila, L. I. Ochola-Oyier, H. Ocholla, A. Oduro, I. Omedo, M. A. Onyamboko, J. B. Ouedraogo, K. Oyebola, W. A. Oyibo, R. Pearson, N. Peshu, A. P. Phyoo, C. V. Plowe, R. N. Price, S. Pukritayakamee, H. H. Quang, M. Randrianarivelojosia, J. C. Rayner, P. Ringwald, A. Rosanas-Urgell, E. Rovira-Vallbona, V. Ruano-Rubio, L. Ruiz, D. Saunders, A. Shayo, P. Siba, V. J. Simpson, M. S. Sissoko, C. Smith, X. Z. Su, C. Sutherland, S. Takala-Harrison, A. Talman, L. Tavul, N. V. Thanh, V. Thathy, A. M. Thu, M. Toure, A. Tshefu, F. Verra, J. Vinetz, T. E. Wellems, J. Wendler, N. J. White, G. Whitton, W. Yavo, and R. W. van der Pluijm. Pf7: an open dataset of *Plasmodium falciparum* genome variation in 20,000 worldwide samples. *Wellcome Open Research*, 8(22), 2023. doi: 10.12688/wellcomeopenres.18681.1.

- [38] K. Galinsky, C. Valim, A. Salmier, B. de Thoisy, L. Musset, E. Legrand,

- A. Faust, M. L. Baniecki, D. Ndiaye, R. F. Daniels, D. L. Hartl, P. C. Sabeti, D. F. Wirth, S. K. Volkman, and D. E. Neafsey. Coil: a methodology for evaluating malarial complexity of infection using likelihood from single nucleotide polymorphism data. *Malaria Journal*, 14, 2015. doi: 10.1186/1475-2875-14-4.
- [39] K. Tanabe, N. Sakihama, O. Kaneko, A. Saito-Ito, and M. Kimura. A PCR method for molecular epidemiology of *Plasmodium falciparum* msp-1. *Tokai Journal of Experimental and Clinical Medicine*, 23(6):375–381, 1998.
- [40] M. Wanaguru, W. M. Liu, B. H. Hahn, J. C. Rayner, and G. J. Wright. Rh5-basigin interaction plays a major role in the host tropism of *Plasmodium falciparum*. *Proceedings of the National Academy of Sciences of the United States of America*, 110:20735–20740, 2013. doi: 10.1073/pnas.1320771110.
- [41] C. Crosnier, L. Y. Bustamante, S. J. Bartholdson, A. K. Bei, M. Theron, M. Uchikawa, S. Mboup, O. Ndir, D. P. Kwiatkowski, M. T. Duraisingh, J. C. Rayner, and G. J. Wright. Basigin is a receptor essential for erythrocyte invasion by *Plasmodium falciparum*. *Nature*, 480:534–U158, 2011. doi: 10.1038/nature10606.
- [42] L. Y. Bustamante, S. J. Bartholdson, C. Crosnier, M. G. Campos, M. Wanaguru, C. Nguon, D. P. Kwiatkowski, G. J. Wright, and J. C. Rayner. A full-length recombinant *Plasmodium falciparum* PfRH5 protein induces inhibitory antibodies that are effective across common PfRH5 genetic variants. *Vaccine*, 31:373–379, 2013. doi: 10.1016/j.vaccine.2012.10.106.
- [43] S. E. Silk, W. F. Kalinga, J. Salkeld, I. M. Mtaka, S. Ahmed, F. Milando, A. Diouf, C. K. Bundi, N. Balige, O. Hassan, C. G. Mkindi, S. Rwezaula, T. Athumani, S. Mswata, N. S. Lilolime, B. Simon, H. Msami, M. Mohamed, D. M. David, L. Mohammed, G. Nyaulingo, B. Mwalimu, O. Juma, T. G. Mwamlima, I. A. Sasamalo, R. P. Mkumbange, J. J. Kamage, J. R. Barrett, L. D. W. King, M. M. Hou, D. Pulido, C. Carnrot, A. M. Lawrie, R. E. Cowan, F. L. Nugent, R. Roberts, J. S. Cho, C. A. Long, C. M. Nielsen, K. Miura, S. J. Draper, A. I. Olotu, and A. M. Minassian. Blood-stage

malaria vaccine candidate RH5.1/Matrix-M in healthy Tanzanian adults and children; an open-label, non-randomised, first-in-human, single-centre, phase 1b trial. *The Lancet Infectious Diseases*, 24(10):1105–1117, 2024. doi: 10.1016/S1473-3099(24)00312-8.

- [44] H. M. Natama, J. Salkeld, A. Somé, S. Soremekun, S. Diallo, O. Traoré, T. Rouamba, F. Ouédraogo, E. Ouédraogo, K. C. S. Daboné, N. A. Koné, Z. M. J. Compaoré, M. Kafando, M. D. A. Bonko, F. Konaté, H. Sorgho, C. M. Nielsen, D. Pipini, A. Diouf, L. D. W. King, U. Shaligram, C. A. Long, J. S. Cho, A. M. Lawrie, K. Skinner, R. Roberts, K. Miura, J. Bradley, S. E. Silk, S. J. Draper, H. Tinto, and A. M. Minassian. Safety and efficacy of the blood-stage malaria vaccine RH5.1/Matrix-M in Burkina Faso: interim results of a double-blind, randomised, controlled, phase 2b trial in children. *Lancet Infect Dis*, 2024. doi: 10.1016/S1473-3099(24)00752-7.
- [45] J. Cheaveau, D. C. Mogollon, M. A. Mohon, L. Golassa, D. Yewhalaw, and D. R. Pillai. Asymptomatic malaria in the clinical and public health context. *Expert Review of Anti-infective Therapy*, 17:997–1010, 2019. doi: 10.1080/14787210.2019.1693259.
- [46] I. Sagara, S. A. Healy, M. H. Assadou, M. Kone, B. J. Swihart, J. L. Kwan, J. Fintzi, K. Sissoko, B. Kamate, Y. Samake, M. A. Guindo, M. Doucoure, K. Niare, A. Dolo, B. Diarra, K. M. Rausch, D. L. Narum, D. S. Jones, N. J. Macdonald, D. M. Zhu, J. P. Gorres, A. Imeru, R. Mohan, I. Thera, I. Zaidi, F. Salazar-Miralles, J. H. Duan, J. Neal, R. D. Morrison, O. Muratova, D. Sylla, E. M. O’Connell, Y. M. Wu, J. C. C. Hume, M. B. Coulibaly, C. F. Anderson, S. F. Traore, O. K. Doumbo, and P. E. Duffy. Malaria transmission-blocking vaccines Pfs230D1-EPA and Pfs25-EPA in alhydrogel in healthy Malian adults; a phase 1 randomised, controlled trial. *The Lancet Infectious Diseases*, 23:1266–1279, 2023. doi: 10.1016/S1473-3099(23)00276-1.
- [47] *World Malaria Report 2021*. World Health Organization, Geneva, 2021. Licence: CC BY-NC-SA 3.0 IGO.

- [48] D. E. Neafsey, A. R. Taylor, and B. L. MacInnis. Advances and opportunities in malaria population genomics. *Nature Reviews Genetics*, 22(8):502–517, 2021. ISSN 1471-0056. doi: 10.1038/s41576-021-00349-5.
- [49] Z. Yu, T. H. H. Coorens, M. M. Uddin, K. G. Ardlie, N. Lennon, and P. Natarajan. Genetic variation across and within individuals. *Nature Reviews Genetics*, 25:548–562, 2024. doi: 10.1038/s41576-024-00709-x.
- [50] J. Shendure and H. Ji. Next-generation DNA sequencing. *Nature Biotechnology*, 26(10):1135–1145, 2008.
- [51] W. L. Hamilton, A. Claessens, T. D. Otto, M. Kekre, R. M. Fairhurst, J. C. Rayner, and D. Kwiatkowski. Extreme mutation bias and high AT content in *Plasmodium falciparum*. *Nucleic Acids Research*, 45(4):1889–1901, 2017. ISSN 0305-1048. doi: 10.1093/nar/gkw1259.
- [52] A. V. S. Hill, C. E. M. Allsopp, D. Kwiatkowski, N. M. Anstey, P. Twumasi, P. A. Rowe, S. Bennett, D. Brewster, A. J. McMichael, and B. M. Greenwood. Common West African HLA antigens are associated with protection from severe malaria. *Nature*, 352(6336):595–600, 1991. ISSN 0028-0836. doi: 10.1038/352595a0.
- [53] J. H. Sul, L. S. Martin, and E. Eskin. Population structure in genetic studies: Confounding factors and mixed models. *PLoS Genet*, 14(12), Dec 2018. doi: 10.1371/journal.pgen.1007309.
- [54] K. A. Rockett, G. M. Clarke, K. Fitzpatrick, C. Hubbart, A. E. Jeffreys, K. Rowlands, R. Craik, M. Jallow, D. J. Conway, K. A. Bojang, M. Pinder, S. Usen, F. Sisay-Joof, G. Sirugo, O. Toure, M. A. Thera, S. Konate, S. Sissoko, A. Niangaly, B. Poudiougou, V. D. Mangano, E. C. Bougouma, S. B. Sirima, D. Modiano, L. N. Amenga-Etego, A. Ghansah, K. A. Koram, M. D. Wilson, A. Enimil, J. Evans, O. Amodu, S. Olaniyan, T. Apinjoh, R. Mugri, A. Ndi, C. M. Ndila, S. Uyoga, A. Macharia, N. Peshu, T. N. Williams, A. Manjurano, E. Riley, C. Drakeley, H. Reyburn, V. Nyirongo, D. Kachala, M. Molyneux, S. J. Dunstan, N. H. Phu, N. T. N. Quyen, C. Q. Thai, T. T. Hien, L. Manning, M. Laman, P. Siba, H. Karunajeewa, S. Allen, A. Allen,

- T. M. E. Davis, P. Michon, I. Mueller, A. Green, S. Molloy, K. J. Johnson, A. Kerasidou, V. Cornelius, L. Hart, A. Vanderwal, M. SanJoaquin, G. Band, S. Q. Le, M. Pirinen, N. Sepúlveda, C. C. A. Spencer, T. G. Clark, T. Agbenyega, E. Achidi, O. Doumbo, J. Farrar, K. Marsh, T. Taylor, D. P. Kwiatkowski, and Network of Malaria Genomic Epidemiology. Reappraisal of known malaria resistance loci in a large multicenter study. *Nature Genetics*, 46(11):1197–1204, 2014. ISSN 1061-4036. doi: 10.1038/ng.3107.
- [55] D. Damena, A. Denis, L. Golassa, and E. R. Chimusa. Genome-wide association studies of severe *P. falciparum* malaria susceptibility: progress, pitfalls and prospects. *BMC Medical Genomics*, 12(1), 2019. doi: 10.1186/s12920-019-0564-x.
- [56] S. N. Kariuki and T. N. Williams. Human genetics and malaria resistance. *Human Genetics*, 139:801–811, 2020. doi: 10.1007/s00439-020-02142-6.
- [57] F. Flinter. Should we sequence everyone’s genome? no. *BMJ : British Medical Journal*, 346:f3132, 2013. doi: 10.1136/bmj.f3132. URL <https://www.bmj.com/content/bmj/346/bmj.f3132.full.pdf>.
- [58] J. A. M. Verlouw, E. Clemens, J. H. de Vries, O. Zolk, A. J. M. H. Verkerk, A. A. Zehnhoff-Dinnesen, C. Medina-Gomez, C. Lanvers-Kaminsky, F. Rivadeneira, T. Langer, J. B. J. van Meurs, M. M. van den Heuvel-eibrink, A. G. Uitterlinden, and L. Broer. A comparison of genotyping arrays. *European Journal of Human Genetics*, 29:1611–1624, 2021. doi: 10.1038/s41431-021-00917-7.
- [59] B. E. Slatko, A. F. Gardner, and F. M. Ausubel. Overview of next-generation sequencing technologies. *Current Protocols in Molecular Biology*, 122(1):e59, 2018. doi: 10.1002/cpmb.59.
- [60] J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, A. Bibillo, K. Bjornson, B. Chaudhuri, F. Christians, R. Cicero, S. Clark, R. Dalal, A. Dewinter, J. Dixon, M. Foquet, A. Gaertner, P. Hardenbol, C. Heiner, K. Hester, D. Holden, G. Kearns, X. X. Kong, R. Kuse, Y. Lacroix, S. Lin, P. Lundquist, C. C. Ma, P. Marks,

- M. Maxham, D. Murphy, I. Park, T. Pham, M. Phillips, J. Roy, R. Sebra, G. Shen, J. Sorenson, A. Tomaney, K. Travers, M. Trulson, J. Vieceli, J. Wegener, D. Wu, A. Yang, D. Zaccarin, P. Zhao, F. Zhong, J. Korlach, and S. Turner. Real-time DNA sequencing from single polymerase molecules. *Science*, 323:133–138, 2009. doi: 10.1126/science.1162986.
- [61] Y. H. Wang, Y. Zhao, A. Bollas, Y. R. Wang, and K. F. Au. Nanopore sequencing technology, bioinformatics and applications. *Nature Biotechnology*, 39:1348–1365, 2021. doi: 10.1038/s41587-021-01108-x.
- [62] M. Eisenstein. Illumina swallows PacBio in long shot for market domination. *Nature Biotechnology*, 37:2–3, 2019.
- [63] M. J. Godin, A. Sebastian, I. Albert, and S. E. Lindner. Long-read genome assembly and gene model annotations for the rodent malaria parasite *Plasmodium yoelii* 17XNL. *Journal of Biological Chemistry*, 299, 2023. doi: 10.1016/j.jbc.2023.104871.
- [64] Revio system: Reveal more with accurate long-read sequencing at scale. Pacific Biosciences, 2022. URL <https://www.pacb.com/wp-content/uploads/Revio-brochure.pdf>. Accessed: 2024-07-09.
- [65] J. Ha, R. Martinson, S. K. Iwamoto, and A. Nishi. Hemoglobin E, malaria and natural selection. *Evolution, Medicine, and Public Health*, pages 232–241, 2019. doi: 10.1093/emph/eoz034.
- [66] F. B. Piel, A. P. Patil, R. E. Howes, O. A. Nyangiri, P. W. Gething, M. Dewi, W. H. Temperley, T. N. Williams, D. J. Weatherall, and S. I. Hay. Global epidemiology of sickle haemoglobin in neonates: a contemporary geostatistical model-based map and population estimates. *Lancet*, 381:142–151, 2013. doi: 10.1016/S0140-6736(12)61229-X.
- [67] F. B. Piel, R. E. Howes, A. P. Patil, O. A. Nyangiri, P. W. Gething, S. Bhatt, T. N. Williams, D. J. Weatherall, and S. I. Hay. The distribution of haemoglobin C and its prevalence in newborns in Africa. *Scientific Reports*, 3, 2013. ISSN 2045-2322. doi: 10.1038/srep01671.

- [68] C. M. Ndila, S. Uyoga, A. W. Macharia, G. Nyutu, N. Peshu, J. Ojal, M. Shebe, K. O. Awuondo, N. Mturi, B. Tsofa, N. Sepúlveda, T. G. Clark, G. Band, G. Clarke, K. Rowlands, C. Hubbart, A. Jeffreys, S. Kariuki, K. Marsh, M. Mackinnon, K. Maitland, D. P. Kwiatkowski, K. A. Rockett, T. N. Williams, and Network of Malaria Genomic Epidemiology. Human candidate gene polymorphisms and risk of severe malaria in children in Kilifi, Kenya: a case-control association study. *Lancet Haematology*, 5:E333–E345, 2018. doi: 10.1016/S2352-3026(18)30107-8.
- [69] D. J. Weatherall and J. B. Clegg. Inherited haemoglobin disorders: an increasing global health problem. *Bulletin of the World Health Organization*, 79:704–712, 2001.
- [70] D. H. Opi, L. B. Ochola, M. Tendwa, B. R. Siddondo, H. Ocholla, H. Fanjo, A. Ghumra, D. J. P. Ferguson, J. A. Rowe, and T. N. Williams. Mechanistic studies of the negative epistatic malaria-protective interaction between sickle cell trait and alpha+ thalassemia. *eBioMedicine*, 1:29–36, 2014. doi: 10.1016/j.ebiom.2014.10.006.
- [71] T. N. Williams, T. W. Mwangi, S. Wambua, N. D. Alexander, M. Kortok, R. W. Snow, and K. Marsh. Sickle cell trait and the risk of *Plasmodium falciparum* malaria and other childhood diseases. *Journal of Infectious Diseases*, 192(1):178–186, 2005. ISSN 0022-1899. doi: 10.1086/430744.
- [72] V. Introini, A. Marin-Menendez, G. Nettesheim, Y. C. Lin, S. N. Kariuki, A. L. Smith, L. Jean, J. N. Brewin, D. C. Rees, P. Cicuta, J. C. Rayner, and B. S. Penman. The erythrocyte membrane properties of beta thalassaemia heterozygotes and their consequences for *Plasmodium falciparum* invasion. *Scientific Reports*, 12, 2022. doi: 10.1038/s41598-022-12060-4.
- [73] P. Gagneux and A. Varki. Evolutionary considerations in relating oligosaccharide diversity to biological function. *Glycobiology*, 9:747–755, 1999.
- [74] M. T. Hamblin and A. Di Rienzo. Detection of the signature of natural selection in humans: Evidence from the Duffy blood group locus. *American Journal of Human Genetics*, 66:1669–1679, 2000.

- [75] N. Kermarrec, F. Roubinet, P. A. Apoil, and A. Blancher. Comparison of allele o sequences of the human and non-human primate ABO system. *ImmunoGenetics*, 49:517–526, 1999.
- [76] J. A. Rowe, I. G. Handel, M. A. Thera, A. M. Deans, K. E. Lyke, A. Koné, D. A. Diallo, A. Raza, O. Kai, K. Marsh, C. V. Plowe, O. K. Doumbo, and J. M. Moulds. Blood group O protects against severe *Plasmodium falciparum* malaria through the mechanism of reduced rosetting. *Proceedings of the National Academy of Sciences of the United States of America*, 104:17471–17476, 2007. doi: 10.1073/pnas.0705390104.
- [77] G. Band, K. A. Rockett, C. C. A. Spencer, D. P. Kwiatkowski, Q. S. Le, G. M. Clarke, K. Kivinen, E. M. Leffler, V. Cornelius, D. J. Conway, T. N. Williams, T. Taylor, K. A. Bojang, O. Doumbo, M. A. Thera, D. Modiano, S. B. Sirima, M. D. Wilson, K. A. Koram, T. Agbenyega, E. Achidi, K. Marsh, H. Reyburn, C. Drakeley, E. Riley, M. Molyneux, M. Jallow, M. Pinder, O. B. Toure, S. Konate, S. Sissoko, E. C. Bougouma, V. D. Mangano, L. N. Amenga-Etego, A. K. Ghansah, A. V. O. Hodgson, D. Ansong, A. Enimil, J. Evans, T. O. Apinjoh, A. Macharia, C. M. Ndila, C. Newton, N. Peshu, S. Uyoga, A. Manjurano, D. Kachala, V. Nyirongo, D. Mead, E. Drury, S. Auburn, S. G. Campino, B. MacInnis, J. Stalker, E. Gray, C. Hubbart, A. E. Jeffreys, K. Rowlands, A. Mendy, R. Craik, K. Fitzpatrick, S. Molloy, L. Hart, R. Hutton, A. Kerasidou, K. J. Johnson, and Network of Malaria Genomic Epidemiology. A novel locus of resistance to severe malaria in a region of ancient balancing selection. *Nature*, 526(7572): 253, 2015. ISSN 0028-0836. doi: 10.1038/nature15390.
- [78] E. M. Leffler, Z. Y. Gao, S. Pfeifer, L. Segurel, A. Auton, O. Venn, R. Bowden, R. Bontrop, J. D. Wall, G. Sella, P. Donnelly, G. McVean, and M. Przeworski. Multiple instances of ancient balancing selection shared between humans and chimpanzees. *Science*, 339(6127):1578–1582, 2013. ISSN 0036-8075. doi: 10.1126/science.1234070.
- [79] S. N. Kariuki, A. Marin-Menendez, V. Introini, B. J. Ravenhill, Y. C. Lin, A. Macharia, J. Makale, M. Tendwa, W. Nyamu, J. Kotar, M. Carrasquilla,

- J. A. Rowe, K. Rockett, D. Kwiatkowski, M. P. Weekes, P. Cicuta, T. N. Williams, and J. C. Rayner. Red blood cell tension protects against severe malaria in the dantu blood group. *Nature*, 585:579, 2020. doi: 10.1038/s41586-020-2726-6.
- [80] N. J. White. Severe malaria. *Malaria Journal*, 21(1), 2022. doi: 10.1186/s12936-022-04301-8.
- [81] J. A. Watson, S. J. Leopold, J. A. Simpson, N. P. Day, A. M. Dondorp, and N. J. White. Collider bias and the apparent protective effect of glucose-6-phosphate dehydrogenase deficiency on cerebral malaria. *eLife*, 2019. doi: 10.7554/eLife.43154.
- [82] P. C. Sabeti, D. E. Reich, J. M. Higgins, H. Z. P. Levine, D. J. Richter, S. F. Schaffner, S. B. Gabriel, J. V. Planko, N. J. Patterson, G. J. McDonald, H. C. Ackerman, S. J. Campbell, D. Altshuler, R. Cooper, D. Kwiatkowski, R. Ward, and E. S. Lander. Detecting recent positive selection in the human genome from haplotype structure. *Nature*, 419(6909):832–837, 2002. ISSN 0028-0836. doi: 10.1038/nature01140.
- [83] C. Timmann, T. Thye, M. Vens, J. Evans, J. May, C. Ehmen, J. Sievertsen, B. Muntau, G. Ruge, W. Loag, D. Ansong, S. Antwi, E. Asafo-Adjei, S. B. Nguah, K. O. Kwakye, A. O. Y. Akoto, J. Sylverken, M. Brendel, K. Schuldt, C. Loley, A. Franke, C. G. Meyer, T. Agbenyega, A. Ziegler, and R. D. Horstmann. Genome-wide association study indicates two novel resistance loci for severe malaria. *Nature*, 489(7416):443–446, 2012. ISSN 0028-0836. doi: 10.1038/nature11334.
- [84] S. Lessard, E. S. Gatof, M. Beaudoin, P. G. Schupp, F. Sher, A. Ali, S. Prehar, R. Kurita, Y. Nakamura, E. Baena, J. Ledoux, D. Oceandy, D. E. Bauer, and G. Lettre. An erythroid-specific *ATP2B4* enhancer mediates red blood cell hydration and malaria susceptibility. *Journal of Clinical Investigation*, 127:3065–3074, 2017. doi: 10.1172/JCI94378.
- [85] A. Jamwal, C. F. Constantin, S. Hirschi, S. Henrich, W. Bildl, B. Fakler, S. J. Draper, U. Schulte, and M. K. Higgins. Erythrocyte invasion-neutralising

- antibodies prevent *Plasmodium falciparum* RH5 from binding to basigin-containing membrane protein complexes. *eLife*, 12:e83681, 2023. doi: 10.7554/eLife.83681.
- [86] J. K. Kulski, S. Suzuki, and T. Shiina. Human Leukocyte Antigen super-locus: nexus of genomic supergenes, SNPs, INDELs, transcripts, and haplotypes. *Human Genome Variation*, 9, 2022. doi: 10.1038/s41439-022-00226-5.
- [87] A. J. Kwok, A. Mentzer, and J. C. Knight. Host genetics and infectious disease: new tools, insights and translational opportunities. *Nature Reviews Genetics*, 22:137–153, 2021. doi: 10.1038/s41576-020-00297-6.
- [88] A. V. S. Hill, J. Elvin, A. C. Willis, M. Aidoo, C. E. M. Allsopp, F. M. Gotch, X. M. Gao, M. Takiguchi, B. M. Greenwood, A. R. M. Townsend, A. J. McMichael, and H. C. Whittle. Molecular analysis of the association of HLA-B*53 and resistance to severe malaria. *Nature*, 360(6403):434–439, 1992. ISSN 0028-0836. doi: 10.1038/360434a0.
- [89] M. Aidoo, A. Lalvani, S. C. Gilbert, J. T. Hu, P. Daubersies, N. Hurt, H. C. Whittle, P. Druihle, and A. V. S. Hill. Cytotoxic T-lymphocyte epitopes for HLA-B53 and other HLA types in the malaria vaccine candidate liver-stage antigen 3. *Infection and Immunity*, 68(1):227–232, 2000. ISSN 0019-9567. doi: 10.1128/iai.68.1.227-232.2000.
- [90] A. V. S. Hill, S. N. R. Yates, C. E. M. Allsopp, S. Gupta, S. C. Gilbert, A. Lalvani, M. Aidoo, M. Davenport, and M. Plebanski. Human-leukocyte antigens and natural selection by malaria. *Philosophical Transactions of the Royal Society B-Biological Sciences*, 346(1317):379–385, 1994. ISSN 0962-8436. doi: 10.1098/rstb.1994.0155.
- [91] K. E. Lyke, M. A. Fernandez-Vina, K. Cao, J. Hollenbach, D. Coulibaly, A. K. Kone, A. Guindo, L. A. Burdett, R. J. Hartzman, A. R. Wahl, W. H. Hildebrand, O. K. Doumbo, C. V. Plowe, and M. B. Sztein. Association of HLA alleles with *Plasmodium falciparum* severity in Malian children. *Tissue Antigens*, 77(6):562–571, 2011. ISSN 0001-2815. doi: 10.1111/j.1399-0039.2011.01661.x.

- [92] J. Fellay, K. V. Shianna, D. L. Ge, S. Colombo, B. Ledergerber, M. Weale, K. L. Zhang, C. Gumbs, A. Castagna, A. Cossarizza, A. Cozzi-Lepri, A. De Luca, P. Easterbrook, P. Francioli, S. Mallal, J. Martinez-Picado, J. M. Miro, N. Obel, J. P. Smith, J. Wyniger, P. Descombes, S. E. Antonarakis, N. L. Letvin, A. J. McMichael, B. F. Haynes, A. Telenti, and D. B. Goldstein. A whole-genome association study of major determinants for host control of HIV-1. *Science*, 317:944–947, 2007. doi: 10.1126/science.1143767.
- [93] J. G. Beeson, F. H. A. Osier, and C. R. Engwerda. Recent insights into humoral and cellular immune responses against malaria. *Trends in Parasitology*, 24:578–584, 2008. doi: 10.1016/j.pt.2008.08.008.
- [94] M. J. Gardner, N. Hall, E. Fung, O. White, M. Berriman, R. W. Hyman, J. M. Carlton, A. Pain, K. E. Nelson, S. Bowman, I. T. Paulsen, K. James, J. A. Eisen, K. Rutherford, S. L. Salzberg, A. Craig, S. Kyes, M. S. Chan, V. Nene, S. J. Shallom, B. Suh, J. Peterson, S. Angiuoli, M. Pertea, J. Allen, J. Selengut, D. Haft, M. W. Mather, A. B. Vaidya, D. M. A. Martin, A. H. Fairlamb, M. J. Fraunholz, D. S. Roos, S. A. Ralph, G. I. McFadden, L. M. Cummings, G. M. Subramanian, C. Mungall, J. C. Venter, D. J. Carucci, S. L. Hoffman, C. Newbold, R. W. Davis, C. M. Fraser, and B. Barrell. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature*, 419(6906):498–511, 2002. ISSN 0028-0836. doi: 10.1038/nature01097.
- [95] I. Kozarewa, Z. M. Ning, M. A. Quail, M. J. Sanders, M. Berriman, and D. J. Turner. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G plus C)-biased genomes. *Nature Methods*, 6(4):291–295, 2009. ISSN 1548-7091. doi: 10.1038/nmeth.1311.
- [96] Improved library preparation with the NEBNext Ultra II DNA Library Prep Kit for Illumina. New England Biolabs. URL <https://www.neb-online.de/wp-content/uploads/2020/03/Improved-library-preparation-with-the-NEBNext-Ultra-II-DNA-Library-Prep-Kit-for-Illumina-E7645.pdf>. Accessed: 2024-07-10.

- [97] C. da Silva, S. Boene, D. Datta, E. Rovira-Vallbona, A. Aranda-Díaz, P. Cisteró, N. Hathaway, S. Tessema, A. Chidimatembue, G. Matambisso, A. Nhama, E. Macete, A. Pujol, L. Nhamussua, B. Galatas, C. Guinovart, S. Enosse, E. De Carvalho, E. Rogier, M. M. Plucinski, J. Colborn, R. Zulliger, A. Saifodine, P. L. Alonso, B. Candrinho, B. Greenhouse, P. Aide, F. Saute, and A. Mayor. Targeted and whole-genome sequencing reveal a north-south divide in *P. falciparum* drug resistance markers and genetic structure in Mozambique. *Communications Biology*, 6, 2023. doi: 10.1038/s42003-023-04997-7.
- [98] N. B. Teyssier, A. N. Chen, E. M. Duarte, R. Sit, B. Greenhouse, and S. K. Tessema. Optimization of whole-genome sequencing of *Plasmodium falciparum* from low-density dried blood spot samples. *Malaria Journal*, 20(1), 2021. doi: 10.1186/s12936-021-03630-4.
- [99] S. O. Oyola, C. V. Ariani, W. L. Hamilton, M. Kekre, L. N. Amenga-Etego, A. Ghansah, G. G. Rutledge, S. Redmond, M. Manske, D. Jyothi, C. G. Jacob, T. D. Otto, K. Rockett, C. I. Newbold, M. Berriman, and D. P. Kwiatkowski. Whole genome sequencing of *Plasmodium falciparum* from dried blood spots using selective whole genome amplification. *Malaria Journal*, 15, 2016. ISSN 1475-2875. doi: 10.1186/s12936-016-1641-7.
- [100] C. Alkan, B. P. Coe, and E. E. Eichler. Applications of next-generation sequencing: genome structural variation discovery and genotyping. *Nature Reviews Genetics*, 12:363–375, 2011. doi: 10.1038/nrg2958.
- [101] T. J. Treangen and S. L. Salzberg. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics*, 13:36–46, 2012. doi: 10.1038/nrg3117.
- [102] A. Miles, Z. Iqbal, P. Vauterin, R. Pearson, S. Campino, M. Theron, K. Gould, D. Mead, E. Drury, J. O’Brien, V. R. Rubio, B. MacInnis, J. Mwangi, U. Samarakoon, L. Ranford-Cartwright, M. Ferdig, K. Hayton, X. Z. Su, T. Wellems, J. Rayner, G. McVean, and D. Kwiatkowski. Indels,

structural variation, and recombination drive genomic diversity in *Plasmodium falciparum*. *Genome Research*, 26(9):1288–1299, 2016. ISSN 1088-9051. doi: 10.1101/gr.203711.115.

- [103] T. D. Otto, U. Böhme, M. Sanders, A. Reid, E. I. Bruske, C. W. Duffy, P. C. Bull, R. D. Pearson, A. Abdi, S. Dimonte, L. B. Stewart, S. Campino, M. Kekre, W. L. Hamilton, A. Claessens, S. K. Volkman, D. Ndiaye, A. Amambua-Ngwa, M. Diakite, R. M. Fairhurst, D. J. Conway, M. Franck, C. I. Newbold, and M. Berriman. Long read assemblies of geographically dispersed *Plasmodium falciparum* isolates reveal highly structured subtelomeres. *Wellcome Open Res*, 3:52, 2018. doi: 10.12688/wellcomeopenres.14571.1.
- [104] M. de Cesare, M. Mwenda, A. E. Jeffreys, J. Chirwa, C. Drakeley, K. Schneider, B. Mambwe, K. Glanz, C. Ntalla, M. Carrasquilla, S. Portugal, R. J. Verity, J. A. Bailey, I. Ghinai, G. B. Busby, B. Hamainza, M. Hawela, D. J. Bridges, and J. A. Hendry. Flexible and cost-effective genomic surveillance of *P. falciparum* malaria with targeted nanopore sequencing. *Nature Communications*, 15, 2024. doi: 10.1038/s41467-024-45688-z.
- [105] L. S. Ross and D. A. Fidock. Elucidating mechanisms of drug-resistant *Plasmodium falciparum*. *Cell Host & Microbe*, 26:35–47, 2019. doi: 10.1016/j.chom.2019.06.001.
- [106] A. A. Escalante, D. L. Smith, and Y. Kim. The dynamics of mutations associated with anti-malarial drug resistance in *Plasmodium falciparum*, 2009.
- [107] A. M. Talman, O. Domarle, F. E. McKenzie, F. Ariey, and V. Robert. Gametocytogenesis: the puberty of *Plasmodium falciparum*. *Malaria Journal*, 3, 2004. doi: 10.1186/1475-2875-3-24.
- [108] A. Amambua-Ngwa, L. Amenga-Etego, E. Kamau, R. Amato, A. Ghansah, L. Golassa, M. Randrianarivelojosia, D. Ishengoma, T. Apinjoh, O. Maiga-Ascofare, B. Andagalu, W. Yavo, M. Bouyou-Akotet, O. Kolapo, K. Mane, A. Worwui, D. Jeffries, V. Simpson, U. D’Alessandro, D. Kwiatkowski, and

- A. A. Djimde. Major subpopulations of *Plasmodium falciparum* in sub-Saharan Africa. *Science*, 365(6455):813, 2019. ISSN 0036-8075. doi: 10.1126/science.aav5427.
- [109] S. Pou, R. W. Winter, A. Nilsen, J. X. Kelly, Y. X. Li, J. S. Doggett, E. W. Riscoe, K. W. Wegmann, D. J. Hinrichs, and M. K. Riscoe. Sontochin as a guide to the development of drugs against chloroquine-resistant malaria. *Antimicrobial Agents and Chemotherapy*, 56:3475–3480, 2012. doi: 10.1128/AAC.00100-12.
- [110] A. Djimdé, O. K. Doumbo, J. F. Cortese, K. Kayentao, S. Doumbo, Y. Diourté, A. Dicko, X. Z. Su, T. Nomura, D. A. Fidock, T. E. Wellem, C. V. Plowe, and D. Coulibaly. A molecular marker for chloroquine-resistant *falciparum* malaria. *New England Journal of Medicine*, 344:257–263, 2001.
- [111] A. Ecker, A. M. Lehane, J. Clain, and D. A. Fidock. PfCRT and its role in antimalarial drug resistance. *Trends in Parasitology*, 28:504–514, 2012. doi: 10.1016/j.pt.2012.08.002.
- [112] A. Amambua-Ngwa, K. A. Button-Simons, X. Li, S. Kumar, K. V. Breneman, M. Ferrari, L. A. Checkley, M. T. Haile, D. A. Shoue, M. McDew-White, S. M. Tindall, A. Reyes, E. Delgado, H. Dalhoff, J. K. Larbalestier, R. Amato, R. D. Pearson, A. B. Taylor, F. H. Nosten, U. D’Alessandro, D. Kwiatkowski, I. H. Cheeseman, S. H. I. Kappe, S. V. Avery, D. J. Conway, A. M. Vaughan, M. T. Ferdig, and T. J. C. Anderson. Chloroquine resistance evolution in *Plasmodium falciparum* is mediated by the putative amino acid transporter *aat1*. *Nature Microbiology*, 8(7):1213–+, 2023. ISSN 2058-5276. doi: 10.1038/s41564-023-01377-z.
- [113] A. E. P. Frosch, M. K. Laufer, D. P. Mathanga, S. Takala-Harrison, J. Skarbinski, C. W. Claassen, F. K. Dzinjalama, and C. V. Plowe. Return of widespread chloroquine-sensitive *Plasmodium falciparum* to Malawi. *Journal of Infectious Diseases*, 210:1110–1114, 2014. doi: 10.1093/infdis/jiu216.

- [114] C. Wongsrichanalai, A. L. Pickard, W. H. Wernsdorfer, and S. R. Meshnick. Epidemiology of drug-resistant malaria. *Lancet Infectious Diseases*, 2:209–218, 2002.
- [115] K. X. Si, X. S. He, L. P. Chen, A. Q. Zhang, C. Y. Guo, and M. H. Li. The structure of *Plasmodium falciparum* multidrug resistance protein 1 reveals an N-terminal regulatory domain. *Proceedings of the National Academy of Sciences of the United States of America*, 120, 2023. doi: 10.1073/pnas.2219905120.
- [116] S. Borges, P. Cravo, A. Creasey, R. Fawcett, K. Modrzynska, L. Rodrigues, A. Martinelli, and P. Hunt. Genomewide scan reveals amplification of *MDR1* as a common denominator of resistance to mefloquine, lumefantrine, and artemisinin in *Plasmodium chabaudi* malaria parasites. *Antimicrobial Agents and Chemotherapy*, 55:4858–4865, 2011. doi: 10.1128/AAC.01748-10.
- [117] A. Hott, D. Casandra, K. N. Sparks, L. C. Morton, G. G. Castanares, A. Rutter, and D. E. Kyle. Artemisinin-resistant *Plasmodium falciparum* parasites exhibit altered patterns of development in infected erythrocytes. *Antimicrobial Agents and Chemotherapy*, 59:3156–3167, 2015. doi: 10.1128/AAC.00197-15.
- [118] M. M. Ippolito, K. A. Moser, J. B. B. Kabuya, C. Cunningham, and J. J. Juliano. Antimalarial drug resistance and implications for the WHO global technical strategy. *Current Epidemiology Reports*, 8:46–62, 2021. doi: 10.1007/s40471-021-00266-5.
- [119] J. Straimer, P. Gandhi, K. C. Renner, and E. K. Schmitt. High prevalence of plasmodium falciparum k13 mutations in rwanda is associated with slow parasite clearance after treatment with artemether-lumefantrine. *J Infect Dis*, 225(8):1411–1414, Apr 2022. doi: 10.1093/infdis/jiab352.
- [120] D. S. Ishengoma, C. I. Mandara, C. Bakari, A. A. Fola, R. A. Madebe, M. D. Seth, F. Francis, C. C. Buguzi, R. Moshi, I. Garimo, S. Lazaro, A. Lusasi, S. Aaron, F. Chacky, A. Mohamed, R. J. A. Njau, J. Kitau, C. Rasmussen, J. A. Bailey, J. J. Juliano, and M. Warsame. Evidence of

artemisinin partial resistance in northwestern tanzania: clinical and molecular markers of resistance. *Lancet Infect Dis*, 24(11):1225–1233, Nov 2024. doi: 10.1016/S1473-3099(24)00362-1.

- [121] L. Ndwiga, K. M. Kimenyi, K. Wamae, V. Osoti, M. Akinyi, I. Omedo, D. S. Ishengoma, N. Duah-Quashie, B. Andagalu, A. Ghansah, A. Amambua-Ngwa, S. Tukwasibwe, S. K. Tessema, C. Karema, A. A. Djimde, A. M. Dondorp, J. Raman, R. W. Snow, P. Bejon, and L. I. Ochola-Oyier. A review of the frequencies of *Plasmodium falciparum* kelch 13 artemisinin resistance mutations in Africa. *International Journal for Parasitology: Drugs and Drug Resistance*, 16:155–161, 2021. doi: 10.1016/j.ijpddr.2021.06.001.
- [122] L. Hviid and A. T. R. Jensen. *PfEMP1 – A Parasite Protein Family of Key Importance in Plasmodium falciparum Malaria Immunity and Pathogenesis*, volume 88, pages 51–84. Academic Press, 2015. ISBN 0065-308X. doi: <https://doi.org/10.1016/bs.apar.2015.02.004>.
- [123] W. C. Lee, B. Russell, and L. Renia. Sticking for a cause: The *falciparum* malaria parasites cytoadherence paradigm. *Frontiers in Immunology*, 10, 2019. doi: 10.3389/fimmu.2019.01444.
- [124] L. B. Ochola, B. R. Siddondo, H. Ocholla, S. Nkya, E. N. Kimani, T. N. Williams, J. O. Makale, A. Liljander, B. C. Urban, P. C. Bull, T. Szeszak, K. Marsh, and A. G. Craig. Specific receptor usage in *Plasmodium falciparum* cytoadherence is associated with disease outcome. *PLOS ONE*, 6, 2011. doi: 10.1371/journal.pone.0014741.
- [125] J. A. Rowe, J. M. Moulds, C. I. Newbold, and L. H. Miller. *P. falciparum* rosetting mediated by a parasite-variant erythrocyte membrane protein and complement-receptor 1. *Nature*, 388:292–295, 1997.
- [126] N. D. Pasternak and R. Dzikowski. Pfemp1: An antigen that plays a key role in the pathogenicity and immune evasion of the malaria parasite *Plasmodium falciparum*. *International Journal of Biochemistry & Cell Biology*, 41:1463–1466, 2009. doi: 10.1016/j.biocel.2008.12.012.

- [127] M. Rottmann, T. Lavstsen, J. P. Mugasa, M. Kaestli, A. T. R. Jensen, D. Müller, T. Theander, and H. P. Beck. Differential expression of *var* gene groups is associated with morbidity caused by *Plasmodium falciparum* infection in Tanzanian children. *Infection and Immunity*, 74:3904–3911, 2006. doi: 10.1128/IAI.02073-05.
- [128] J. Gill and A. Sharma. Structural and genomic analysis of single nucleotide polymorphisms in human host factor endothelial protein c receptor (EPCR) reveals complex interplay with malaria parasites. *Infection, Genetics and Evolution*, 110, 2023. doi: 10.1016/j.meegid.2023.105413.
- [129] A. Sakoguchi and H. Arase. Mechanisms for host immune evasion mediated by *Plasmodium falciparum*-infected erythrocyte surface antigens. *Frontiers in Immunology*, 13, 2022. doi: 10.3389/fimmu.2022.901864.
- [130] G. Band, E. M. Leffler, M. Jallow, F. Sisay-Joof, C. M. Ndila, A. W. Macharia, C. Hubbart, A. E. Jeffreys, K. Rowlands, T. Nguyen, S. Gonçalves, C. V. Ariani, J. Stalker, R. D. Pearson, R. Amato, E. Drury, G. Sirugo, U. d’Alessandro, K. A. Bojang, K. Marsh, N. Peshu, J. W. Saelens, M. Diakité, S. M. Taylor, D. J. Conway, T. N. Williams, K. A. Rockett, and D. P. Kwiatkowski. Malaria protection due to sickle haemoglobin depends on parasite genotype. *Nature*, 602(7895):106–111, 2022. 1476-4687.
- [131] W. L. Hamilton, G. Band, Y. Aniweh, V. Asoala, E. Drury, K. Rockett, G. A. Awandare, D. P. Kwiatkowski, and L. N. Amenga-Etego. A fourth locus in the *Plasmodium falciparum* genome associated with sickle haemoglobin, 2023.
- [132] T. D. Otto, J. C. Rayner, U. Böhme, A. Pain, N. Spottiswoode, M. Sanders, M. Quail, B. Ollomo, F. Renaud, A. W. Thomas, F. Prugnolle, D. J. Conway, C. Newbold, and M. Berriman. Genome sequencing of chimpanzee malaria parasites reveals possible pathways of adaptation to human hosts. *Nature Communications*, 5, 2014. ISSN 2041-1723. doi: 10.1038/ncomms5754.
- [133] V. D. Mangano, Y. Kabore, E. C. Bougouma, F. Verra, N. Sepulveda, C. Bisseye, F. Santolamazza, P. Avellino, A. B. Tiono, A. Diarra, I. Nebie, K. A.

- Rockett, S. B. Sirima, D. Modiano, and MalariaGEN Consortium. Novel insights into the protective role of hemoglobin S and C against *Plasmodium falciparum* parasitemia. *Journal of Infectious Diseases*, 212:626–634, 2015. doi: 10.1093/infdis/jiv098.
- [134] N. Ponts, L. J. Fu, E. Y. Harris, J. Zhang, D. W. D. Chung, M. C. Cervantes, J. Prudhomme, V. Atanasova-Penichon, E. Zehraoui, E. M. Bunnik, E. M. Rodrigues, S. Lonardi, G. R. Hicks, Y. S. Wang, and K. G. Le Roch. Genome-wide mapping of DNA methylation in the human malaria parasite *Plasmodium falciparum*, 2013-12-11 2013.
- [135] S. A. Sundararaman, L. J. Plenderleith, W. M. Liu, D. E. Loy, G. H. Learn, Y. Y. Li, K. S. Shaw, A. Ayoub, M. Peeters, S. Speede, G. M. Shaw, F. D. Bushman, D. Brisson, J. C. Rayner, P. M. Sharp, and B. H. Hahn. Genomes of cryptic chimpanzee *Plasmodium* species reveal key evolutionary events leading to human malaria. *Nature Communications*, 7, 2016. ISSN 2041-1723. doi: 10.1038/ncomms11078.
- [136] A. F. Somé, T. Bazié, I. Zongo, R. S. Yerbanga, F. Nikiéma, C. Neya, L. K. Taho, and J. B. Ouédraogo. *Plasmodium falciparum* MSP1 and MSP2 genetic diversity and allele frequencies in parasites isolated from symptomatic malaria patients in Bobo-Dioulasso, Burkina Faso. *Parasites & Vectors*, 11 (1):323, 2018. doi: 10.1186/s13071-018-2895-4.
- [137] S. Lambie, E. Batty, M. Attar, D. Buck, R. Bowden, G. Lunter, D. Crook, B. El-Fahmawi, and P. Piazza. Improved workflows for high throughput library preparation using the transposome-based Nextera system. *BMC Biotechnology*, 13, 2013. ISSN 1472-6750. doi: 10.1186/1472-6750-13-104.
- [138] R. W. Davies, M. Kucka, D. W. Su, S. N. Shi, M. Flanagan, C. M. Cunliffe, Y. F. Chan, and S. Myers. Rapid genotype imputation from sequence with reference panels. *Nature Genetics*, 53:1104, 2021. doi: 10.1038/s41588-021-00877-0.
- [139] T. G. Mamedov, E. Pienaar, S. E. Whitney, J. R. TerMaat, G. Carvill, R. Goliath, A. Subramanian, and H. J. Viljoen. A fundamental study of the

- PCR amplification of GC-rich DNA templates. *Computational Biology and Chemistry*, 32:452–457, 2008. doi: 10.1016/j.compbiolchem.2008.07.021.
- [140] J. C. Knight, I. Udalova, A. V. S. Hill, B. M. Greenwood, N. Peshu, K. Marsh, and D. Kwiatkowski. A polymorphism that affects oct-1 binding to the *TNF* promoter region is associated with severe malaria. *Nature Genetics*, 22:145–150, 1999.
- [141] W. McGuire, A. V. S. Hill, C. E. M. Allsopp, B. M. Greenwood, and D. Kwiatkowski. Variation in the TNF-alpha promoter region associated with susceptibility to cerebral malaria. *Nature*, 371:508–511, 1994.
- [142] C. Ruwende, S. C. Khoo, A. W. Snow, S. N. R. Yates, D. Kwiatkowski, S. Gupta, P. Warn, C. E. M. Allsopp, S. C. Gilbert, N. Peschu, C. I. Newbold, B. M. Greenwood, K. Marsh, and A. V. S. Hill. Natural selection of hemizygotes and heterozygotes for G6PD deficiency in Africa by resistance to severe malaria. *Nature*, 376:246–249, 1995.
- [143] J. Koster and S. Rahmann. Snakemake - a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522, 2012. ISSN 1367-4803. doi: 10.1093/bioinformatics/bts480.
- [144] Li Heng. Seqtk: a fast and lightweight tool for processing fasta or fastq sequences, 2018. URL <https://github.com/lh3/seqtk>.
- [145] H. B. Xu, X. Luo, J. Qian, X. H. Pang, J. Y. Song, G. R. Qian, J. H. Chen, and S. L. Chen. Fastuniq: A fast *De Novo* duplicates removal tool for paired short reads. *PLoS One*, 7(12), 2012. ISSN 1932-6203. doi: 10.1371/journal.pone.0052249.
- [146] A. M. Bolger, M. Lohse, and B. Usadel. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 2014. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu170.

- [147] P. Ewels, M. Magnusson, S. Lundin, and M. Kaller. Multiqc: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32(19):3047–3048, 2016. ISSN 1367-4803. doi: 10.1093/bioinformatics/btw354.
- [148] H. Li and R. Durbin. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp324.
- [149] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and Proc Genome Project Data. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp352.
- [150] P. P. S. Wang, W. T. Parker, S. Branford, and A. W. Schreiber. BAM-matcher: a tool for rapid NGS sample matching. *Bioinformatics*, 32:2699–2701, 2016. doi: 10.1093/bioinformatics/btw239.
- [151] P. Cingolani, A. Platts, L. L. Wang, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Y. Lu, and D. M. Ruden. A program for annotating and predicting the effects of single nucleotide polymorphisms, snpeff: Snps in the genome of *Drosophila melanogaster*. *Fly*, 6(2):80–92, 2012. ISSN 1933-6934. doi: 10.4161/fly.19695.
- [152] G. A. Van der Auwera, M. O. Carneiro, C. Hartl, R. Poplin, G. Del Angel, A. Levy-Moonshine, T. Jordan, K. Shakir, D. Roazen, J. Thibault, E. Banks, K. V. Garimella, D. Altshuler, S. Gabriel, and M. A. DePristo. From fastq data to high confidence variant calls: the genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics*, 43(1110):11.10.1–11.10.33, 2013. ISSN 1934-3396. doi: 10.1002/0471250953.bi1110s43.
- [153] M. S. Hasan, X. W. Wu, and L. Q. Zhang. Performance evaluation of INDEL calling tools using real short-read data. *Human Genomics*, 9, 2015. ISSN 1473-9542. doi: 10.1186/s40246-015-0042-2.

- [154] A. Tan, G. R. Abecasis, and H. M. Kang. Unified representation of genetic variants. *Bioinformatics*, 31(13):2202–2204, 2015. ISSN 1367-4803. doi: 10.1093/bioinformatics/btv112.
- [155] E. Garrison, Z. N. Kronenberg, E. T. Dawson, B. S. Pedersen, and P. Prins. A spectrum of free software tools for processing the vcf variant call format: vcflib, bio-vcf, cyvcf2, hts-nim and slivar. *PLOS Computational Biology*, 18(5), 2022. ISSN 1553-734X. doi: 10.1371/journal.pcbi.1009123.
- [156] F. B. Piel, A. P. Patil, R. E. Howes, O. A. Nyangiri, P. W. Gething, T. N. Williams, D. J. Weatherall, and S. I. Hay. Global distribution of the sickle cell gene and geographical confirmation of the malaria hypothesis. *Nature Communications*, 1, 2010. doi: 10.1038/ncomms1104.
- [157] M. Manske, O. Miotto, S. Campino, S. Auburn, J. Almagro-Garcia, G. Maslen, J. O’Brien, A. Djimde, O. Doumbo, I. Zongo, J. B. Ouedraogo, P. Michon, I. Mueller, P. Siba, A. Nzila, S. Borrmann, S. M. Kiara, K. Marsh, H. Jiang, X. Z. Su, C. Amaratunga, R. Fairhurst, D. Socheat, F. Nosten, M. Imwong, N. J. White, M. Sanders, E. Anastasi, D. Alcock, E. Drury, S. Oyola, M. A. Quail, D. J. Turner, V. Ruano-Rubio, D. Jyothi, L. Amenga-Etego, C. Hubbart, A. Jeffreys, K. Rowlands, C. Sutherland, C. Roper, V. Mangano, D. Modiano, J. C. Tan, M. T. Ferdig, A. Amambua-Ngwa, D. J. Conway, S. Takala-Harrison, C. V. Plowe, J. C. Rayner, K. A. Rockett, T. G. Clark, C. I. Newbold, M. Berriman, B. MacInnis, and D. P. Kwiatkowski. Analysis of *Plasmodium falciparum* diversity in natural infections by deep sequencing. *Nature*, 487:375–379, 2012. doi: 10.1038/nature11174.
- [158] D. B. Zhong, C. Koepfli, L. W. Cui, and G. Y. Yan. Molecular approaches to determine the multiplicity of *Plasmodium* infections. *Malaria Journal*, 17, 2018. doi: 10.1186/s12936-018-2322-5.
- [159] A. Paschalidis, O. J. Watson, O. Aydemir, R. Verity, and J. A. Bailey. *coiaf*: Directly estimating complexity of infection with allele frequencies. *PLOS Computational Biology*, 19(6), 2023. ISSN 1553-734X. doi: 10.1371/journal.pcbi.1010247.

- [160] G. Coop, D. Witonsky, A. Di Rienzo, and J. K. Pritchard. Using environmental correlations to identify loci underlying local adaptation. *Genetics*, 185(4):1411–1423, 2010. ISSN 0016-6731. doi: 10.1534/genetics.110.114819.
- [161] P. D. Keightley and B. C. Jackson. Inferring the probability of the derived vs. the ancestral allelic state at a polymorphic site. *Genetics*, 209(3):897–906, 2018. ISSN 0016-6731. doi: 10.1534/genetics.118.301120.
- [162] B. F. Voight, S. Kudaravalli, X. Q. Wen, and J. K. Pritchard. A map of recent positive selection in the human genome. *PLOS Biology*, 4:446–458, 2006. doi: 10.1371/journal.pbio.0040072.
- [163] S. C. Nkhoma, S. G. Trevino, K. M. Gorena, S. Nair, S. Khoswe, C. Jett, R. Garcia, B. Daniel, A. Dia, D. J. Terlouw, S. A. Ward, T. J. C. Anderson, and I. H. Cheeseman. Co-transmission of related malaria parasite lineages shapes within-host parasite diversity. *Cell Host & Microbe*, 27(1):93, 2020. ISSN 1931-3128. doi: 10.1016/j.chom.2019.12.001.
- [164] S. J. Zhu, J. Almagro-Garcia, and G. McVean. Deconvolution of multiple infections in *Plasmodium falciparum* from high throughput sequencing data. *Bioinformatics*, 34:9–15, 2018. doi: 10.1093/bioinformatics/btx530.
- [165] M. Gautier and R. Vitalis. rehh: an R package to detect footprints of selection in genome-wide SNP data from haplotype structure. *Bioinformatics*, 28(8):1176–1177, 2012. ISSN 1367-4803. doi: 10.1093/bioinformatics/bts115.
- [166] H. Iriko, O. Kaneko, H. Otsuki, T. Tsuboi, X. Z. Su, K. Tanabe, and M. Torii. Diversity and evolution of the *rhoph1/clag* multigene family of *Plasmodium falciparum*, 2008-03-01 2008.
- [167] MalariaGEN. An open dataset of *Plasmodium falciparum* genome variation in 7,000 worldwide samples. *Wellcome Open Research*, 6, 2021. doi: <https://doi.org/10.12688/wellcomeopenres.16168.2>.
- [168] Z. Shah, M. Adams, K. A. Moser, B. Shrestha, E. M. Stucke, M. K. Laufer, D. Serre, J. C. Silva, and S. Takala-Harrison. Optimization of parasite DNA

enrichment approaches to generate whole genome sequencing data for *Plasmodium falciparum* from low parasitaemia samples. *Malaria Journal*, 19(1), 2020. doi: 10.1186/s12936-020-03195-8.

- [169] A. Piovesan, M. C. Pelleri, F. Antonaros, P. Strippoli, M. Caracausi, and L. Vitale. On the length, weight and GC content of the human genome. *European Journal of Human Genetics*, 27:583–583, 2019. ISSN 1018-4813.
- [170] C. F. Yang, Y. P. Shi, V. Udhayakumar, M. P. Alpers, M. M. Pova, W. A. Hawley, W. E. Collins, and A. A. Lal. Sequence variations in the nonrepetitive regions of the liver stage-specific antigen-1 (LSA-1) of *Plasmodium falciparum* from field isolates. *Molecular and Biochemical Parasitology*, 71(2): 291–294, 1995. ISSN 0166-6851. doi: 10.1038/nature11174.
- [171] M. T. W. Ebbert, M. E. Wadsworth, L. A. Staley, K. L. Hoyt, B. Pickett, J. Miller, J. Duce, J. S. K. Kauwe, P. G. Ridge, and Initi Alzheimer’s Dis Neuroimaging. Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. *BMC Bioinformatics*, 17, 2016. ISSN 1471-2105. doi: 10.1186/s12859-016-1097-3.
- [172] N. Tangpukdee, S. Krudsood, S. Kano, and P. Wilairatana. *falciparum* malaria parasitemia index for predicting severe malaria. *International Journal of Laboratory Hematology*, 34(3):320–327, 2012. ISSN 1751-5521. doi: 10.1111/j.1751-553X.2011.01398.x.
- [173] L. Huang, Y. Li, A. B. Singleton, J. A. Hardy, G. Abecasis, N. A. Rosenberg, and P. Scheet. Genotype-imputation accuracy across worldwide human populations. *American Journal of Human Genetics*, 84(2):235–250, 2009. ISSN 0002-9297. doi: 10.1016/j.ajhg.2009.01.013.
- [174] K. K. Esoh, T. O. Apinjoh, A. Amambua-Ngwa, S. G. Nyanjom, E. R. Chimusa, L. Amenga-Etego, A. Wonkam, and E. A. Achidi. Genome-wide association study identifies novel candidate malaria resistance genes in Cameroon. *Human Molecular Genetics*, 32(12):1946–1958, 2023. ISSN 0964-6906. doi: 10.1093/hmg/ddad026.

- [175] B. C. Young, S. G. Earle, S. Soeng, P. Sar, V. Kumar, S. Hor, V. Sar, R. Bousfield, N. D. Sanderson, L. Barker, N. Stoesser, K. R. W. Emery, C. M. Parry, E. K. Nickerson, P. Turner, R. Bowden, D. W. Crook, D. H. Wyllie, N. P. J. Day, D. J. Wilson, and C. E. Moore. Panton-valentine leucocidin is the key determinant of *Staphylococcus aureus* pyomyositis in a bacterial gwas. *eLife*, 8, 2019. doi: 10.7554/eLife.42486.
- [176] H. H. Chang, E. L. Moss, D. J. Park, D. Ndiaye, S. Mboup, S. K. Volkman, P. C. Sabeti, D. F. Wirth, D. E. Neafsey, and D. L. Hartl. Malaria life cycle intensifies both natural selection and random genetic drift. *Proceedings of the National Academy of Sciences of the United States of America*, 110(50): 20129–20134, 2013. ISSN 0027-8424. doi: 10.1073/pnas.1319857110.
- [177] F. Arieu, B. Witkowski, C. Amaratunga, J. Beghain, A. C. Langlois, N. Khim, S. Kim, V. Duru, C. Bouchier, L. Ma, P. Lim, R. Leang, S. Duong, S. Sreng, S. Suon, C. M. Chuor, D. M. Bout, S. Ménard, W. O. Rogers, B. Genton, T. Fandeur, O. Miotto, P. Ringwald, J. Le Bras, A. Berry, J. C. Barale, R. M. Fairhurst, F. Benoit-Vical, O. Mercereau-Puijalon, and D. Ménard. A molecular marker of artemisinin-resistant *Plasmodium falciparum* malaria. *Nature*, 505(7481):50, 2014. ISSN 0028-0836. doi: 10.1038/nature12876.
- [178] C. Roper, R. Pearce, S. Nair, B. Sharp, F. Nosten, and T. Anderson. Inter-continental spread of pyrimethamine-resistant malaria. *Science*, 305(5687): 1124–1124, 2004. ISSN 0036-8075. doi: 10.1126/science.1098876.
- [179] J. C. Wootton, X. R. Feng, M. T. Ferdig, R. A. Cooper, J. B. Mu, D. I. Baruch, A. J. Magill, and X. Z. Su. Genetic diversity and chloroquine selective sweeps in *Plasmodium falciparum*. *Nature*, 418(6895):320–323, 2002. ISSN 0028-0836. doi: 10.1038/nature00813.
- [180] A Amambua-Ngwa, D Jeffries, R Amato, A Worwui, M Karim, S Ceesay, H Nyang, D Nwakanma, J Okebe, D Kwiatkowski, DJ Conway, and U D’Alessandro. Consistent signatures of selection from genomic analysis

of pairs of temporal and spatial *Plasmodium falciparum* populations from The Gambia, 2018-06-26 2018.

- [181] T. J. Sargeant, M. Marti, E. Caler, J. M. Carlton, K. Simpson, T. P. Speed, and A. F. Cowman. Lineage-specific expansion of proteins exported to erythrocytes in malaria parasites, 2006.
- [182] M. S. Kiwuwa, U. Ribacke, K. Moll, J. Byarugaba, K. Lundblom, A. Färnert, K. Fred, and M. Wahlgren. Genetic diversity of *Plasmodium falciparum* infections in mild and severe malaria of children from Kampala, Uganda. *Parasitology Research*, 112(4):1691–1700, 2013. ISSN 0932-0113. doi: 10.1007/s00436-013-3325-3.
- [183] X. W. Liu, Y. Y. Wang, J. Liang, L. J. Wang, N. Qin, Y. Zhao, and G. Zhao. In-depth comparative analysis of malaria parasite genomes reveals protein-coding genes linked to human disease in *Plasmodium falciparum* genome, 2018-05-02 2018.
- [184] T. D. Otto, S. A. Assefa, U. Böhme, M. J. Sanders, D. Kwiatkowski, M. Beriman, and C. Newbold. Evolutionary analysis of the most polymorphic gene family in *falciparum* malaria. *Wellcome Open Research*, 4:193, 12 2019. doi: 10.12688/wellcomeopenres.15590.1.
- [185] M. Kaestli, I. A. Cockburn, A. Cortes, K. Baea, J. A. Rowe, and H. P. Beck. Virulence of malaria is associated with differential expression of *Plasmodium falciparum* var gene subgroups in a case-control study, 2006-06-01 2006.
- [186] O. Mercereau-Puijalon, J. C. Barale, and E. Bischoff. Three multigene families in *Plasmodium* parasites: facts and questions, 2002-10-01 2002.
- [187] T. Hussain, M. Yogavel, and A. Sharma. Inhibition of protein synthesis and malaria parasite development by drug targeting of methionyl-tRNA synthetases. *Antimicrobial Agents and Chemotherapy*, 59:1856–1867, 2015. doi: 10.1128/AAC.02220-13.

- [188] A. Ashley-Koch, Q. Yang, and R. S. Olney. Sick cell hemoglobin (HbS) allele and sick cell disease: A HuGE review. *American Journal of Epidemiology*, 151(9):839–845, 2000. ISSN 0002-9262.
- [189] M. Aidoo, A. Lalvani, C. E. M. Allsopp, M. Plebanski, S. J. Meisner, P. Krausa, M. Browning, S. Morrisjones, F. Gotch, D. A. Fidock, M. Takiguchi, K. J. H. Robson, B. M. Greenwood, P. Druilhe, H. C. Whittle, and A. V. S. Hill. Identification of conserved antigenic components for a cytotoxic T-lymphocyte-inducing vaccine against malaria. *Lancet*, 345(8956):1003–1007, 1995. ISSN 0140-6736. doi: 10.1016/s0140-6736(95)90754-8.
- [190] J. M. Blackwell, S. E. Jamieson, and D. Burgner. HLA and infectious diseases. *Clinical Microbiology Reviews*, 22(2):370, 2009. ISSN 0893-8512. doi: 10.1128/cmr.00048-08.
- [191] S. C. Gilbert, M. Plebanski, S. J. Harris, C. E. M. Allsopp, R. Thomas, G. T. Layton, and A. V. S. Hill. A protein particle vaccine containing multiple malaria epitopes. *Nature Biotechnology*, 15(12):1280–1284, 1997. ISSN 1087-0156. doi: 10.1038/nbt1197-1280.
- [192] M. Skwarczynski, S. Chandrudu, B. Rigau-Planella, M. T. Islam, Y. S. Cheong, G. Liu, X. Wang, I. Toth, and W. M. Hussein. Progress in the development of subunit vaccines against malaria. *Vaccines (Basel)*, 8(3), 2020. ISSN 2076-393X. doi: 10.3390/vaccines8030373.
- [193] P. Bejon, J. Mwacharo, O. Kai, T. Mwangi, P. Milligan, S. Todryk, S. Keating, T. Lang, B. Lowe, C. Gikonyo, C. Molyneux, G. Fegan, S. C. Gilbert, N. Peshu, K. Marsh, and A. V. S. Hill. A phase 2b randomised trial of the candidate malaria vaccines FP9 ME-TRAP and MVA ME-TRAP among children in Kenya. *PLOS Clinical Trials*, 1(6), 2006. ISSN 1555-5887. doi: 10.1371/journal.pctr.0010029.
- [194] L. B. Jorde. Linkage disequilibrium and the search for complex disease genes. *Genome Research*, 10(10):1435–1444, Oct 2000. doi: 10.1101/gr.144500.

- [195] C. Zapata. On the uses and applications of the most commonly used measures of linkage disequilibrium from the comparative analysis of their statistical properties. *Human Heredity*, 71(3):186–195, 2011. ISSN 0001-5652. doi: 10.1159/000327732.
- [196] S Mira-Martínez, N Rovira-Graells, VM Crowley, LM Altenhofen, M Llinás, and A Cortés. Epigenetic switches in *clag3* genes mediate blasticidin S resistance in malaria parasites, 2013-11-01 2013.
- [197] T. K. Jonsdottir, M. Gabriela, B. S. Crabb, T. F. de Koning-Ward, and P. R. Gilson. Defining the essential exportome of the malaria parasite. *Trends in Parasitology*, 37(7):664–675, 2021. ISSN 1471-4922. doi: 10.1016/j.pt.2021.04.009.
- [198] J. Marchini, L. R. Cardon, M. S. Phillips, and P. Donnelly. The effects of human population structure on large genetic association studies. *Nature Genetics*, 36:512–517, 2004. doi: 10.1038/ng1337.
- [199] E. Uffelmann, Q. Q. Huang, N. S. Munung, J. de Vries, Y. Okada, A. R. Martin, H. C. Martin, T. Lappalainen, and D. Posthuma. Genome-wide association studies. *Nature Reviews Methods Primers*, 1(1):59, 2021. ISSN 2662-8449. doi: 10.1038/s43586-021-00056-9.
- [200] E. Thorsby. A short history of HLA. *Tissue Antigens*, 74(2):101–116, 2009. doi: 10.1111/j.1399-0039.2009.01291.x.
- [201] Z. L. Feng, D. L. Smith, F. E. McKenzie, and S. A. Levin. Coupling ecology and evolution: malaria and the *S*-gene across time scales. *Mathematical Biosciences*, 189:1–19, 2004. doi: 10.1016/j.mbs.2004.01.005.
- [202] G. Q. Tonkin-Hill, L. Trianty, R. Noviyanti, H. H. T. Nguyen, B. F. Sebayang, D. A. Lampah, J. Marfurt, S. A. Cobbold, J. S. Rambhatla, M. J. McConville, S. J. Rogerson, G. V. Brown, K. P. Day, R. N. Price, N. M. Anstey, A. T. Papenfuss, and M. F. Duffy. The *Plasmodium falciparum* transcriptome in severe malaria reveals altered expression of genes involved

- in important processes including surface antigen-encoding *var* genes. *Plos Biology*, 16, 03 2018. doi: 10.1371/journal.pbio.2004328.
- [203] A. Gupta, G. Thiruvengadam, and S. A. Desai. The conserved *clag* multigene family of malaria parasites: Essential roles in host-pathogen interaction. *Drug Resistance Updates*, 18:47–54, 2015. doi: 10.1016/j.drug.2014.10.004.
- [204] P. Sharma, K. Wollenberg, M. Sellers, K. Zainabadi, K. Galinsky, E. Moss, W. Nguitragool, D. Neafsey, and S. A. Desai. An epigenetic antimalarial resistance mechanism involving parasite genes linked to nutrient uptake. *Journal of Biological Chemistry*, 288:19429–19440, 2013. doi: 10.1074/jbc.M113.468371.
- [205] H. Davies, H. Belda, M. Broncel, X. D. Ye, C. Bisson, V. Introini, D. Dorin-Semlat, J. P. Semlat, M. Tibúrcio, B. Gamain, M. Kaforou, and M. Treeck. An exported kinase family mediates species-specific erythrocyte remodelling and virulence in human malaria. *Nature Microbiology*, 5:848, 2020. doi: 10.1038/s41564-020-0702-4.
- [206] P. Ward, L. Equinet, J. Packer, and C. Doerig. Protein kinases of the human malaria parasite *Plasmodium falciparum*: the kinome of a divergent eukaryote. *BMC Genomics*, 5:79, 10 2004. doi: 10.1186/1471-2164-5-79.
- [207] M. C. Nunes, J. P. Goldring, C. Doerig, and A. Scherf. A novel protein kinase family in *Plasmodium falciparum* is differentially transcribed and secreted to various cellular compartments of the host cell. *Molecular Microbiology*, 63(2):391–403, 1 2007. doi: 10.1111/j.1365-2958.2006.05521.x.
- [208] A. Agarwal, A. Guindo, Y. Cissoko, J. G. Taylor, D. Coulibaly, A. Koné, K. Kayentao, A. Djimde, C. V. Plowe, O. Doumbo, T. E. Wellems, and D. Diallo. Hemoglobin C associated with protection from severe malaria in the Dogon of Mali, a West African population with a low prevalence of hemoglobin S. *Blood*, 96:2358–2363, 2000.
- [209] B. M. Peter, E. Huerta-Sanchez, and R. Nielsen. Distinguishing between selective sweeps from standing variation and from a *de novo* mutation. *PLOS Genetics*, 8, 2012. doi: 10.1371/journal.pgen.1003011.

- [210] M. Mahmoud, N. Gobet, D. I. Cruz-Davalos, N. Mounier, C. Dessimoz, and F. J. Sedlazeck. Structural variant calling: the long and the short of it. *Genome Biology*, 20(1), 2019. ISSN 1474-760X. doi: 10.1186/s13059-019-1828-7.
- [211] J. T. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander, G. Getz, and J. P. Mesirov. Integrative genomics viewer. *Nature Biotechnology*, 29(1):24–26, 2011. ISSN 0733-222X. doi: 10.1038/nbt.1754.
- [212] M. Michel, E. Skourtanioti, F. Pierini, E. K. Guevara, A. Mötsch, A. Kocher, R. Barquera, R. A. Bianco, S. Carlhoff, L. C. Bove, S. Freilich, K. Giffin, T. Hermes, A. Hiss, F. Knolle, E. A. Nelson, G. U. Neumann, L. Pappac, S. Penske, A. B. Rohrlach, N. Salem, L. Semerau, V. Villalba-Mouco, I. Abadie, M. Aldenderfer, J. F. Beckett, M. Brown, F. G. R. Campus, T. Chenghwa, M. C. Berrocal, L. Damasek, K. S. D. Carlson, R. Durand, M. Ernée, C. Fântaneanu, H. Frenzel, G. G. Atiénzar, S. Guillén, E. Hsieh, M. Karwowski, D. Kelvin, N. Kelvin, A. Khokhlov, R. L. Kinstanton, A. Korolev, K. L. Krettek, M. Küssner, L. Lai, C. Look, K. Majander, K. Mandl, V. Mazzarello, M. McCormick, P. D. Ibáñez, R. Murphy, R. E. Németh, K. Nordqvist, F. Novotny, M. Obenaus, L. Olmo-Enciso, P. Onkamo, J. Orschiedt, V. Patrushev, S. Peltola, A. Romero, S. Rubino, A. Sajantila, D. C. Salazar-García, E. Serrano, S. Shaydullaev, E. Sias, M. Slaus, L. Stanco, T. Swanston, M. Teschler-Nicola, F. Valentin, K. Van de Vijver, T. L. Varney, A. V. E. Guirado, C. K. Waters, E. Weiss-Krejci, E. Winter, T. C. Lamnidis, K. Prüfer, K. Nägele, M. Spyrou, S. Schiffels, P. W. Stockhammer, W. Haak, C. Posth, C. Warinner, K. I. Bos, A. Herbig, and J. Krause. Ancient *Plasmodium* genomes shed light on the history of human malaria. *Nature*, 631:125, 2024. doi: 10.1038/s41586-024-07546-2.
- [213] E. R. Lucas, S. C. Nagi, A. Egyir-Yawson, J. Essandoh, S. Dadzie, J. Chabi, L. S. Djogbenou, A. A. Medjigbodo, C. V. Edi, G. K. Ketoh, B. G. Koudou, A. E. Van’t Hof, E. J. Rippon, D. Pipini, N. J. Harding, N. A. Dyer, L. T. Cerdeira, C. S. Clarkson, D. P. Kwiatkowski, A. Miles, M. J. Donnelly, and D. Weetman. Genome-wide association studies reveal novel loci associated

with pyrethroid and organophosphate resistance in *Anopheles gambiae* and *Anopheles coluzzii*. *Nature Communications*, 14, 2023. doi: 10.1038/s41467-023-40693-0.

- [214] E. M. Leffler, G. Band, G. B. J. Busby, K. Kivinen, Q. S. Le, G. M. Clarke, K. A. Bojang, D. J. Conway, M. Jallow, F. Sisay-Joof, E. C. Bougouma, V. D. Mangano, D. Modiano, S. B. Sirima, E. Achidi, T. O. Apinjoh, K. Marsh, C. M. Ndila, N. Peshu, T. N. Williams, C. Drakeley, A. Manjurano, H. Reyburn, E. Riley, D. Kachala, M. Molyneux, V. Nyirongo, T. Taylor, N. Thornton, L. Tilley, S. Grimsley, E. Drury, J. Stalker, V. Cornelius, C. Hubbard, A. E. Jeffreys, K. Rowlands, K. A. Rockett, C. C. A. Spencer, D. P. Kwiatkowski, and Netwo Malaria Genomic Epidemiology. Resistance to malaria through structural variation of red blood cell invasion receptors. *Science*, 356(6343), 2017. ISSN 0036-8075. doi: 10.1126/science.aam6393.
- [215] N. Lu, Y. Qiao, Z. H. Lu, and J. Tu. Chimera: The spoiler in multiple displacement amplification. *Computational and Structural Biotechnology Journal*, 21:1688–1696, 2023. doi: 10.1016/j.csbj.2023.02.034.
- [216] Q. M. Geiman and M. J. Meagher. Susceptibility of a new world monkey to *Plasmodium falciparum* from man. *Nature*, 215:437, 1967.
- [217] J. W. Saelens, J. E. V. Petersen, E. Freedman, R. C. Moseley, D. Konaté, S. A. S. Diakité, K. Traoré, N. Vance, R. M. Fairhurst, M. Diakité, S. B. Haase, and S. M. Taylor. Impact of sickle cell trait hemoglobin on the intraerythrocytic transcriptional program of *Plasmodium falciparum*. *mSphere*, 6(5), 2021. doi: 10.1128/mSphere.00755-21.
- [218] M. T. Watzlowik, S. Das, M. Meissner, and G. Langst. Peculiarities of *Plasmodium falciparum* gene regulation and chromatin structure. *International Journal of Molecular Sciences*, 22(10), 2021. doi: 10.3390/ijms22105168.
- [219] D. K. Nam, S. Lee, G. L. Zhou, X. H. Cao, C. Wang, T. Clark, J. J. Chen, J. D. Rowley, and S. M. Wang. Oligo(dT) primer generates a high frequency of truncated cDNAs through internal poly(A) priming during reverse transcription. *Proceedings of the National Academy of Sciences of the*

United States of America, 99(9):6152–6156, 2002. ISSN 0027-8424. doi:
10.1073/pnas.092140899.