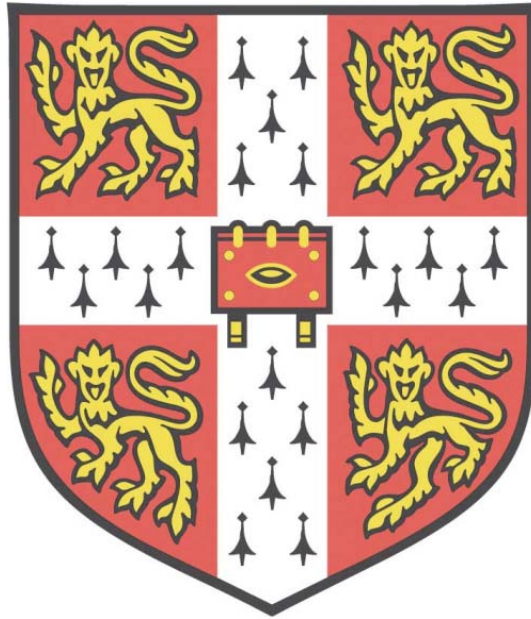


# The Influence of Genetics on Gamma-Herpesvirus Infections



Neneh Sallah

University of Cambridge  
Wellcome Trust Sanger Institute

This dissertation is submitted for the degree of Doctor of Philosophy

September 2016

Murray Edwards College



## Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the contributions section within each chapter and/or specified in the text. It is not being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University of similar institution. It does not exceed the prescribed word limit for the Faculty of Biology.

Neneh Sallah

September 2016

## Abstract

Gamma-herpesviruses are double stranded DNA lymphotropic viruses that include Epstein-Barr Virus (EBV) and Kaposi's sarcoma herpesvirus (KSHV). They establish life-long infections in the human host and have been associated with a variety of malignant tumours. Upon exposure to an infectious pathogen, both host and pathogen genetic differences influence variation in an individual's immune response including potential disease outcome. Although both viruses have been studied extensively, genetic and environmental influences on susceptibility to infection in individuals and potential disease outcome as a result remain unclear. While EBV is nearly ubiquitous globally, KSHV displays striking geographic variation with highest prevalence in sub-Saharan Africa, particularly in Uganda. Thus, this thesis investigates how host and virus genetics influence pathogenesis in EBV and KSHV infections, particularly the contribution of human genetic variation, using the Ugandan General Population Cohort (GPC) as a study population. The GPC provides a phenotype rich dataset and the availability of human genomic data in a large subset of individuals provides the opportunity to investigate the genetics of infection.

In chapter 2, I characterised the seroprevalence of oncogenic viral infections, assessed the influence of co-infection on EBV and KSHV serological traits in the GPC, and assessed the genetic population structure and heritability of Immunoglobulin G (IgG) antibody response traits.

In chapters 3 and 4, I explored the influence of host genetic variation on EBV and KSHV IgG antibody levels respectively, as a proxy for infection and potential disease risk. I performed the first genome-wide association analysis of anti-EBV IgG traits and anti-KSHV IgG traits in Africa, using a combined approach including array genotyping, whole-genome sequencing and imputation to a panel with African sequence data to extensively capture genetic variation and aid locus discovery. For EBV infection, I identified novel loci and through trans-ethnic meta-analysis with a cohort of European ancestry I uncovered

distinct variants contributing to variation in immune responses in Uganda. For KSHV infection multiple putative candidate loci were identified with modest effect sizes potentially contributing to infection.

As Uganda sustains such high levels of KSHV seroprevalence compared to the rest of the world, in chapter 5, I also explored the viral genetic diversity of KSHV in the GPC by whole genome sequencing of viral DNA isolated from saliva of asymptomatic individuals, and analysed the population structure comparing it to published KSHV genomes from around the world. This analysis showed a greater appreciation of variation of genes in the central region of the genome, some of which are under positive selection, contributing to the clustering of genomes by geography, thus, suggesting the use of whole-genomes in KSHV viral characterisation.

Together, the findings described in this thesis reinforce the importance of conducting genetic studies of infectious disease in African populations to uncover functionally relevant loci associated with traits of interest, and independent of environmental factors. Furthermore, the ability to obtain both host and viral genomes from the same individuals, allows for a comprehensive assessment of factors underlying the course of infection. The development of African resources to fully capture genetic diversity across the continent and building of research capacity will be fundamental to facilitate large scale studies and uncover meaningful biological insights.

## Acknowledgements

I would like to thank my supervisors Dr. Inês Barroso and Prof. Paul Kellam immensely for their unwavering support and continued guidance throughout my PhD. My research was possible as result of extensive collaboration and thus, I would like to acknowledge: Dr. Rob Newton and his team at the MRC/UVRI in Uganda for setting up and granting me access to the GPC, Dr. Manj Sandhu and his team at the Sanger for curating the genetic data and Dr. Denise Whitby and her team at the FNLRCR in the US for providing the serological data for analyses. I would also like to thank the GPC study participants for their generosity. I thank the members of my thesis committee, Dr. Stacey Efstathiou and Dr. Arthur Kaser for their useful discussions.

Many thanks to those who've shared their expertise with me: Dr. Carl Anderson and Dr. Chris Franklin for introducing me to the analysis of human genetic data, UNIX and R; Dr. Eleanor Wheeler for all the useful Stats guidance; Dr. Anne Palser and Dr. Simon Watson for their guidance on virus genomic analysis. Thanks to all the members of teams 146 and 35, in particular, Carol Dunbar for all her admin support and always finding a way, and Fernando for all the good humour in N-333. Thanks to the Wellcome Trust Sanger Institute for generously funding my studies and the graduate programme for their support.

On a more personal note, I thank my many friends at Sanger, especially, Pinky, Carmen, Mia, Eva, Sophia, Michal, Tomi, Martin and the rest of PhD12, for sharing the peaks and troughs of research and writing-up, the motivational and therapeutic conversations and making the past four years an enjoyable experience! Thanks to the members of the "Murrays Breakfast Table" for the morning banter, ensuring a good start to every day!

Finally, I would like to thank my family, especially my parents, Zahra and O.G for their encouragement, constantly believing in me and being my inspiration. Thanking Alima for her invaluable company in Cambridge and keeping my life exciting! Huge thanks to Gibril,

Jamil and Alpha for being a source of comfort and reminding me that there's more to life than 'work'! Thanks to Louis, for his patience and embarking on this journey with me. Thanks to Uncle Halim for delivering me safely to all my destinations and reminding me to relax. Thanks to all my 'Kusineras' for keeping my life interesting and my second mothers, Sheriffa and Granny for their encouragement! Thanks to my MRC Gambia family, in particular, Pa Tamba Ngom, Bouke de Jong, Harr Njai (R.I.P.) and Martin Antonio for taking me under their wings as a young, aspiring scientist.

This PhD is dedicated to my mother and all the great, persevering women in my family who've been my motivation and supported me endlessly! I thank God for keeping me determined and helping me overcome my challenges.

## Table of Contents

<b>Declaration</b> .....	<b>i</b>
<b>Abstract</b> .....	<b>ii</b>
<b>Acknowledgements</b> .....	<b>iv</b>
<b>List of Tables</b> .....	<b>ix</b>
<b>List of Figures</b> .....	<b>ix</b>
<b>List of Abbreviations</b> .....	<b>xii</b>
<b>1 Introduction</b> .....	<b>1</b>
<b>1.1 Gamma-herpesviruses</b> .....	<b>1</b>
<b>1.2 Epstein-Barr Virus (EBV)</b> .....	<b>6</b>
1.2.1 Epidemiology of EBV Infection & Associated Diseases .....	6
1.2.2 The Biology of Infection.....	9
1.2.3 Host Immune Responses to Infection.....	13
<b>1.3 Kaposi's Sarcoma-Associated Herpesvirus (KSHV)</b> .....	<b>17</b>
1.3.1 Epidemiology of KSHV Infection & Associated Diseases.....	19
1.3.2 The Biology of Infection .....	23
1.3.3 Host Immune Response to Infection .....	26
<b>1.4 The Influence of Host Genetics on Infectious Diseases</b> .....	<b>28</b>
1.4.1 Genome-Wide Association as a Tool to Study Infectious Diseases .....	29
1.4.2 Genome-Wide Association Studies in African Populations .....	33
<b>1.5 Thesis Aims</b> .....	<b>36</b>
<b>2 Chapter 2: The General Population Cohort, a Platform to Study the Genetic Architecture of Host Response to Gamma-Herpesvirus Infections</b> .....	<b>37</b>
<b>2.1 Introduction</b> .....	<b>37</b>
2.1.1 Chapter Aims .....	43
<b>2.2 Methods</b> .....	<b>44</b>
2.2.1 Sample Collection .....	44
2.2.2 Ethics .....	44
2.2.3 Serology and Quality Control of Phenotypic Data .....	44
2.2.4 Statistical Analysis of Quantitative Antibody Traits.....	47
2.2.5 SNP Genotyping and Quality Control.....	47
2.2.6 Principal Components Analysis.....	48
2.2.7 Heritability of Antibody Response Traits in The GPC.....	51
<b>2.3 Results</b> .....	<b>52</b>
2.3.1 Seroprevalence of Infectious Traits in The GPC.....	52
2.3.2 Inter-Individual Variation in IgG Antibody Responses to EBV and KSHV Infections .	54
2.3.3 Predictors of IgG response levels to EBV and KSHV infection .....	58
2.3.4 Genetic Population Structure in The GPC.....	60
2.3.5 Heritability of IgG Antibody Response Traits in The GPC .....	64
<b>2.4 Discussion</b> .....	<b>67</b>
<b>3 Chapter 3: The Influence of Host Genetics on Epstein-Barr Virus Infection</b> .....	<b>71</b>



<b>3.1</b>	<b>Introduction</b> .....	<b>71</b>
3.1.1	Chapter Aims .....	79
<b>3.2</b>	<b>Methods</b> .....	<b>80</b>
3.2.1	Sample Selection .....	80
3.2.2	Whole-Genome Sequencing and Quality Control.....	80
3.2.3	Imputation .....	80
3.2.4	Association Analyses.....	81
3.2.5	Trans-Ethnic Meta-Analysis .....	83
3.2.6	Fine Mapping.....	83
3.2.7	Functional Annotation of Candidate Variants .....	83
<b>3.3</b>	<b>Results</b> .....	<b>85</b>
3.3.1	Discovery of Novel African-Specific Anti-VCA IgG Loci .....	87
3.3.2	Replicating a Known Anti-EBNA-1 IgG Response Locus.....	91
3.3.3	Multivariate Quantitative Association Boosts HLA Signal .....	95
3.3.4	Distinct Association Signals in the HLA Class II Region for Anti-EBNA-1 IgG Response 97	
<b>3.4</b>	<b>Discussion</b> .....	<b>101</b>
<b>4</b>	<b>Chapter 4: The Influence of Host Genetics on Kaposi's Sarcoma-Associated Herpesvirus Infection</b> .....	<b>105</b>
<b>4.1</b>	<b>Introduction</b> .....	<b>105</b>
4.1.1	Chapter Aims .....	115
<b>4.2</b>	<b>Methods</b> .....	<b>116</b>
4.2.1	Sample Selection and Quality Control .....	116
4.2.2	Imputation .....	118
4.2.3	Association Analyses.....	118
4.2.4	Functional Annotation of Candidate Variants .....	119
<b>4.3</b>	<b>Results</b> .....	<b>121</b>
4.3.1	Discovery of Candidate Loci Associated with Latent KSHV Infection .....	123
4.3.2	Discovery of Candidate Loci Associated with Increased Lytic Antigen Levels .....	128
4.3.3	Multivariate Association Analyses of IgG response to KSHV infection .....	133
4.3.4	Associations with Previously Identified Candidate Variants in This Study .....	137
<b>4.4</b>	<b>Discussion</b> .....	<b>138</b>
<b>5</b>	<b>Chapter 5: Characterizing the Genetic Diversity of KSHV in The Uganda GPC.....</b>	<b>145</b>
<b>5.1</b>	<b>Introduction</b> .....	<b>145</b>
5.1.1	Chapter Aims .....	148
<b>5.2</b>	<b>Methods</b> .....	<b>149</b>
5.2.1	Sample Selection and Collection .....	149
5.2.2	DNA Extraction, Purification and Quantification .....	149
5.2.3	Quantitative PCR for Viral DNA Detection .....	150
5.2.4	KSHV Whole-Genome Sequencing .....	151
5.2.5	Guided Assembly of KSHV Whole-Genomes .....	151
5.2.6	Comparative and Phylogenetic Sequence Analysis .....	152
<b>5.3</b>	<b>Results</b> .....	<b>154</b>
5.3.1	KSHV Shedding and Viral Load in the GPC.....	154
5.3.2	KSHV Viral Load Correlates with Whole-Genome Sequencing Quality.....	156
5.3.3	KSHV Genome Variability .....	159
5.3.4	Virus Population Structure and Geographic Variability.....	165

5.3.5	Genotypic Diversity of Strains in the GPC.....	170
5.4	<b>Discussion .....</b>	<b>175</b>
6.	<b>Conclusions and Future Outlook .....</b>	<b>180</b>
6.1	Inferring the Causality of Variants .....	181
6.2	The Contribution of Low-Frequency and Rare Variants to Infectious Disease .....	182
6.3	Genome-to-Genome Analysis.....	183
	<b>References .....</b>	<b>184</b>

## List of Tables

Table 1.1 The Human Herpesviruses.....	2
Table 1.2 EBV latency programs associated with infection.....	12
Table 2.1 Characteristics of individuals in the GPC .....	46
Table 2.2. Distribution of Ethno-linguistic Groups Genotyped* in the GPC.....	48
Table 2.3. Distribution of samples included in Principal Components Analysis .....	50
Table 2.4 Seroprevalence of co-infection with EBV or KSHV .....	53
Table 2.5 Covariates/Predictors of IgG response levels for EBV and KSHV infection.....	59
Table 2.6 Heritability estimates of EBV and KSHV IgG antibody traits in the GPC .....	66
Table 3.1 Putative candidate loci associated with EBV and associated diseases identified by candidate gene approaches.....	75
Table 3.2 Summary of Genome-wide Significant Association Results in The GPC .....	96
Table 3.3 Loci with strong evidence of association with anti-EBNA-1 IgG levels after trans- ethnic meta-analysis of Ugandan and European ancestry GWAS .....	99
Table 3.4 Conditional analysis of lead Ugandan and European SNPs .....	100
Table 4.1 Putative Candidate Loci Associated with KSHV infection and Diseases .....	111
Table 4.2 Characteristics of individuals in the GPC used in this study .....	117
Table 4.3 Summary of significant linear regression coefficients .....	122
Table 4.4 Summary of lead anti-LANA IgG response level association results ( $p < 1 \times 10^{-6}$ ) .....	125
Table 4.5 Summary of lead anti-K8.1 IgG response level association results ( $p < 1 \times 10^{-6}$ )	130
Table 4.6 Summary of lead anti-KSHV IgG response level multivariate association results ( $p < 1 \times 10^{-6}$ ) .....	135
Table 4.7 Associations with previously identified candidate variants.....	137
Table 5.1 qPCR Primer and probe sequences .....	151
Table 5.2 Summary of KSHV samples used in this study .....	159
Table 5.3 Eighty-four annotated KSHV genes based on the GK18 sequence .....	160
Table 5.4 Characteristics of Ugandan GPC Samples .....	173

## List of Figures

Fig. 1.1 Schematic alignment of EBV and KSHV genomes. ....	2
Fig. 1.2 General Life Cycle of EBV and KSHV. ....	4
Fig. 1.3 Summary of the range of diseases associated with EBV and/or KSHV infections..	5
Fig. 1.4 EBV antibody dynamics in the immunocompetent host following primary infection.. ....	15
Fig. 1.5 The KSHV Episome.. ....	18
Fig. 1.6 KSHV gene expression dynamics following primary infection. ....	24
Fig. 2.1. Maps showing The GPC study area in context of Uganda and Africa.. ....	42
Fig. 2.2 Map of Africa showing the location of samples in the AGVP used for PCA.. ....	49

Fig. 2.3 Seroprevalence of viral infections tested in the GPC between 2008-2011 .....	52
Fig. 2.4 The number of seropositive reactions to viruses for all participants in The GPC between 2008-2011. ....	53
Fig. 2.5 Inter-individual variability in IgG antibody responses to EBV.....	55
Fig. 2.6 Distribution of anti-EAD IgG Mean Fluorescence Intensity (MFI).....	56
Fig. 2.7 Inter-individual variability in IgG antibody responses to KSHV.....	57
Fig. 2.8 Genetic population structure of individuals within the GPC ethnolinguistic groups. .....	61
Fig. 2.9 Genetic population structure of the GPC in the context of AGVP African populations. ....	62
Fig. 2.10 Genetic population structure of GPC in the context of AGVP and global 1000G populations.. ....	63
Fig. 2.11 Heritability of IgG antibody traits for EBV and KSHV infections. ....	65
Fig. 3.1 Genome-wide association workflow for EBV serological traits in the Uganda GPC .....	84
Fig. 3.2 Statistical power (%) to identify genetic variants at $p < 5 \times 10^{-9}$ , given different allele frequencies (%) and effect sizes ( $\beta$ ) (N=1567).....	86
Fig. 3.3. Genome-wide association results of anti-VCA IgG response.....	88
Fig. 3.4 Regional association plots for VCA serostatus genome-wide (GW) significant associations, N=1567, Pos=1350, Neg=217, threshold= $p < 5 \times 10^{-9}$ . ....	89
Fig. 3.5 Comparison of allele frequencies of lead VCA GWAS SNPs between 1000 Genomes phase 3 populations and the GPC. ....	90
Fig. 3.6 Genome-wide association results of anti-EBNA-1 IgG response. ....	92
Fig. 3.7 Regional association plot for anti-EBNA-1 IgG response levels in 1473 individuals. .....	93
Fig. 3.8 Comparison of allele frequencies of lead EBNA-1 GWAS SNP, rs9272371 in HLA- DQA1 on chromosome 6 - between 1000 Genomes phase 3 populations and the GPC.....	93
Fig. 3.9 The effect of rs9272371 genotypes on HLA-DQA1 gene expression, cis-eQTL data from the GTEx database.....	94
Fig. 3.10 Multivariate genome-wide association results of anti-EBV IgG response levels.....	95
Fig. 3.11 Trans-ethnic meta-analysis association for EBNA-1 IgG response levels in 3635 individuals of Ugandan and European ancestry (EUR) (threshold= $\log_{10}BF > 6$ ). ....	98
Fig. 4.1 Genome-wide association workflow for KSHV serological traits in the Uganda GPC .....	120
Fig. 4.2 Statistical power to identify genetic variants at $p < 5 \times 10^{-9}$ , given different allele frequencies (%) and different effect sizes ( $\beta$ ) (N=4466).....	122
Fig. 4.3 Genome-wide association results of anti-LANA IgG response levels.....	124
Fig. 4.4 Regional association plots for SNPs associated with anti-LANA IgG ( $p < 1 \times 10^{-6}$ , N=4466.).....	126
Fig. 4.5 Regional association plots for SNPs associated with anti-LANA IgG levels ( $p < 1 \times 10^{-6}$ , N=4466.) (continued). ....	127
Fig. 4.6 Genome-wide association results of anti-K8.1 IgG response levels. ....	129

Fig. 4.7 Regional association plots for SNPs associated with anti-K8.1 IgG response levels, N=4466, threshold= $p < 1 \times 10^{-6}$ .....	131
Fig. 4.8 Regional association plots for SNP on chromosome 3 associated with anti-K8.1 IgG response levels, N=4466, threshold= $p < 1 \times 10^{-6}$ .....	132
Fig. 4.9 Multivariate Genome-wide Association results of anti-KSHV IgG response levels. ....	134
Fig. 4.10 Regional association plots for multivariate anti-KSHV IgG levels, N=4466, threshold= $p < 1 \times 10^{-6}$ .....	136
Fig. 5.1 KSHV genome analysis workflow. ....	153
Fig. 5.2 KSHV ORF73 gene qPCR for BCBL-1 DNA dilution series from 30 to $3 \times 10^6$ viral copies/ml.....	155
Fig. 5.3 Correlation matrix of Viral load (copies/ml), KSHV mapped reads (%) and mean sequencing depth of 200x. ....	157
Fig. 5.4 Map showing GPC the study area in Uganda.....	158
Fig. 5.5 Genome variability of 83 KSHV genomes. ....	162
Fig. 5.6 SNP variation across coding region.....	163
Fig. 5.7 Non-synonymous to synonymous change (dN/dS) analysis across KSHV coding region. ....	164
Fig. 5.8 KSHV whole-genome phylogenetic analysis of 83 samples.. ....	166
Fig. 5.9 KSHV whole-genome phylogeographic analysis of 83 samples. ....	167
Fig. 5.10 KSHV genome phylogenetic analysis of central region minus K1 and K15 genes in 83 samples.....	169
Fig. 5.11 KSHV K15 gene phylogenetic analysis of 83 samples.. ....	171
Fig. 5.12 KSHV K1 gene phylogeographic analysis of 83 samples.. ....	172

## List of Abbreviations

<b>Abbreviation</b>	<b>Full Name</b>
1000G	1000 Genomes
AGVP	Africa Genome Variation Project
BL	Burkitt's lymphoma
CAEBV	Chronic active EBV
EAD	Early antigen D
EAF	Effect allele frequency
EBNA	EBV nuclear antigen
EBV	Epstein-Barr Virus
GPC	General Population Cohort
GWAS	Genome-wide association study
HAART	Highly active anti-retroviral therapy
HBV	Hepatitis B Virus
HCV	Hepatitis C Virus
HHV	Human Herpesvirus
HIV	Human Immunodeficiency Virus
HL	Hodgkin's lymphoma
HLA	Human leukocyte antigen
IBD	Identity-by-Descent
IFN	Interferon
Ig	Immunoglobulin
IL	Interleukin
IM	Infectious mononucleosis
KICS	Kaposi's Sarcoma inflammatory cytokine syndrome
KS	Kaposi's Sarcoma
KSHV	Kaposi's sarcoma-associated herpesvirus
LANA	Latency-associated nuclear antigen
LCL	Lymphoblastoid cell line
LD	Linkage disequilibrium
LMM	Linear mixed model
LMP	Latency-associated membrane protein
MAF	Minor allele frequency
MCD	Multicentric castelman's disease
MFI	Mean fluorescence intensity
MHC	Major histocompatibility complex
NPC	Nasopharyngeal Carcinoma
OD	Optical density

OR	Odds ratio
ORF	Open reading frame
PBMCs	Peripheral blood mononuclear cells
PCA	Principal components analysis
PCR	Polymerase chain reaction
PEL	Primary effusion lymphoma
SNP	Single nucleotide polymorphism
UG2G	Uganda 2000 genomes
UGWAS	Uganda GWAS
VCA	Viral capsid antigen
WGS	Whole-genome sequencing

