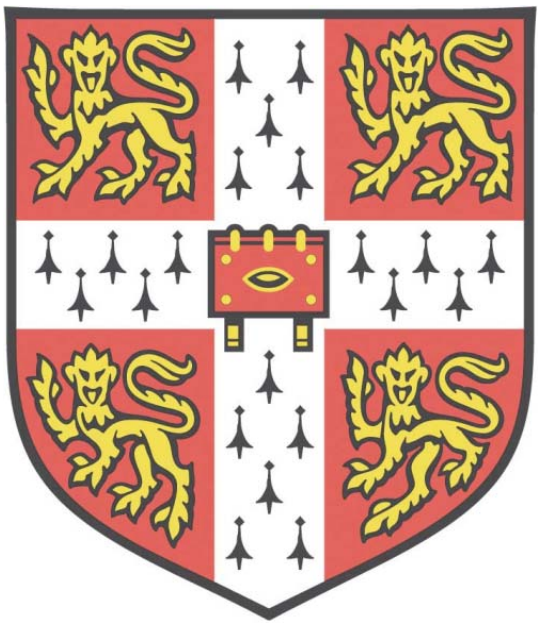
The Influence of Genetics on Gamma-Herpesvirus Infections



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 FNLCR 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

List of Tables

List of Figures

List of Abbreviations

Abbreviation	Full Name
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 castleman'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

1 Introduction

1.1 Gamma-herpesviruses

Herpesviruses are large double stranded (ds) DNA viruses that are amongst the most successful and ubiquitous pathogens in nature. Based on genome sequence, organization and biological functions, herpesviruses are divided into three main families, the alpha (α)-, beta (β)– and gamma (γ)- herpesvirinae. Even though more than 130 herpesviruses have been reported in the animal kingdom only eight have been reported to infect humans to date (Table 1.1). Herpesviruses are thought to share common evolutionary origin, as supported by high amino acid sequence similarity between their viral gene products and have co-evolved with, and adapted to, their hosts during speciation[1]. Unlike other herpesviruses, γ-herpesviruses have a tropism for B-lymphocytes (i.e. lymphotropic), they have oncogenic potential and are associated with a number of malignancies in a subset of infected individuals[2]. They are subdivided into two families: lymphocryptovirus (γ1) and rhadinovirus (γ2). The two human γ-herpesviruses are, Epstein-Barr (EBV) also known as human herpesvirus 4 (HHV-4) which is a lymphocryptovirus, and Kaposi's Sarcoma herpesvirus (KSHV) also known as Human herpesvirus-8 (HHV-8) and which belongs to the rhadinovirus sub-family[2]. The rhadinovirus sub-family also includes herpesvirus samiri (HVS) that infects squirrel monkeys and can also induce neoplasm. The γ-herpesviruses are more closely related to each other than to any other members of the herpesvirus family and also have relatively similar genome architecture (Fig. 1.1)[1].

Table 1.1 The Human Herpesviruses

Family	Common Name	Taxon Name
Alphaherpesvirinae	Herpes simplex virus 1	*Human herpesvirus 1*
	Herpes simplex virus 2	*Human herpesvirus 2*
	Varicella-zoster virus	*Human herpesvirus 3*
Betaherpesvirinae	Human cytomegalovirus	*Human herpesvirus 5*
	HHV-6	*Human herpesvirus 6*
	HHV-7	*Human herpesvirus 7*
Gammaherpesvirinae	Epstein-Barr virus	*Human herpesvirus 4*
	Kaposi's Sarcoma herpesvirus	*Human herpesvirus 8*

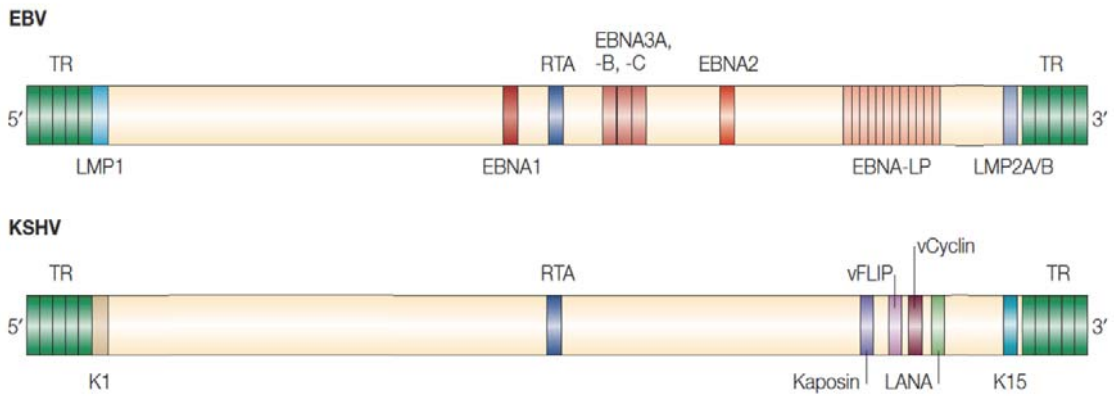


Fig. 1.1 Schematic alignment of EBV and KSHV genomes. The alignment is shown with respect to genes important in modulating important biological process in different stages of the infection life cycle, discussed in detail later. Adapted from Damania, 2004[1]

A signature of herpesviruses is their ability to establish a chronic, persistent infection in immunocompetent hosts. EBV and KSHV utilise a biphasic life cycle with a latent stage where the viral genome exists as a closed circular episome and is characterised by a restricted pattern of gene expression in infected B-cells; and a lytic stage with intermittent periods of replication to facilitate their spread of infectious progeny into other cells and transmission to other hosts (Fig. 1.2). Both viruses have evolved similar strategies to strike a balance with the host immune response, this includes the manipulation of key signalling pathways and molecular piracy/mimicry of important host proteins to promote lifelong survival, immune evasion and tumorigenesis. As EBV and KSHV are restricted to humans, insights into their pathogenesis have been gained by using Murine γherpesvirus-68 (MHV-68) which was isolated from a bank vole as an experimental model in mice[3]. As will be discussed in detail later, both viruses have been linked to a number of cancers and other diseases (Fig. 1.3).

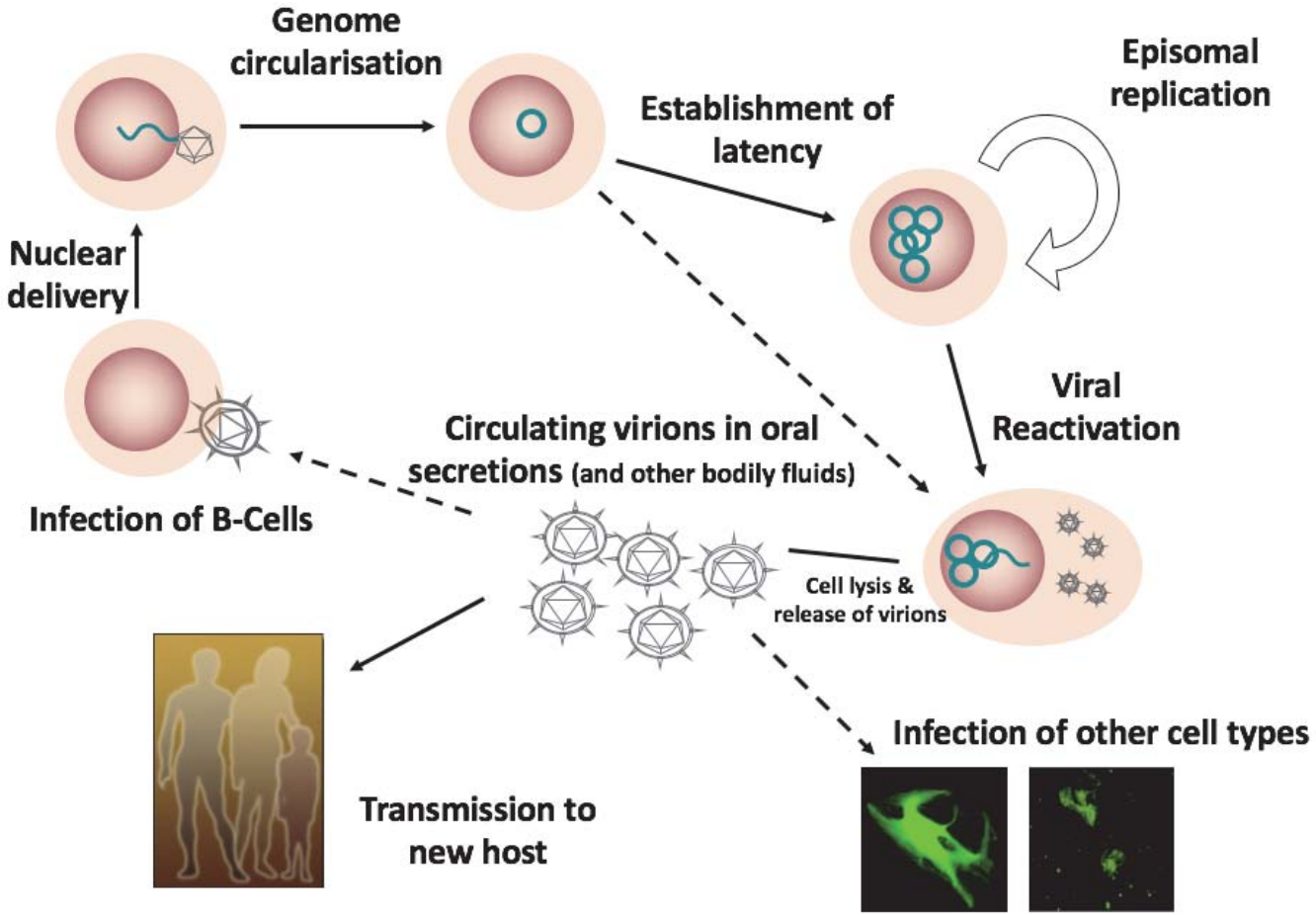


Fig. 1.2 General Life Cycle of EBV and KSHV. Following B-cell infection, typically via saliva, the virus genome is delivered to the nucleus where its' genomic DNA exists as a circular episome with a restricted pattern of gene expression and establishes latency. The virus can also reactivate from latency; whereby lytic replication occurs allowing spread of the virus progeny to other cells or be transmitted to other hosts. Adapted from Dalton-Griffin, L, 2010[4]

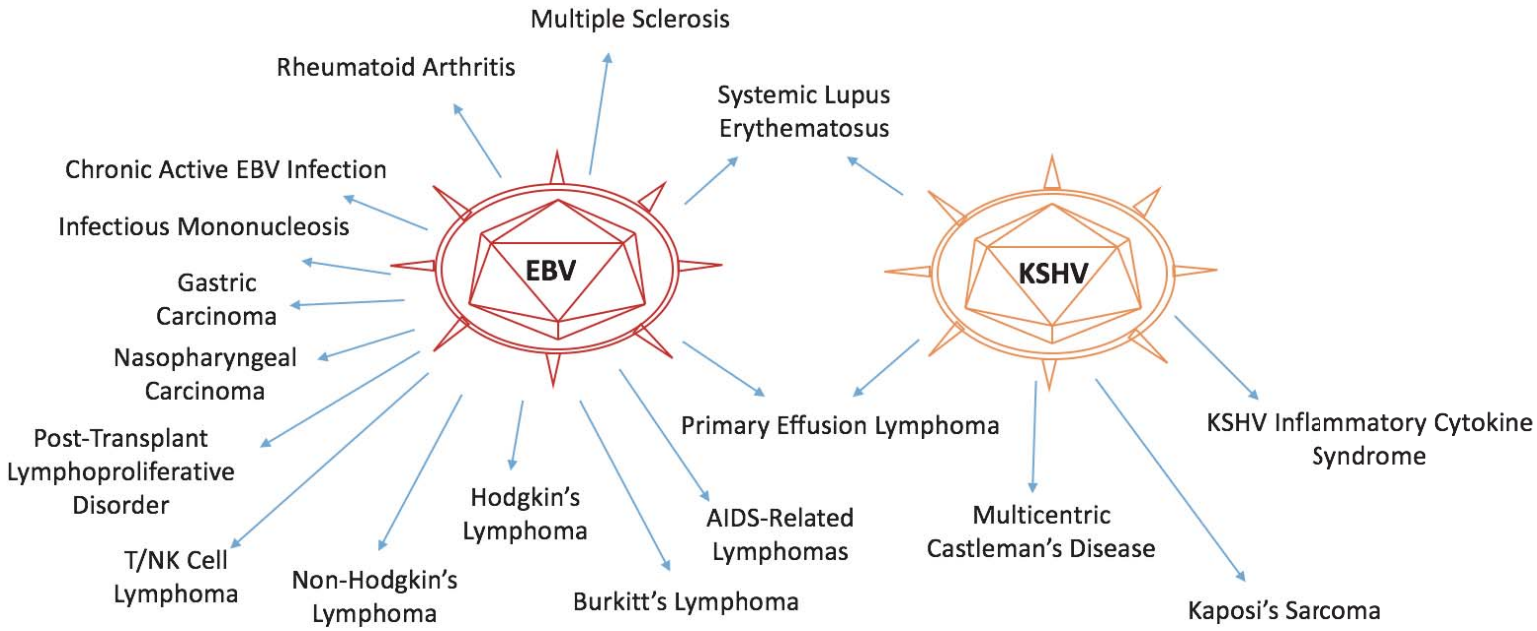


Fig. 1.3 Summary of the range of diseases associated with EBV and/or KSHV infections. EBV and KSHV have been associated with a spectrum of diseases including: chronic infection, tumours, immunosuppression-associated lymphoproliferative disorders and autoimmune diseases. This list is by no means exhaustive.

1.2 Epstein-Barr Virus (EBV)

EBV was the first γ-herpesvirus to be discovered, by Anthony Epstein, Yvonne Barr and Burt Achong in 1964, who examined electron micrographs of cells cultured from a tumour affecting children in sub-Saharan Africa, now known as Burkitt's Lymphoma, from a study which had been initiated in Uganda over 10 years prior by the Irish surgeon, Denis Burkitt[5,6]. Subsequently, Henle and colleagues discovered that EBV could transform B cells in umbilical cord lymphocytes to become continuously proliferating, immortalised lymphoblastoid cell lines (LCLs)[7]. This discovery also provided an invaluable experimental system for the future study of EBV infection. Shortly after, they found EBV to be the causative agent for infectious mononucleosis (IM)[8] and in 1970, another group detected EBV DNA in tissues isolated from patients with Nasopharyngeal Carcinoma (NPC)[9]. In the 1980's, EBV was linked to other cancers including non-Hodgkin's Lymphoma and oral hairy leukoplakia in individuals with Acquired Immunodeficiency Syndrome (AIDS)[10,11]; and subsequently EBV DNA was also detected in T-cell Lymphoma and Hodgkin's Disease Reeds-Stenberg cells[12,13]. EBV became the first human virus aetiologically linked to the development of cancers.

1.2.1 Epidemiology of EBV Infection & Associated Diseases

Globally, 95% of the adult population are infected with EBV as diagnosed by detection of antibodies to the EBV nuclear antigen -1 (EBNA-1) latent antigen or the viral capsid antigen (VCA) lytic antigen. EBV is predominantly transmitted via contact with oral secretions early in childhood and nearly all infected individuals actively shed virus in saliva[14]. In developing countries, seroconversion has been observed before the age of two[15,16], whereas in developed countries infection usually occurs in adolescence. While transmission in children is mainly via saliva exchange from parent-child, in adolescence or young adulthood, acquisition from intimate partners via kissing is thought to be a likely route[17-19]. Other potential sources of transmission are breast milk[20], blood transfusions and organ transplantations[21,22], in addition

sexual transmission has also been reported, however the evidence to support sexual transmission is limited[23,24]. Socioeconomic conditions have also been reported as risk factors for early EBV infection, for example low income, crowded living conditions, in addition to social patterns associated with poorer conditions such as the chewing of food by mothers prior to feeding children, or exposure to saliva by playing with unclean toys in nurseries[15,25-27].

Two major types of EBV, type 1 and type 2 (originally A and B) have been characterised based on sequence variability in genes encoding the latent nuclear antigens: EBNA-2 and the EBNA-3's (A, B and C)[28]. Type 1 is the most prevalent and is detected globally, while type 2 is rarely detected, and is more prevalent in parts of central Africa, Papua New Guinea and Alaska, and is also found commonly in individuals with the Human Immunodeficiency Virus (HIV)[24,29,30]. Type 2 reportedly has poor transforming ability of cells compared to type 1 EBV strains[31], nonetheless, it is still unclear whether the two EBV types contribute differently to pathogenesis and the development of disease.

The majority of people live with EBV infection with the absence of clinical symptoms throughout life. However, infection in young adults is associated with the self-limiting condition, Infectious Mononucleosis (IM) in >50% of cases. In addition, EBV is associated with 200,000 new cases of cancer including Burkitt's Lymphoma (BL), Hodgkin's Lymphoma (HL), Nasopharyngeal Carcinoma (NPC) and some gastric cancers (Fig. 1.3) and more than 140,000 deaths annually[32,33]. EBV is also reported to be a risk factor for the development of autoimmune diseases such as Systemic Lupus Erythematosus (SLE), Rheumatoid Arthritis and Multiple Sclerosis[34,35]. Unlike infection which is ubiquitous, some cancers associated with EBV have a varied geographic distribution in incidence[16]. BL is more common in equatorial Africa, particularly Malaria endemic regions, and it is thought that early exposure to EBV increases susceptibility[36-38]. NPC has higher incidences in southern China compared to the rest of the world[39].

Below is a brief description of four of the commonly studied diseases associated with EBV.

1.2.1.1 Infectious Mononucleosis (IM)

When primary EBV infection is delayed to adolescence or young adulthood, a self-limiting, benign lymphoproliferative disease known as Infectious Mononucleosis (IM) (also known as Glandular Fever) occurs in up to 70% of individuals, as a result of an acute infection associated with a large T-cell expansion[40]. IM is more common in western countries where exposure to EBV is found to occur mainly in adolescence compared to in developing countries. Symptoms of IM occur after an incubation period of 4-7 weeks and include fever, lymphadenopathy and pharyngitis[41]. The majority of patients do not need hospitalisation and infection is rarely fatal.

1.2.1.2 Burkitt's Lymphoma (BL)

Burkitt's Lymphoma (BL) is an aggressive malignant tumour of small, non-cleaved B-cells. Early studies by Denis Burkitt and others in equatorial Africa investigating a childhood tumour presenting in the jaw, later called BL, found it to be associated with the aetiological agent for Malaria, *Plasmodium falciparum* and subsequently discovered that EBV was present in >90% of cases[5,6]. This type of BL is considered endemic BL and mainly affects children in holoendemic (i.e. ubiquitous) malaria regions such as sub-Saharan Africa, Papua New Guinea and parts of Central America. Malaria co-infection is thought to enhance B-cell proliferation by inhibiting the T-cell response. The other type of BL is sporadic, usually occurring in western countries, such as the USA, and presents as an abdominal tumour, EBV is only associated with ~20% of sporadic BL cases. Epidemiological studies of children in Uganda have reported that elevated titres of antibody to the EBV structural protein, VCA, increases the risk of BL development[37,38,42].

1.2.1.3 Hodgkin's Lymphoma (HL)

Shortly after the discovery of EBV, Weiss and colleagues fist discovered the presence of the EBV genome in Reed-Sternberg cells of Hodgkin's Lymphoma[13]. Since then, EBV

has been associated with 40-60% of Hodgkin's lymphoma cases in the USA and patients have shown high antibody titres to EBV prior to the onset of, or with the development, of lymphoma in comparison to the general population[43,44]

1.2.1.4 Nasopharyngeal Carcinoma (NPC)

EBV is associated with ~100% of Nasopharyngeal Carcinoma (NPC) cases, with the genome being present in the epithelial cells of the nasopharynx. While the global incidence is ~2/100,000, NPC has the highest incidences reported in southern China of 20-30/100,000, and it is also reported to be prevalent in northern Africa and among Inuits from Alaska and Greenland. It has also been found to occur sporadically in the US and parts of western Europe. Serological detection of immunoglobulin A (IgA) antibodies to EBV has been useful in screening NPC patients for early detection in southern China as IgA antibodies are elevated in NPC patients and high titres following treatment for NPC have been associated with a poor prognosis compared to a declining or constant level[45-47].

1.2.2 The Biology of Infection

In 1984, EBV was the first herpesvirus to have its genome sequenced, using a prototype strain called B95.8[48]. The ~172kb genome encodes >80 genes which are subdivided into latent and lytic based on the stage in the life cycle they are expressed. EBV is mainly transmitted via contact with oral secretions and nearly all infected individuals actively shed virus in saliva[14]. While early studies suggested oral epithelial cells were the primary site of infection followed by B-cells[49,50], others suggest that B-cells in the oropharynx are the primary site of infection[51,52]. For viral entry into the cell, the EBV major envelope glycoprotein, gp350 binds to the CD21 receptor molecule on the B-cell surface, other factors including HLA class II molecules are also important for this process[53,54]. Primary infection results in short term "abortive" lytic replication and proliferation of B cells to avoid detection and control by the host immune response[55]. Memory B-cells are the reservoir of latently infected cells where the virus persists for life in the face of an active immune system, however, EBV can also infect T, NK and other cell types[56].

1.2.2.1 Latent Infection

Studies in LCLs have shown that in the latent stage of infection, EBV limits its gene expression program to encode less than 10 proteins: six nuclear antigens (EBNAs), two latency associated membrane proteins (LMPs) (Fig. 1.1) and two encoded small RNAs (EBERs), to manipulate host processes, keep the cell immunologically silent and maintain survival[57]. EBNA-1 plays a critical role in the maintenance of viral episomes during latency, it interacts with the DNA replication origin, *oriP,* on the viral genome to ensure efficient DNA replication and segregation occurs at each cell division[58]. EBNA-1 has also been shown to function as an immunomodulatory protein by interfering with Major Histocompatibility Complex (MHC) class I presentation and thus inhibiting T-cell mediated host responses[59,60]. Although its oncogenic potential has been debated, studies have observed that EBNA-1 is capable of transforming B-cells in transgenic mice resulting in B-cell lymphoma[61,62]; an EBNA-1-deleted mutant virus also had an attenuated ability to immortalise cells[63]; and EBNA-1 expression in EBV-negative NPC cells increased its tumorigenicity[64]. EBNA-2 is capable of modulating both cellular and viral gene expression, it reportedly mimics the Notch signalling pathway allowing it to bind to the host DNA binding protein CBF1, inhibiting differentiation and inducing B-cell proliferation[65-67]. EBNA-2 also upregulates LMP-1 and -2 to inhibit reactivation from latency[68,69]. The EBNA-3 proteins (A, B and C) are involved in the regulation of cellular gene expression and -3A and -3B are essential for B-cell transformation, and the EBNA leader protein (LP) is involved in augmenting EBNA-2's function[69-71].

The LMPs, LMP-1 and -2A/B are membrane bound proteins with distinct biological functions. LMP1 is a potent oncogene with several pleiotropic effects, its expression in mice leads to B-cell lymphomas[72,73]. It induces B-cell activation by mimicking a constitutively active form of the host CD40 receptor and regulates antibody production, isotype switching and the clonal expansion of B-cells[74-77]. LMP-1 can also activate nuclear factor-κB (NF-κβ) transcription factor by interacting with Tumour

Necrosis Factor (TNF) receptor associated factors and the Jun kinase (JNK) pathways for sustained proliferation of EBV-infected B-cells[78-83]. LMP-1 has also been shown to activate the phosphatidylinositol 3-kinase (PI3K) pathways, which play a role in suppressing apoptosis and promoting cell survival[84]. EBV LMP-2 blocks reactivation from latency via the inhibition of tyrosine kinase phosphorylation and its expression in transgenic mice allowed B cell survival in the absence of B-cell receptor signalling[85-87].

The role of the EBERs in B-cell transformation have been contradictory, however they have been found to induce the production of pro-inflammatory cytokines such as Il-10, important for cell growth[57]. To date, three latency programs: I, II and III can be established by EBV characterised by differential latent antigen expression and are associated with different malignancies, summarised in Table 1.2. EBNA-1 is the only protein expressed by all latency programs and also expressed during the lytic cycle, highlighting its indispensable function in EBV pathogenesis[88].

Table 1.2 EBV latency programs associated with infection

Latency Program	Latent Antigens Expressed	Malignancies
0	EBERs (LMP2)	None - Asymptomatic
I	EBNA-1	Burkitt's Lymphoma, Primary Effusion Lymphoma, AIDS-related DLBCL
II	EBNA-1, LMP-1 & LMP-2	Nasopharyngeal Carcinoma, Hodgkin's Lymphoma, T/NK-Cell lymphoma
III	EBNA-1, -2, -3, -LP and LMP-1 & LMP-2	Lymphoproliferative diseases, Infectious Mononucleosis

DLBCL – Diffused Large B-cell Lymphoma

1.2.2.2 Lytic Reactivation

EBV can be spontaneously reactivated from latency in a subset of infected B-cells, although this latent-lytic switch is poorly understood[57]. Studies in LCLs have shown that environmental stress induced by agents such as phorbol esters, sodium butyrate or by cross-linking immunoglobulin (Ig) on the cell surface can influence reactivation; in addition, environmental factors such as immune suppression can contribute to this process[57]. In the viral lytic cycle, EBV expresses all its lytic genes in a sequential order which are further categorised into immediate-early, early and late genes, based on the timing of their expression. The viral genome thereby linearizes and amplifies >100 fold allowing the production of infectious virions, facilitating spread and onward transmission[89]. Immediate early gene expression can be directly induced by B-cell receptor signalling *in vivo* with the simultaneous expression of BZLF and BRLF usually within ~2h[90-93]. BZLF and BRLF encode ZEBRA and RTA respectively that function as transactivators of lytic gene expression, and their functions include DNA replication, late gene expression and virion assembly[94-99]. Early lytic genes have a wider range of functions including DNA replication, metabolism and inhibition of antigen processing.

The late lytic genes encode structural proteins such as the glycoprotein gp350 and the viral capsid antigen, VCA.

The lytic cycle has been reported to upregulate the expression of viral/cellular cytokines and growth factors, in a subset of cells, thereby stimulating the proliferation of latently-infected neighbouring cells[100]. While malignancies of EBV occur mostly in the latent state of infection, the lytic cycle is a key driver in oncogenesis as shown by *in vitro* and *in vivo* studies[101-103].

1.2.3 Host Immune Responses to Infection

EBV can successfully co-exist with its human host throughout life without the presence of clinical symptoms. The rare occurrence of diseases in the immunocompetent host in contrast to the frequency and severity in the immunocompromised highlights the indispensable role of the immune system in controlling EBV infection.

1.2.3.1 Antibody Response

In immunocompetent individuals, following primary infection with EBV, the virus stimulates a strong humoral response by the host which produces a repertoire of antibodies against antigens marking the different stages of infection (Fig. 1.4). EBV–specific antibody assays have been developed for EBV serodiagnosis and typically rely on the detection of anti-VCA IgG, a marker of recent/active infection and anti-EBNA-1 IgG, a marker of infection history with EBV[104]. IgG antibodies reflect the cumulative exposure to EBV and are the most abundant type of antibody elicited by plasma cells in response to an infectious agent. In the initial stage of infection, high levels of maternal (IgM) antibodies against VCA are produced, which disappear within a few weeks following exposure. Within the first week of infection when most individuals show no clinical symptoms, anti-VCA IgG levels rapidly increase, peaking ~2 weeks during acute infection. Antibodies against VCA IgG decline slowly during the

convalescence period, remaining steady throughout life. During primary infection, IgG antibodies to Early Antigen (EA) also peak and disappears during convalescence, however, it may be detected at low levels in some individuals periodically throughout life. In the immunocompetent host, anti-EBNA-1 IgG gradually rises and peaks during the convalescence period, remaining fairly constant throughout life.

A relationship between EBV seroconversion and IM was first established by Henle and colleagues in 1968. Following primary infection, IM patients developed IgM followed by IgG antibodies specific for the VCA[8]. Transmission studies observed that EBV infection elicited the production of heterophile antibodies and IM, providing a causal association between EBV and IM[105]. Immunoglobulin antibody levels have also been reported to influence the development of certain EBV-associated diseases, however the strength of the evidence is variable, with the most conclusive evidence observed for the development of Nasopharyngeal Carcinoma (NPC) and Hodgkin's Lymphoma (HL) (reviewed extensively by Coghill and Hildesheim, 2014[106]). Multiple case-control studies conducted in Chinese populations, which are at a risk for NPC, have demonstrated that elevated antibody titres, particularly IgA against VCA, precede the development of NPC[107-111]. Unlike IgG antibodies, IgA antibodies are expressed at mucosal surfaces such as the oral or nasopharyngeal epithelium and reflect more recent/active infection and its utility as potential marker for defining risk of NPC is currently being assessed[110,112,113]. Strong evidence also exists for the association between IgG antibody response and risk of HL with elevated antibody titres against VCA and EA-D and EBNA antigens in HL patients[114-118]. While the evidence supporting antibody responses to BL is not as robust as that for NPC and HL, early sero-epidemiological studies conducted in Uganda established a causal link between EBV and BL and showed that EBV infected individuals with high antibody titres against VCA and EBNA compared to the mean control group were 30 times more at risk of developing BL[42,119]. More recently, two hospital based case-control studies observed elevated anti-VCA IgG titres in children with BL compared to controls[37,38]. Antibody measures however have limited utility in the diagnosis of malignancies aside from NPC[42,119-127].

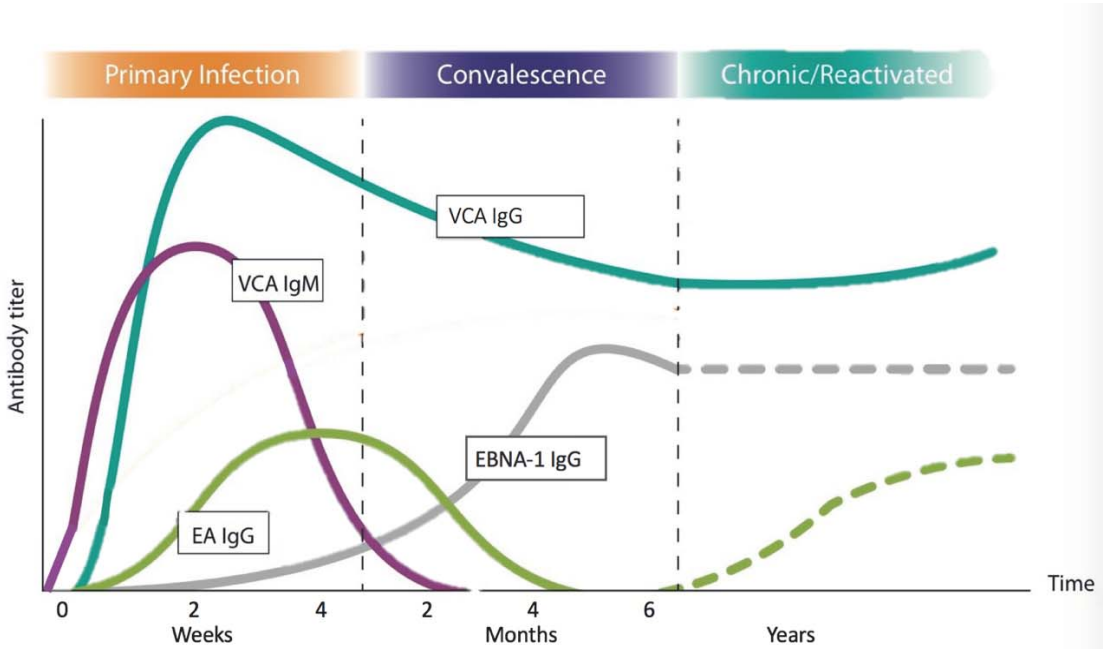


Fig. 1.4 EBV antibody dynamics in the immunocompetent host following primary infection. The antibody repertoire produced by the host response changes with time as primary infection with EBV progresses and marks different stages of infection. Adapted from katkars.com/pages/vidas-ebv.htm.

1.2.3.2 T-Cell Response

The cellular immune response to EBV is not very well studied, however it is essential in controlling primary and persistent infection[128]. In the immunocompetent host, CD4+ T helper and CD8+ cytotoxic T cells are readily detected in EBV positive transformed B-cells and target them for destruction and thus controlling EBV infection[129]. Following primary infection, CD8+ T cells are highly activated by the presentation of antigens (particularly early lytic antigens BRLF1 and BZLF1) by HLA class I molecules to target cells for destruction[130]. When primary infection occurs post-childhood in >50% cases an over-expansion CD8+ T cells have been reported resulting in Infectious Mononucleosis (IM). Studies have shown that while viral loads at this stage of infection are similar in children, an over production of CD8+ T cells has not been reported[130-132]. The reasons for CD8+ T cell expansion leading to IM in adults vs in children is still not known. CD4+ T cells play a role in maintaining CD8+ T cell responses and also responds to a repertoire of latent and lytic antigens following presentation by HLA class II molecules[129,133,134]. CD4+ T cell deficiency has been linked to viral reactivation in immunocompromised individuals with EBV and/or KSHV infections[134,135], highlighting the importance of CD4+ T cells in controlling persistent infection.

1.3 Kaposi's Sarcoma-Associated Herpesvirus (KSHV)

Kaposi's sarcoma-associated herpesvirus (KSHV) also known as human herpesvirus-8 (HHV-8) was first isolated by Chang and colleagues in 1994 as the aetiological agent of Kaposi's Sarcoma (KS), an endothelial tumour originally defined by Mauritz Kaposi in 1872 as an "idiopathic multiple pigmented sarcoma of the skin"[136,137]. Within two years of its discovery, the fist KSHV genome was sequenced using Sanger sequencing by Russo and colleagues, revealing a ~165kb dsDNA genome with a ~140kb long unique coding region (LUR) (Fig. 1.5)[138]. KSHV is also found to be associated with other lymphoproliferative disorders particularly, Primary effusion lymphoma (PEL) and Multicentric Castleman's disease (MCD)[136,139,140]. More recently, KSHV Inflammatory cytokine syndrome (KICS) has been reported in individuals with KSHV and HIV co-infection, resulting in elevated levels of IL-6 production[141]. KSHV-associated disease predominantly occurs in immunosuppressed individuals[142], thus widespread HIV infection has fuelled the KS epidemic in sub-Saharan Africa. KSHV accounts for up to 10% of cancers in African men and is the leading cause of HIV-associated cancer[143,144]. Virus transmission is mainly via saliva[145,146], however detection of viral DNA in peripheral blood mononuclear cells (PBMCs) suggest blood-borne transmission can also occur. Studies have also reported sexual transmission in homosexual men from the USA[147,148]. Despite high seroprevalence, only a small proportion of people infected develop tumours, in addition, uneven geographical distribution in seroprevalence along with familial clustering of cancer are suggestive of environmental and genetic factors associated with disease[145,149-152].

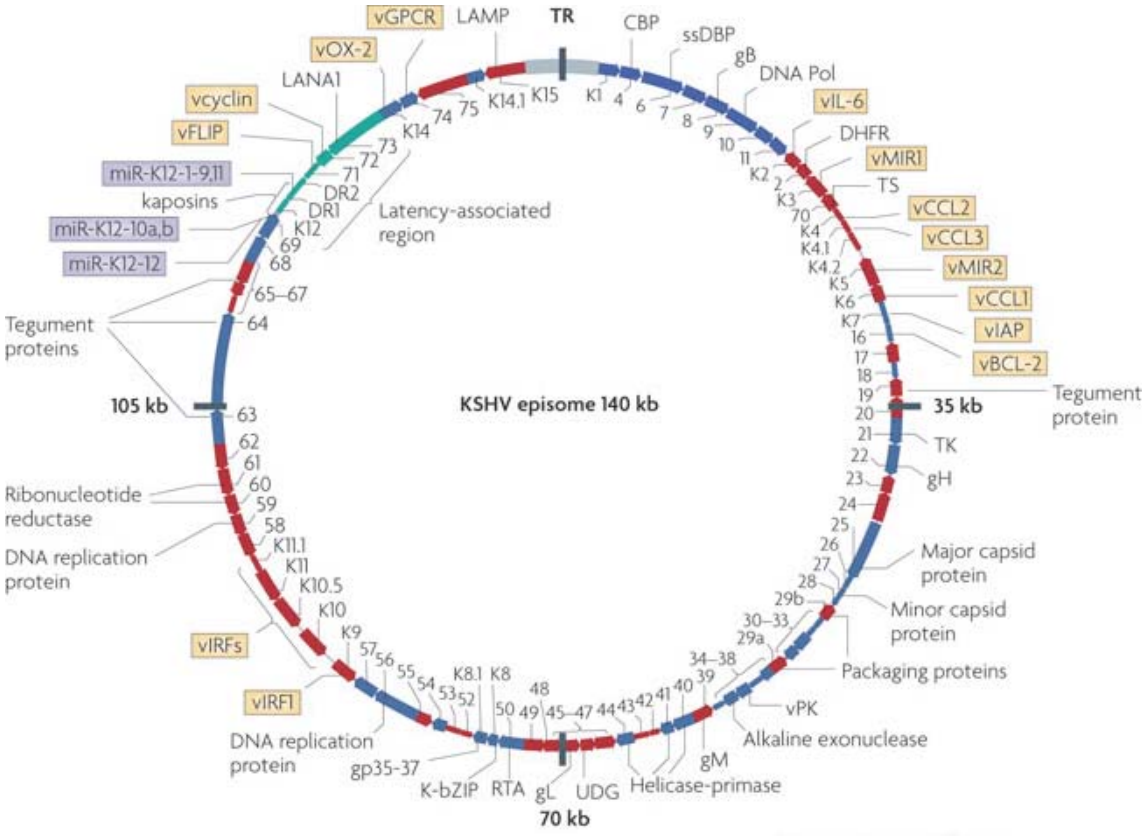


Fig. 1.5 The KSHV Episome. The ~140kb KSHV long unique coding region is flanked by ~800bp GC-rich terminal repeat (TR) regions. KSHV encodes ~87 open reading frames (ORFS) including viral homologues of cellular proteins (yellow boxes) and > 17 microRNAs (purple boxes). Following viral infection, the KSHV genome exist as a circular episome within the host nucleus. Latency is the default mode of infection and is necessary for immune evasion, established by the ORF73 protein, LANA. Latency associated transcripts are in green. ORF50 is the master regulator switch for the viral lytic replication programme. In the lytic stage the early lytic genes (red) are expressed for DNA replication and gene expression and the late lytic genes (blue) encode structural proteins for virion assembly. From Mesri, Cesarman & Boshoff, 2010[153].

1.3.1 Epidemiology of KSHV Infection & Associated Diseases

Unlike EBV infection, infection with KSHV is not ubiquitous, seroprevalence of KSHV varies greatly and its global distribution is found to parallel the incidence of Kaposi's Sarcoma. As KSHV DNA is not readily detected in all individuals, diagnosis of infection is made predominantly by the detection of antibodies to latent or lytic antigens. However, epidemiological studies estimating and comparing seroprevalence globally, between research groups, has been hindered by the lack of high sensitivity and specificity, universally accepted, assays for KSHV serodiagnosis. Nevertheless, three major patterns of seroprevalence have been consistently identified: high endemic (seroprevalence >30%), intermediate endemic (seroprevalence ~10-30%) and non-endemic (seroprevalence <10%). The highest seroprevalence rates have been reported in sub-Saharan Africa of up to >80% in adult populations. In Uganda where most adults are KSHV seropositive, variation in seropositivity has been reported ranging from 34% in a population based study in Mukono district to ~88% in individuals at the Uganda Cancer Institute in Kampala[154-156]. Other countries with high seroprevalence rates in sub-Saharan Africa include, Botswana (76-87%), The Gambia (29-84%), South Africa (35%), Cameroon (28-62%) and Zambia (47-58%)[157-163]. In the Mediterranean regions where classic KS is prevalent, intermediate seroprevalence of KSHV has been reported[150,164,165], whereas in the US and parts of Europe and Asia seroprevalence is low in the general population but relatively higher in homosexual men, suggestive of a sexual transmission route[147,148]. The San Francisco Men's Health Study reported the prevalence of KSHV infection amongst homosexual men to be 37.6%, which was also correlated with number of sexual partners[166] . Interestingly, in China where KSHV seroprevalence is generally low, classic and AIDS associated KS are found predominantly in the Uyghurs and Kazakhs ethnic groups in southern parts of China with KSHV seroprevalence of 48% and 67% respectively, and they have been found to acquire KSHV in childhood[167,168]. The Amerindians in Brazil have also been categorised a KSHV hyperendemic population with seroprevalence in children >70% and increasing with age up to 90% compared to non-Amerindian ethnic groups[169]. These trends are similar to that reported in sub-Saharan Africa, with exposure and seroconversion occurring early in childhood, highly suggestive of vertical transmission

of KSHV DNA in saliva. The seroprevalence of KSHV infection in children from endemic areas in sub-Saharan Africa has been shown to correlate with mothers shedding a high number of viral DNA copies/ml[170-174]. KSHV transmission studies conducted in Zambia have shown that differences in KSHV genotypes within household suggest that children and adults can contract KSHV infection from sources outside the family unit[175]. As the prevalence of childhood infection in children from developed countries is low, very little is known about the modes of transmission[176].

The risk factors associated with KSHV acquisition and development of disease are not well understood. The striking geographic variation in seroprevalence with increased seroprevalence in certain ethnic groups and disparities in disease incidence suggest host genetic and/or environmental cofactors are involved. Similarly to EBV, specific child feeding practices in endemic areas in Zambia have been reported to be correlated with early childhood infection and substantiate previous findings for mother-child transmission of KSHV[177]. Immunosuppression by HIV coinfection or organ transplantation is known risk factor for all KSHV associated malignancies; furthermore, individuals co-infected with HIV have been reported to have increased viral shedding in saliva. A prospective longitudinal cohort study in Zambia showed that KSHV acquisition was 5-fold higher in HIV positive children than in HIV negative children[178]. A study conducted in Ugandan mother-child pairs showed that along with HIV coinfection, malaria parasitaemia was also associated with KSHV seroprevalence; a subsequent study then showed that in individuals with malaria coinfection antibody titres to LANA and K8.1 were elevated compared to those who didn't have malaria[179,180]. Studies conducted in KSHV endemic regions have reported other factors such as chronic Schistosome and other parasitic infections, licking of wounds as a result of insect bites, exposure to natural plant extracts used in traditional medicines by some cultures, sources of water and living near volcanic soils can influence seropositivity, KSHV viral reactivation and lytic replication[181-185]. However, it is unknown whether these factors are associated with seroprevalence of infection or developing disease.

The whole genome diversity of KSHV strains remains largely uncharacterised with whole genome data published from only three different countries, Greece, USA and most recently Zambia reflecting a large gap in the global characterisation of KSHV whole genomes (described in detail in chapter 5). While characterisation of EBV strains has been mainly achieved using the whole genome, KSHV strains have been categorised predominantly based on variation in K1 and K15 variable genes (discussed in chapter 5). K1 is located at 5' termini of the KSHV genome (Fig. 1.1 and Fig. 1.5) and encodes a cell surface glycoprotein involved in signal transduction, cell transformation, stimulation of inflammatory cytokines and down regulation of the B-cell receptor[186-188]. The K15 gene also known as Latency associated membrane protein (LAMP) is on the 3' termini of the genome (Fig. 1.1 and Fig. 1.5) and has been involved in induction of inflammatory cytokines and chemokines as well as the inhibition of B-cell receptor signalling[189-191]. Molecular epidemiological studies have revealed a distinct genotypic distribution of K1 genotypes and this has been reported to reflect the co-evolution of KSHV with its human host through time[144,192]. A relationship between KSHV genotypes and susceptibility to KSHV or pathogenesis of associated malignancies (described below), however remains to be elucidated.

1.3.1.1 Kaposi's Sarcoma (KS)

Kaposi's Sarcoma (KS) is an angioproliferative, spindle cell endothelial tumour of lymphatic origin that was first discovered by the Hungarian dermatologist, Dr. Moritz Kaposi in 1872[137,193,194]. Clinically, KS presents as highly pigmented dermatological lesions that can be found cutaneously, mucosally or viscerally. With the AIDS epidemic from 1980's the incidence of KS dramatically increased particularly in homosexual HIV-positive men, suggesting the involvement of an infectious agent. In 1994, Chang and Moore discovered the DNA of a novel human gammaherpesvirus in biopsies of AIDS patients that was associated with 95% of infected cells, they named it KSHV[136]. KS has since been categorised in to 4 different sub-types based on epidemiological and clinical outcomes: classic (sporadic), endemic (African), epidemic (AIDS-associated) & iatrogenic (post-transplant)[195]. Classic KS was the first to be described, it occurs with a slow progressing course and has been typically found in elderly men of Mediterranean and Eastern European descent[196-198]. Endemic KS can

be more aggressive than classic KS and has been typically common across equatorial Africa (referred to as the KS belt) such as Uganda, Kenya, Tanzania, and Cameroon, prior to the AIDS epidemic, and is also common in parts of southern Africa such as Malawi, Zambia, South Africa and Zimbabwe[157,199]. Epidemic or AIDS-associated KS is the most common type of KS in individuals and its occurrence peaked during the HIV epidemic; thus was also a marker for HIV disease in the 1980's, and with the roll out of highly active antiretroviral therapy (HAART) in 1990's a decline in AIDS-KS was observed[200]. Iatrogenic KS is associated with immune suppression following therapy used to prevent allograft rejection during organ transplantation and is most common in Renal transplant patients[153].

1.3.1.2 Primary Effusion Lymphoma

Primary Effusion Lymphoma (PEL) also known as body cavity based lymphoma (BCBL) is a B-cell non-Hodgkin's Lymphoma primarily associated with KSHV infection and is predominantly found in patients with AIDS[139,201]. PEL is distinguishable by lymphomatous effusions in serous cavities with an absence of a solid tumour mass. Frequent co-infection with EBV occurs in PEL; both viruses promote latency by subverting the host immune response and use B-cells as a reservoir of infection[202]. EBV mimics host cell signalling pathways stimulating a memory B cell phenotype whereas KSHV infection drives B-cell development towards a 'long lived' plasma cell by 'pirating' host genes involved in signalling, proliferation and apoptosis inhibition and has been linked to its transformation *in vivo*[202,203]. While PEL only accounts for ~3% of AIDS associated lymphomas its discovery has driven KSHV research, as they were developed as the first KSHV positive cell lines with 40-80 KSHV genome copies per cell and the first KSHV whole genome was derived from the PEL cell line BC-1. In addition, PEL cell lines have been a useful tool in serodiagnosis of KSHV, virus purification and studies of the KSHV gene expression and stages in the viral life cycle[204,205].

1.3.1.3 Multicentric Castleman's Disease

Multicentric Castleman's Disease (MCD) is a non-neoplastic, reactive lymphoadenopathy of which the plasmablastic variant was found to be associated with KSHV[206]. While MCD is poorly understood, it is characterised by immune dysregulation, marked by an overproduction of IL-6[207]. Like KS and PEL, KSHV DNA is detected in nearly all HIV-associated cases of MCD and ~50% of HIV negative cases[206,208].

1.3.2 The Biology of Infection

KSHV is predominantly transmitted via oral shedding of viral DNA in saliva, however other modes of transmission do exist. Following infection, attachment and entry of KSHV in to target cells via endocytosis is a complex, multi-stage process, mediated by the interaction between multiple KSHV envelope glycoproteins particularly, gB, gH, K8.1 and KSHV complement control protein, extensively reviewed by Chandran[209]. KSHV primarily targets B-cells, however may spread to endothelial cells and other cell types such as dendritic cells, monocytes and macrophages (Fig. 1.2)[210]. Like all herpesviruses, KSHV establishes a lifelong infection via two phases of the viral life cycle: latent and lytic replication cycles, each with distinct patterns of gene expression (Fig. 1.6). In the immunocompetent host, following viral entry, KSHV exists as a linear genome, upregulates genes involved in B-cell activation usually within the first 10 hours of infection, this lytic infection is transient and quickly followed by the establishment of latency to avoid host immune surveillance[211,212]. Similarly to EBV, in the latent stage of infection KSHV exists as a circularised episome with a restricted pattern of viral gene expression allowing genome maintenance, the evasion of host immune responses and promoting persistence in peripheral CD19+ B-cells[203].

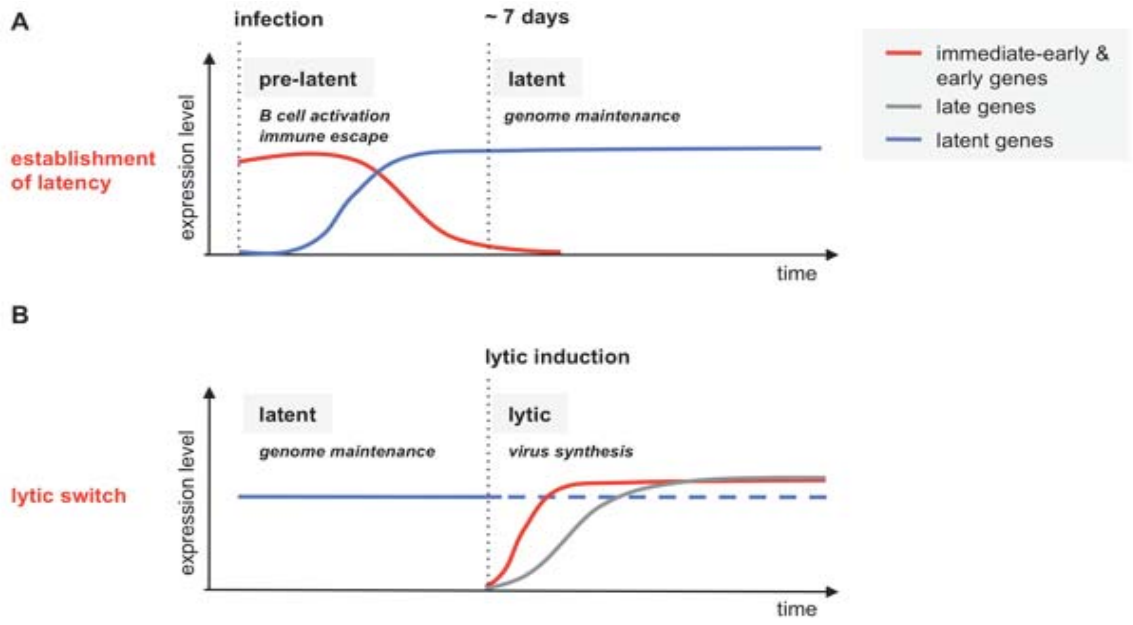


Fig. 1.6 KSHV gene expression dynamics following primary infection. The initial stages of KSHV infection result in B-cell activation and abortive lytic replication to facilitate immune escape. KSHV then circularises and down regulates it gene expression to establish latency throughout life. Intermittent lytic reactivation may occur throughout the course of infection resulting in the upregulation of its lytic gene repertoire. From http://www.helmholtz-munchen.de

1.3.2.1 Latent Infection

Latency is the default infection pathway used by the virus to evade immune surveillance, also all KSHV-associated malignancies occur in the latent stage of infection. Most studies on KSHV latency have been performed using PEL cell lines, which are able to maintain stable viral episomes and allow for the detection of latent transcripts in all cells. KSHV exhibits a restricted viral gene expression program during latency just as EBV, however only encoding four proteins: the viral trans element, latency-associated nuclear antigen (LANA), encoded by Open Reading Frame (ORF) 73, v-Cyclin, encoded by ORF72, vFLIP (also known as K13) encoded by ORF71 and the Kaposin family encoded by K12[213-218]. LANA is a multifunctional, versatile oncogenic protein which is known to interact with various cellular and host proteins and is crucial for DNA replication, efficient viral genome segregation and maintenance of latency. LANA is the most consistent antigen expressed in all KSHV-infected cells, it

binds to the terminal repeat (TR) motifs at the termini of the genome via its C-terminal domain, thus allowing genome maintenance during cell division[219,220]. LANA has also been shown to bind to tumour suppressor genes such as p53, transcription factors, chromatin binding proteins and signal transducers to inhibit apoptosis, drive spindle cell proliferation and contribute to tumourigenesis; it also maintains latency by binding to viral promoters and inhibiting lytic replication[221,222]. vCyclin is the viral homologue of the host protein Cyclin D, while its role is not fully understood, it is reported to contribute to regulating the cell cycle and sustaining proliferation of cells; and also modulates the latent–lytic switch[223-225]. vFLIP is the homolog of the cellular FLICE (Fas-associated death domain (FADD) like interleukin-1 beta-converting enzyme) inhibitory protein and plays a role in promoting cell proliferation and providing a survival advantage by blocking apoptotic pathways and inhibition of the NF-κβ pathway, which has also been shown to suppress lytic replication in latently infected PEL cells[226-229]. The Kaposin locus are the most abundantly expressed viral transcripts in latent infection encoding for three proteins Kaposin A, B and C. Kaposin B has been found to modulate cytokine expression and influence the proinflammatory microenvironment in KS tumours, however, their functionality is still poorly understood[217,230].

1.3.2.2 Lytic Reactivation

The KSHV latent-lytic switch is a complex and tightly regulated process that is highly sensitive and responsive to different cellular and physiological cues including: viral co-infection, immune suppression, cellular oxidative stress, hypoxia, inflammatory cytokines and treatment with chromatin modifying agents such as sodium butyrate[231]. The lytic phase displays an ordered cascade of expression of at least 80 transcripts categorised into immediate early, early and late genes based on the timing of their expression. In the lytic stage the virus expresses both 'pirated' host genes involved in signal transduction, cell cycle regulation, apoptosis inhibition and immune modulation to manipulate host processes and drive active replication of linear viral genomes utilizing virus replication and structural genes, thereby facilitating virus production, spread and pathogenesis[204,232-235]. The molecular switch

that controls KSHV reactivation is the immediate early gene product, Replication and Transcription Activator (RTA), encoded by ORF50, this is found to be the key protein that is sufficient and necessary for KSHV reactivation[10,236]. RTA auto-activates the expression of its own promoter followed by the activation of downstream immediate early genes that are expressed within the first 10 hours of infection and are important for gene transcription and cellular modifications for viral replication. The Early genes are then expressed from 10h-24h post induction of the lytic cycle for efficient DNA replication and gene expression, the late genes follow at ~48h post infection and encode structural proteins including gB and K8.1, essential for virion assembly and maturation[204].

An understanding of the lytic cycle is important because although KSHV-associated disease occurs in the latent stage, background lytic reactivation has been found to occur stochastically and may play a role in driving oncogenesis. Recurrent reactivation is also found to be partly attributable to the expansion of the reservoir of infected B-cells and also plays a role in the terminal differentiation of B-cells into antibody-secreting plasma cells as shown in PEL cell lines[237,238].

1.3.3 Host Immune Response to Infection

The host immune response to KSHV parallels that of EBV, with the humoral response characterising different infection stages and antibody responses used potentially as a marker for serodiagnosis and the cellular immune response is important in controlling infection. The rare occurrence of KS and other malignancies in the general population in contrast to prevalence in HIV and organ transplant immunocompromised individuals highlights the importance of host immune response in controlling KSHV infection.

1.3.3.1 Antibody Response

Following infection with KSHV, the host elicits a repertoire of specific antibodies against antigens marking the different stages of the infection life cycle. KSHV–specific antibody assays have been developed for serodiagnosis and typically rely on the

detection of anti-LANA IgG, a marker of infection history with KSHV and of anti-K8.1 IgG, a marker of recent/active infection[104]. Antibodies against the minor capsid protein ORF65 have also been used a marker of infection. The antibody response dynamics for LANA and K8.1 are highly similar to that of EBV EBNA-1 and VCA responses (Fig. 1.4). Very recently, Labo and colleagues have developed a multiplex bead based assay for KSHV detection of 72 expressed KSHV antigenic proteins in the plasma of adults with KSHV-associated malignancies, allowing them to systematically characterise the KSHV proteome[239]. While seroreactivity patterns were highly varied amongst patients, antibodies against a set of antigens were consistently produced in all patients, with K8.1 eliciting the strongest response, followed by ORF65, LANA, ORF38, ORF71, ORF59 and K5[239]. This new multiplex assay expanded on data provided by previous serological studies, and provides a powerful new tool for serodiagnosis. The influence of antibody response in development of disease is unclear. Nonetheless, a few studies have observed that KS risk is increased in individuals with high lytic K8.1 or latent LANA antibody titre[163,240-243].

1.3.3.2 T-Cell Response

The host T-cell response to KSHV infection is even less well studied in comparison to EBV[244]. Unlike EBV, KSHV primary infection is not marked by an altered expansion of T cells and is not associated with any obvious primary infection syndrome such as Infectious Mononucleosis[131,245,246]. Cytotoxic CD8+ T cells produced in response to primary infection target lytic antigens, presented by HLA class I molecules, have been reported to peak within the first two months of infection and decline along with antibody titres, suggestive of reduced viral replication[246]. A higher frequency CD8+ T cell responses has been observed in asymptomatic KSHV positive individuals compared to immunocompromised patients with KS[247-249]. In addition, immune reconstitution in HIV infected individuals following highly active retroviral therapy (HAART) has been shown to correlate with a decline in KSHV viral load and increase in CD8+ T-Cell responses, suggesting CD8+ T-cells are important in viral clearance[250]. CD4+ T helper cell proliferation has also been reported in response to HLA class II presentation of KSHV lytic antigens following infection, and has been found to control viral lytic replication in B-cells[251-253].

1.4 The Influence of Host Genetics on Infectious Diseases

Following exposure to an infectious pathogen, a combination of host and pathogen genetics, in addition to other environmental exposures, are contributing factors in the susceptibility to infection and to the phenotype that arises in a particular individual, which ranges from asymptomatic carriage or mild symptoms, to progressive disease or death. The host and pathogen counterparts co-evolve, co-adapt, are both genetically variable and their interaction is a point of genetic selective pressure. Evolutionary selective pressures exerted by infectious diseases were evident since the 1950's with the classic example of sickle cell anaemia in individuals homozygous for haemoglobin S (HbS) allele, whereas carriers who were heterozygous had a selective advantage being highly protected against *Plasmodium falciparum* Malaria[254]. Early studies in twins also highlighted a substantial host genetic component contributing to the susceptibility to infectious diseases such as Tuberculosis (TB), leprosy, Poliomyelitis and Hepatitis B[255-258]. In the late 1980's, a landmark study was conducted in adopted children whose relative risk (RR) of death from infection was significantly higher (RR=5.81) if they had a biological parent that died prematurely as a result of infection, in contrast to the death of an adoptive parent that had no significant influence on risk (RR=~1); this provided evidence that host genetics contributed significantly to risk of death from infectious diseases[259]. These early studies paved the way for the investigation of specific candidate genes with known or hypothesised biological function that influenced susceptibility to infectious diseases.

In the 1990's, a number of studies discovered mutations in single genes that led to susceptibility/resistance to disease. A study led by Tournamille and colleagues identified a point mutation in the gene encoding the Duffy antigen receptor for chemokines (DARC) which led to the inhibition of expression on erythrocytes and conferred resistance to *Plasmodium vivax* Malaria[260]. For HIV infection, Samson *et al*, discovered a 32 bp deletion in the chemokine receptor 5 (CCR5) gene in HIV negative Caucasian individuals that made target cells resistant to HIV infection, thus conferring protection[261]. Subsequent studies also identified high penetrance mutations in genes

involved in immune function such as interleukin-12 (IL-12) and interferon-gamma (IFNγ) that led to increased susceptibility to Mycobacterial disease (e.g. TB) and Salmonella caused by intracellular pathogens[262-264]. More recently, Everitt *et al*, conducted a siRNA screen of candidate genes involved in susceptibility to influenza viruses and identified interferon-induced transmembrane 3 (IFITM3) which reduced the morbidity due to influenza, and mice deficient in *IFITM3* succumbed to severe viral pneumonia when infected with influenza strains with low pathogenicity[265]. They also identified a single nucleotide polymorphism (SNP) in *IFITM3* (rs12252-C) which encodes a truncated protein and was found in 5.7% of hospitalised European individuals with severe influenza compared to 0.3% of population controls[265]. Subsequently, the rs12252-C *IFITM3* mutation was replicated in an independent cohort and present in 69% Chinese severe influenza patients compared to 25% controls with mild illness[266].

While these discoveries highlight single genes with major effects, these variants tend to be rare, and at the population level most individuals do not succumb to infection/disease as result of Mendelian defects, rather infectious diseases are predominantly complex traits that occur as a result of the interaction between host-pathogen genomes and environmental factors.

1.4.1 Genome-Wide Association as a Tool to Study Infectious Diseases

1.4.1.1 Genome-Wide Association Study (GWAS)

In the early 2000's, shortly after the completion of the Human Genome Project, a turning point in the field of human genetics occurred driven by the HapMap project[267] which mapped thousands of variants across the genome and identified their correlation through linkage disequilibrium (LD) and thus became the driving force for the development of high resolution genotyping arrays, and the use of the genome-wide association study (GWAS) as a tool to investigate genetic risk factors associated with common diseases. GWAS typically scan markers i.e. single-nucleotide polymorphisms (SNPs) across the entire genome to identify statistically significant

differences in allele, or genotype, frequencies in individuals with the phenotype under study compared to unaffected control individuals, or associated with a quantitative trait being studied[268]. Due to patterns of LD between SNPs within the same genetic region a SNP associated with a given trait may be "tagging" a true causal variant as result of high correlation[268]. While correlation makes inferring the causality of variants challenging because it is difficult to distinguish between the effects of highly correlated markers, it is leveraged by imputation algorithms that can infer genotypes that were not directly genotyped. Imputation algorithms infer missing genotypes based on a representative reference panel of haplotypes from the HapMap Project or 1000 Genomes Project[267,269,270]. Imputation has been greatly beneficial for GWAS by increasing the power to detect associated loci as reference panels are more likely to contain causal variants (or have a better tag) than the genotype array, in addition it increases the resolution of an association signal of the locus allowing for fine mapping and facilitates the meta-analyses of studies performed using different genotyping platforms[271].

In comparison to candidate gene studies that test for statistically significant differences in allele frequencies between cases and controls in markers within a gene of interest selected based on presumed biological function, GWAS has a number of strengths. The main success of GWAS was the appreciation of larger sample sizes and the need to adjust for multiple testing, leaving behind lenient p-values and using more stringent thresholds to identify and declare robust association signals. GWAS provides a high throughput, agnostic approach that is relatively unbiased (allowing for the discovery of completely unexpected new biological findings), it is not restricted to families or sibling pairs (unlike linkage studies), has greater statistical power to identify loci of small to moderate effect; and in addition, technological advances have made GWAS an increasingly cost-effective tool to investigate the genetic contribution to traits of interest.

There are two main GWAS study designs: Case-control and quantitative trait studies. Case-control study designs investigate binary traits such as diseased vs healthy and look for enrichment of particular variants in cases versus controls that predispose

them to the phenotype under study. Quantitative trait study designs typically investigate continuous variables that are predictors of a phenotype of interest (e.g. cholesterol levels as a biomarker for heart disease) and assume that multiple genetic variants of small to moderate effect sizes in addition to gene-environment interactions contribute to phenotypic variation.

In 2005, the first published GWAS compared ~116,000 single nucleotide polymorphisms (SNPs) between just 96 cases and 50 controls leading to a discovery of a SNP in the Complement Factor H gene that increased the risk of age-related macular degeneration seven-fold in homozygous individuals[272]. Since then, GWAS has been used commonly as a tool to study the genetic architecture of complex disease traits such as inflammatory bowel disease (IBD), diabetes and obesity, and has successfully uncovered thousands of novel genetic loci associated with disease[273-275].

1.4.1.2 Insights from GWAS of infectious diseases

In the context of infectious disease research, GWAS has been used to probe different aspects of the host response to pathogen infections using both case-control and quantitative study designs[276]. Case-control designs have investigated susceptibility or resistance to infection, or pathogen clearance, or presence of severe disease, while quantitative traits studies have used antibody titres, cell counts or viral load as markers of clinical outcome. In addition, GWAS has been used to investigate response to drug therapy and vaccination[277]. Below is a brief review of the literature of some of the successes of GWAS in infectious disease research.

In 2007, the first GWAS for an infectious disease was conducted by Fellay and colleagues investigating HIV-1 control. The authors investigated inter-individual variability in set point viral load (spVL), a commonly used marker of disease progression following infection, in 486 HIV positive individuals[278]. Following GWAS across ~500,000 SNPs, they identified a SNP in strong LD with *HLA-B*5701,* and another independently associated SNP, upstream of *HLA-C* that accounted for close to 15% of the variation in HIV-1 spVL[278]. They subsequently extended their study to

2,554 European patients, replicating their findings and confirming that *HLA-B* ($p=4.5x10^{-35}$) and *–C* ($p=5.9x10^{-32}$) play a central role in HIV-1 control[279]. This analysis was repeated in 515 HIV positive African-Americans replicating the association signal in *HLA-B* ($p=5.6x10^{-32}$) with spVL[280], and another GWAS in a large multi-ethnic cohort replicated previous findings and also identified novel variants in the MHC region[281]. The variant upstream of *HLA-C* has been found to result in altered *HLA-C* expression and a study in 1,698 HIV positive individuals of European ancestry revealed that high *HLA-C* expression correlated with a slower progression to AIDS indicative of more effective viral control than in individuals with low *HLA-C* expression[282]. A subsequent study in >5000 African- and European- Americans also showed an influence of *HLA-C* expression in HIV outcomes including viral load and CD4+ T cell counts, however high *HLA-C* expression was predisposing to Crohn's disease[283]. A very recent GWAS in 6,315 European HIV positive individuals observed strong association signals in HLA region and were able to map the top associations to amino acid positions in the peptide-binding groove of HLA-B and HLA-A[284]. This study is also the first GWAS to identify statistically significant variants in the *CCR-5* gene which together with variants in HLA region explained ~25% of the variability in viral load, thus, concluding that the control of HIV-1 outcome is mainly attributed to common variants with large effect[284]. In addition to these findings, two independent studies found that hypersensitivity to the HIV-1 drug Abacavir was associated with the HLA-B*5701 polymorphism in Caucasian individuals[285,286], highlighting the utility of GWAS findings in the clinic for screening patients prior to treatment.

Success has also been achieved in the study of chronic Hepatitis C (HCV) and Hepatitis B (HBV) infections that can lead to chronic inflammation of the liver, cirrhosis and in some cases hepatocellular carcinoma. A GWAS in a Japanese and Thai cohort identified variants in the *HLA-DP* locus strongly associated with chronic HBV infection, this has been replicated by other studies[287,288]. In European patients with chronic HCV infection, a SNP (rs12979860-C) upstream of the *IL-28b* gene encoding for IFNγ3 was associated with a 2-fold response to PegIFN2a/b combined with Ribavirin treatment resulting in viral clearance. This variant occurs at a much lower frequency in individuals of African ancestry compared to European and East Asian ancestries,

thus, explaining heterogeneity in responses between different ancestries[289]. This is also an example of an association that may translate into a clinical benefit. Variation in *HLA* genes have also been associated with viral outcomes following HCV infection[290].

Variants in HLA class II genes have also been associated with other infectious disease phenotypes including susceptibility to visceral Leishmaniasis, seropositivity to Human Papilloma Virus (HPV) infection, resistance to enteric fever caused by *Salmonella* infection and very recently has also been associated with protection from TB[291-294]. These associations highlight the influence of the MHC loci in response to pathogenic infection, confirming the importance of the adaptive immune response in the control of infection. The influence of the HLA on infectious diseases has been extensively reviewed by Blackwell *et al*,[295].

1.4.2 Genome-Wide Association Studies in African Populations

Currently, most GWAS have been performed for non-communicable diseases predominantly in populations of European ancestry and with less than 5% of GWASs conducted in African populations, the contribution of human genetic variation to disease traits in such diverse populations remains largely uncharacterized[296,297]. Owing to a long history of evolution and adaptation to varying environments, diet, demographic changes and exposure to infectious disease, African populations have the highest level of human genetic variation and are more phenotypically diverse compared to non-African populations[267,270,298,299]. Therefore, the distribution of genetic risk factors and contribution within Africa and among other populations globally may differ. In 2009, a GWAS investigating the susceptibility to Malaria in The Gambia provided invaluable insights into the challenges of conducting genetic association studies in African populations. Some of these challenges include, shorter haplotype blocks (i.e. lower correlation between markers) in African populations which may lead to a loss of power to detect genome-wide significant association signals particularly when combining data from multiple ancestries; and lower LD between alleles at neighbouring loci which result in lower coverage and weaker

'tagging' of causal variants and thus dense genotyping arrays were required to boost statistical power for loci detection[300]. However, weaker LD patterns can also be advantageous in African populations as they allow for better resolution and refinement of association signals[299-301]. In addition, the high level of genomic diversity in African populations also means that population structure and genetic relatedness (overt or cryptic) need to be effectively accounted for as they can lead to false positive associations[299].

Other studies have also reported differences in environmental factors, heterogeneity in allele frequencies and effect sizes across different populations, thus loci identified in European or other populations may not replicate in African populations and vice-versa. Nonetheless, performing GWAS in African populations will benefit from uncovering functionally relevant, African-specific loci. For example, a very recent GWAS in 5,000 Kenyan children identified a low frequency, African-specific SNP (i.e. monomorphic in non-African populations) in the long intergenic non-coding RNA (lincRNA) gene *AC011288.2* that is associated with a 2-fold risk of pneumococcal bacteraemia ($p=1.69x10^{-9}$) in homozygous carrier of the risk allele; they also further observed its expression in neutrophils which are found to influence pneumococcal clearance[302].

These findings show that exploring host genetic variation is essential to enhance our understanding of factors that underlie infectious disease outcomes, particularly in African populations that bear a great burden of infectious diseases. The influence of host genetics on EBV infection and KSHV infection has not been very well studied and will be described in detail in chapter 3 and chapter 4 respectively.

As current genotyping platforms and imputation reference panels are skewed toward European ancestry populations, GWAS of African populations using these platforms will fail to capture up to 40% of genomic variation[303]. Thus, initiatives such as Human Heredity and Health in Africa (H3Africa) which seeks to build genomic research capacity in Africa to overcome these limitations and the African Genome Variation Project (AGVP) with dense genotype data from 1481 individuals representing 18

ethno-linguistic groups and whole genome sequence data for 320 individuals from sub-Saharan Africa provide a resource to capture genomic variation in Africa and improve the understanding of the genetic landscape and genetic architecture of traits in African populations[299,304].

1.5 Thesis Aims

In this thesis, I describe four distinct chapters that aim to shed light on how the genetics of host-virus interactions influences pathogenesis, in particular, the contribution of host genetic variation to EBV and KSHV infections in a rural African population cohort - the Ugandan General Population Cohort (GPC). As the projects have different aims (briefly outlined below) more specific introduction and discussion are contained within each chapter.

1. The aim of chapter 2 is to characterise the prevalence of infection, the inter-individual variability of serological phenotypes, and the genetic diversity of individuals in the Ugandan GPC.

2. The aim of chapter 3 is to explore the influence of host genetic variation on EBV Immunoglobulin G (IgG) antibody levels as a proxy for infection and potential disease risk using a GWAS approach.

3. The aim of chapter 4 is to explore the influence of host genetic variation on KSHV IgG antibody levels as a proxy for infection and potential disease risk, using a GWAS approach.

4. The aim of chapter 5 is to characterise the genetic diversity of KSHV whole-genomes isolated from asymptomatic individuals and assess the virus genetic population structure in a global context.

2 Chapter 2: The General Population Cohort, a Platform to Study the Genetic Architecture of Host Response to Gamma-Herpesvirus Infections

2.1 Introduction

Infectious agents are estimated to cause ~2 million new cases of cancer each year which is ~16.1% of the total number of cancer cases as diagnosed in 2008[143] of which 80% (1.6 million) occurs in less developed countries (Table.2.1). Sub-Saharan Africa bears the highest burden of cancers due to infectious agents (32.7%), Hepatitis B virus (HBV) and Hepatitis C virus (HCV) are associated with 84% of liver cancers, EBV is associated with >95% of Burkitt's Lymphoma (BL) and KSHV is associated with 100% cases of Kaposi's Sarcoma (KS)[32,305]. New incident cases of cancer attributed to these viruses are also much higher in this part of the world compared to more developed regions[32,305,306]. While EBV is nearly ubiquitous globally, the prevalence of and deaths caused by associated malignancies, display extensive geographic variation. BL is endemic in sub-Saharan Africa, Nasopharyngeal Carcinoma (NPC) is endemic in East Asia and Hodgkin's Lymphoma (HL) has a higher prevalence in Western Europe[307,308]. Unlike EBV, KSHV is not ubiquitous and its associated malignancies also display striking geographic variability[151,153]; in sub-Saharan Africa prevalence is 35%-60% or higher[181], in the Mediterranean region it is ~10-30%[309-311] and prevalence is lowest in Europe and North America (<10%), with higher prevalence occurring in homosexual men[148,312]. Other infectious causes of cancer are: Human papillomavirus which is associated with cervical cancer, Helicobacter pylori which is associated with gastric cancer, human T-cell lymphotropic virus which is associated with non-HL and the parasites *Opisthorchis viverrini* and *Clonorchis sinensis* which are also associated with liver cancers[32].

Table 2.1 Number of new cancer cases in 2008, stratified by infection and region*

	Less developed regions	More developed regions	World
Hepatitis B and C viruses	520 000 (32·0%)	80 000 (19·4%)	600 000 (29·5%)
Human papillomavirus	490 000 (30·2%)	120 000 (29·2%)	610 000 (30·0%)
Helicobacter pylori	470 000 (28·9%)	190 000 (46·2%)	660 000 (32·5%)
Epstein-Barr virus	96 000 (5·9%)	16 000 (3·9%)	110 000 (5·4%)
Human herpes virus type 8	39 000 (2·4%)	4100 (1·0%)	43 000 (2·1%)
Human T-cell lymphotropic virus type 1	660 (0·0%)	1500 (0·4%)	2100 (0·1%)
Opisthorchis viverrini and *Clonorchis sinensis*	2000 (0·1%)	0 (0·0%)	2000 (0·1%)
Schistosoma haematobium	6000 (0·4%)	0 (0·0%)	6000 (0·3%)
Total	1 600 000 (100·0%)	410 000 (100·0%)	2 000 000 (100·0%)

Data are number of new cancer cases attributed to a particular infectious agent (proportion of the total number of new cases attributed to infection that is attributable to a specific agent). *Numbers are rounded to two significant digits.

*All countries in Europe, North America as well as Australia, New Zealand and Japan were considered as more developed regions and all other countries were considered as less developed regions (according to UN definitions). From de Martel *et al,* 2012[32].

The inter and intra-continental heterogeneity in the distribution of infection as assessed by seroprevalence, and associated malignancies, while owing to a number of factors including differences in pathogen prevalence, and access to prophylactic/therapeutic measures, may also be influenced by other environmental factors, host genetics and gene-environment interactions in different populations. Many studies have reported that even though infection with oncogenic virus is necessary, it is insufficient to cause tumour development, suggesting other co factors are involved.

The immunosuppressive effect of HIV co-infection has been shown to promote tumourigenesis by oncogenic viruses[33]. HIV-1 co-infection is the most prominent co-factor for KSHV-associated malignancies[313-317]. In individuals who are dually infected with KSHV and HIV-1, the risk of developing KS has been reported to be significantly higher and disease is more aggressive than in HIV seronegative individuals[318-320]. Incidences of Primary effusion lymphoma (PEL) and Multicentric Castleman's disease (MCD) are also higher in HIV-1 seropositive co-infected individuals compared to HIV-1 seronegative individuals[321-326]. HIV can induce viral reactivation from latency either directly or indirectly via the production of inflammatory cytokines and chemokines allowing lytic replication which is important in KSHV transmission and

pathogenesis[315,327-329]. The major deregulation of both host innate and adaptive immune responses caused by sustained and uncontrolled HIV infection, can also create a favourable microenvironment for cellular proliferation and angiogenesis, playing an essential role in inducing KSHV-associated malignancies, such as KS[330-334]. In addition, active bidirectional talks between both viruses in the same environment has also been reported to stimulate reciprocal gene expression, worsening the prognosis for both infections[335-340].

Similarly to KSHV, HIV seropositivity is also associated with promoting EBV-associated lymphomas[38,321,341-344] with a 60-200 fold and 8-10 fold higher relative risk reported for developing non-HL and HL, respectively, and also a higher incidence and poorer prognosis of BL compared to HIV seronegative individuals[345-351]. Mechanisms of viral cooperation and immune modulation by HIV are similar to that described for KSHV-HIV co-infection. In the 1990's, the roll out of highly active anti-retroviral therapy (HAART) which suppresses HIV viral load and allows reconstitution of the immune system in HIV infected individuals, resulted in a decline in the incidence of AIDS-associated malignancies mainly in the developed world[200,352-354]. However, in resource limited settings such as sub-Saharan Africa the impact of treatment on outcomes and incidence is less clear and associated malignancies still remain a public health burden[355-357].

Synergistic interactions have also been reported for KSHV-EBV co-infection. While KSHV is necessary for developing PEL, EBV coinfection exists in ~70% of cases and some studies have reported that EBV-positive PEL cell lines are more tumourigenic than EBV-negative cell lines[358,359]. Both viruses use B-lymphocytes as a reservoir of infection and have been found to promote latency by subverting the host immune response and inhibiting lytic reactivation of each other in dually infected cells[202,360-364]. A number of studies have also found that co-infection with EBV and the Malaria parasite *Plasmodium falciparum* is associated with endemic BL[37,38,365-369]; and more recently *P. falciparum* has also been associated with KSHV seropositivity and increase in antibody responses in endemic regions[180,181,370]. HCV and HBV have also been

reported to be associated with EBV and KSHV infections and reported to trigger viral reactivation of the viruses indirectly by inducing chronic inflammation[371-374].

Viral factors associated with both viruses have been extensively studied, however, host genetic factors and their interactions with the environment leading to potential disease outcomes are largely unknown and require investigation, particularly in Africa. In Uganda, like in most sub-Saharan African countries, the leading causes of morbidity and mortality are attributed to infectious diseases, particularly HIV/AIDs and Malaria[375-377]. Uganda is a BL endemic region and has the highest seroprevalence of KSHV, and incidence of KS, compared to the rest of the world[365,378,379]. Hence, most studies have investigated the epidemiology of EBV and KSHV here[37,121,125,149,154,241,357,365,366,379-387]. Why this population sustains a higher prevalence of BL, KSHV and associated tumours, even after the roll out of HAART compared to the rest of the world still remains unknown.

In 1989, the General Population Cohort (GPC), a rural population-based cohort was set up by the Medical Research Council (UK) In collaboration with the Uganda Virus Research Institute (UVRI) to investigate trends in HIV infection prevalence and incidence and their determinants[377,388,389]. The study area is located ~120 Km from Entebbe town, in the Kyamulibwa sub-county of the Kalungu district in south-western Uganda with one half of the sub county lying ~40 Km from the shores of Lake Victoria[389,390]. The area comprises 25 neighbouring villages defined by administrative boundaries (Fig. 2.1) with a few concentrated in small trading centres. The villages consist of ~22,000 residents in total, ranging in size from 300-1500 residents, including families living within households[388]. The residents are mostly peasant farmers who grow bananas as a subsistence crop, cultivate coffee for trade and also raise livestock. Uganda has a diversity of ethno-linguistic (i.e. tribal) groups as result of a long history of migration and mixing between populations over the last 150 years (and also more recently) from surrounding regions and within the country influenced by factors including civil conflict within those regions and also attracted by labour and other economic incentives[389,391]. The Baganda are the predominant ethno-linguistic group constituting ~70% of the population, and a substantial number of migrants

who settled from neighbouring Rwanda, Burundi and Tanzania and make up the Banyarwanda ethno-linguistic group. There at least nine other minor ethno-linguistic groups that make up the rest of the population[389].

The GPC is dynamic with new births, deaths and migration reported at each round of follow-up, and with less than half of the population under survey being ≥13 years of age. Data are collected through an annual census, with questionnaire data including details of sexual behaviour, medical and socio-demographic factors. Blood specimens are also obtained at each survey for serological testing. More recently, research activity has leveraged the GPC as a platform to investigate the epidemiology and genetics of non-communicable diseases including cancer, cardiovascular disease and diabetes[389]; in addition to infectious disease traits. With the wealth of data available and abundance of circulating pathogens in Uganda, the GPC presents an opportunity to further investigate environmental and genetic factors associated with EBV, KSHV and other oncogenic viral infections.

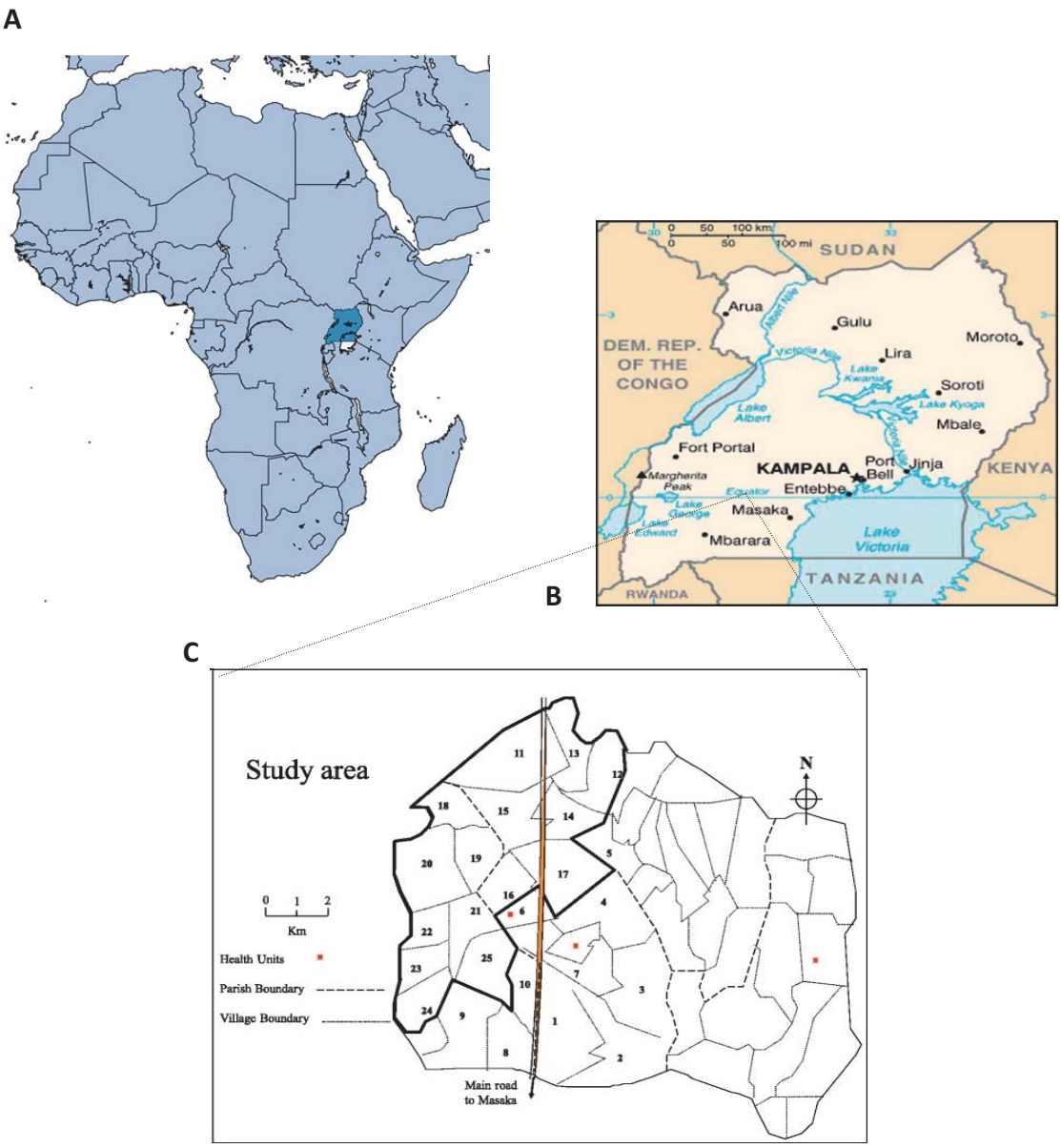


Fig. 2.1. Maps showing The GPC study area in context of Uganda and Africa. **A**. Map of Africa with Uganda shaded in dark blue. **B**. Map of Uganda and its bordering countries. **C**. The GPC study area encompassing 25 villages (labelled numerically) in the south-western region of Uganda. From Asiki et al, 2013[389].

2.1.1 Chapter Aims

The main aim of this chapter is to characterise the GPC in Uganda (the study population that will be used for my entire thesis) and assess its suitability to address the knowledge gap in the contribution of host genetics to EBV and KSHV infections. I use serological antibody response measures to infection, and genotype data of individuals in the cohort, to:

i. Describe the seroprevalence of EBV, KSHV and other oncogenic viruses circulating in this region and also the burden and influence of co-infection on antibody response levels.
ii. Investigate the genetic population structure of the Ugandan individuals in the cohort, in the context of Africa and the rest of the world using publically available datasets.
iii. Explore the heritable component of IgG response traits to EBV and KSHV.

Contributions

The GPC study team in Uganda coordinated sample collection and DNA extraction. Denise Whitby and Rachel Bagni's groups at the Frederick National Laboratory for Cancer Research (FNLCR) conducted serology and ascertained serostatus for all infectious disease traits investigated here. The Wellcome Trust Sanger Institute (WTSI) sequencing pipelines conducted genotyping and whole genome sequencing. The Global Health and populations team led by Manj Sandhu at WTSI performed curation of the Ugandan human genetic data including: Variant calling, SNP and sample quality control (QC), estimation of relatedness in individuals and haplotype phasing. All other analyses unless stated were performed by myself.

2.2 Methods

2.2.1 Sample Collection

Blood samples from 7000 GPC study participants, representing 11 self-reported ethno-linguistic groups, were collected during medical survey sampling rounds conducted in the study area between 1998-2011, as described previously[389]. Details of sexual behaviour, medical, socio-demographic and geographic factors were also recorded. Serum was tested for HIV-1 and the remainder was stored at -80 degrees Celsius in freezers in Entebbe prior to further serological testing.

2.2.2 Ethics

Informed consent in conjunction with parental/guardian consent for under 18 year olds was obtained from participants either with signature or a thumb print if the individual was unable to write. The GPC study was approved by the MRC/UVRI, Research Ethics committee (UVRI-REC) (Ref. GC/127/10/10/25), the Uganda National Council for Science and Technology (UNCST), and the UK National Research Ethics Service, Research Ethics Committee (UK NRES REC) (Ref. 11/H0305/5).

2.2.3 Serology and Quality Control of Phenotypic Data

As part of a larger investigation of oncogenic infections in the GPC, antibodies against EBV, KSHV, HBV and HCV were measured from a cross-sectional sample of people at three time points between 1991 and 2008, and an additional subset of samples were collected and assayed for KSHV in 2011. The sample was age and sex stratified to provide a 1:1 sex ratio and to increase the proportion of participants >15 years old. Of the original 7000 people sampled, 1570 had phenotype data for EBV (mean age ± SD = 34 ± 19.6 years, 54% female) and 4900 had data for analyses of KSHV traits (mean age ± SD = 34 ± 19.6 years, 58% female). Table 2.1 shows the characteristics of study participants in the GPC from samples collected in round 3, 11, 19 (1990-2008) and round 22 (2010/11).

EBV

2187 of blood samples collected from the GPC during sampling rounds 3 (1991/92), 11(1999/00) and 19(2007/08) were assayed for IgG antibody responses, EBNA-1, VCA and EAD antigens using a multiplex flow immunoassay on the Luminex® platform based on glutathione-S-transferase (GST) fusion capture immunosorbent assays combined with fluorescent bead technology as previously described[392]. The mean fluorescence intensity (MFI) across all beads was computed for each sample, and recorded after subtracting the background fluorescence. MFI cut-offs for seropositivity for each plate were defined as the average of the negative controls. Seropositivity was determined based on the presence of detectable IgG MFI > 519, >165 and >117 cutoffs for EBNA-1, VCA or EAD respectively. After removal of duplicate sample ID's, selecting for the most recent sampling round, 1570 unique individuals from round 3, 11 and 19 were available for the analyses of EBV antibody response traits.

Hepatitis B and C

2187 blood samples collected from the GPC during sampling rounds 3 (1991/92), 11(1999/00) and 19(2007/08) and additional 4437 collected in round 22 (2010/11) were assayed for antibody responses against Hepatitis B, HepB core antigen (HBcAG) and HepB surface antigen (HBsAG) and Hepatitis C core antigens (PepC1 and PepC2) and structural antigens (NS4 and NS5) using a multiplex flow immunoassay as described above for EBV antibody responses. Seropositivity was defined as being seropositive to all antigens tested for each virus. The criteria used to categorize specimens as seropositive were based on conventional antibody profiles used by the FNCLR and described in the literature[104,369,393-396].

KSHV

2187 blood samples collected from the GPC during sampling rounds 3 (1991/92), 11 (1999/00) and 19 (2007/08) and additional 4437 collected in round 22 (2010/11) were assayed for IgG antibody responses against LANA (ORF73) and K8.1 antigens using enzyme-linked immunosorbent assay (ELISA) based on recombinant proteins as previously described[393]. OD cutoffs for seropositivity for each plate were defined as

the average of negative controls plus 0.75 for the K8.1 ELISA and the average of the negative controls plus 0.35 for the LANA ELISA, to account for plate-to-plate variability. After removal of duplicate sample ID's, selecting for the most recent sampling round 4900 unique individuals across all rounds (3, 11, 19 and 22) were available for KSHV analyses.

Table 2.1 Characteristics of individuals in the GPC

		EBV Analyses		KSHV Analyses	
Characteristic		**N=1570**	**(%)**	**N=4900**	**(%)**
Sex	Male	725	46.2	2082	42.5
	Female	845	53.8	2818	57.5
Age Group	<15	335	21.3	76	1.6
	15-24	293	18.7	1902	38.8
	25-44	466	29.7	1600	32.6
	>44	474	30.2	1322	26.8
EBV	Positive	1473	93.8	N.T	N.T
	Negative	97	6.2	N.T	N.T
KSHV	Positive	1449	92.3	4466	91.0
	Negative	121	7.7	434	9.0
HIV	Positive	105	6.7	332	6.7
	Negative	1465	93.3	4566	93.3
HBV	Positive	143	9.1	287	6.0
	Negative	1427	90.9	4613	94.0
HCV	Positive	94	6.0	266	5.5
	Negative	1476	94.0	4634	94.5
Sampling Round (Year)	3 (1991/92)	193	12.3	71	1.4
	11 (1999/00)	388	24.7	115	2.4
	19 (2007/08)	989	63.0	277	5.7
	22 (2010/11)	-	-	4437	90.5
Human Genetic Data*	Genotype	949	60.4	3461	70.6

N= The number of unique individuals
N.T= Not tested
*Generated from samples collected in Round 22 in 2010/11 and described below.

2.2.4 Statistical Analysis of Quantitative Antibody Traits

To investigate factors influencing IgG antibody response levels to EBV and KSHV infections, residuals were obtained following multi-variate linear regression of EBV and KSHV IgG antibody traits on age, sex, sampling round, EBV/KSHV, HIV, HBV and HCV infection serostatus (treated as binary variables) using R statistical package[397]. To ensure normalisation of MFI and OD values for analyses, I performed a log transformation of IgG antibody traits in R.

2.2.5 SNP Genotyping and Quality Control

Of the 7000 samples, 5000 samples collected in 2011 (round 22) were densely genotyped on the Illumina HumanOmni 2.5M BeadChip array (UGWAS). A total of 2,314,174 autosomal and 55,208 X-chromosome markers were genotyped on the HumanOmni2.5-8 chip. Of these, 39,368 autosomal markers were excluded because they did not pass the quality thresholds for the SNP called proportion (<97%, 25,037 SNPs) and Hardy Weinberg Equilibrium (HWE) ($p<10\text{-}8$, 14,331 SNPs). HWE testing was only carried out on the founders for autosomes, and female unrelated individuals for the X chromosome defined by an Identity by descent (IBD) threshold <0.10 as estimated by PLINK[398]. An additional 91 samples were excluded during sample QC as they did not pass the quality thresholds for proportion of samples called (>97%) or heterozygosity (outliers: mean±3SD), or the gender inferred from the X-chromosome data was discordant with the supplied gender. Three additional samples were dropped because of high relatedness, IBD >0.90. 2,230,258 autosomal markers and 4,778 samples (Table 2.2) remained following SNP and sample QC respectively for downstream analyses.

Table 2.2. Distribution of Ethno-linguistic Groups Genotyped* in the GPC

Ethno-linguistic Group	No. of Samples
Baganda	3585
Banyarwanda	422
Rwandese Ugandan	202
Batanzania	44
Bakiga	60
Banyankole	147
Barundi	191
Batooro	10
Basoga	16
Bafumbira	5
Other Ugandan	45
Unknown	51
Total	4778

*This represents the samples that passed QC that were used for downstream genetic analyses

2.2.6 Principal Components Analysis

To explore the population structure of ethno-linguistic groups in The GPC (Table 2.2), I performed principal components analyses (PCA) using SMARTPCA in Eigensoft v4.2[399]. I also performed PCA in 1,753 unrelated individuals to contextualize the GPC in Africa with African Genome Variation Project (AGVP)[299] populations (Fig. 2.2) and in a global context including 1000 Genomes phase III[270] populations as a reference panel (

Table ***2.3***). PCA was done including markers with MAF>1% after LD pruning to a pairwise threshold of r^2=0.5 using a sliding window approach with a window size of 200kb, sliding 5 SNPs sequentially.

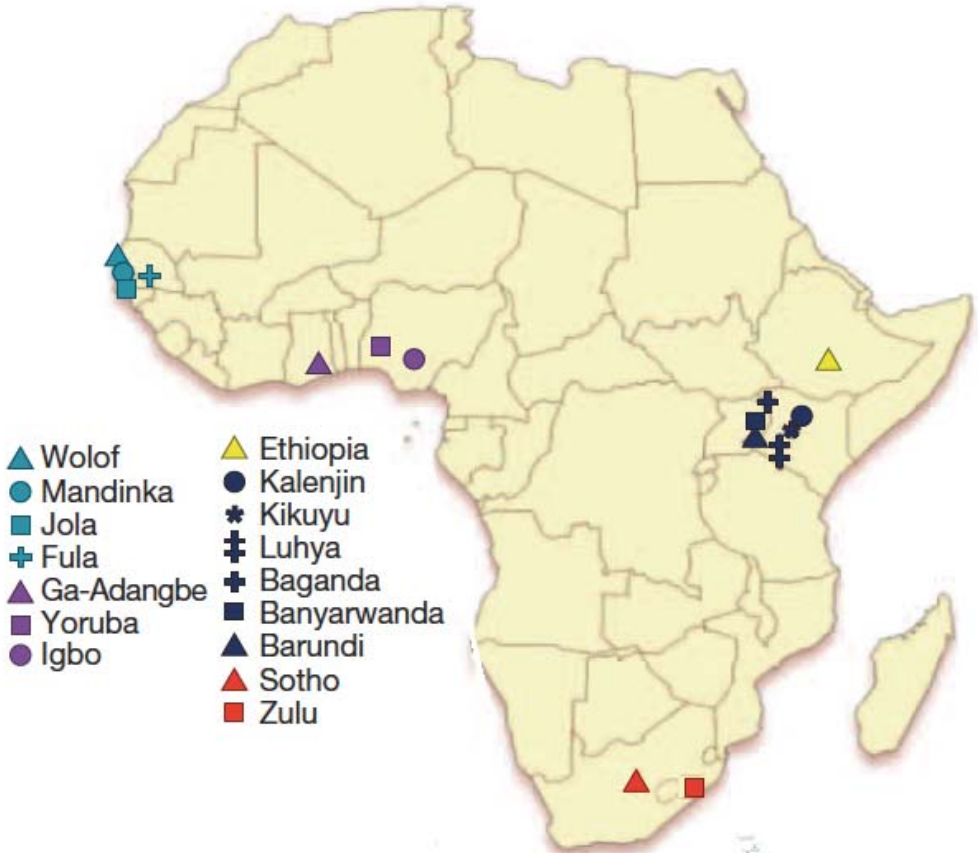


Fig. 2.2 Map of Africa showing the location of samples in the AGVP used for PCA. Representing ethno-linguistic groups from The Gambia, Ghana, Nigeria, Ethiopia, Kenya, Uganda and South Africa. Adapted from Gurdasani et al, 2014[299].

Table 2.3. Distribution of samples included in Principal Components Analysis

Ancestry	Population	Population Group	N
East African	Uganda	GPC	1753
East African	Uganda	Baganda	45
		Banyarwanda	75
		Barundi	26
	Kenya	Kikuyu	99
		Kalenjin	100
		LWK* - Luyha	74
	Ethiopia	Amhara	42
		Oromo	26
		Somali	39
South African	South Africa	Sotho	86
		Zulu	100
West African	Nigeria	Igbo	99
		YRI* - Yoruba from Ibadan	100
	Ghana	Ga-adangbe	100
	Gambia	Fula	74
		Wolof	78
		Jola	79
		Mandinka	88
African (AFR)	USA (Southwest)	ASW - Americans of African Ancestry	49
	Barbados	ACB - African Caribbeans	72
European (EUR)	England & Scotland	GBR - British	91
	Finland	FIN - Finnish	97
	Spain	IBS – Iberian population	99
	Italy	TSI – Toscani	92
	USA	CEU – Utah Residents (CEPH)	95
Admixed American (AMR)	USA	MXL - Mexican Ancestry, Los Angeles	47
	Colombia	CLM – Colombians from Medellin	65
	Peru	PEL – Peruvians from Lima	50
	Puerto Rico	PUR – Puerto Ricans	72
East Asian (EAS)	China	CDX - Chinese Dai in Xishuangbanna	83
South Asian (SAS)	China	CHB – Han Chinese in Beijing	98
	China	CHS – Southern Han Chinese	86
	Japan	JPT – Japanese in Tokyo	96
	Vietnam	KHV - Kinh in Ho Chi Minh City	96
	Texas	GIH – Gujarati Indian, Houston	95
		Total	4466

Uganda GPC unrelated individuals highlighted in red. AVGP populations (N=1330) highlighted in Yellow. 1000 Genomes Phase III (N=1383) highlighted in Green.

* Also part of 1000 Genomes phase III populations

2.2.7 Heritability of Antibody Response Traits in The GPC

I estimated narrow-sense heritability (h^2), which represents the proportion of phenotypic variation attributed to additive genetic variation, with FaST-LMM, using a linear mixed model (LMM) with two random effects, one based on genetic effects and the other on environmental effects[400]. Spatial location recorded as Global Position System (GPS) coordinates collected during sampling rounds was used as a proxy for environmental effects. Prior to the analyses KING (http://people.virginia.edu/~wc9c/KING/) was run to create an accurate set of pedigrees from the genotype data and remove highly related individuals. Haplotypes were phased with SHAPEIT2[401] and an Identity-by-descent (IBD) matrix was generated using methods previously described[402]. Heritability was calculated as the proportion of phenotypic variance explained by the IBD matrix. Gene-environment interactions were also explored as detailed in Heckerman *et al*, 2016 using Fast-LMM[400].

2.3 Results

2.3.1 Seroprevalence of Infectious Traits in The GPC

To investigate seroprevalence of infections, serological measures from 1570 individuals in the GPC taken during rounds 3, 11 and 19 were examined for 5 viruses: KSHV, EBV, HBV, HCV and HIV. I assessed the serological evidence of exposure based on seroreactivity to the antigens for each virus, showing 1536 individuals (~99%) are infected with at least 1 virus. EBV and KSHV are both nearly ubiquitous in this population and have the highest seroprevalence at >90% (Fig. 2.3). Chronic HBV seropositivity is 9.1% and chronic HCV infection has the lowest seroprevalence among the pathogens examined at 6% (Fig. 2.3). HIV infection seroprevalence in this study is 6.7%.

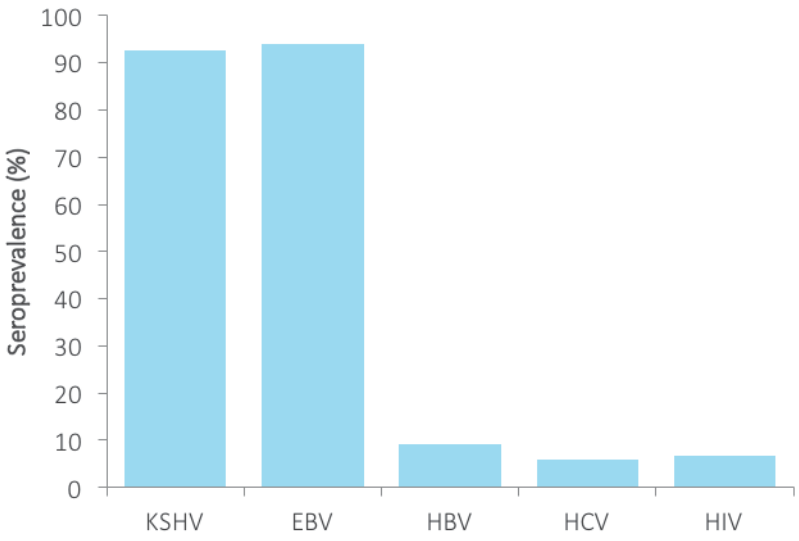


Fig. 2.3 Seroprevalence of viral infections tested in the GPC between 2008-2011

To investigate the pathogen burden in study participants, I calculated the number of infections participants were seropositive for (Fig. 2.4) and also assessed the seroprevalence of co-infection (Table 2.4). The majority of participants, 1052 (67%) were seropositive to at least 2 of the viruses tested, only 4 (0.25%) participants were seropositive for all 5 viruses and 14 (0.89%) participants were seronegative for all

viruses (Fig. 2.4). Co-infection was highest for EBV and KSHV with 93% of individuals seropositive to antigens for both pathogens (Table 2.4). Co-infections of other pathogens with EBV or KSHV was similar and mirrors the seroprevalence estimates seen in the cohort (Fig. 2.3 and Table 2.4).

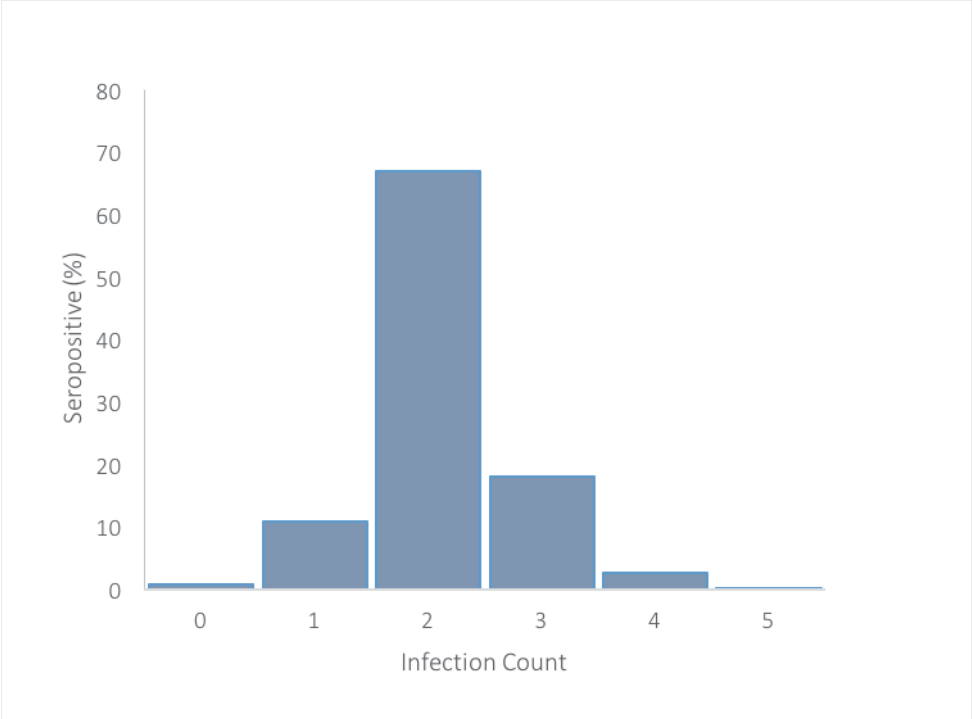


Fig. 2.4 The number of seropositive reactions to viruses for all participants in The GPC between 2008-2011. The infection count represents the minimum number of infections participants are seropositive for

Table 2.4 Seroprevalence of co-infection with EBV or KSHV

Infection	EBV (N=1473)	KSHV (N=1449)
EBV	-	1352 (93%)
KSHV	1352 (92%)	-
HIV	91 (6.2%)	91 (6.2%)
HBV	139 (9.4%)	133 (9.4%)
HCV	94 (6.3%)	93 (6.4%)

N represents the number of seropositive individuals

2.3.2 Inter-Individual Variation in IgG Antibody Responses to EBV and KSHV Infections

To investigate the inter-individual variation of IgG antibody responses to EBV and KSHV infections, seropositivity to the latent and lytic antigens were measured and distributions of antibody levels were determined.

In 1570 individuals tested for EBV antibodies, anti-VCA IgG, anti-EBNA-1 IgG and anti-EAD IgG seropositivity was 82%, 92% and 28% respectively. All antibody response measures were highly variable across individuals as shown by the wide MFI range and displayed a skew to left (Fig. 2.5 and Fig. 2.6). MFI Values for anti-VCA IgG ranged from 15 to 11321 (mean ± S.D = 1834.5 ± 2035.6) (Fig. 2.5.A) and all individuals with MFI >165 were considered seropositive for VCA. MFI values for anti-EBNA-1 IgG displayed a similar range to anti-VCA IgG albeit higher, ranging from 17 to 19794 (mean ± SD = 3164.2 ± 3360.3) (Fig. 2.5.B) and all individuals with MFI >519 were considered seropositive for EBNA-1. Anti-EAD IgG response ranged from 15.5 to 5618.5 (mean ± S.D = 196.8 ± 369.6) (Fig. 2.6) and all individuals with MFI >117 were considered seropositive for EAD.

In 4930 individuals tested for KSHV antibodies from rounds 3, 11, 19 and 22, LANA and K8.1 seropositivity was 91% and 96% respectively. KSHV ELISAs captured OD ranging from 0 to 4 for LANA and K8.1. Whilst anti-LANA IgG responses displayed a U-shaped distribution, anti-K8.1 IgG responses displayed a skew to the right (Fig. 2.7. A). LANA OD had a mean ± S.D = 1.83 ± 1.22 (Fig. 2.7.A) and K8.1 mean ± S.D = 2.41 ± 1.05 (Fig. 2.7.B).

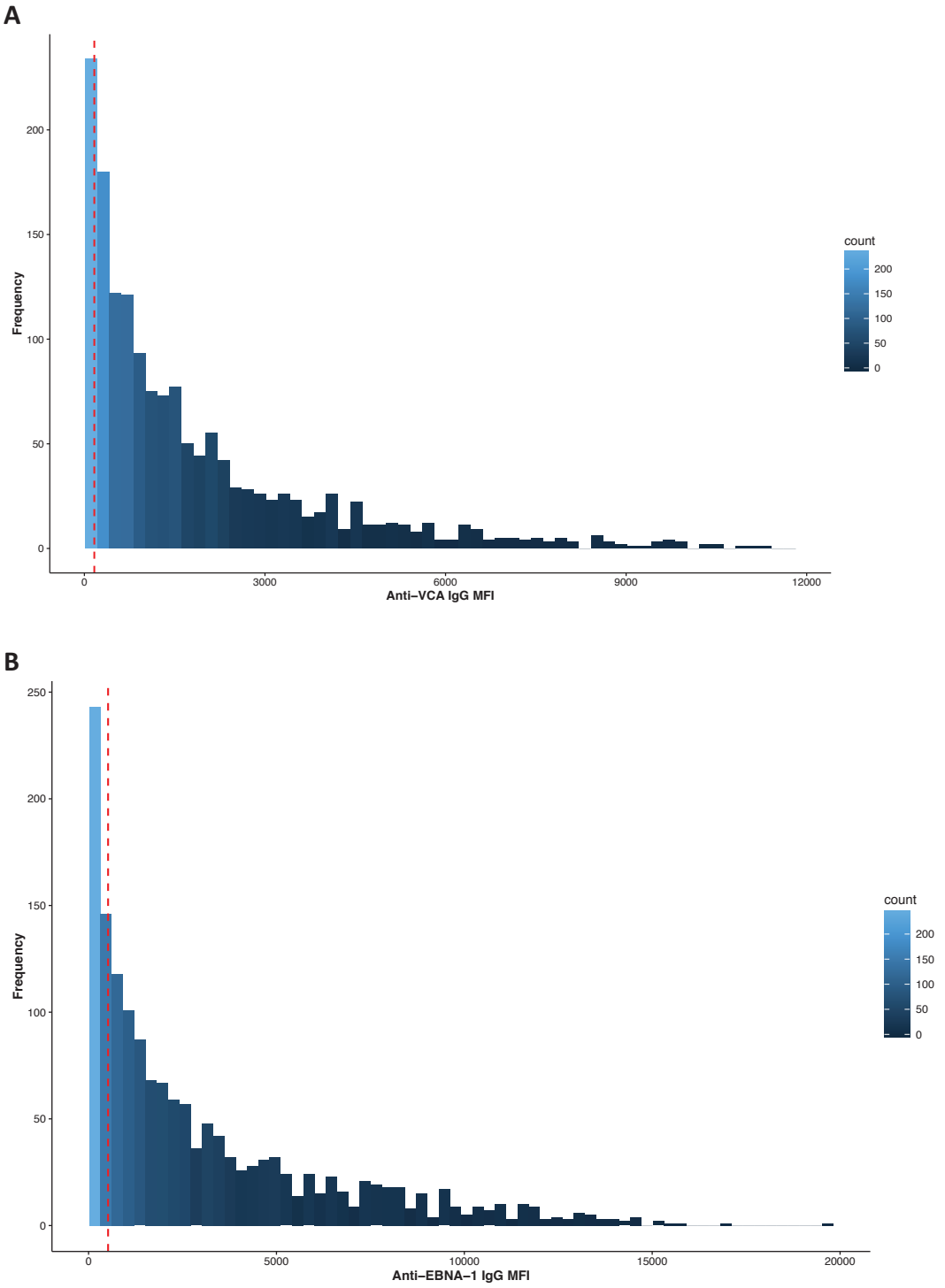


Fig. 2.5 Inter-individual variability in IgG antibody responses to EBV. **A.** Distribution of anti-VCA IgG mean fluorescence intensity (MFI). Red dotted line represents MFI cut-off=165. Seropositive=1350, Seronegative=217. **B.** Distribution of anti-EBNA-1 IgG MFI. Red dotted line represents MFI cut-off =519. Seropositive=1206, Seronegative=361

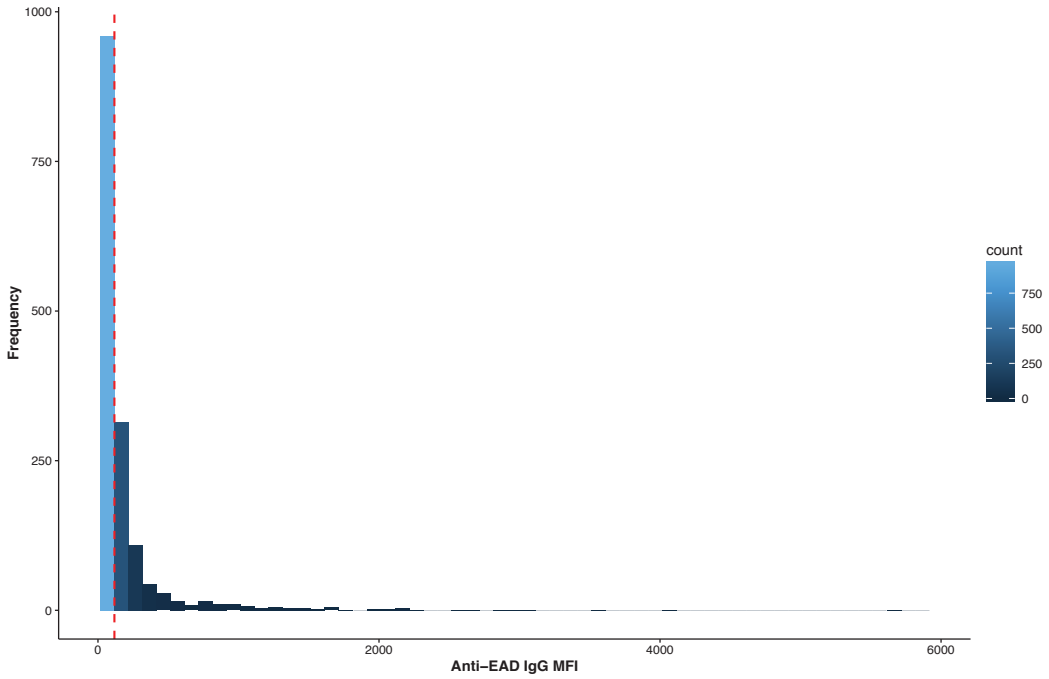


Fig. 2.6 Distribution of anti-EAD IgG Mean Fluorescence Intensity (MFI). Red dotted line represents MFI cut-off =117. Seropositive=406, Seronegative=1170

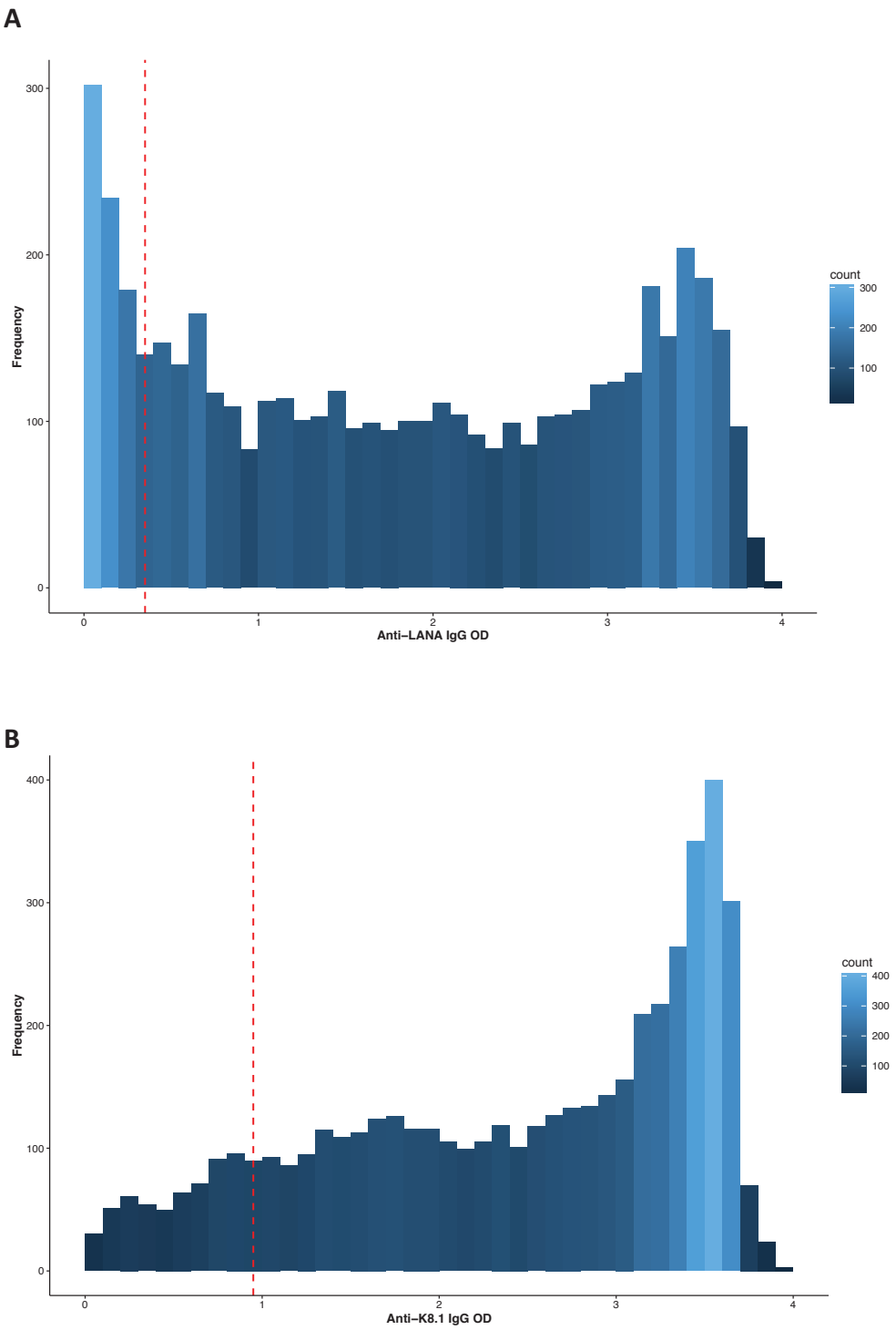


Fig. 2.7 Inter-individual variability in IgG antibody responses to KSHV.

A. Distribution of anti-LANA IgG optical density (OD). Red dotted line represents mean OD cutoff =0.35. Seropositive=4100, Seronegative=830. B. Distribution of anti-K8.1 OD. Red dotted line represents OD cutoff =0.95. Seropositive=4194, Seronegative=706

2.3.3 Predictors of IgG response levels to EBV and KSHV infection

I then investigated the factors that could potentially influence the variation in IgG antibody levels to EBV and KSHV infection in 1570 individuals from round 3, 11 and 19 with data on age, sex, sampling round, seropositivity to EBV, KSHV, HIV, HBV and HCV using a multi-variate linear regression model in R (Table 2.5).

For EBV (Table 2.5), no significant differences in IgG levels were observed between sexes for all traits. Increase in age had a significantly lowering effect on anti-EBNA-1 IgG levels which was not observed for other traits. More recent sampling rounds had significantly higher responses to anti-EBNA-1 and anti-VCA IgG levels, showing year of sample collection had an effect on antibody response. HIV seropositivity significantly lowered responses to anti-VCA IgG, whereas KSHV seropositivity resulted in significantly higher responses to anti-VCA and anti-EAD IgG. HBV and HCV seropositivity also resulted in significantly higher responses to all EBV IgG traits.

For KSHV (Table 2.5) IgG antibody traits, increase in age significantly increased IgG levels for both LANA and K8.1, while OD values were significantly higher in males than females for LANA. Significantly higher OD values for LANA and K8.1 were also observed for EBV and HCV seropositive individuals, whereas HIV seronegative individuals displayed higher IgG levels for K8.1.

Table 2.5 Covariates/Predictors of IgG response levels for EBV and KSHV infection

		Linear Regression Coefficients (p-value)							
	IgG	**Age**	**Sex[a]**	**Sampling round[b]**	**EBV Status[c]**	**KSHV Status[c]**	**HIV Status[c]**	**HBV Status[c]**	**HCV Status[c]**
EBV	**EBNA-1**	**-0.0088 ($4.4x10^{-6}$)**	-0.0083 (0.91)	**0.0196 (0.002)**	-	0.323 (0.019)	-0.036 (0.79)	**0.47 (0.000274)**	**0.59 (0.0002)**
	VCA	0.002 (0.18)	0.10 (0.119)	**0.02 ($2.67 x10^{-6}$)**	-	**0.479 (0.00016)**	**-0.68 ($3.52 x10^{-7}$)**	**0.58 ($1.21x10^{-6}$)**	**0.5 (0.0005)**
	EAD	0.002 (0.08)	-0.04 (0.347)	-0.005 (0.16)	-	**0.47 ($8.37 x10^{-9}$)**	-0.16 (0.052)	**0.44 ($1.4x10^{-8}$)**	**0.57 ($2.63x10^{-6}$)**
KSHV	**LANA**	**0.014 ($9.24x10^{-15}$)**	**-0.15 (0.004)**	-0.003 (0.96)	**0.298 (0.003)**	-	-0.24 (0.025)	0.04 (0.56)	**0.55 (2.93E-06)**
	K8.1	**0.006 ($2.19 x10^{-6}$)**	-0.138 (0.01)	-0.017 (0.83)	**0.39 (0.00014)**	-	**-0.2966 (0.00674)**	0.093 (0.25)	**0.478 ($6.58 x10^{-5}$)**

All p-values in bold remain statistically significant after correcting for multiple testing using Bonferroni correction $p<0.007$.

[a] Positive regression coefficient relates to higher MFI/OD values in females than males.

[b] Positive regression coefficient relates to higher MFI/OD values in later sampling rounds.

[c] Positive regression coefficient relates to higher MFI/OD values in seropositive than seronegative individuals.

2.3.4 Genetic Population Structure in The GPC

To investigate the genetic diversity of individuals in the GPC, I used PCA to infer the axes of genetic variation and explore the population structure in the Ugandan GPC ethno-linguistic groups in the context of the AGVP populations, and globally, including 1000 Genomes phase 3 populations as a reference panel (Table 2.3). PCA ascertained homogeneity in the cohort with no clear separation observed in unrelated individuals from the different ethnolinguistic groups (Fig. 2.8). In the context of African populations in AGVP, the Ugandan GPC samples cluster with the other Ugandan populations of the AGVP, also there's separation based on geographic origin and a cline along PC1 (Fig. 2.9), as previously reported[299] (Gurdasani *et al*. 2016, in review). In a global context, the GPC samples clustered well with other African populations (Fig. 2.10) as expected with a clear cline towards European populations.

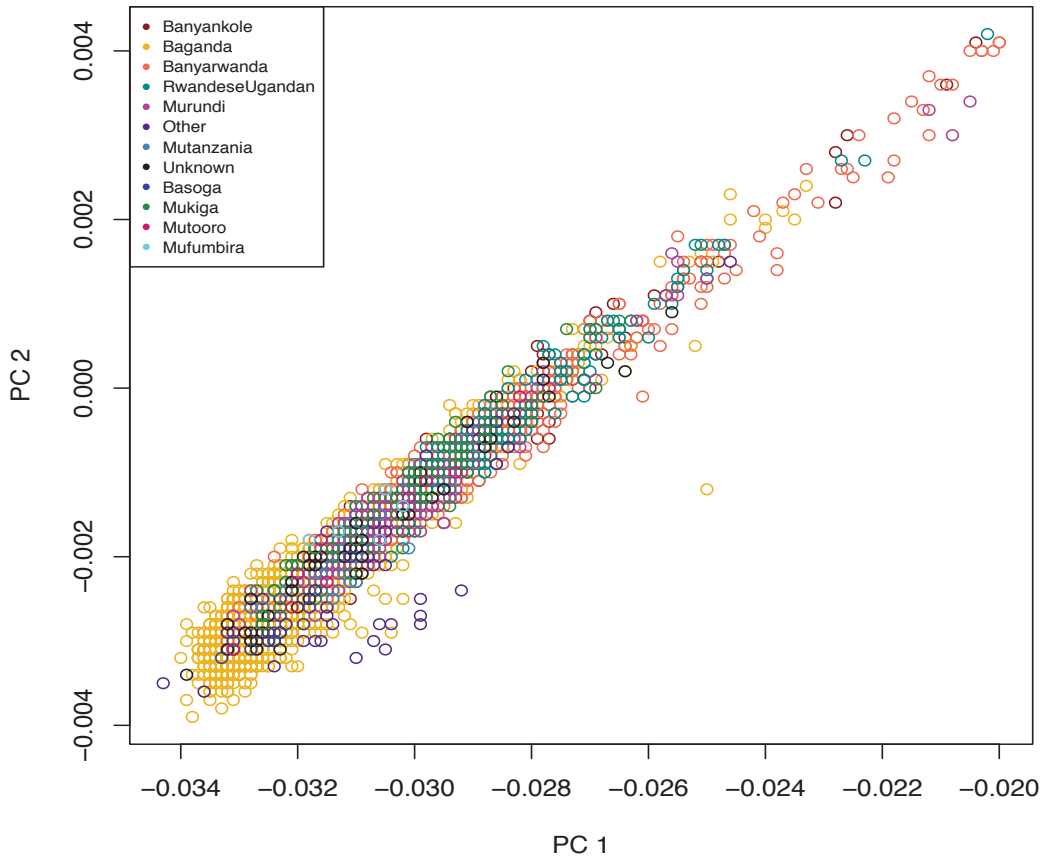


Fig. 2.8 Genetic population structure of individuals within the GPC ethnolinguistic groups. No clear separation observed for ethno-linguistic groups.

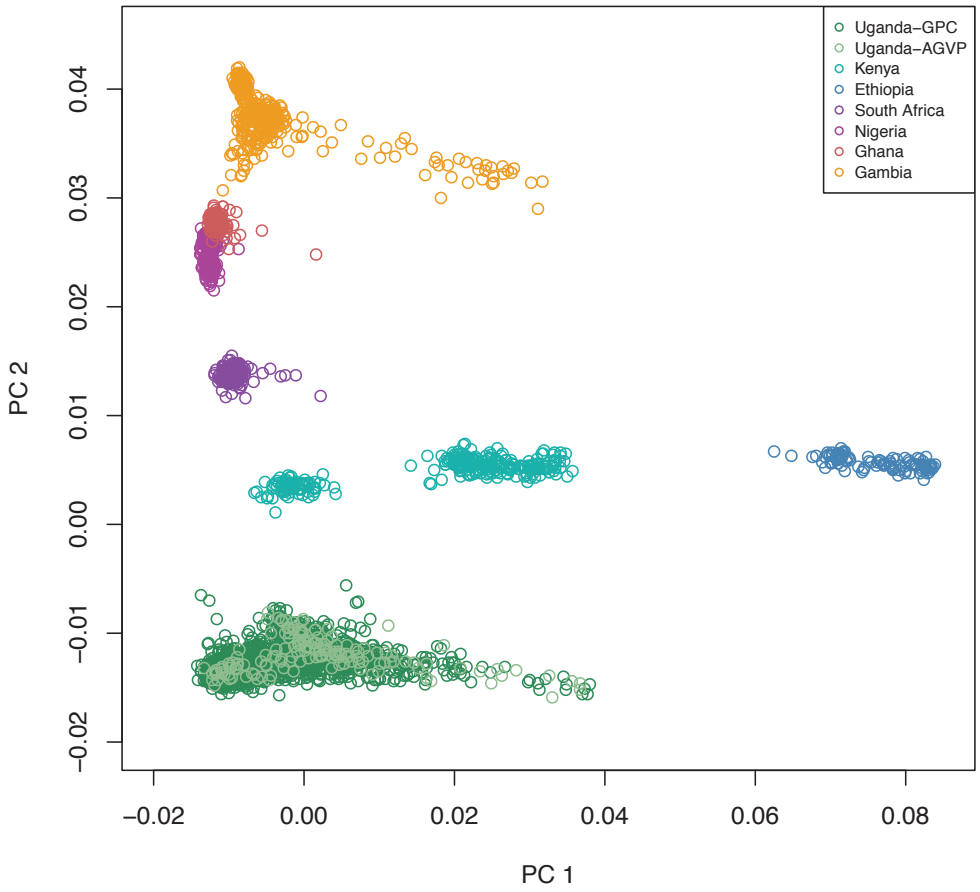


Fig. 2.9 Genetic population structure of the GPC in the context of AGVP African populations. PC1 shows cline seen among East and West Africans. PC2 shows separation by populations from different regions.

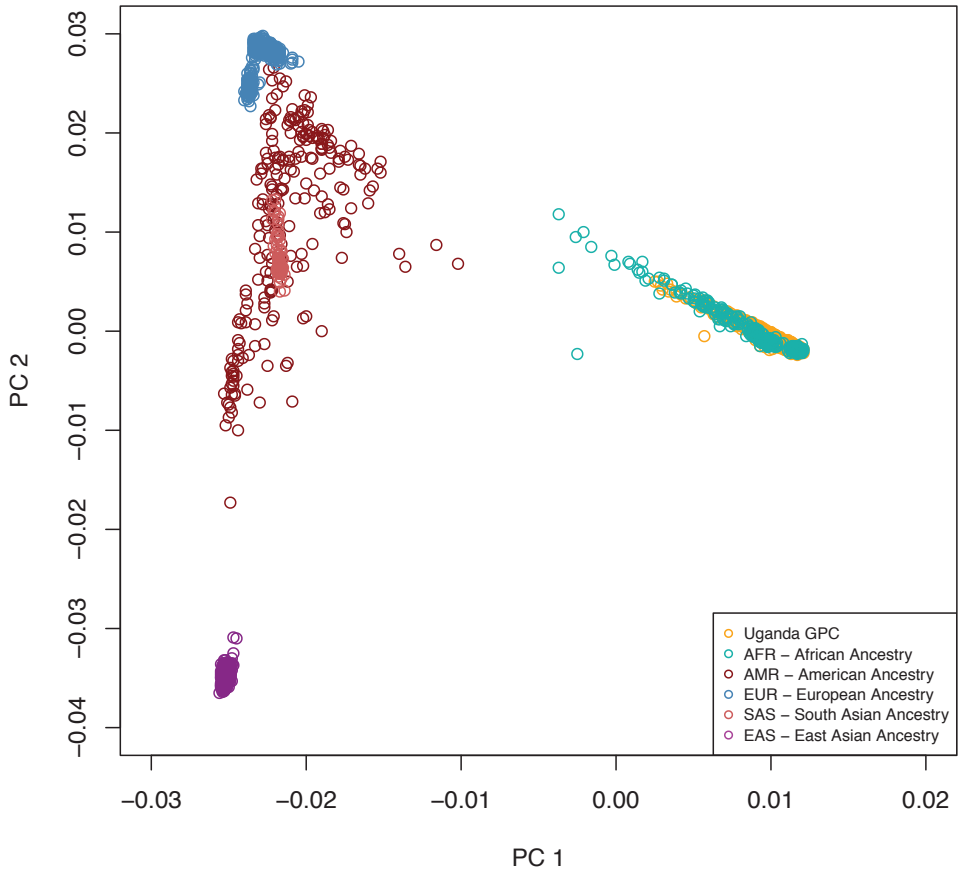


Fig. 2.10 Genetic population structure of GPC in the context of AGVP and global 1000G populations. PC1 shows a cline seen among Ugandan GPC and other African ancestry populations (include both AGVP and 1000 Genomes) extending towards Europeans.

2.3.5 Heritability of IgG Antibody Response Traits in The GPC

To investigate whether variation in IgG responses could be explained by host genetics i.e. whether they had a heritable component. I modelled genetic relatedness and estimated narrow-sense heritability (h^2) based on directly observed genotypes of individuals in the GPC, similar to other methods performed in recent years[403,404]. I also adjusted for environmental correlation (h^2adjusted) between individuals by using spatial distances based on GPS coordinates as a proxy for shared environment in the linear mixed model in FaST-LMM as detailed in Heckerman *et al,*. In this context and for the remainder of the thesis, I will refer to narrow-sense heritability as 'heritability'.

Estimates of heritability for IgG antibody response measures to EBV and KSHV were variable and attenuated after adjustment for environmental correlation (Fig. 2.11 and Table 2.6). For EBV, anti-EBNA-1, anti-VCA and anti-EAD IgG responses were heritable after adjusting for environmental effects, h^2adjusted= 11%, 7.7%, 14%, respectively. Antibody responses to KSHV anti-LANA IgG (h^2adjusted = 27%) and anti-K8.1 IgG (h^2adjusted = 25%) were also significantly heritable in this population. Lower sample sizes for EBV antibody responses (N=949) in comparison to KSHV antibody responses (N=3461) mirror larger standard errors (S.E), and estimates for gene-environmental interaction may not be as reliable with such small sample sizes, and thus, this analysis is exploratory.

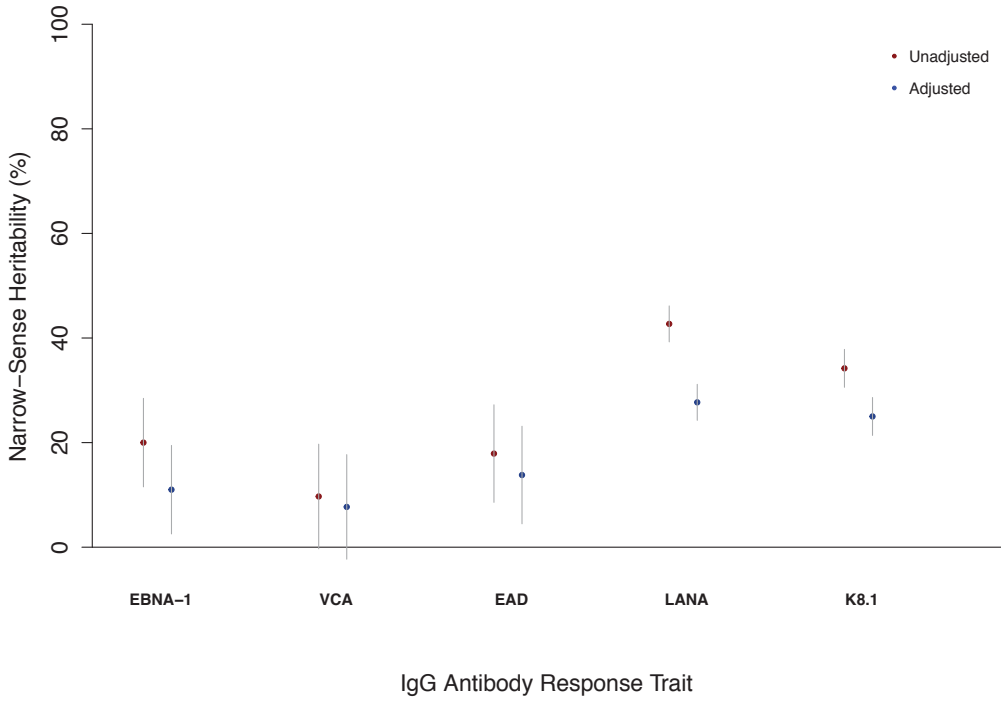


Fig. 2.11 Heritability of IgG antibody traits for EBV and KSHV infections. Fast-LMM narrow-sense heritability (h^2%) estimates with (blue) and without (red) adjustment for environmental effects accounted for by GPS coordinates. Error bars represent standard error (S.E).

Table 2.6 Heritability estimates of EBV and KSHV IgG antibody traits in the GPC

Phenotype			Heritability and environmental variance			Gene-environment interaction
Infection	**N**	**IgG**	**h^2Unadjusted(±SE)**	**h^2Adjusted (±SE)**	**e^2(±SE)**	***gxe*^{2}(P)**
EBV	949	EBNA-1	0.200 (0.09)	0.110 (0.09)	0.06 (0.1)	0.23 (0.13)
		VCA	0.097 (0.1)	0.077 (0.1)	0.02 (0.07)	0.21(0.23)
		EAD	0.180 (0.1)	0.140 (0.1)	0.03 (0.07)	-
KSHV	3461	LANA	0.430 (0.04)	0.270 (0.04)	0.12 (0)	-
		K8.1	0.340 (0.04)	0.250 (0.04)	0.05 (0.07)	0.07 (0.08)

N represents the number of individuals with genotype data seropositive for each infection
e^2 represents environmental effects, gxe^2 represents gene-environment interaction

2.4 Discussion

In this chapter, I assessed the seroprevalence of EBV, KSHV, HIV, HBV and HCV infections in the Ugandan GPC along with the burden of co-infections. I investigated the variability of serological IgG antibody response measures to EBV and KSHV and explored the factors influencing variation in IgG antibody levels. In addition, the availability of human genotype data on a subset of individuals allowed for the assessment of the suitability of the cohort for use in investigating the human genetic contribution to IgG antibody response traits by analysing the genetic population structure and heritability of IgG antibody responses.

Serological diagnoses of infections are useful in clinical management decisions, and improving our understanding of prevalence and transmission of infection, in addition to understanding virology and host immunity. The presence of antibodies representing host immune response are commonly used as diagnostic markers for current infection, history of infection or monitoring the outcome of vaccination against infection[104]. In addition, quantifying host antibody responses to viral antigens is particularly beneficial for infections whereby the virus is dormant i.e. in the latent stage of infection with minimal viral replication occurring preventing detection of viral nucleic acids by other methods. Therefore, seroprevalence is widely used as a measure of the frequency of infections in a population. The challenge however, in comparing seroprevalence estimates is the fact that assay designs and thresholds (used to categorise individuals as seropositive, seronegative or indeterminate) may vary between studies. Furthermore, while serology is useful for infection diagnosis it cannot be used to diagnose associated malignancies.

Seroprevalence estimates of EBV (94%), HBV (9%)[405], HCV (6%)[396,406] and HIV (6.7%) are consistent with previous findings in Uganda. However, KSHV (93%) is nearly ubiquitous, like EBV, and seroprevalence is higher than published estimates. While the majority of individuals are infected with KSHV by adulthood, heterogeneity in seropositivity rates from 34%-88% has been observed in a number of studies from different regions of the country[179,181,378,385,407]. It is also evident in this study that a

high number of individuals harbour multiple infections, and nearly everyone is dually infected with EBV and KSHV. Environmental factors such as co-infection with other pathogens have been studied and found to influence variation in seroprevalence estimates, and antibody responses associated with infection. This is consistent with the results presented in this study (Table 2.5) with co-infection with HCV and HBV influencing IgG antibody response levels to EBV EBNA-1 latent antigen, and co-infection with KSHV influencing lytic antigen VCA response to infection. For KSHV, co-infection with EBV and HCV influences antibody responses to both latent (LANA) and lytic (K8.1) antigens, whereas HIV co-infection only influences anti-K8.1 IgG response levels.

In the GPC, the burden of infection and co-infection is high as described above (and Fig. 2.3 and Table 2.4) and differences in environmental variation here compared to non-African populations likely contribute to the underlying phenotypic variance. In addition to environmental factors such as co-infection, host genetic factors (which might also be influenced by the environment) can also contribute to phenotypic variation in infectious disease traits. The host genetic contribution to infectious disease traits have been gaining interest in the recent years and a number of studies exist that have used candidate gene, linkage and genome-wide association approaches for investigation[276]. However, as most studies have focused on European populations, a paucity of data exists in African populations. In this thesis, I aimed to address this gap by using data from a population cohort in Uganda to undertake genetic analysis of EBV and KSHV antibody response traits. To undertake genetic association studies it is essential that any systematic differences between individuals is first assessed as this can influence results. This can include the local environment, in this setting, it is evident that co-infection with other pathogens such as HIV, HBV, HCV influence antibody response levels traits (Table 2.5 and described above) and thus, this environmental correlation represents a covariate which would need to be adjusted for in genetic analyses. In addition, population structure arises when large scale systematic differences exist in individuals, due to either differences in immigrant ancestry of individuals or more shared recent ancestry or relatedness in individuals than one would expect, which is usually in alignment with geographic

region. Population structure and close relatedness can confound association studies by leading to spurious associations or reducing power to detect true associations, by increasing statistical errors[408]. Analysis of the genetic population structure in The GPC shows homogeneity between ethno-linguistic groups (Fig. 2.8); the population structure is minimal, suggesting it might not be an issue for downstream genetic analyses. Analysis of the genetic population structure in this cohort shows homogeneity between ethno-linguistic groups which is advantageous in studying the host genetics of disease traits. However, given that recruitment of individuals in the GPC is done within households and the fact that families co-habit, and thus, as expected, the cohort has increased levels of cryptic relatedness that could confound downstream association analyses, this would need to accounted for to avoid false positive associations in genetic analyses.

To explore the proportion of variance in antibody responses attributable to the host genetics, I investigated their heritable component. Here, I estimated narrow-sense heritability (h^2), a commonly used metric in determining the genetic basis of complex disease traits as it represents the fraction of phenotypic variation attributable to additive genetic variation. This represents the extent to which an individual's phenotype is determined by their parents and estimates have been based on relatedness in families, particularly in twin studies[409]. For infectious disease traits most studies of heritability have focused on response to vaccination and/or clearance of infection with sparse information generated from African populations. Therefore, I explored the heritability of five IgG response traits to EBV and KSHV antigens in the GPC, in seropositive individuals with genotype data, adjusting for environmental correlation using spatial distances and similar methodology to previous studies[410]. These analyses show that while EBV and KSHV antibody response traits are heritable there is overestimation in heritability prior to accounting for shared environment. The heritability estimates for EBV anti-EBNA-1 and anti-VCA IgG traits in this cohort are much lower (~2-20%) in comparison to those reported in populations of Mexican-American and European descent with reported heritability at ~30-43% for IgG antibodies against EBNA-1 and 32-48% for VCA, respectively[121,411-413]. The heritability estimates for KSHV (~21-32%) are also lower than the reported estimate of 37% in a

Mexican-American ancestry cohort[414]. Differences in the heritability estimates in the GPC compared to other populations could be attributed to a number of factors including differences in study and assay design, locus or allelic heterogeneity influencing traits, not fully adjusting for environmental effects and differences in gene-environment interactions. The differences in heritability estimates for different serological traits might be due to underlying differences in genetic architecture for the infections. A limitation of this analysis, however, is the small sample size for some of the traits thus environmental variance estimates may be less accurate.

In summary, this study confirmed previous reports that EBV and KSHV infections are highly prevalent in Uganda, and showed that antibody responses to infection are variable between individuals and can be influenced by co-infection with other pathogens. This study also showed that the GPC is genetically homogenous and antibody responses traits to EBV and KSHV are partly heritable. Owing to the richness in phenotypic detail and availability of host genotype data in a large fraction of individuals, the GPC opens up avenues of research to greatly improve our understanding of how host genetics contribute to inter-individual variability in immune responses to EBV and KSHV infections, particularly in a population that sustains such a high transmission of infection and bears a great burden of associated disease. In addition, the data on co-infection by other circulating pathogens provides the opportunity to explore the genetic architecture of EBV and KSHV infections in the context of the environment. Lastly, another advantage of the GPC is the ability to return to individuals to collect more samples to study other questions, such as the viral genomic or genotypic diversity of EBV and KSHV, facilitating the study of both host and virus from the same individuals.

3 Chapter 3: The Influence of Host Genetics on Epstein-Barr Virus Infection

3.1 Introduction

While EBV has been extensively studied[29], the influence of host genetics on potential disease outcome and susceptibility to infection are still unclear. Antibodies against EBV nuclear antigen-1 (EBNA-1) and those against viral capsid antigen (VCA) together are widely used as markers to study the stages of infection. Early life infection and high antibody titres have been strongly linked to development of certain cancers[42,119-127]. In a single individual, antibody titres have been found to remain fairly constant throughout life in the absence of immunosuppression, and intense stress[42]. In addition, inter-individual variability in IgG responses to EBNA-1 and VCA have been found to be 32-48% heritable traits [121,412] and thus suggestive of host genetic influence. Familial clustering of diseases also suggests shared genetic risk factors underlying pathogenesis[415,416].

Candidate Gene Approaches

Studies in the 1990s through the early 2000s have relied on candidate gene approaches to investigate associations between variation in host genes and EBV infection and associated diseases[417]. Genes of interest were selected based on *a priori* knowledge of biological function and sets of markers were genotyped and association was tested between cases and controls looking for statistically significant differences in allele frequencies. While most of the studies highlighted interesting putative associations with genes involved in EBV immune response (reviewed below), lack of replication and consistency between studies, probably attributable to low samples sizes and statistically lenient p-value thresholds of between 0.01 to 0.05 (to provide evidence of association), suggests most findings were likely false positives (Table 3.1).

Cytokine Genes (results summarised in Table 3.1)

Cytokine genes and their receptors play a role in cell-mediated immune response, and have been popular candidates owing to their immune regulatory and inflammatory response functions. Hurme and Helminen reported differences in IL1-β allele frequencies in EBV seronegative vs seropositive individuals[418] suggesting immunological differences play a role in EBV defense mechanisms. The IL1-β polymorphism was not statistically significantly associated in Japanese individuals with Infectious Mononucleosis (IM), haemophagocytic lymphocytosis (HLH) and chronic active EBV infection (CAEBV). However, the IL1-α -889C and the TGF-β1 codon 10C allele were significantly lower and higher respectively, in EBV seropositive individuals with IM or HLH than in controls[419]. The TGF-β1 10C polymorphism increases levels of mRNA and protein expression[420,421] which consequently inhibits immune response to EBV following exposure, thereby, the lack of viral control may promote cell proliferation leading to disease as has been shown in EBV-positive Burkitt's Lymphoma (BL) cell lines and B cells[422]. The Hurme group also investigated promoter polymorphisms in IL-10 at positions -1082A/G, -812C/T and -592C/A and showed the ATA haplotype had a protective effect against primary EBV infection[423,424]. They showed this was possibly mediated by production of high IL-10 levels which control viral infection via anti-inflammatory responses[424]. Another study reported a higher frequency in the IL-10 promoter GCC haplotype in EBV seropositive-BL children compared to controls, with the -1082GG genotypes associated with a higher risk of BL development[425]. The IL-10 -819C promoter polymorphism was found to be associated with high anti-VCA IgG titres in in Japanese women, but not in men[426]. These genetic findings, however, failed to replicate in another study of HL[427], a study of EBV-positive individuals with gastric cancers[428] and a study of endemic BL in EBV positive children from Kenya[429].

HLA Genes (results summarised in Table 3.1)

Genetic variation in the HLA locus of the major histocompatibility complex (MHC) on chromosome 6p21.3 has also been of much interest as class I and II alleles participate in presentation of viral antigenic peptides to CD8+ and CD4+ T cells respectively, and thus modulate immune responses to control infection. An analysis of seronegative

adults >60 years who have never seroconverted, suggested long term protection was associated with *HLA-C* and *HLA-Bw4* variants[430]. Microsatelllite markers (D6S256 and D6S510) in the HLA class I region and SNPs in the 80kb region near *HLA-A* and *HcG9* (HLA complex group 9) genes have been found to be associated with EBV-positive classical Hodgkin's Lymphoma (cHL)[431,432]. They also found *HLA-A*02* associated with a decreased risk and *HLA-A*01* with increased risk of developing cHL in EBV positive patients[433]. *HLA-A*02* has been shown to facilitate presentation of EBV antigenic peptides to T-cells and thus may explain its protective phenotype, *HLA-A*01,* however lacks the ability to evoke an immunogenic cytotoxic T lymphocyte responses thus resulting in increased susceptibility as shown in another study[434]. A similar study reported the same microsatellite associations in the HLA class I locus along with the SNPs rs253088 and rs6457110 on chromosome 5 and 6 respectively, with the development of IM[435]. In contrast, the *HLA-A* SNPs rs253088 and rs6457110 failed to reproduce association with anti-EBNA-1 titres or in patients with history of IM in a recent study of MS by Simon *et al*[436]. The HLA-DR2 haplotype has been found to be positively correlated with anti-VCA IgG titres and a risk factor for MS in a Danish study with 517 healthy individuals[437]. A HLA genetic screening study for cHL showed distinct genetic variants between EBV positive vs EBV negative cHL and while the HLA-DR2 polymorphism was not statistically significant in EBV positive cHL patients, an increased susceptibility was associated with *HLA-A1*, and *HLA-A2* was associated with a ~40% reduced risk[438,439]. Using directly typed HLA alleles, these findings were also extended by an alternative study which also identified *HLA* class II alleles, HLA-DRB15*01:01 and HLA–DPB1*01:01 as associated with a reduced risk of EBV positive cHL[440].

Other Putative Candidate Genes

Mannose-binding lectin (MBL) which plays a role in the innate immune defence, has been reported to influence EBV infection in infants <4 years in a Greenland cohort[441]. This study showed that MBL2 genotypes leading to MBL-insufficiency were associated with seronegativity and lower VCA-IgG response compared to MBL-sufficient infants, thereby, resulting in a delayed primary infection[441]. In a study with 755 asymptomatic Cantonese individuals from a Nasopharyngeal Carcinoma (NPC) endemic region,

variation in the homologous recombination repair (HRR) genes (*MDC1, RAD54L, TP53BP1, RPA1, LIG3* and *RFC1*) which are involved in lytic reactivation and viral reactivation were found to be associated with high anti-VCA IgA responses and influence EBV seropositivity[442].

While these studies highlight the potential role of immunogenetic variation in the control of EBV infection, results should be interpreted with caution as candidate gene studies have been criticized for their lack of thoroughness and with small sample sizes the studies are not as robust, leading to more false positive and less convincing associations.

Table 3.1 Putative candidate loci associated with EBV and associated diseases identified by candidate gene approaches

Trait	N	Subjects	Gene	Variants(s)	P	OR (95% C.I)	Ancestry	Ref
EBV Serostatus	400	380 EBV seropositive, 20 EBV seronegative	*IL-1β*	-511 promoter polymorphism	<0.05	N.R	Finnish	[418]
Infectious Mononucleosis (IM)	111	30 cases,81 seropositive controls	*TGF-β, IL-1α*	10C -899C	<0.001 <0.05	N.R	Japanese	[419]
Haemophagocytic Lymphohistiocytosis	109	28 cases, 81 seropositive controls	*TGF-β*	10C	<0.001	N.R	Japanese	
EBV Serostatus	108	36 patients acute EBV infection, 52 seropositive, 20 seronegative	*IL-10*	-1082A	<0.001	N.R	Finnish	[424]
Primary EBV infection	116	44 Seropositive, 72 seronegative	*IL-10*	-1082A, -819T, -592A	0.04	2.6 (1.04-6.7)	Finnish	[424]
Burkitt's Lymphoma (BL)	278	62 paediatric cases, 216 controls	*IL-10*	-1082GG	0.008	2.62 (1.25-5.51)	Brazilian	[425]
VCA IgG antibody response levels	123	123 Females	*IL-10*	-819CC	0.037	4.31 (1.09-29.79)	Japanese	[426]
EBV Serostatus	56	17 Seronegative, 39 seropositive (age>60y)	*HLA-C* *HLA-B*	35TT Bw4	0.03 0.04	N.R	European	[430]
Hodgkin's Lymphoma (HL)	402	54 cases, 292 family controls	*HLA-A*	D6S265 (126bp) D6S510(284bp)	0.0006 0.005	8.25(2.49-27.4) 7.14(1.94-26.3)	Dutch	[431,432]
Hodgkin's Lymphoma	198	81 cases, 117 family controls	*HLA-A* *HCG9*	rs471326 GG, rs2523972 AA	$6.58x10^{-6}$ $1.13x10^{-5}$	9.78(2.74-34.89) 9.28(3.27-26.37)	Dutch	441
Hodgkin's Lymphoma	160	70 EBV+ cases, 31 EBV- cases, 59 controls	*HLA-A*	*HLA-A*01* *HLA-A*02*	<0.001 <0.001	N.R	Dutch	[433]
Hodgkin's Lymphoma	934	278 EBV positive cases, 656 EBV negative cases	*HLA-A*	*HLA-A*01* *HLA-A*02*	<0.001 <0.001	2.15(1.60-2.88) 0.70(0.52-0.97)	Dutch	[434]
Infectious Mononucleosis	286	97 EBV positive IM, 140 EBV positive no IM, 49 EBV negative	*HLA-A*	Rs2530388-A Rs6457110-A	0.011 0.038	N.R N.R	European	[435]
VCA IgG antibody response levels	517	316 male, 201 female healthy subjects	*HLA-DR2*	*HLA-DR2*	0.03	N.R	Danish	[437]
Hodgkin's Lymphoma	156	84 EBV positive cases, 72 EBV negative cases	*HLA-A*	*HLA-A*02:07*	0.0003	6.34(2.33-17.28)	China	[438,439]

Trait	N	Subjects	Gene	Variants(s)	P	OR (95% C.I)	Ancestry	Ref
Hodgkin's Lymphoma	600	156 EBV positive cases, 464 EBV negative cases	*HLA-A*	*HLA-A1* *HLA-A2*	$9.2x10^{-5}$ $5.2x10^{-5}$	2.23(1.12-4.42) 0.39(0.18-0.85)	Dutch	448
Hodgkin's Lymphoma	502 455 378	155 EBV positive cases, 347 EBV controls 144 EBV positive cases, 311 EBV controls 73 EBV positive cases, 305 EBV controls	*HLA-A* *HLA-B* *HLA-DRB1* *HLA-DPB1*	*A*01:01* *B*37:01* *DRB1*15:01* *DPB1*01:01*	$2.5x10^{-7}$ 0.024 0.0019 0.004	2.49(1.75-3.59) 2.58(1.13-6.04) 0.45(0.26-0.75) 0.22(0.06-0.65)	Scottish	[440]
VCA IgA antibody seropositivity	755	128 seropositive, 627 seronegative	*MDC1* *RAD54L* *TP53BP1* *RPA1* *LIG3* *RFC1*	Rs10947087-GA Rs17102086-CC Rs12592757-GT Rs11078676-AA Rs1052536-CT Rs2306597-GA	0.00027 0.00087 0.00581 0.00637 0.00768 0.00854	3.99(1.89-8.46) 2.67(1.51-4.69) 2.19(1.15-4.19) 2.52(1.27-5.00) 1.79(1.02-3.44) 1.94(1.12-3.37)	Chinese	[442]

N.R- Not reported

Genome-wide approaches

With the availability of high resolution genotyping platforms genome-wide association studies (GWAS) have been employed as an agnostic approach to identify variants associated with many different diseases and traits[443]. This approach increases thoroughness, and with the larger sample sizes normally used increases power for discovery of novel loci, and possibly for validation of findings from smaller candidate gene studies. The majority of GWASs performed, have however focused on diseases associated with EBV, and have not taken into account EBV status, making it difficult to tease out genetic factors associated with the underlying infection. Only five GWASs have been performed for EBV infection using a quantitative trait association approach in asymptomatic individuals with antibodies to EBV or viral load as a phenotype[412,413,444-446].

Urayama and colleagues performed the first GWAS of cHL stratified by EBV status in 1200 patients and 6417 controls of European ancestry. They identified two HLA class I SNPS, rs2734986 ($p=1.2x10^{-15}$, OR=2.45) near *HLA-A* and rs6904029 ($p=5.5x10^{-10}$, OR=0.46) located 124kb downstream *HCG9* as independently associated with EBV positive cHL, and in strong LD with HLA-A*01 and HLA-A*02 allelic groups, respectively[447]. Two SNPs were associated with cHL showing no heterogeneity in effect irrespective of EBV status, rs2248462 ($p=1x10^{-13}$, OR=0.61) near the *MICB* gene and rs2395185 ($p=8.3x10^{-25}$, OR=0.56) in *HLA-DRA*[447]. A case-control GWAS was also performed for NPC in individuals of southern Chinese ancestry identifying a lead SNP, rs417162 in the *HLA-A* locus ($p=1x10^{-11}$) and amino acids in the peptide biding groove, in combined discovery and replication studies[448]. They then compared HLA allele frequencies in 1405 NPC patients, 1288 EBV positive and 1352 EBV negative controls (as determined by the presence/absence of antibodies to VCA IgA). They found statistically significant differences in allele frequencies between NPC cases and EBV negative controls but not EBV positive controls for alleles in the HLA-A locus: 02:07 and 33:03, in the HLA-B locus: 27:04, 46:01, 58:01 and in the HLA-C locus 01:02, 03:02 and 12:02[448]. These differences suggest this might be attributable to mechanisms underlying EBV infection.

Rubicz and colleagues, conducted the first genome-wide study investigating antibody responses to EBV infection. They performed a combined genome-wide linkage and association study of anti-EBNA IgG phenotype in 1367 individuals of Mexican American descent, which showed significant evidence of association in the HLA class II region. Two lead SNPs in complete LD with each other, rs477515 and rs2516049 (combined discovery and replication $p=3x10^{-13}$ β=-0.28) were mapped to *HLA-DRB1* with effect alleles T and G, respectively, associated with reduced IgG antibody response levels to the EBNA-1 antigen. They also identified an additional independent SNP, rs2854275-T ($p=2.3x10^{-10}$, β=-0.45) in *HLA-DQB1* associated with anti-EBNA-1 IgG levels. In addition, they reported an overlap of genetic loci between EBNA-1 traits and previously published NPC, HL, SLE and MS susceptibility loci in the HLA region, further suggesting development of disease and viral control are linked[412]. In relation to this, GWAS for anti-EBNA-1 IgG was also performed in 3599 individuals from Australian twin families and meta-analysis with the Mexican American cohort replicated SNPs in the *HLA* class II region and identified genetic overlap with MS risk SNPs[413]. Similarly, through linkage and GWAS in 417 French individuals from 86 families, Pedergnana and colleagues replicated rs477515 and rs2516049 as significantly associated with anti-EBNA-1 IgG responses, showing they were in substantial LD (r^2>0.6) with their lead SNPs rs9268403 and rs9268454 (r^2=1) located in the HLA class II region[444]. The major allele (T) for rs9268403 was found to be associated with high anti–EBNA-1 levels and also associated with HL[444]. However, their study failed to identify any associations through linkage with anti-VCA IgG response levels and thus, did not perform as GWAS for VCA response. Recently, another study of anti-EBNA-1 IgG responses was conducted in 2162 EBV seropositive individuals of European ancestry. Through imputation to 1000 Genomes dataset this study identified strong associations in the HLA class II region, with the lead SNP rs6927022-A ($p=7.35x10^{-26}$) mapping to *HLA-DRB1* and associated with increased levels of IgG[445]. They imputed the HLA region and pinpointed amino acid positions 11 and 26 of *HLA-DRB1* as independent SNPs accounting for the association; amongst HLA alleles, *HLA-DRB1**07:01 ($p=1.01x10^{-14}$),and *HLA-DRB1**03:01 ($p=2.6x10^{-9}$),were the strongest associations, however these alleles could not fully explain their top GWAS signal[445].

Together, these findings show stronger associations with EBV infection (than those identified by candidate gene studies), and potential disease outcome, with genes in the HLA region, suggesting variation in immune response genes play a role in controlling viral infection and pathogenesis. No non-*HLA* loci have been convincingly associated with EBV infection. Large well-powered studies are essential in reliably identifying genetic variants contributing to the risk of EBV infection and associated diseases. However, as none of the studies described above have been performed in individuals of African descent, GWAS in such diverse populations is essential to identify functionally relevant loci in the context of the environment.

3.1.1 Chapter Aims

The overarching aim of this chapter is to bridge the gap in the understanding of host genetic factors that contribute to EBV immune response serological traits in an African population cohort. I use whole-genome sequence data, dense genotyping array data and imputation to a panel with African sequence data to:

I. Identify novel genetic loci associated with EBV infection.
II. Attempt to replicate known EBV associated genetic loci.
III. Investigate the portability of genetic findings between populations of different ancestry.

Contributions

The GPC study team in Uganda coordinated sample collection and DNA extraction. Denise Whitby's group at the Frederick National Laboratory for Cancer Research (FNLCR) conducted serology of all infectious disease traits investigated here. The Wellcome Trust Sanger Institute (WTSI) sequencing pipelines conducted genotyping and whole-genome sequencing. The Global Health and Populations team led by Manj Sandhu at WTSI performed curation of the Ugandan human genetic data including: sequence assembly, alignment and variant calling, SNP and sample quality control (QC), haplotype phasing, generation of the merged 1000G+AGV+UG2G imputation reference panel and provided scripts for imputation. All other analyses unless otherwise stated were performed by myself.

3.2 Methods

3.2.1 Sample Selection

The samples used in this chapter have been described in detail in chapter 2. Briefly, 5000 samples were genotyped on the Illumina HumanOmni 2.5M BeadChip array data (described in chapter 2) and 2000 samples sequenced on the Illumina HiSeq 2000 platform and subject to stringent QC (described below). Following QC, 1567 samples were selected based on the complete availability of EBV antibody response phenotype data and corresponding human genetic data i.e. genotyped or sequenced. Participants' ages ranged from 2-90 years (mean age ± SD = 34 ± 19.6 years, 54% female). For the remainder of this chapter I focus on the genetic association analyses of anti-EBNA-1 IgG and anti-VCA IgG traits, the workflow is summarized in Fig. 3.1.

3.2.2 Whole-Genome Sequencing and Quality Control

Two thousand individuals (UG2G) in the GPC of which 343 individuals had already been genotyped (see chapter 2 section 2.2.5) were subjected to 100 base-paired end sequencing at 4x coverage on the Illumina HiSeq 2000 platform following the manufacturer's protocol. Variant calling was performed with GATK unified Genotyper 3.3. Variant filtering was performed with GATK VariantRecalibrator 3.2 using variant quality score recalibration (VQSR). Stringent sample and variant QC filtering was performed. Low quality variants that mapped to multiple regions within the human genome or did not map to any region, and duplicate variants genotyped on the chip were removed. Samples with a call rate <97% and heterozygosity >3 SD from the mean, discordant genetic sex and reported sex, and sites deviating from Hardy Weinberg Equilibrium ($p<10^{-8}$) were also excluded. Following this, 1632 samples with whole genome sequence (WGS) data that were non-overlapping with the genotype data and ~9.5M SNPs were available for analyses.

3.2.3 Imputation

A merged reference panel consisting of, 1000 Genomes phase III dataset, 320 individuals from the African Genome Variation Project (AGVP)[299], and UG2G sequence data from 1071 unrelated individuals in the GPC, generated following

refinement with Beagle4 and haplotype phasing with SHAPEIT2[401] was used for imputation into the chip data. The reference panel consisted of 3895 unrelated individuals and 104.3M SNPs. Data was phased with SHAPEIT2 and then IMPUTE2[449] was used to estimate unobserved genotypes. I imputed 40.5M SNPs across autosomal markers and X-chromosome of which only high quality sites (info score >0.3 and r^2>0.6) with minor allele frequency (MAF) >= 0.5% were included for analysis. Pooled imputed genotypes (UGWAS) and UG2G sequence data (UG2G) following QC resulted in 6410 samples and 17,619,938 SNPs across autosomes and X-chromosome that were available for genome-wide association analyses.

3.2.4 Association Analyses

For genetic association, 1567 individuals had EBV phenotypes available for analyses. The statistical power to identify genetic variants of genome-wide significance (see below) and with different effect sizes given the sample size was estimated using QUANTO software (http://biostats.usc.edu/software). To control for cryptic relatedness and population structure within the GPC, GWAS was performed using kinship estimation and the standard mixed model approach in GEMMA[450]. For each trait I conducted a quantitative trait and discrete serostatus analysis across ~17M SNPs with MAF >0.5% from pooled UGWAS + UG2G dosages including a kinship matrix analysis described below. To account for lower LD between common variants in African populations and correcting for multiple testing a more stringent threshold of $p<5x10^{-9}$ was used to declare statistical significance, previously determined by Gurdasani *et al* (in review).

3.2.4.1 Kinship Estimation and Mixed-Modelling

To model random effects, I generated a kinship matrix to define pairwise genetic relatedness among individuals using UGWAS and UG2G data for all autosomes and X-chromosome using the k=1 option in GEMMA[450]. The data was LD pruned (r^2=0.2) using dosages from both datasets and a MAF threshold of 1% was applied. The kinship matrix is also useful in modelling phenotypic variance accounting for correlation among individuals.

3.2.4.2 Quantitative Trait Association Analyses

To ensure normalisation of mean fluorescence intensity (MFI) values for statistical analyses, I performed a rank based inverse normal transformation of trait residuals in R statistical package[397]. Residuals obtained following multi-variate linear regression of MFI values for anti-EBNA-1 IgG and anti-VCA IgG responses for 1567 individuals were used for association analysis. For anti-EBNA-1 IgG analysis age, sampling round, HBV and HCV status were adjusted for as significant covariates. For anti-VCA IgG analysis KSHV and HIV statuses were also adjusted for as significant covariates. To account for batch effects, genotyping or sequencing method was adjusted for during association analysis in GEMMA. To boost power to detect association signals, I also conducted a multivariate analysis of both anti-EBNA-1 and anti-VCA IgG traits (r^2=0.3) in GEMMA[450].

3.2.4.3 Discrete Serostatus Association Analysis

Based on MFI cutoffs for anti-EBNA-1 and anti-VCA IgG (previously described in chapter 2), 1567 individuals were classified as seropositive or seronegative and coded 1 and 0 respectively for association analyses. Significant covariates as described above for the quantitative analysis were adjusted for both traits. For anti-EBNA-1 analysis 1206 individuals were seropositive and 361 were seronegative. For anti-VCA 1350 individuals were seropositive and 217 were seronegative.

3.2.4.4 Identification of Secondary Association Signals

Following association analyses, to identify secondary association signals, I performed a conditional analysis in GEMMA[450]. Each SNP within 1MB of the lead association SNP was conditioned. If any SNP was statistically significant it was added stepwise onto the mixed model and analysed jointly, this was done until no SNPs with $p<5x10^{-9}$ remained. All SNPs remaining statistically significant were considered distinct association signals. For conditional analysis where genotype data was unavailable, association summary statistics were obtained, and I performed approximate conditional analysis, as described above, using GCTA[451].

3.2.5 Trans-Ethnic Meta-Analysis

I used MANTRA[452] to perform a genome-wide trans-ethnic meta-analysis of anti-EBNA-1 IgG responses with association summary statistics of 1473 individuals from the Ugandan analysis, combined with publically available 1000 Genomes-imputed GWAS data from 2162 individuals of European ancestry from a previous study[453], across ~4.1M overlapping SNPs. The MANTRA approach leverages differences in LD structures across populations to account for differences in genetic architecture and accommodates heterogeneity of allelic effects between distantly related populations within a Bayesian partition framework. A $\log_{10}$ Bayes Factor (BF) >6 which is comparable to a $p<5x10^{-8}$, previously determined by Wang *et al*[454], is used to show association of a trait with a variant. I determined the heterogeneity of allelic affect sizes using Cochran's *Q*-test for heterogeneity in METAL[455].

3.2.6 Fine Mapping

To refine association signals for anti-EBNA IgG responses, I used MANTRA results to generate and compare fine mapping intervals for each associated lead SNP in the Ugandan and combined Ugandan + European datasets. 99% credible sets most likely to drive association signals and contain causal variants (or tagging unobserved causal variants) were generated by analysing the variants 500kb upstream and downstream of the lead SNP. For this, posterior probabilities were calculated for SNPs and then ranked in decreasing order according to BF, proceeding down the rank until the cumulative posterior probability exceeded 99% as described previously[456,457]. All SNPs >=0.99 were included in the credible set. The credible interval is defined as the length in base pairs spanned by the SNPs.

3.2.7 Functional Annotation of Candidate Variants

To functionally annotate the most significant associations I used the Ensembl Variant Effect Predictor (VEP) and the gene/tissue expression database (GTEx)[458] to access data on expression quantitative trait loci (eQTL) from tissues. GTEx contains information on the relationship between human genetic variation and gene expression levels across multiple tissues[458].

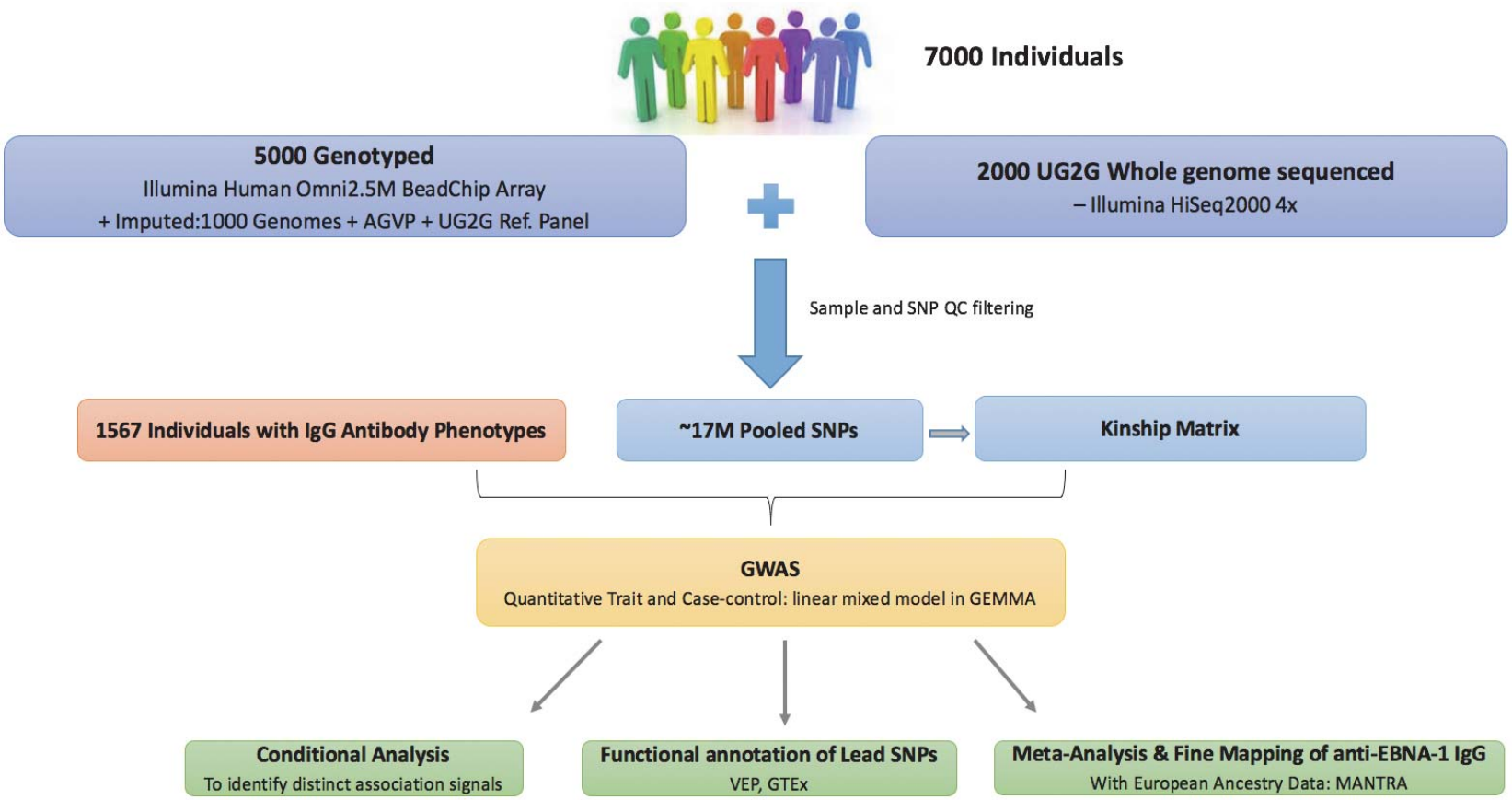


Fig. 3.1 Genome-wide association workflow for EBV serological traits in the Uganda GPC

3.3 Results

Following QC of SNPs for pooled UGWAS and UG2G datasets, ~17M SNPs of MAF>=0.5% were available for GWAS across autosomes and X-chromosome. In 1,567 individuals with corresponding EBV antibody response phenotypes: anti-EBNA-1 and anti-VCA IgG traits, markers of history of infection and viral reactivation, respectively, were available for analyses. With 1,567 samples and using a genome-wide significance threshold of $p<5x10^{-9}$, this study had >80% power to detect common variants with allele frequencies of at least 30% and with moderate to large effect sizes (β>0.25) (Fig. 3.2). For low-frequency variants of 5% or lower, 80% power was only achieved for large effect sizes (β>0.4) (Fig. 3.2). As population structure and genetic relatedness between individuals can confound association studies, systematic differences in the GPC were previously analysed in chapter 2 which showed that the population was homogenous with minimal structure between ethnolinguistic groups. Therefore, using kinship estimation and linear mixed modelling employed in GEMMA controlled well for any inflation due to cryptic relatedness and any residual population substructure, with genome inflation factor (lambda (λ)) for all traits <=1.01 (Fig. 3.3, Fig. 3.6 and Fig. 3.10). This is consistent with association results reported by Gurdasani et al, (in review) which showed no significant difference in λs before and after adjusting for the first 10 principal components as covariates in the linear mixed model using the same dataset. Analysis of the infectious diseases burden in the GPC in chapter 2 also showed that environmental factors i.e. KSHV, HBV, HCV and HIV infections statuses influenced antibody response levels, thus adjustment was also made for significant environmental covariates, to further account for potential confounding that may bias SNP effect estimates and may also improve statistical power by decreasing residual variance.

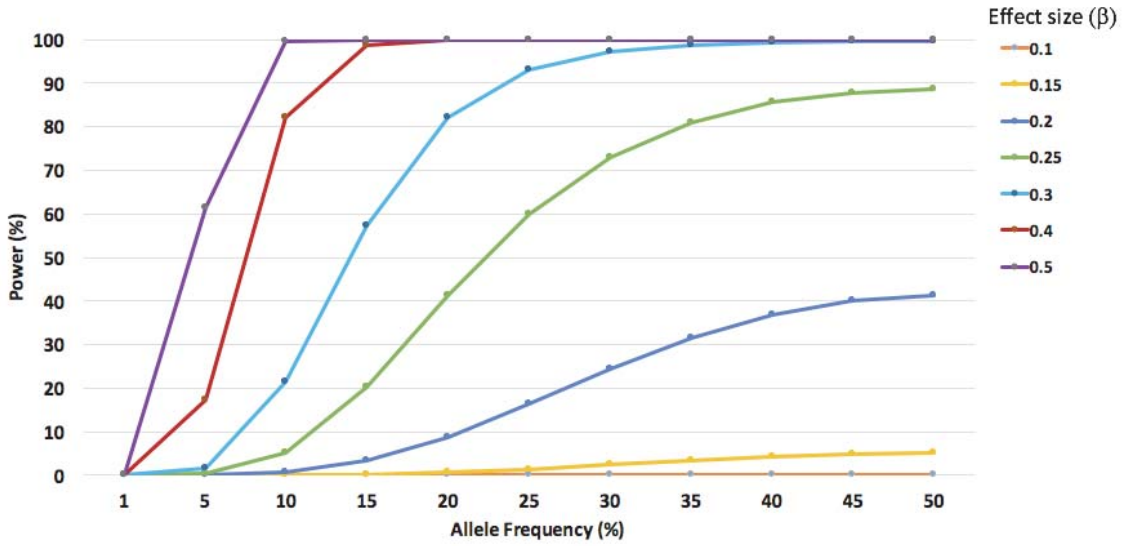


Fig. 3.2 Statistical power (%) to identify genetic variants at $p<5x10^{-9}$, given different allele frequencies (%) and effect sizes (β) (N=1567).

.

3.3.1 Discovery of Novel African-Specific Anti-VCA IgG Loci

To identify genetic variants associated with response to the lytic antigen VCA, IgG antibody responses were quantified and 1350 individuals were categorised seropositive and 217 individuals as seronegative based on VCA MFI cutoffs (see chapter 2, section 2.2.3 and Fig. 2.5). Quantitative analyses of anti-VCA IgG levels did not yield any genome-wide significant associations (Fig. 3.3.A). However, using a case-control analysis for discrete serostatus (i.e seropositive vs. seronegative), I identified four novel genome-wide significant loci (Fig. 3.3.B and Fig. 3.4). rs183816209-T ($p=4.5x10^{-9}$, OR=0.59) an intronic variant in THADA on chromosome 2p21 (Fig. 3.4A), rs190139255-G ($p= 4.0x10^{-10}$, OR=0.57) an intergenic variant on chromosome 7p21.3 with the nearest gene a non-coding RNA U3, 17kb upstream (Fig. 3.3B), rs115256851-C ($p= 6.8x10^{-10}$, OR=0.69) an intronic variant in GALC on chromosome 14q31.3 (Fig. 3.4.C) and rs114576416-G ($p=2.2x10^{-9}$, OR=0.86) an intronic variant in CACGN5 on chromosome 17q24.2 (Fig. 3.4.D). All SNPs passed variant filtering QC post imputation and genotypes were concordant in individuals with overlapping genotype and sequence data, giving confidence to the associations. All SNPs were also associated with seronegativity and were low frequency variants (minor allele frequency<10%) (Table 3.2). rs183816209 and rs115256851 were monomorphic in other 1000 Genomes populations besides African ancestry (Fig. 3.5A, Fig. 3.5B and Table 3.2), suggesting that they are African-specific. rs114676416 was also monomorphic in all populations except Africans and at <1% in admixed Americans (AMR) (Fig. 3.5.C). rs190139255 had no allele frequency data reported in 1000 Genomes populations. No eQTL data was available for these SNPs in the GTEx database.

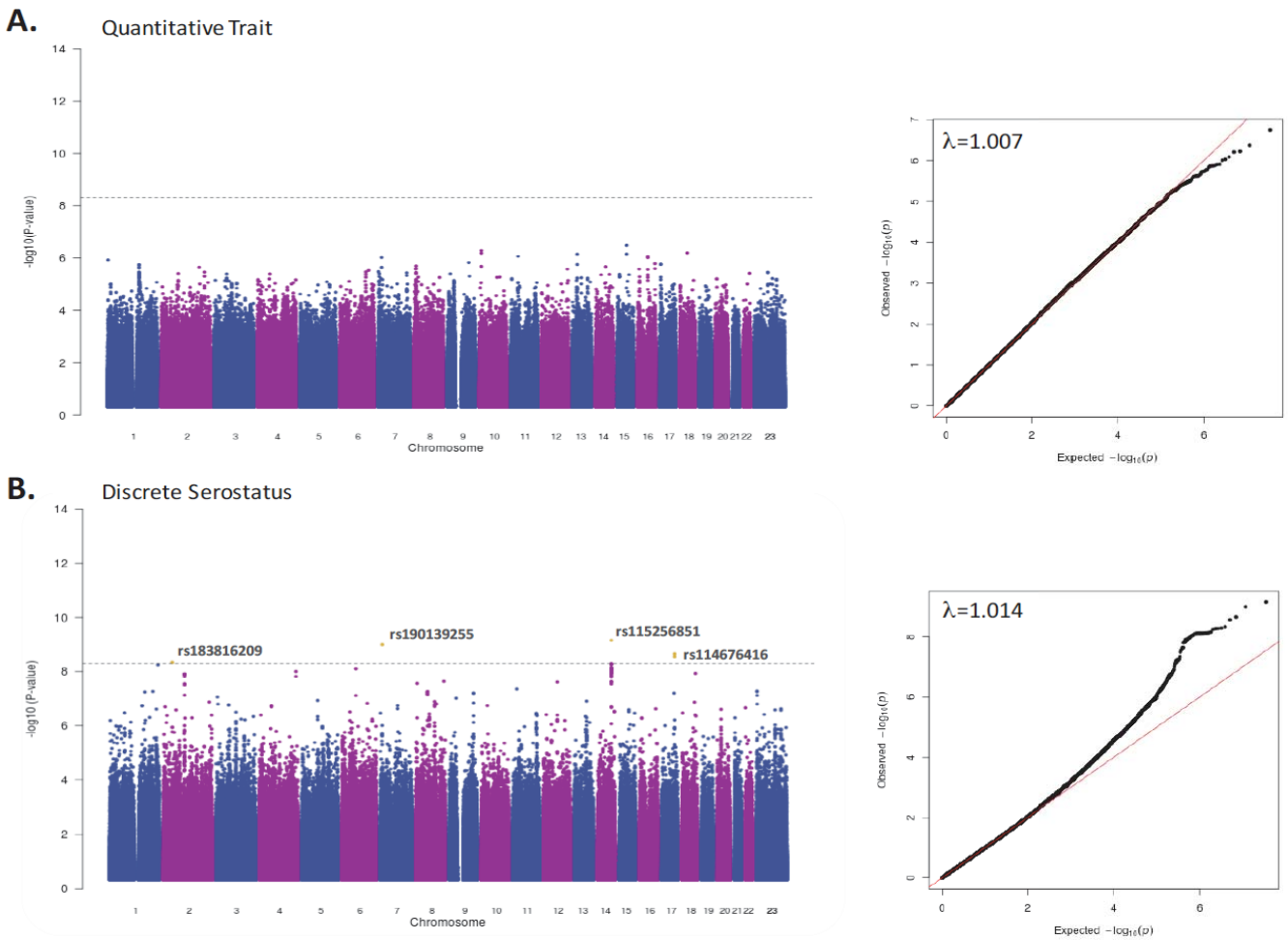


Fig. 3.3. Genome-wide association results of anti-VCA IgG response. Manhattan Plots (Left Panel), Grey dashed line: Genome-wide significance threshold ($p<5x10^{-9}$) and QQ Plots (Right panel). **A.** Quantitative IgG response levels of EBV seropositive individuals (N=1473). **B.** Discrete Serostatus (Pos=1350, Neg=217). 23=X-Chromosome

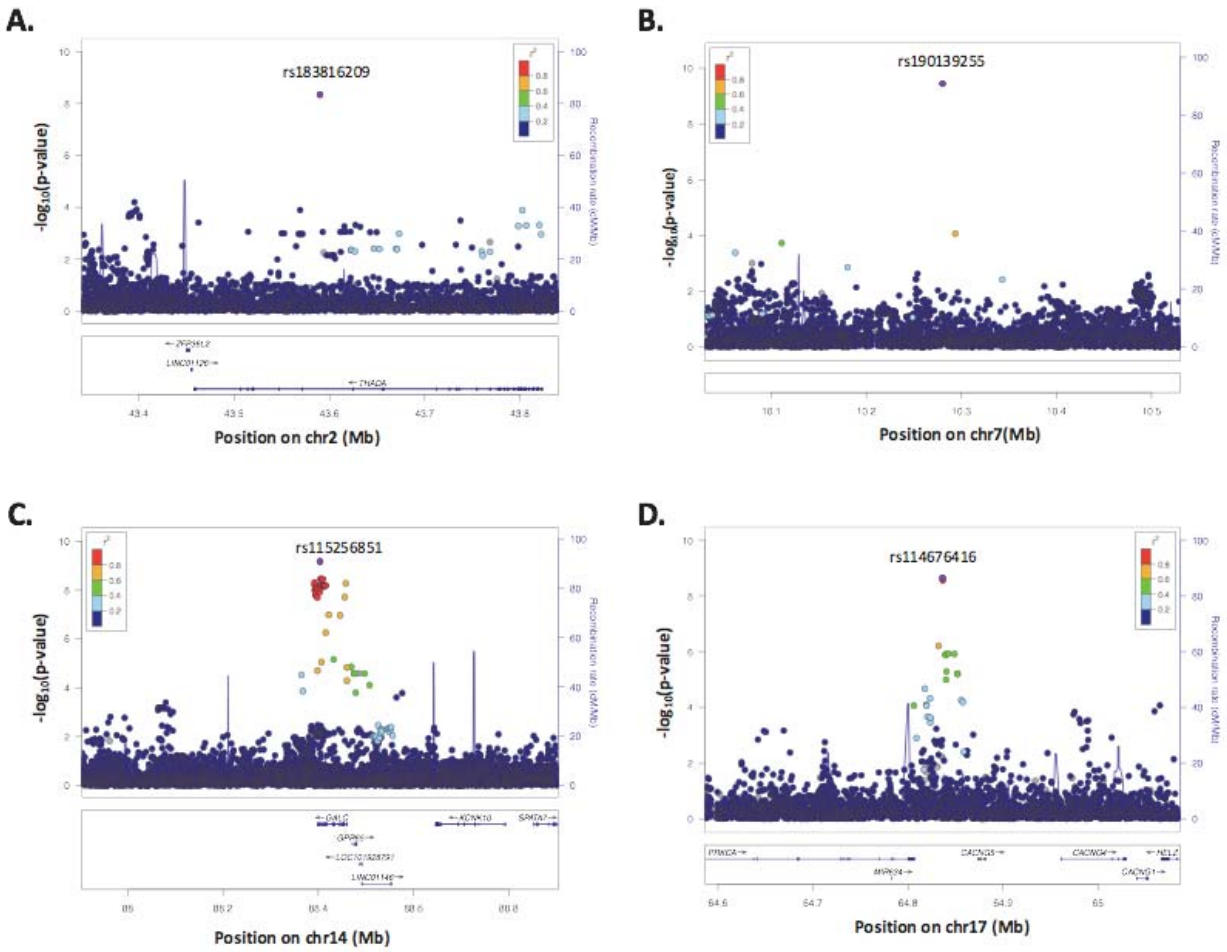


Fig. 3.4 Regional association plots for VCA serostatus genome-wide (GW) significant associations, N=1567, Pos=1350, Neg=217, threshold=p<5x10^{-9}. A. GW significant association on Chromosome 2 in the THADA region. **B.** GW significant association on Chromosome 7. **C.** GW significant association on Chromosome 14 in the GALC region. **D.** GW significant association on Chromosome 17 in the CACNG5 region. The lead SNPs are labelled and coloured in purple. LD (r^2) was calculated based on SNP genotypes.

Fig. 3.5 Comparison of allele frequencies of lead VCA GWAS SNPs between 1000 Genomes phase 3 populations and the GPC. A. rs183816209 in *THADA* on chromosome 2. **B.** rs115256851 in *GALC* on chromosome 14. **C.** rs114676416 in *CACNG5* on chromosome 17. Ancestry codes: AFR-African, AMR-American, EAS-East Asian, EUR, European, SAS-South Asian, ACB-African Caribbean, ASW-African Americans, ESN-Esan (Nigeria), LWK- Luyha (Kenya), GWD-Gambian, MSL-Mende (Sierra Leone), YRI- Yoruba (Nigeria)

3.3.2 Replicating a Known Anti-EBNA-1 IgG Response Locus

To identify genetic variants associated with response to the antigen EBNA-1, IgG antibody responses were quantified and 1206 individuals were categorised seropositive and 361 individuals as seronegative (see chapter 2, section 2.2.3 and Fig. 2.5). Following quantitative GWAS for anti-EBNA-1 IgG response levels to infection, the peak association was in the MHC region, with 404 SNPs meeting the genome-wide significance threshold of $p<5x10^{-9}$ (Fig. 3.6.A). Consistent with previous findings the lead SNP rs9272371 was in the HLA class II region and in *HLA-DQA1* ($p=2.6x10^{-17}$) (Fig. 3.7) with the minor allele (C) being associated with low antibody response levels (β=-0.36) (Table 3.2). Following GWAS of discrete serostatus (i.e seropositive vs. seronegative), only the lead SNP from the quantitative GWAS, rs92732371 ($p=1x10^{-9}$) met the threshold (Fig. 3.6.B), with a lower significance however compared to the quantitative analysis. rs9272371 was in moderate but not strong LD ($r^2>0.8$) with any SNP (Fig. 3.7) and no secondary associations were identified following conditional analysis. rs9272371-C is a common variant and occurs at a frequency of 30.5% in the GPC, in 1000 genomes populations a global allele frequency of 28% was reported, with the lowest frequency of 14% seen in East-Asian ancestry (EAS) (Fig. 3.8). The expression of 10 genes (*C4A, HLA-DQA1, HLA-DQB1-AS1, HLA-DQB1, HLA-DQB2, HLA-DRB1, HLA-DRB5, XXbac-BPG254F23.6, NOTCH4, HLA-DMA*) in 34 tissues were found to be affected by rs9272371 in the GTEx database. All of these genes are known to mediate immune function. The expression of *HLA-DQA1* was significantly down regulated in all tissues including whole blood (eQTL $p=5.2x10^{-36}$, β=-0.75) (Fig. 3.9.A) and EBV transformed lymphocytes (eQTL $p=9x10^{-12}$, β=-0.94) (Fig. 3.9.B), which is consistent with the direction of our associations (Table 3.2).

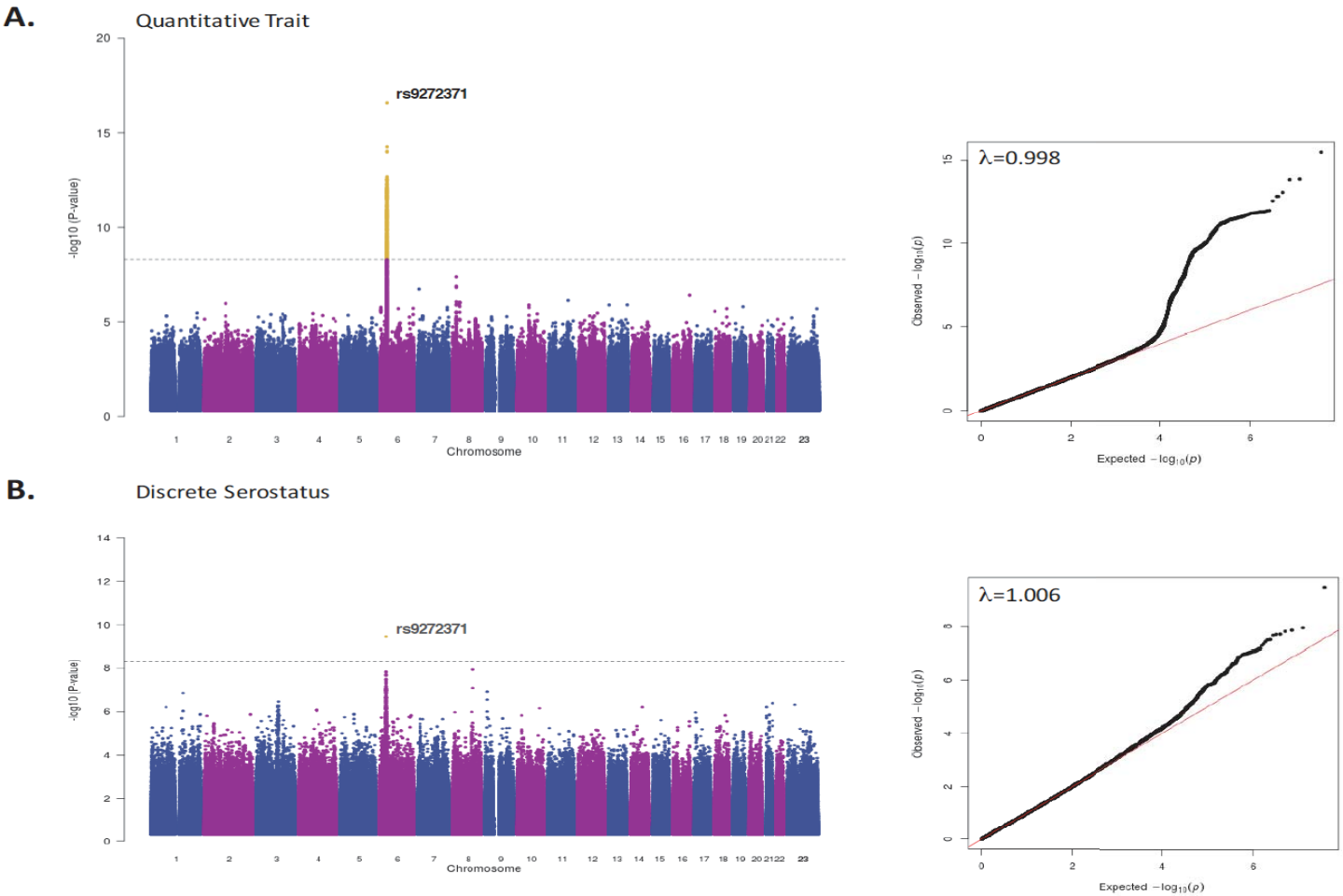


Fig. 3.6 Genome-wide association results of anti-EBNA-1 IgG response. Manhattan Plots (Left Panel), grey dashed line: Genome-wide significance threshold ($p<5x10^{-9}$) and QQ Plots (Right panel). **a.** Quantitative IgG response levels of EBV seropositive individuals (N=1473). **b.** Discrete Serostatus (Pos=1206, Neg=361). 23=X-Chromosome

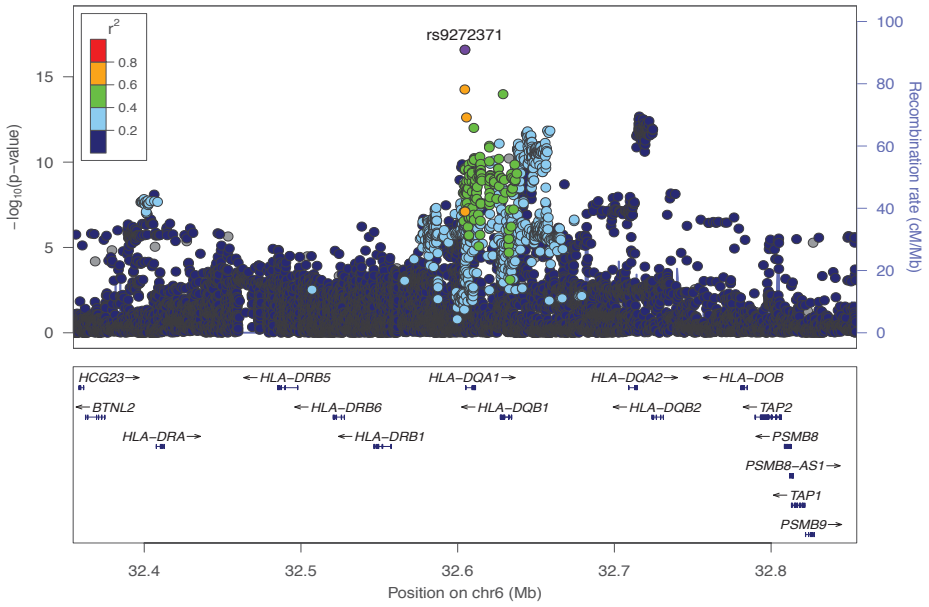


Fig. 3.7 Regional association plot for anti-EBNA-1 IgG response levels in 1473 individuals. (Genome-wide significance threshold= $p<5x10^{-9}$). The lead SNP is labelled and coloured in purple. LD (r^2) was calculated based on SNP genotypes.

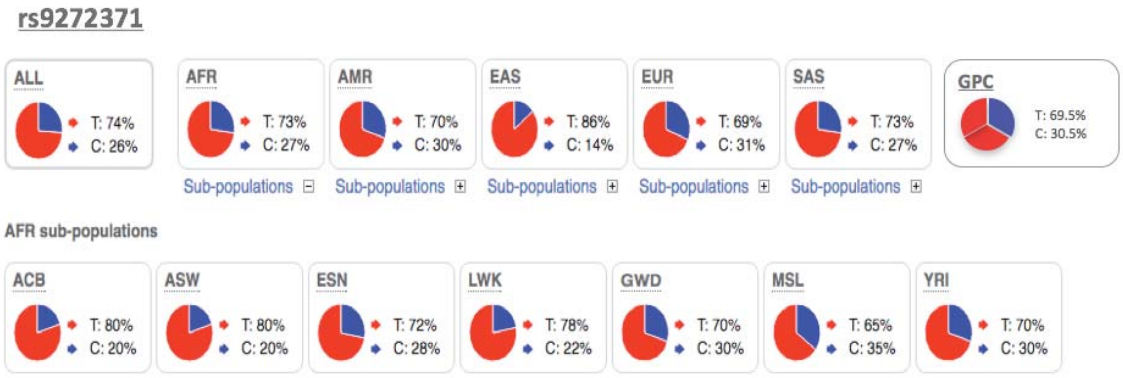


Fig. 3.8 Comparison of allele frequencies of lead EBNA-1 GWAS SNP, rs9272371 in HLA-DQA1 on chromosome 6 - between 1000 Genomes phase 3 populations and the GPC. Ancestry codes: AFR-African, AMR-American, EAS-East Asian, EUR, European, SAS-South Asian, ACB-African Caribbean, ASW- African Americans, ESN-Esan (Nigeria), LWK- Luyha (Kenya), GWD-Gambian, MSL-Mende (Sierra Leone), YRI-Yoruba (Nigeria)

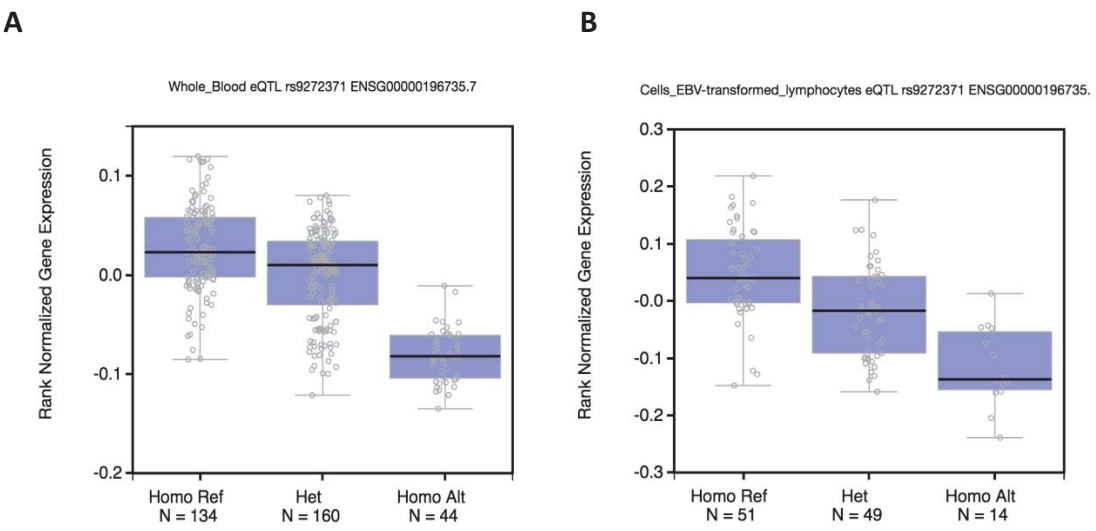


Fig. 3.9 The effect of rs9272371 genotypes on *HLA-DQA1* gene expression, cis-eQTL data from the GTEx database. A. Whole blood (eQTL $p=5.2x10^{-36}$, $\beta=-0.75$). **B.** EBV-transformed lymphocytes (eQTL $p=9x10^{-12}$, $\beta=-0.94$).

3.3.3 Multivariate Quantitative Association Boosts HLA Signal

As multivariate analysis of quantitative traits has been found to increase statistical power for variant detection by exploiting the correlation between phenotypes[459], I combined both anti-EBNA-1 and anti-VCA IgG quantitative traits (r^2=0.3) in a multitrait GWAS. I identified 526 SNPs reaching the genome-wide significance threshold in the HLA region with the lead SNP for anti-EBNA-1 in *HLA-DQA1* rs9272371-C ($p=5.8x10^{-21}$) (Fig. 3.10 and Table 3.2) remaining the lead SNP in this analysis and achieving a stronger significance compared to the univariate analysis. No secondary associations were identified following conditional analyses on rs9272371. This analysis also did not yield any additional statistically significant loci.

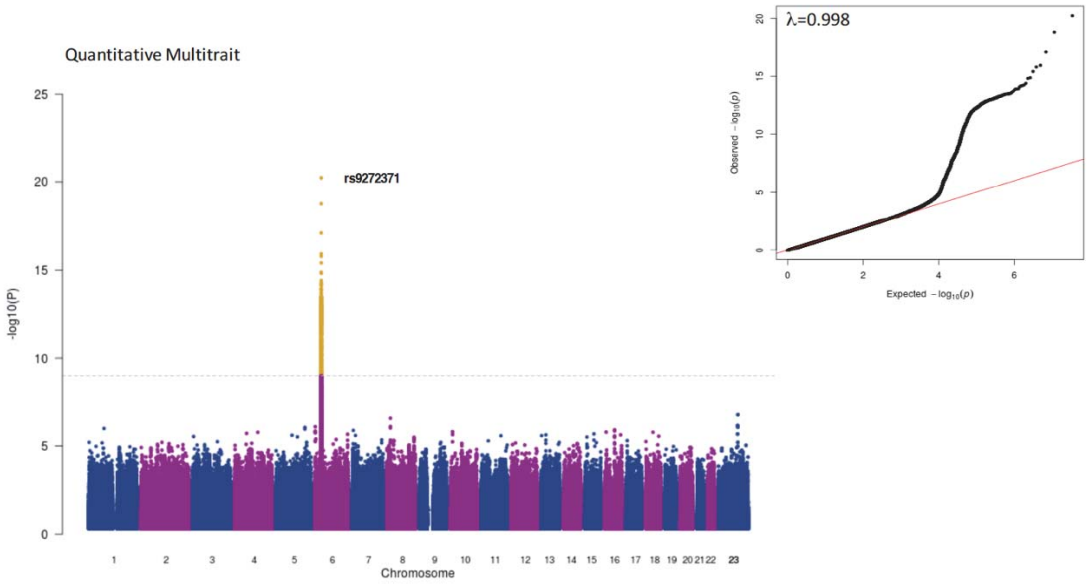


Fig. 3.10 Multivariate genome-wide association results of anti-EBV IgG response levels. Manhattan plot (Left Panel) and QQ Plot (right panel). N=1473. Grey dashed line: Genome-wide significance threshold ($p<5x10^{-9}$). 23=X-Chromosome

Table 3.2 Summary of Genome-wide Significant Association Results in The GPC

Trait	Chr:Pos (b37)	SNP	Gene	Consequence	EA	EAF (%)	P, β_{SNP}/OR (95% CI)
VCA Serostatus	2:43590060	rs183816209	*THADA*	Intronic	T	0.5	$4.5x10^{-9}$ 0.59 (0.41,0.77)
VCA Serostatus	7:10280129	rs190139255	-	Intergenic	G	0.5	$1.0x10^{-9}$ 0.57 (0.39,0.76)
VCA Serostatus	14:88403492	rs115256851	*GALC*	Intronic	C	1.1	$6.9x10^{-10}$ 0.69 (0.57,0.81)
VCA Serostatus	17:64836303	rs114676416	*CACNG5*	Intronic	G	8.1	$2.2x10^{-9}$ 0.86 (0.82,0.91)
EBNA-1 QT	6:32604654	rs9272371	*HLA-DQA1*	Intronic	C	30.5	$2.6x10^{-17}$ -0.36 (-0.26, -0.42)
EBNA-1 Serostatus	6:32604654	rs9272371	*HLA-DQA1*	Intronic	C	30.5	$3.5 x10^{-10}$, 0.89 (0.86,0.93)
EBV Multitrait	6:32604654	rs9272371	*HLA-DQA1*	Intronic	C	30.5	$5.8x10^{-21}$, -0.36 (-0.27, -0.44)

EA=Effect Allele, EAF=Effect Allele Frequency, QT=Quantitative Trait.

OR <1 associated with seronegativity, >1 associated with seropositivity.

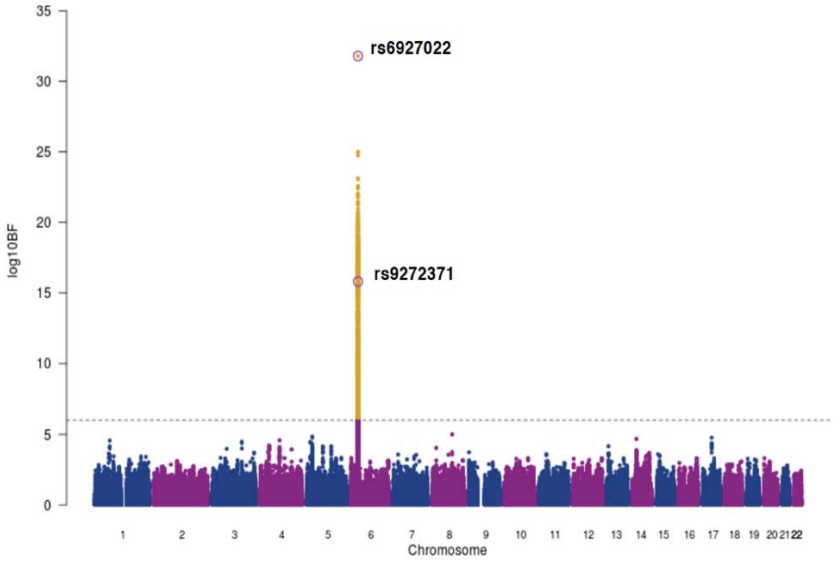
β >0 associated with higher antibody levels, <0 associated with lower antibody levels.

3.3.4 Distinct Association Signals in the HLA Class II Region for Anti-EBNA-1 IgG Response

To investigate whether the lead SNPs identified for anti-EBNA-1 IgG response levels were unique or overlapping with studies conducted in other populations, I compared my association results for the GPC with summary statistics of European and Mexican American ancestries. In the European ancestry GWAS, rs9272371 showed no evidence of significant association (p=0.139)[445] and was absent in the Mexican American GWAS. I also assessed the significance of the lead SNPs, rs6927022 and rs477515 identified in individuals of European and Mexican American descent, respectively, in this study. While rs6927022 was significant in the GPC ($p=2.01x10^{-9}$) and in moderate LD with the lead SNP rs9272371 ($r^2=0.32$) (Table 3.3), the association was markedly attenuated when conditioned on rs9272371 ($p_{cond}=0.0065$) (Table 3.3). In contrast, rs477515 was not statistically significant (p=0.02) in the GPC (Table 3.3), and was in low LD with rs9272371 ($r^2=0.12$). In the European GWAS, rs477515 was not statistically significant after conditioning on rs6927022 ($p_{cond}=0.01$) and thus, was not likely to be an independent signal. As differences in statistical significance between the Ugandan and European association signals may be owing to allelic heterogeneity or differences in LD structure in these populations, further investigation was needed to refine this signal (see below).

I used MANTRA to perform a genome-wide trans-ethnic meta-analysis for anti-EBNA IgG responses, with association summary statistics of 1473 EBV seropositive individuals from the Ugandan GWAS combined with 2162 seropositive individuals from the 1000 Genomes imputed European ancestry GWAS[445], giving a total of 3635 individuals with ~4.9 million shared SNPs for analysis. I excluded genotype data from the Mexican American GWAS[412] as the SNP density was not comparable. Using a threshold of $\log_{10}BF > 6$[454] I found strong evidence of association in the *HLA* class II region (Fig. 3.11), the lead SNP, rs6927022 ($\log_{10}BF=31.8$) was previously identified as the lead association SNP in the European ancestry study, whilst the Ugandan lead SNP rs9272371 ($\log_{10}BF=15.8$) displayed heterogeneity in effect sizes in the two studies ($P_Q=3.56x10^{-8}$) (Table 3.3). rs6927022 is similarly associated with the expression of 9 out of the 10 genes affected by rs9272371 (see section 3.3.2).

A



B

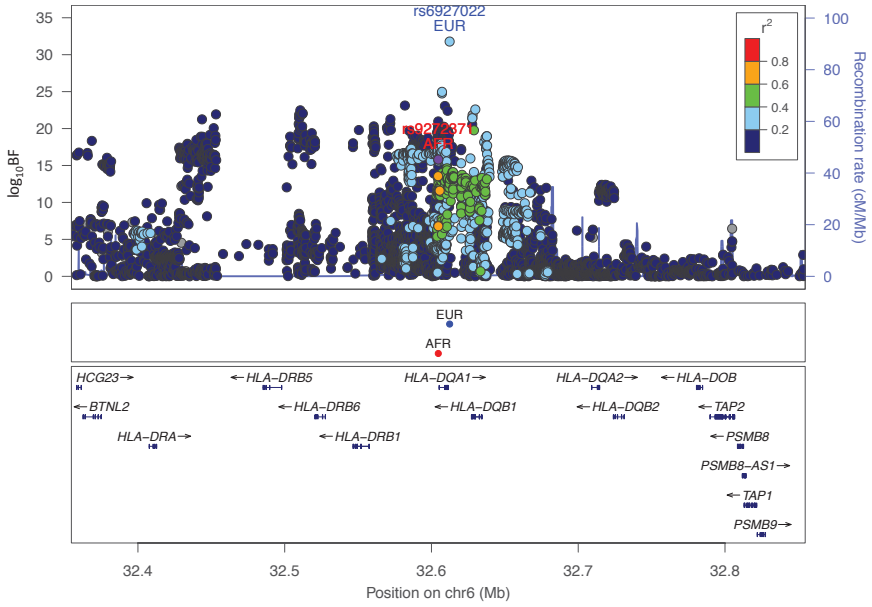


Fig. 3.11 Trans-ethnic meta-analysis association for EBNA-1 IgG response levels in 3635 individuals of Ugandan and European ancestry (EUR) (threshold= log_{10}BF >6). **A. Manhattan plot.** Grey dashed line: threshold= log_{10}BF > 6). The lead SNPs for EUR (rs6927022) and Uganda (rs9272371) GWASs are labelled and circled in purple. Yellow: SNPs that meet the statistical significance threshold. **B. Regional association plot.** The Ugandan lead SNP (AFR) is labelled in red and coloured in purple. LD (r^2) was calculated based on SNP genotypes in the Ugandan dataset. The European (EUR) lead SNP is labelled in blue

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

		Allele	European (N=2162)			Ugandan (N=1473)			MANTRA EUR +UG (N=3635)	
Lead SNP	**Reported gene**	**Effect/ Other**	**EAF**	**Beta**	**P**	**EAF**	**Beta**	**P**	**log_{10}BF**	**P_Q***
rs6927022[a]	HLA-DRB1	A/G	0.59	0.16	$7.35x10^{-26}$	0.73	0.26	$1.93x10^{-09}$	31.8	0.06
rs9272371[b]	HLA-DQA1	C/T	0.37	-0.02	0.14	0.30	-0.36	$2.80x10^{-17}$	15.8	$3.56x10^{-8}$
rs477515[c]	HLA-DRB1	G/A	0.26	-0.15	$1.75x10^{-21}$	0.15	-0.12	0.02	20.1	$4.77x10^{-5}$

EAF - Effect Allele Frequency
N.R – Not reported
[a]European (EUR) lead SNP
[b]Ugandan (UG) lead SNP
[c]Mexican American lead SNP
*P_Q – Cochran's *Q*-test for heterogeneity

To further investigate whether the signals were distinct or partially tagging an un-typed functional variant contributing to both underlying association signals, I performed reciprocal conditional analysis of rs6927022 on the Uganda GWAS in GEMMA and conditioned on the Uganda lead SNP rs9272371 in the European GWAS with association summary statistics using GCTA. Both lead SNPs remained genome-wide significant after adjusting for the effect of the other SNP in the respective cohorts. Together, these findings suggest rs9272371 and rs6927022 are likely to be potentially distinct variants in the *HLA* class II region, with a single signal in Europeans (rs6927022) and a signal mostly driven by rs9272371 in Uganda (Table 3.4). No other locus was found to be in association with anti-EBNA-1 response.

Table 3.4 Conditional analysis of lead Ugandan and European SNPs

	***p*(GWAS)**		**Condition on rs9272371 *p*(cond)**		**Condition on rs6927022 *p*(cond)**	
SNP	**Uganda**	**EUR**	**Uganda**	**EUR***	**Uganda**	**EUR**
rs9272371[a]	2.8×10^{-17}	0.139	-	-	5.9×10^{-10}	0.316
rs6927022[b]	1.93×10^{-09}	7.35×10^{-26}	0.0065	4.5×10^{-26}	-	-

[a]Uganda Lead SNP
[b]EUR (European) Lead SNP
*Conditional analysis performed with association summary statistics in GCTA

The availability of whole-genome sequence data and smaller LD blocks in African populations are advantageous for the refinement of genetic association signals. In line with this, using MANTRA results I generated 99% credible sets most likely to drive association signals and contain variants (or tagging unobserved causal variants) and compared fine mapping intervals for each associated lead SNP by analysing the variants 500kb up and downstream of the lead SNP in the Ugandan and combined Ugandan + European datasets as described previously[456,457] resulting in only one SNP in each credible set, rs6927022 and rs9272371 for the Ugandan + European and Ugandan GWASs respectively, further suggesting that rs927022 does not fully drive associations in the Ugandan population.

3.4 Discussion

In this study, I assessed the host genetic contribution to anti-EBV IgG responses in a rural African population cohort and highlight the utility of dense genotyping combined with whole-genome sequencing and imputation of genotypes to a combined reference panel with African sequence data to aid locus discovery and refinement of causal variants. As sample size is limiting particularly in African populations to conduct well powered GWASs for diseases such as Burkitt's Lymphoma, IgG response traits provide a good intermediate phenotype, indicating the strength of the humoral immune response and control of infection. EBV infection is nearly ubiquitous in Africa, with infection occurring early in childhood[15] and thus seronegativity based on cutoffs which are arbitrarily determined most likely reflect a low immune response as opposed to lack of exposure to EBV. Previous studies have shown correlation of IgG levels with Burkitt's and Hodgkin's Lymphoma and are hypothesised to be potentially predictive of disease risk[42,119,120].

It is interesting that despite the fact that both anti-EBV IgG traits display low heritability in this population after accounting for shared environment, h^2=12% and 7% for anti-EBNA and anti-VCA IgG levels, respectively (described in chapter 2), I still identified strong associations with SNPs contributing to variability in immune responses. In this setting, exposure to other pathogens are cofactors influencing these traits (see chapter 2, section 2.3.3) and thus, have been adjusted for in the GWAS.

Previously, no GWASs had been done for anti-VCA IgG responses and one linkage analysis had been performed, without success in identifying statistically significant associations. For the first time, I have identified novel genetic loci associated with anti-VCA IgG serostatus, which reflect viral reactivation, and are African specific. On chromosome 2, rs183816209-T in *THADA* was associated with VCA seronegativity (OR=0.59). The *THADA* gene encodes for thyroid adenoma associated, which has been observed in thyroid adenomas, a tumour of epithelial origin and evidence suggests it plays a role in the death receptor pathway and apoptosis[460]. In the infected cell, a

strategy EBV has used to evade detection and elimination by the host immune system for viral persistence is the inhibition of apoptotic pathways. This has been demonstrated in B-cell lymphomas that showed resistance to death receptor-mediated apoptosis by Fas/Fas ligand and TRAIL receptors and is dependent on LMP1 signalling[461-463]. Variants in the *THADA* gene have also been associated with a range of diseases including Multiple Sclerosis[34], Type 2 diabetes[464], Polycystic Ovary Syndrome[465], Inflammatory Bowel disease[273,466] and Prostrate cancer[467]. On chromosome 14, rs115256851-C in *GALC* was also associated with seronegativity (0R=0.69). The *GALC* gene encodes the enzyme galactosylceramidase. Mutations in this gene have been associated with Krabbe disease, also known as globoid cell leukodystrophy a progressive, often fatal neurodegenerative disorder that affects the myelin sheath of the nervous system[26-28]. Zhao and colleagues recently showed tumour suppressive effects of GALC expression in EBV-associated nasopharyngeal carcinoma[468]. On chromosome 17, rs114676416-G was identified in *CACNG5* (OR=0.86). *CACNG5* is a member of the family of gamma subunits of voltage dependent calcium channels, and as such is involved in calcium flux[469]. Calcium signalling is a key target for viral proteins as it regulates fundamental cellular processes for EBV entry, B-cell lymphocyte survival and activation[470,471]. Variants in *CACNG5* have been reported to be associated with susceptibility to Bipolar Disorder and Schizophrenia[472]. The variants identified in this study are in genes that may play a role in modulating EBV evasion from host defence, and are associated with VCA seronegativity i.e. a protective effect from EBV viral replication, and thus, allowing viral persistence in infected individuals. At this stage it cannot be established whether these SNPs are causal or tagging causal variants and/or directly regulating the genes they map to, replication in larger sample sizes will be essential to validate these findings taking into account that the SNPs are not common and are African specific. Furthermore, functional investigation of how rs190139255, rs115256851 and rs114676416 affect expression and regulation of genes such as *THADA*, *GALC and CACGN5,* or other genes nearby, will help us further understand how variation in these loci modulate EBV viral reactivation and potential tumour development.

I also successfully replicated an association locus in the HLA class II region for anti-EBNA-1 IgG responses, reflecting infection history, identified in individuals of Mexican American and European descent; and through trans-ethnic meta-analysis of European and African individuals with fine-mapping identify distinct association signals in the *HLA* class II region. Disentangling signals in the HLA region to pinpoint causal alleles is nontrivial owing to the poor representation of African ancestry data on HLA imputation reference panels that are heavily skewed towards European populations. In the European GWAS, Hammer and colleagues were able to achieve resolution of 4 digit classical *HLA* alleles and amino acids in *HLA-DRB1* through imputation using the Type 1 diabetes genetics consortium (T1DGC) Immunochip/HLA reference panel which is predominantly European[445,473]. Current efforts are being made to type *HLA* alleles in the GPC, and imputation of the *HLA* region using this reference panel may help to further resolve the association signals. *HLA* class II molecules present peptides to CD4+ T cells (T helper cells) eliciting both cell mediated and antibody responses to control viral infection. EBV has also been found to use HLA class II molecules as a co-factor mediating entry into B cell lymphocytes[29,30]. Given that *HLA* haplotypes are highly polymorphic and display geographic variability, conducting host genetic studies in diverse populations will allow us to capture variation and understand its contribution to EBV's ability to modulate the immune response to persist and cause disease.

Unsurprisingly none of the candidate gene association signals were replicated in my study, and thus could either be false-positives in the previous results or owing to differences in study design (e.g. quantitative antibody trait measured here vs. case-control), or if they have small-modest effects they could have been missed.

In summary, I describe the first whole-genome sequence and genome-wide association analysis performed for EBV anti-VCA and anti-EBNA-1 IgG traits in an African population. I highlight the combination of whole-genome sequencing and imputing genotypes to a

panel with additional African sequence data to aid discovery of novel loci, low-frequency and population specific variants which might have been missed by using only the 1000 Genomes reference panel for imputation. Furthermore, the availability of data on covariates such as infection with other pathogens allows us to capture genetic variation independently of the environment. I identified four novel loci associated with anti-VCA IgG serostatus and replicated variants in the *HLA* class II region contributing to anti-EBNA IgG response levels. Trans-ethic meta-analysis and fine-mapping of anti-EBNA-1 IgG response with an additional cohort of European ancestry, revealed distinct causal variants driving associations in the two populations. Future studies should include replication of the novel loci associated with anti-VCA IgG responses in larger sample sizes of African descent, particularly as the majority (>90%) of individuals are infected with EBV (i.e. Cases) and thus the number of controls is relatively small. To refine signals in the HLA region it is essential that HLA typing of Ugandan individuals are performed and an imputation panel is generated with African populations well represented to capture genetic diversity. In addition, while GWAS still remains a leading tool to identify variants, functional validation to fully understand their biological effects is crucial.

4 Chapter 4: The Influence of Host Genetics on Kaposi's Sarcoma-Associated Herpesvirus Infection

4.1 Introduction

KSHV seroprevalence displays striking geographic variation that parallels Kaposi sarcoma incidence of disease caused by the virus, with highest prevalence reported in sub-Saharan Africa[474]. In addition to varying geographic distribution in seroprevalence, the finding that KSHV is necessary but insufficient for KSHV disease pathogenesis, the strong correlation of serological status between siblings (not explained by known risk factors), and the familial clustering of KSHV disease[170-172,475,476]are all highly suggestive of host genetic influence. Host genetic variation and its influence on KSHV infection and disease pathogenesis is an emerging field of study and remains largely unexplored. Using candidate gene, whole-exome and next-generation sequencing methods, researchers have reported Mendelian causes of Kaposi's Sarcoma (KS) in children as a result of inborn errors in immunity[477]. Less convincing results have been reported for acquired immunodeficiency in adults as a result of variation in genes modulating the immune system and related pathways[477]. Unlike EBV, no genome-wide association study (GWAS) has been performed for KSHV traits or associated diseases. Below is a review of the findings reported by studies exploring the association between host genetic variation and KSHV infection and associated diseases and summarised in Table 4.1.

Inborn Errors of Immunity

Early case reports in two unrelated children, hypothesised that single-gene inborn errors of immunity were underlying their disease. They were both from the Mediterranean basin, with cases of classic KS which is extremely rare in childhood[478-480]. The first child was born to consanguineous parents and had Tuberculosis, was previously diagnosed at age 9 with autosomal recessive complete IFN-γR1 deficiency as a result of the inheritance of two copies of C77Y IFN-γR1 allele, leading to the surface expression of non-functional

receptors[478]. The second child was previously diagnosed at 23 months with Wiskott-Aldrich Syndrome (WAS), a rare, X-linked immunodeficiency disorder leading to thrombocytopenia, eczema, susceptibility to recurrent infections, and increased risk to autoimmunity and malignancies, due to a deletion (422del6) in the *WAS*, and had other clinical phenotypes including EBV-driven lymphoma[479]. The same group observed that three additional unrelated Turkish children born to consanguineous parents had classic KS and no other clinical phenotype or evidence of immunodeficiency, further hypothesising that heterogenous monogenic defects in children impair KSHV immunity[480], however, at the time did not perform any genetic analysis. Recently, Byun and colleagues used whole-exome sequencing methods combined with biochemical and cellular characterisation to identify and follow up candidate mutations associated with a case of disseminated cutaneous and systemic KS in a 2-year old female Turkish child born to consanguineous parents. This led to the discovery of a homozygous 538-1G<A (rs397515390), loss- of- function splice site mutation in Stromal interaction molecule-1 (*STIM-1*) which was absent in 100 healthy Turkish control subjects. STIM-1 is an ER-resident transmembrane protein involved in regulating store-operated calcium entry and its deficiency results in primary functional T-cell immunodeficiency[481]. Subsequently, in a 14-year old female diagnosed with classic KS and also born to Turkish consanguineous parents, whole-exome sequencing revealed a homozygous C93T missense variant in *TNFRFS4* that conferred an R56C amino acid substitution that was absent in 185 healthy Turkish controls and 974 individuals in the HGDP-CEPH Human Diversity Panel[482]. *TNFRFS4* encodes OX40 a co-stimulatory receptor expressed on activated T-cells and they found that OX40 ligand is found highly expressed on KS lesions, and thus they suggested that R65C mutation resulted in a lack of binding and OX40 deficiency, confirming OX40 is necessary for CD4+ T-cell memory and has a protective effect to KSHV immunity[482]. Together, these studies provide proof-of principle that inherited single gene defects, especially in T-cell immunity genes underlie childhood classic KS. In addition to this, studies by Plancoulaine and colleagues, used segregation analysis and genome-wide linkage scans in families of African descent to provide evidence of a recessive locus on

chromosome 3p22, highlighting a broad linkage peak containing eight genes, that predisposes to KSHV infection in children (LOD score =3.83, $p=1x10^{-5}$)[483,484]. They suggested that this locus does not control infection in adults which is consistent with the hypothesis of age-dependent genetic architecture of infectious diseases[485].

Acquired Immunodeficiency

Cytokine Genes

Cytokine genes have gathered much interest owing to their pivotal role in host immunity and surveillance of tumour cells. In particular, cytokines link cell-mediated and humoral immunity by modulating the Th1/Th2 balance of T-lymphocytes[486]. Studies have shown that a predominant pro-inflammatory response or an altered balance in favour of Th1 cytokines, is associated with viral reactivation and KS pathogenesis[487,488]. In addition, KSHV is known to pirate host proinflammatory genes to facilitate evasion from host innate and adaptive immune defences and promote cell survival and latency[489,490]. Early studies identified associations with KS or KSHV seropositivity with genes such as *FCγRIIIA,* the interleukins*,* however as the studies used lenient P-value thresholds ($p<0.01$-0.05), none of the associations are robust by today's standards and most failed to replicate[491] (summary of results in Table 4.1). In addition, most associations required extensive post-hoc analysis to generate nominal p-values only[*492*]. Below are some examples of the findings from candidate gene studies (summary of results in Table 4.1).

A candidate gene study of host cytokine genes in HIV positive Caucasian American men reported the *IL6* G174C promoter polymorphism to be associated with susceptibility to KS ($p=0.0035$)[493]. They found the homozygous GG genotype, previously found to be associated with decreased plasma IL6 production[494], was overrepresented ($p=0.0046$) in KS cases, and the CC genotype ($p=0.0062$) associated with increased IL6 production[494] was underrepresented, respectively[493]. *IL6* is a proinflammatory cytokine that can

stimulate a Th-2 type T-cell dependent humoral immune response, and thus, as progression to AIDS is associated with immune dysregulation and a Th1-Th2 imbalance, this suggests that in AIDS-KS this polymorphism might be in favour of KSHV, facilitating cell-mediated immune escape from the host[493]. Moreover, KSHV encodes a viral homologue of the *IL6* gene substantiating the importance of *IL6* in pathogenesis[234]. Studies by another group investigating the *IL8* A251T polymorphism in 64 AIDS-KS cases and 89 AIDS with no KS controls reported marginal associations with the TT genotype having a protective effect on severe KS development (p=0.038, OR=0.4)[495]. *IL8* is involved in growth and angiogenesis in a number of tumours including KS, prior studies have shown increased serum IL8 levels in KS and HIV positive patients[496-498]. More recently, in 133 Italian KS cases and 172 KSHV positive controls, SNPs in *IL8RB* C1235T/-1010G diplotype (p=0.003, OR=0.49) and *IL13* G98A promoter region (p=0.01, OR=1.88) were correlated with decreased and increased classic KS risk respectively[499]. Similarly, Brown and colleagues investigated 28 common variants in 14 host immune genes in 172 KSHV seropositive adults from Italy without KS[500]. They found a 3-locus *IL4* haplotype containing the 1098G allele overrepresented in individuals with high lytic K8.1 antibody titres (p=0.02, OR=2.8) and an *IL6* promoter variant also overrepresented in individuals in with high (the upper tertile) compared to low (the two lowest tertiles) antibody titres[500]. In individuals with a high LANA latent antibody titre, they found an overrepresentation of inferred *IL12A* -798T/277A haplotype (p=0.006, OR=2.4) compared to those with low antibody titres[500]. These findings were preliminary and associations were generally weak-moderate, nonetheless, they raised the possibility that host immunogenetics plays a role in controlling KSHV infection.

HLA Genes

The Human leukocyte antigen (HLA) complex has been found to play a crucial role in immunity to infectious disease with different alleles having been associated with susceptibility or resistance to range of infections, including KSHV[54,283,293-295,501-507]. A summary of association results in *HLA* genes are presented in Table 4.1. The first studies

investigating genetic association of *HLA* with classic KS cases were conducted prior to the discovery of KSHV, and with only 62 cases in the largest study, they were also very underpowered, and reported no significant findings[508-514]. Preliminary evidence that variation in *HLA* genes was associated with classic KS was provided by studies led by Masala that identified *HLA* alleles *DRB1*1104* and *DQB1*0604* as predisposing to KS, and *HLA-B58* associated with a protective effect, in 62 cases and 220 controls from a KSHV endemic Sardinian cohort[515]. Further studies by Dorak *et al*., in 147 matched HIV-infected KS cases and controls, and by Guerini *et al.,* in 62 KS cases and 285 healthy controls, also reported *DRB1*1302* and *DQB1*0604* as moderately associated with predisposition to HIV-associated KS[516,517]. More recently, Aissani and colleagues screened 467 candidate susceptibility SNPs in the MHC region in 348 Caucasian American KS cases and controls and reported their strongest signal, rs6902982, an intronic SNP, in *HLA-DMB* associated with a four-fold increase of risk to KS in HIV-infected individuals compared to HIV-infected individuals without KS (p=0.0003, OR=4.04)[518]. Alkharsah and colleagues conducted a study to identify the determinants of viral shedding in ~240 mothers, in a rural South African population and reported that HLA alleles, *HLA-A*6801, HLA-A*4301*, and *HLA-DRB1*04* contribute to increased viral shedding in saliva among KSHV positive subjects[173]. This study suggested that variation in *HLA* genes was associated with impaired viral control, and consequently increased shedding, facilitating KSHV transmission. The most recent study compared the frequencies of *HLA* and their NK cell immunoglobulin-like receptors (*KIR*) allele frequencies and assessed whether they influenced the risk of KSHV seroprevalence and classic KS in an Italian cohort consisting of 250 KS cases, 280 KSHV seropositive and 576 seronegative controls[519]. They found that risk of classic KS was increased in individuals with *HLA-C*0701* (p=0.002, OR=1.6) and reduced in individuals with HLA-A*1101[519]. The *KIR3SD1+HLA-B Bw480I* variant had significant opposing effects i.e. while KS risk was increased 2-fold (p=0.002, OR=2.1), KSHV seroprevalence was 40% lower (p=0.01, OR=0.6); thus, they suggested that *KIR*-mediated NK cell activation may reduce the risk of infection, but if infection occurs, it enhances progression to KS[519]. While this seems an interesting biological proposal, given

unimpressive p-values and small sample sizes, it is a possibility that both results are false-positives.

Other Putative Candidate Genes

A recent candidate gene study of three Finnish familial classic KS cases used whole-genome sequencing, SNP genotyping and linkage analysis to identify a heterozygous C1337T mutation in the DNA-binding domain of the *STAT4* gene conferring an Thr446Ile amino acid change, predicted to be damaging, and absent in 242 Finnish control genomes[520]. *STAT4* belongs to the 7 member STAT family of genes that are highly expressed in myeloid cells, T-lymphocytes and spermatozoa and are involved in the regulation of immunomodulatory genes such as IFNγ[521,522]. Functional follow up in carriers of this variant showed that IFNγ responses in activated T-helper cells were attenuated and thus suggested that *STAT4* is a putative classic KS predisposing gene[520]. Yang and colleagues, used targeted next-generation sequencing of the X-chromosome in 16 PEL cell lines and identified 34 tumour-specific missense variants including a Phe196Ser in *IRAK-1* which was absent in normal tissue from two patients[523]. *IRAK-1* is part of a multicomponent complex and is activated by MyD88, together they mediate toll-like receptor (TLR) immune signalling which is important for controlling KSHV reactivation[524]. The *IRAK-1* Phe196Ser variant was found to be constitutively phosphorylated and necessary for cell survival and a driver for growth[523].

Table 4.1 Putative Candidate Loci Associated with KSHV infection and Diseases

Phenotype	N	Subjects	Nearest Gene	Variant(s)/rsID	P	OR (95% C.I)	Ancestry	Other phenotypes	Ref
Classic KS	1	1 case	*IFN-γR1*[a]	rs104893974	N.R	N.R	Turkish	Tuberculosis	478
Classic KS	1	1 case	*WAS*[a]	422del6 ($_{130}$Asp-Glu$_{131}$)	N.R	N.R	Tunisian	EBV Lymphoma	479
Classic KS	101	1 case, 100 controls	*STIM-1*[b]	rs397515390	N.R	N.R	Turkish	N.R	481
Classic KS	1160	1 case, 1159 controls	*TNFRSF4*[b]	rs587777075	N.R	N.R	Turkish	N.R	482
Classic KS	240	128 KS cases, 112 No KS controls	*FCγRIIIA*	V158F- rs396991	0.00028	N.R	Caucasian American	HIV Positive	491
KSHV Seropositivity	223	130 cases, 93 controls	*FCγRIIIA*	V158F- rs396991	0.071	N.R	Caucasian American	HIV positive	
Classic KS	282	94 cases, 188 controls	*FCγRIIIA*	V158F- rs396991	0.02	0.4 (0.2-0.8)	Italian	HIV Negative	492
AIDS KS	41	115 cases, 126 Controls	*IL6*	rs1800795	0.035	N.R	Caucasian American	HIV Positive	493
AIDS KS	153	84 AIDS-KS cases, 69 AIDS no KS controls	*IL8*	rs4073	0.039	0.49 (0.25-0.97)	Dutch	HIV Positive	495
Classic KS	305	133 cases, 172 controls	*IL8RB*	rs1126579 / rs1126580	0.003	0.49 (0.3-0.78)	Italian	HIV Negative	500
Classic KS	305	133 cases, 172 controls	*IL13*	rs20541	0.01	1.88 (1.15-3.08)	Italian	HIV Negative	
K8.1 Antibody Titre	172	172 KSHV seropositive	*IL4*	rs2243248	0.05	2.8 (1.1–7.0)	Italian	HIV Negative	
K8.1 Antibody Titre	172	172 KSHV seropositive	*IL6*	rs1800795	0.05	3.7 (1.1–12.8)	Italian	HIV Negative	
LANA Antibody Titre	172	172 KSHV seropositive	*IL12A*	rs568408	0.02	2.4 (1.1–5.4)	Italian	HIV Negative	
Classic KS	262	62 Cases, 200 controls	*HLA-A*	*A30*	0.019	0.48 (0.25–0.90)	Italian	N.R	515
Classic KS	262	62 Cases, 200 controls	*HLA-C*	*Cw5*	0.0006	0.32 (0.16 – 0.64)	Italian	N.R	
Classic KS	262	62 Cases, 200 controls	*HLA-C*	*Cw7*	0.01	2.4 (1.2–4.7)	Italian	N.R	

Phenotype	N	Subjects	Nearest Gene	Variant(s)/rsID	P	OR (95% C.I)	Ancestry	Other phenotypes	Ref
Classic KS	262	62 Cases, 200 controls	*HLA-B*	*B58*	0.00001	0.03 (0.002–0.58)	Italian	N.R	[515]
Classic KS	262	62 Cases, 200 controls	*HLA-DRB1*	*DRB1*1104*	0.0473	2.1 (1.05–4.25)	Italian	N.R	
Classic KS	262	62 Cases, 200 controls	*HLA-DRB1*	*DRB1*1302*	0.0037	5.82 (1.73–19.83)	Italian	N.R	
Classic KS	262	62 Cases, 200 controls	*HLA-DRB1*	*DRB1*1601*	0.0425	0.5 (0.25–0.95)	Italian	N.R	
Classic KS	262	62 Cases, 200 controls	*HLA-DQA1*	*DQA1*0302*	0.0189	11.97 (1.26–103.36)	Italian	N.R	
Classic KS	262	62 Cases, 200 controls	*HLA-DQB1*	*DQB1*0502*	0.0465	0.519 (0.27–0.97)	Italian	N.R	
Classic KS	262	62 Cases, 200 controls	*HLA-DQB1*	*DQB1*0604*	0.0017	7.74 (2.02–29.70)	Italian	N.R	
Classic KS	294	147 cases and 147 controls	*HLA-DRB1/-DQB1*	*DRB1*1302 / DQB1*0604*	0.02	6.12 (1.29–28.9)	Italian	HIV Positive	[516]
Classic KS	326	41 cases, 285 controls	*HLA-DRB1*	*DRB1*1302*	<0.001	2.4 (1.26-4.54)	Italian	N.R	[517]
Classic KS	326	41 cases, 285 controls	*HLA-DQB1*	*DQB1*0604*	<0.05	2.6 (1.07-6.43)	Italian	N.R	
Viral Load	217	217 mothers	*HLA-A*	*A*6801*	0.02	3.1 (1.1–8.6)	South African	HIV Positive	[173]
Viral Load	217	217 mothers	*HLA-A*	*A*4301*	0.009	4.4 (1.2–17.2)	South African	HIV Positive	
Viral Load	243	243 mothers	*HLA-DRB1*	*DRB1*04*	0.02	3.4 (1.1–9.7)	South African	HIV positive & negative	
Classic KS	692	348 cases, 348 No KS controls	*HLA-DMB*	rs6902982	0.0003	4.09 (1.9-8.8)	Caucasian American	HIV positive	[518]
Classic KS	1106	250 cases, 280 KSHV seropositive and 576 seronegative controls	*HLA-A*	*HLA-A*11:01*	0.002	0.4 (0.2-0.7)	Italian	HIV Negative	[519]
			HLA-C	*HLA-C*07:01*	0.002	1.6 (1.2-2.1)	Italian	HIV Negative	

Phenotype	N	Subjects	Nearest Gene	Variant(s)/rsID	P	OR (95% C.I)	Ancestry	Other phenotypes	Ref
Classic KS	530	250 cases, 280 KSHV seropositive	*KIR +HLA-B*	*3DS1 / Bw4-80I*	0.002	2.1 (1.1-3.4)	Italian	HIV Negative	[519]
KSHV Seropositivity	856	280 KSHV seropositive and 576 seronegative controls	*KIR +HLA-B*	*3DS1 / Bw4-80I*	0.01	0.6 (0.4-0.9)	Italian	HIV Negative	[519]
Classic KS	245	3 cases, 242 controls	*STAT4* [c]	rs141331848	N.R	N.R	Finnish	N.R	[520]
PEL	16	16 samples	*IRAK1* [d]	rs1059702	N.R	N.R	N.R	N.R	[523]

N- Sample Size, N.R- Not reported

[a] Case report - diagnosis

[b] Whole-exome sequencing identification

[c] Linkage analysis identification

[d] X-Chromosome next generation sequencing identification

While stronger evidence has been provided to support the hypothesis that inborn errors of immunity can underlie disease in childhood, disease in adults and the control of infection in asymptomatic individuals is less clear. Most of the studies reviewed above use statistically lenient p-value thresholds of between 0.01 to 0.05 providing marginal evidence of association for variants (Table 4.1) in immunomodulatory genes associated with virus biological function, pathogenesis of KSHV and development of tumours. It should be noted that the studies have a number of limitations that include: very small sample sizes, failing to adjust for environmental factors such as co-infection with other pathogens, or confounding by strong HLA associations with HIV and AIDS, and mostly not correcting p-values for multiple testing. In addition, some of the above studies did not stratify controls by KSHV serostatus or include KSHV seronegative controls for comparison and all lack replication in independent samples. Lastly, all but one study has been conducted in non-African populations. Therefore, while the field has taken steps in the right direction, findings should be interpreted with caution and replication in large sample sizes and conducting such studies in populations were KSHV and associated diseases are endemic is essential.

4.1.1 Chapter Aims

To overcome the limitations of previous studies and attempt to convincingly identify associations with KSHV immune response traits, I performed a GWAS in >4000 individuals from an African population cohort, where KSHV and KS are endemic, using antibody responses as markers for latent and active infection. I used whole-genome sequence data, dense genotyping array data and imputation to a panel with African sequence data to:

I. Identify novel genetic loci associated with KSHV infection
II. Attempt to replicate previously identified genetic loci

Contributions

The GPC study team in Uganda coordinated sample collection and DNA extraction. Denise Whitby's group at the Frederick National Laboratory for Cancer Research (FNLCR) conducted serology of all infectious disease traits investigated here. The Wellcome Trust Sanger Institute (WTSI) sequencing pipelines conducted genotyping and whole-genome sequencing. The Global Health and Populations team led by Manj Sandhu at WTSI performed curation of the Ugandan human genetic data including: sequence assembly, alignment and variant calling, SNP and sample quality control (QC), haplotype phasing, generation of the merged 1000G+AGV+UG2G imputation reference panel and provided scripts for imputation. All other analyses unless otherwise stated were performed by myself.

4.2 Methods

4.2.1 Sample Selection and Quality Control

4900 samples from the GPC were selected based on the availability of both KSHV antibody response phenotype data and corresponding genotype and sequence data (described in detail in chapter 2 and 3). Briefly, 3641 samples were genotyped on the Illumina HumanOmni 2.5M BeadChip array and 1259 samples sequenced on the Illumina HiSeq 2000 platform and subject to stringent quality control (QC). Participants' ages ranged from 3-97 years (mean age ± SD = 34 ± 19.6 years, 57.5% female). Optical density (OD) values of antibody responses were measured by Enzyme linked immunosorbent assay (ELISA) (as described in detail in chapter 2) and 4466 samples (91%) were classified as KSHV seropositive based on the detection of LANA or K8.1 antigen and used for downstream association analysis (Table 4.2).

Table 4.2 Characteristics of individuals in the GPC used in this study

Characteristic		N=4900	(%)
Sex	Male	2082	42.5
	Female	2818	57.5
Age Group	<15	76	1.6
	15-24	1902	38.8
	25-44	1600	32.6
	>44	1322	26.8
KSHV	Positive	4466	91
	Negative	434	9
HIV	Positive	332	6.8
	Negative	4566	93.2
HBV	Positive	287	6
	Negative	4613	94
HCV	Positive	306	6.3
	Negative	4594	93.7
Round (Year)	3 (1991/92)	71	1.4
	11 (1999/00)	115	2.3
	19 (2007/08)	277	5.6
	22 (2010/11)	4437	90.6
Human Genetic Data*	Genotype	3641	74.3
	Sequence	1259	24.7

*The genotype data is described in detail chapter 2 and the sequence data is described in chapter 3

4.2.2 Imputation

The imputation data is described in detail in chapter 3 (section 3.2.3). Briefly, a merged reference panel consisting 1000 Genomes phase III dataset, 320 individuals from the African Genome Variation Project (AGVP)[299], and UG2G sequence data from 1071 unrelated individuals in the GPC, generated following refinement with Beagle4 and haplotype phasing with SHAPEIT2[401] was used for imputation into the UGWAS chip data (as described in chapter 2 and 3). Following QC, 17,619,938 SNPs across autosomes and X-chromosome remained for analysis.

4.2.3 Association Analyses

For genetic association, pooled UGWAS imputed genotypes and UG2G sequence data post-QC resulted in 4466 samples and 17,619,938 SNPs across autosomes and X-chromosome. The genetic association analyses using pooled UGWAS and UG2G data will be referred to as GWAS and the workflow is summarised in Fig. 4.1. To ensure normalisation of optical density (OD) values for statistical analyses, I performed a rank based inverse normal transformation of trait residuals following linear regression of OD values for anti-LANA-1 IgG and anti-K8.1 IgG responses adjusting for age, sex, sampling round, HIV and HCV statuses in R statistical package[397]. The statistical power to identify genetic variants of genome-wide significance (see below) and with different effect sizes given the sample size was estimated using QUANTO software (http://biostats.usc.edu/software). To control for cryptic relatedness and population structure within the GPC, GWAS was performed using the standard mixed model approach in GEMMA[450]. To account for batch effects genotyping or sequencing method was adjusted for later during association analysis in GEMMA. For each trait I conducted a quantitative trait analysis across ~17M SNPs with MAF >0.5% from pooled UGWAS + UG2G dosages including a kinship matrix analysis (described in chapter 3). To boost power to detect association signals, I conducted a multivariate analysis of both traits (r^2=0.62) also in GEMMA[450]. To account for lower LD between common variants in African populations and correcting for multiple testing a more stringent threshold of

$p<5x10^{-9}$ was used to declare statistical genome-wide significance, previously determined by Gurdasani *et al* and a less stringent threshold of $p<1x10^{-6}$ was used to determine suggestive significance. To identify distinct signals within a locus, SNPs within 1MB of the lead SNP were conditioned in GEMMA and were considered distinct if they met the $p<1x10^{-6}$ threshold.

4.2.4 Functional Annotation of Candidate Variants

To functionally annotate the most significant associations I used the Ensembl Variant Effect Predictor (VEP) and the gene/tissue expression database (GTEx)[458] to access data on expression quantitative trait loci (eQTLs) from tissues. GTEx contains information on the relationship between human genetic variation and gene expression levels across multiple tissues[458].

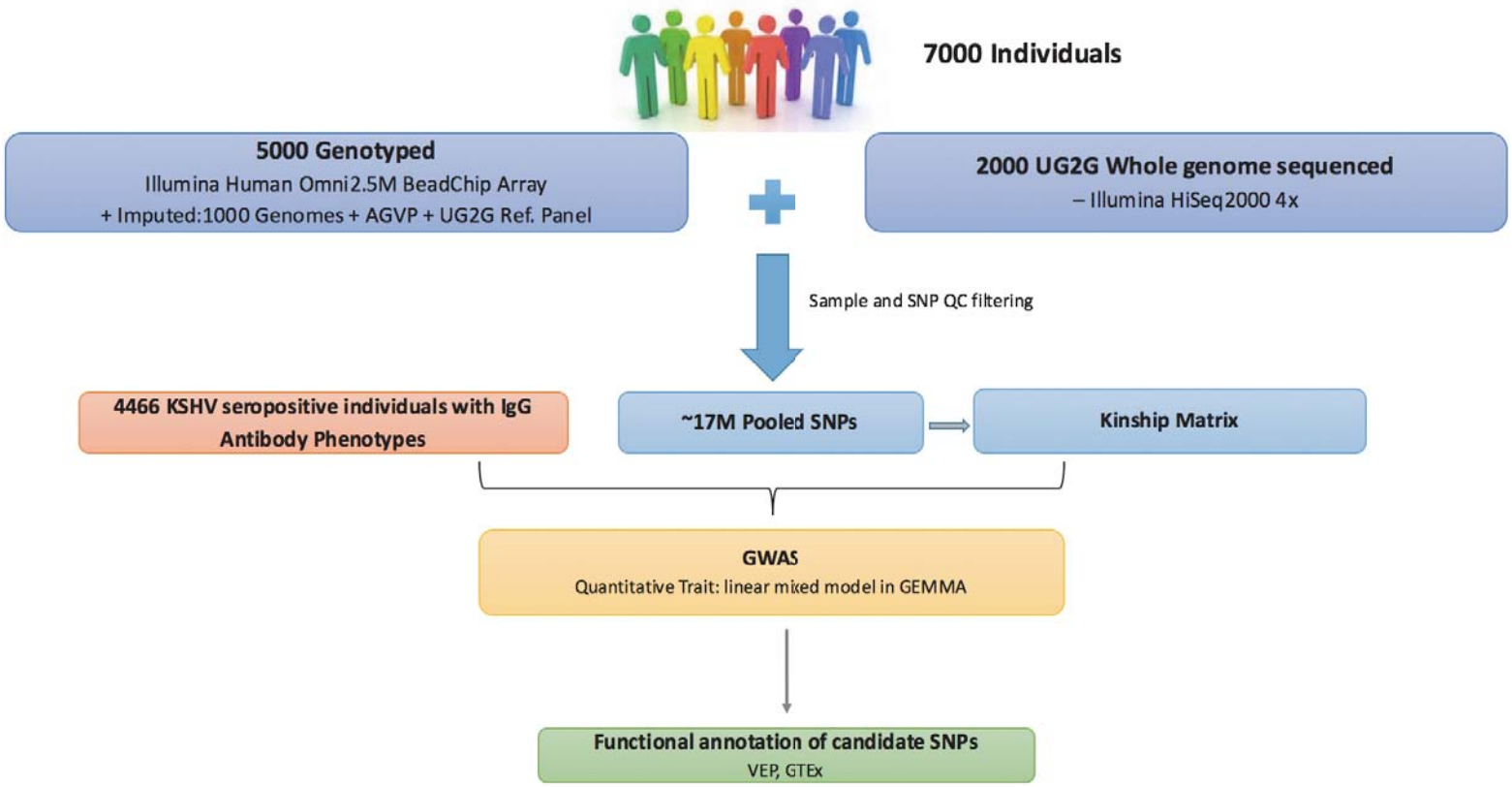


Fig. 4.1 Genome-wide association workflow for KSHV serological traits in the Uganda GPC

4.3 Results

Following QC of SNPs for pooled UGWAS and UG2G datasets, ~17M SNPs of MAF>=0.5% were available for GWAS across autosomes and X-chromosome. For association analyses, 4,466 KSHV seropositive individuals with corresponding KSHV antibody response phenotypes: anti-LANA and anti-K8.1 IgG traits, as markers of latent infection and active infection, respectively, were available. With 4,466 samples and using a genome-wide significance threshold of $p<5x10^{-9}$, this study had >80% power to detect common variants with minor allele frequencies of at least 25% with moderate effect sizes (β=0.15) (Fig. 4.2). For lower frequency variants with allele frequencies ~5%, this study had 80% power to detect moderate-large effect sizes (β=>0.2) (Fig. 4.2). As population structure and genetic relatedness between individuals can confound association studies, systematic differences in the GPC were previously analysed in chapter 2 and showed that the population was homogenous with minimal structure between ethnolinguistic groups. Therefore, using kinship estimation and linear mixed modelling employed in GEMMA controlled well for any inflation due to cryptic relatedness and any residual population substructure, with genome inflation factor (lambda (λ)) for all traits <=1.01 (Fig. 4.3, Fig. 4.6 and Fig. 4.9). This is consistent with results reported for the EBV GWAS in chapter 3 which also had λs close to 1. Adjustment was also made for age, sex, sampling round and significant environmental covariates i.e. HCV and HIV infections status (Table 4.3), to further account for potential confounding that may bias SNP effect estimates and may also improve statistical power by decreasing residual variance.

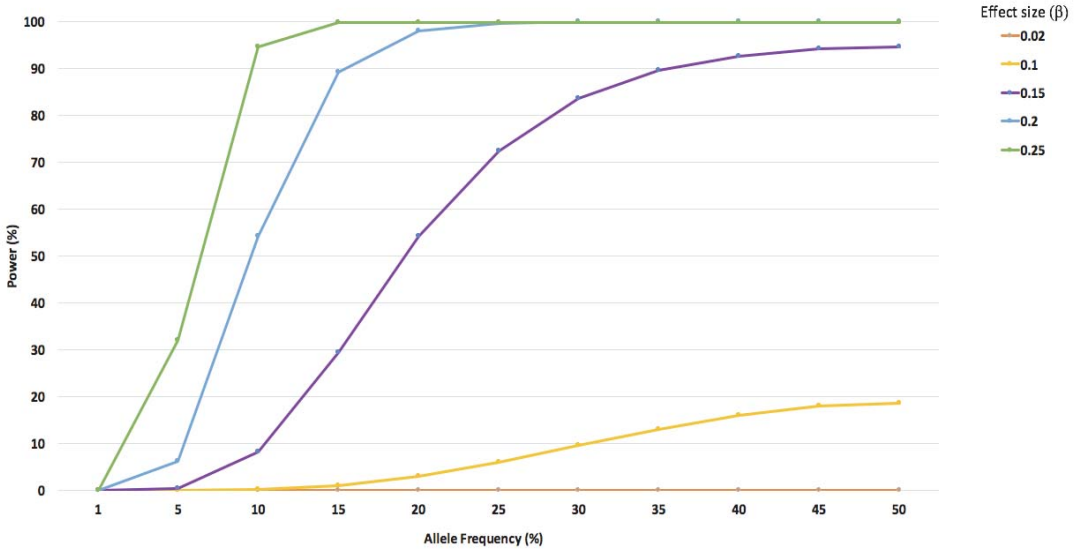


Fig. 4.2 Statistical power to identify genetic variants at $p<5x10^{-9}$, given different allele frequencies (%) and different effect sizes (β) (N=4466).

Table 4.3 Summary of significant linear regression coefficients

Anti-IgG	Age	Sex[a]	Sampling Round[b]	HIV[c]	HCV[c]
LANA	0.009 ($< 2x10^{-16}$)	0.247 ($5.80x10^{-13}$)	0.545 ($8.34x10^{-5}$)	-0.218 (0.0015)	-0.341 ($7.05x10^{-5}$)
K8.1	0.002 (0.005)	0.168 ($9.04x10^{-10}$)	-0.011 (0.03)	-0.990 (0.037)	-0.211 (0.03)

[a] Positive regression coefficient relates to higher OD values in males than females.
[b] Positive regression coefficient relates to higher OD values in Round 22 than other rounds.
[c] Positive regression coefficient relates to higher OD values in seropositive than seronegative individuals.

4.3.1 Discovery of Candidate Loci Associated with Latent KSHV Infection

Following GWAS of quantitative anti-LANA IgG levels for 4466 KSHV seropositive individuals, no SNPs reached the genome-wide significance threshold of $p<1x10^{-9}$, however, using the less stringent the suggestive significance threshold of $p<1x10^{-6}$, SNPs were identified in/nearby seven candidate loci (Fig. 4.3 and Table 4.4). The effect alleles of the lead SNPs in five loci were associated with elevated antibody responses, rs9273255-G ($p=5.19x10^{-7}$, $\beta=0.15$) ~9kb downstream of *HLA-DQA1* on chromosome 6 (Fig.4.4A) and rs71545585-A ($p=2.46x10^{-8}$, $\beta=0.19$) an intronic SNP in *PTPRN2* on chromosome 7 (Fig. 4.4.B), rs111286220-T ($p=4.49x10^{-7}$, $\beta=0.49$) an intergenic variant on chromosome 2 nearby the long intergenic non-coding RNA (lincRNA) gene *LINC01159* (Fig. 4.4.C), rs142363697-T ($p=4.58x10^{-7}$, $\beta=0.45$) an intergenic variant in a gene desert on chromosome 4 (Fig. 4.4.D), and rs138111114-T ($p=8.35x10^{-7}$, $\beta=0.41$) an intronic SNP in *TMEM184b* on chromosome 22 (Fig. 4.5.A). The effect alleles of lead SNPs in two loci were associated with lowered antibody responses: rs4534832-C ($p=7.26x10^{-7}$, $\beta=-0.15$) an intergenic variant nearby *LINC00311* on chromosome 16 (Fig. 4.5.B) and on chromosome 20 an intergenic SNP, rs143267425-T ($p=7.17x10^{-7}$, $\beta=-0.28$) nearby the *SAMHD1* gene (Fig. 4.5.C). The only SNP with expression data in the GTEx database was rs9273255 and it affected the expression of nine genes that mediate immune function (*C4A, HLA-DQA1, HLA-DQA2, HLA-DQB1, HLA-DQB1-AS1, HLA-DQB2, HLA-DRB1, HLA-DRB5, HLA-DRB6, XXbac-BPG254F23.6*) in 26 tissues. The expression of *HLA-DQA1* was significantly down regulated in individuals who were homozygous for the effect allele in all tissues including whole blood (eQTL $p=5.9x10^{-55}$, $\beta=-0.59$) and EBV transformed lymphocytes (eQTL $p=1.5x10^{-24}$, $\beta=-1.01$). All SNPs were also present in other 1000 genomes populations.

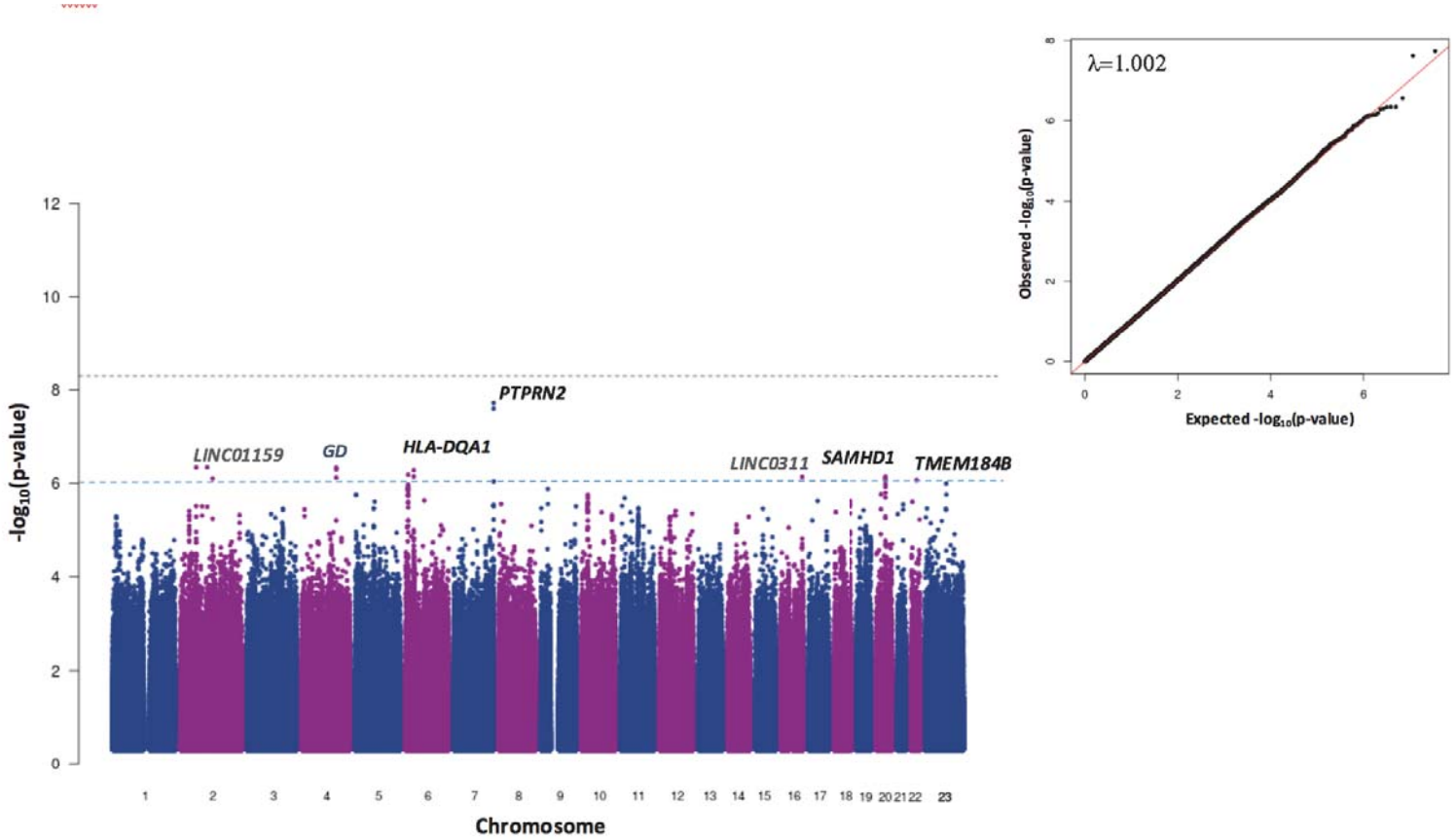


Fig. 4.3 Genome-wide association results of anti-LANA IgG response levels. Manhattan Plot (Left), grey dashed line: genome-wide significance threshold ($p<5x10^{-9}$), blue-dashed line: suggestive significance threshold (p<1x10-6). 23=X-Chromosome. Genes labelled in black (protein-coding), grey (long non-coding RNAs), GD= Gene desert. QQ Plot (Top right).

Table 4.4 Summary of lead anti-LANA IgG response level association results ($p<1x10^{-6}$)

Chr	Position	SNP	Gene*	Consequence	EA	NEA	EAF (%)	P	β (95% C.I)
2	105501955	rs111286220	*LINC01159*	Intergenic	T	C	1.5	4.49E-07	0.49 (0.28 – 0.70)
4	136385432	rs142363697	-	Intergenic	T	C	1.0	4.58E-07	0.45 (0.23 – 0.67)
6	32614228	rs9273255	*HLA-DQA1*	Downstream	G	A	22.3	5.19E-07	0.15 (0.09 – 0.21)
7	158118672	rs71545585	*PTPRN2*	Intron	A	G	10.9	2.46E-08	0.19 (0.12 – 0.26)
16	85273356	rs4534832	*LINC00311*	Intergenic	C	G	12	7.26E-07	-0.15 (-0.22 – -0.08)
20	35600894	rs143267425	*SAMHD1*	Intergenic	T	C	3.4	7.17E-07	-0.28 (-0.0.4 – -0.16)
22	38664909	rs138111114	*TMEM184B*	Intron	T	A	1.6	8.35E-07	0.41 (0.23 – 0.59)

*Mapped gene

EA= Effect allele, NEA= Non effect allele, EAF=Effect allele frequency

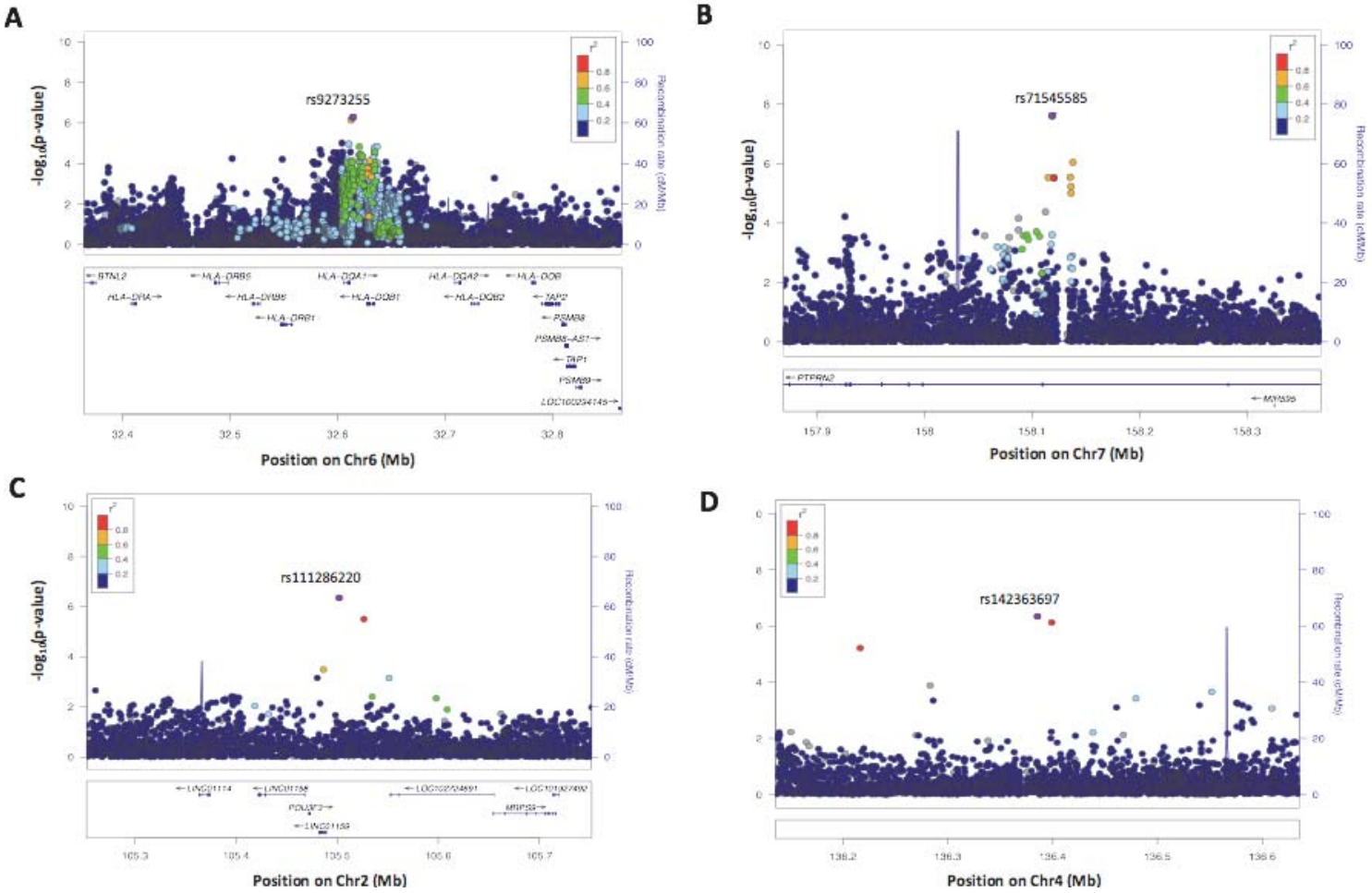


Fig. 4.4 Regional association plots for SNPs associated with anti-LANA IgG ($p<1x10^{-6}$, N=4466.).

A. Association on Chromosome 6 in the *HLA-DQA1* region. **B.** Association on Chromosome 7 in the *PTPRN2* region. **C.** Association on Chromosome 2 nearby *LINC01159*. **D.** Association on Chromosome 4 in gene desert region. The lead SNPs are labelled and coloured in purple. LD (r^2) was calculated based on SNP genotypes

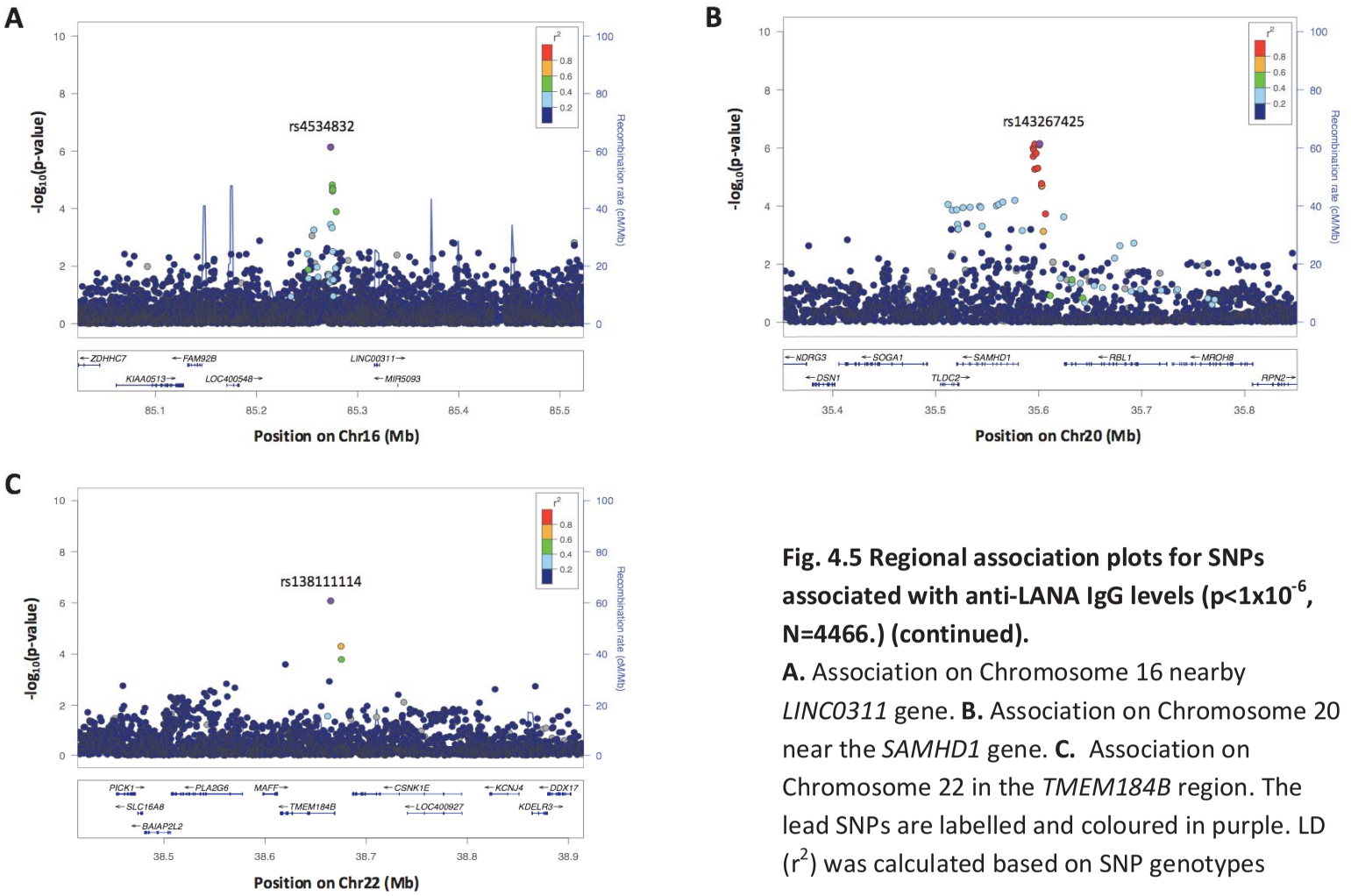


Fig. 4.5 Regional association plots for SNPs associated with anti-LANA IgG levels ($p<1x10^{-6}$, N=4466.) (continued).

A. Association on Chromosome 16 nearby *LINC0311* gene. **B.** Association on Chromosome 20 near the *SAMHD1* gene. **C.** Association on Chromosome 22 in the *TMEM184B* region. The lead SNPs are labelled and coloured in purple. LD (r^2) was calculated based on SNP genotypes

4.3.2 Discovery of Candidate Loci Associated with Increased Lytic Antigen Levels

Following GWAS of quantitative anti-K8.1 IgG levels for 4466 KSHV seropositive individuals, no SNPs reached the genome-wide significance threshold of $p<1x10^{-9}$, however, SNPs were identified in five candidate loci which met the suggestive significance threshold of $p<1x10^{-6}$. All SNPs except for one were in or nearby protein-coding genes on chromosomes 1, 6, 15 and 17; the SNP on chromosome 3 was in a gene desert (Fig. 4.6 and Table 4.5). The effect alleles of SNPs in three loci were associated with elevated K8.1 lytic antibody responses: rs62422641-A ($p=3.22x10^{-7}$, $\beta=0.21$) an intronic variant on chromosome 6 in the *ARID1B* gene (Fig. 4.7.A), rs183160271-A ($p=2.79x10^{-8}$, $\beta=0.35$) an intergenic variant ~50kb downstream of the *DCAKD* gene on chromosome 17 (Fig. 4.7.B). Two loci were associated with low antibody responses, rs1005442-A ($p=3.92x10^{-7}$, $\beta=-0.26$) (Fig. 4.7.C) an intronic variant in the *TNR* gene on chromosome 2 and rs72738070-T ($p=2.63x10^{-7}$, $\beta=-0.73$) an intronic variant in *CGNL* on chromosome 15 (Fig. 4.7.D). The effect allele of SNP 3:20993842 in a gene desert was also associated with elevated antibody responses (Fig. 4.8). None of the lead SNPs in any of the candidate genes had available GTEx data. All SNPs were also present in other 1000 genomes populations.

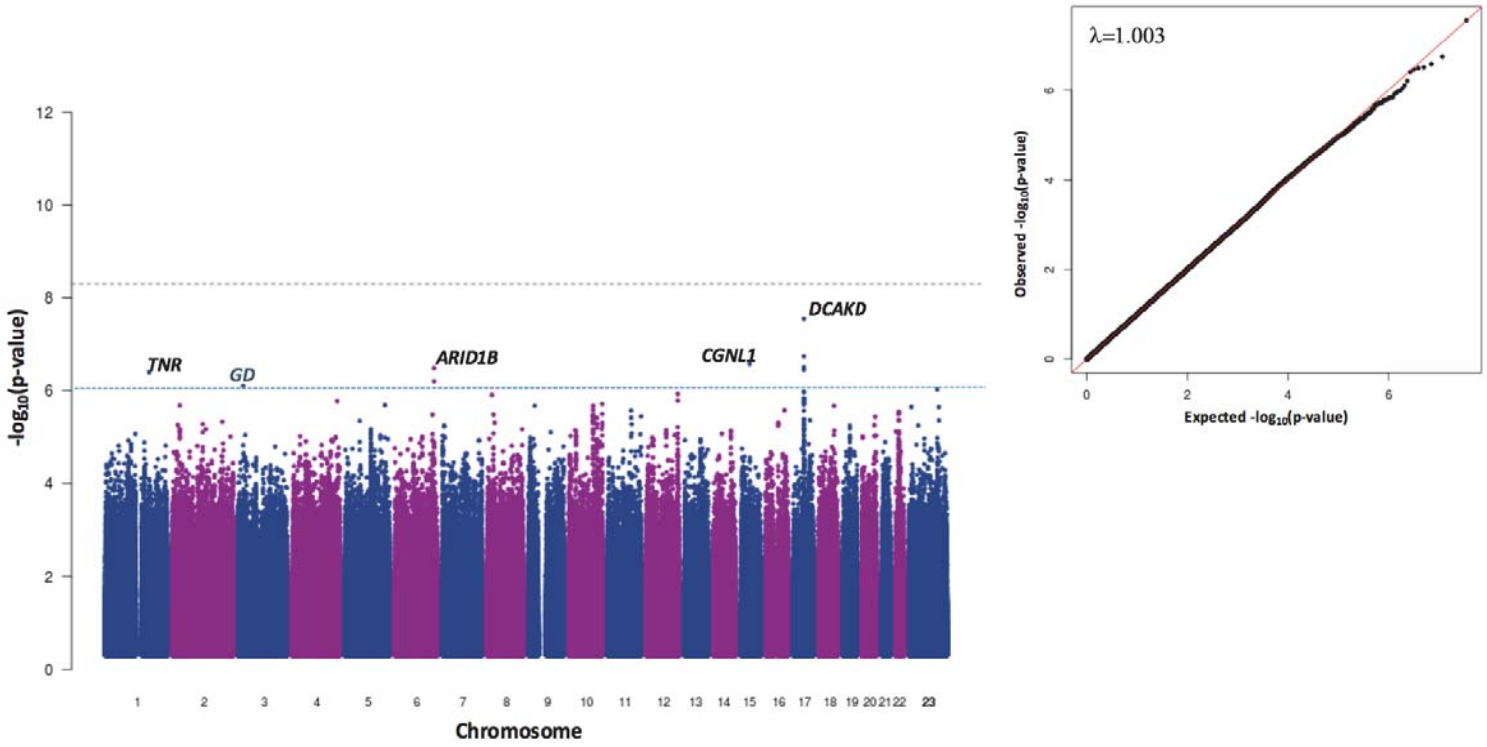


Fig. 4.6 Genome-wide association results of anti-K8.1 IgG response levels. Manhattan Plot (Left), grey dashed line: Genome-wide significance threshold ($p<5x10^{-9}$), blue-dashed line: Suggestive significance threshold ($p<1x10^{-6}$), GD= Gene desert, 23=X-Chromosome. QQ Plot (Top right).

Table 4.5 Summary of lead anti-K8.1 IgG response level association results ($p<1x10^{-6}$)

Chr	Pos	SNP	Gene*	Consequence	EA	NEA	EAF (%)	P	β (95% C.I)
1	175585184	rs1005442	*TNR*	Intron	A	G	4.8	3.92E-07	-0.26 (-0.36 – -0.16)
3	20993842	3:20993842	-	Intergenic	T	C	3.3	7.88E-07	0.34 (0.20 – 0.49)
6	157302035	rs62422641	*ARID1B*	Intron	A	G	8.2	3.22E-07	0.21 (0.13 – 0.29)
15	57708690	rs72738070	*CGNL1*	Intron	T	C	0.6	2.63E-07	-0.73 (-1.02 – -0.43)
17	43053764	rs183160271	*DCAKD*	Intergenic	A	G	3.6	2.79E-08	0.35 (0.22 – 0.48)

*Mapped or closest gene
EA= Effect allele, NEA= Non effect allele, EAF=Effect allele frequency

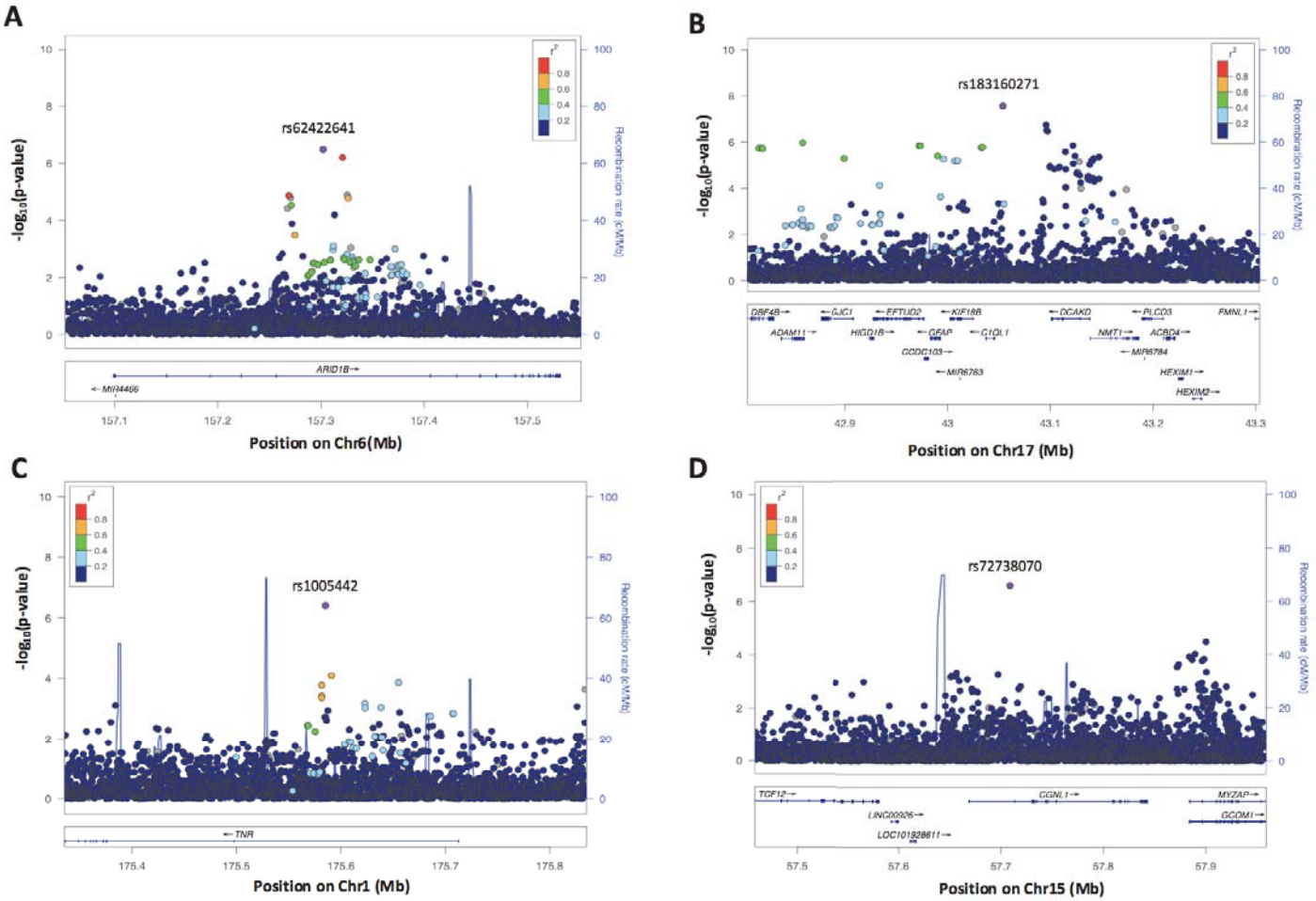


Fig. 4.7 Regional association plots for SNPs associated with anti-K8.1 IgG response levels, N=4466, threshold=p<1x10^{-6}. A. Association on Chromosome 6 in the ARID1B region. **B.** Association on Chromosome 17 in the DCAKD region. **C.** Association on Chromosome 1 in the TNR region. **D.** Association on Chromosome 15 in the CGNL1 region. The lead SNPs are labelled and coloured in purple. LD (r^2) was calculated based on SNP genotypes.

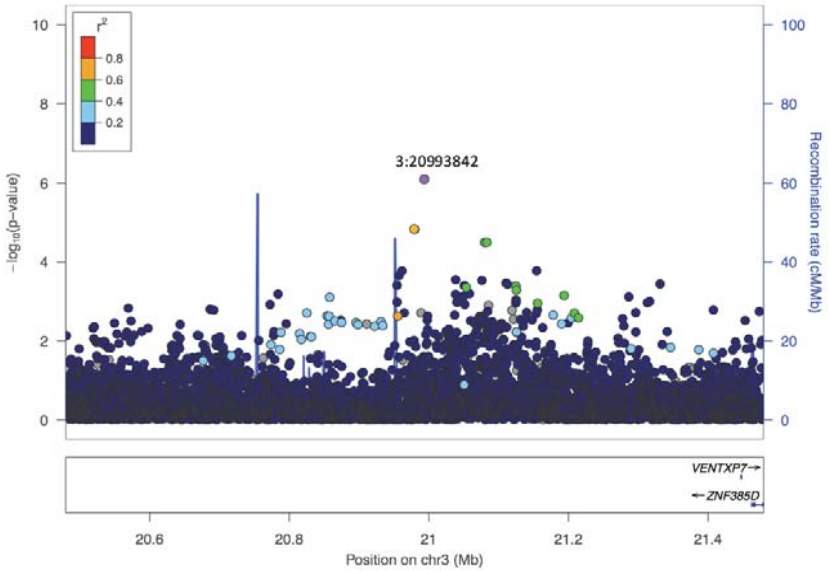


Fig. 4.8 Regional association plots for SNP on chromosome 3 associated with anti-K8.1 IgG response levels, N=4466, threshold=p<1x10^{-6}. The lead SNP is labelled and coloured in purple. LD (r^2) was calculated based on SNP genotypes.

4.3.3 Multivariate Association Analyses of IgG response to KSHV infection

As multivariate analysis of quantitative traits has been found to increase statistical power for variant detection by exploiting the correlation between phenotypes[459], I combined both anti-LANA and anti-K8.1 IgG quantitative traits (r^2=0.68) in a multitrait GWAS. I identified six candidate loci which met the suggestive significance threshold of $p<1x10^{-6}$ (Fig. 4.9 and Table 4.6). Out of the six candidate loci, three loci previously identified as associated with anti-LANA IgG response levels in the univariate analysis (Table 4.4) also remained significant, with the same lead SNPs for rs9273255 ($p=2.61x10^{-7}$) downstream of *HLA-DQA1* on chromosome 6 and rs71545585 an intronic SNP in *PTPRN2* ($p=4.43x10^{-8}$) on chromosome 7, while rs142363697 ($p=9.44x10^{-7}$) an intergenic variant on chromosome 4 was in strong LD (r^2>0.8) with rs142363697 (the lead SNP identified for anti-LANA IgG) (Table 4.6). While the signal in chromosome 4 was attenuated compared to the anti-LANA IgG univariate analysis, the *HLA-DQA1* signal was slightly stronger and the *PTPRN2* signal was similar. The three new candidate loci not previously identified in any of the univariate analyses mapped to chromosomes 2, 17 and 19; rs6752274 an intronic SNP in the lincRNA gene, *AC096554.1* (Fig. 4.10.A), rs151232332 ($p=4.14x10^{-7}$) an intronic SNP in *DUSP14* (Fig. 4.10.B) and rs79312437 ($p=6.43x10^{-7}$) an intronic SNP in *GMFG* (Fig. 4.10.C), respectively. None of the novel SNPs identified have available GTEx data.

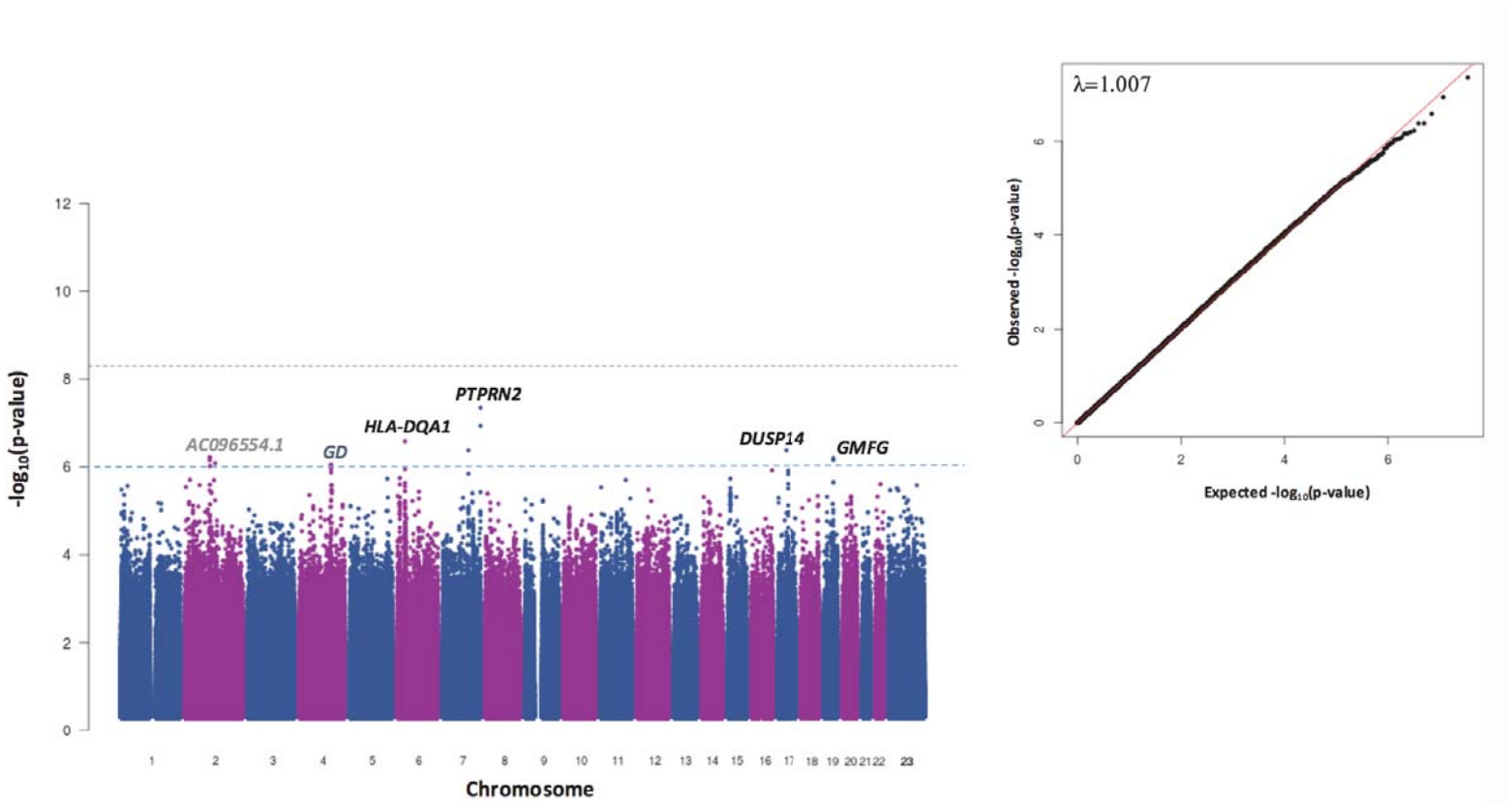


Fig. 4.9 Multivariate Genome-wide Association results of anti-KSHV IgG response levels. Manhattan Plot (Left), grey dashed line: genome-wide significance threshold ($p<5x10^{-9}$), blue-dashed line: suggestive significance threshold ($p<1x10^{-6}$). Genes labelled in black (protein-coding), grey (long non-coding RNAs), GD= Gene desert, 23=X-Chromosome. QQ Plot (Top right).

Table 4.6 Summary of lead anti-KSHV IgG response level multivariate association results ($p<1x10^{-6}$)

Chr	Position	SNP	Gene*	Consequence	EA	NEA	EAF (%)	P	β_{LANA}	$\beta_{K8.1}$
2	**104753193**	**rs6752274**	***AC096554.1***	**Intron**	**A**	**G**	**16.3**	**5.98E-07**	**0.10**	**0.15**
4	133267070	rs78315860	-	Intergenic	G	A	3.6	8.96E-07	-0.02	0.23
6	32614228	rs9273255	*HLA-DQA1*	Downstream	G	A	22.3	2.61E-07	0.15	0.02
7	158118672	rs71545585	*PTPRN2*	Intron	A	G	27.9	4.43E-08	0.19	0.05
17	**35867803**	**rs151232332**	***DUSP14***	**Intron**	**T**	**C**	**5.7**	**4.14E-07**	**-0.22**	**-0.22**
19	**39826332**	**rs79312437**	***GMFG***	**Intron**	**A**	**C**	**14.4**	**6.43E-07**	**-0.11**	**0.05**

*Mapped or closest gene

EA= Effect allele, NEA= Non effect allele, EAF=Effect allele frequency

SNPs in bold are novel associations not identified in the previous univariate analysis of traits

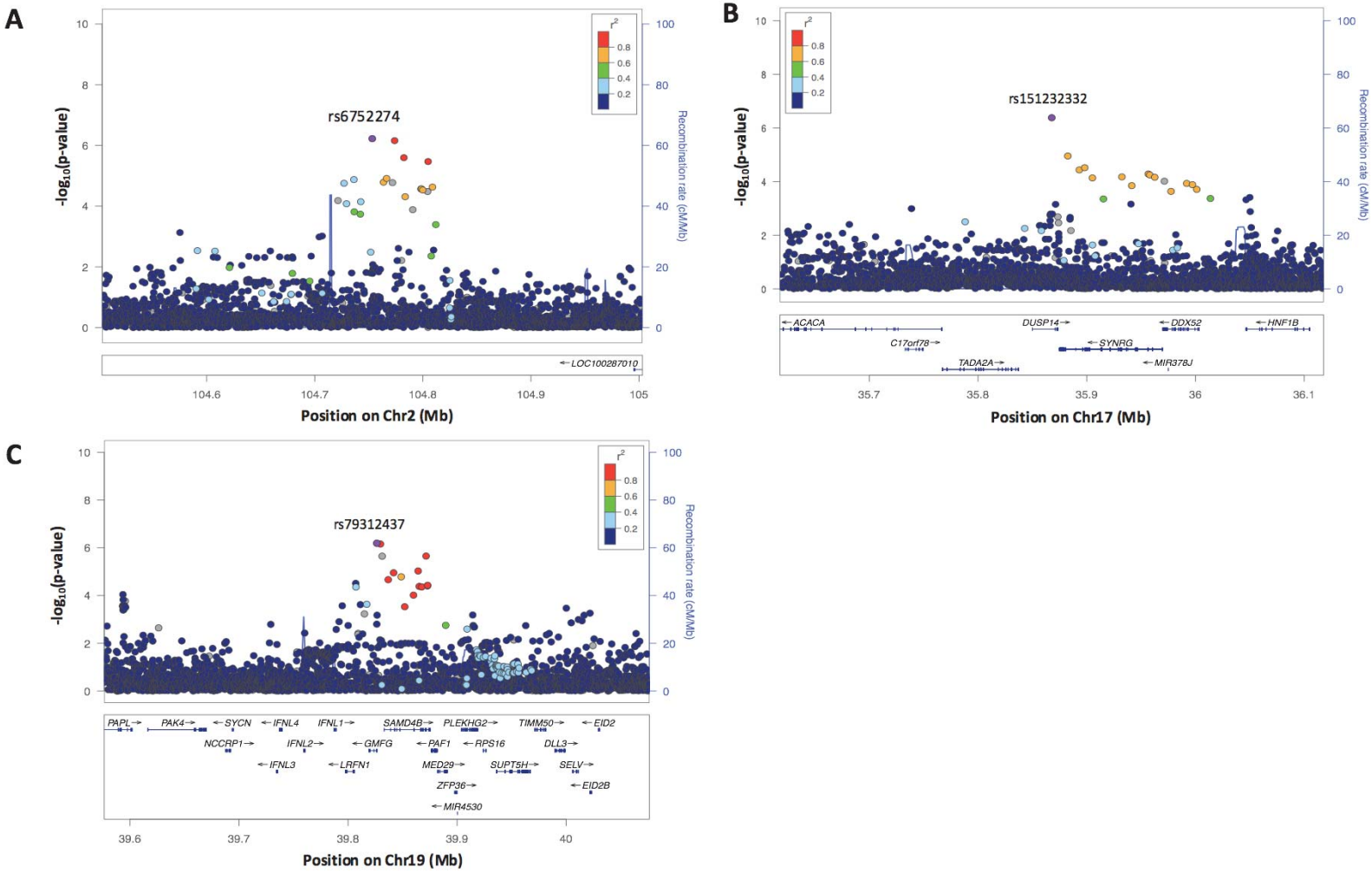


Fig. 4.10 Regional association plots for multivariate anti-KSHV IgG levels, N=4466, threshold=p<1x10^{-6}. A. Association on Chromosome 2 nearby *AC096554.1*. **B.** Association on Chromosome 17 in the *DUSP14* region. **C.** Association on Chromosome 19 in the *GMFG* region.

4.3.4 Associations with Previously Identified Candidate Variants in This Study

I assessed whether the thirty-one variants within 22 genetic loci that been previously identified with marginal significance as associated with classic KS, KSHV seropositivity, KSHV viral load, antibody response or PEL (Table 4.1) were present in this study and had plausible signals. In this study only 8/31 variants were typed/imputed in this study and were in *IL6*, *IL8RB*, *IL13*, *IL4*, *IL12A*, and *IRAK1.* None of the variants were statistically significantly associated with any of the traits in this study, with p values between 0.9 to 0.04 (Table 4.7).

Table 4.7 Associations with previously identified candidate variants

Gene	SNP	P_1 (OR)	$P_{UG.LANA}$ (β)	$P_{UG.K8.1}$ (β)
IL12A	rs568408	0.02 (2.4)	0.04 (-0.05)	0.04 (-0.05)
IL6	rs1800795	0.04 (N.R)	0.79 (-0.02)	0.79 (-0.02)
IL4	rs2243248	0.05 (2.8)	0.77 (0.01)	0.77 (0.01)
IL13	rs20541	0.01 (1.88)	0.80 (0.03)	0.30 (0.03)
IL8RB	rs1126579	0.003(0.49)	0.90 (0.05)	0.12 (0.05)
IL6	rs1800795	0.05 (3.7)	0.71 (-0.01)	0.71 (-0.01)
FCγRIIIA	rs396991	0.00028 (N.R)	0.10 (-0.04)	0.10 (-0.04)
IRAK1	rs1059702	N.R	0.18 (-0.05)	0.40 (-0.05)

P_1 – P-value from original study (Table 4.1)

$P_{UG.LANA}$– P-value from Uganda GPC anti-LANA IgG GWAS

$P_{UG.K8.1}$ – P-value from Uganda GPC anti-K8.1 IgG GWAS

4.4 Discussion

In this study, I assessed the host genetic contribution to anti-KSHV IgG responses in a rural African population cohort of >4000 individuals. Previously, no GWAS had been done for any KSHV phenotype and as sample size is limiting to conduct well powered GWASs for KSHV-associated diseases such as Kaposi's Sarcoma, IgG response traits provide a good intermediate phenotype, indicating the strength of the humoral immune response and control of infection; moreover, previous studies have shown correlation of IgG levels with the development of KS[163,525]. This cohort is KSHV endemic and both anti-KSHV IgG traits are partly heritable after accounting for shared environment, h^2=27.7% and 25% for anti-LANA and anti-K8.1 IgG levels, respectively (described in chapter 2, Fig. 2.11). However, no genome-wide significant SNPs were identified contributing to inter-individual variability in immune responses despite having >80% to detect signals of moderate to large effect sizes (β>0.15, for MAFs>25%) (Fig. 4.2). Nonetheless, using a less stringent threshold of $p<1x10^{-6}$, candidate loci were identified with suggestive associations for both quantitative traits, of which none outside the MHC locus have been previously implicated in KSHV pathogenesis, and are discussed below. In the Uganda GPC KSHV infection is nearly ubiquitous (>90%), with infection reportedly occurring early in childhood[146,174], and thus sero-negativity most likely reflect either a low immune response or lack of antibody detection owing to assay sensitivity issues as opposed to lack of exposure to KSHV. As this cohort is mainly consisting of adults (~99%), virtually everyone has been exposed to KSHV, and thus a binary analysis of traits would not have been as informative and was not performed.

The variants identified in these analyses were mapped to multiple promising candidate genes with modest effect sizes. However, these results would need to be replicated to ensure the signals are robust and fine-mapped to identify the causal variants; mapping close to a gene or within its intron does not mean necessarily that this is the effector transcript, in fact, the SNPs (or the causal variant once fine-mapped) could be affecting completely different genes. Besides the MHC locus none of the genes identified have

been previously implicated in KSHV infection or pathogenesis, thus, at this stage it is speculative that they potentially play roles in fundamental KSHV cellular processes including, viral entry, spread, and the role of T-cell immunity in viral control and evasion from host defence for lifelong survival. None of the previously identified candidate loci (Table 4.1 and Table 4.7) were significantly associated in this study, given the lower MAF threshold of 10% and my sample size, I would have had 80% power to detect sizes of 0.2, suggesting that the lack of significant results is either because of differences in the study design (e.g. quantitative antibody response measure here vs case-control) or because the previous results were false positive associations.

For anti-LANA IgG response levels reflecting history of infection, relatively strong associations were in *HLA-DQA1* (rs9273255, $p=5.19x10^{-7}$) and *PTPRN2* (rs71545585, $p=2.46x10^{-8}$) in chromosomes 6 and 7, respectively (Fig. 4.3, Fig. 4.4.A and Fig. 4.4.B) and associated with elevated antibody responses to infection. The HLA class II region has been implicated in the pathogenesis of KSHV[511], however, previous genetic association studies have failed to identify convincing associations (Table 4.1) and none of the previously identified SNPs were replicated in this study. Activation of CD4+ T-cells is particularly important for anti-KSHV immunity and *in vitro* studies have shown the CD4+ T cells can inhibit viral replication in KSHV-infected tonsillar B cells[252]. Like EBV, KSHV has evolved strategies to evade immune detection, including negatively regulating the process by which HLA class II molecules present antigens to CD4+ T-cells, thereby promoting its survival. LANA, like it's EBV homolog EBNA-1, is expressed in the immunologically silent stage of the KSHV life cycle and actively plays roles in modulating host innate and adaptive immune responses. Recently, LANA and another KSHV latent protein, v-IRF3 have been reported to inhibit MHC class II peptide presentation by blocking the transcription of the class II transactivator (CIITA), a master regulator of class II expression[526-528]; this is consistent with the association signal (β=0.15), an increase in anti-LANA IgG is associated with decreased HLA class II expression based on eQTL data from GTEx (see section 4.3.1). On chromosome 7, rs71545585-A in the *PTPRN2* gene was

associated with elevated antibody responses to LANA. *PTPRN2* encodes a protein tyrosine phosphatase receptor, type 2, and a recent study characterizing epigenetic variation in CD4+T cells showed hypomethylation of *PTPRN2* in Lupus patients of African-American descent resulting in increased risk of autoimmunity and other T-cell related diseases[529], suggesting a potential role in T-cell immunity. Other genes identified were: the long intergenic non coding RNA (lincRNA) genes *LINC01159* and *LINC00311* on chromosome 2 and 16 respectively; *SAMHD1* on chromosome 20 and *TMEM184b* on chromosome 22 (Fig. 4.4.C, and Fig. 4.5). LincRNA genes are transcribed by RNA polymerase II, however do not have the ability to code proteins. While the role of lincRNAs is still not well understood, they have been reported by a number of studies to play a role in various cellular process including innate immune regulation, T-cell development, differentiation and adaptation in adaptive immune responses and have been reported to interact with KSHV infected cells to modulate their functions[530,531]. The *SAMHD1* gene was first identified as a mouse orthologue expressed as a result INF-γ induction following a viral infection, and mutations in the gene are associated with Aicardi-Goutières syndrome, a rare early-onset genetic encephalopathy, characterised by dysregulated inflammatory responses with symptoms resembling a congenital viral infection[532]. More recently, SAMHD1 was found to restrict HIV-1 viral replication in dendritic cells, myeloid cells and resting CD4+ T cells[533-535]. While not previously implicated in KSHV pathogenesis, this potentially suggests a role for *SAMHD1* in modulating inflammatory responses. The potential role of *TMEM184b* that encodes an uncharacterised transmembrane protein[536], in KSHV pathogenesis is less clear. These findings suggest that variation in genes that play a role in modulating the immune response, particularly, T-cell immunity could contribute to inter-individual differences in anti-LANA IgG response levels to control KSHV viral infection, however, given lack of genome-wide significance, replication, fine-mapping and functional follow-up would be essential to confirm these hypotheses.

Candidate loci associated with anti-K8.1 IgG antibody levels representing active viral

infection and replication were also identified. The strongest association, rs183160271-A ($p=2.79x10^{-8}$, $\beta=0.35$) nearby the *DCAKD* gene on chromosome 17 (Fig. 4.6 and Fig. 4.7.B), was associated with elevated anti-K8.1 IgG responses. Very little is known about *DCAKD* which encodes Dephospo-CoA Kinase domain containing, expression has been associated with risk of Parkinson's Disease[537], however its putative role in KSHV pathogenesis is unclear. Also associated with high antibody responses is rs62422641-A ($p=3.22x10^{-7}$, $\beta=0.21$) in the *ARID1B* gene in chromosome 6 (Fig. 4.7.A). *ARID1B* is a chromatin remodeling gene and haploinsuffiency of *ARID1B* is causally associated with intellectually disability[538-540], more recently this gene has been reportedly associated with peripheral T-cell lymphoma (PTCL)[541,542], and in addition, somatic mutations have been associated with viral liver cancer[543]. Chromatin remodeling is an important aspect for maintaining KSHV stable gene expression and latency, studies have shown that chromatin modification of the KSHV ORF50 lytic master switch promotes reactivation from latency[544-546]. On chromosome 15, rs72738070-T ($p=2.63x10^{-7}$, $\beta=-0.73$) in *CGNL1* (Fig. 4.7.D) was associated with lowering anti-K8.1 IgG responses. *CGNL1* encodes for Cingulin-like 1, which regulates small GTPases RhoA and Rac1, that play a role in a variety of cellular processes including cytoskeleton organization, proliferation, differentiation cytoplasmic transport, endocytic vesicle trafficking and gene expression[547,548]. The Rho-GTPases have also been involved in facilitating KSHV viral entry in adherent cells, mediating nuclear translocation and viral cell to cell spread[549,550]. On chromosome 2, rs1005442-A ($p=3.92x10^{-7}$, $\beta=-0.26$) in the *TNR* gene (Fig. 4.7.C) is also associated with low antibody responses to K8.1. *TNR* encodes Tenascin R a member of the tenascin family of extracellular matrix (ECM) glycoproteins, while TNR has not been functionally implicated in KSHV infection, it has been reported that KSHV CD138 binds to components of the ECM including Tenascin to drive B-cell differentiation in pre-B and plasma cells[551]. These findings suggest genes that play a role in regulating KSHV biological function and cellular processes could contribute to inter-individual differences in anti-K8.1 IgG response levels, limiting viral replication to evade the host defence or facilitating it to promote spread.

Furthermore, multivariate GWAS combining anti-LANA and -K8.1 IgG responses, suggest the presence of variants with pleiotropic effects[459,552]. *HLA-DQA1* and *PTPRN2* previously identified as associated with anti-LANA IgG responses remained significant (Fig. 4.9 and Table 4.6). In addition, three new candidate loci were identified, *AC096554.1, DUSP14* and *GMFG* with marginally significant associations (Table 4.6 and Fig. 4.10). *AC096554.*1 is a long intergenic non-coding RNA gene whose role may be similar to that for *LINC01159* as described above. *DUSP14* belongs to the dual-specificity phostaphase family that deactivate kinases and has been found to inhibit *TNF*- and *IL1*- induced NF-κβ activation[553]. In KSHV infected cells, NF-κβ activation was found to promote latency and suppress viral reactivation[554,555]. *GMFG* is highly expressed in the thymus, spleen, T-lymphocytes, macrophages and fibroblasts, and it has been reported recently to be necessary for the migration and chemotaxis for T-cells which is associated with cellular adhesion[556]. Another study has also reported that high *GMFG* expression correlates with poor prognosis of ovarian cancer by promoting cell migration and invasion[557].

The fact that no genome-wide significant associations were discovered despite having >80% power to detect common variants of moderate to large effect sites (β>0.15) (Fig. 4.2), in addition to having greater heritability (h^2LANA= 27%, h^2K8.1=25%) (chapter 2) in comparison to EBV traits (h^2EBNA= 11%, h^2K8.1=7.7%), whereby GWAS in ~1500 individuals from the same cohort had <5% power to detect common variants of the same effect sizes (β>0.15) revealed strong associations in the MHC region ($p<1x10^{-9}$) (see chapter 3, Fig. 3.2), suggests differences in underlying genetic architectures between the traits. It is also possible that variants with very small effect sizes or rare genetic variants (i.e. EAF<0.5%) influence inter-individual variability in KSHV immune responses, however, this study is under-powered to detect low-frequency variants with small effect sizes and the genetic data used here, nor the tools used are optimal for accurate rare variant detection.

Another possibility is that differences in study design could contribute to the lack of significant findings. For example, a major difference between the EBV GWAS design and this GWAS is the serological assay used for antibody detection and quantification. Here, the ELISA assay was used to quantify responses to the recombinant LANA and K8.1 antigens with 89% and 98% sensitivity, respectively, and specificity of >98%[558]. While, the ELISA has been reported to produce reliable and reproducible results compared to other assays such as the immunofluorescence assay (IFA), the narrow dynamic range (see chapter 2, section 2.3.2, Fig. 2.5 and Fig. 2.7) might present a drawback for this study, limiting the use of ODs as a marker for antibody levels, which could lead to loss of power for quantitative trait GWAS. Very recently, Labo and colleagues developed a multiplex bead-based assay that can simultaneously detect six antigens and enable more complete characterisation of KSHV immune responses[239]. In addition, they reported several advantages over the ELISA, specifically a wider dynamic range (mean fluorescence intensity up to 100,000 compared to OD<4) which decreases the need for the dilution of samples with very high antibody levels. In addition, the multiplex assay can also be advantageous as it measures multiple antigens that can be used for performing a multivariate GWAS of correlated quantitative traits to boost discovery power.

In summary, I describe the first GWAS performed for KSHV traits and in >4000 individuals the largest KSHV host genetic study, leveraging the combination of whole-genome sequencing and imputing genotypes to a panel with additional African sequence data to aid discovery of novel candidate loci. The availability of data on environmental covariates such as co-infection with other pathogens allows for the capture genetic variation independently of the environment. As Uganda is also a Malaria endemic area and studies have reported co-infection with *P.falciparum* as having an influence on antibody titre[179-181,243], it might be useful to also incorporate this in future study design. I identified seven putative candidate gene regions associated with anti-LANA IgG response levels, five putative candidate gene regions associated with anti-K8.1 IgG levels and an additional three candidate regions were identified following multivariate analysis of both

phenotypes. Together, the findings suggest common variants of moderate effect sizes in multiple genes are involved in modulating important biological pathways to control KSHV infection and manipulate the immune system. While none of the findings reach the more stringent genome-wide significance threshold ($p<5x10^{-9}$), variants in *PTPRN2* and *DCAKD* had $p<5x10^{-8}$, which is the widely accepted threshold for European ancestry GWAS. While replication of these findings is crucial, future studies will need to address the possible limitation in this study by using optimised antibody detection assays that maximise the dynamic range, such as the multiplex bead-based assay, which also allows the detection of multiple antigens. To follow up significant GWAS findings, replication of novel loci is essential, in addition pathway analysis tools have been developed and could be used to reliably identify gene enrichments in pathways and protein interaction networks. However, as most rely on genetic data provided by HapMap or 1000 Genomes and have been designed for predominantly European ancestry data, might not be optimal to analyse the African genotype/sequence data used here. Lastly, fine-mapping to refine casual variants and functional validation will be key to understand how the variants identified modulate gene expression and fundamental KSHV biological processes.

5 Chapter 5: Characterizing the Genetic Diversity of KSHV in The Uganda GPC

5.1 Introduction

The understanding of sequence diversity of KSHV at the whole-genome level and its relation to pathogenesis and disease is limited. In particular, if and how genetic variation in virus genes contributes to the control of KSHV infection and the potential development of disease is not very well understood. The first KSHV genomes to be sequenced used Sanger sequencing technology and were all derived from tumours or cell lines using cosmid clones or DNA inserted into bacterial artificial chromosome (BAC) technology. The large dsDNA ~165 kb genome sequence of KSHV was first determined by Russo and colleagues in 1996[138]. Using phage and cosmid libraries from the KSHV infected primary effusion lymphoma (PEL) cell line, BC-1, they revealed a long unique coding region (LUR) of ~140kb with 81 ORFs some of which had functional homologues in *herpesvirus samiri* (HVS), a related non-human primate gamma-herpesvirus, and 801 bp long terminal repeats (TR) at the ends of the genome[138]. Following this, a nearly complete genome was sequenced using shotgun sequencing of fragments following partial digestion of DNA isolated from AIDS-associated-KS biopsies[559]. The first complete and most extensively annotated KSHV genome sequence, GK18 was isolated from a Greek individual with classic KS and sequenced from a cosmid clone[560]. Two additional genomes were generated from KSHV-infected PEL cell lines isolated in the USA: JSC-1 and BCBL-1 were sequenced from DNA cloned into a BAC16 and BAC36 cassette, respectively[138,561,562]. A more recently sequenced whole-genome, DG-1, was sequenced using Illumina sequencing technology (unlike the five genomes described above) and was the first genome to be generated from actively replicating virus isolated from blood plasma of a patient infected with KSHV inflammatory cytokine syndrome (KICS) and also co-infected with HHV-6[563]. A very recent study, conducted by Olp and colleagues, used Illumina paired-end SureSelect deep-sequencing, and generated 16 whole-genomes

directly from skin lesions of Zambian KS patients, actively enriching for viral DNA, this represents the first non-Western genomes to be generated from sub-Saharan Africa, and a KSHV and KS endemic region[564].

The KSHV genome map has changed little since its discovery, with the annotation of the GK18 sequence revealing 86 genes of which 22 encode putative immunomodulatory proteins (see chapter 1 Fig. 1.5). Extensive transcriptional and proteomic profiling has increased our understanding of KSHV gene expression at different stages in its lifecycle and improved annotations of non-coding regions[204,565,566].

The KSHV genome displays high conservation with 99% sequence similarity between viral strains, however both 5' and 3' ends of the genome have displayed high sequence variability and as such have been used to characterize viral strains[187]. ORF K1 located at the 5' termini of the genome (see chapter 1, Fig. 1.5) encodes highly glycosylated transmembrane proteins, with hypervariable regions (V1 and V2) that display up to 30% amino acid variability, such that K1 variants are used for viral genotyping[186]. Seven major KSHV subtypes, A-E and more recently F and Z have been identified by K1 subtyping (and phylogenetics) and display considerable geographic variation[187,567,568]. The P (Predominant), M (Minor) and N alleles from the K15 locus at the 3' termini of the genome (see chapter 1, Fig. 1.5), that encodes an integral membrane protein with up to 70% inter-allele divergence at the amino acid level, has also been used to characterize viral variants[192,569,570]. While the central region of the KSHV genome is highly conserved, nine discrete loci: K12, K2, K3 ORF18/19, ORF26, K8, ORF73 and two loci within ORF75 (which make up ~5.6% of the genome) with low level variation have also been used in a number of phylogenetic studies for characterisation of subtypes[192]. For example, ORF26, encodes a minor capsid protein and has been used for variant characterisation to identify KSHV subtypes A/C, J, K/M, D/E, B, Q, R or N which are also heterogeneously distributed amongst different populations although distribution is found to parallel K1 subtypes[571]. The remaining >90% of the genome has not been taken into account due to lack of high

coverage whole-genome sequencing data. While the early sequencing studies provided insights into KSHV genome architecture, the variability of the K1 and K15 genes and the coding capacity of its genome, the Zambian KS whole-genome study is the first to show that low level genetic variation in the central conserved region contributes to a unique phylogenetic structure; showing distinct genomic variants of Zambian isolates compared to Western (USA and Greece) isolates, and therefore, are indeed important for accurate viral characterization[564]. With a small number of genomes from only three countries (USA, Greece and Zambia) and all from diseased individuals, whether genomic diversity contributes to the distribution of KSHV seroprevalence and incidence of associated diseases remains unclear.

Uganda presents a good candidate to study KSHV molecular epidemiology and phylogeography as it is inhabited by different ethno-linguistic groups with divergent historic origins as a result of migration over several hundred years from surrounding regions[391,572,573], and in addition the population sustains the highest seroprevalence of KSHV in the world[150,574,575]. In the GPC, the seroprevalence of KSHV is >90% (see chapters 2 and 4). Several studies conducted in Uganda have provided invaluable insights into KSHV seroepidemiology and transmission[145,146,378,382,385,576], therefore, characterising genetic diversity on a whole-genome scale will bridge gaps in the understanding of the co-evolution of host and virus and its implications in disease pathogenesis.

5.1.1 Chapter Aims

With only 21 full genome sequences of KSHV published to date from three different countries and all isolated from KSHV associated diseases, the understanding of KSHV genomic diversity in relation to disease pathogenesis is not clear. No genomes have been isolated from asymptomatic persistently infected individuals and thus the 'wild-type', non-tumour associated KSHV genome has never been characterised. Therefore, the aims of this chapter are to:

I. Generate whole-genome sequences of KSHV isolated from saliva of KSHV disease free individuals and assess the variability between KSHV genomes isolated from different sources and of diseased individuals.
II. Assess the population structure of KSHV strains within the Uganda GPC and in a global context.

Contributions

Sample collection, storage and shipment was conducted by the GPC team in Uganda. RNA bait libraries for target enrichment were developed by the Virus genomics team. Whole-genome sequencing was conducted by the Illumina high-throughput sequencing team at Sanger. All other analyses unless otherwise stated were performed by myself.

5.2 Methods

5.2.1 Sample Selection and Collection

The Uganda GPC and ethics are described in detail in chapter 2. Briefly, the GPC is a population-based cohort in rural south west Uganda consisting of 25 neighbouring villages mainly inhabited by peasant farmers who grow bananas as a subsistence crop, cultivate coffee for trade and also raise livestock[388,389]. Household are scattered with some concentrated in the trading centres. The Baganda are the predominant tribal group constituting ~70% of the population with a substantial number of migrants who settled from neighbouring Rwanda. To assess the determinants of viral shedding and diversity of KSHV whole-genomes, 2036 saliva samples were collected from individuals during medical survey round 24 between January to July 2015. 2ml of saliva was collected with the Oragene® DNA self-collection kit, OMNIgene®.ORAL OM-505 (DNA Genotek Inc., ON, Canada) following manufacturer's instructions by the GPC team in Uganda and stored at -80^{o}C prior to shipment on dry ice to the Sanger Institute.

5.2.2 DNA Extraction, Purification and Quantification

I conducted all sample preparation in class II biosafety cabinets using aseptic techniques. Saliva samples were transferred to Corning™ Costar™ 96 well plates (ThermoFisher scientific, UK) in 1 ml aliquots, and lysed. RNA was removed with proteinase K (600mAU/ml) Buffer VXL solution and RNase A (100mg/ml) treatment (Qiagen™, UK). 200 ul aliquots of lysates were then aliquoted into 96-well S-blocks (Qiagen™, UK) for DNA extraction using the QIAamp 96 DNA QIAcube®HT robot following the manufacturer's protocol, and the remainder stored at -80^{o}C. Briefly, 96 samples were processed simultaneously, optimal DNA binding and filtering of contaminants were achieved through buffering steps followed by ethanol wash steps to remove residual contamination and enzyme inhibitors. DNA was eluted in buffer AE under vacuum and then enhanced by overlaying elution buffer TE fluid to achieve a final volume of 80 ul of

DNA per sample. Quantification of total genomic DNA was performed using the Quanti-iT™ PicoGreen® dsDNA assay kit (ThermoFisher scientific, UK).

5.2.3 Quantitative PCR for Viral DNA Detection

I used quantitative PCR (qPCR) for viral genome detection and determination of viral genome load measured by determining the viral copy number relative to a control PEL DNA sample. Out of the 2000 samples, 746 were processed in duplicates using the QuantiTect Muliplex PCR kit (Qiagen, UK) on a Stratagene Mx3005P (Agilent Technologies, UK) owing to time-constraints. Primers and probes targeting KSHV *ORF73* were designed for viral detection using sequences from Lallemand *et al*.,[577] (Table 5.1). GAPDH was used as a normalizing assay with sequences from Pardieu *et al*,[578] (Table 5.1). All primers and probes were synthesised by Metabion international AG, Germany. Primer-probe mixes were diluted to a 20X solution, for KSHV this consisted of 10 pmol/ul of each primer and 1.25pmol/ul for the probe; for GAPDH 2.5pmol/ul for each primer and 1.25 pmol/ul for the probe for each reaction. The master mix for qPCR in a 25 ul/reaction was as follows: 1.25 ul of KSHV primer-probe 20X mix, 2 ul GAPDH primer-probe 20X mix, 12.5 ul QuantiTect multiplex mastermix, 4.25 ul nuclease-free water and 5 ul of DNA. DNA from the BCBL-1 cell line was used as a positive control and used to generate a standard curve with 10-fold serial dilutions from $3x10^6$ to 30. The qPCR conditions were as follows: Initial denaturation at 95^oC for 15 mins and 45 cycles of denaturation at 95^oC for 15s and annealing at 60^oC for 1min. The fluorescence data was captured during the annealing step. The Ct values were compared to the standard curve to assign a copy number per ml. Data analysis was performed using MxPro v4.10 qPCR software (Agilent Technologies).

Table 5.1 qPCR Primer and probe sequences

Primer/Probe	Sequence (5' -> 3')
ORF73 - Forward	TTGCCACCCACGCAGTCT
ORF73 - Reverse	GGACGCATAGGTGTTGAAGAGTCT
ORF73 - Probe	6-FAM-TCTTCTCAAAGGCCACCGCTTTCAAGTC-TAMRA
GAPDH- Forward	GGCTGAGAACGGGAAGCTT
GAPDH - Reverse	AGGGATCTCGCTCCTGGAA
GAPDH - Probe	HEX-TCATCAATGGAAATCCCATCACCA-BHQ-2

5.2.4 KSHV Whole-Genome Sequencing

I selected 240 samples for whole-genome sequencing based on viral DNA detection (Ct values <36) following qPCR. For KSHV target enrichment, overlapping 120mer RNA baits spanning the length of KSHV GK18 and BC1 reference sequences (Accession numbers: NC_00933 and NC_003409, KSHV Type P and Type M respectively) were designed by Matt Cotten using eArray software (Agilent Technologies). The Illumina High Throughput sequencing team at the Wellcome Trust Sanger Institute performed target enrichment and genome sequencing following the SureSelect™ protocol (version 1.1). Briefly, 1-3ug of each DNA sample was sheared to 200-500bp fragments followed by end-repair, non-template addition of 3'-A, adaptor ligation, hybridisation, enrichment PCR, index tagging and sample pooling. Samples were multiplexed on an 8 lane flow cell with 24 samples per lane, cluster generation and sequencing was performed on an Illumina HiSeq 2000 sequencer. Sequencing reads were 250bp paired-ends in FASTQ format with per base Phred quality scores.

5.2.5 Guided Assembly of KSHV Whole-Genomes

I used the QUASR QC pipeline (http://sourceforge.net/projects/quasr)[579] to retain high quality full length reads. Duplicate reads and paired reads with a raw median Phred quality score Q<32 were either filtered out or trimmed from the 3' end until Q>32. Any

reads less then 100bp in length post trimming were also excluded. High quality paired-end reads post-QC were then mapped back to GK18 and BC1 reference sequences using Burrows-Wheeler Aligner (BWA)[580] and the average depth and coverage calculated using SAMTools[581] with an in-house script written by Anne Palser. To investigate whether viral load influenced sequencing quality, I generated a pairwise-correlation matrix for qPCR viral load, KSHV mapped reads (%) and sequencing depth of coverage using Pearson's correlation in R.

5.2.6 Comparative and Phylogenetic Sequence Analysis

For comparative and phylogenetic analysis I selected 62 consensus sequences with an average sequencing depth of at least 10x and coverage of >90% across the genome, generated following BWA mapping and aligned them with 21 publically available KSHV Genomes from Greece, USA and Zambia using MAFFT[582] (v7.0) and viewed using AliView software. I masked repeat regions across the alignment with coordinates retrieved from the GK18 reference sequence annotation in Genbank (). SNPs between genomes were counted relative to consensus sequence generated from the multiple sequence alignment and genome-wide mutations were visualised in a 1000 nucleotide scanning window and the number of codon changes per gene (synonymous and non-synonymous) were calculated using an in-house script written by Simon Watson. I then used the multiple sequence alignment to generate whole-genome trees using maximum-likelihood methods implemented in RAxML (v8) with the general time reversible (GTR) model of nucleotide substitution including a Gamma distribution for among site rate variation[583]. Tree topology was assessed using 1000 bootstrap replicates in RAxML. I also generated whole-genome trees removing the K1 and K15 variable genes. To investigate genotypic diversity, I performed phylogenetic analysis following alignment of the coding sequences of the K15 gene and K1 gene along with representative sequences for the following genotypes (Genbank accession number): A1 (AF133038), A2 (AF130305), A3 (U86667), A4 (AF133039), A5 (AF178823), B1 (AF133040), B2 (AY042947), B3

(AY042941), B4 (DQ309754), C1 (AF133041), C3 (AF133042), D1 (AF133043), D2 (AF133043), E (AF220292) and F (FJ884616). All trees were midpoint rooted.

An overview of the workflow used in this chapter from sample collection to analysis is presented in Fig. 5.1.

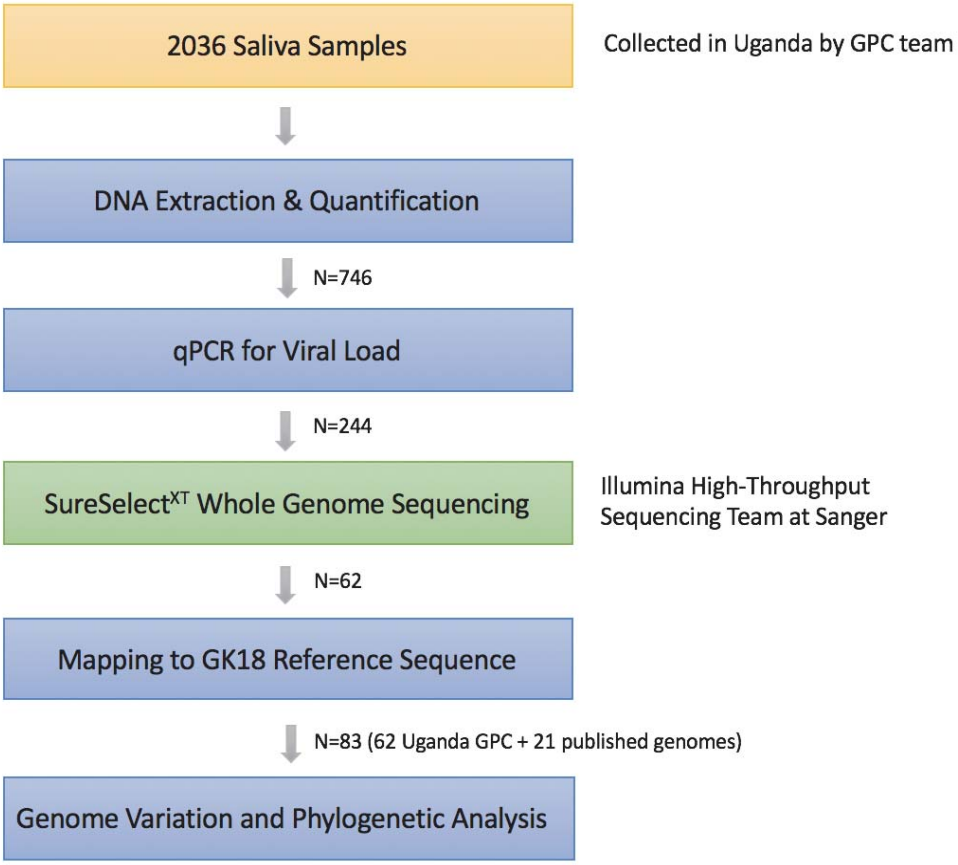


Fig. 5.1 KSHV genome analysis workflow. N represents the number of samples processed at each stage of the analyses.

5.3 Results

5.3.1 KSHV Shedding and Viral Load in the GPC

In this study, 2036 saliva samples were collected from asymptomatic individuals in a cluster of neighbouring villages in the GPC (previously described in chapter 2, Fig. 2.1) from January to July 2015 to characterise and assess the genetic diversity of wild-type KSHV whole-genome sequence. Following DNA extraction, I screened 746 randomly selected samples for KSHV positivity and viral load based on qPCR targeting the ORF73 gene and using a ten-fold dilution of BCBL-1 DNA to ascertain viral copy numbers against a standard curve with a detection range of $3x10^6$ – 10 copies/ml (C.t of 15 to 43) (Fig. 5.2). While most of the individuals in the GPC (>90%) are seropositive to KSHV (see chapter 2 and 4), qPCR positivity reflects individuals shedding virus and therefore with viral DNA in saliva, given that viral DNA is only detectable during the lytic stage of infection. Following qPCR, 244 (32.8%) individuals had detectable KSHV viral DNA above the qPCR threshold with C.t values ranging from 21.55 to 41.5 representing a viral load from $5.35x10^5$ to 1.5 copies/ml (mean=9186 copies/ml), respectively. Out of the 244 samples, 80 (32.7%) samples had viral loads from $>10^4$ – 10^5 (C.t <31), 70 (28.6 %) samples had viral loads from 10^2 – 10^4 (Ct>31-35) and 94 samples (38.7%) had viral loads from <10 – 10^2 (Ct >35).

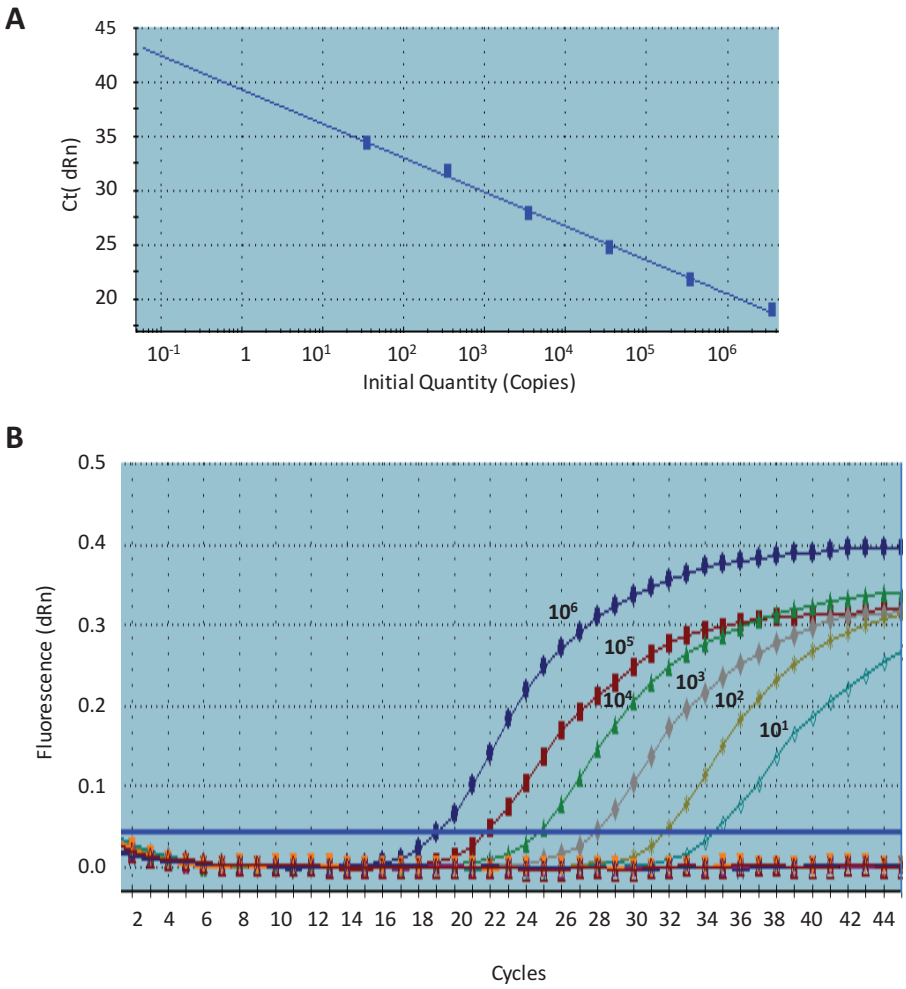


Fig. 5.2 KSHV ORF73 gene qPCR for BCBL-1 DNA dilution series from 30 to 3x10^6 viral copies/ml. A. Standard Curve. B. Amplification plot. Cycles represent cycles of PCR amplification and the blue solid line is the detection threshold for the FAM fluorescence channel. The amplification of standards is represented accordingly, all negative control samples (H_20) are below the threshold.

5.3.2 KSHV Viral Load Correlates with Whole-Genome Sequencing Quality

The low abundance of viral DNA compared to the host DNA in addition to the large KSHV genome, makes sequencing of KSHV quite the challenge, thus, to maximise the chances of success all 244 samples with detectable viral DNA were submitted for whole-genome sequencing using the SureSelect^{XT} method to allow selective capture of KSHV DNA using custom biotinylated RNA baits, the success of this approach has been demonstrated using cell lines and clinical samples by other studies[584]. Following this, all samples were mapped to the KSHV GK18 reference sequences for type P and type M KSHV strains (NC_009333 and NC_003409 respectively) using BWA. The GK18 sequence was selected as it was used for the SureSelect RNA bait library design and also represents the most comprehensively annotated whole-genome sequence.

To assess whether viral load influenced the number of KSHV reads that were mapped and the mean sequencing depth of coverage >90% and thus generated better quality reads, I generated a correlation matrix calculated using Pearson's correlation in R. This confirmed that qPCR viral load were strongly positively correlated with the percentage of mapped reads (r^2=0.84) and also showed that a high viral load was positively correlated with achieving good mean sequencing depth with at least 90% coverage across the genome (Fig. 5.3). Out of the 244 samples, 62 (25.4%) had at least a 10x mean sequencing depth of >90% coverage across the genome. Of these 62 samples, depth ranged between 10x-1000x and 55% of samples had an average 25x depth across the genome. As these 62 represent the samples that also have a higher percentage of KSHV mapped reads with greater accuracy (up to ~77% mapped reads) and thus, better genome quality, I used this sample set for downstream analysis. The 62 samples corresponded to the samples with the highest viral loads (10^4 – 10^5 copies/ml) and were collected from 8 neighbouring villages (12-19) in the GPC (Fig. 5.4); they consisted of 32 males and 30 females between the ages of 16-91 (Mean ± S.D = 41.65 ± 20.69), 5 individuals were also HIV positive.

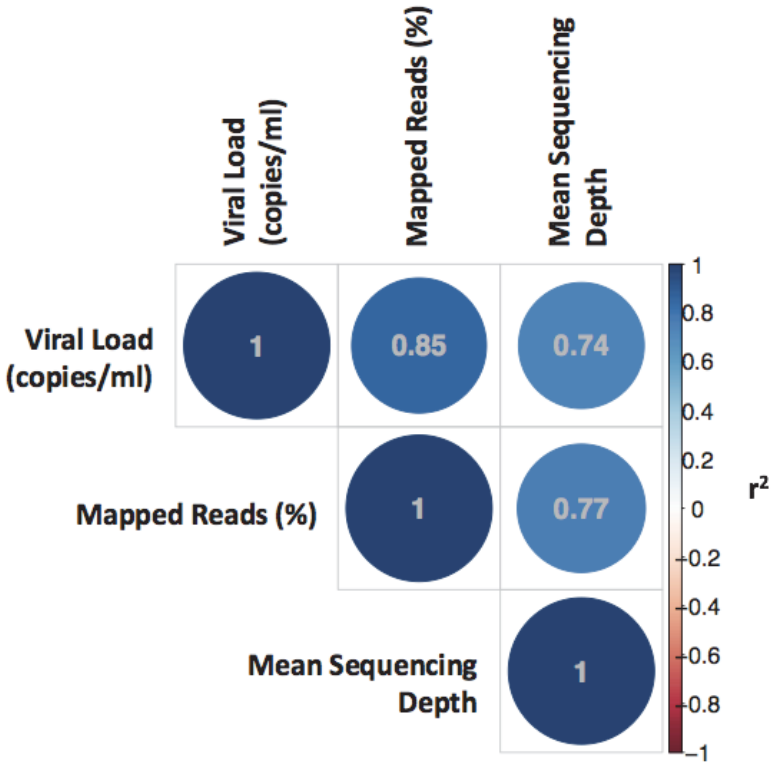


Fig. 5.3 Correlation matrix of Viral load (copies/ml), KSHV mapped reads (%) and mean sequencing depth of 200x. Correlating viral load as estimated by qPCR in copies/ml with % KSHV mapped reads and with a mean depth of 200X at >=90% coverage. Positive correlations are in blue, negative correlations are in red, intensity and size of the circle are proportional to the correlation coefficients (r^2) labelled in the circles and indicated on the right hand side of the correlogram. All tests meet Pearson's significance threshold of $p<0.01$.

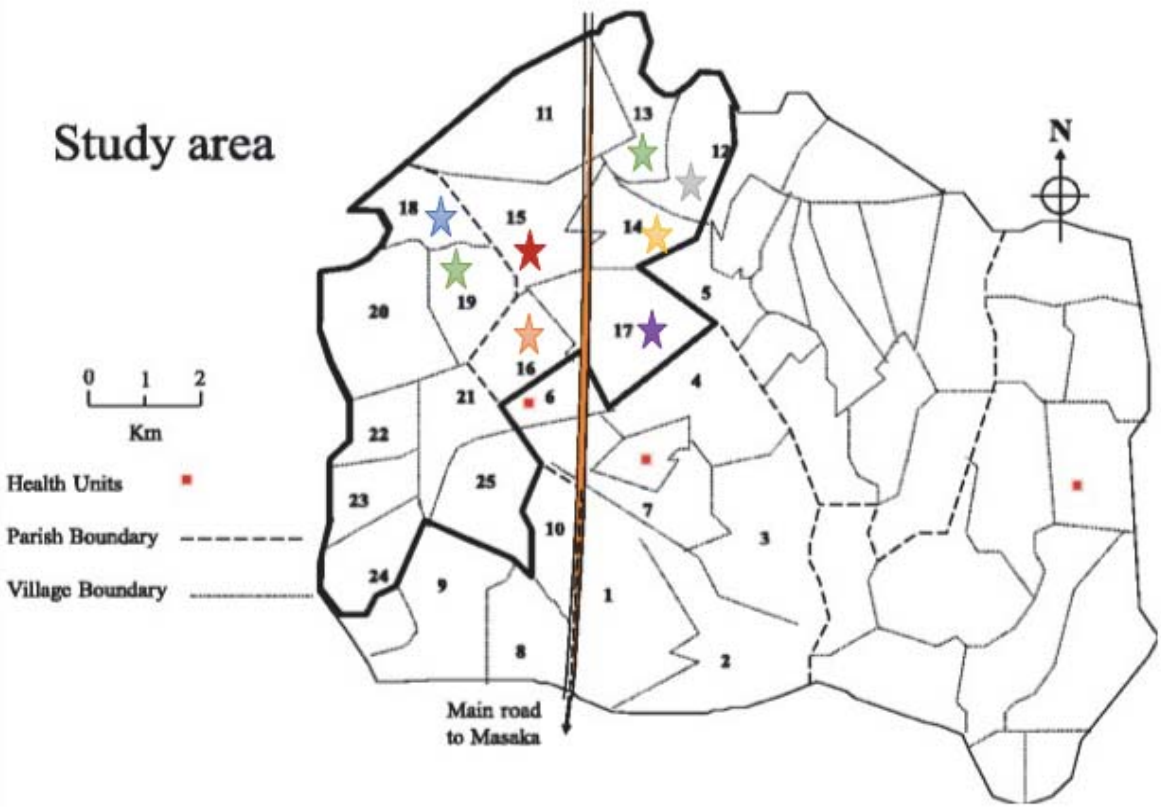


Fig. 5.4 Map showing GPC the study area in Uganda. The stars correspond to the 8 villages (12-19) where the 62 samples were collected from and the colours represent the number of samples in each village (Grey=2, blue=3, purple = 5, green=6, yellow=9, orange=13 and red=19).

5.3.3 KSHV Genome Variability

To determine how variable the 62 new saliva (wild-type) KSHV genomes were and explore which parts of the genome were contributing to the most variation, I performed a multiple sequence alignment including the 21 previously published KSHV genome sequences from Greece, USA and Zambia (Table 5.2) and used the consensus sequence generated from the alignment as reference for single nucleotide polymorphism (SNP) calling. Genes were annotated, and gaps and repeat regions were masked based on the GK18 sequence coordinates retrieved from Genbank (Table 5.3).

Table 5.2 Summary of KSHV samples used in this study

Sample Name	No.	Geographic Origin	Clinical Presentation	Source
GK18[560]	1	Greece	KS	KS Tumour
BCBL1[562]	1	USA	PEL	B Cell line
JSC1[561]	1	USA	PEL (EBV+)	B Cell line
BC1[138]	1	USA	PEL (EBV+)	B Cell line
DG1[563]	1	USA	KICS (HHV-6+)	Blood
ZM*[564]	16	Zambia	KS	KS Tumour
UG*	62	Uganda	Asymptomatic carrier	Saliva

KS-Kaposi's Sarcoma, PEL – Primary effusion lymphoma, KICS- KSHV inflammatory cytokine syndrome

Table 5.3 Eighty-four annotated KSHV genes based on the GK18 sequence

Gene	Start-Stop[a]	Timing[b]	Cycle[b]	Putative Function[b]
K1	105-944	Latent	Latent	Glycoprotein
ORF4	1112-2764	24h	Lytic	Complement binding protein
ORF6	3179-6577	24-48h	Lytic	ssDNA Binding Protein
ORF7	6594-8681	N.R	N.R	Virion Protein
ORF8	8665-11202	48-72h	Lytic	Glycoprotein B
ORF9	11329-14367	48-72h	Lytic	DNA Polymerase
ORF10	14485-15741	48-72h	Lytic	Regulator of Interferon Function
ORF11	15756-16979	8h	Lytic	Predicted UTPase
K2	17227-17841	Latent	Latent	vIL6 homolog
ORF2	17887-18519	N.R	N.R	Dihydrofolate Reductase
K3	18574-19542	24h	Lytic	Immune Modulator
ORF70	20023-21036	N.R	N.R	
K4	21480-21764	8h	Lytic	vMIP-II
K5	25865-26635	8h	Lytic	RING-CH E3 Ubiquitin Ligase
K6	27289-27576	8h	Lytic	vMIP-IA
K7	28774-29154	8h	Lytic	Late gene expression (overlaps with PAN)
ORF16	30242-30769	8h	Lytic	Bcl2 Homolog
ORF17	30920-32524	48h	Lytic	Protease
ORF18	32523-33296	24h	Lytic	Late gene regulation
ORF19	33293-34942			
ORF20	34710-35483			
ORF21	35482-37224	48-72h	Lytic	Thymidine Kinase
ORF22	37212-39404	48-72h	Lytic	Glycoprotein H
ORF23	39401-40615	48-72h	Lytic	Glycoprotein(predicted)
ORF24	40619-42877	48-72h	Lytic	Essential for replication (MHV68)
ORF25	42876-47006	48-72h	Lytic	Major Capsid Protein
ORF26	47032-47949	48-72h	Lytic	Minor Capsid Protein
ORF27	47973-48845	48-72h	Lytic	Glycoprotein (MHV68)
ORF28	49091-49399	48-72h	Lytic	BDLF3 EBV Homolog
ORF29	49462-50604, 53855-54775	72h	Lytic	Packaging Protein
ORF30	50723-50956	48-72h	Lytic	Late Gene Regulation (MHV68)
ORF31	50953-51537	48-72h	Lytic	Nuclear and Cytoplasmic (MHV68)
ORF32	51504-52868	48-72h	Lytic	Tegument Protein
ORF33	52861-53865	48-72h	Lytic	Tegument Protein (MHV68)
ORF34	54774-55757	24-48h	Lytic	N/A
ORF35	55738-56190	24-48h	Lytic	N/A
ORF36	56075-57409	24-48h	Lytic	Serine Protein Kinase
ORF37	57372-58832	24-48h	Lytic	Sox
ORF38	58787-58972	24-48h	Lytic	Myristylated Protein
ORF39	59072-60274	24-48h	Lytic	Glycoprotein M
ORF40	60407-61756, 61884-62543	48-72h	Lytic	Helicase Primase
ORF42	62535-63371	48-72h	Lytic	Tegument Protein
ORF43	63235-65052	48-72h	Lytic	Portal Protein (capsid)
ORF44	64991-67357	48-72h	Lytic	Helicase
ORF45	67452-68675	8h	Lytic	RSK activator
ORF46	68736-69503	24h	Lytic	Uracil deglycosylase

ORF47	69511-70014	24h	Lytic	Glycoprotein L
ORF48	70272-71480			N/A
ORF50	71695-71712, 72671-74728	8-24h	Lytic	RTA
ORF49	71729-72637			Activates JNK/p38
K8	74949-75662, 75744-75890	8-24h	Lytic	bZIP
K8.1	76014-76437, 76532-76794	48h	Lytic	Glycoprotein
ORF52	76901-77296	48-72h	Lytic	Tegument protein
ORF53	77432-77764	48-72h	Lytic	Glycoprotein N
ORF54	77835-78722	48-72h	Lytic	dUTPase/Immunomodulator
ORF55	78864-79547	48-72h	Lytic	Tegument Protein
ORF56	79535-82066	48-72h	Lytic	DNA Replication
ORF57	82169-82217, 82326-83644	8h	Lytic	mRNA Export/Splicing
vIRF-1	83960-85309	48-72h	Lytic	K9
vIRF-4	86174-88442, 88544-89010	48-72h	Lytic	
vIRF-3	89700-90945, 91042-91496	48-72h	Lytic	
vIRF-2	92066-93620, 93742-94229	48-72h	Lytic	
ORF58	94577-95650	24h	Lytic	N/A
ORF59	95655-96845	24h	Lytic	Processivity Factor
ORF60	96976-97893	24-48h	Lytic	Ribonucleoprotein Reductase
ORF61	97922-100300	24-48h	Lytic	Ribonucleoprotein Reductase
ORF62	100305-101300	72h	Lytic	N/A
ORF63	101314-104100	N.R	N.R	NLR Homolog
ORF64	104106-112013	N.R	N.R	Deubiquitinase
ORF65	112037-112549	48-72h	Lytic	Capsid
ORF66	112576-113865	48-72h	Lytic	Capsid
ORF67	113799-114614	48-72h	Lytic	Nuclear Egress Complex
ORF67A	114669-114911	48-72h	Lytic	N/A
ORF68	115108-116511	48-72h	Lytic	Glycoprotein
ORF69	116544-117452	48-72h	Lytic	BRLF2 Nuclear Egress
K12	118025-118207	Latent	Latent	Kaposin
ORF71	122393-122959	Latent	Latent	vFLIP
ORF72	123042-123815	Latent	Latent	vCyclin
ORF73	124057-127446	Latent	Latent	LANA
K14	128264-129079	24-48h	Lytic	vOX2
ORF74	129520-130548	24-48h	Lytic	vGPCR
ORF75	130699-134589	48-72h	Lytic	FGARAT
K15	134824- 136899	N.R	N.R	LAMP

[a] Genomic position from Genbank (GK18)
[b] Annotation from Arias et al, 2014[566]
N.R= Not reported

To get an overview of genomic variation across all 83 genomes, the proportion of variants to the consensus sequence were determined within a 1000 nucleotide sliding window. This showed a high proportion of nucleotide changes at the left termini of the genome (~35%), which corresponds to the K1 gene and at the right termini of the genome (~40%) corresponding to the K15 gene with modest variation observed across the central regions of the genome (Fig. 5.5). To further resolve the variation in the 84 individual genes across the genome, I calculated the number of synonymous and non-synonymous changes for each KSHV gene normalising for gene length (Fig. 5.6). This confirmed the presence of the highest variation in the K1 and K15 genes with a total of 39.1% and 39.6% base changes respectively. Other genes with high levels of non-synonymous changes include ORF73 (13.7%) and K12 (11.5%). The K1 gene had the highest ratio of non-synonymous to synonymous (dN/dS) changes of 7.8 and the K15 gene had a dN/dS ratio of 3.6. Other genes with a high dN/dS ratio (>=2) include, ORF73, K4.2, K8.1, ORF34, v-IRF2, ORF38, ORF67A with dN/dS ratios of 3.2, 2.8, 3.0, 2.2, 2.1, 2.0 and 2.0 respectively suggesting that these genes are evolving under selection (Fig. 5.7).

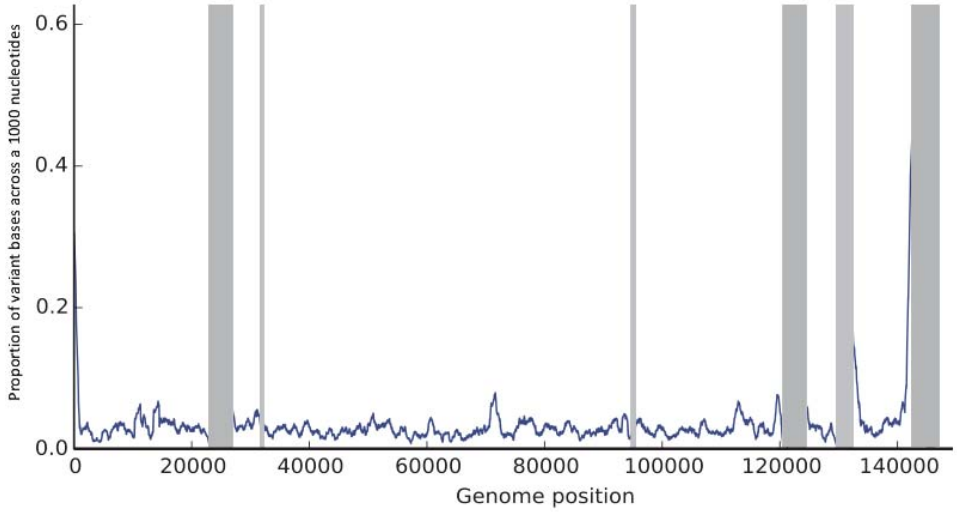


Fig. 5.5 Genome variability of 83 KSHV genomes. Line graph plotted across the genome showing the proportion of variant bases in a 1000 nucleotide sliding window where at least one KSHV genome sequence has a SNP relative to the consensus sequence generated from the alignment. Grey bars: Masked repeat regions.

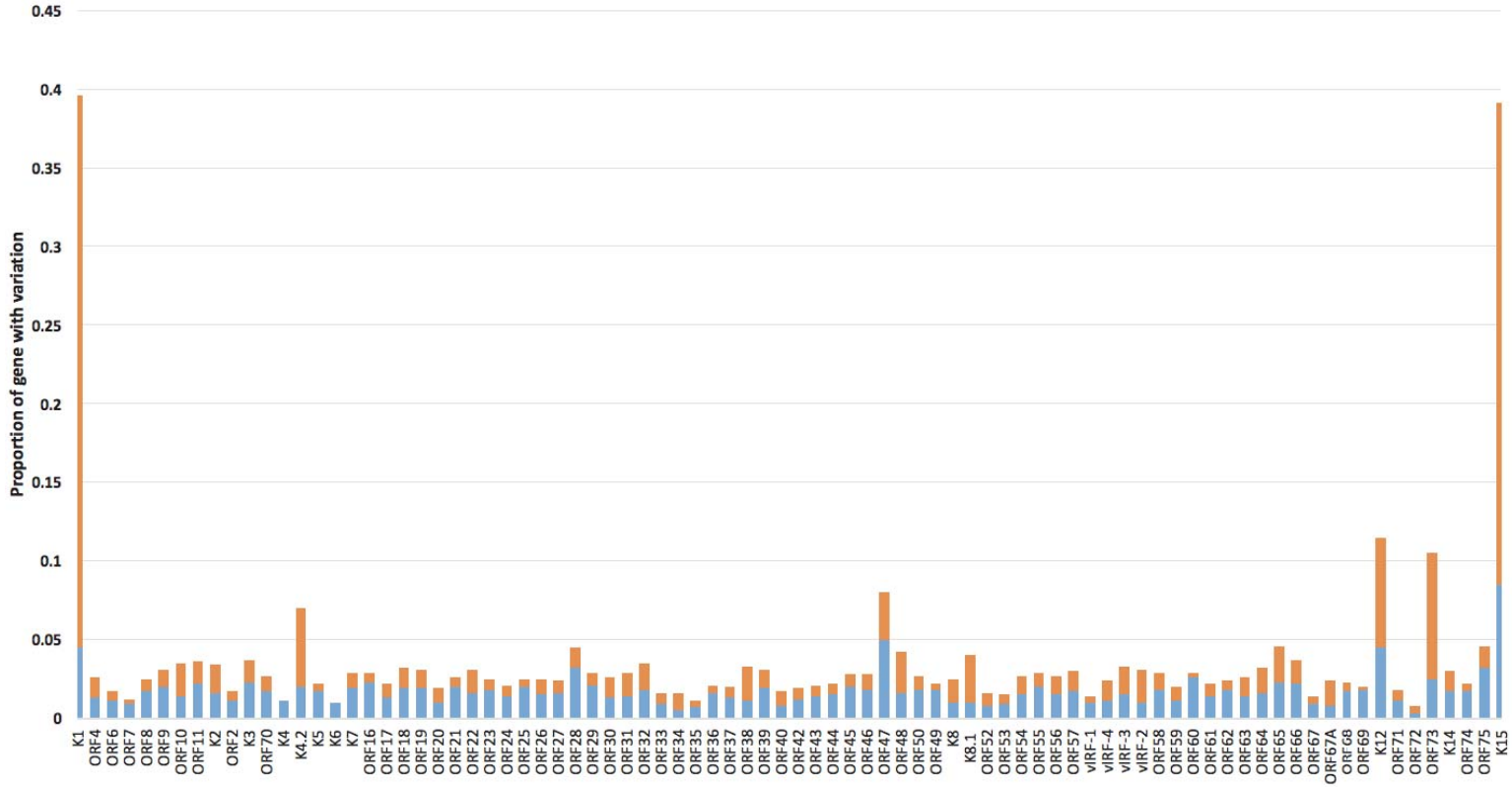


Fig. 5.6 SNP variation across coding region. Number of codon changes per gene across the genome relative to the consensus, presented as the proportion of the gene to normalise for gene length. Blue bars: Synonymous changes. Orange bars: Non-synonymous changes.

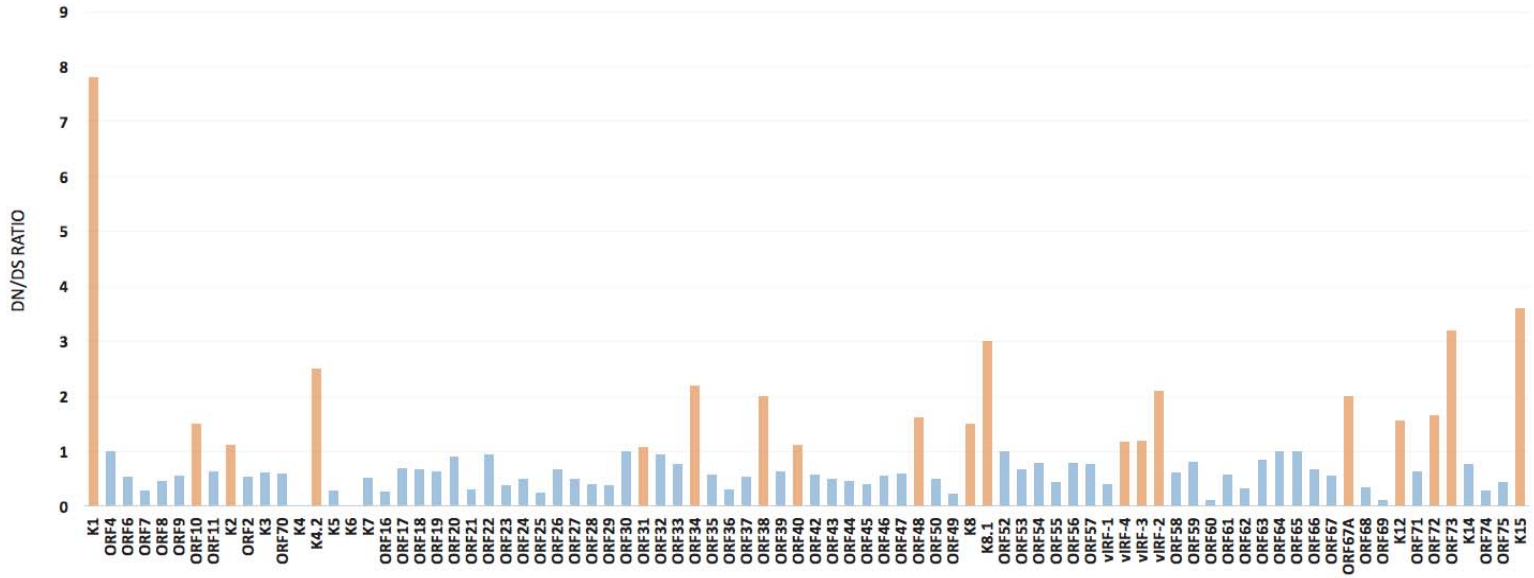


Fig. 5.7 Non-synonymous to synonymous change (dN/dS) analysis across KSHV coding region. Number of non-synonymous to synonymous mutations per gene across the genome relative to the consensus sequence generated from the alignment. Orange bars: dN/dS ratio >1. Blue bars: dN/dS ratio <=1

5.3.4 Virus Population Structure and Geographic Variability

To investigate the population structure of our 62 new wild-type KSHV whole-genomes from Uganda in a global context, I generated phylogenetic trees using a maximum likelihood method following a multiple sequence alignment of all 83 whole-genomes. The whole-genome analysis showed two distinct clades (Fig. 5.8), which have been previously classified as the type P and type M strains based on variation in the K15 gene. Two distinct sub-clades are also observed for both type P and type M strains. In addition, within each type, the Western samples (i.e. Greece and USA) cluster separately from the African samples (i.e. Zambia and Uganda) which show no distinct separation by country.

I also explored whether genomes had any patterns of distribution within villages in the GPC and observed no distinct clustering of samples by strain in the respective villages (Fig. 5.9). In addition, based on the tree there were no major differences between genomes isolated from saliva compared to other sources; and also no major differences between samples from asymptomatic vs. diseased individuals or cell lines (Fig. 5.8).

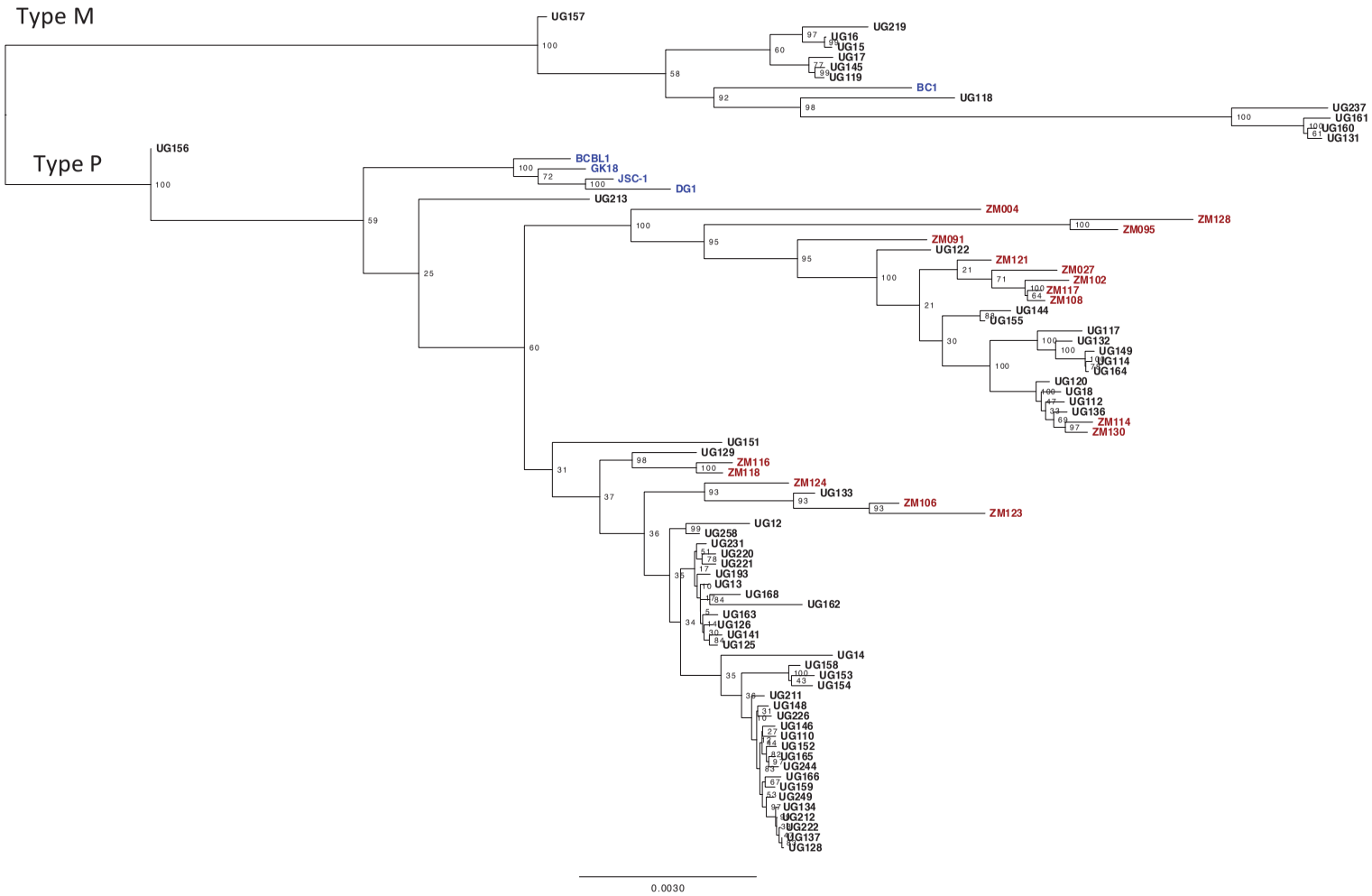


Fig. 5.8 KSHV whole-genome phylogenetic analysis of 83 samples. Midpoint-rooted maximum-likelihood tree of 21 published and 62 new KSHV genomes generated with 1000 bootstrap replicates. Sample names labelled in blue (Western), red (Zambian) and black (Ugandan GPC). Numbers on the nodes correspond to bootstrap values. The scale bar represents 0.003 nucleotide substitutions per site.

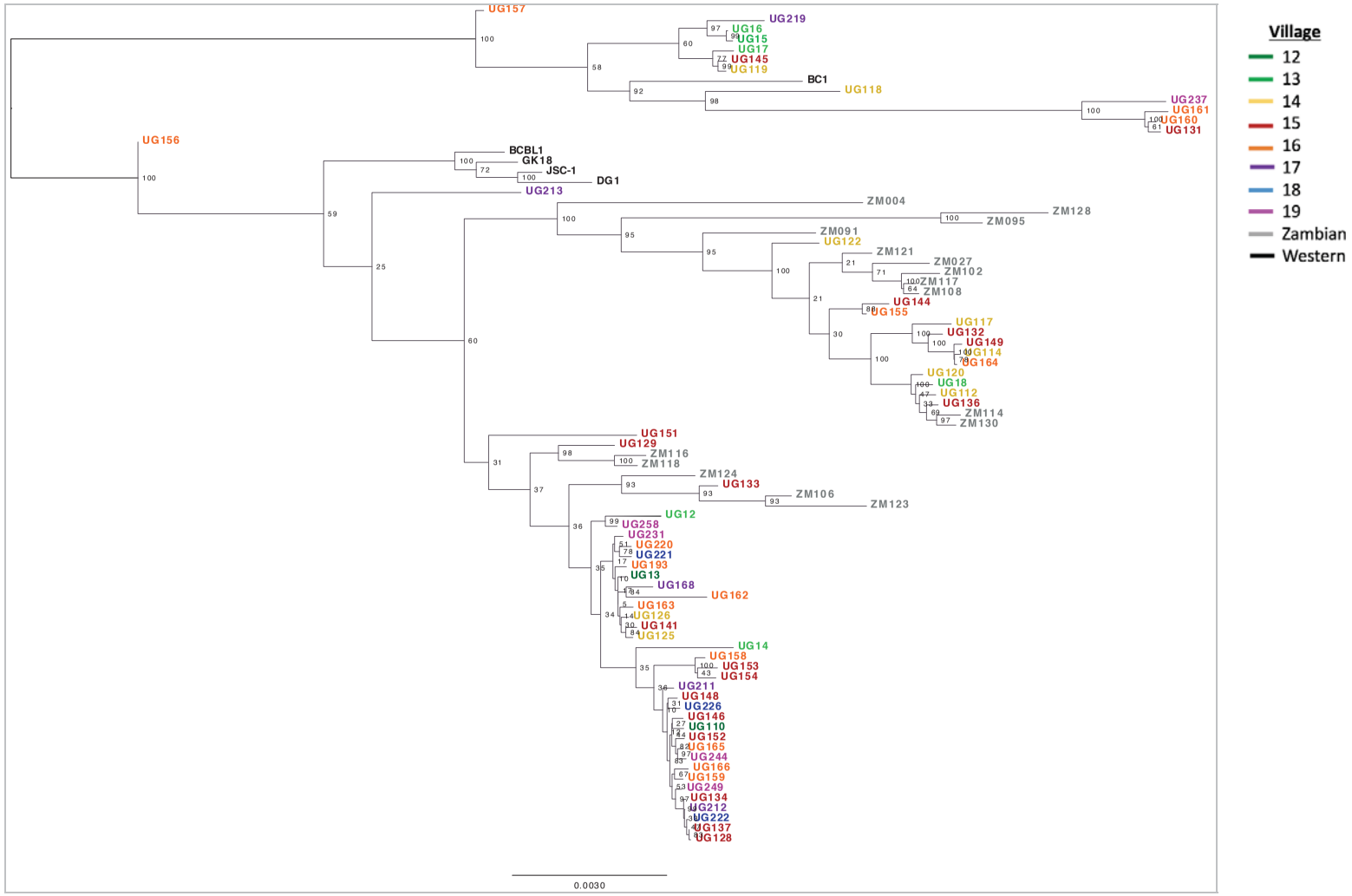


Fig. 5.9 KSHV whole-genome phylogeographic analysis of 83 samples. Midpoint-rooted maximum-likelihood tree of 21 published and 62 new KSHV genomes generated with 1000 bootstrap replicates. Sample names are coloured by village. Numbers on the nodes correspond to bootstrap values. The scale bar represents 0.003 nucleotide substitutions per site.

To assess whether the tree topology was driven by the most variable genes, K1 and K15, I realigned the genomes of all the samples removing the K1 and K15 genes and generated a new tree. While the clustering by Type (P vs M) was lost, two distinct clades still remained and the Western isolates all remained clustered (Fig. 5.10). This substantiated that most of the variation was driven by K1 and K15, however, suggest that genes in the central region are also contributing to the diversity of genomes and thus the geographical clustering, this is consistent with the SNP analysis conducted previously (Fig. 5.5 and Fig. 5.6).

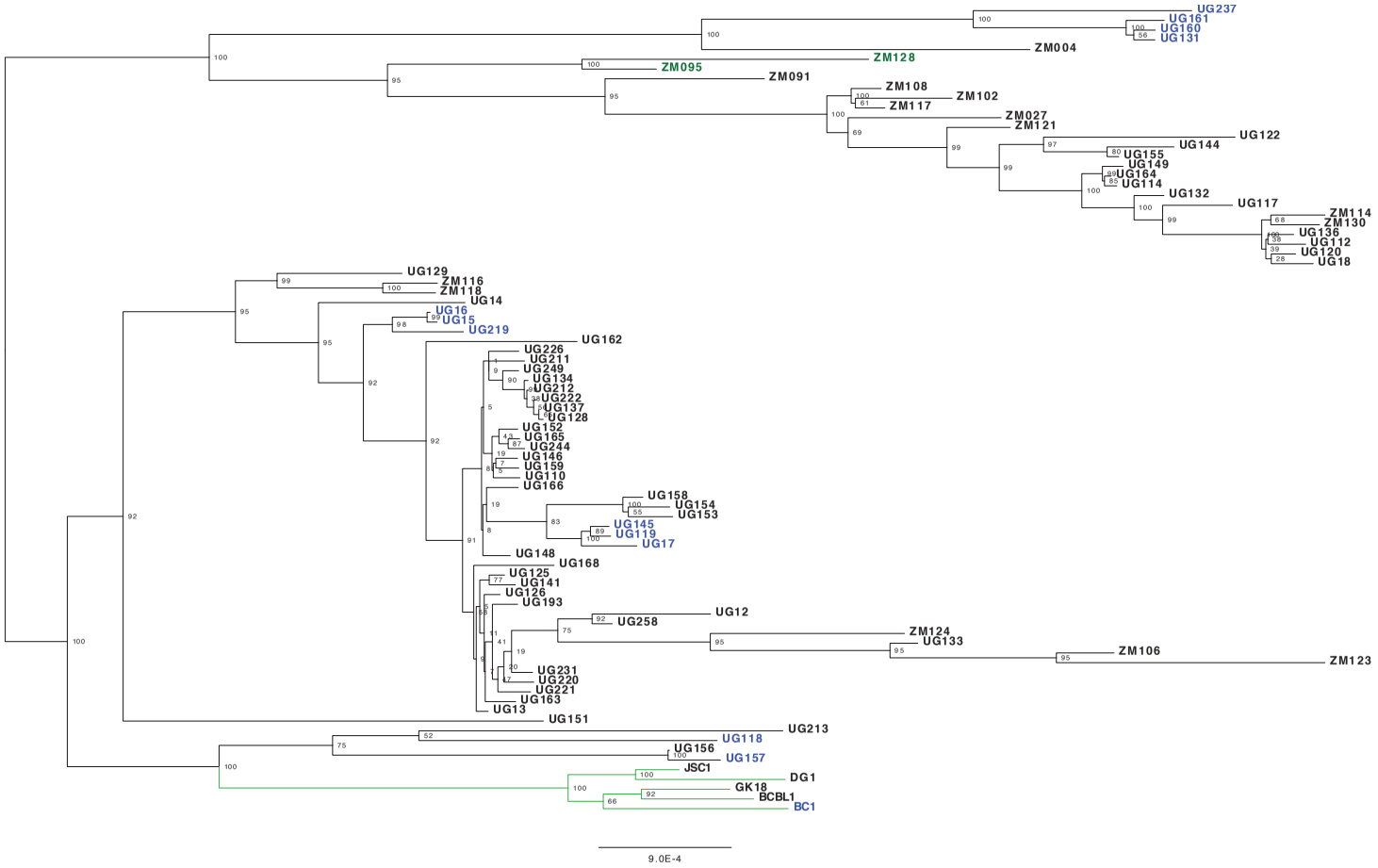


Fig. 5.10 KSHV genome phylogenetic analysis of central region minus K1 and K15 genes in 83 samples. Midpoint-rooted maximum-likelihood tree of 21 published and 62 new KSHV genomes generated with 1000 bootstrap replicates. Sample names are coloured by Type: in black (P), blue (M) and green (N). Branches are also coloured by region: Black (Africa) and Green (Western). Numbers on the nodes correspond to bootstrap values. The scale bar represents 0.0009 nucleotide substitutions per site.

5.3.5 Genotypic Diversity of Strains in the GPC

Since the K1 and K15 genes are the most variable in the genome in this analysis and consistently with previous findings, I generated trees for each gene to determine the genotypes circulating in the 62 Ugandan samples and confirm the genotypes of the 21 samples that have been previously published. For the K15 phylogenetic analysis clear separation was observed between the strains types P Vs M, the majority of the Ugandan samples (50, 80%) remained clustered with the Type P strain and 12 (20%) samples cluster with the type M and none of the GPC samples belonged to the type N strain (Fig. 5.11). To date, these 12 samples from Uganda are the only type M genomes sequenced aside from the only published BC-1 sample which was sequenced from a PEL cell line isolated in the USA (Table 5.2). The major clades observed in the K15 phylogenetic tree are also consistent with that identified in the whole-genome tree (Fig. 5.8).

For the K1 phylogenetic analysis, I aligned the 83 genomes with representative K1 genes for the following genotypes: A1, A2, A3, A4, A5, B1, B2, B3, B4, C1, C3, D1, D2, E and F. Of the 62 Ugandan GPC samples, 30 (48.4%) clustered with B genotypes, 28 (45.1%) clustered with the A genotype and 4 (6.5%) samples clustered with the C genotype (Fig. 5.12). While the B genotypes displayed heterogeneity in subtypes, clustering mainly with B1 and B3, all the A genotypes clustered with the A5 subtype. The C genotypes clustered with the C1 subtype. All the Zambian samples but one, which belonged to the A5 genotype, clustered with the B genotypes and the Western samples clustered with the C and A genotypes as expected. No samples in this study clustered with the D, E or F genotypes. In the GPC, no clustering of genotypes by village was observed but rather a mixed distribution of types across all villages; only two samples were from the same family and living in the same household, UG129 and UG131, both belonged to the B genotype, but were K15 Type P and M respectively. A characterisation of all 62 samples in the GPC is presented in Table 5.4.

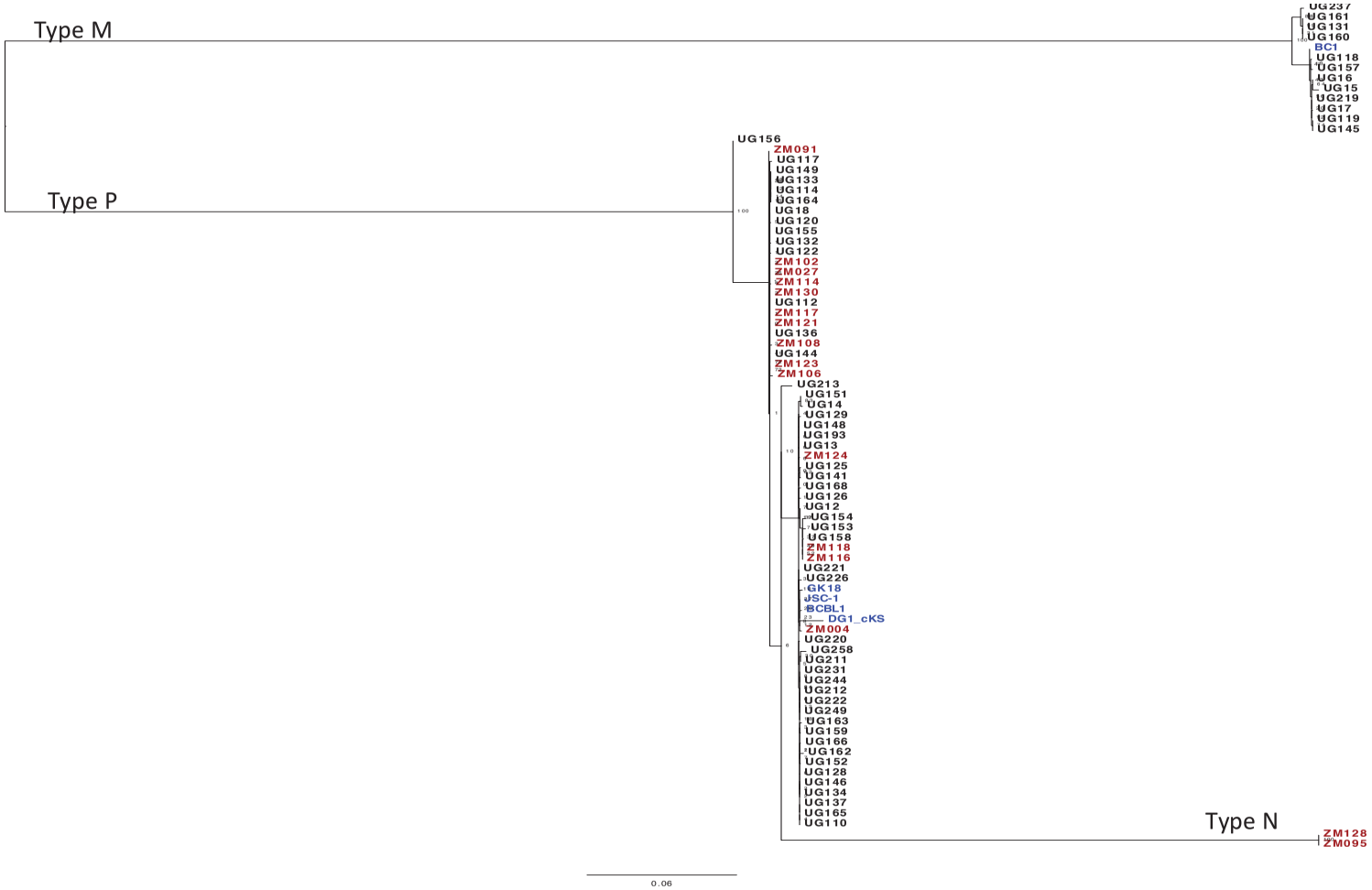


Fig. 5.11 KSHV K15 gene phylogenetic analysis of 83 samples. Midpoint-rooted maximum-likelihood tree of 21 published and 62 new KSHV genomes generated with 1000 bootstrap replicates. Sample names a labelled in blue (Western), red (Zambian) and black (Ugandan GPC). Numbers on the nodes correspond to bootstrap values. The scale bar represents 0.09 nucleotide substitutions per site.

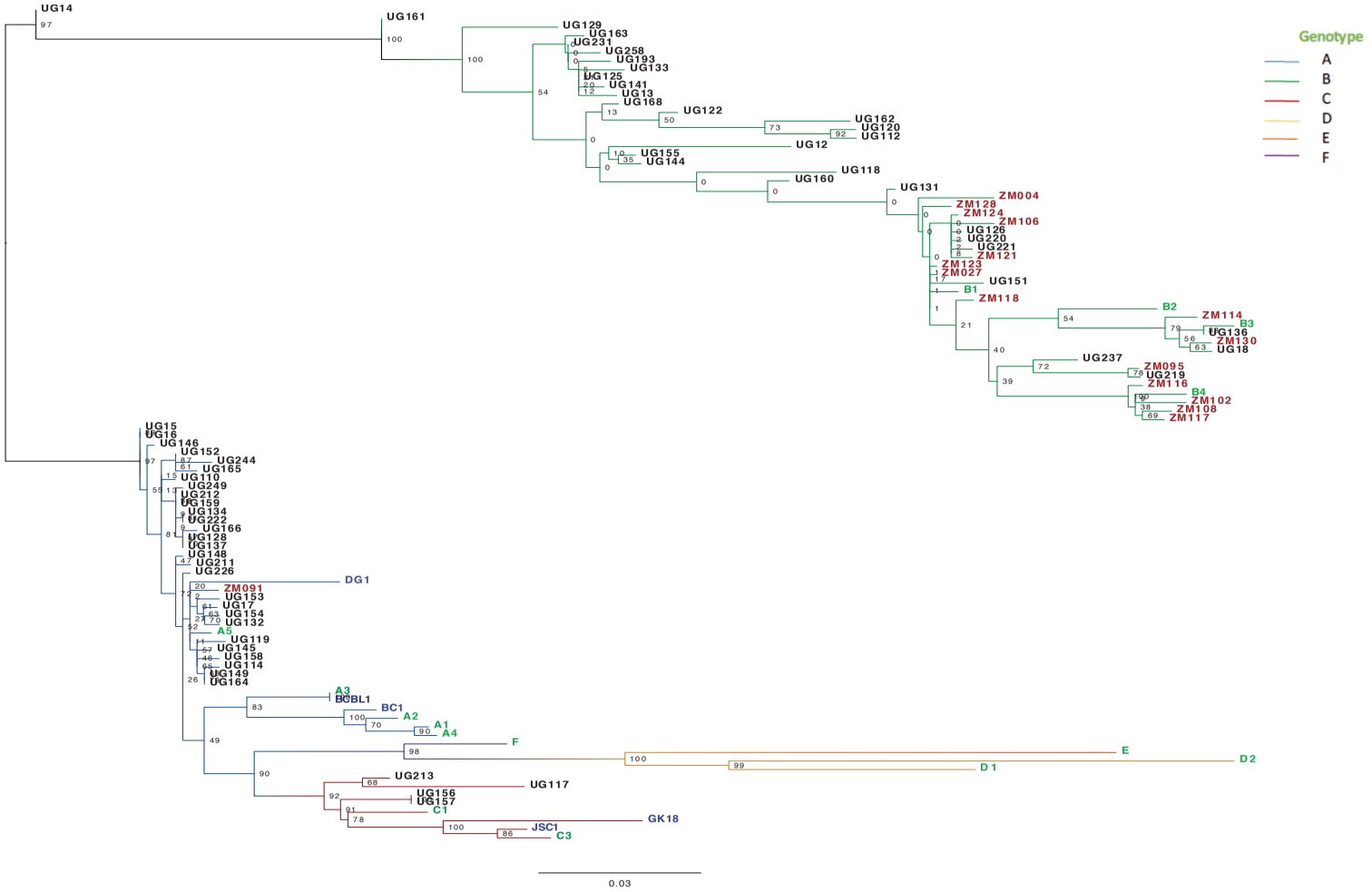


Fig. 5.12 KSHV K1 gene phylogeographic analysis of 83 samples. Midpoint-rooted maximum-likelihood tree of KSHV genomes generated with 1000 bootstrap replicates. Sample names are labelled in blue (Western), red (Zambian), black (Ugandan GPC) and green (representative K1 genotypes). The branches are coloured by genotype as labelled in the figure. Numbers on the nodes correspond to bootstrap values. The scale bar represents 0.03 nucleotide substitutions per site.

Table 5.4 Characteristics of Ugandan GPC Samples

Name	Village No.	Sex	Age	HIV status	Viral Load (Copies/ml)	Mean Depth	Mapped Reads (%)	K15 Type	K1 Genotype
UG110	12	1	40	0	2980	25X	23.3	P	A
UG13	12	2	16	0	16000	200X	29.7	P	B
UG12	13	2	53	1	33600	400X	40.3	P	B
UG14	13	1	59	0	1540	10X	4.4	P	B
UG15	13	1	85	0	7040	25X	21.4	M	A
UG16	13	2	69	0	2150	25X	13.9	M	A
UG17	13	1	22	0	784	10X	7.5	M	A
UG18	13	2	31	0	1960	10X	5.2	P	B
UG112	14	1	91	0	3930	10X	9.3	P	B
UG114	14	1	18	0	45200	25X	33.5	P	A
UG117	14	2	43	0	24700	25X	35.1	P	C
UG118	14	1	21	0	6000	25X	8.6	M	B
UG119	14	2	36	0	11900	25X	19.6	M	A
UG120	14	1	77	0	15500	25X	15.6	P	B
UG122	14	1	32	0	3390	10X	7.1	P	B
UG125	14	2	20	0	15300	25X	35.2	P	B
UG126	14	1	18	0	535000	1000X	70.8	P	B
UG128	15	1	21	0	67700	500X	65.2	P	A
UG129*	15	2	50	0	23400	25X	42.6	P	B
UG131*	15	2	52	0	93700	500X	62.2	M	B
UG132	15	2	46	0	18600	25X	11.7	P	A
UG133	15	2	25	0	26600	25X	35.7	P	B
UG134	15	2	19	0	5730	25X	12.5	P	A
UG136	15	1	79	0	4730	25X	15.6	P	B
UG137	15	1	30	0	13000	25X	22.5	P	A
UG141	15	1	23	0	24200	25X	34.4	P	B
UG144	15	1	33	0	1090	10X	5.2	P	B
UG145	15	1	59	0	7260	25X	21.8	M	A
UG146	15	2	42	0	20200	25X	26.1	P	A
UG148	15	2	17	0	1560	25X	10.7	P	A
UG149	15	2	86	0	53500	25X	40.6	P	A
UG151	15	1	79	0	5630	25X	11.4	P	B
UG152	15	1	19	0	3720	25X	1.7	P	A
UG153	15	1	62	0	19100	10X	16.6	P	A
UG154	15	1	39	0	9200	10X	13.0	P	A
UG155	16	2	29	0	6410	25X	30.3	P	B
UG156	16	2	31	1	104000	1000X	36.6	P	C

Name	Village No.	Sex	Age	HIV status	Viral Load (Copies/ml)	Mean Depth	Mapped Reads (%)	K15 Type	K1 Genotype
UG157	16	2	30	1	37000	1000X	34.9	M	C
UG158	16	2	37	0	9140	20X	1.9	P	A
UG159	16	2	56	0	8960	25X	15.6	P	A
UG160	16	2	40	0	8670	25X	19.8	M	B
UG161	16	1	60	0	7920	10X	4.6	M	B
UG162	16	1	22	0	9590	25X	30.2	P	B
UG163	16	1	30	0	22000	25X	23.9	P	B
UG164	16	1	21	0	135000	750X	76.8	P	A
UG165	16	1	72	0	15200	25X	15.2	P	A
UG166	16	1	18	0	9140	25X	24.9	P	A
UG193	16	2	20	0	1324	10X	2.8	P	B
UG220	16	1	38	0	4250	10X	4.3	P	B
UG168	17	2	50	1	24900	25X	26.5	P	B
UG211	17	1	34	0	2230	10X	3.4	P	A
UG212	17	1	20	0	31900	25X	16.9	P	A
UG213	17	2	67	0	3050	10X	4.4	P	C
UG219	17	2	42	0	16800	25X	11.0	M	B
UG221	18	2	77	0	3300	10X	5.3	P	B
UG222	18	1	22	0	19900	25X	10.0	P	A
UG226	18	1	51	1	8220	20X	1.5	P	A
UG231	19	2	50	0	2210	10X	8.0	P	B
UG237	19	1	75	0	11700	25X	18.0	M	B
UG244	19	2	43	0	2910	20X	5.3	P	A
UG249	19	1	19	0	4830	10X	8.1	P	A
UG258	19	2	33	0	2060	10X	6.3	P	B

[a] Sex: 1=Male, 2=Female
[b] HIV Status: 0=Negative, 1=Positive
*Belong to the same household
Coloured by village number

5.4 Discussion

Whole-genome sequence analyses of viruses are crucial to enhance our understanding of viral phenotypes, define virus population structure, transmission chains, elucidate variants under selection and explore relationships between genomic diversity and disease pathogenesis, which remains largely unknown for KSHV. In this study, I present for the first time 62 new wild-type genomes isolated from saliva of non-diseased adults, and performed a genomic variation and phylogenetic analyses in combination with 21 previously published genomes from Greece, USA and Zambia, thus representing the largest KSHV whole-genome study to date. In addition, this study presents a unique dataset with adults from eight neighbouring villages, as the majority of KSHV molecular epidemiology studies have focused on children and/or looked at transmission dynamics between mother-child or within hospitals[145,146,158,170,172,174,175,585].

Previous genetic analyses of whole-genomes generated from KS, PEL and KICS samples have provided invaluable insights into KSHV genomic architecture and viral epidemiology[138,560,563], however, they may not be representative of those found in the general population, particularly in KSHV endemic regions such as Uganda where oral transmission is the most likely route of infection. Studies using saliva pose a significant challenge given the virus is difficult to detect particularly in asymptomatic individuals unless they're shedding virus i.e. during the lytic stage of infection, and viral levels are much lower in saliva and blood compared to in tumour biopsies or cell lines[563]. With such low quantity viral DNA within a greater pool of complex human DNA recent studies have demonstrated that target enrichment technology is better suited for capturing viral genomes[584,586]. Therefore, here, I isolated KSHV DNA from saliva of asymptomatic carriers in Uganda, assessed the prevalence of viral shedding by qPCR and successfully sequenced multiple whole-genomes from a population-based cohort using a target enrichment approach with Illumina paired-end sequencing technology. qPCR analysis of DNA isolated from 746 adults showed that in the GPC ~33% (244) of individuals are

actively shedding KSHV (i.e. detection of KSHV genome in saliva), with inter-individual variability in viral loads from 1.5 to 5.35x10^5 copies/ml which is within range of previous findings[146,316,587-590]. Following whole-genome sequencing of all 244 samples, I sought to investigate whether viral load influenced sequencing quality and found that viral load was highly correlated with achieving good sequencing depth and percentage of KSHV mapped reads, this is reflected by the 62 samples (25%) with viral loads of >10^4 copies/ml, which had a mean sequencing depth of at least 10x of coverage >90% across the genome (Fig. 5.3 and Table 5.4). In addition, viral load was also correlated with the number of reads that mapped to the KSHV genome (Fig. 5.3). This is an important observation for future studies that wish to sequence KSHV DNA directly from clinical samples such as saliva, these data suggest one should focus on samples with a viral load >10^4 copies/ml. As these 62 samples had good sequencing coverage, I retained them for further genetic variation and phylogenetic analyses with an additional 21 published whole-genomes from diseased individuals from Greece, USA and Zambia.

Multiple sequence alignment of the 83 genomes showed high levels of sequence conservation, with a higher level of genetic variation in the 5' and 3' genome ends, corresponding to the K1 and K15 genes, respectively (Fig. 5.5). A low level of genetic variation was found across the central region of the genome, consistent with previous findings[563,564]. This was confirmed by analysis of SNPs across the coding region of the genome (Fig. 5.6). While it has been found that the K1 gene has been evolving under strong host selective pressure in association with cytotoxic T-lymphocyte recognition[145,591,592], few data exist for other genes under selection driven by the host. Along with the K1 and K15 genes, other genes of interest from the SNP analysis across the KSHV coding region are ORF73 (encoding LANA), K12 (encoding Kaposin), K4.2, K8.1 and v-IRF2 in the central region (Fig. 5.7), which have a higher proportion of non-synonymous polymorphisms suggesting that they're under positive selection. It is interesting that these genes are all encoding immunomodulatory proteins[593] and thus this selective pressure may facilitate KSHV's evasion of the immune response. All the

genes above, except ORF73 were identified with a high number of polymorphisms in the Zambian study, highlighting the advantage of having more genomes to comprehensively identify genes under selection.

Similarly to the well-established Type 1/ Type 2 classification and whole-genome clustering of EBV based on the divergent EBNA-3 allele[586,594], for KSHV, the Type P / Type M based on variation in the K15 gene remains the major form of variation correlating with whole-genome clustering (Fig. 5.8). KSHV Types based on the multiple sequence alignment of the K15 gene confirmed the Type P/M split observed in the whole-genome tree with 50 of the GPC samples belonging to the Type P and 12 to the Type M (Fig. 5.11). Whole-genome clustering also showed similarity in genomes irrespective of derivation from different clinical presentations i.e. asymptomatic vs diseased or source of sample isolation i.e. saliva vs biopsy/cell line. Distinct phylogenetic clustering was observed between the African samples (Uganda and Zambia) and the Western samples, as previously observed in the Zambian KS whole-genome study[564]. Giving evidence to this, the Type P/M clustering was lost following removal of the K15 gene from the genome alignment, however, geographical clustering of samples still remained (Fig. 5.10). The addition of more samples to the Zambian study refined the phylogenetic relationships in the tree and two distinct sub-clades were observed for types P and M, potentially arising as a result of variation in the central region which would need to be further resolved using ancestral reconstruction methods.

Geographic association of K1 genotypes has been reported by several studies globally. Hayward hypothesised that KSHV is an old human virus and the distribution of its' subtypes arose as a result of ancient human migration >100,000 years ago out of Africa[144,192]. The A (particularly A1-A4) and C subtypes are found to predominate in Europe, USA, Australia, the Middle East and Asia; the A5 and B genotypes are typical for populations of African descent and more recently the F genotype was identified in Ugandans; the D and E genotypes are more common in the Pacific Islands and Brazilian

Amerindians, respectively[186,187,567,570,571,576,595-603]. Genotypic analysis of the 62 samples based on the K1 gene revealed a heterogeneous distribution of subtypes throughout the villages in the GPC (Fig. 5.12), consistent with previous studies the B and A5 subtypes predominate at 48.4% and 45.1% respectively with a few samples belonging to the C1 (6.5%) genotype. Interestingly, while the A genotypes all clustered with A5, the B genotypes were heterogenous with more subgroups compared to the A, suggesting the K1 genotyping is not fully capturing variation of these genomes. No subgroups based on villages were observed. Conflicting data exists on whether different genotypes are attributed to pathogenic or tumourigenic properties of KSHV, a very recent study conducted in a South African population reported that the A5 genotype is associated with extensive disease in AIDS-KS[604] and a Zambian study found it to be associated with childhood KS[596], however, an earlier study showed that the A5 genotype is more prevalent in African children than mothers and thus represents more efficient viral transmission[145]. Thus, genotypic diversity and its relation to pathogenesis remains unclear, however, it might be further resolved by taking the whole genome into account.

In summary, in this study I assessed the prevalence of KSHV shedding (i.e. detection of KSHV genome in saliva) in the Ugandan GPC and identified a viral load threshold for the successful sequencing of KSHV whole-genomes isolated from clinical samples by target enrichment. I present the largest KSHV whole-genome analyses to date with 62 new wild-type whole-genomes from Uganda which are the first to be generated from saliva of asymptomatic individuals and extend the analysis conducted recently with 16 Zambian KS genomes and including 5 previously published genomes from Greece and USA. This study confirmed the presence of high level variation at the 5' and 3' ends of the genome that drives major variation between KSHV strains, in addition to low level variation in the central conserved region, with genes involved in modulating host response and under selective pressure contributing to distinct phylogenetic clustering between Western and African samples. The heterogenous distribution of KSHV strains, with a variety of genotypes observed throughout all villages, suggesting cross-ethnic and cross-village

transmission is not surprising given how well connected the villages are, with relaxed administrative boundaries enabling ease of access and movement between the villages[388]. Two adults who belonged to the same family/household had different K15 genotypes (Table 5.4), suggesting transmission is not horizontal between these adults, this is consistent with previous findings that KSHV is predominantly transmitted vertically (i.e. from mother-child)[145,174,585,605]. However, to reliably identify transmission patterns in this study more familial and household samples across all age groups would be required. It is also worth noting that as result of mapping the 62 new genomes to the GK18 reference sequence a bias may exist, for example, missing out insertions and deletions, and thus, underestimating genetic diversity.

In conclusion, despite extensive genotypic characterization worldwide, how selection pressure is driving genomic variation and whether specific genotypes are linked to pathogenesis and disease remains unclear and thus will require further investigation. From this study and the previous study of Zambian KS patients, while a high level of similarity exists between genomes, it is evident that K1 and K15 genotyping insufficiently capture genetic variation and whole-genome variation is greater than previously appreciated. The addition of genomes from Uganda to the Zambian dataset identified additional genes under selection and refined phylogenetic relationships between genomes. Therefore, whole-genome sequences from other parts of the world providing a more comprehensive global dataset would be essential to substantiate these findings. Viral characterisation based on whole-genome diversity needs to be considered coupled with a revision of the nomenclature.

6. Conclusions and Future Outlook

In this thesis, I have described four distinct results chapters with the aim of understanding how the genetics of host-virus interactions influences pathogenesis, in particular, the contribution of host genetic variation to EBV and KSHV infections in a rural African population cohort, the Ugandan General Population Cohort (GPC).

Chapter 2 focused on characterising the GPC in rural southwest Uganda and the systematic differences such as environmental factors, population structure, in addition to the heritability of IgG antibody response traits, confirming the suitability of the GPC for use in host genetic studies of gamma-herpesvirus antibody response traits. This study revealed that EBV and KSHV infections were ubiquitous (>90%) and also showed a high burden of co-infection(s) with other viruses that influenced inter-individual variation in Immunoglobulin G (IgG) antibody responses traits. Furthermore, both EBV and KSHV IgG antibody response traits were partly heritable after adjusting for environmental correlation. Thus, the GPC data allowed for the host genetic studies of EBV and KSHV infections independently of the environment described in chapter 3 and chapter 4 respectively, in addition a subset of individuals were resampled for the study of KSHV viral genomic diversity described in chapter 5.

For chapters 3 and 4, GWAS was performed 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. This approach has overcome limitations that previous studies had using genotype arrays and/or imputation panels developed based on European ancestry genetic data. Chapter 3 showed variation to anti-EBNA-1 IgG levels is mainly influenced by variants in the *HLA* Class II region, while response to anti-VCA IgG is regulated by multiple genes involved in pathways that might limit EBV replication and thus together facilitate the evasion of host defences. While the GWAS in chapter 4 did not reveal strong associations with KSHV antibody response traits

despite greater power for variant detection than the EBV GWAS, this suggests differences in genetic architecture underlying responses to the two infections. It is also possible that a combination of multiple variants with small to modest effect sizes are underlying KSHV phenotypic variability. To follow up significant GWAS findings, replication of novel loci is essential, in addition, the development of pathway analysis tools with African populations well represented would be necessary to reliably identify gene enrichments in pathways and protein interaction networks. For both these studies, fine-mapping to infer causality and functional validation to fully understand how these variants affect biological function to potentially cause disease in individuals is crucial. Furthermore, the development of African resources such as HLA imputation reference panels based on African genetic data and gene expression data from Africans will be crucial to be able to leverage approaches such as GWAS.

In chapter 5, individuals were resampled for saliva to isolate KSHV whole genomes and attempt to understand whether variation in viral genomes could explain differences in high seroprevalence in Uganda compared to the rest of the world. Viral genomes clustered largely based on previously defined K15 gene sub types (P and M), in addition within the types, samples clustered based on geography (i.e African Vs Western). It is highly likely that variation driven by central region of the genome is also driving geographical clustering. Genomic data from other parts of the world would be required to refine these findings.

6.1 Inferring the Causality of Variants

Despite identifying thousands of loci through GWAS, inferring causality of variants, their potential effector transcripts and biological mechanisms remains a challenge, nevertheless an advantage African populations such as the GPC present are short LD blocks which make refining multiple signals down to a single causal variant easier compared to European ancestry populations. In chapters 3 and 4 multiple novel candidate loci were identified and potential roles in EBV or KSHV pathogenesis were

described, however it is worth noting that an associated locus often contains numerous SNPs in correlation, and spanning across multiple genes, therefore, the variants may be affecting the expression of completely different genes. For example, intronic SNPs in *FTO* locus were identified as strongly associated with body mass index and obesity[606]. Subsequently, the *FTO* gene was shown to be expressed in hypothalamic neurons that control appetite and energy[607] and early rodent models suggested that the genetic association with adiposity was mediated by a direct effect of the *FTO* gene. However, subsequent studies by other groups provided compelling evidence that *FTO* interacted with and was involved in the expression of the distant genes *IRX3* and *IRX5* but not *FTO* itself[608]. The causal variant was also recently identified and was in strong LD with the lead SNP and found to alter *ARID5B* repressor binding leading to stimulation of *IRX3* and *IRX5*[609]. This study is a classic example of the value of fine-mapping and functional follow up of GWAS findings to gain biological insights. The challenge will be to conduct such studies in a powered, high quality and scalable fashion to keep up with the pace of new genetic discovery.

6.2 The Contribution of Low-Frequency and Rare Variants to Infectious Disease

Consistent with the demographic history of African populations, African populations carry the largest number of variants compared to Europeans with the majority being rare. Thus exploring the contribution of rare genetic variants in infectious disease risk or trait variability is highly important. While whole-genome sequencing has greatly improved the ability to detect low-frequency and rare genetic variants, in this study the statistical power to detect such variants are low, for example in the KSHV GWAS of ~4500 individuals, the power to detect variants of 1% with an effect size of at least 0.6 is ~50%, thus for variants less than 1%, unless the effect sizes are large, larger sample sizes would be required. To overcome such challenges in statistical power and sample sizes, methods have been recently developed that aggregate and evaluate association signals for multiple variants in a gene, rather than single variant testing as performed in GWAS[610]. In addition, very recently, the haplotype reference consortium (HRC) has built

a large reference panel, however, again this is predominantly for Europeans. Therefore, developing further large-scale imputation reference panels expanding the current 1000G+AGV+UG2G reference panel (used here) to include much large sample numbers from diverse African populations would further enable the greater capture of the genetic diversity across the continent and facilitate large-scale studies without the need for whole-genome sequencing approaches which are still prohibitively expensive to be done at the scale required to have power, i.e, in the order of 20,000 cases and >100,000 controls[611].

6.3 Genome-to-Genome Analysis

An insightful approach to bridge the gap in host-virus interactions is by conducting a genome-genome analysis as proposed by Bartha and colleagues[612] using host and viral genomic data available from individuals in the GPC. This method compares viral genome to that of the infected host to elucidate selection pressures imposed by host genomic factors that suppress viral function to those that are overcome by the pathogen. Using paired host and virus genomic data from 1071 HIV-infected individuals, Bartha and colleagues performed genome-wide scans across ~7 million variants and used HIV amino acid variation as an intermediate phenotype for association. They identified significant associations of SNPs in the HLA class I region with a total of 48 HIV amino acid variants ($p=2.4x10^{-12}$); this association was also stronger than when they used viral load as a phenotype. For EBV or KSHV, viral genome sequence diversity could be used as intermediate phenotypes. The challenge in using viral genomes for KSHV analyses, however, would be the availability of genome sequence data from a large sample size, in chapter 5, viral DNA was only detected from ~30% of individuals who were presumably shedding KSHV in saliva at the time of sample collection, thus large sample sizes would be required to achieve enough power to perform such analysis. Thus, to expand human genetic studies of infection across the world incorporating the contribution of the pathogen genome, it would be beneficial for future studies to invest in collecting both host and pathogen genetic data simultaneously.

References

1. Damania, B. Oncogenic gamma-herpesviruses: comparison of viral proteins involved in tumorigenesis. *Nat Rev Microbiol* **2**, 656-68 (2004).
2. Longnecker, R. & Neipel, F. Introduction to the human gamma-herpesviruses. in *Human Herpesviruses: Biology, Therapy, and Immunoprophylaxis* (eds. Arvin, A. *et al.*) (Cambridge, 2007).
3. Barton, E., Mandal, P. & Speck, S.H. Pathogenesis and host control of gammaherpesviruses: lessons from the mouse. *Annu Rev Immunol* **29**, 351-97 (2011).
4. Dalton-griffin, L. The role of host cell factors in the lytic reactivation of Kaposi ' s sarcoma-associated herpesvirus from latency. 1-279 (2010).
5. Epstein, M.A., Achong, B.G. & Barr, Y.M. VIRUS PARTICLES IN CULTURED LYMPHOBLASTS FROM BURKITT'S LYMPHOMA. *The Lancet* **283**, 702-703 (1964).
6. Burkitt, D. A sarcoma involving the jaws in African children. *Br J Surg* **46**, 218-23 (1958).
7. Henle, W., Diehl, V., Kohn, G., Zur Hausen, H. & Henle, G. Herpes-type virus and chromosome marker in normal leukocytes after growth with irradiated Burkitt cells. *Science* **157**, 1064-5 (1967).
8. Henle, G., Henle, W. & Diehl, V. Relation of Burkitt's tumor-associated herpes-ytpe virus to infectious mononucleosis. *Proc Natl Acad Sci U S A* **59**, 94-101 (1968).
9. zur Hausen, H. *et al.* EBV DNA in biopsies of Burkitt tumours and anaplastic carcinomas of the nasopharynx. *Nature* **228**, 1056-8 (1970).
10. Ziegler, J.L. *et al.* Outbreak of Burkitt's-like lymphoma in homosexual men. *Lancet* **2**, 631-3 (1982).
11. Greenspan, J.S. *et al.* Replication of Epstein-Barr virus within the epithelial cells of oral "hairy" leukoplakia, an AIDS-associated lesion. *N Engl J Med* **313**, 1564-71 (1985).
12. Jones, J.F. *et al.* T-cell lymphomas containing Epstein-Barr viral DNA in patients with chronic Epstein-Barr virus infections. *N Engl J Med* **318**, 733-41 (1988).
13. Weiss, L.M., Movahed, L.A., Warnke, R.A. & Sklar, J. Detection of Epstein-Barr viral genomes in Reed-Sternberg cells of Hodgkin's disease. *N Engl J Med* **320**, 502-6 (1989).
14. Yao, Q.Y., Rickinson, A.B. & Epstein, M.A. A re-examination of the Epstein-Barr virus carrier state in healthy seropositive individuals. *Int J Cancer* **35**, 35-42 (1985).
15. Biggar, R.J. *et al.* Primary Epstein-Barr virus infections in African infants. I. Decline of maternal antibodies and time of infection. *Int J Cancer* **22**, 239-43 (1978).
16. de-The, G. Epstein-Barr virus behavior in different populations and implications for control of Epstein-Barr virus-associated tumors. *Cancer Res* **36**, 692-5 (1976).
17. Crawford, D.H. *et al.* A cohort study among university students: identification of risk factors for Epstein-Barr virus seroconversion and infectious mononucleosis.

Clinical infectious diseases : an official publication of the Infectious Diseases Society of America **43**, 276-82 (2006).
18. Mbulaiteye, S.M. *et al.* High levels of Epstein-Barr virus DNA in saliva and peripheral blood from Ugandan mother-child pairs. *The Journal of infectious diseases* **193**, 422-6 (2006).
19. Hjalgrim, H., Friborg, J. & Melbye, M. The epidemiology of EBV and its association with malignant disease. in *Human Herpesviruses: Biology, Therapy, and Immunoprophylaxis* (eds. Arvin, A. *et al.*) (Cambridge, 2007).
20. Daud, I.I. *et al.* Breast Milk as a Potential Source of Epstein-Barr Virus Transmission Among Infants Living in a Malaria-Endemic Region of Kenya. *The Journal of infectious diseases* **212**, 1735-42 (2015).
21. Alfieri, C. *et al.* Epstein-Barr virus transmission from a blood donor to an organ transplant recipient with recovery of the same virus strain from the recipient's blood and oropharynx. *Blood* **87**, 812-7 (1996).
22. Scheenstra, R. *et al.* The value of prospective monitoring of Epstein-Barr virus DNA in blood samples of pediatric liver transplant recipients. *Transpl Infect Dis* **6**, 15-22 (2004).
23. Pagano, J.S. Is Epstein-Barr virus transmitted sexually? *J Infect Dis* **195**, 469-70 (2007).
24. van Baarle, D. *et al.* High prevalence of Epstein-Barr virus type 2 among homosexual men is caused by sexual transmission. *J Infect Dis* **181**, 2045-9 (2000).
25. Chang, R.S., Rosen, L. & Kapikian, A.Z. Epstein-Barr virus infections in a nursery. *Am J Epidemiol* **113**, 22-9 (1981).
26. Lang, D.J., Garruto, R.M. & Gajdusek, D.C. Early acquisition of cytomegalovirus and Epstein-Barr virus antibody in several isolated Melanesian populations. *Am J Epidemiol* **105**, 480-7 (1977).
27. Dowd, J.B., Palermo, T., Brite, J., McDade, T.W. & Aiello, A. Seroprevalence of Epstein-Barr virus infection in U.S. children ages 6-19, 2003-2010. *PLoS One* **8**, e64921 (2013).
28. Sample, J. *et al.* Epstein-Barr virus types 1 and 2 differ in their EBNA-3A, EBNA-3B, and EBNA-3C genes. *J Virol* **64**, 4084-92 (1990).
29. Gratama, J.W. & Ernberg, I. Molecular epidemiology of Epstein-Barr virus infection. *Adv Cancer Res* **67**, 197-255 (1995).
30. Yao, Q.Y. *et al.* Epidemiology of infection with Epstein-Barr virus types 1 and 2: lessons from the study of a T-cell-immunocompromised hemophilic cohort. *J Virol* **72**, 4352-63 (1998).
31. Cancian, L., Bosshard, R., Lucchesi, W., Karstegl, C.E. & Farrell, P.J. C-terminal region of EBNA-2 determines the superior transforming ability of type 1 Epstein-Barr virus by enhanced gene regulation of LMP-1 and CXCR7. *PLoS Pathog* **7**, e1002164 (2011).
32. de Martel, C. *et al.* Global burden of cancers attributable to infections in 2008: a review and synthetic analysis. *The Lancet Oncology* **13**, 607-615 (2012).

33. Group, I.W. *IARC monograph on the evaluation of carcinogenic risks to Humans. Volume 100b, Review of human carcinogens: biological agents. IARC, Lyon, France, 2011.*, 1-441 (2012).
34. Patsopoulos, N.A. *et al.* Genome-wide meta-analysis identifies novel multiple sclerosis susceptibility loci. *Ann Neurol* **70**, 897-912 (2011).
35. Ulff-Moller, C.J., Nielsen, N.M., Rostgaard, K., Hjalgrim, H. & Frisch, M. Epstein-Barr virus-associated infectious mononucleosis and risk of systemic lupus erythematosus. *Rheumatology (Oxford)* **49**, 1706-12 (2010).
36. Biggar, R.J. *et al.* Primary Epstein-Barr virus infections in African infants. II. Clinical and serological observations during seroconversion. *Int J Cancer* **22**, 244-50 (1978).
37. Carpenter, L.M. *et al.* Antibodies against malaria and Epstein-Barr virus in childhood Burkitt lymphoma: a case-control study in Uganda. *International journal of cancer* **122**, 1319-23 (2008).
38. Mutalima, N. *et al.* Associations between Burkitt lymphoma among children in Malawi and infection with HIV, EBV and malaria: results from a case-control study. *PLoS One* **3**, e2505 (2008).
39. Neel, H.B., 3rd, Pearson, G.R. & Taylor, W.F. Antibodies to Epstein-Barr virus in patients with nasopharyngeal carcinoma and in comparison groups. *Ann Otol Rhinol Laryngol* **93**, 477-82 (1984).
40. Silins, S.L. *et al.* Asymptomatic primary Epstein-Barr virus infection occurs in the absence of blood T-cell repertoire perturbations despite high levels of systemic viral load. *Blood* **98**, 3739-44 (2001).
41. Chang, R.S. & Chang, Y.Y. Activation of lymphocytes from Epstein-Barr virus-seronegative donors by autologous Epstein-Barr virus-transformed cells. *J Infect Dis* **142**, 156-62 (1980).
42. de-Thé, G. *et al.* Epidemiological evidence for causal relationship between Epstein-Barr virus and Burkitt's lymphoma from Ugandan prospective study. *Nature* **274**, 756-761 (1978).
43. Glaser, S.L. *et al.* Epstein-Barr virus-associated Hodgkin's disease: epidemiologic characteristics in international data. *Int J Cancer* **70**, 375-82 (1997).
44. Pallesen, G., Hamilton-Dutoit, S.J. & Zhou, X. The association of Epstein-Barr virus (EBV) with T cell lymphoproliferations and Hodgkin's disease: two new developments in the EBV field. *Adv Cancer Res* **62**, 179-239 (1993).
45. Halprin, J. *et al.* Enzyme-linked immunosorbent assay of antibodies to Epstein-Barr virus nuclear and early antigens in patients with infectious mononucleosis and nasopharyngeal carcinoma. *Ann Intern Med* **104**, 331-7 (1986).
46. Sam, C.K., Prasad, U. & Pathmanathan, R. Serological markers in the diagnosis of histopathological types of nasopharyngeal carcinoma. *Eur J Surg Oncol* **15**, 357-60 (1989).
47. Zeng, Y. *et al.* Serological mass survey for early detection of nasopharyngeal carcinoma in Wuzhou City, China. *Int J Cancer* **29**, 139-41 (1982).
48. Baer, R. *et al.* DNA sequence and expression of the B95-8 Epstein-Barr virus genome. *Nature* **310**, 207-11 (1984).

49. Sixbey, J.W. *et al.* Replication of Epstein-Barr virus in human epithelial cells infected in vitro. *Nature* **306**, 480-3 (1983).
50. Allday, M.J. & Crawford, D.H. Role of epithelium in EBV persistence and pathogenesis of B-cell tumours. *Lancet* **1**, 855-7 (1988).
51. Anagnostopoulos, I., Hummel, M., Kreschel, C. & Stein, H. Morphology, immunophenotype, and distribution of latently and/or productively Epstein-Barr virus-infected cells in acute infectious mononucleosis: implications for the interindividual infection route of Epstein-Barr virus. *Blood* **85**, 744-50 (1995).
52. Niedobitek, G. *et al.* Epstein-Barr virus (EBV) infection in infectious mononucleosis: virus latency, replication and phenotype of EBV-infected cells. *J Pathol* **182**, 151-9 (1997).
53. Fingeroth, J.D. *et al.* Epstein-Barr virus receptor of human B lymphocytes is the C3d receptor CR2. *Proc Natl Acad Sci U S A* **81**, 4510-4 (1984).
54. Li, Q. *et al.* Epstein-Barr virus uses HLA class II as a cofactor for infection of B lymphocytes. *Journal of virology* **71**, 4657-62 (1997).
55. Kalla, M. & Hammerschmidt, W. Human B cells on their route to latent infection--early but transient expression of lytic genes of Epstein-Barr virus. *Eur J Cell Biol* **91**, 65-9 (2012).
56. Babcock, G.J., Decker, L.L., Volk, M. & Thorley-Lawson, D.A. EBV persistence in memory B cells in vivo. *Immunity* **9**, 395-404 (1998).
57. Kieff, E. Epstein-Barr virus and its replication. in *Fields Virology. 3rd Ed.*, Vol. 2 (eds. Knipe, D.M. & Howley, P.M.) 2343-96 (Lippincott-Raven, Philadelphia, 1996).
58. Yates, J.L., Warren, N. & Sugden, B. Stable replication of plasmids derived from Epstein-Barr virus in various mammalian cells. *Nature* **313**, 812-5 (1985).
59. Levitskaya, J. *et al.* Inhibition of antigen processing by the internal repeat region of the Epstein-Barr virus nuclear antigen-1. *Nature* **375**, 685-8 (1995).
60. Levitskaya, J., Sharipo, A., Leonchiks, A., Ciechanover, A. & Masucci, M.G. Inhibition of ubiquitin/proteasome-dependent protein degradation by the Gly-Ala repeat domain of the Epstein-Barr virus nuclear antigen 1. *Proc Natl Acad Sci U S A* **94**, 12616-21 (1997).
61. Wilson, J.B., Bell, J.L. & Levine, A.J. Expression of Epstein-Barr virus nuclear antigen-1 induces B cell neoplasia in transgenic mice. *EMBO J* **15**, 3117-26 (1996).
62. Wilson, J.B. & Levine, A.J. The oncogenic potential of Epstein-Barr virus nuclear antigen 1 in transgenic mice. *Curr Top Microbiol Immunol* **182**, 375-84 (1992).
63. Humme, S. *et al.* The EBV nuclear antigen 1 (EBNA1) enhances B cell immortalization several thousandfold. *Proc Natl Acad Sci U S A* **100**, 10989-94 (2003).
64. Sheu, L.F. *et al.* Enhanced malignant progression of nasopharyngeal carcinoma cells mediated by the expression of Epstein-Barr nuclear antigen 1 in vivo. *J Pathol* **180**, 243-8 (1996).
65. Ling, P.D., Hsieh, J.J., Ruf, I.K., Rawlins, D.R. & Hayward, S.D. EBNA-2 upregulation of Epstein-Barr virus latency promoters and the cellular CD23 promoter utilizes a common targeting intermediate, CBF1. *J Virol* **68**, 5375-83 (1994).

66. Grossman, S.R., Johannsen, E., Tong, X., Yalamanchili, R. & Kieff, E. The Epstein-Barr virus nuclear antigen 2 transactivator is directed to response elements by the J kappa recombination signal binding protein. *Proc Natl Acad Sci U S A* **91**, 7568-72 (1994).
67. Kaiser, C. *et al.* The proto-oncogene c-myc is a direct target gene of Epstein-Barr virus nuclear antigen 2. *J Virol* **73**, 4481-4 (1999).
68. Johannsen, E. *et al.* Epstein-Barr virus nuclear protein 2 transactivation of the latent membrane protein 1 promoter is mediated by J kappa and PU.1. *J Virol* **69**, 253-62 (1995).
69. Wang, F., Tsang, S.F., Kurilla, M.G., Cohen, J.I. & Kieff, E. Epstein-Barr virus nuclear antigen 2 transactivates latent membrane protein LMP1. *J Virol* **64**, 3407-16 (1990).
70. Wang, F. *et al.* Epstein-Barr virus latent membrane protein (LMP1) and nuclear proteins 2 and 3C are effectors of phenotypic changes in B lymphocytes: EBNA-2 and LMP1 cooperatively induce CD23. *J Virol* **64**, 2309-18 (1990).
71. Tomkinson, B., Robertson, E. & Kieff, E. Epstein-Barr virus nuclear proteins EBNA-3A and EBNA-3C are essential for B-lymphocyte growth transformation. *J Virol* **67**, 2014-25 (1993).
72. Dawson, C.W., Rickinson, A.B. & Young, L.S. Epstein-Barr virus latent membrane protein inhibits human epithelial cell differentiation. *Nature* **344**, 777-80 (1990).
73. Kaye, K.M., Izumi, K.M. & Kieff, E. Epstein-Barr virus latent membrane protein 1 is essential for B-lymphocyte growth transformation. *Proc Natl Acad Sci U S A* **90**, 9150-4 (1993).
74. Gires, O. *et al.* Latent membrane protein 1 of Epstein-Barr virus mimics a constitutively active receptor molecule. *EMBO J* **16**, 6131-40 (1997).
75. Kilger, E., Kieser, A., Baumann, M. & Hammerschmidt, W. Epstein-Barr virus-mediated B-cell proliferation is dependent upon latent membrane protein 1, which simulates an activated CD40 receptor. *EMBO J* **17**, 1700-9 (1998).
76. Klein, E., Teramoto, N., Gogolak, P., Nagy, N. & Bjorkholm, M. LMP-1, the Epstein-Barr virus-encoded oncogene with a B cell activating mechanism similar to CD40. *Immunol Lett* **68**, 147-54 (1999).
77. Uchida, J. *et al.* Mimicry of CD40 signals by Epstein-Barr virus LMP1 in B lymphocyte responses. *Science* **286**, 300-3 (1999).
78. Eliopoulos, A.G., Blake, S.M., Floettmann, J.E., Rowe, M. & Young, L.S. Epstein-Barr virus-encoded latent membrane protein 1 activates the JNK pathway through its extreme C terminus via a mechanism involving TRADD and TRAF2. *J Virol* **73**, 1023-35 (1999).
79. Eliopoulos, A.G. *et al.* TRAF1 is a critical regulator of JNK signaling by the TRAF-binding domain of the Epstein-Barr virus-encoded latent infection membrane protein 1 but not CD40. *J Virol* **77**, 1316-28 (2003).
80. Eliopoulos, A.G. & Young, L.S. Activation of the cJun N-terminal kinase (JNK) pathway by the Epstein-Barr virus-encoded latent membrane protein 1 (LMP1). *Oncogene* **16**, 1731-42 (1998).

81. Huen, D.S., Henderson, S.A., Croom-Carter, D. & Rowe, M. The Epstein-Barr virus latent membrane protein-1 (LMP1) mediates activation of NF-kappa B and cell surface phenotype via two effector regions in its carboxy-terminal cytoplasmic domain. *Oncogene* **10**, 549-60 (1995).
82. Kaye, K.M. *et al.* Tumor necrosis factor receptor associated factor 2 is a mediator of NF-kappa B activation by latent infection membrane protein 1, the Epstein-Barr virus transforming protein. *Proc Natl Acad Sci U S A* **93**, 11085-90 (1996).
83. Paine, E., Scheinman, R.I., Baldwin, A.S., Jr. & Raab-Traub, N. Expression of LMP1 in epithelial cells leads to the activation of a select subset of NF-kappa B/Rel family proteins. *J Virol* **69**, 4572-6 (1995).
84. Dawson, C.W., Tramountanis, G., Eliopoulos, A.G. & Young, L.S. Epstein-Barr virus latent membrane protein 1 (LMP1) activates the phosphatidylinositol 3-kinase/Akt pathway to promote cell survival and induce actin filament remodeling. *J Biol Chem* **278**, 3694-704 (2003).
85. Miller, C.L. *et al.* Integral membrane protein 2 of Epstein-Barr virus regulates reactivation from latency through dominant negative effects on protein-tyrosine kinases. *Immunity* **2**, 155-66 (1995).
86. Caldwell, R.G., Brown, R.C. & Longnecker, R. Epstein-Barr virus LMP2A-induced B-cell survival in two unique classes of EmuLMP2A transgenic mice. *J Virol* **74**, 1101-13 (2000).
87. Caldwell, R.G., Wilson, J.B., Anderson, S.J. & Longnecker, R. Epstein-Barr virus LMP2A drives B cell development and survival in the absence of normal B cell receptor signals. *Immunity* **9**, 405-11 (1998).
88. Kang, M.S. & Kieff, E. Epstein-Barr virus latent genes. *Exp Mol Med* **47**, e131 (2015).
89. Tsurumi, T., Fujita, M. & Kudoh, A. Latent and lytic Epstein-Barr virus replication strategies. *Rev Med Virol* **15**, 3-15 (2005).
90. Mellinghoff, I. *et al.* Early events in Epstein-Barr virus genome expression after activation: regulation by second messengers of B cell activation. *Virology* **185**, 922-8 (1991).
91. Sinclair, A.J., Brimmell, M., Shanahan, F. & Farrell, P.J. Pathways of activation of the Epstein-Barr virus productive cycle. *J Virol* **65**, 2237-44 (1991).
92. Packham, G., Brimmell, M., Cook, D., Sinclair, A.J. & Farrell, P.J. Strain variation in Epstein-Barr virus immediate early genes. *Virology* **192**, 541-50 (1993).
93. Takada, K. & Ono, Y. Synchronous and sequential activation of latently infected Epstein-Barr virus genomes. *J Virol* **63**, 445-9 (1989).
94. Flemington, E. & Speck, S.H. Epstein-Barr virus BZLF1 trans activator induces the promoter of a cellular cognate gene, c-fos. *J Virol* **64**, 4549-52 (1990).
95. Flemington, E. & Speck, S.H. Autoregulation of Epstein-Barr virus putative lytic switch gene BZLF1. *J Virol* **64**, 1227-32 (1990).
96. Speck, S.H., Chatila, T. & Flemington, E. Reactivation of Epstein-Barr virus: regulation and function of the BZLF1 gene. *Trends Microbiol* **5**, 399-405 (1997).

97. Lieberman, P.M. & Berk, A.J. In vitro transcriptional activation, dimerization, and DNA-binding specificity of the Epstein-Barr virus Zta protein. *J Virol* **64**, 2560-8 (1990).
98. Ragoczy, T., Heston, L. & Miller, G. The Epstein-Barr virus Rta protein activates lytic cycle genes and can disrupt latency in B lymphocytes. *J Virol* **72**, 7978-84 (1998).
99. Ragoczy, T. & Miller, G. Role of the epstein-barr virus RTA protein in activation of distinct classes of viral lytic cycle genes. *J Virol* **73**, 9858-66 (1999).
100. Murata, T. & Tsurumi, T. Switching of EBV cycles between latent and lytic states. *Rev Med Virol* **24**, 142-53 (2014).
101. Katsumura, K.R., Maruo, S. & Takada, K. EBV lytic infection enhances transformation of B-lymphocytes infected with EBV in the presence of T-lymphocytes. *J Med Virol* **84**, 504-10 (2012).
102. Kalla, M., Schmeinck, A., Bergbauer, M., Pich, D. & Hammerschmidt, W. AP-1 homolog BZLF1 of Epstein-Barr virus has two essential functions dependent on the epigenetic state of the viral genome. *Proc Natl Acad Sci U S A* **107**, 850-5 (2010).
103. Ma, S.D. *et al.* A new model of Epstein-Barr virus infection reveals an important role for early lytic viral protein expression in the development of lymphomas. *J Virol* **85**, 165-77 (2011).
104. Morrison, B.J., Labo, N., Miley, W.J. & Whitby, D. Serodiagnosis for tumor viruses. *Seminars in oncology* **42**, 191-206 (2015).
105. Evans, A.S., Wanat, J. & Niederman, J.C. Failure to demonstrate concomitant antibody changes to viral antigens other than Epstein-Barr virus (EBV) during or after infectious mononucleosis. *Yale J Biol Med* **56**, 203-9 (1983).
106. Coghill, A.E. & Hildesheim, A. Epstein-Barr virus antibodies and the risk of associated malignancies: review of the literature. *American journal of epidemiology* **180**, 687-95 (2014).
107. Cao, S.M. *et al.* Fluctuations of epstein-barr virus serological antibodies and risk for nasopharyngeal carcinoma: a prospective screening study with a 20-year follow-up. *PLoS One* **6**, e19100 (2011).
108. Hildesheim, A. & Wang, C.P. Genetic predisposition factors and nasopharyngeal carcinoma risk: a review of epidemiological association studies, 2000-2011: Rosetta Stone for NPC: genetics, viral infection, and other environmental factors. *Semin Cancer Biol* **22**, 107-16 (2012).
109. Ji, M.F. *et al.* Sustained elevation of Epstein-Barr virus antibody levels preceding clinical onset of nasopharyngeal carcinoma. *Br J Cancer* **96**, 623-30 (2007).
110. Yu, K.J. *et al.* Prognostic utility of anti-EBV antibody testing for defining NPC risk among individuals from high-risk NPC families. *Clin Cancer Res* **17**, 1906-14 (2011).
111. Chien, Y.C. *et al.* Serologic markers of Epstein-Barr virus infection and nasopharyngeal carcinoma in Taiwanese men. *N Engl J Med* **345**, 1877-82 (2001).
112. Liu, Z. *et al.* Two Epstein-Barr virus-related serologic antibody tests in nasopharyngeal carcinoma screening: results from the initial phase of a cluster

randomized controlled trial in Southern China. *Am J Epidemiol* **177**, 242-50 (2013).
113. Hildesheim, A. Invited commentary: Epstein-Barr virus-based screening for the early detection of nasopharyngeal carcinoma: a new frontier. *Am J Epidemiol* **177**, 251-3 (2013).
114. Mueller, N. *et al.* Hodgkin's disease and Epstein-Barr virus. Altered antibody pattern before diagnosis. *The New England journal of medicine* **320**, 689-95 (1989).
115. EVANS, A. PRESENCE OF ELEVATED ANTIBODY TITRES TO EPSTEIN-BARR VIRUS BEFORE HODGKIN'S DISEASE. *The Lancet* **317**, 1183-1186 (1981).
116. Levin, L.I. *et al.* Atypical prediagnosis Epstein-Barr virus serology restricted to EBV-positive Hodgkin lymphoma. *Blood* **120**, 3750-5 (2012).
117. Mueller, N.E., Lennette, E.T., Dupnik, K. & Birmann, B.M. Antibody titers against EBNA1 and EBNA2 in relation to Hodgkin lymphoma and history of infectious mononucleosis. *Int J Cancer* **130**, 2886-91 (2012).
118. Evans, A.S. & Gutensohn, N.M. A population-based case-control study of EBV and other viral antibodies among persons with Hodgkin's disease and their siblings. *Int J Cancer* **34**, 149-57 (1984).
119. Geser, A., de Thé, G., Lenoir, G., Day, N.E. & Williams, E.H. Final case reporting from the Ugandan prospective study of the relationship between EBV and Burkitt's lymphoma. *International journal of cancer* **29**, 397-400 (1982).
120. Besson, C. *et al.* Positive correlation between Epstein-Barr virus viral load and anti-viral capsid immunoglobulin G titers determined for Hodgkin's lymphoma patients and their relatives. *Journal of clinical microbiology* **44**, 47-50 (2006).
121. Besson, C. *et al.* Strong correlations of anti-viral capsid antigen antibody levels in first-degree relatives from families with Epstein-Barr virus-related lymphomas. *The Journal of infectious diseases* **199**, 1121-7 (2009).
122. Liu, M.T. & Yeh, C.Y. Prognostic value of anti-Epstein-Barr virus antibodies in nasopharyngeal carcinoma (NPC). *Radiation medicine* **16**, 113-7.
123. Nystad, T.W. & Myrmel, H. Prevalence of primary versus reactivated Epstein-Barr virus infection in patients with VCA IgG-, VCA IgM- and EBNA-1-antibodies and suspected infectious mononucleosis. *Journal of clinical virology : the official publication of the Pan American Society for Clinical Virology* **38**, 292-7 (2007).
124. Okano, M. *et al.* Frequent association of Epstein-Barr virus in Japanese patients with Burkitt's lymphoma. *Japanese journal of clinical oncology* **22**, 320-4 (1992).
125. Orem, J. *et al.* Epstein-Barr virus viral load and serology in childhood non-Hodgkin's lymphoma and chronic inflammatory conditions in Uganda: implications for disease risk and characteristics. *Journal of medical virology* **86**, 1796-803 (2014).
126. Piriou, E. *et al.* Early age at time of primary Epstein-Barr virus infection results in poorly controlled viral infection in infants from Western Kenya: clues to the etiology of endemic Burkitt lymphoma. *J Infect Dis* **205**, 906-13 (2012).
127. Weinreb, M. *et al.* The consistent association between Epstein-Barr virus and Hodgkin's disease in children in Kenya. *Blood* **87**, 3828-36 (1996).

128. Callan, M.F. The immune response to Epstein-Barr virus. *Microbes Infect* **6**, 937-45 (2004).
129. Landais, E., Saulquin, X. & Houssaint, E. The human T cell immune response to Epstein-Barr virus. *Int J Dev Biol* **49**, 285-92 (2005).
130. Pudney, V.A., Leese, A.M., Rickinson, A.B. & Hislop, A.D. CD8+ immunodominance among Epstein-Barr virus lytic cycle antigens directly reflects the efficiency of antigen presentation in lytically infected cells. *J Exp Med* **201**, 349-60 (2005).
131. Hislop, A.D. & Sabbah, S. CD8+ T cell immunity to Epstein-Barr virus and Kaposi's sarcoma-associated herpes virus. *Semin Cancer Biol* **18**, 416-22 (2008).
132. Jayasooriya, S. *et al.* Early virological and immunological events in asymptomatic Epstein-Barr virus infection in African children. *PLoS Pathog* **11**, e1004746 (2015).
133. Long, H.M. *et al.* MHC II tetramers visualize human CD4+ T cell responses to Epstein-Barr virus infection and demonstrate atypical kinetics of the nuclear antigen EBNA1 response. *J Exp Med* **210**, 933-49 (2013).
134. Long, H.M. *et al.* CD4+ T-cell responses to Epstein-Barr virus (EBV) latent-cycle antigens and the recognition of EBV-transformed lymphoblastoid cell lines. *J Virol* **79**, 4896-907 (2005).
135. Freeman, M.L. *et al.* Cutting edge: activation of virus-specific CD4 T cells throughout gamma-herpesvirus latency. *J Immunol* **187**, 6180-4 (2011).
136. Chang, Y. *et al.* Identification of herpesvirus-like DNA sequences in AIDS-associated Kaposi's sarcoma. *Science* **266**, 1865-9 (1994).
137. Kaposi, M. Idiopathic Multiple Pigmented Sarcoma of the Skin. *CA: A Cancer Journal for Clinicians* **32**, 342-347 (1982).
138. Russo, J.J. *et al.* Nucleotide sequence of the Kaposi sarcoma-associated herpesvirus (HHV8). *Proc Natl Acad Sci U S A* **93**, 14862-7 (1996).
139. Cesarman, E. & Mesri, E.A. Kaposi sarcoma-associated herpesvirus and other viruses in human lymphomagenesis. *Curr Top Microbiol Immunol* **312**, 263-87 (2007).
140. Cesarman, E., Chang, Y., Moore, P.S., Said, J.W. & Knowles, D.M. Kaposi's sarcoma-associated herpesvirus-like DNA sequences in AIDS-related body-cavity-based lymphomas. *N Engl J Med* **332**, 1186-91 (1995).
141. Uldrick, T.S. *et al.* An interleukin-6-related systemic inflammatory syndrome in patients co-infected with Kaposi sarcoma-associated herpesvirus and HIV but without Multicentric Castleman disease. *Clin Infect Dis* **51**, 350-8 (2010).
142. Moore, P.S. & Chang, Y. Detection of herpesvirus-like DNA sequences in Kaposi's sarcoma in patients with and without HIV infection. *N Engl J Med* **332**, 1181-5 (1995).
143. Ferlay, J. *et al.* Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. *Int J Cancer* **127**, 2893-917 (2010).
144. Hayward, G.S. KSHV strains: the origins and global spread of the virus. *Semin Cancer Biol* **9**, 187-99 (1999).
145. Mbulaiteye, S. *et al.* Molecular evidence for mother-to-child transmission of Kaposi sarcoma-associated herpesvirus in Uganda and K1 gene evolution within the host. *J Infect Dis* **193**, 1250-7 (2006).

146. Mbulaiteye, S.M. *et al.* Detection of kaposi sarcoma-associated herpesvirus DNA in saliva and buffy-coat samples from children with sickle cell disease in Uganda. *J Infect Dis* **190**, 1382-6 (2004).
147. Engels, E.A. *et al.* Risk factors for human herpesvirus 8 infection among adults in the United States and evidence for sexual transmission. *J Infect Dis* **196**, 199-207 (2007).
148. Martro, E. *et al.* Risk factors for human Herpesvirus 8 infection and AIDS-associated Kaposi's sarcoma among men who have sex with men in a European multicentre study. *Int J Cancer* **120**, 1129-35 (2007).
149. Chang, Y. *et al.* Kaposi's sarcoma-associated herpesvirus and Kaposi's sarcoma in Africa. Uganda Kaposi's Sarcoma Study Group. *Arch Intern Med* **156**, 202-4 (1996).
150. Gao, S.J. *et al.* KSHV antibodies among Americans, Italians and Ugandans with and without Kaposi's sarcoma. *Nat Med* **2**, 925-8 (1996).
151. Chatlynne, L.G. & Ablashi, D.V. Seroepidemiology of Kaposi's sarcoma-associated herpesvirus (KSHV). *Semin Cancer Biol* **9**, 175-85 (1999).
152. Malope, B.I. *et al.* Transmission of Kaposi sarcoma-associated herpesvirus between mothers and children in a South African population. *J Acquir Immune Defic Syndr* **44**, 351-5 (2007).
153. Mesri, E.A., Cesarman, E. & Boshoff, C. Kaposi's sarcoma and its associated herpesvirus. *Nat Rev Cancer* **10**, 707-19 (2010).
154. Butler, L.M. *et al.* Human herpesvirus 8 infection in children and adults in a population-based study in rural Uganda. *J Infect Dis* **203**, 625-34 (2011).
155. Johnston, C. *et al.* Impact of HIV infection and Kaposi sarcoma on human herpesvirus-8 mucosal replication and dissemination in Uganda. *PLoS One* **4**, e4222 (2009).
156. de-The, G., Bestetti, G., van Beveren, M. & Gessain, A. Prevalence of human herpesvirus 8 infection before the acquired immunodeficiency disease syndrome-related epidemic of Kaposi's sarcoma in East Africa. *J Natl Cancer Inst* **91**, 1888-9 (1999).
157. Dollard, S.C. *et al.* Substantial regional differences in human herpesvirus 8 seroprevalence in sub-Saharan Africa: insights on the origin of the "Kaposi's sarcoma belt". *Int J Cancer* **127**, 2395-401 (2010).
158. Gessain, A. *et al.* Human herpesvirus 8 primary infection occurs during childhood in Cameroon, Central Africa. *Int J Cancer* **81**, 189-92 (1999).
159. Rezza, G. *et al.* Prevalence and risk factors for human herpesvirus 8 infection in northern Cameroon. *Sex Transm Dis* **27**, 159-64 (2000).
160. Engels, E.A. *et al.* Latent class analysis of human herpesvirus 8 assay performance and infection prevalence in sub-saharan Africa and Malta. *Int J Cancer* **88**, 1003-8 (2000).
161. He, J. *et al.* Seroprevalence of human herpesvirus 8 among Zambian women of childbearing age without Kaposi's sarcoma (KS) and mother-child pairs with KS. *J Infect Dis* **178**, 1787-90 (1998).

162. Olsen, S.J., Chang, Y., Moore, P.S., Biggar, R.J. & Melbye, M. Increasing Kaposi's sarcoma-associated herpesvirus seroprevalence with age in a highly Kaposi's sarcoma endemic region, Zambia in 1985. *AIDS* **12**, 1921-5 (1998).
163. Sitas, F. *et al.* Antibodies against human herpesvirus 8 in black South African patients with cancer. *N Engl J Med* **340**, 1863-71 (1999).
164. Cattani, P. *et al.* Age-specific seroprevalence of Human Herpesvirus 8 in Mediterranean regions. *Clin Microbiol Infect* **9**, 274-9 (2003).
165. Larocca, L. *et al.* Prevalence of antibodies to HHV-8 in the general population and in individuals at risk for sexually transmitted and blood-borne infections in Catania, Eastern Sicily. *Infez Med* **13**, 79-85 (2005).
166. Martin, J.N. *et al.* Sexual transmission and the natural history of human herpesvirus 8 infection. *N Engl J Med* **338**, 948-54 (1998).
167. Cao, Y. *et al.* High prevalence of early childhood infection by Kaposi's sarcoma-associated herpesvirus in a minority population in China. *Clin Microbiol Infect* **20**, 475-81 (2014).
168. Wang, H. *et al.* Seroprevalence and risk factors of Kaposi's sarcoma-associated herpesvirus infection among the general Uygur population from south and north region of Xinjiang, China. *Virol J* **8**, 539 (2011).
169. de Souza, V.A. *et al.* Human herpesvirus-8 infection and oral shedding in Amerindian and non-Amerindian populations in the Brazilian Amazon region. *J Infect Dis* **196**, 844-52 (2007).
170. Mbulaiteye, S.M. *et al.* Human herpesvirus 8 infection within families in rural Tanzania. *J Infect Dis* **187**, 1780-5 (2003).
171. Plancoulaine, S. *et al.* Respective roles of serological status and blood specific antihuman herpesvirus 8 antibody levels in human herpesvirus 8 intrafamilial transmission in a highly endemic area. *Cancer Res* **64**, 8782-7 (2004).
172. Plancoulaine, S. *et al.* Human herpesvirus 8 transmission from mother to child and between siblings in an endemic population. *Lancet* **356**, 1062-5 (2000).
173. Alkharsah, K.R., Dedicoat, M., Blasczyk, R., Newton, R. & Schulz, T.F. Influence of HLA alleles on shedding of Kaposi sarcoma-associated herpesvirus in saliva in an African population. *J Infect Dis* **195**, 809-16 (2007).
174. Dedicoat, M. *et al.* Mother-to-child transmission of human herpesvirus-8 in South Africa. *J Infect Dis* **190**, 1068-75 (2004).
175. Olp, L.N. *et al.* Early childhood infection of Kaposi's sarcoma-associated herpesvirus in Zambian households: a molecular analysis. *Int J Cancer* **132**, 1182-90 (2013).
176. Martro, E. *et al.* Comparison of human herpesvirus 8 and Epstein-Barr virus seropositivity among children in areas endemic and non-endemic for Kaposi's sarcoma. *Journal of medical virology* **72**, 126-31 (2004).
177. Crabtree, K.L. *et al.* Risk factors for early childhood infection of human herpesvirus-8 in Zambian children: the role of early childhood feeding practices. *Cancer Epidemiol Biomarkers Prev* **23**, 300-8 (2014).

178. Minhas, V. *et al.* Early childhood infection by human herpesvirus 8 in Zambia and the role of human immunodeficiency virus type 1 coinfection in a highly endemic area. *Am J Epidemiol* **168**, 311-20 (2008).
179. Nalwoga, A. *et al.* Association between malaria exposure and Kaposi's sarcoma-associated herpes virus seropositivity in Uganda. *Trop Med Int Health* **20**, 665-672 (2015).
180. Wakeham, K. *et al.* Risk factors for seropositivity to Kaposi sarcoma-associated herpesvirus among children in Uganda. *J Acquir Immune Defic Syndr* **63**, 228-33 (2013).
181. Wakeham, K. *et al.* Parasite infection is associated with Kaposi's sarcoma associated herpesvirus (KSHV) in Ugandan women. *Infect Agent Cancer* **6**, 15 (2011).
182. Whitby, D. *et al.* Reactivation of Kaposi's sarcoma-associated herpesvirus by natural products from Kaposi's sarcoma endemic regions. *Int J Cancer* **120**, 321-8 (2007).
183. Lin, C.J. *et al.* Intestinal parasites in Kaposi sarcoma patients in Uganda: indication of shared risk factors or etiologic association. *Am J Trop Med Hyg* **78**, 409-12 (2008).
184. Ziegler, J.L. Endemic Kaposi's sarcoma in Africa and local volcanic soils. *Lancet* **342**, 1348-51 (1993).
185. Mbulaiteye, S.M. *et al.* Water, socioeconomic factors, and human herpesvirus 8 infection in Ugandan children and their mothers. *J Acquir Immune Defic Syndr* **38**, 474-9 (2005).
186. Zong, J.C. *et al.* High-level variability in the ORF-K1 membrane protein gene at the left end of the Kaposi's sarcoma-associated herpesvirus genome defines four major virus subtypes and multiple variants or clades in different human populations. *J Virol* **73**, 4156-70 (1999).
187. Cook, P.M. *et al.* Variability and evolution of Kaposi's sarcoma-associated herpesvirus in Europe and Africa. International Collaborative Group. *AIDS* **13**, 1165-76 (1999).
188. Brinkmann, M.M. & Schulz, T.F. Regulation of intracellular signalling by the terminal membrane proteins of members of the Gammaherpesvirinae. *J Gen Virol* **87**, 1047-74 (2006).
189. Brinkmann, M.M., Pietrek, M., Dittrich-Breiholz, O., Kracht, M. & Schulz, T.F. Modulation of host gene expression by the K15 protein of Kaposi's sarcoma-associated herpesvirus. *J Virol* **81**, 42-58 (2007).
190. Brinkmann, M.M. *et al.* Activation of mitogen-activated protein kinase and NF-kappaB pathways by a Kaposi's sarcoma-associated herpesvirus K15 membrane protein. *J Virol* **77**, 9346-58 (2003).
191. Choi, J.K., Lee, B.S., Shim, S.N., Li, M. & Jung, J.U. Identification of the novel K15 gene at the rightmost end of the Kaposi's sarcoma-associated herpesvirus genome. *J Virol* **74**, 436-46 (2000).
192. Hayward, G.S. & Zong, J.C. Modern evolutionary history of the human KSHV genome. *Curr Top Microbiol Immunol* **312**, 1-42 (2007).

193. Beckstead, J.H., Wood, G.S. & Fletcher, V. Evidence for the origin of Kaposi's sarcoma from lymphatic endothelium. *Am J Pathol* **119**, 294-300 (1985).
194. Wang, H.W. *et al.* Kaposi sarcoma herpesvirus-induced cellular reprogramming contributes to the lymphatic endothelial gene expression in Kaposi sarcoma. *Nat Genet* **36**, 687-93 (2004).
195. Friedman-Kien, A.E. & Saltzman, B.R. Clinical manifestations of classical, endemic African, and epidemic AIDS-associated Kaposi's sarcoma. *J Am Acad Dermatol* **22**, 1237-50 (1990).
196. Iscovich, J., Boffetta, P. & Brennan, P. Classic Kaposi's sarcoma as a first primary neoplasm. *Int J Cancer* **80**, 173-7 (1999).
197. Iscovich, J., Boffetta, P. & Brennan, P. Classic Kaposi's sarcoma in Arabs living in Israel, 1970-1993: a population-based incidence study. *Int J Cancer* **77**, 319-21 (1998).
198. Iscovich, J., Boffetta, P., Winkelmann, R., Brennan, P. & Azizi, E. Classic Kaposi's sarcoma in Jews living in Israel, 1961-1989: a population-based incidence study. *AIDS* **12**, 2067-72 (1998).
199. Oettle, A.G. Geographical and racial differences in the frequency of Kaposi's sarcoma as evidence of environmental or genetic causes. *Acta Unio Int Contra Cancrum* **18**, 330-63 (1962).
200. Biggar, R.J. AIDS-related cancers in the era of highly active antiretroviral therapy. *Oncology (Williston Park)* **15**, 439-48; discussion 448-9 (2001).
201. Gaidano, G. & Carbone, A. Primary effusion lymphoma: a liquid phase lymphoma of fluid-filled body cavities. *Adv Cancer Res* **80**, 115-46 (2001).
202. Horenstein, M.G. *et al.* Epstein-Barr virus latent gene expression in primary effusion lymphomas containing Kaposi's sarcoma-associated herpesvirus/human herpesvirus-8. *Blood* **90**, 1186-91 (1997).
203. Liang, C., Lee, J.S. & Jung, J.U. Immune evasion in Kaposi's sarcoma-associated herpes virus associated oncogenesis. *Semin Cancer Biol* **18**, 423-36 (2008).
204. Jenner, R.G., Alba, M.M., Boshoff, C. & Kellam, P. Kaposi's sarcoma-associated herpesvirus latent and lytic gene expression as revealed by DNA arrays. *J Virol* **75**, 891-902 (2001).
205. Jenner, R.G. *et al.* Kaposi's sarcoma-associated herpesvirus-infected primary effusion lymphoma has a plasma cell gene expression profile. *Proc Natl Acad Sci U S A* **100**, 10399-404 (2003).
206. Soulier, J. *et al.* Kaposi's sarcoma-associated herpesvirus-like DNA sequences in multicentric Castleman's disease. *Blood* **86**, 1276-80 (1995).
207. Polizzotto, M.N. *et al.* Human and viral interleukin-6 and other cytokines in Kaposi sarcoma herpesvirus-associated multicentric Castleman disease. *Blood* **122**, 4189-98 (2013).
208. Larroche, C. *et al.* Castleman's disease and lymphoma: report of eight cases in HIV-negative patients and literature review. *Am J Hematol* **69**, 119-26 (2002).
209. Chandran, B. Early events in Kaposi's sarcoma-associated herpesvirus infection of target cells. *J Virol* **84**, 2188-99 (2010).

210. Bagni, R. & Whitby, D. Kaposi's sarcoma-associated herpesvirus transmission and primary infection. *Curr Opin HIV AIDS* **4**, 22-6 (2009).
211. Miller, G., El-Guindy, A., Countryman, J., Ye, J. & Gradoville, L. Lytic cycle switches of oncogenic human gammaherpesviruses. *Adv Cancer Res* **97**, 81-109 (2007).
212. Moore, P.S. & Chang, Y. Molecular virology of Kaposi's sarcoma-associated herpesvirus. *Philos Trans R Soc Lond B Biol Sci* **356**, 499-516 (2001).
213. Kellam, P. *et al.* Identification of a major latent nuclear antigen, LNA-1, in the human herpesvirus 8 genome. *J Hum Virol* **1**, 19-29 (1997).
214. Komatsu, T., Ballestas, M.E., Barbera, A.J. & Kaye, K.M. The KSHV latency-associated nuclear antigen: a multifunctional protein. *Front Biosci* **7**, d726-30 (2002).
215. Dittmer, D. *et al.* A cluster of latently expressed genes in Kaposi's sarcoma-associated herpesvirus. *J Virol* **72**, 8309-15 (1998).
216. Kedes, D.H., Lagunoff, M., Renne, R. & Ganem, D. Identification of the gene encoding the major latency-associated nuclear antigen of the Kaposi's sarcoma-associated herpesvirus. *J Clin Invest* **100**, 2606-10 (1997).
217. Sadler, R. *et al.* A complex translational program generates multiple novel proteins from the latently expressed kaposin (K12) locus of Kaposi's sarcoma-associated herpesvirus. *J Virol* **73**, 5722-30 (1999).
218. Rainbow, L. *et al.* The 222- to 234-kilodalton latent nuclear protein (LNA) of Kaposi's sarcoma-associated herpesvirus (human herpesvirus 8) is encoded by orf73 and is a component of the latency-associated nuclear antigen. *J Virol* **71**, 5915-21 (1997).
219. Verma, S.C., Lan, K. & Robertson, E. Structure and function of latency-associated nuclear antigen. *Curr Top Microbiol Immunol* **312**, 101-36 (2007).
220. Fejer, G. *et al.* The latency-associated nuclear antigen of Kaposi's sarcoma-associated herpesvirus interacts preferentially with the terminal repeats of the genome in vivo and this complex is sufficient for episomal DNA replication. *J Gen Virol* **84**, 1451-62 (2003).
221. Fakhari, F.D., Jeong, J.H., Kanan, Y. & Dittmer, D.P. The latency-associated nuclear antigen of Kaposi sarcoma-associated herpesvirus induces B cell hyperplasia and lymphoma. *J Clin Invest* **116**, 735-42 (2006).
222. Friborg, J., Jr., Kong, W., Hottiger, M.O. & Nabel, G.J. p53 inhibition by the LANA protein of KSHV protects against cell death. *Nature* **402**, 889-94 (1999).
223. Van Dross, R. *et al.* Constitutively active K-cyclin/cdk6 kinase in Kaposi sarcoma-associated herpesvirus-infected cells. *J Natl Cancer Inst* **97**, 656-66 (2005).
224. Direkze, S. & Laman, H. Regulation of growth signalling and cell cycle by Kaposi's sarcoma-associated herpesvirus genes. *Int J Exp Pathol* **85**, 305-19 (2004).
225. Sarek, G. *et al.* Nucleophosmin phosphorylation by v-cyclin-CDK6 controls KSHV latency. *PLoS Pathog* **6**, e1000818 (2010).
226. Field, N. *et al.* KSHV vFLIP binds to IKK-gamma to activate IKK. *J Cell Sci* **116**, 3721-8 (2003).
227. Guasparri, I., Keller, S.A. & Cesarman, E. KSHV vFLIP is essential for the survival of infected lymphoma cells. *J Exp Med* **199**, 993-1003 (2004).

228. Liu, L. *et al.* The human herpes virus 8-encoded viral FLICE inhibitory protein physically associates with and persistently activates the Ikappa B kinase complex. *J Biol Chem* **277**, 13745-51 (2002).
229. Ye, F.C. *et al.* Kaposi's sarcoma-associated herpesvirus latent gene vFLIP inhibits viral lytic replication through NF-kappaB-mediated suppression of the AP-1 pathway: a novel mechanism of virus control of latency. *J Virol* **82**, 4235-49 (2008).
230. McCormick, C. & Ganem, D. The kaposin B protein of KSHV activates the p38/MK2 pathway and stabilizes cytokine mRNAs. *Science* **307**, 739-41 (2005).
231. Purushothaman, P., Uppal, T. & Verma, S.C. Molecular biology of KSHV lytic reactivation. *Viruses* **7**, 116-53 (2015).
232. Sun, R. *et al.* Kinetics of Kaposi's sarcoma-associated herpesvirus gene expression. *J Virol* **73**, 2232-42 (1999).
233. Katano, H., Sato, Y., Itoh, H. & Sata, T. Expression of human herpesvirus 8 (HHV-8)-encoded immediate early protein, open reading frame 50, in HHV-8-associated diseases. *J Hum Virol* **4**, 96-102 (2001).
234. Moore, P.S., Boshoff, C., Weiss, R.A. & Chang, Y. Molecular mimicry of human cytokine and cytokine response pathway genes by KSHV. *Science* **274**, 1739-44 (1996).
235. Ganem, D. KSHV-induced oncogenesis. in *Human Herpesviruses: Biology, Therapy, and Immunoprophylaxis* (eds. Arvin, A. *et al.*) (Cambridge, 2007).
236. Lukac, D.M., Kirshner, J.R. & Ganem, D. Transcriptional activation by the product of open reading frame 50 of Kaposi's sarcoma-associated herpesvirus is required for lytic viral reactivation in B cells. *J Virol* **73**, 9348-61 (1999).
237. Yu, F. *et al.* B cell terminal differentiation factor XBP-1 induces reactivation of Kaposi's sarcoma-associated herpesvirus. *FEBS Lett* **581**, 3485-8 (2007).
238. Wilson, S.J. *et al.* X box binding protein XBP-1s transactivates the Kaposi's sarcoma-associated herpesvirus (KSHV) ORF50 promoter, linking plasma cell differentiation to KSHV reactivation from latency. *J Virol* **81**, 13578-86 (2007).
239. Labo, N. *et al.* Heterogeneity and Breadth of Host Antibody Response to KSHV Infection Demonstrated by Systematic Analysis of the KSHV Proteome. *PLoS Pathogens* **10**(2014).
240. Gao, S.J. *et al.* Seroconversion to antibodies against Kaposi's sarcoma-associated herpesvirus-related latent nuclear antigens before the development of Kaposi's sarcoma. *N Engl J Med* **335**, 233-41 (1996).
241. Ziegler, J. *et al.* Risk factors for Kaposi's sarcoma: a case-control study of HIV-seronegative people in Uganda. *Int J Cancer* **103**, 233-40 (2003).
242. Laney, A.S. *et al.* Human herpesvirus 8 presence and viral load are associated with the progression of AIDS-associated Kaposi's sarcoma. *AIDS* **21**, 1541-5 (2007).
243. Wakeham, K. *et al.* Trends in Kaposi's sarcoma-associated Herpesvirus antibodies prior to the development of HIV-associated Kaposi's sarcoma: a nested case-control study. *Int J Cancer* **136**, 2822-30 (2015).
244. Robey, R.C., Mletzko, S. & Gotch, F.M. The T-Cell Immune Response against Kaposi's Sarcoma-Associated Herpesvirus. *Adv Virol* **2010**, 340356 (2010).

245. Hermans, P. Epidemiology, etiology and pathogenesis, clinical presentations and therapeutic approaches in Kaposi's sarcoma: 15-year lessons from AIDS. *Biomed Pharmacother* **52**, 440-6 (1998).
246. Wang, Q.J. *et al.* Primary human herpesvirus 8 infection generates a broadly specific CD8(+) T-cell response to viral lytic cycle proteins. *Blood* **97**, 2366-73 (2001).
247. Guihot, A. *et al.* Low T cell responses to human herpesvirus 8 in patients with AIDS-related and classic Kaposi sarcoma. *J Infect Dis* **194**, 1078-88 (2006).
248. Barozzi, P. *et al.* Changes in the immune responses against human herpesvirus-8 in the disease course of posttransplant Kaposi sarcoma. *Transplantation* **86**, 738-44 (2008).
249. Lambert, M. *et al.* Differences in the frequency and function of HHV8-specific CD8 T cells between asymptomatic HHV8 infection and Kaposi sarcoma. *Blood* **108**, 3871-80 (2006).
250. Bourboulia, D. *et al.* Short- and long-term effects of highly active antiretroviral therapy on Kaposi sarcoma-associated herpesvirus immune responses and viraemia. *AIDS* **18**, 485-93 (2004).
251. Strickler, H.D. *et al.* Human herpesvirus 8 cellular immune responses in homosexual men. *J Infect Dis* **180**, 1682-5 (1999).
252. Myoung, J. & Ganem, D. Active lytic infection of human primary tonsillar B cells by KSHV and its noncytolytic control by activated CD4+ T cells. *J Clin Invest* **121**, 1130-40 (2011).
253. Robey, R.C. *et al.* The CD8 and CD4 T-cell response against Kaposi's sarcoma-associated herpesvirus is skewed towards early and late lytic antigens. *PLoS One* **4**, e5890 (2009).
254. Allison, A.C. Protection afforded by sickle-cell trait against subtertian malareal infection. *Br Med J* **1**, 290-4 (1954).
255. Comstock, G.W. Tuberculosis in twins: a re-analysis of the Prophit survey. *Am Rev Respir Dis* **117**, 621-4 (1978).
256. Herndon, C.N. & Jennings, R.G. A twin-family study of susceptibility to poliomyelitis. *Am J Hum Genet* **3**, 17-46 (1951).
257. Lin, T.M. *et al.* Hepatitis B virus markers in Chinese twins. *Anticancer Res* **9**, 737-41 (1989).
258. Misch, E.A., Berrington, W.R., Vary, J.C., Jr. & Hawn, T.R. Leprosy and the human genome. *Microbiol Mol Biol Rev* **74**, 589-620 (2010).
259. Sorensen, T.I., Nielsen, G.G., Andersen, P.K. & Teasdale, T.W. Genetic and environmental influences on premature death in adult adoptees. *N Engl J Med* **318**, 727-32 (1988).
260. Tournamille, C., Colin, Y., Cartron, J.P. & Le Van Kim, C. Disruption of a GATA motif in the Duffy gene promoter abolishes erythroid gene expression in Duffy-negative individuals. *Nat Genet* **10**, 224-8 (1995).
261. Samson, M. *et al.* Resistance to HIV-1 infection in caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene. *Nature* **382**, 722-5 (1996).

262. Newport, M.J. *et al.* A mutation in the interferon-gamma-receptor gene and susceptibility to mycobacterial infection. *N Engl J Med* **335**, 1941-9 (1996).
263. Dorman, S.E. & Holland, S.M. Mutation in the signal-transducing chain of the interferon-gamma receptor and susceptibility to mycobacterial infection. *J Clin Invest* **101**, 2364-9 (1998).
264. Altare, F. *et al.* Inherited interleukin 12 deficiency in a child with bacille Calmette-Guerin and Salmonella enteritidis disseminated infection. *J Clin Invest* **102**, 2035-40 (1998).
265. Everitt, A.R. *et al.* IFITM3 restricts the morbidity and mortality associated with influenza. *Nature* **484**, 519-23 (2012).
266. Zhang, Y.H. *et al.* Interferon-induced transmembrane protein-3 genetic variant rs12252-C is associated with severe influenza in Chinese individuals. *Nat Commun* **4**, 1418 (2013).
267. International HapMap, C. The International HapMap Project. *Nature* **426**, 789-96 (2003).
268. Bush, W.S. & Moore, J.H. Chapter 11: Genome-wide association studies. *PLoS Comput Biol* **8**, e1002822 (2012).
269. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G.R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet* **44**, 955-9 (2012).
270. Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).
271. Howie, B., Marchini, J. & Stephens, M. Genotype imputation with thousands of genomes. *G3 (Bethesda)* **1**, 457-70 (2011).
272. Klein, R.J. *et al.* Complement factor H polymorphism in age-related macular degeneration. *Science* **308**, 385-9 (2005).
273. Jostins, L. *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119-24 (2012).
274. Liu, J.Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat Genet* **47**, 979-86 (2015).
275. Replication, D.I.G. *et al.* Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat Genet* **46**, 234-44 (2014).
276. Chapman, S.J. & Hill, A.V. Human genetic susceptibility to infectious disease. *Nat Rev Genet* **13**, 175-88 (2012).
277. Mentzer, A.J., O'Connor, D., Pollard, A.J. & Hill, A.V.S. Searching for the human genetic factors standing in the way of universally effective vaccines. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **370**(2015).
278. Fellay, J. *et al.* A whole-genome association study of major determinants for host control of HIV-1. *Science* **317**, 944-7 (2007).
279. Fellay, J. *et al.* Common genetic variation and the control of HIV-1 in humans. *PLoS Genet* **5**, e1000791 (2009).

280. Pelak, K. *et al.* Host determinants of HIV-1 control in African Americans. *J Infect Dis* **201**, 1141-9 (2010).
281. International, H.I.V.C.S. *et al.* The major genetic determinants of HIV-1 control affect HLA class I peptide presentation. *Science* **330**, 1551-7 (2010).
282. Thomas, R. *et al.* HLA-C cell surface expression and control of HIV/AIDS correlate with a variant upstream of HLA-C. *Nat Genet* **41**, 1290-4 (2009).
283. Apps, R. *et al.* Influence of HLA-C expression level on HIV control. *Science* **340**, 87-91 (2013).
284. McLaren, P.J. *et al.* Polymorphisms of large effect explain the majority of the host genetic contribution to variation of HIV-1 virus load. *Proceedings of the National Academy of Sciences of the United States of America* (2015).
285. Hetherington, S. *et al.* Genetic variations in HLA-B region and hypersensitivity reactions to abacavir. *Lancet* **359**, 1121-2 (2002).
286. Mallal, S. *et al.* Association between presence of HLA-B*5701, HLA-DR7, and HLA-DQ3 and hypersensitivity to HIV-1 reverse-transcriptase inhibitor abacavir. *Lancet* **359**, 727-32 (2002).
287. Kamatani, Y. *et al.* A genome-wide association study identifies variants in the HLA-DP locus associated with chronic hepatitis B in Asians. *Nat Genet* **41**, 591-5 (2009).
288. Matsuura, K., Isogawa, M. & Tanaka, Y. Host genetic variants influencing the clinical course of Hepatitis B virus infection. *Journal of medical virology* (2015).
289. Ge, D. *et al.* Genetic variation in IL28B predicts hepatitis C treatment-induced viral clearance. *Nature* **461**, 399-401 (2009).
290. Fitzmaurice, K. *et al.* Additive effects of HLA alleles and innate immune genes determine viral outcome in HCV infection. *Gut* **64**, 813-9 (2015).
291. Chen, D. *et al.* A systematic investigation of the contribution of genetic variation within the MHC region to HPV seropositivity. *Human molecular genetics* **24**, 2681-8 (2015).
292. Leish, G.E.N.C. *et al.* Common variants in the HLA-DRB1-HLA-DQA1 HLA class II region are associated with susceptibility to visceral leishmaniasis. *Nat Genet* **45**, 208-13 (2013).
293. Dunstan, S.J. *et al.* Variation at HLA-DRB1 is associated with resistance to enteric fever. *Nature genetics* **46**, 1333-6 (2014).
294. Sveinbjornsson, G. *et al.* HLA class II sequence variants influence tuberculosis risk in populations of European ancestry. *Nature genetics* **48**, 318-322 (2016).
295. Blackwell, J.M., Jamieson, S.E. & Burgner, D. HLA and infectious diseases. *Clin Microbiol Rev* **22**, 370-85, Table of Contents (2009).
296. Burdett T (EBI), H.P.N., Hastings E (EBI), Hindorff LA (NHGRI), Junkins HA (NHGRI), Klemm AK (NHGRI), MacArthur J (EBI), Manolio TA (NHGRI), Morales J (EBI), Parkinson H (EBI) and Welter D (EBI). The NHGRI-EBI Catalog of published genome-wide association studies. Available at: http://www.ebi.ac.uk/gwas Accessed: 12/04/2016.

297. Peprah, E., Xu, H., Tekola-Ayele, F. & Royal, C.D. Genome-wide association studies in Africans and African Americans: expanding the framework of the genomics of human traits and disease. *Public health genomics* **18**, 40-51 (2015).
298. Campbell, M.C. & Tishkoff, S.A. The evolution of human genetic and phenotypic variation in Africa. *Curr Biol* **20**, R166-73 (2010).
299. Gurdasani, D. *et al.* The African Genome Variation Project shapes medical genetics in Africa. *Nature* **advance on**(2014).
300. Jallow, M. *et al.* Genome-wide and fine-resolution association analysis of malaria in West Africa. *Nat Genet* **41**, 657-65 (2009).
301. Band, G. *et al.* Imputation-based meta-analysis of severe malaria in three African populations. *PLoS Genet* **9**, e1003509 (2013).
302. Kenyan Bacteraemia Study, G. *et al.* Polymorphism in a lincRNA Associates with a Doubled Risk of Pneumococcal Bacteremia in Kenyan Children. *Am J Hum Genet* **98**, 1092-100 (2016).
303. Spencer, C.C., Su, Z., Donnelly, P. & Marchini, J. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS Genet* **5**, e1000477 (2009).
304. Rotimi, C. *et al.* Research capacity. Enabling the genomic revolution in Africa. *Science (New York, N.Y.)* **344**, 1346-8 (2014).
305. De Flora, S. & La Maestra, S. Epidemiology of cancers of infectious origin and prevention strategies. *Journal of preventive medicine and hygiene* **56**, E15-20 (2015).
306. Jemal, A. *et al.* Cancer burden in Africa and opportunities for prevention. *Cancer* **118**, 4372-84 (2012).
307. Khan, G. & Hashim, M.J. Global burden of deaths from Epstein-Barr virus attributable malignancies 1990-2010. *Infect Agent Cancer* **9**, 38 (2014).
308. Bouvard, V. *et al.* A review of human carcinogens--Part B: biological agents. *Lancet Oncol* **10**, 321-2 (2009).
309. Brown, E.E. *et al.* Correlates of Human Herpesvirus-8 DNA detection among adults in Italy without Kaposi sarcoma. *Int J Epidemiol* **34**, 1110-7 (2005).
310. Pelser, C. *et al.* Socio-economic and other correlates of Kaposi sarcoma-associated herpesvirus seroprevalence among older adults in Sicily. *J Med Virol* **81**, 1938-44 (2009).
311. Vitale, F. *et al.* Kaposi's sarcoma herpes virus and Kaposi's sarcoma in the elderly populations of 3 Mediterranean islands. *Int J Cancer* **91**, 588-91 (2001).
312. Engels, E.a. *et al.* Risk factors for human herpesvirus 8 infection among adults in the United States and evidence for sexual transmission. *The Journal of infectious diseases* **196**, 199-207 (2007).
313. Aoki, Y. & Tosato, G. Interactions between HIV-1 Tat and KSHV. *Curr Top Microbiol Immunol* **312**, 309-26 (2007).
314. Boshoff, C., Whitby, D., Talbot, S. & Weiss, R.A. Etiology of AIDS-related Kaposi's sarcoma and lymphoma. *Oral Dis* **3 Suppl 1**, S129-32 (1997).

315. Merat, R. *et al.* HIV-1 infection of primary effusion lymphoma cell line triggers Kaposi's sarcoma-associated herpesvirus (KSHV) reactivation. *Int J Cancer* **97**, 791-5 (2002).
316. Whitby, D. *et al.* Detection of Kaposi sarcoma associated herpesvirus in peripheral blood of HIV-infected individuals and progression to Kaposi's sarcoma. *Lancet* **346**, 799-802 (1995).
317. Bhutani, M., Polizzotto, M.N., Uldrick, T.S. & Yarchoan, R. Kaposi sarcoma-associated herpesvirus-associated malignancies: epidemiology, pathogenesis, and advances in treatment. *Semin Oncol* **42**, 223-46 (2015).
318. Newton, R. *et al.* Infection with Kaposi's sarcoma-associated herpesvirus (KSHV) and human immunodeficiency virus (HIV) in relation to the risk and clinical presentation of Kaposi's sarcoma in Uganda. *British journal of cancer* **89**, 502-504 (2003).
319. Rohner, E. *et al.* HIV and human herpesvirus 8 co-infection across the globe: Systematic review and meta-analysis. *Int J Cancer* **138**, 45-54 (2016).
320. Ariyoshi, K. *et al.* Kaposi's sarcoma in the Gambia, West Africa is less frequent in human immunodeficiency virus type 2 than in human immunodeficiency virus type 1 infection despite a high prevalence of human herpesvirus 8. *J Hum Virol* **1**, 193-9 (1998).
321. Pinzone, M.R., Berretta, M., Cacopardo, B. & Nunnari, G. Epstein-barr virus- and Kaposi sarcoma-associated herpesvirus-related malignancies in the setting of human immunodeficiency virus infection. *Seminars in oncology* **42**, 258-71 (2015).
322. Engels, E.A. *et al.* Cancer risk in people infected with human immunodeficiency virus in the United States. *Int J Cancer* **123**, 187-94 (2008).
323. Krause, J.R., Robinson, S.D. & Vance, E.A. Multicentric Castleman's disease and HIV. *Proc (Bayl Univ Med Cent)* **27**, 28-30 (2014).
324. Powles, T. *et al.* The role of immune suppression and HHV-8 in the increasing incidence of HIV-associated multicentric Castleman's disease. *Ann Oncol* **20**, 775-9 (2009).
325. Okada, S., Goto, H. & Yotsumoto, M. Current status of treatment for primary effusion lymphoma. *Intractable Rare Dis Res* **3**, 65-74 (2014).
326. Suda, T. *et al.* HHV-8 infection status of AIDS-unrelated and AIDS-associated multicentric Castleman's disease. *Pathol Int* **51**, 671-9 (2001).
327. Greene, W. *et al.* Molecular biology of KSHV in relation to AIDS-associated oncogenesis. *Cancer Treat Res* **133**, 69-127 (2007).
328. Mercader, M. *et al.* Induction of HHV-8 lytic cycle replication by inflammatory cytokines produced by HIV-1-infected T cells. *Am J Pathol* **156**, 1961-71 (2000).
329. Varthakavi, V., Smith, R.M., Deng, H., Sun, R. & Spearman, P. Human immunodeficiency virus type-1 activates lytic cycle replication of Kaposi's sarcoma-associated herpesvirus through induction of KSHV Rta. *Virology* **297**, 270-80 (2002).

330. Mbulaiteye, S.M., Biggar, R.J., Goedert, J.J. & Engels, E.A. Immune deficiency and risk for malignancy among persons with AIDS. *J Acquir Immune Defic Syndr* **32**, 527-33 (2003).
331. Zeng, Y. *et al.* Intracellular Tat of human immunodeficiency virus type 1 activates lytic cycle replication of Kaposi's sarcoma-associated herpesvirus: role of JAK/STAT signaling. *J Virol* **81**, 2401-17 (2007).
332. Biggar, R.J., Chaturvedi, A.K., Goedert, J.J., Engels, E.A. & Study, H.A.C.M. AIDS-related cancer and severity of immunosuppression in persons with AIDS. *J Natl Cancer Inst* **99**, 962-72 (2007).
333. Gallo, R.C. The enigmas of Kaposi's sarcoma. *Science* **282**, 1837-9 (1998).
334. Douglas, J.L., Gustin, J.K., Moses, A.V., Dezube, B.J. & Pantanowitz, L. Kaposi Sarcoma Pathogenesis: A Triad of Viral Infection, Oncogenesis and Chronic Inflammation. *Transl Biomed* **1**(2010).
335. Huang, L.M. *et al.* Reciprocal regulatory interaction between human herpesvirus 8 and human immunodeficiency virus type 1. *J Biol Chem* **276**, 13427-32 (2001).
336. Caselli, E. *et al.* Human herpesvirus 8 enhances human immunodeficiency virus replication in acutely infected cells and induces reactivation in latently infected cells. *Blood* **106**, 2790-7 (2005).
337. Caselli, E., Galvan, M., Cassai, E. & Di Luca, D. Transient expression of human herpesvirus-8 (Kaposi's sarcoma-associated herpesvirus) ORF50 enhances HIV-1 replication. *Intervirology* **46**, 141-9 (2003).
338. Caselli, E. *et al.* Human herpesvirus-8 (Kaposi's sarcoma-associated herpesvirus) ORF50 interacts synergistically with the tat gene product in transactivating the human immunodeficiency virus type 1 LTR. *J Gen Virol* **82**, 1965-70 (2001).
339. Hyun, T.S., Subramanian, C., Cotter, M.A., 2nd, Thomas, R.A. & Robertson, E.S. Latency-associated nuclear antigen encoded by Kaposi's sarcoma-associated herpesvirus interacts with Tat and activates the long terminal repeat of human immunodeficiency virus type 1 in human cells. *J Virol* **75**, 8761-71 (2001).
340. Karijolich, J., Zhao, Y., Peterson, B., Zhou, Q. & Glaunsinger, B. Kaposi's sarcoma-associated herpesvirus ORF45 mediates transcriptional activation of the HIV-1 long terminal repeat via RSK2. *J Virol* **88**, 7024-35 (2014).
341. Carbone, A., Cesarman, E., Spina, M., Gloghini, A. & Schulz, T.F. HIV-associated lymphomas and gamma-herpesviruses. *Blood* **113**, 1213-24 (2009).
342. Carbone, A. & Gloghini, A. AIDS-related lymphomas: from pathogenesis to pathology. *Br J Haematol* **130**, 662-70 (2005).
343. Pietersma, F., Piriou, E. & van Baarle, D. Immune surveillance of EBV-infected B cells and the development of non-Hodgkin lymphomas in immunocompromised patients. *Leuk Lymphoma* **49**, 1028-41 (2008).
344. Wabinga, H.R., Parkin, D.M., Wabwire-Mangen, F. & Mugerwa, J.W. Cancer in Kampala, Uganda, in 1989-91: changes in incidence in the era of AIDS. *Int J Cancer* **54**, 26-36 (1993).
345. Grogg, K.L., Miller, R.F. & Dogan, A. HIV infection and lymphoma. *J Clin Pathol* **60**, 1365-72 (2007).

346. Biggar, R.J. *et al.* Hodgkin lymphoma and immunodeficiency in persons with HIV/AIDS. *Blood* **108**, 3786-91 (2006).
347. Bibas, M. & Antinori, A. EBV and HIV-Related Lymphoma. *Mediterr J Hematol Infect Dis* **1**, e2009032 (2009).
348. Orem, J., Maganda, A., Mbidde, E.K. & Weiderpass, E. Clinical characteristics and outcome of children with Burkitt lymphoma in Uganda according to HIV infection. *Pediatric blood & cancer* **52**, 455-8 (2009).
349. Carbone, A. AIDS-related non-Hodgkin's lymphomas: from pathology and molecular pathogenesis to treatment. *Hum Pathol* **33**, 392-404 (2002).
350. Beral, V., Peterman, T., Berkelman, R. & Jaffe, H. AIDS-associated non-Hodgkin lymphoma. *Lancet* **337**, 805-9 (1991).
351. Ziegler, J.L. *et al.* Non-Hodgkin's lymphoma in 90 homosexual men. Relation to generalized lymphadenopathy and the acquired immunodeficiency syndrome. *N Engl J Med* **311**, 565-70 (1984).
352. Polesel, J. *et al.* Non-Hodgkin lymphoma incidence in the Swiss HIV Cohort Study before and after highly active antiretroviral therapy. *AIDS* **22**, 301-6 (2008).
353. Cote, T.R. *et al.* Non-Hodgkin's lymphoma among people with AIDS: incidence, presentation and public health burden. AIDS/Cancer Study Group. *Int J Cancer* **73**, 645-50 (1997).
354. Diamond, C., Taylor, T.H., Aboumrad, T. & Anton-Culver, H. Changes in acquired immunodeficiency syndrome-related non-Hodgkin lymphoma in the era of highly active antiretroviral therapy: incidence, presentation, treatment, and survival. *Cancer* **106**, 128-35 (2006).
355. Orem, J., Otieno, M.W. & Remick, S.C. AIDS-associated cancer in developing nations. *Curr Opin Oncol* **16**, 468-76 (2004).
356. Rohner, E. *et al.* HHV-8 seroprevalence: a global view. *Syst Rev* **3**, 11 (2014).
357. Coghill, A.E. *et al.* Contribution of HIV infection to mortality among cancer patients in Uganda. *AIDS* **27**, 2933-42 (2013).
358. Trivedi, P. *et al.* Infection of HHV-8+ primary effusion lymphoma cells with a recombinant Epstein-Barr virus leads to restricted EBV latency, altered phenotype, and increased tumorigenicity without affecting TCL1 expression. *Blood* **103**, 313-6 (2004).
359. Boshoff, C. *et al.* Establishing a KSHV+ cell line (BCP-1) from peripheral blood and characterizing its growth in Nod/SCID mice. *Blood* **91**, 1671-9 (1998).
360. Krithivas, A., Young, D.B., Liao, G., Greene, D. & Hayward, S.D. Human herpesvirus 8 LANA interacts with proteins of the mSin3 corepressor complex and negatively regulates Epstein-Barr virus gene expression in dually infected PEL cells. *J Virol* **74**, 9637-45 (2000).
361. Xu, D. *et al.* Epstein-Barr virus inhibits Kaposi's sarcoma-associated herpesvirus lytic replication in primary effusion lymphomas. *J Virol* **81**, 6068-78 (2007).
362. Jiang, Y., Xu, D., Zhao, Y. & Zhang, L. Mutual inhibition between Kaposi's sarcoma-associated herpesvirus and Epstein-Barr virus lytic replication initiators in dually-infected primary effusion lymphoma. *PLoS One* **3**, e1569 (2008).

363. Spadavecchia, S., Gonzalez-Lopez, O., Carroll, K.D., Palmeri, D. & Lukac, D.M. Convergence of Kaposi's sarcoma-associated herpesvirus reactivation with Epstein-Barr virus latency and cellular growth mediated by the notch signaling pathway in coinfected cells. *J Virol* **84**, 10488-500 (2010).
364. Kempkes, B. & Robertson, E.S. Epstein-Barr virus latency: current and future perspectives. *Current opinion in virology* **14**, 138-44 (2015).
365. Mbulaiteye, S.M. Burkitt Lymphoma: beyond discoveries. *Infect Agent Cancer* **8**, 35 (2013).
366. Morrow, R.H. Epidemiological evidence for the role of falciparum malaria in the pathogenesis of Burkitt's lymphoma. *IARC scientific publications*, 177-86 (1985).
367. Njie, R. *et al.* The effects of acute malaria on Epstein-Barr virus (EBV) load and EBV-specific T cell immunity in Gambian children. *The Journal of infectious diseases* **199**, 31-8 (2009).
368. Orem, J., Mbidde, E.K., Lambert, B., de Sanjose, S. & Weiderpass, E. Burkitt's lymphoma in Africa, a review of the epidemiology and etiology. *African health sciences* **7**, 166-75 (2007).
369. Piriou, E. *et al.* Serological evidence for long-term Epstein-Barr virus reactivation in children living in a holoendemic malaria region of Kenya. *J Med Virol* **81**, 1088-93 (2009).
370. Conant, K.L., Marinelli, A. & Kaleeba, J.A. Dangerous liaisons: molecular basis for a syndemic relationship between Kaposi's sarcoma and P. falciparum malaria. *Front Microbiol* **4**, 35 (2013).
371. Beltra, J.-C. & Decaluwe, H. Cytokines and persistent viral infections. *Cytokine* (2016).
372. Chang, K.C. *et al.* Pathogenesis of virus-associated human cancers: Epstein-Barr virus and hepatitis B virus as two examples. *J Formos Med Assoc* **113**, 581-90 (2014).
373. Mbulaiteye, S.M. *et al.* Seroprevalence and risk factors for human herpesvirus 8 infection, rural Egypt. *Emerg Infect Dis* **14**, 586-91 (2008).
374. Mbulaiteye, S.M. *et al.* Risk factors for human herpesvirus 8 seropositivity in the AIDS Cancer Cohort Study. *J Clin Virol* **35**, 442-9 (2006).
375. Kamali, A. *et al.* Seven-year trends in HIV-1 infection rates, and changes in sexual behaviour, among adults in rural Uganda. *AIDS* **14**, 427-34 (2000).
376. KENGEYA-KAYONDO, J.-F. *et al.* Incidence of HIV-1 Infection in Adults and Socio-Demographic Characteristics of Seroconverters in a Rural Population in Uganda: 1990–1994. *International Journal of Epidemiology* **25**, 1077-1082 (1996).
377. Nalwoga, A. *et al.* Nutritional status of children living in a community with high HIV prevalence in rural Uganda: a cross-sectional population-based survey. *Tropical medicine & international health : TM & IH* **15**, 414-22 (2010).
378. Pfeiffer, R.M. *et al.* Geographic heterogeneity of prevalence of the human herpesvirus 8 in sub-Saharan Africa: clues about etiology. *Ann Epidemiol* **20**, 958-63 (2010).
379. Emmanuel, B. *et al.* African Burkitt lymphoma: age-specific risk and correlations with malaria biomarkers. *Am J Trop Med Hyg* **84**, 397-401 (2011).

380. Biggar, R.J. AIDS in subsaharan Africa. *Cancer Detect Prev Suppl* **1**, 487-91 (1987).
381. Biryahwaho, B. *et al.* Sex and geographic patterns of human herpesvirus 8 infection in a nationally representative population-based sample in Uganda. *J Infect Dis* **202**, 1347-53 (2010).
382. Butler, L.M. *et al.* Kaposi sarcoma-associated herpesvirus (KSHV) seroprevalence in population-based samples of African children: evidence for at least 2 patterns of KSHV transmission. *J Infect Dis* **200**, 430-8 (2009).
383. Kalungi, S., Wabinga, H. & Bostad, L. Reactive lymphadenopathy in Ugandan patients and its relationship to EBV and HIV infection. *APMIS* **117**, 302-7 (2009).
384. Mbulaiteye, S.M. *et al.* Spectrum of cancers among HIV-infected persons in Africa: the Uganda AIDS-Cancer Registry Match Study. *Int J Cancer* **118**, 985-90 (2006).
385. Newton, R. *et al.* The sero-epidemiology of Kaposi's sarcoma-associated herpesvirus (KSHV/HHV-8) in adults with cancer in Uganda. *International journal of cancer. Journal international du cancer* **103**, 226-32 (2003).
386. Ogwang, M.D., Bhatia, K., Biggar, R.J. & Mbulaiteye, S.M. Incidence and geographic distribution of endemic Burkitt lymphoma in northern Uganda revisited. *Int J Cancer* **123**, 2658-63 (2008).
387. Parkin, D.M., Nambooze, S., Wabwire-Mangen, F. & Wabinga, H.R. Changing cancer incidence in Kampala, Uganda, 1991-2006. *Int J Cancer* **126**, 1187-95 (2010).
388. Nunn, A.J. *et al.* Mortality associated with HIV-1 infection over five years in a rural Ugandan population: cohort study. *BMJ* **315**, 767-771 (1997).
389. Asiki, G. *et al.* The general population cohort in rural south-western Uganda: a platform for communicable and non-communicable disease studies. *Int J Epidemiol* **42**, 129-41 (2013).
390. Mbulaiteye, S.M., Mahe, C., Ruberantwari, A. & Whitworth, J.A. Generalizability of population-based studies on AIDS: a comparison of newly and continuously surveyed villages in rural southwest Uganda. *Int J Epidemiol* **31**, 961-7 (2002).
391. Richards, A.I. *Economic development and tribal change : a study of immigrant labour in Buganda*, xix, 319 p., 8 leaves of plates (Oxford University Press, Nairobi, 1973).
392. Binnicker, M.J., Jespersen, D.J., Harring, J.A., Rollins, L.O. & Beito, E.M. Evaluation of a multiplex flow immunoassay for detection of epstein-barr virus-specific antibodies. *Clin Vaccine Immunol* **15**, 1410-3 (2008).
393. Mbisa, G.L. *et al.* Detection of antibodies to Kaposi's sarcoma-associated herpesvirus: A new approach using K8.1 ELISA and a newly developed recombinant LANA ELISA. *Journal of Immunological Methods* **356**, 39-46 (2010).
394. Krajden, M. Hepatitis C virus diagnosis and testing. *Can J Public Health* **91 Suppl 1**, S34-9, S36-42 (2000).
395. Krajden, M., McNabb, G. & Petric, M. The laboratory diagnosis of hepatitis B virus. *Can J Infect Dis Med Microbiol* **16**, 65-72 (2005).
396. Quesada, P. *et al.* Hepatitis C virus seroprevalence in the general female population from 8 countries. *J Clin Virol* **68**, 89-93 (2015).

397. R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
398. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-75 (2007).
399. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics* **38**, 904-9 (2006).
400. Heckerman, D. *et al.* Linear mixed model for heritability estimation that explicitly addresses environmental variation. *Proceedings of the National Academy of Sciences* **113**, 7377-7382 (2016).
401. O'Connell, J. *et al.* A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS genetics* **10**, e1004234 (2014).
402. O'Connell, J. *et al.* A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS Genet* **10**, e1004234 (2014).
403. Visscher, P.M. *et al.* Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS Genet* **2**, e41 (2006).
404. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* **42**, 565-9 (2010).
405. Schweitzer, A., Horn, J., Mikolajczyk, R.T., Krause, G. & Ott, J.J. Estimations of worldwide prevalence of chronic hepatitis B virus infection: a systematic review of data published between 1965 and 2013. *Lancet* **386**, 1546-55 (2015).
406. Karoney, M.J. & Siika, A.M. Hepatitis C virus (HCV) infection in Africa: a review. *Pan Afr Med J* **14**, 44 (2013).
407. Shebl, F.M. *et al.* Population-based assessment of kaposi sarcoma-associated herpesvirus DNA in plasma among Ugandans. *J Med Virol* **85**, 1602-10 (2013).
408. Zeng, P. *et al.* Statistical analysis for genome-wide association study. *J Biomed Res* **29**, 285-97 (2015).
409. Chen, C.J. *et al.* Genetic variance and heritability of temperament among Chinese twin infants. *Acta Genet Med Gemellol (Roma)* **39**, 485-90 (1990).
410. Zaitlen, N. *et al.* Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. *PLoS Genet* **9**, e1003520 (2013).
411. Rubicz, R. *et al.* Genetic factors influence serological measures of common infections. *Hum Hered* **72**, 133-41 (2011).
412. Rubicz, R. *et al.* A genome-wide integrative genomic study localizes genetic factors influencing antibodies against Epstein-Barr virus nuclear antigen 1 (EBNA-1). *PLoS Genet* **9**, e1003147 (2013).
413. Zhou, Y. *et al.* Genetic loci for Epstein-Barr virus nuclear antigen-1 are associated with risk of multiple sclerosis. *Multiple sclerosis (Houndmills, Basingstoke, England)* (2016).
414. Rubicz, R. *et al.* Genome-wide genetic investigation of serological measures of common infections. *Eur J Hum Genet* **23**, 1544-8 (2015).
415. Bei, J.X., Jia, W.H. & Zeng, Y.X. Familial and large-scale case-control studies identify genes associated with nasopharyngeal carcinoma. *Semin Cancer Biol* **22**, 96-106 (2012).

416. Hjalgrim, H. *et al.* Familial clustering of Hodgkin lymphoma and multiple sclerosis. *J Natl Cancer Inst* **96**, 780-4 (2004).
417. Houldcroft, C.J. & Kellam, P. Host genetics of Epstein-Barr virus infection, latency and disease. *Rev Med Virol* **25**, 71-84 (2015).
418. Hurme, M. & Helminen, M. Polymorphism of the IL-1 gene complex in Epstein-Barr virus seronegative and seropositive adult blood donors. *Scand J Immunol* **48**, 219-22 (1998).
419. Hatta, K. *et al.* Association of transforming growth factor-beta1 gene polymorphism in the development of Epstein-Barr virus-related hematologic diseases. *Haematologica* **92**, 1470-4 (2007).
420. Suthanthiran, M. *et al.* Transforming growth factor-beta 1 hyperexpression in African-American hypertensives: A novel mediator of hypertension and/or target organ damage. *Proc Natl Acad Sci U S A* **97**, 3479-84 (2000).
421. Yamada, Y. *et al.* Association of a polymorphism of the transforming growth factor-beta1 gene with genetic susceptibility to osteoporosis in postmenopausal Japanese women. *J Bone Miner Res* **13**, 1569-76 (1998).
422. Blomhoff, H.K., Smeland, E., Mustafa, A.S., Godal, T. & Ohlsson, R. Epstein-Barr virus mediates a switch in responsiveness to transforming growth factor, type beta, in cells of the B cell lineage. *Eur J Immunol* **17**, 299-301 (1987).
423. Helminen, M., Lahdenpohja, N. & Hurme, M. Polymorphism of the interleukin-10 gene is associated with susceptibility to Epstein-Barr virus infection. *The Journal of infectious diseases* **180**, 496-9 (1999).
424. Helminen, M.E., Kilpinen, S., Virta, M. & Hurme, M. Susceptibility to primary Epstein-Barr virus infection is associated with interleukin-10 gene promoter polymorphism. *The Journal of infectious diseases* **184**, 777-80 (2001).
425. Minnicelli, C. *et al.* Relationship of Epstein-Barr virus and interleukin 10 promoter polymorphisms with the risk and clinical outcome of childhood Burkitt lymphoma. *PLoS One* **7**, e46005 (2012).
426. Yasui, Y. *et al.* Association of Epstein-Barr virus antibody titers with a human IL-10 promoter polymorphism in Japanese women. *Journal of autoimmune diseases* **5**, 2 (2008).
427. Munro, L.R. *et al.* Polymorphisms in the interleukin-10 and interferon gamma genes in Hodgkin lymphoma. *Leuk Lymphoma* **44**, 2083-8 (2003).
428. Wu, M.S. *et al.* Interleukin-10 genotypes associate with the risk of gastric carcinoma in Taiwanese Chinese. *Int J Cancer* **104**, 617-23 (2003).
429. Oduor, C.I. *et al.* Interleukin-6 and interleukin-10 gene promoter polymorphisms and risk of endemic Burkitt lymphoma. *Am J Trop Med Hyg* **91**, 649-54 (2014).
430. Durovic, B. *et al.* Epstein-Barr virus negativity among individuals older than 60 years is associated with HLA-C and HLA-Bw4 variants and tonsillectomy. *J Virol* **87**, 6526-9 (2013).
431. Diepstra, A. *et al.* Association with HLA class I in Epstein-Barr-virus-positive and with HLA class III in Epstein-Barr-virus-negative Hodgkin's lymphoma. *Lancet* **365**, 2216-24 (2005).

432. Niens, M. *et al.* The human leukocyte antigen class I region is associated with EBV-positive Hodgkin's lymphoma: HLA-A and HLA complex group 9 are putative candidate genes. *Cancer Epidemiol Biomarkers Prev* **15**, 2280-4 (2006).
433. Niens, M. *et al.* HLA-A*02 is associated with a reduced risk and HLA-A*01 with an increased risk of developing EBV+ Hodgkin lymphoma. *Blood* **110**, 3310-5 (2007).
434. Hjalgrim, H. *et al.* HLA-A alleles and infectious mononucleosis suggest a critical role for cytotoxic T-cell response in EBV-related Hodgkin lymphoma. *Proc Natl Acad Sci U S A* **107**, 6400-5 (2010).
435. McAulay, K.A. *et al.* HLA class I polymorphisms are associated with development of infectious mononucleosis upon primary EBV infection. *The Journal of clinical investigation* **117**, 3042-8 (2007).
436. Claire Simon, K., Schmidt, H., Loud, S. & Ascherio, A. Epstein-Barr virus candidate genes and multiple sclerosis. *Multiple sclerosis and related disorders* **4**, 60-4 (2015).
437. Nielsen, T.R., Pedersen, M., Rostgaard, K., Frisch, M. & Hjalgrim, H. Correlations between Epstein-Barr virus antibody levels and risk factors for multiple sclerosis in healthy individuals. *Mult Scler* **13**, 420-3 (2007).
438. Huang, X. *et al.* HLA-A*02:07 is a protective allele for EBV negative and a susceptibility allele for EBV positive classical Hodgkin lymphoma in China. *PLoS One* **7**, e31865 (2012).
439. Huang, X. *et al.* HLA associations in classical Hodgkin lymphoma: EBV status matters. *PLoS One* **7**, e39986 (2012).
440. Johnson, P.C.D. *et al.* Modeling HLA associations with EBV-positive and -negative Hodgkin lymphoma suggests distinct mechanisms in disease pathogenesis. *International journal of cancer. Journal international du cancer* (2015).
441. Friborg, J.T. *et al.* Mannose-binding lectin genotypes and susceptibility to epstein-barr virus infection in infancy. *Clin Vaccine Immunol* **17**, 1484-7 (2010).
442. Shen, G.-P. *et al.* Human genetic variants of homologous recombination repair genes first found to be associated with Epstein-Barr virus antibody titers in healthy Cantonese. *International journal of cancer. Journal international du cancer* **129**, 1459-66 (2011).
443. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* **42**, D1001-6 (2014).
444. Pedergnana, V. *et al.* Combined linkage and association studies show that HLA class II variants control levels of antibodies against Epstein-Barr virus antigens. *PLoS One* **9**, e102501 (2014).
445. Hammer, C. *et al.* Amino Acid Variation in HLA Class II Proteins Is a Major Determinant of Humoral Response to Common Viruses. *Am J Hum Genet* **97**, 738-43 (2015).
446. Houldcroft, C.J. *et al.* Host genetic variants and gene expression patterns associated with Epstein-Barr virus copy number in lymphoblastoid cell lines. *PLoS One* **9**, e108384 (2014).

447. Urayama, K.Y. *et al.* Genome-wide association study of classical Hodgkin lymphoma and Epstein-Barr virus status-defined subgroups. *J Natl Cancer Inst* **104**, 240-53 (2012).
448. Tang, M. *et al.* The principal genetic determinants for nasopharyngeal carcinoma in China involve the HLA class I antigen recognition groove. *PLoS Genet* **8**, e1003103 (2012).
449. Howie, B.N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS genetics* **5**, e1000529 (2009).
450. Zhou, X. & Stephens, M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature methods* **11**, 407-9 (2014).
451. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nature genetics* **44**, 369-75, S1-3 (2012).
452. Morris, A.P. Transethnic meta-analysis of genomewide association studies. *Genetic epidemiology* **35**, 809-22 (2011).
453. Hammer, C. *et al.* Amino Acid Variation in HLA Class II Proteins Is a Major Determinant of Humoral Response to Common Viruses. *American journal of human genetics* (2015).
454. Wang, X. *et al.* Comparing methods for performing trans-ethnic meta-analysis of genome-wide association studies. *Human molecular genetics* **22**, 2303-11 (2013).
455. Willer, C.J., Li, Y. & Abecasis, G.R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190-1 (2010).
456. Charles, B.A. *et al.* A genome-wide association study of serum uric acid in African Americans. *BMC Med Genomics* **4**, 17 (2011).
457. Maller, J.B. *et al.* Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nature genetics* **44**, 1294-301 (2012).
458. Consortium, G.T. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580-5 (2013).
459. Zhou, X. & Stephens, M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nat Methods* **11**, 407-9 (2014).
460. Rippe, V. *et al.* Identification of a gene rearranged by 2p21 aberrations in thyroid adenomas. *Oncogene* **22**, 6111-4 (2003).
461. Hatton, O.L., Harris-Arnold, A., Schaffert, S., Krams, S.M. & Martinez, O.M. The interplay between Epstein-Barr virus and B lymphocytes: implications for infection, immunity, and disease. *Immunol Res* **58**, 268-76 (2014).
462. Snow, A.L. *et al.* Resistance to Fas-mediated apoptosis in EBV-infected B cell lymphomas is due to defects in the proximal Fas signaling pathway. *J Immunol* **167**, 5404-11 (2001).
463. Snow, A.L. *et al.* EBV can protect latently infected B cell lymphomas from death receptor-induced apoptosis. *J Immunol* **177**, 3283-93 (2006).
464. Zeggini, E. *et al.* Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet* **40**, 638-45 (2008).

465. Chen, Z.J. *et al.* Genome-wide association study identifies susceptibility loci for polycystic ovary syndrome on chromosome 2p16.3, 2p21 and 9q33.3. *Nat Genet* **43**, 55-9 (2011).
466. Franke, A. *et al.* Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat Genet* **42**, 1118-25 (2010).
467. Eeles, R.A. *et al.* Identification of seven new prostate cancer susceptibility loci through a genome-wide association study. *Nat Genet* **41**, 1116-21 (2009).
468. Zhao, Y. *et al.* GALC gene is downregulated by promoter hypermethylation in Epstein-Barr virus-associated nasopharyngeal carcinoma. *Oncology Reports* **34**, 1369-1378 (2015).
469. Chu, P.J., Robertson, H.M. & Best, P.M. Calcium channel gamma subunits provide insights into the evolution of this gene family. *Gene* **280**, 37-48 (2001).
470. Dugas, B. *et al.* Activation and infection of B cells by Epstein-Barr virus. Role of calcium mobilization and of protein kinase C translocation. *J Immunol* **141**, 4344-51 (1988).
471. Dugas, B., Mencia-Huerta, J.M., Braquet, P., Galanaud, P. & Delfraissy, J.F. Extracellular but not intracellular calcium mobilization is required for Epstein-Barr virus-containing supernatant-induced B cell activation. *Eur J Immunol* **19**, 1867-71 (1989).
472. Curtis, D. *et al.* Case-case genome-wide association analysis shows markers differentially associated with schizophrenia and bipolar disorder and implicates calcium channel genes. *Psychiatr Genet* **21**, 1-4 (2011).
473. Jia, X. *et al.* Imputing amino acid polymorphisms in human leukocyte antigens. *PloS one* **8**, e64683 (2013).
474. de Sanjose, S. *et al.* Geographic variation in the prevalence of Kaposi sarcoma-associated herpesvirus and risk factors for transmission. *The Journal of infectious diseases* **199**, 1449-56 (2009).
475. Guttman-Yassky, E. *et al.* Familial clustering of classic Kaposi sarcoma. *J Infect Dis* **189**, 2023-6 (2004).
476. Kaasinen, E. *et al.* Nationwide registry-based analysis of cancer clustering detects strong familial occurrence of Kaposi sarcoma. *PLoS One* **8**, e55209 (2013).
477. Jackson, C.C. *et al.* Kaposi Sarcoma of Childhood: Inborn or Acquired Immunodeficiency to Oncogenic HHV-8. *Pediatr Blood Cancer* **63**, 392-7 (2016).
478. Camcioglu, Y. *et al.* HHV-8-associated Kaposi sarcoma in a child with IFNgammaR1 deficiency. *J Pediatr* **144**, 519-23 (2004).
479. Picard, C. *et al.* Kaposi's sarcoma in a child with Wiskott-Aldrich syndrome. *Eur J Pediatr* **165**, 453-7 (2006).
480. Sahin, G. *et al.* Classic Kaposi Sarcoma in 3 Unrelated Turkish Children Born to Consanguineous Kindreds. *Pediatrics* **125**, e704-e708 (2010).
481. Byun, M. *et al.* Whole-exome sequencing-based discovery of STIM1 deficiency in a child with fatal classic Kaposi sarcoma. *J Exp Med* **207**, 2307-12 (2010).
482. Byun, M. *et al.* Inherited human OX40 deficiency underlying classic Kaposi sarcoma of childhood. *J Exp Med* **210**, 1743-59 (2013).

483. Pedergnana, V. *et al.* A major locus on chromosome 3p22 conferring predisposition to human herpesvirus 8 infection. *Eur J Hum Genet* **20**, 690-5 (2012).
484. Plancoulaine, S., Gessain, A., van Beveren, M., Tortevoye, P. & Abel, L. Evidence for a recessive major gene predisposing to human herpesvirus 8 (HHV-8) infection in a population in which HHV-8 is endemic. *J Infect Dis* **187**, 1944-50 (2003).
485. Alcais, A. *et al.* Life-threatening infectious diseases of childhood: single-gene inborn errors of immunity? *Ann N Y Acad Sci* **1214**, 18-33 (2010).
486. Delves, P.J. & Roitt, I.M. The immune system. Second of two parts. *N Engl J Med* **343**, 108-17 (2000).
487. Chang, J., Renne, R., Dittmer, D. & Ganem, D. Inflammatory cytokines and the reactivation of Kaposi's sarcoma-associated herpesvirus lytic replication. *Virology* **266**, 17-25 (2000).
488. Ensoli, B., Barillari, G. & Gallo, R.C. Cytokines and growth factors in the pathogenesis of AIDS-associated Kaposi's sarcoma. *Immunol Rev* **127**, 147-55 (1992).
489. Brulois, K. & Jung, J.U. Interplay between Kaposi's sarcoma-associated herpesvirus and the innate immune system. *Cytokine Growth Factor Rev* **25**, 597-609 (2014).
490. Coscoy, L. Immune evasion by Kaposi's sarcoma-associated herpesvirus. *Nat Rev Immunol* **7**, 391-401 (2007).
491. Lehrnbecher, T.L. *et al.* Variant genotypes of FcgammaRIIIA influence the development of Kaposi's sarcoma in HIV-infected men. *Blood* **95**, 2386-90 (2000).
492. Brown, E.E. *et al.* A common genetic variant in FCGR3A-V158F and risk of Kaposi sarcoma herpesvirus infection and classic Kaposi sarcoma. *Cancer Epidemiol Biomarkers Prev* **14**, 633-7 (2005).
493. Foster, C.B. *et al.* An IL6 promoter polymorphism is associated with a lifetime risk of development of Kaposi sarcoma in men infected with human immunodeficiency virus. *Blood* **96**, 2562-7 (2000).
494. Fishman, D. *et al.* The effect of novel polymorphisms in the interleukin-6 (IL-6) gene on IL-6 transcription and plasma IL-6 levels, and an association with systemic-onset juvenile chronic arthritis. *J Clin Invest* **102**, 1369-76 (1998).
495. van der Kuyl, A.C. *et al.* An IL-8 gene promoter polymorphism is associated with the risk of the development of AIDS-related Kaposi's sarcoma: a case-control study. *AIDS* **18**, 1206-8 (2004).
496. Lane, B.R. *et al.* Interleukin-8 and growth-regulated oncogene alpha mediate angiogenesis in Kaposi's sarcoma. *J Virol* **76**, 11570-83 (2002).
497. Masood, R. *et al.* Interleukin 8 is an autocrine growth factor and a surrogate marker for Kaposi's sarcoma. *Clin Cancer Res* **7**, 2693-702 (2001).
498. Matsumoto, T. *et al.* Elevated serum levels of IL-8 in patients with HIV infection. *Clin Exp Immunol* **93**, 149-51 (1993).
499. Brown, E.E. *et al.* Associations of classic Kaposi sarcoma with common variants in genes that modulate host immunity. *Cancer Epidemiol Biomarkers Prev* **15**, 926-34 (2006).

500. Brown, E.E. *et al.* Host immunogenetics and control of human herpesvirus-8 infection. *J Infect Dis* **193**, 1054-62 (2006).
501. Brunson, M.E., Balakrishnan, K. & Penn, I. HLA and Kaposi's sarcoma in solid organ transplantation. *Hum Immunol* **29**, 56-63 (1990).
502. Contu, L. & Cerimele, P. HLA and classic Kaposi's sarcoma in Sardinia. *J Am Acad Dermatol* **30**, 508-9 (1994).
503. Crosslin, D.R. *et al.* Genetic variation in the HLA region is associated with susceptibility to herpes zoster. *Genes Immun* **16**, 1-7 (2015).
504. Delahaye-Sourdeix, M. *et al.* A Novel Risk Locus at 6p21.3 for Epstein-Barr Virus-Positive Hodgkin Lymphoma. *Cancer Epidemiol Biomarkers Prev* **24**, 1838-43 (2015).
505. Fitzmaurice, K. *et al.* Additive effects of HLA alleles and innate immune genes determine viral outcome in HCV infection. *Gut* **64**, 813-9 (2015).
506. McLaren, P.J. & Carrington, M. The impact of host genetic variation on infection with HIV-1. *Nature immunology* **16**, 577-83 (2015).
507. Tekola Ayele, F. *et al.* HLA class II locus and susceptibility to podoconiosis. *N Engl J Med* **366**, 1200-8 (2012).
508. Melbye, M., Kestens, L., Biggar, R.J., Schreuder, G.M. & Gigase, P.L. HLA studies of endemic African Kaposi's sarcoma patients and matched controls: no association with HLA-DR5. *Int J Cancer* **39**, 182-4 (1987).
509. Robbins, E., Cohen, N., Contu, L. & Dausset, J. A study of the HLA-D region in patients with classic Kaposi's sarcoma. *Immunogenetics* **24**, 115-7 (1986).
510. Papasteriades, C. *et al.* Histocompatibility antigens HLA-A, -B, -DR in Greek patients with Kaposi's sarcoma. *Tissue Antigens* **24**, 313-5 (1984).
511. Contu, L., Cerimele, D., Pintus, A., Cottoni, F. & La Nasa, G. HLA and Kaposi's sarcoma in Sardinia. *Tissue Antigens* **23**, 240-5 (1984).
512. Pollack, M.S. *et al.* Frequencies of HLA and Gm immunogenetic markers in Kaposi's sarcoma. *Tissue Antigens* **21**, 1-8 (1983).
513. Ohara, N. & Chang, S.W. Kaposi's sarcoma and the HLA-Dr5 alloantigen. *Ann Intern Med* **97**, 617 (1982).
514. Ioannidis, J.P., Skolnik, P.R., Chalmers, T.C. & Lau, J. Human leukocyte antigen associations of epidemic Kaposi's sarcoma. *AIDS* **9**, 649-51 (1995).
515. Masala, M.V. *et al.* Classic Kaposi's sarcoma in Sardinia: HLA positive and negative associations. *Int J Dermatol* **44**, 743-5 (2005).
516. Dorak, M.T. *et al.* HLA-B, -DRB1/3/4/5, and -DQB1 gene polymorphisms in human immunodeficiency virus-related Kaposi's sarcoma. *Journal of medical virology* **76**, 302-10 (2005).
517. Guerini, F.R. *et al.* Association of HLA-DRB1 and -DQB1 with Classic Kaposi's Sarcoma in Mainland Italy. *Cancer Genomics Proteomics* **3**, 191-196 (2006).
518. Aissani, B. *et al.* SNP screening of central MHC-identified HLA-DMB as a candidate susceptibility gene for HIV-related Kaposi's sarcoma. *Genes and immunity* **15**, 424-9 (2014).

519. Goedert, J.J. *et al.* Risk of Classic Kaposi Sarcoma With Combinations of Killer Immunoglobulin-Like Receptor and Human Leukocyte Antigen Loci: A Population-Based Case-control Study. *The Journal of infectious diseases* **213**, 432-8 (2016).
520. Aavikko, M. *et al.* Whole-Genome Sequencing Identifies STAT4 as a Putative Susceptibility Gene in Classic Kaposi Sarcoma. *J Infect Dis* **211**, 1842-51 (2015).
521. Takeda, K. & Akira, S. STAT family of transcription factors in cytokine-mediated biological responses. *Cytokine Growth Factor Rev* **11**, 199-207 (2000).
522. Zhang, F. & Boothby, M. T helper type 1-specific Brg1 recruitment and remodeling of nucleosomes positioned at the IFN-gamma promoter are Stat4 dependent. *J Exp Med* **203**, 1493-505 (2006).
523. Yang, D. *et al.* Interleukin 1 receptor-associated kinase 1 (IRAK1) mutation is a common, essential driver for Kaposi sarcoma herpesvirus lymphoma. *Proceedings of the National Academy of Sciences of the United States of America* **111**, E4762-8 (2014).
524. Gregory, S.M. *et al.* Toll-like receptor signaling controls reactivation of KSHV from latency. *Proc Natl Acad Sci U S A* **106**, 11725-30 (2009).
525. Cannon, M.J. *et al.* Risk factors for Kaposi's sarcoma in men seropositive for both human herpesvirus 8 and human immunodeficiency virus. *AIDS* **17**, 215-22 (2003).
526. Butler, L.M. *et al.* Kaposi's sarcoma-associated herpesvirus inhibits expression and function of endothelial cell major histocompatibility complex class II via suppressor of cytokine signaling 3. *J Virol* **86**, 7158-66 (2012).
527. Sabbah, S. *et al.* T-cell immunity to Kaposi sarcoma-associated herpesvirus: recognition of primary effusion lymphoma by LANA-specific CD4+ T cells. *Blood* **119**, 2083-92 (2012).
528. Thakker, S. *et al.* Kaposi's Sarcoma-Associated Herpesvirus Latency-Associated Nuclear Antigen Inhibits Major Histocompatibility Complex Class II Expression by Disrupting Enhanceosome Assembly through Binding with the Regulatory Factor X Complex. *J Virol* **89**, 5536-56 (2015).
529. Coit, P. *et al.* Ethnicity-specific epigenetic variation in naive CD4+ T cells and the susceptibility to autoimmunity. *Epigenetics Chromatin* **8**, 49 (2015).
530. Campbell, M., Kung, H.J. & Izumiya, Y. Long non-coding RNA and epigenetic gene regulation of KSHV. *Viruses* **6**, 4165-77 (2014).
531. Heward, J.A. & Lindsay, M.A. Long non-coding RNAs in the regulation of the immune response. *Trends Immunol* **35**, 408-19 (2014).
532. Rice, G.I. *et al.* Mutations involved in Aicardi-Goutieres syndrome implicate SAMHD1 as regulator of the innate immune response. *Nat Genet* **41**, 829-32 (2009).
533. Laguette, N. *et al.* SAMHD1 is the dendritic- and myeloid-cell-specific HIV-1 restriction factor counteracted by Vpx. *Nature* **474**, 654-7 (2011).
534. Lahouassa, H. *et al.* SAMHD1 restricts the replication of human immunodeficiency virus type 1 by depleting the intracellular pool of deoxynucleoside triphosphates. *Nat Immunol* **13**, 223-8 (2012).

535. Descours, B. *et al.* SAMHD1 restricts HIV-1 reverse transcription in quiescent CD4(+) T-cells. *Retrovirology* **9**, 87 (2012).
536. Bhattacharya, M.R. *et al.* TMEM184b Promotes Axon Degeneration and Neuromuscular Junction Maintenance. *J Neurosci* **36**, 4681-9 (2016).
537. Latourelle, J.C., Dumitriu, A., Hadzi, T.C., Beach, T.G. & Myers, R.H. Evaluation of Parkinson disease risk variants as expression-QTLs. *PLoS One* **7**, e46199 (2012).
538. Grozeva, D. *et al.* Targeted Next-Generation Sequencing Analysis of 1,000 Individuals with Intellectual Disability. *Hum Mutat* **36**, 1197-204 (2015).
539. Halgren, C. *et al.* Corpus callosum abnormalities, intellectual disability, speech impairment, and autism in patients with haploinsufficiency of ARID1B. *Clin Genet* **82**, 248-55 (2012).
540. Hoyer, J. *et al.* Haploinsufficiency of ARID1B, a member of the SWI/SNF-a chromatin-remodeling complex, is a frequent cause of intellectual disability. *Am J Hum Genet* **90**, 565-72 (2012).
541. Margolskee, E. *et al.* Genetic landscape of T- and NK-cell post-transplant lymphoproliferative disorders. *Oncotarget* (2016).
542. Schatz, J.H. *et al.* Targeted mutational profiling of peripheral T-cell lymphoma not otherwise specified highlights new mechanisms in a heterogeneous pathogenesis. *Leukemia* **29**, 237-41 (2015).
543. Fujimoto, A. *et al.* Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators. *Nat Genet* **44**, 760-4 (2012).
544. Lieberman, P.M. Keeping it quiet: chromatin control of gammaherpesvirus latency. *Nat Rev Microbiol* **11**, 863-75 (2013).
545. Lu, F. *et al.* Chromatin remodeling of the Kaposi's sarcoma-associated herpesvirus ORF50 promoter correlates with reactivation from latency. *J Virol* **77**, 11425-35 (2003).
546. Mercier, A., Arias, C., Madrid, A.S., Holdorf, M.M. & Ganem, D. Site-specific association with host and viral chromatin by Kaposi's sarcoma-associated herpesvirus LANA and its reversal during lytic reactivation. *J Virol* **88**, 6762-77 (2014).
547. Van den Broeke, C. & Favoreel, H.W. Actin' up: herpesvirus interactions with Rho GTPase signaling. *Viruses* **3**, 278-92 (2011).
548. Van den Broeke, C., Jacob, T. & Favoreel, H.W. Rho'ing in and out of cells: viral interactions with Rho GTPase signaling. *Small GTPases* **5**, e28318 (2014).
549. Montaner, S. *et al.* The small GTPase Rac1 links the Kaposi sarcoma-associated herpesvirus vGPCR to cytokine secretion and paracrine neoplasia. *Blood* **104**, 2903-11 (2004).
550. Veettil, M.V. *et al.* RhoA-GTPase facilitates entry of Kaposi's sarcoma-associated herpesvirus into adherent target cells in a Src-dependent manner. *J Virol* **80**, 11432-46 (2006).
551. Schulz, T.F. KSHV/HHV8-associated lymphoproliferations in the AIDS setting. *Eur J Cancer* **37**, 1217-26 (2001).

552. Bolormaa, S., Pryce, J.E., Hayes, B.J. & Goddard, M.E. Multivariate analysis of a genome-wide association study in dairy cattle. *J Dairy Sci* **93**, 3818-33 (2010).
553. Zheng, H. *et al.* The dual-specificity phosphatase DUSP14 negatively regulates tumor necrosis factor- and interleukin-1-induced nuclear factor-kappaB activation by dephosphorylating the protein kinase TAK1. *J Biol Chem* **288**, 819-25 (2013).
554. Keller, S.A. *et al.* NF-kappaB is essential for the progression of KSHV- and EBV-infected lymphomas in vivo. *Blood* **107**, 3295-302 (2006).
555. Grossmann, C. & Ganem, D. Effects of NFkappaB activation on KSHV latency and lytic reactivation are complex and context-dependent. *Virology* **375**, 94-102 (2008).
556. Lippert, D.N. & Wilkins, J.A. Glia maturation factor gamma regulates the migration and adherence of human T lymphocytes. *BMC Immunol* **13**, 21 (2012).
557. Zuo, P. *et al.* High GMFG expression correlates with poor prognosis and promotes cell migration and invasion in epithelial ovarian cancer. *Gynecol Oncol* **132**, 745-51 (2014).
558. Mbisa, G.L. *et al.* Detection of antibodies to Kaposi's sarcoma-associated herpesvirus: a new approach using K8.1 ELISA and a newly developed recombinant LANA ELISA. *J Immunol Methods* **356**, 39-46 (2010).
559. Neipel, F., Albrecht, J.C. & Fleckenstein, B. Cell-homologous genes in the Kaposi's sarcoma-associated rhadinovirus human herpesvirus 8: determinants of its pathogenicity? *J Virol* **71**, 4187-92 (1997).
560. Rezaee, S.A., Cunningham, C., Davison, A.J. & Blackbourn, D.J. Kaposi's sarcoma-associated herpesvirus immune modulation: an overview. *J Gen Virol* **87**, 1781-804 (2006).
561. Brulois, K.F. *et al.* Construction and manipulation of a new Kaposi's sarcoma-associated herpesvirus bacterial artificial chromosome clone. *J Virol* **86**, 9708-20 (2012).
562. Yakushko, Y. *et al.* Kaposi's sarcoma-associated herpesvirus bacterial artificial chromosome contains a duplication of a long unique-region fragment within the terminal repeat region. *J Virol* **85**, 4612-7 (2011).
563. Tamburro, K.M. *et al.* Vironome of Kaposi sarcoma associated herpesvirus-inflammatory cytokine syndrome in an AIDS patient reveals co-infection of human herpesvirus 8 and human herpesvirus 6A. *Virology* **433**, 220-5 (2012).
564. Olp, L.N., Jeanniard, A., Marimo, C., West, J.T. & Wood, C. Whole-Genome Sequencing of Kaposi's Sarcoma-Associated Herpesvirus from Zambian Kaposi's Sarcoma Biopsy Specimens Reveals Unique Viral Diversity. *J Virol* **89**, 12299-308 (2015).
565. Dresang, L.R. *et al.* Coupled transcriptome and proteome analysis of human lymphotropic tumor viruses: insights on the detection and discovery of viral genes. *BMC Genomics* **12**, 625 (2011).
566. Arias, C. *et al.* KSHV 2.0: A Comprehensive Annotation of the Kaposi's Sarcoma-Associated Herpesvirus Genome Using Next-Generation Sequencing Reveals Novel Genomic and Functional Features. *PLoS Pathog* **10**, e1003847 (2014).

567. Meng, Y.X. *et al.* Individuals from North America, Australasia, and Africa are infected with four different genotypes of human herpesvirus 8. *Virology* **261**, 106-19 (1999).
568. Lacoste, V. *et al.* Molecular characterization of Kaposi's sarcoma-associated herpesvirus/human herpesvirus-8 strains from Russia. *J Gen Virol* **81**, 1217-22 (2000).
569. Poole, L.J. *et al.* Comparison of genetic variability at multiple loci across the genomes of the major subtypes of Kaposi's sarcoma-associated herpesvirus reveals evidence for recombination and for two distinct types of open reading frame K15 alleles at the right-hand end. *J Virol* **73**, 6646-60 (1999).
570. Lacoste, V. *et al.* Molecular epidemiology of human herpesvirus 8 in africa: both B and A5 K1 genotypes, as well as the M and P genotypes of K14.1/K15 loci, are frequent and widespread. *Virology* **278**, 60-74 (2000).
571. Tornesello, M.L. *et al.* Human herpesvirus type 8 variants circulating in Europe, Africa and North America in classic, endemic and epidemic Kaposi's sarcoma lesions during pre-AIDS and AIDS era. *Virology* **398**, 280-9 (2010).
572. Diamond, J. & Bellwood, P. Farmers and their languages: the first expansions. *Science* **300**, 597-603 (2003).
573. Salas, A. *et al.* The making of the African mtDNA landscape. *Am J Hum Genet* **71**, 1082-111 (2002).
574. Simpson, G.R. *et al.* Prevalence of Kaposi's sarcoma associated herpesvirus infection measured by antibodies to recombinant capsid protein and latent immunofluorescence antigen. *Lancet* **348**, 1133-8 (1996).
575. Kakoola, D.N. *et al.* Recombination in human herpesvirus-8 strains from Uganda and evolution of the K15 gene. *J Gen Virol* **82**, 2393-404 (2001).
576. Kajumbula, H. *et al.* Ugandan Kaposi's sarcoma-associated herpesvirus phylogeny: evidence for cross-ethnic transmission of viral subtypes. *Intervirology* **49**, 133-43 (2006).
577. Lallemand, F., Desire, N., Rozenbaum, W., Nicolas, J.-c. & Marechal, V. Quantitative Analysis of Human Herpesvirus 8 Viral Load Using a Real-Time PCR Assay. **38**, 1404-1408 (2000).
578. Pardieu, C. *et al.* The RING-CH ligase K5 antagonizes restriction of KSHV and HIV-1 particle release by mediating ubiquitin-dependent endosomal degradation of tetherin. *PLoS pathogens* **6**, e1000843 (2010).
579. Watson, S.J. *et al.* Viral population analysis and minority-variant detection using short read next-generation sequencing. *Philos Trans R Soc Lond B Biol Sci* **368**, 20120205 (2013).
580. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-60 (2009).
581. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-9 (2009).
582. Katoh, K. & Standley, D.M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**, 772-80 (2013).

583. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-3 (2014).
584. Depledge, D.P. *et al.* Specific capture and whole-genome sequencing of viruses from clinical samples. *PLoS One* **6**, e27805 (2011).
585. Brayfield, B.P. *et al.* Distribution of Kaposi sarcoma-associated herpesvirus/human herpesvirus 8 in maternal saliva and breast milk in Zambia: implications for transmission. *J Infect Dis* **189**, 2260-70 (2004).
586. Palser, A.L. *et al.* Genome diversity of Epstein-Barr virus from multiple tumor types and normal infection. *J Virol* **89**, 5222-37 (2015).
587. Pauk, J. *et al.* Mucosal shedding of human herpesvirus 8 in men. *N Engl J Med* **343**, 1369-77 (2000).
588. Taylor, M.M. *et al.* Shedding of human herpesvirus 8 in oral and genital secretions from HIV-1-seropositive and -seronegative Kenyan women. *J Infect Dis* **190**, 484-8 (2004).
589. Corey, L., Brodie, S., Huang, M.L., Koelle, D.M. & Wald, A. HHV-8 infection: a model for reactivation and transmission. *Rev Med Virol* **12**, 47-63 (2002).
590. Bender Ignacio, R.A. *et al.* Patterns of human herpesvirus-8 oral shedding among diverse cohorts of human herpesvirus-8 seropositive persons. *Infect Agent Cancer* **11**, 7 (2016).
591. Stebbing, J. *et al.* Kaposi's sarcoma-associated herpesvirus cytotoxic T lymphocytes recognize and target Darwinian positively selected autologous K1 epitopes. *J Virol* **77**, 4306-14 (2003).
592. Greenspan, G. *et al.* Model-based inference of recombination hotspots in a highly variable oncogene [corrected]. *J Mol Evol* **58**, 239-51 (2004).
593. Boshoff, C. & Weiss, R.A. Kaposi's sarcoma-associated herpesvirus. *Adv Cancer Res* **75**, 57-86 (1998).
594. Kwok, H. *et al.* Genomic diversity of Epstein-Barr virus genomes isolated from primary nasopharyngeal carcinoma biopsy samples. *J Virol* **88**, 10662-72 (2014).
595. Cassar, O. *et al.* Divergent KSHV/HHV-8 subtype D strains in New Caledonia and Solomon Islands, Melanesia. *J Clin Virol* **53**, 214-8 (2012).
596. Kasolo, F.C., Spinks, J., Bima, H., Bates, M. & Gompels, U.A. Diverse genotypes of Kaposi's sarcoma associated herpesvirus (KSHV) identified in infant blood infections in African childhood-KS and HIV/AIDS endemic region. *J Med Virol* **79**, 1555-61 (2007).
597. Ramos da Silva, S., Ferraz da Silva, A.P., Bacchi, M.M., Bacchi, C.E. & Elgui de Oliveira, D. KSHV genotypes A and C are more frequent in Kaposi sarcoma lesions from Brazilian patients with and without HIV infection, respectively. *Cancer Lett* **301**, 85-94 (2011).
598. Zhang, Y. *et al.* Distinct distribution of rare us kshv genotypes in south texas. Implications for kshv epidemiology and evolution. *Ann Epidemiol* **10**, 470 (2000).
599. Kouri, V. *et al.* Kaposi's Sarcoma and Human Herpesvirus 8 in Cuba: evidence of subtype B expansion. *Virology* **432**, 361-9 (2012).
600. Ouyang, X. *et al.* Genotypic analysis of Kaposi's sarcoma-associated herpesvirus from patients with Kaposi's sarcoma in Xinjiang, China. *Viruses* **6**, 4800-10 (2014).

601. Jalilvand, S. *et al.* Molecular epidemiology of human herpesvirus 8 variants in Kaposi's sarcoma from Iranian patients. *Virus Res* **163**, 644-9 (2012).
602. Betsem, E. *et al.* Epidemiology and genetic variability of HHV-8/KSHV in Pygmy and Bantu populations in Cameroon. *PLoS Negl Trop Dis* **8**, e2851 (2014).
603. Whitby, D. *et al.* Genotypic characterization of Kaposi's sarcoma-associated herpesvirus in asymptomatic infected subjects from isolated populations. *J Gen Virol* **85**, 155-63 (2004).
604. Isaacs, T., Abera, A.B., Muloiwa, R., Katz, A.A. & Todd, G. Genetic diversity of HHV8 subtypes in South Africa: A5 subtype is associated with extensive disease in AIDS-KS. *J Med Virol* **88**, 292-303 (2016).
605. Bourboulia, D. *et al.* Serologic evidence for mother-to-child transmission of Kaposi sarcoma-associated herpesvirus infection. *JAMA* **280**, 31-2 (1998).
606. Frayling, T.M. *et al.* A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* **316**, 889-94 (2007).
607. Gerken, T. *et al.* The obesity-associated FTO gene encodes a 2-oxoglutarate-dependent nucleic acid demethylase. *Science* **318**, 1469-72 (2007).
608. Herman, M.A. & Rosen, E.D. Making Biological Sense of GWAS Data: Lessons from the FTO Locus. *Cell Metab* **22**, 538-9 (2015).
609. Claussnitzer, M. *et al.* FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. *N Engl J Med* **373**, 895-907 (2015).
610. Lee, S., Abecasis, G.R., Boehnke, M. & Lin, X. Rare-variant association analysis: study designs and statistical tests. *Am J Hum Genet* **95**, 5-23 (2014).
611. Zuk, O. *et al.* Searching for missing heritability: designing rare variant association studies. *Proc Natl Acad Sci U S A* **111**, E455-64 (2014).
612. Bartha, I. *et al.* A genome-to-genome analysis of associations between human genetic variation, HIV-1 sequence diversity, and viral control. *Elife* **2**, e01123 (2013).