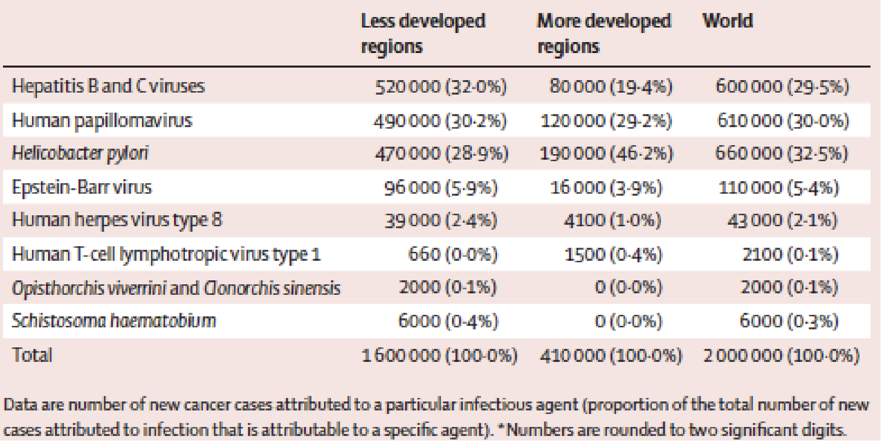
2 Chapter 2: The General Population Cohort, a Platform to Study the Genetic Architecture of Host Response to Gamma-Herpesvirus Infections

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.

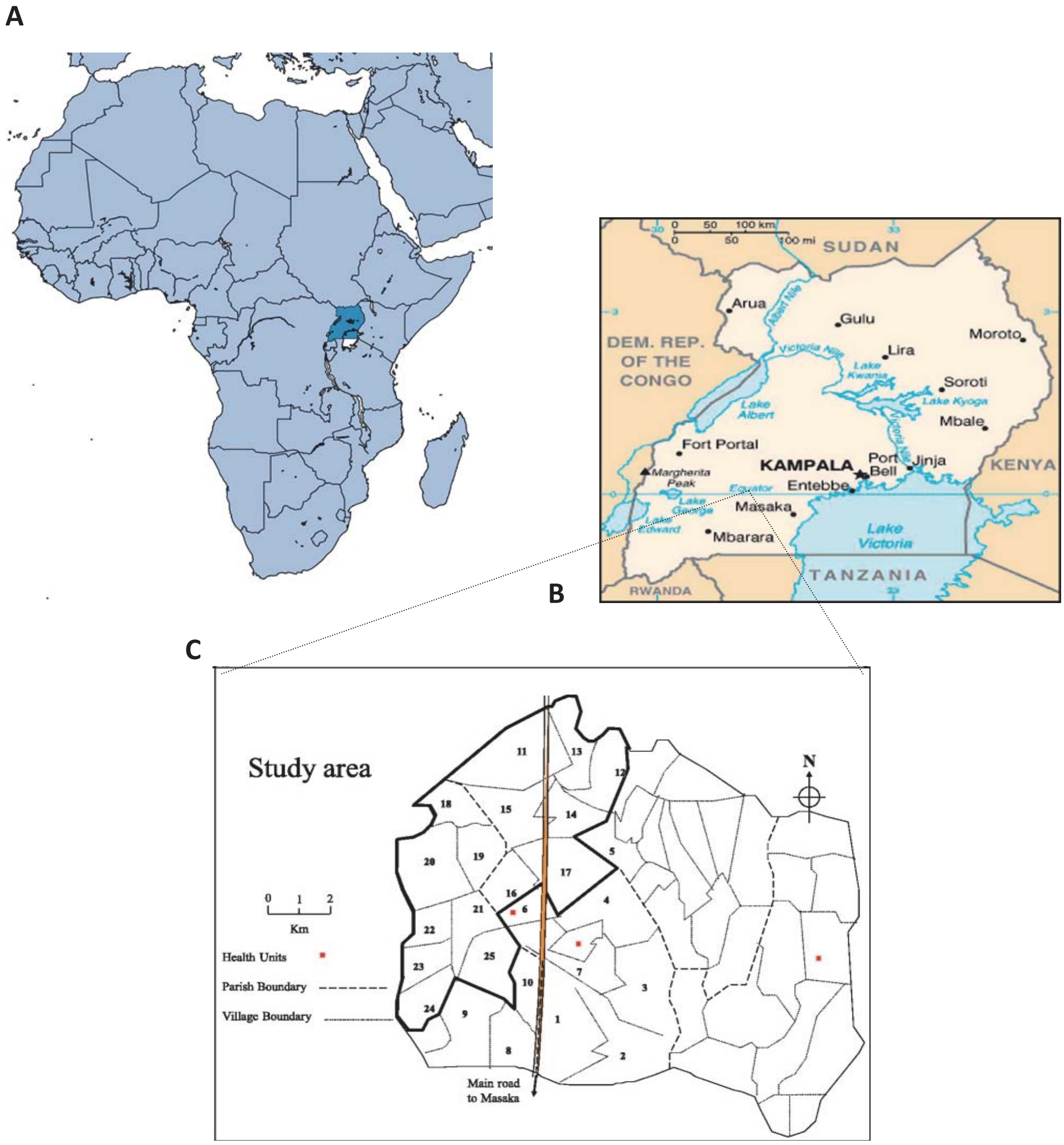


Fig. 2.1. Maps showing The GPC study area in context of Uganda and Africa. **A**. Map of Africa with Uganda shaded in dark blue. **B**. Map of Uganda and its bordering countries. **C**. The GPC study area encompassing 25 villages (labelled numerically) in the south-western region of Uganda. From Asiki 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<10-8, 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.

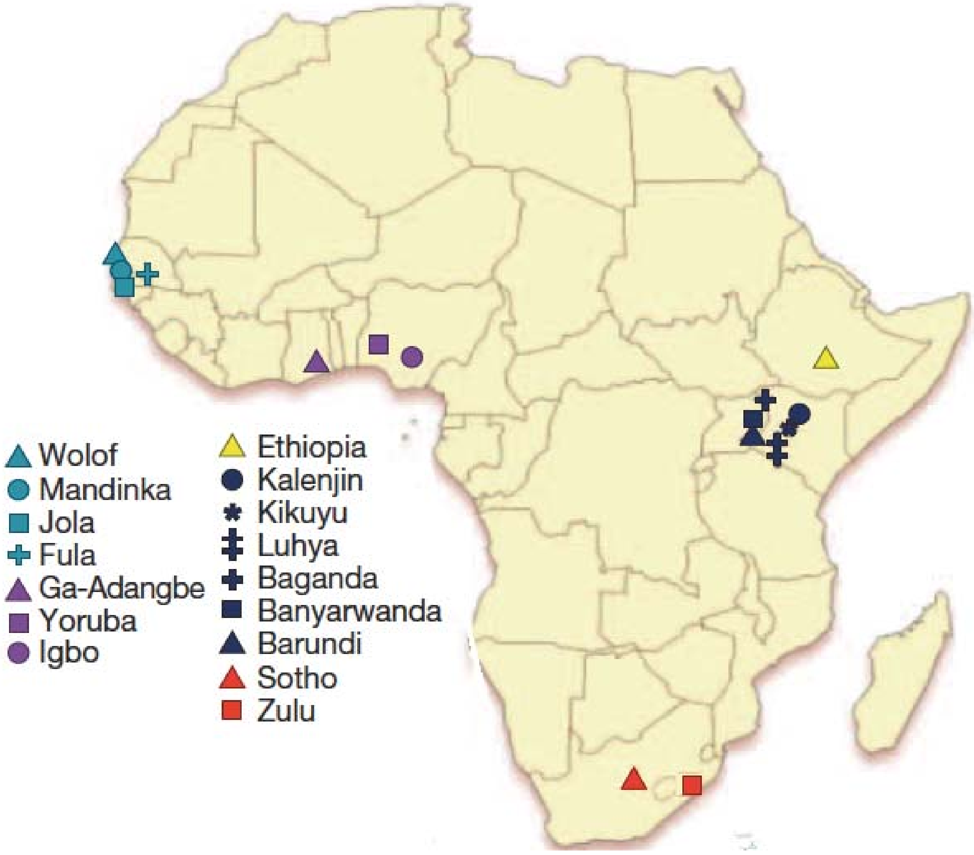


Fig. 2.2 Map of Africa showing the location of samples in the AGVP used for PCA. Representing ethno-linguistic groups from The Gambia, Ghana, Nigeria, Ethiopia, Kenya, Uganda and South Africa. Adapted from Gurdasani et al, 2014[299].

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%.

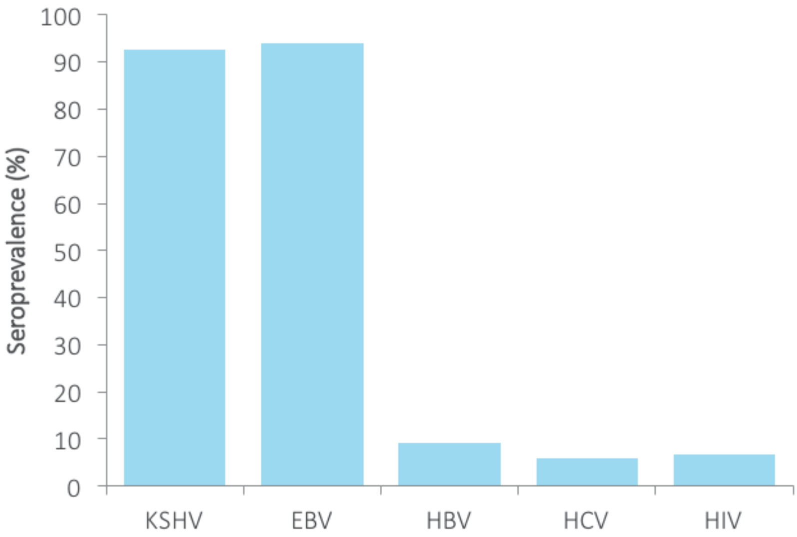


Fig. 2.3 Seroprevalence of viral infections tested in the GPC between 2008-2011

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).

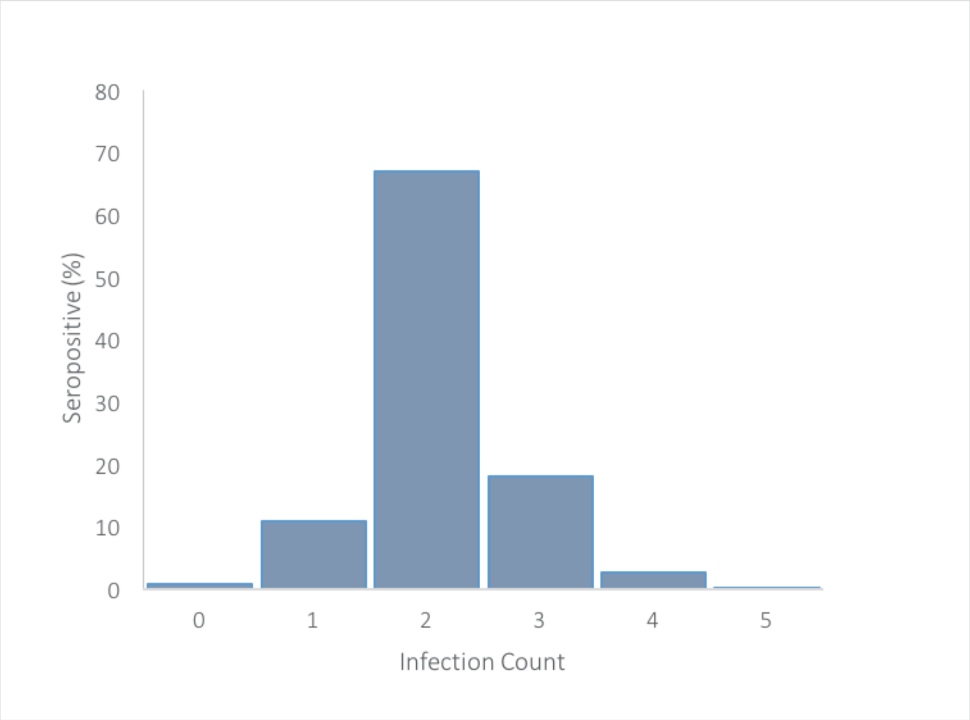


Fig. 2.4 The number of seropositive reactions to viruses for all participants in The GPC between 2008-2011. The infection count represents the minimum number of infections participants are seropositive for

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).

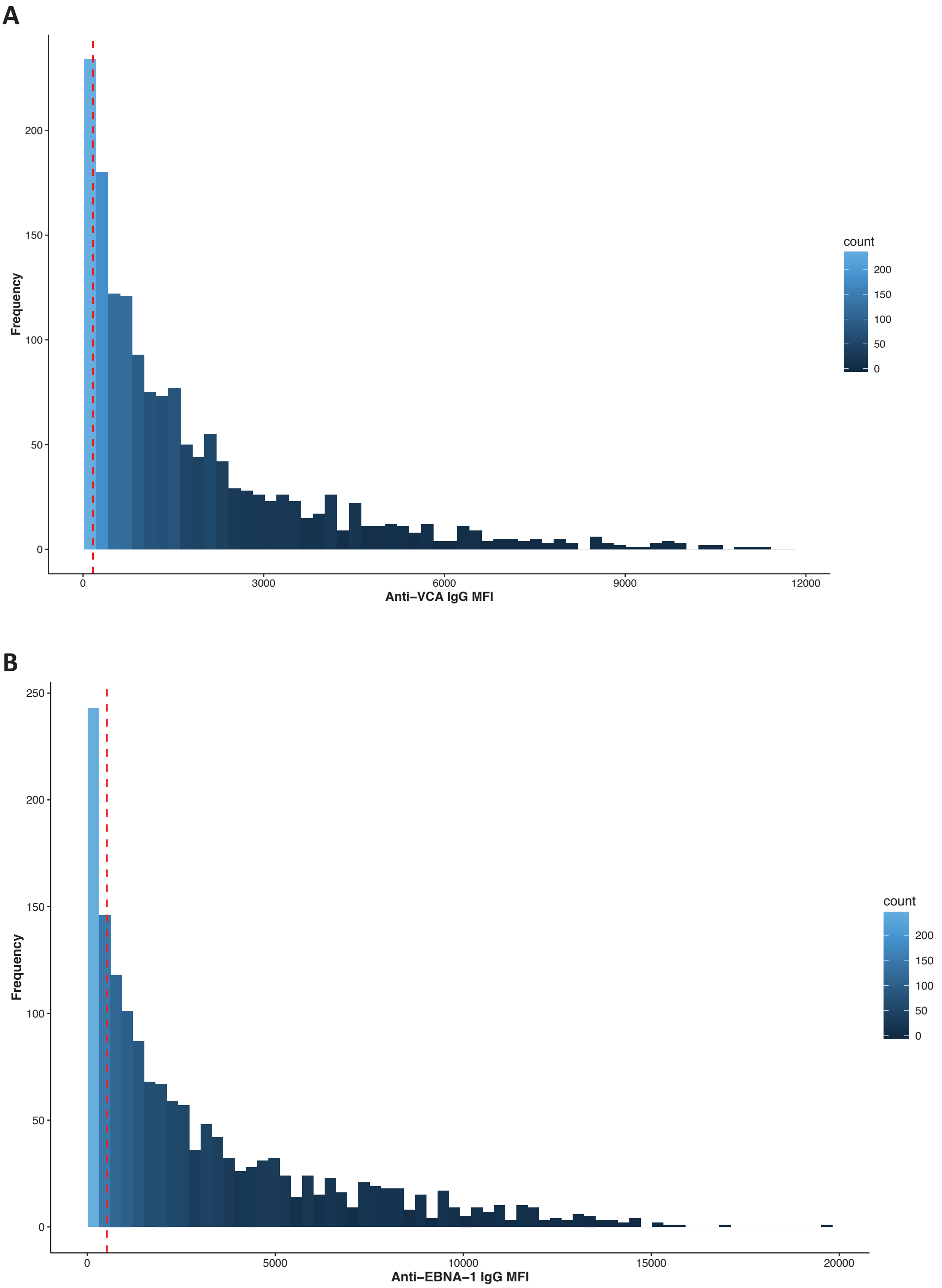


Fig. 2.5 Inter-individual variability in IgG antibody responses to EBV. **A.** Distribution of anti-VCA IgG mean fluorescence intensity (MFI). Red dotted line represents MFI cut-off=165. Seropositive=1350, Seronegative=217. **B.** Distribution of anti-EBNA-1 IgG MFI. Red dotted line represents MFI cut-off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

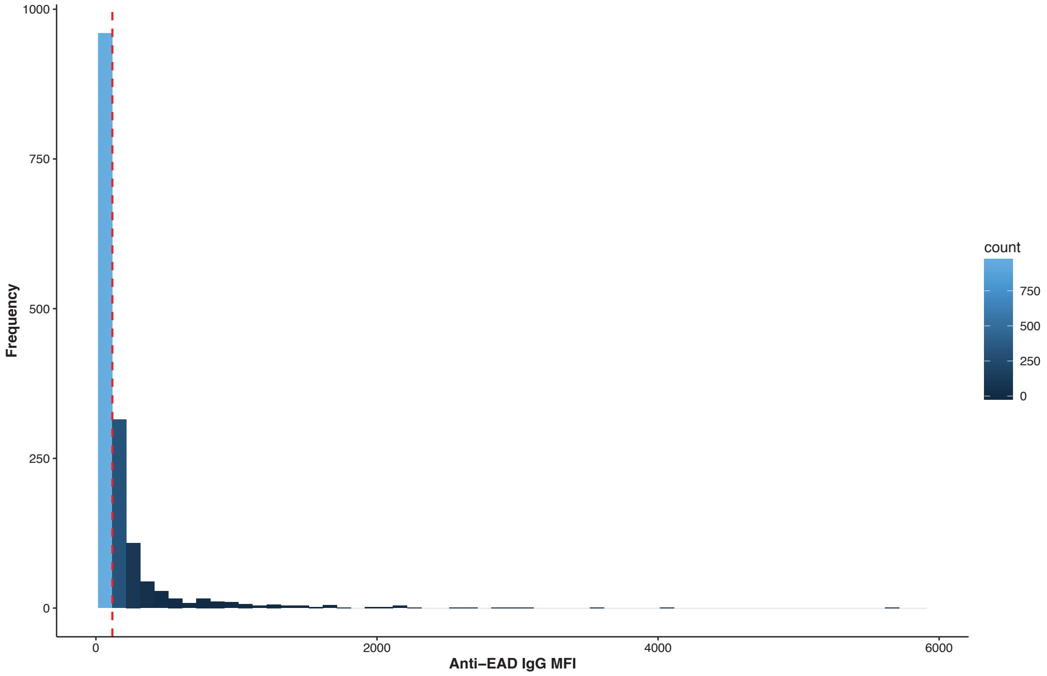


Fig. 2.6 Distribution of anti-EAD IgG Mean Fluorescence Intensity (MFI). Red dotted line represents MFI cut-off =117. Seropositive=406, Seronegative=1170

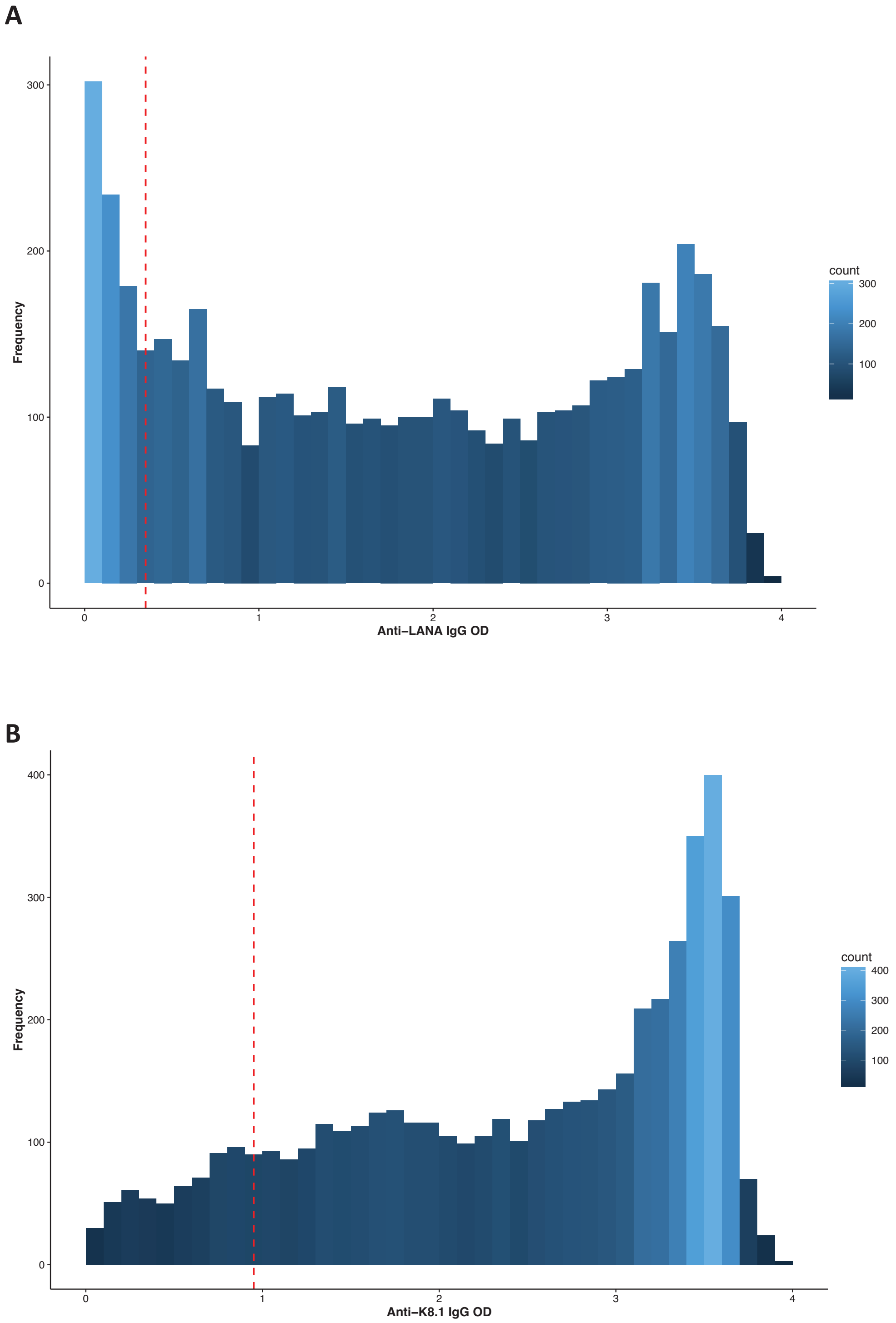


Fig. 2.7 Inter-individual variability in IgG antibody responses to KSHV.

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.

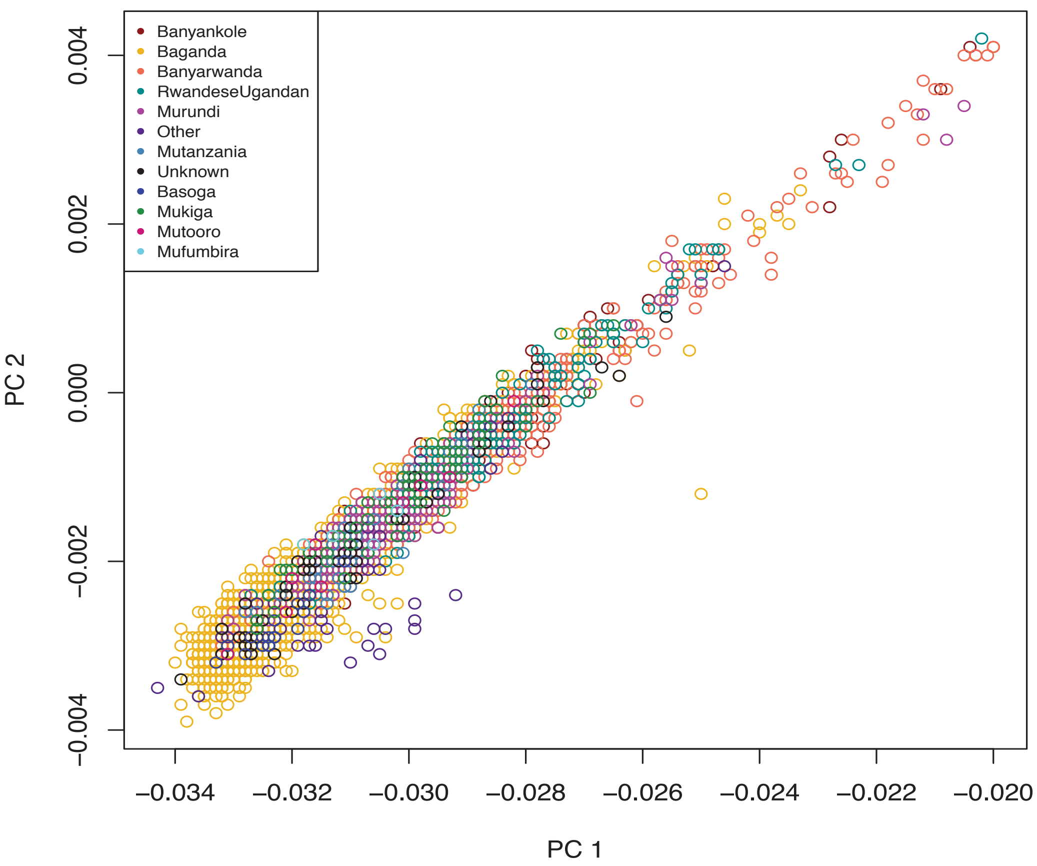


Fig. 2.8 Genetic population structure of individuals within the GPC ethnolinguistic groups. No clear separation observed for ethno-linguistic groups.

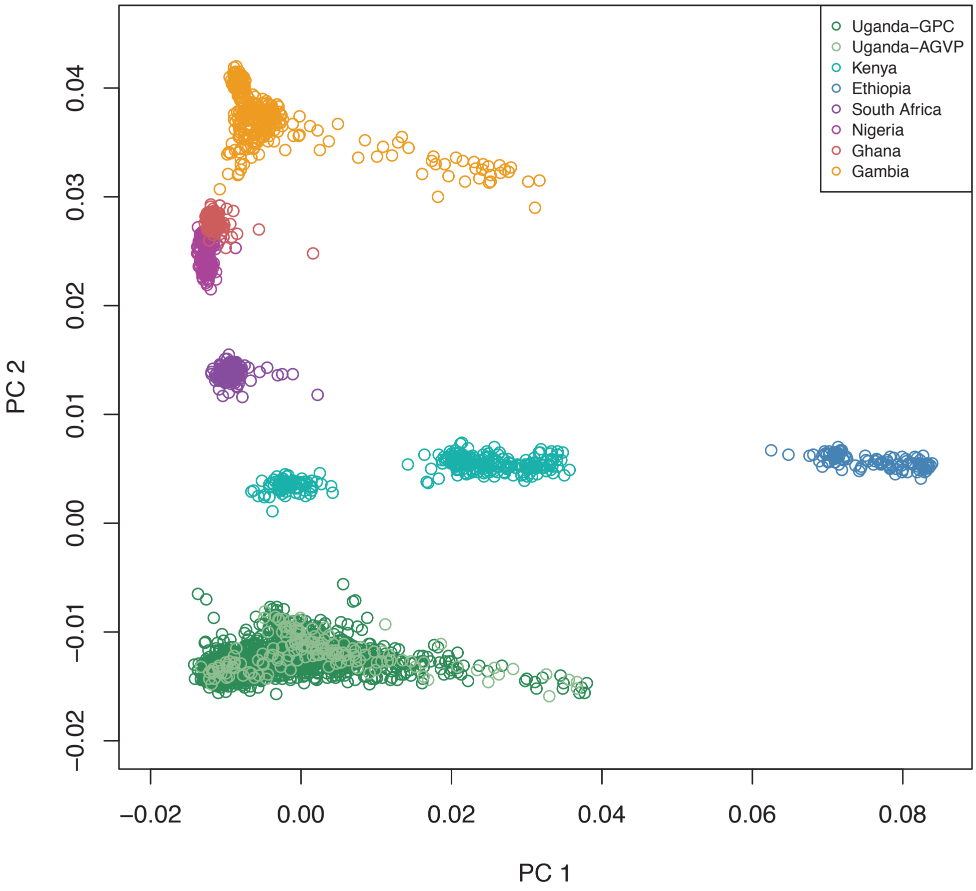


Fig. 2.9 Genetic population structure of the GPC in the context of AGVP African populations. PC1 shows cline seen among East and West Africans. PC2 shows separation by populations from different regions.

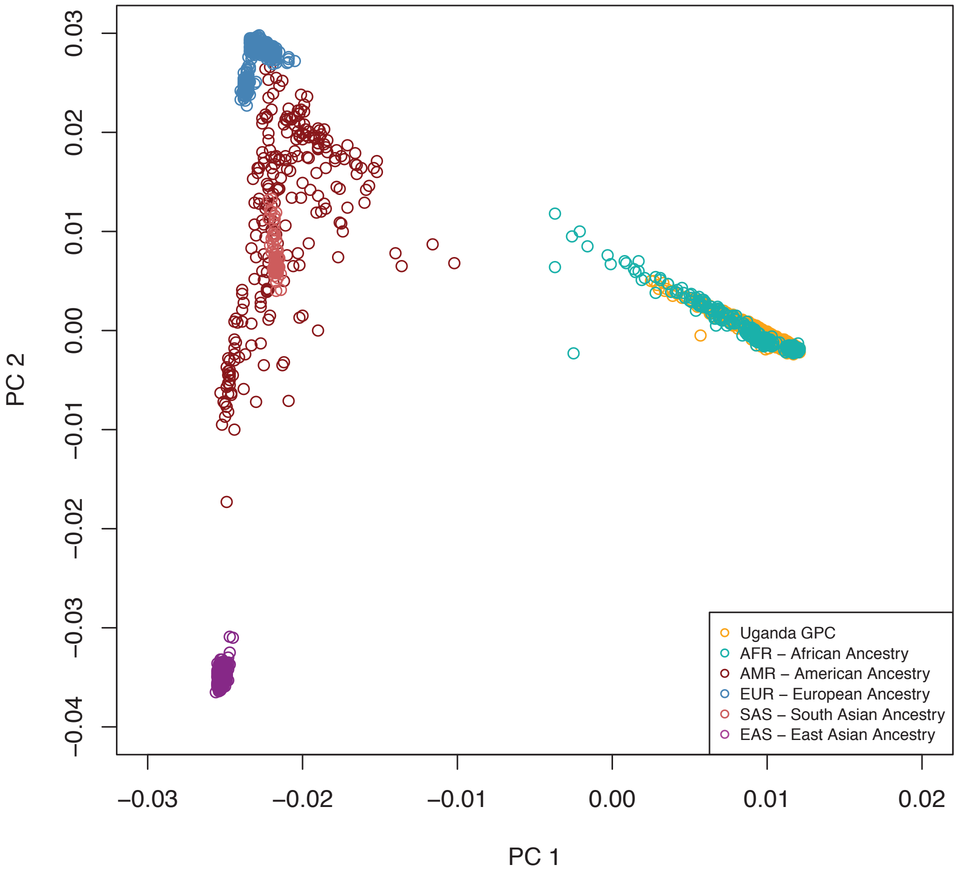


Fig. 2.10 Genetic population structure of GPC in the context of AGVP and global 1000G populations. PC1 shows a cline seen among Ugandan GPC and other African ancestry populations (include both AGVP and 1000 Genomes) extending towards Europeans.

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.

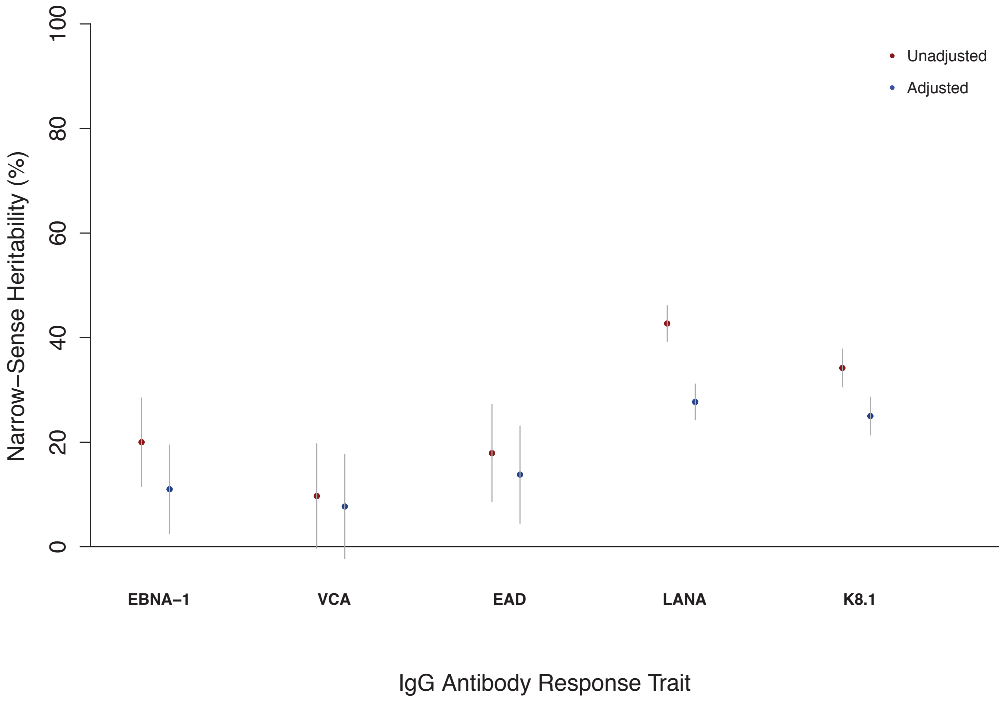


Fig. 2.11 Heritability of IgG antibody traits for EBV and KSHV infections. Fast-LMM narrow-sense heritability (h^2%) estimates with (blue) and without (red) adjustment for environmental effects accounted for by GPS coordinates. Error bars represent standard error (S.E).

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.