# 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].

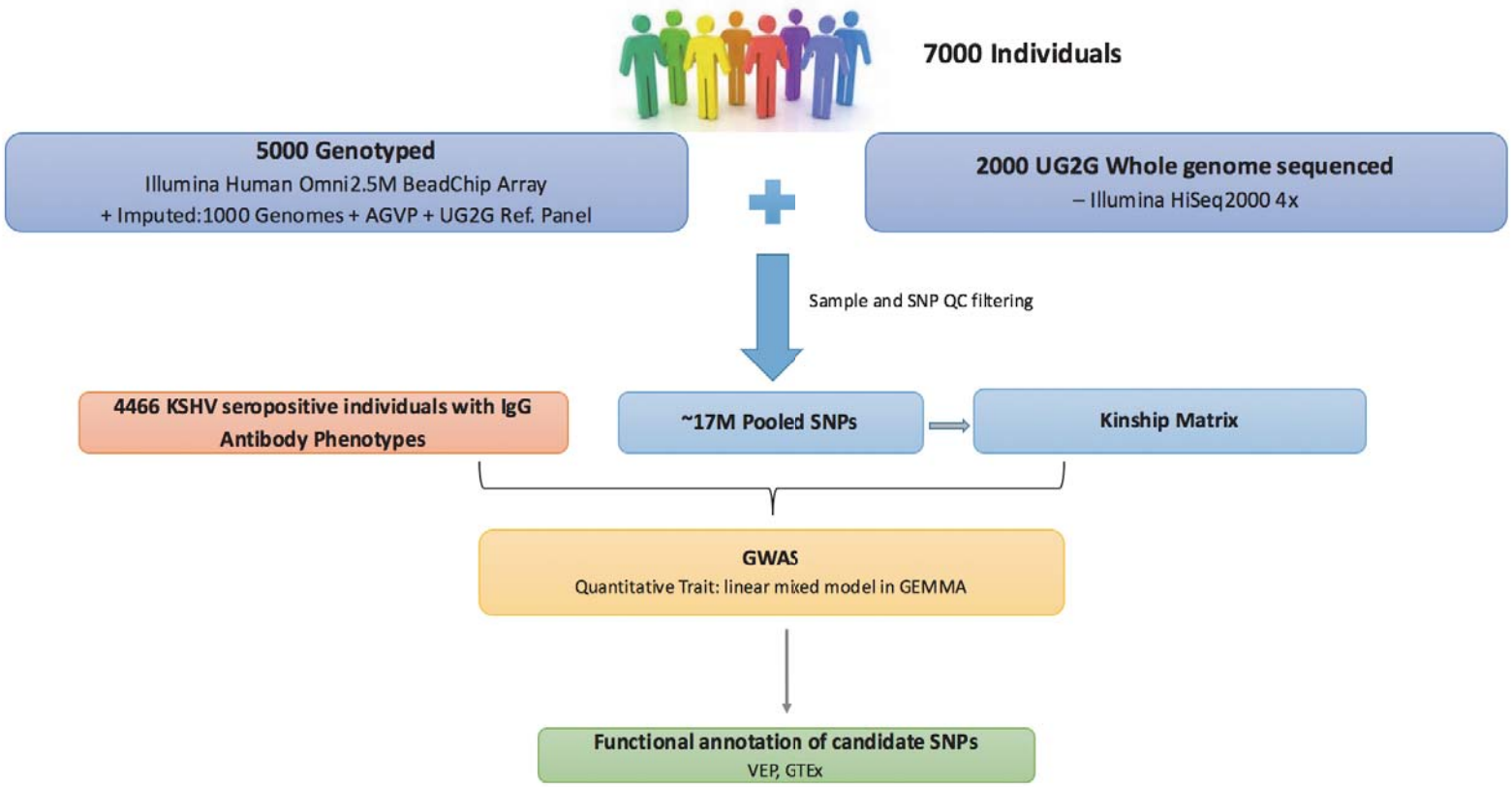


**Fig. 4.1 Genome-wide association workflow for KSHV serological traits in the Uganda GPC**

## 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 ($\beta$=0.15) (Fig. 4.2). For lower frequency variants with allele frequencies ~5%, this study had 80% power to detect moderate-large effect sizes ($\beta$=>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 ($\lambda$)) 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 $\lambda$s 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.

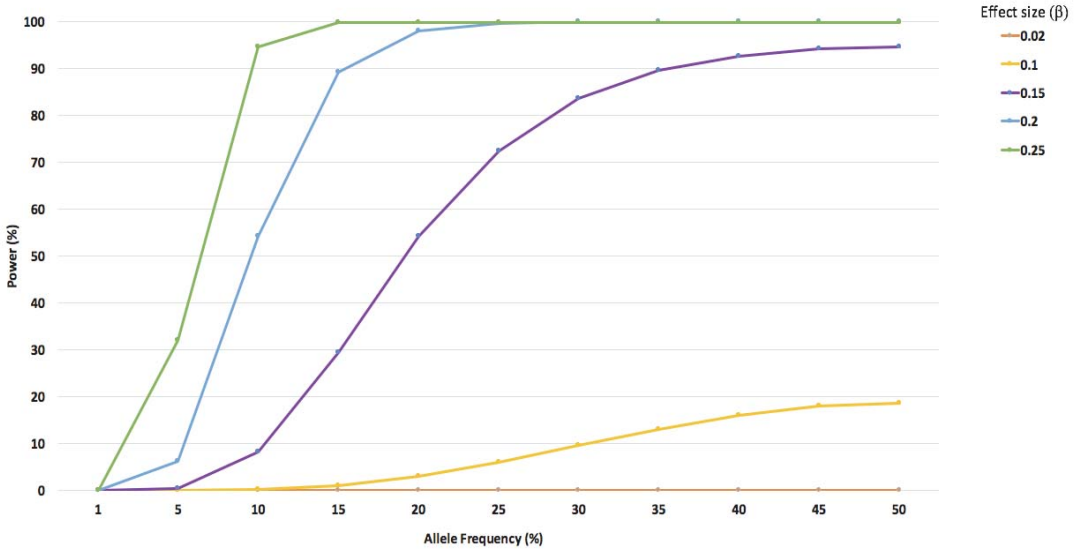


**Fig. 4.2 Statistical power to identify genetic variants at $p<5x10^{-9}$, given different allele frequencies (%) and different effect sizes (β) (N=4466).**

### 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.

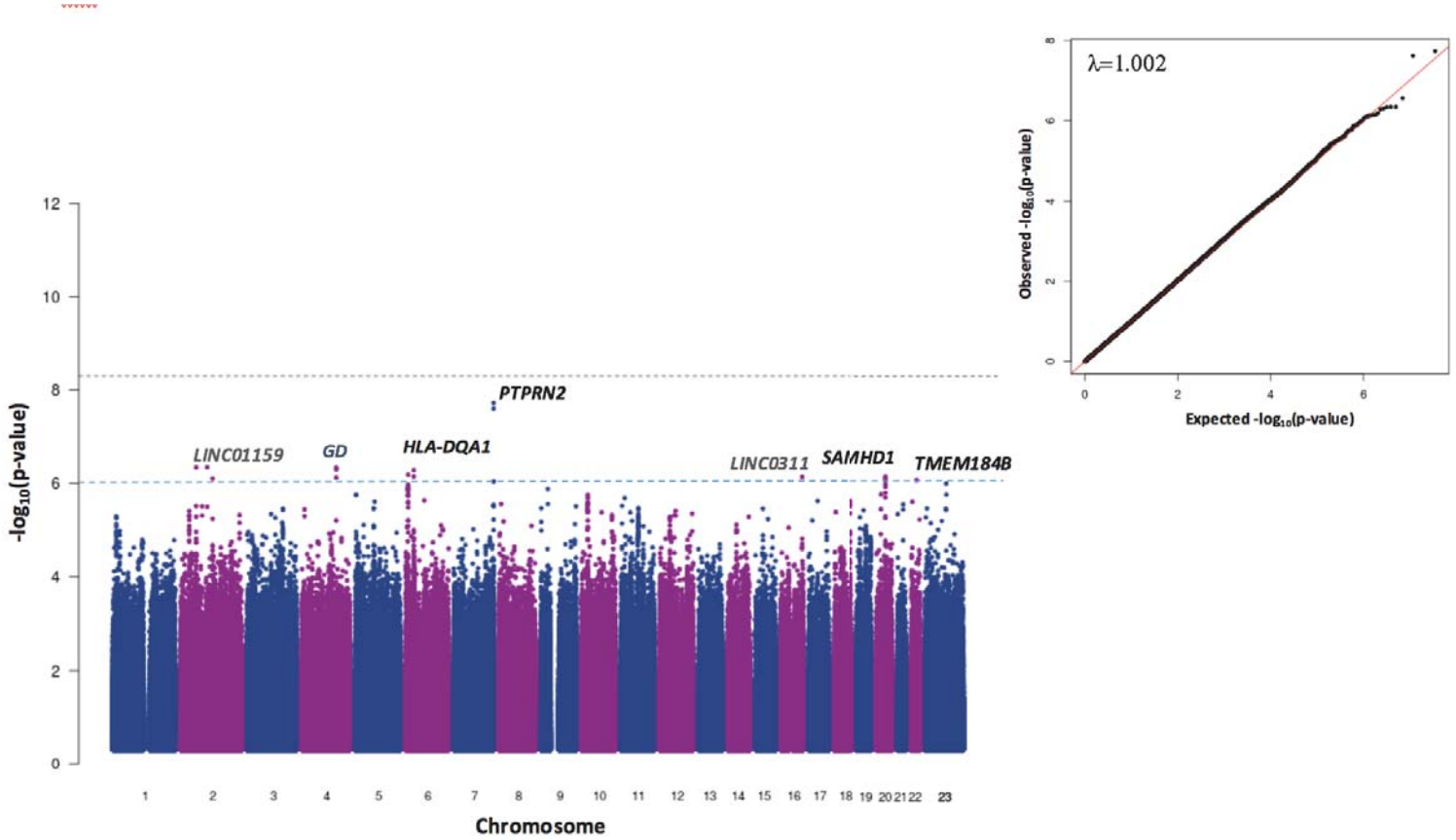


**Fig. 4.3 Genome-wide association results of anti-LANA IgG response levels.** Manhattan Plot (Left), grey dashed line: genome-wide significance threshold ($p<5x10^{-9}$), blue-dashed line: suggestive significance threshold (p<1x10-6). 23=X-Chromosome. Genes labelled in black (protein-coding), grey (long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

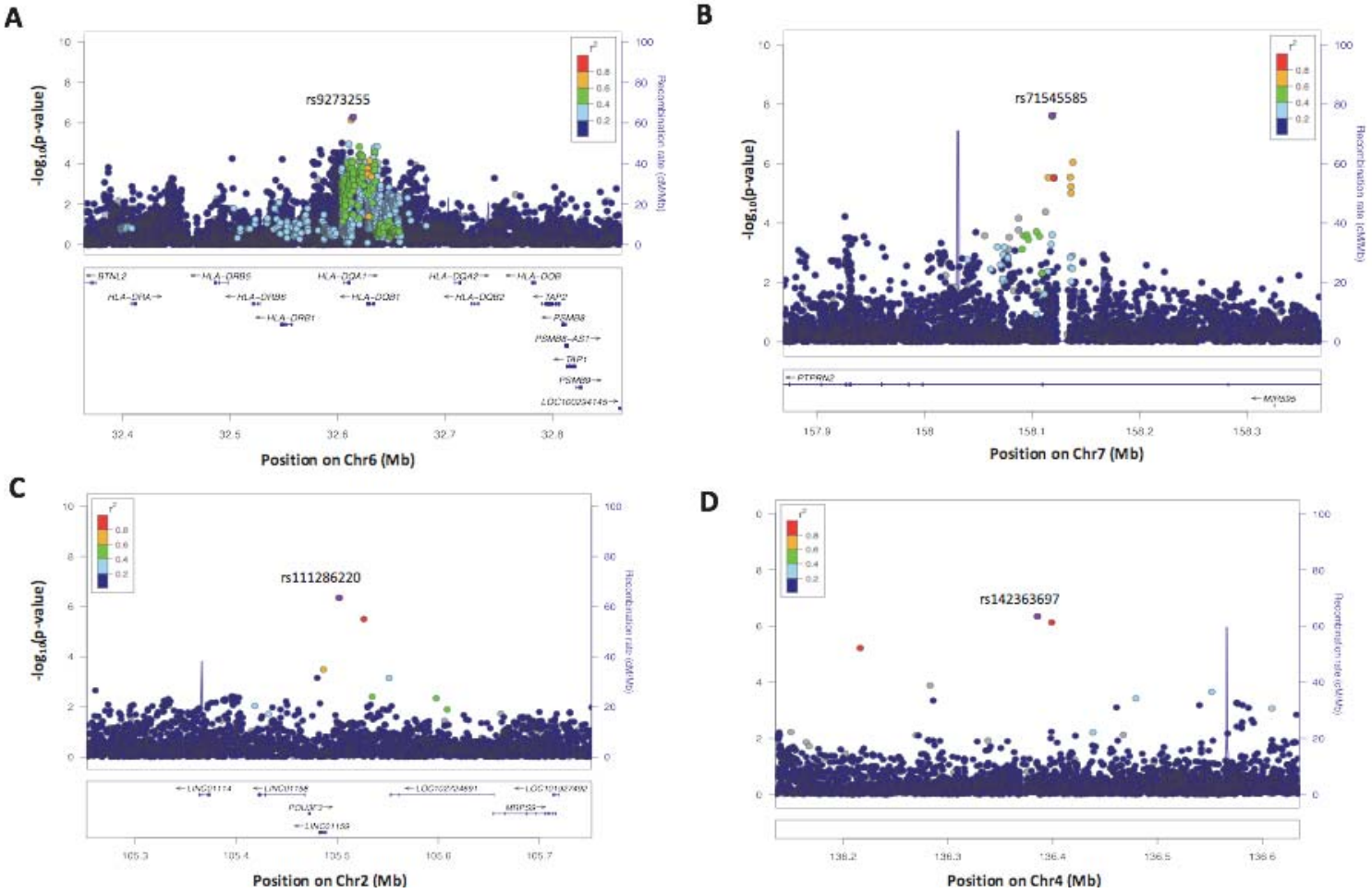


**Fig. 4.4 Regional association plots for SNPs associated with anti-LANA IgG ($p<1x10^{-6}$, N=4466.).**

**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

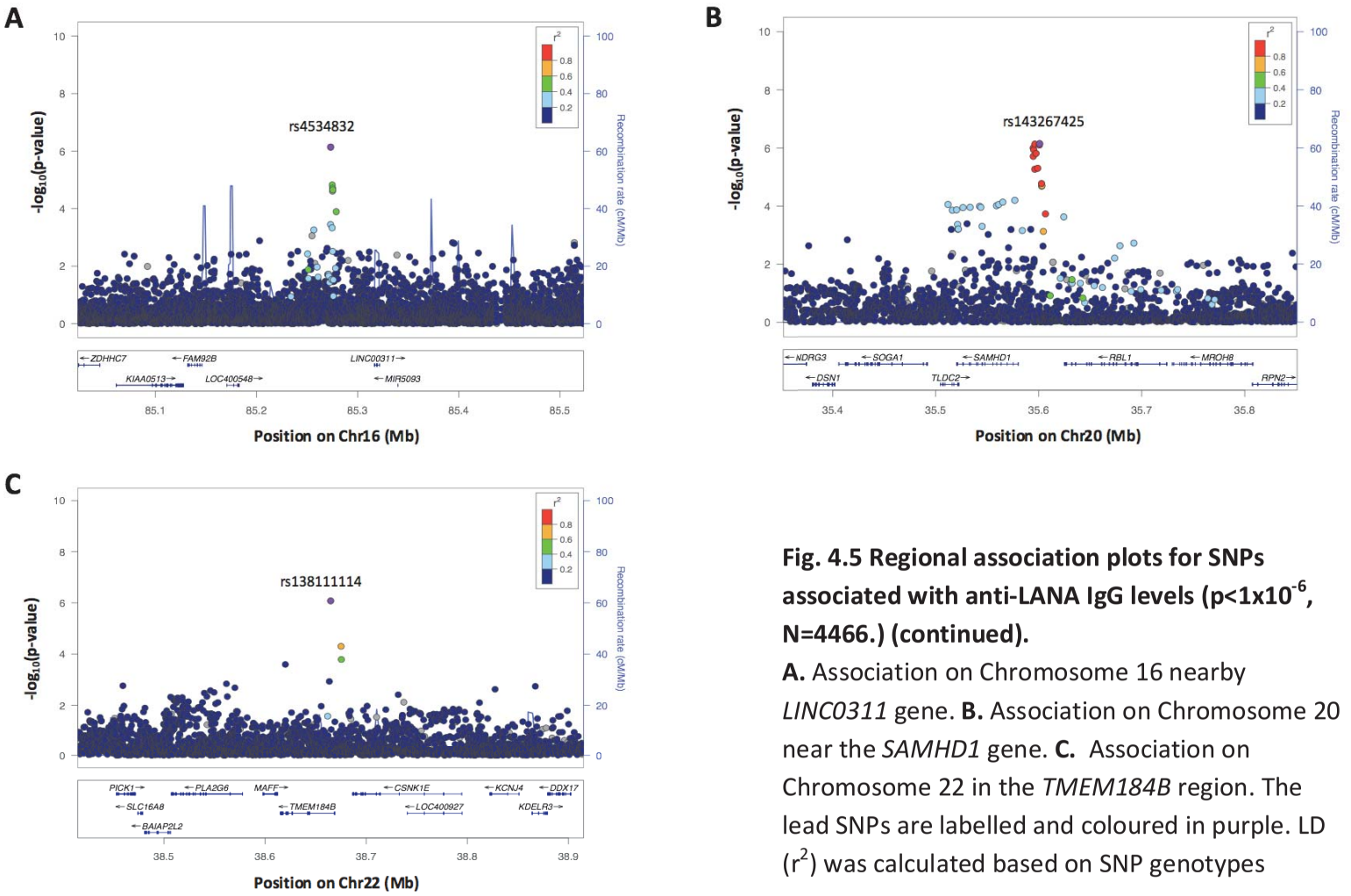


**Fig. 4.5 Regional association plots for SNPs associated with anti-LANA IgG levels ($p<1x10^{-6}$, N=4466.) (continued).**

**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.

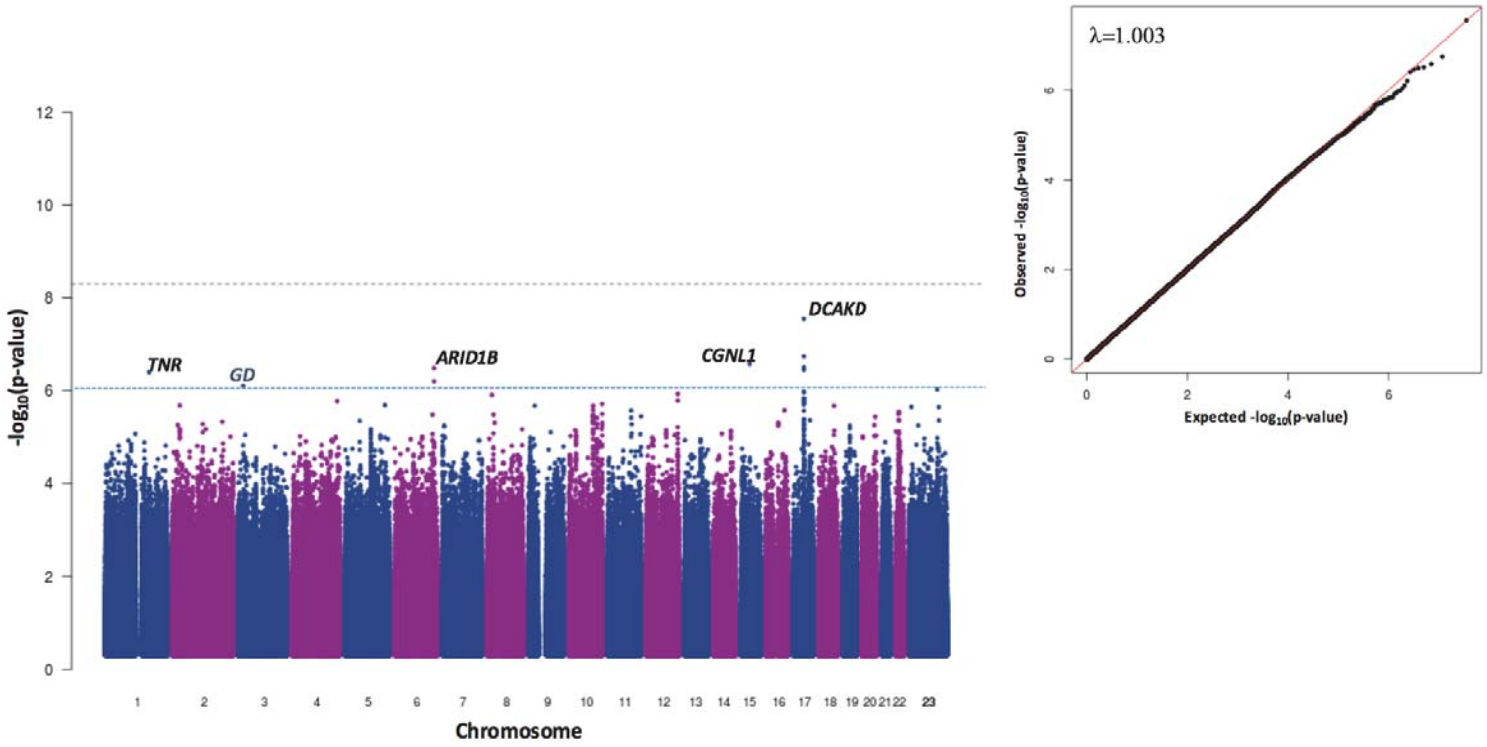


**Fig. 4.6 Genome-wide association results of anti-K8.1 IgG response levels.** Manhattan Plot (Left), grey dashed line: Genome-wide significance threshold ($p<5x10^{-9}$), blue-dashed line: Suggestive significance threshold ($p<1x10^{-6}$), GD= Gene desert, 23=X-Chromosome. QQ Plot (Top right).

**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

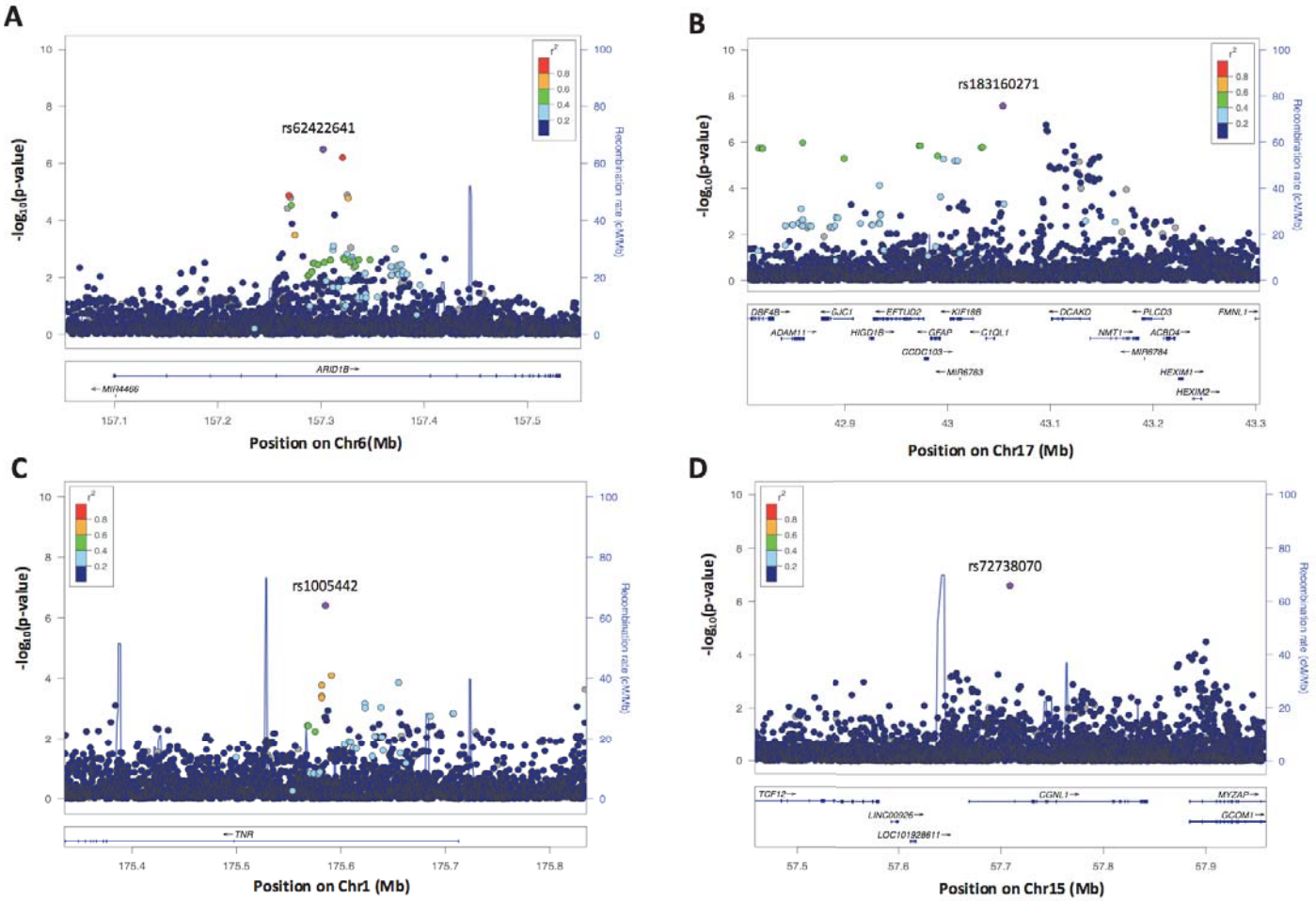


**Fig. 4.7 Regional association plots for SNPs associated with anti-K8.1 IgG response levels, N=4466, threshold=p<1x10$^{-6}$. A.** Association on Chromosome 6 in the ARID1B region. **B.** Association on Chromosome 17 in the DCAKD region. **C.** Association on Chromosome 1 in the TNR region. **D.** Association 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.

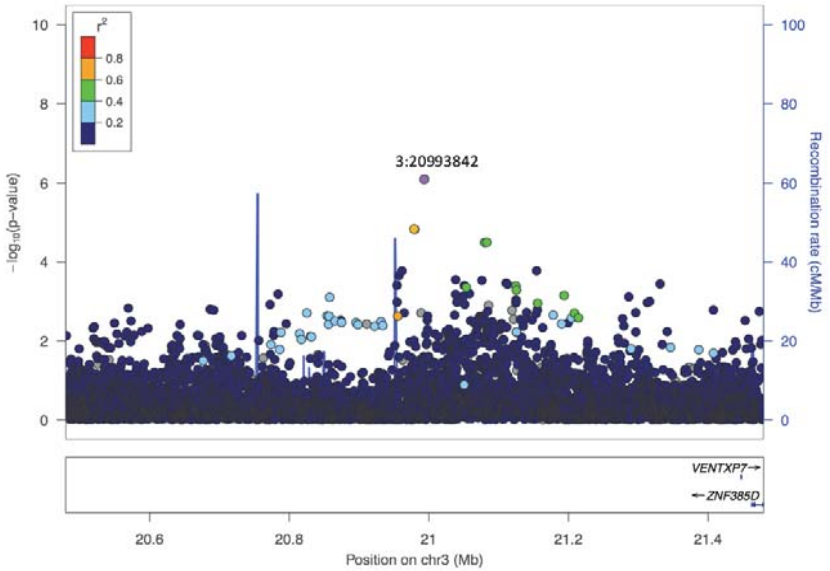


**Fig. 4.8 Regional association plots for SNP on chromosome 3 associated with anti-K8.1 IgG response levels, N=4466, threshold=p<1x10$^{-6}$.** The lead SNP is labelled and coloured in purple. LD ($r^2$) was calculated based on SNP genotypes.

### 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.

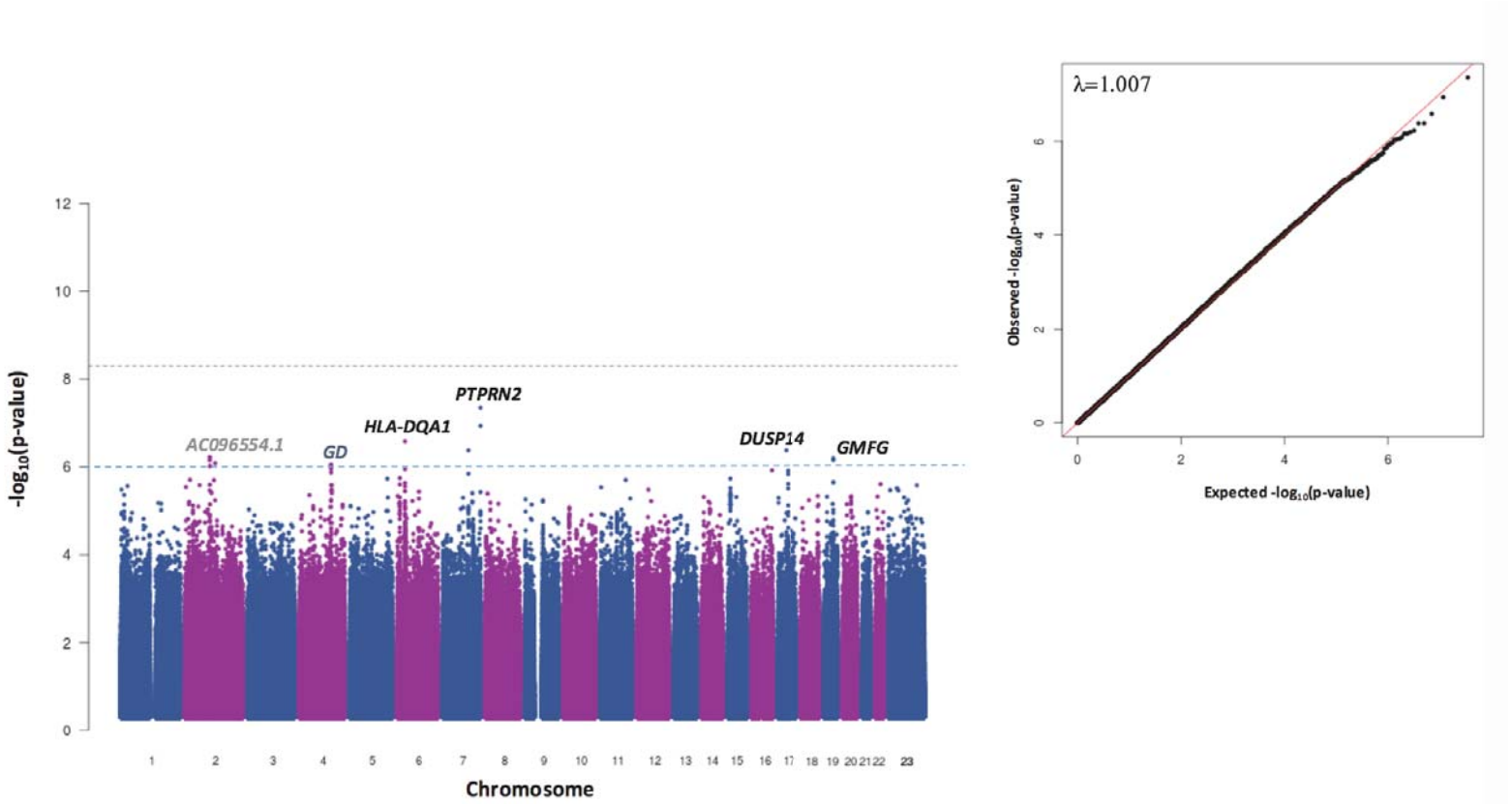


**Fig. 4.9 Multivariate Genome-wide Association results of anti-KSHV IgG response levels.** Manhattan Plot (Left), grey dashed line: genome-wide significance threshold ($p<5x10^{-9}$), blue-dashed line: suggestive significance threshold ($p<1x10^{-6}$). Genes labelled in black (protein-coding), grey (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 | $\beta_{LANA}$ | $\beta_{K8.1}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| **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

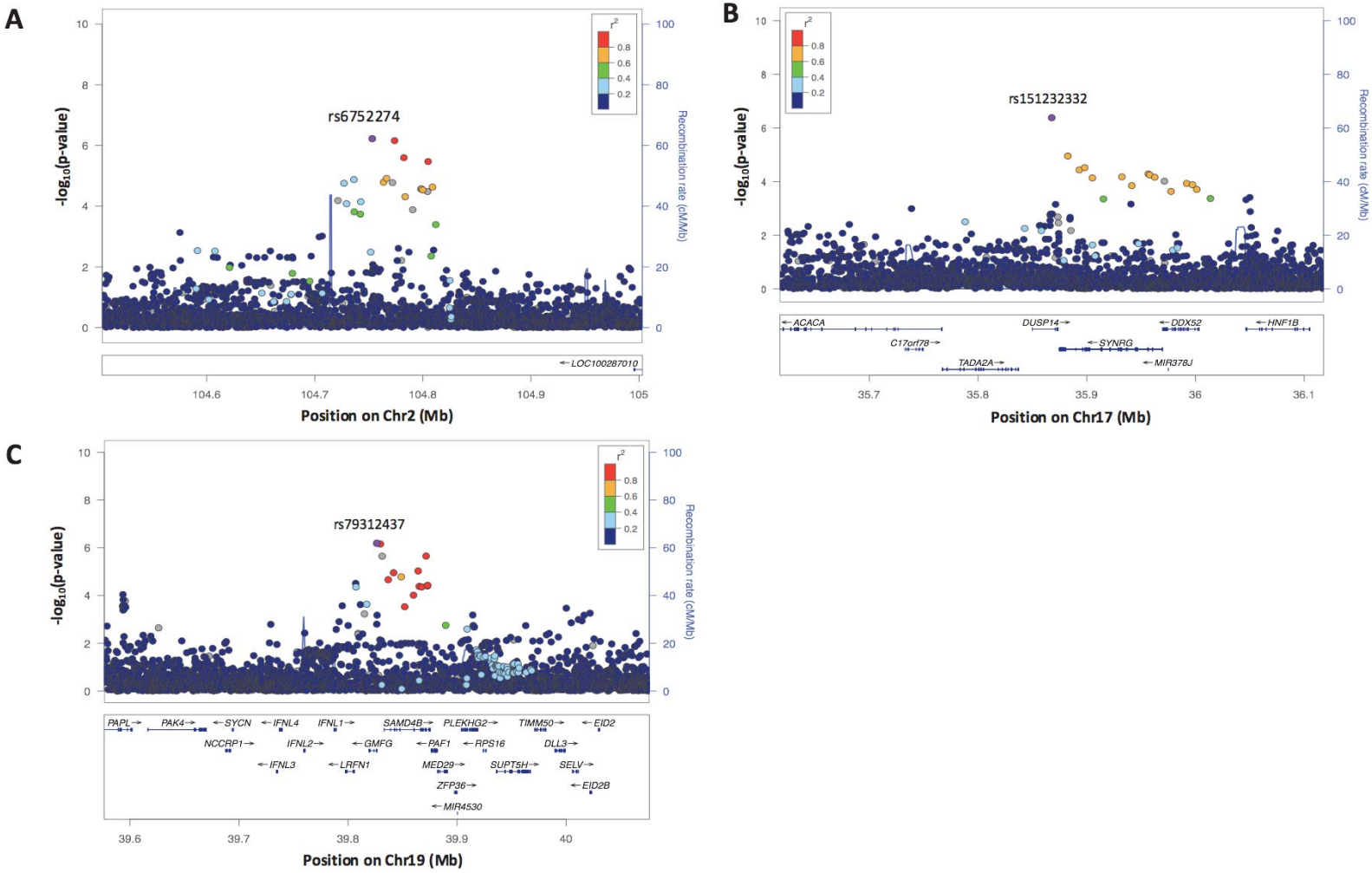


**Fig. 4.10 Regional association plots for multivariate anti-KSHV IgG levels, N=4466, threshold=p<1x10$^{-6}$. A.** Association on Chromosome 2 nearby *AC096554.1*. **B.** Association on Chromosome 17 in the *DUSP14* region. **C.** Association on Chromosome 19 in the *GMFG* region.

### 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 ($\beta$>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 ($\beta$=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 ($h^2$LANA= 27%, $h^2$K8.1=25%) (chapter 2) in comparison to EBV traits ($h^2$EBNA= 11%, $h^2$K8.1=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.