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

3.1 Introduction

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

Candidate Gene Approaches

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

Cytokine Genes (results summarised in Table 3.1)

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

HLA Genes (results summarised in Table 3.1)

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

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

Other Putative Candidate Genes

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

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

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

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

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

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

N.R- Not reported

Genome-wide approaches

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

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

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

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

3.1.1 Chapter Aims

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

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

Contributions

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

3.2 Methods

3.2.1 Sample Selection

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

3.2.2 Whole-Genome Sequencing and Quality Control

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

3.2.3 Imputation

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

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

3.2.4 Association Analyses

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

3.2.4.1 Kinship Estimation and Mixed-Modelling

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

3.2.4.2 Quantitative Trait Association Analyses

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

3.2.4.3 Discrete Serostatus Association Analysis

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

3.2.4.4 Identification of Secondary Association Signals

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

3.2.5 Trans-Ethnic Meta-Analysis

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

3.2.6 Fine Mapping

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

3.2.7 Functional Annotation of Candidate Variants

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

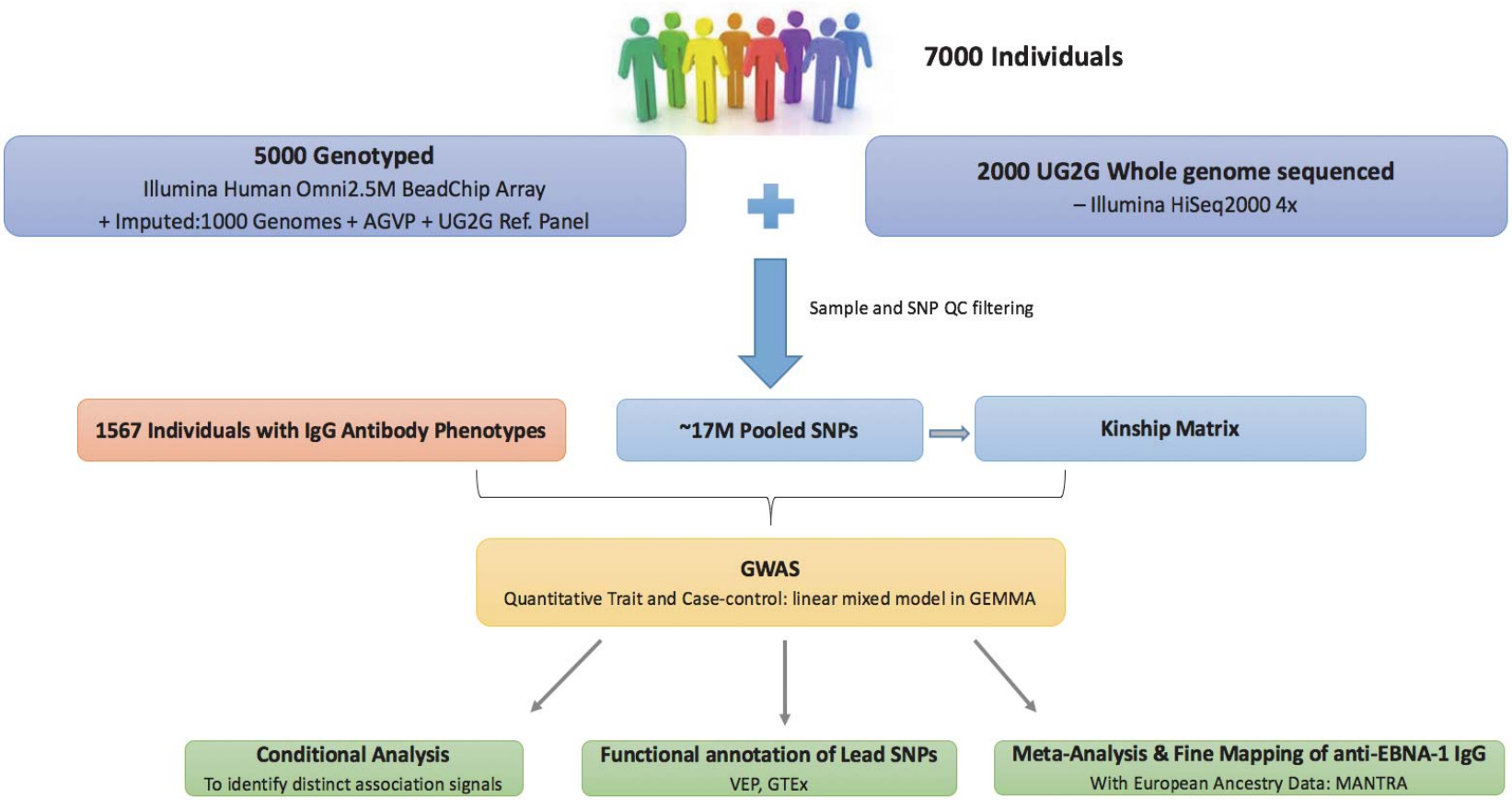


Fig. 3.1 Genome-wide association workflow for EBV serological traits in the Uganda GPC

3.3 Results

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

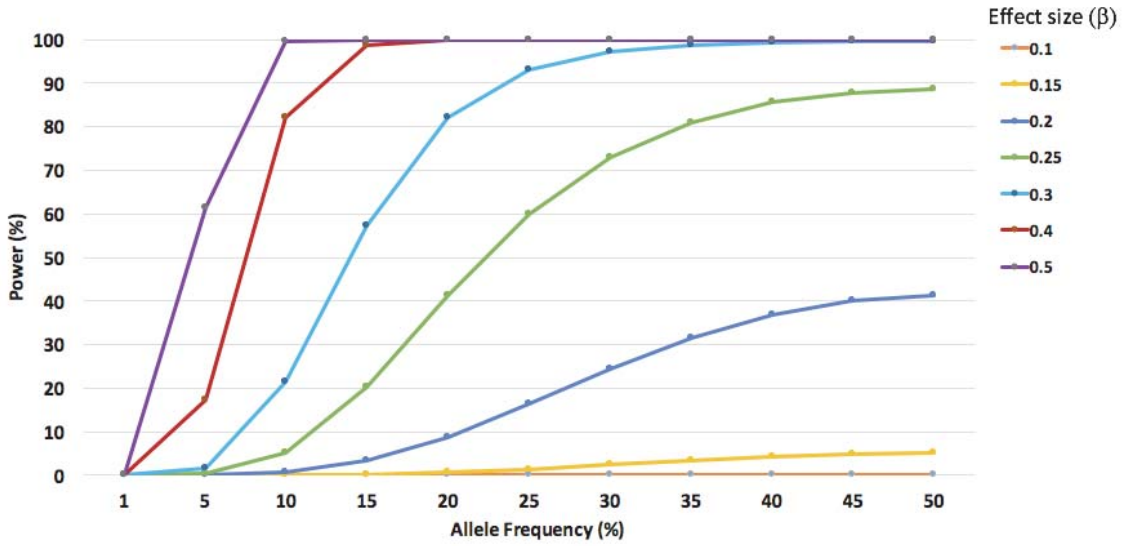


Fig. 3.2 Statistical power (%) to identify genetic variants at $p<5x10^{-9}$, given different allele frequencies (%) and effect sizes (β) (N=1567).

.

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

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

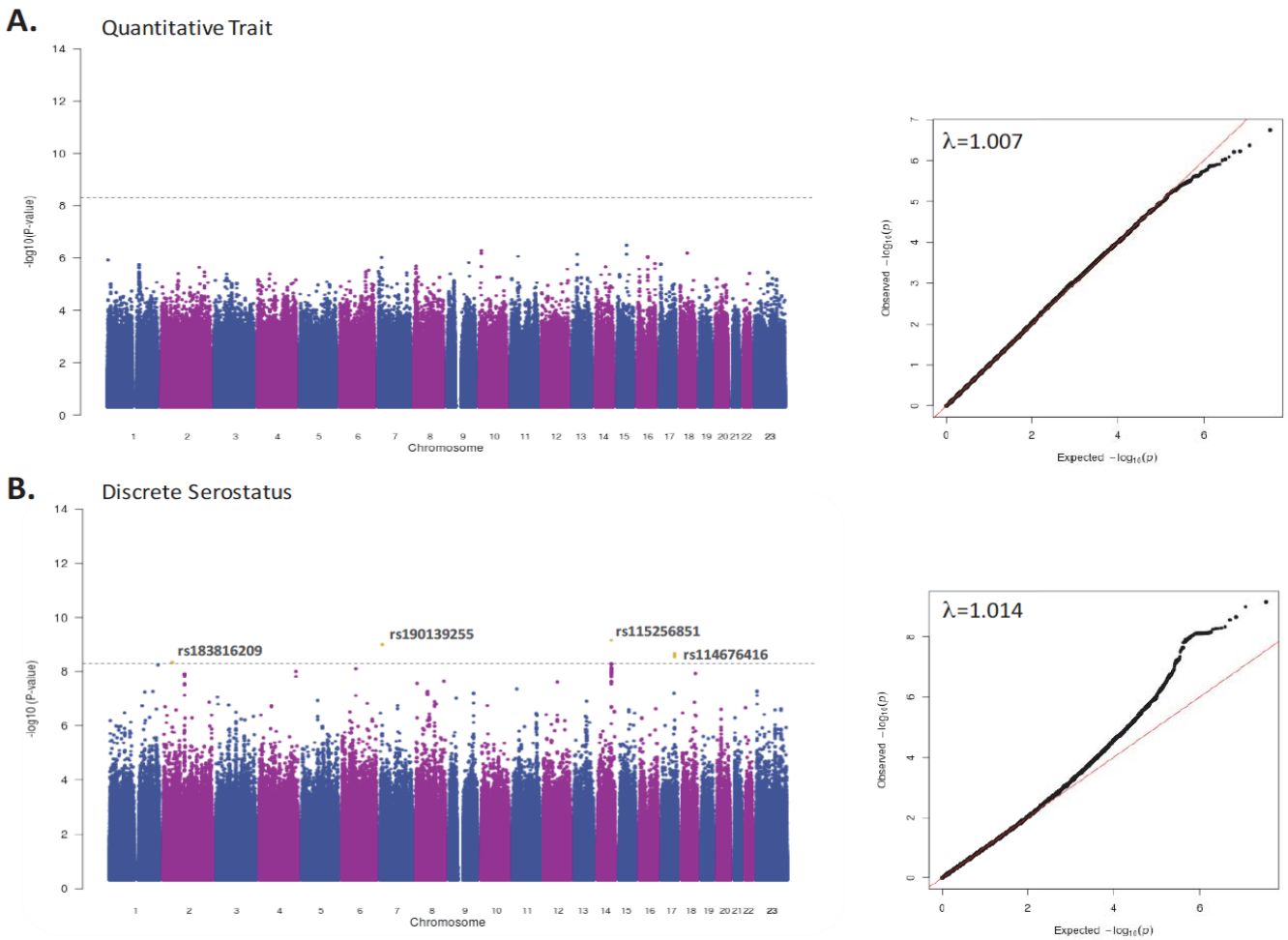


Fig. 3.3. Genome-wide association results of anti-VCA IgG response. Manhattan Plots (Left Panel), Grey dashed line: Genome-wide significance threshold ($p<5x10^{-9}$) and QQ Plots (Right panel). **A.** Quantitative IgG response levels of EBV seropositive individuals (N=1473). **B.** Discrete Serostatus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

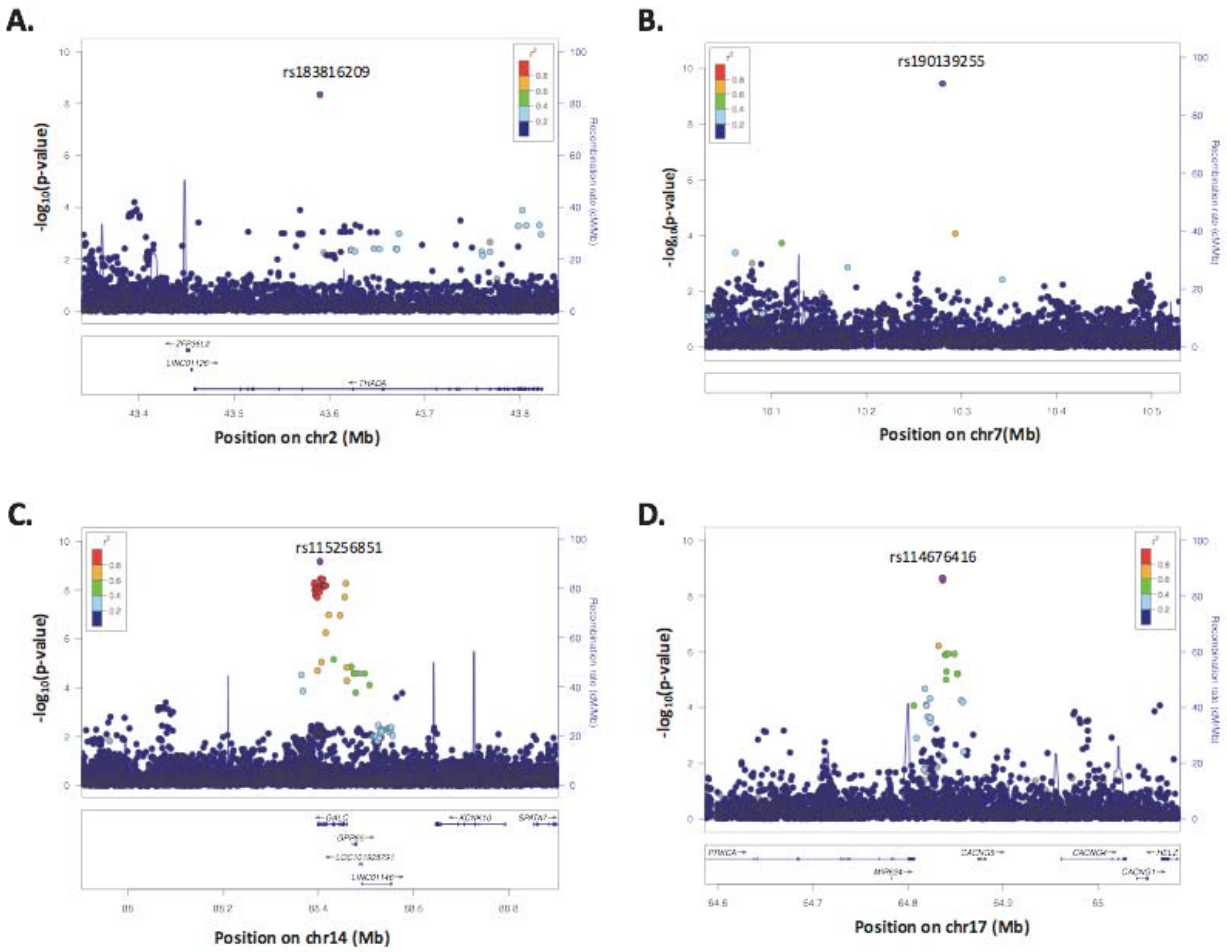


Fig. 3.4 Regional association plots for VCA serostatus genome-wide (GW) significant associations, N=1567, Pos=1350, Neg=217, threshold=p<5x10^{-9}. A. GW significant association on Chromosome 2 in the THADA region. **B.** GW significant association on Chromosome 7. **C.** GW significant association on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

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

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

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

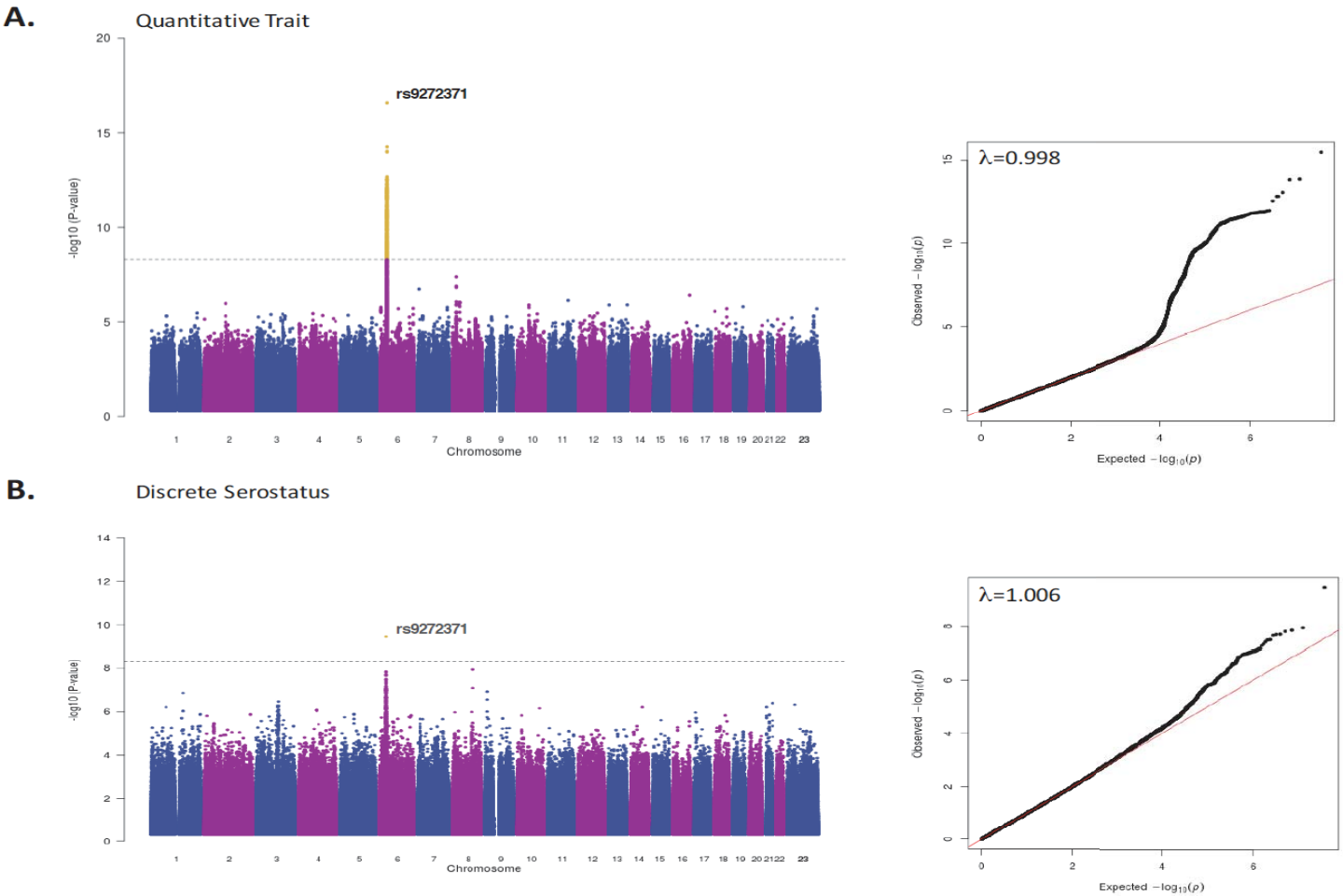


Fig. 3.6 Genome-wide association results of anti-EBNA-1 IgG response. Manhattan Plots (Left Panel), grey dashed line: Genome-wide significance threshold ($p<5x10^{-9}$) and QQ Plots (Right panel). **a.** Quantitative IgG response levels of EBV seropositive individuals (N=1473). **b.** Discrete Serostatus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

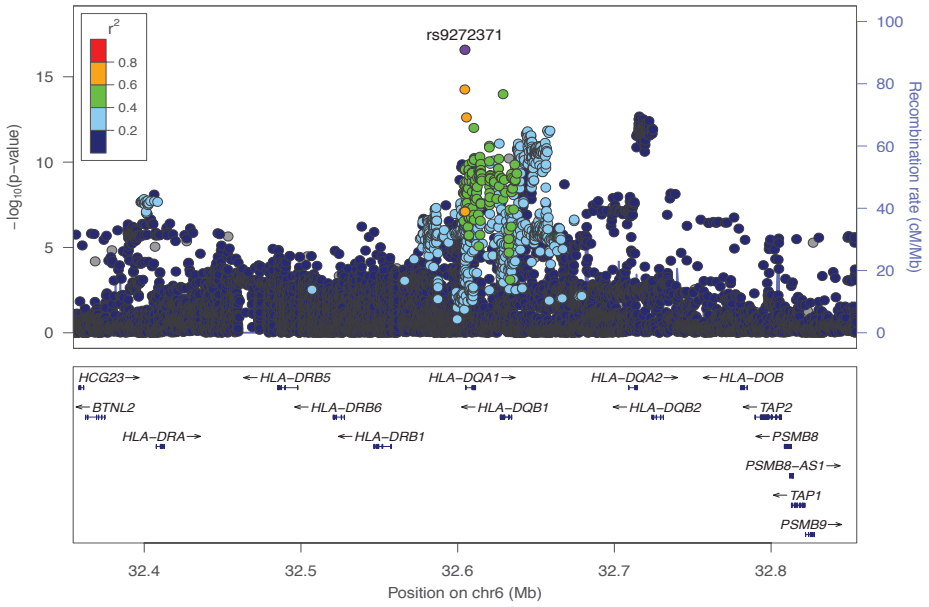


Fig. 3.7 Regional association plot for anti-EBNA-1 IgG response levels in 1473 individuals. (Genome-wide significance threshold= $p<5x10^{-9}$). The lead SNP is labelled and coloured in purple. LD (r^2) was calculated based on SNP genotypes.

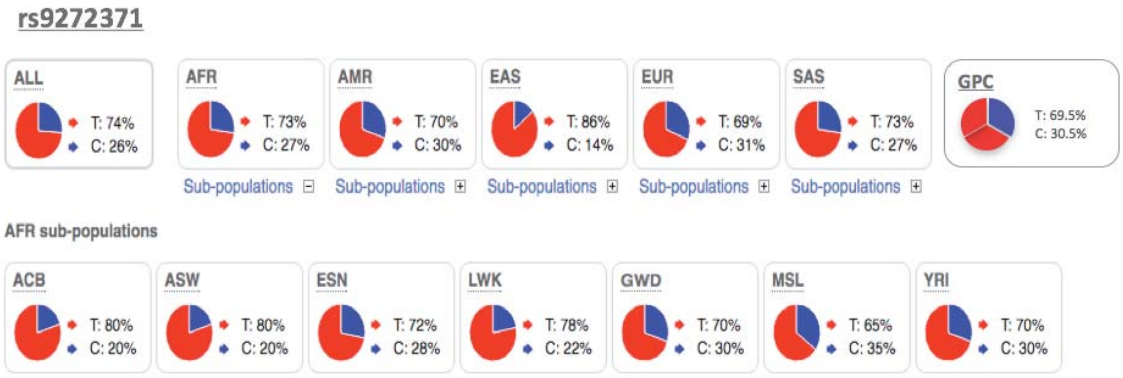


Fig. 3.8 Comparison of allele frequencies of lead EBNA-1 GWAS SNP, rs9272371 in HLA-DQA1 on chromosome 6 - between 1000 Genomes phase 3 populations and the GPC. Ancestry codes: AFR-African, AMR-American, EAS-East Asian, EUR, European, SAS-South Asian, ACB-African Caribbean, ASW- African Americans, ESN-Esan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

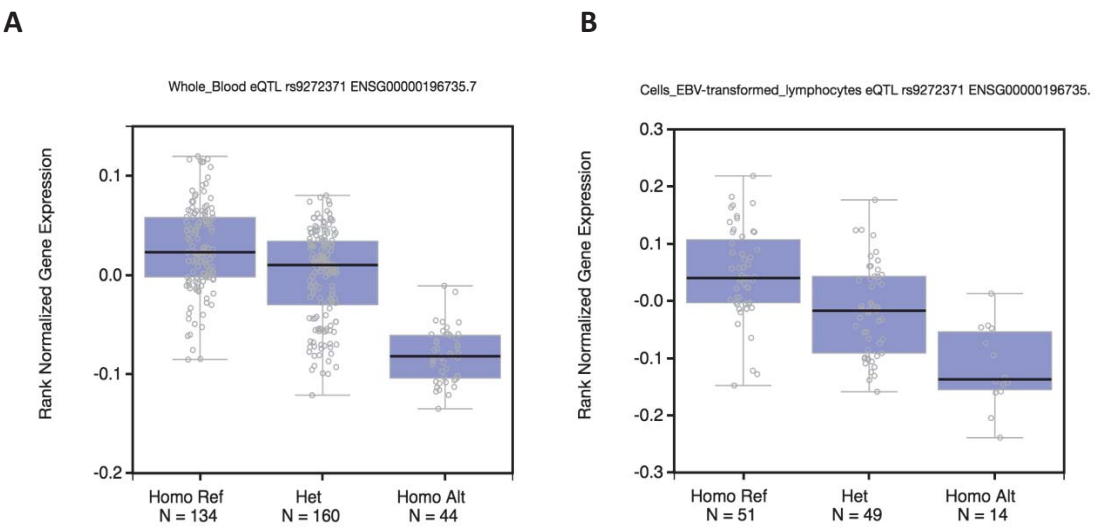


Fig. 3.9 The effect of rs9272371 genotypes on *HLA-DQA1* gene expression, cis-eQTL data from the GTEx database. A. Whole blood (eQTL $p=5.2x10^{-36}$, $\beta=-0.75$). **B.** EBV-transformed lymphocytes (eQTL $p=9x10^{-12}$, $\beta=-0.94$).

3.3.3 Multivariate Quantitative Association Boosts HLA Signal

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

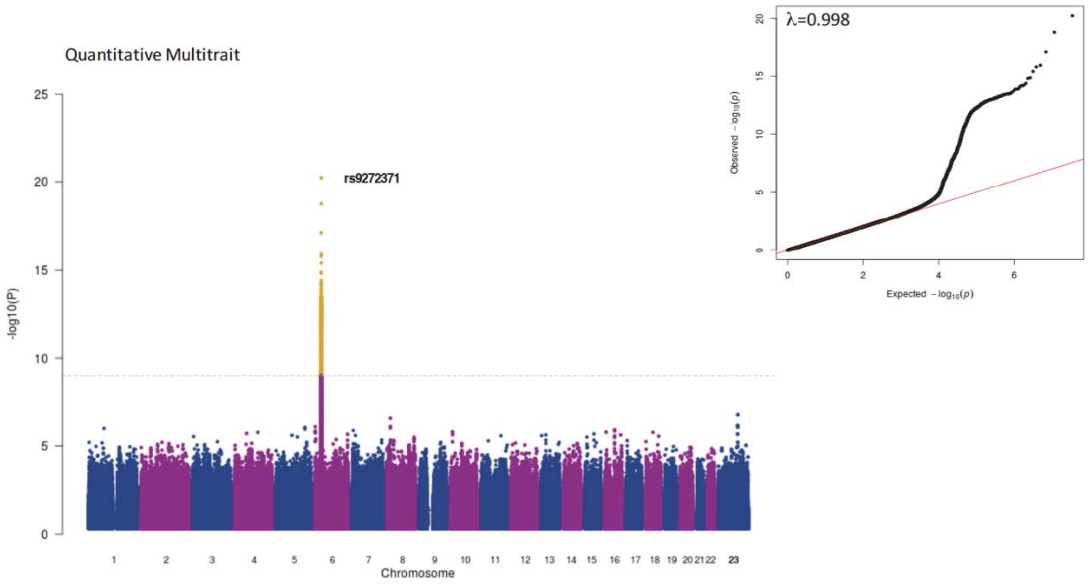


Fig. 3.10 Multivariate genome-wide association results of anti-EBV IgG response levels. Manhattan plot (Left Panel) and QQ Plot (right panel). N=1473. Grey dashed line: Genome-wide significance threshold ($p<5x10^{-9}$). 23=X-Chromosome

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

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

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

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

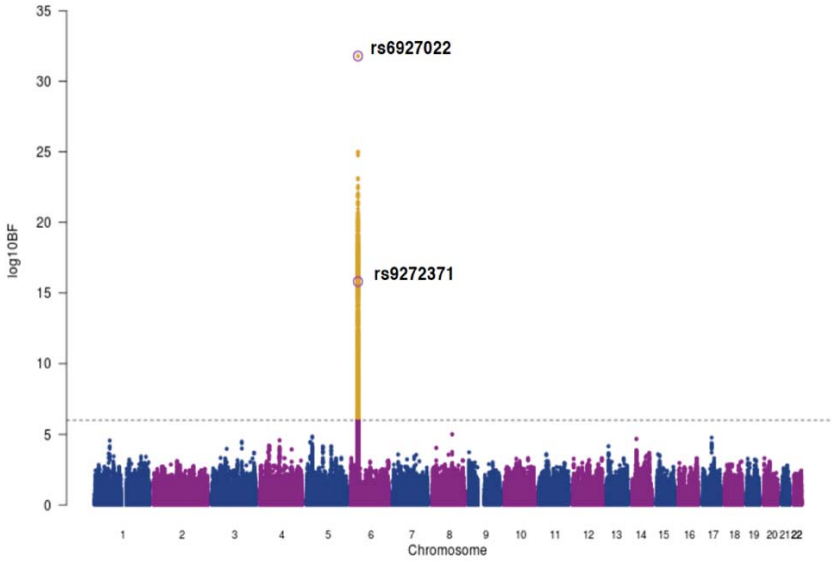
β >0 associated with higher antibody levels, <0 associated with lower antibody levels.

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

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

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

A



B

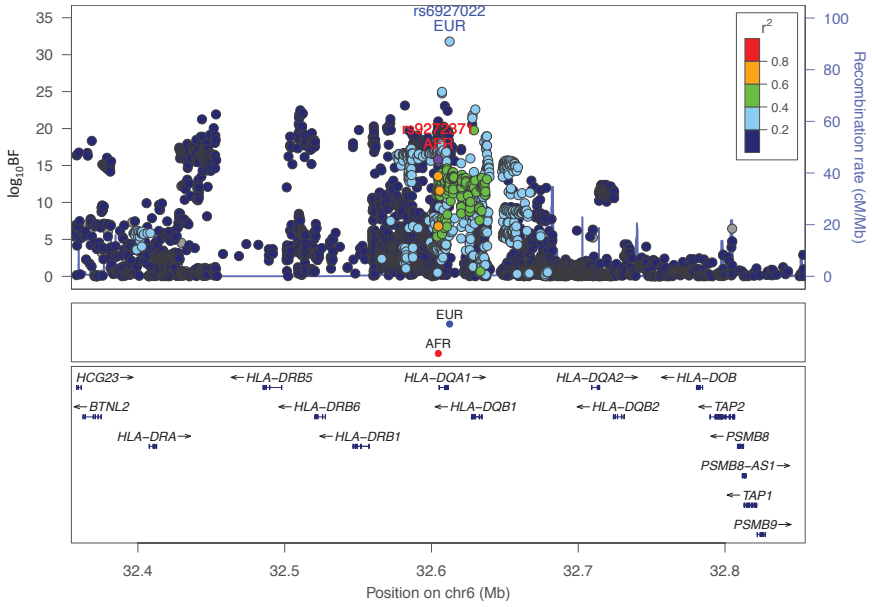


Fig. 3.11 Trans-ethnic meta-analysis association for EBNA-1 IgG response levels in 3635 individuals of Ugandan and European ancestry (EUR) (threshold= log_{10}BF >6). **A. Manhattan plot.** Grey dashed line: threshold= log_{10}BF > 6). The lead SNPs for EUR (rs6927022) and Uganda (rs9272371) GWASs are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

Table 3.3 Loci with strong evidence of association with anti-EBNA-1 IgG levels after trans-ethnic meta-analysis of Ugandan and European ancestry GWAS

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

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

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

Table 3.4 Conditional analysis of lead Ugandan and European SNPs

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

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

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

3.4 Discussion

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

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

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

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

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

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

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

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