# 6. Conclusions and Future Outlook

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

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

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

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

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

## 6.1 Inferring the Causality of Variants

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

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

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

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

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

## 6.3 Genome-to-Genome Analysis

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