Title: Genetic determinants of antibody mediated immune responses to infectious diseases agents: a genome-wide and HLA association study

Authors: *Guillaume Butler-Laporte*¹,², *Devin Kreuzer*¹, Tomoko Nakanishi¹,³,⁴, Adil Harroud⁵,⁶, Vincenzo Forgetta¹, J Brent Richards¹,²,⁷,⁸

*These authors made equal contributions to this manuscript.

Affiliations:

1) Lady Davis Institute, Jewish General Hospital, McGill University, Montréal, Québec, Canada
2) Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal, Québec, Canada
3) Department of Human Genetics, McGill University, Montréal, Canada
4) Kyoto-McGill International Collaborative School in Genomic Medicine, Graduate School of Medicine, Kyoto University, Kyoto, Japan
5) Department of Neurology, University of California San Francisco, San Francisco, CA, USA.
6) Weill Institute for Neurosciences, University of California San Francisco, San Francisco, CA, USA.
7) Department of Human Genetics, McGill University, Montréal, Québec, Canada
8) Department of Twin Research, King’s College London, London, UK

Corresponding authors:

Guillaume Butler-Laporte
Email: guillaume.butler-laporte@mail.mcgill.ca

J Brent Richards
Email: brent.richards@mcgill.ca
**Funding source:** The Richards research group is supported by the Canadian Institutes of Health Research (CIHR), the Lady Davis Institute of the Jewish General Hospital, the Canadian Foundation for Innovation, the NIH Foundation, Cancer Research UK and the Fonds de Recherche Québec Santé (FRQS). JBR is supported by a FRQS Clinical Research Scholarship. TwinsUK is funded by the Welcome Trust, Medical Research Council, European Union, the National Institute for Health Research (NIHR)-funded BioResource, Clinical Research Facility and Biomedical Research Centre based at Guy’s and St Thomas’ NHS Foundation Trust in partnership with King’s College London. These funding agencies had no role in the design, implementation or interpretation of this study.

**Conflict of interest:** JBR has served as an advisor to GlaxoSmithKline and Deerfield Capital.

**Key words:** serology, infections, genome-wide association study, human leukocyte antigen, LASSO
Abstract

Introduction

Infectious diseases are causally related to a large array of non-communicable diseases (NCDs). Identifying genetic determinants of infections and antibody-mediated immune responses may shed light on this relationship and provide therapeutic targets for drug and vaccine development.

Methods

We used the UK biobank cohort of up to 10,000 serological measurements of infectious diseases and genome-wide genotyping. We used data on 13 pathogens to define 46 phenotypes: 15 seropositivity case-control phenotypes, and 31 quantitative antibody measurement phenotypes. For each of these, we performed genome-wide association studies (GWAS) using the fastGWA linear mixed model package, and human leukocyte antigen (HLA) classical allele and amino acid residue associations analyses using Lasso regression for variable selection.

Results

We included a total of 8735 individuals for case-control phenotypes, and an average of 4286 samples per quantitative analyses (range: 276 to 8555). Fourteen of the GWAS yielded a genome-wide significant ($p<5\times10^{-8}$) loci at the major histocompatibility complex (MHC) on chromosome 6. Outside the MHC, we found a total of 60 loci, multiple associated with Epstein-Barr virus (EBV) related NCDs (e.g. RASA3, MED12L, and IRF4). FUT2 was also identified as an important gene for polyomaviridae. HLA analysis highlighted the importance DRB1*09:01, DQB1*02:01, DQA1*01:02, and DQA1*03:01 in EBV serologies, and of DRB1*15:01 in polyomaviridae.

Conclusion

We have identified multiple genetic variants associated with antibody immune response to 13 infections, many of which are biologically plausible therapeutic or vaccine targets. This may help prioritize future research and drug development.
Introduction

Infectious agents have been implicated in the pathogenesis of many non-communicable diseases (NCD)\(^1\). They rank as the third leading cause of cancers worldwide\(^2\), and are associated with multiple chronic conditions such as rheumatoid arthritis\(^3\), Alzheimer’s dementia\(^4\), and multiple sclerosis\(^5\). Measuring the antibody immune response to infections is a common approach to study this relationship, as seropositivity can serve as both a marker of infection exposure and provide clues about the pathophysiology between infections and NCDs (e.g. through molecular mimicry\(^6\)). This could potentially lead to important diagnostic and therapeutic advances, especially in the fields of vaccination and immunotherapy\(^7-10\). Further, given the recent global COVID-19 pandemic, understanding the immune responses to infection could help to identify common pathways that, when perturbed, influence susceptibility to infection and/or immunological response. However, the underlying pathophysiology tying infectious agents and NCDs often remains poorly understood, and their true causal effect also often remains unclear.

One approach to improve our understanding of these associations is through the study of genetic markers of infectious diseases susceptibility or host immune response. Genome-wide association studies (GWAS) have been used to identify the genetic determinants of a large array of diseases, improving our understanding of their pathophysiology, and leading to therapeutic advances\(^11\). The infectious disease immune response is also intrinsically tied to the human leukocyte antigen (HLA) system\(^12\), encoded by the major histocompatibility complex (MHC) gene complex. The MHC is a region of chromosome 6 with a high density of highly polymorphic genes, often in linkage disequilibrium, rendering any disease association study difficult to perform without adequate sample sizes and appropriate statistical methods\(^13\). However, few GWAS and HLA association studies have been published on genetic determinants of infectious diseases, and most were limited by small sample sizes\(^14\), or relied on patient self-reported infectious diseases diagnosis\(^15\), both factors lead to a decrease statistical power and difficulties with clinical interpretation.

Therefore in order to better understand the human immunological response to infectious diseases, we undertook GWASs and HLA association studies using the UK Biobank\(^16\) cohort (UKB) in up to 10,000 serological measurements of 20 infectious diseases

Methods

Phenotype

We performed this analysis using the UKB’s serological measurements of infectious agents. The UKB recruited over half a million British adults between 2006 and 2010, amongst which a subsample of 9724 participants provided serum samples for serological measurements of 20
different microorganisms. Samples were tested for total antibody levels against multiple antigens. These were measured using fluorescent bead-based multiplex serology technology at a dilution of 1:1000 using the Luminex 100 platform (Luminex Corporation, Austin, Texas). This method provides the median fluorescence intensity (MFI), a standardized quantification of the amount of antibody in the sample obtained by measuring the fluorescence emitted by the analyte-capture agent complex. Validation was performed using separate serum samples and a reference gold standard\textsuperscript{17}. This method and the choice of seropositivity threshold was previously validated for multiple infectious agents\textsuperscript{17}. Finally, of the 20 original pathogens, we selected the ones with a seroprevalence of more than 15% for our GWAS,\textsuperscript{1} in order to ensure adequate statistical power to identify associated loci. Table 1 provides more details on the selected infectious agents.

Genome-wide association study

For the GWAS, we used the UKB Version 3 imputed genotype dataset, with genome-wide genotyping data available for 488,000 UK Biobank participants. Details on the collection of this data can be found elsewhere\textsuperscript{16}. For each of the selected microorganisms, we performed three GWAS. First, we performed case-control analyses by splitting participants in seropositive and seronegative samples, based on the UKB suggested seropositivity definitions (Table 1). These GWAS aim to identify genetic variants associated with previous infections to each given pathogen. Second, we performed quantitative analyses using antibody MFI measurements. Given that all serological tests are at risk of low-level cross binding with other non-specific antibodies that are not representative of infection\textsuperscript{18}, we restricted the quantitative analyses to samples above the seropositivity threshold, using thresholds suggested by UKB (Table 1). Therefore, these GWAS aim to identify genetic variants responsible for varying antibody-mediated immune responses within the seropositive population. Third, given the risk of heavily skewed data leading and inflation of variance with higher MFI (in violation of linear regression assumptions), we performed the same quantitative analysis but using a logarithmic transform on antibody MFI.

For all GWAS, we restricted our analysis to white-British individuals as identified using principal component analysis\textsuperscript{19}. This was performed to minimize bias from population stratification, which tends to confound the relationship between the genetic variants and the phenotype of interest\textsuperscript{20}. For the two quantitative analysis, we standardized the data to have a mean of zero and a standard deviation of 1, prior to applying the algorithm. We performed all GWAS using the fastGWA mixed model package\textsuperscript{21} with the following covariates: sex, age, UKB assessment centre visited, and the first 20 principal components. We excluded SNPs with minor allele frequencies lower than 1%. A prespecified p-value threshold of $P<5 \times 10^{-8}$ was then used to identify genome-wide significant SNPs\textsuperscript{22}. Lead genetic variants at each loci were identified using the plink software\textsuperscript{23,24} (v1.90b6.10) clump function with linkage disequilibrium threshold of 0.2, and a physical distance threshold of 500 kilo-base pairs.
HLA association study

Next, we performed HLA association studies on the same three phenotypes as those defined above for our GWASs. This was also done for each of the selected pathogens. HLA classical alleles imputation was performed by the UKB using the HLA*IMP:02 algorithm. The following HLA class I and II genes were studied: A, B, C, DRB1, DRB3, DRB4, DRB5, DPA1, DPB1, DQA1, and DQB1. As recommended by the UKB, we set alleles with a posterior probability call of less than 0.7 to a copy number of 0. Similarly, the UKB used the 99:01 suffix to indicate that no DRB3-4-5 alleles were present and recommends setting them to 0. Given the high degree of collinearity between HLA gene alleles, and the large number of statistical tests planned, we used Lasso regression for variable selection. Lasso imposes a penalty on models selecting larger variable numbers and is therefore less likely to uncover false associations than stepwise regression methods. The analysis used all HLA allele copy numbers, age, sex, and the first 20 principal components as variables from which to select a sparse model. Analyses was also restricted to white British individuals. To further reduce false associations, we performed 10-fold cross-validation 100 times, and only selected variables that were chosen by Lasso in at least 95 of our analyses, using the 1-standard-error selection rule. In cases of repeated measurements from the same individuals, we average their results (for quantitative MFI phenotypes), and an individual was called seropositive if at least one serology test was above seropositivity threshold (for case-control phenotypes). Analyses were performed using the glmnet package (v2.0.16) on R (v3.5.0).

Amino acid residue association study

Given the highly polymorphic HLA genes, different alleles may encode similar sequences of amino acids, and studying the association between diseases and those amino acid residue sequences can be a more powerful and informative statistical analysis. To do this, we first used the IMGT/HLA database to translate each HLA allele copy number (as defined above) to amino acid copy numbers. In cases where the specific amino acid residue was unknown (i.e. an asterisk in the IMGT/HLA database), we set it to 0. We used the same Lasso analysis pathway described above, using amino acid copy numbers, age, sex, and the first twenty principal components. Repeated serology measurements were handled as above. Note that given that some alleles and amino acid residues are not commonly found in every population, amino acid residues with 100% correlation were analyzed as the same and were reported as such in the results.

Review of previously reported associations

We used the NHGRI-EBI GWAS catalog and the Phenoscanner tool to review previously reported disease associations with the genome-wide significant SNPs found in our GWASs, and with the alleles selected using the HLA alleles Lasso analyses. Note that the
Phenoscanner uses SNP rsID rather than alleles identifiers for the HLA region. We restricted our results to genome-wide significant associations.

**Patient consent statement**

Consent was obtained by the UKB for every enrolled participant.

**Results**

A total of 13 pathogens were chosen from the original list of infectious disease agents (Table 1). Most microbes were viruses of the Herpesviridae family, though the three Polyomaviridae viruses most commonly associated with human NCDs were also selected. Seroprevalence was highly variable, with Epstein-Barr Virus (EBV) antibodies being found in more than 94% of individuals, and *Chlamydia trachomatis* in less than 20%. There were a total of 8735 white British individuals participants (55.9% female) providing a total of 8984 individual samples. There was an average of 4286 samples used for quantitative analyses (range: 276 to 8555). Median age was 58 (interquartile range: 51-64) at time of enrolment to UKB.

As suspected prior to performing the analyses, log transformed MFI analyses had more stable estimates than their untransformed counterparts. Visual inspection of the untransformed antibody MFI GWAS Manhattan plots showed a large amount of genome-wide significant loci, even in analyses with smaller sample sizes (results not shown). This suggests that the variance stabilizing logarithmic transform was able to reduce most of the noise observed in the signals. Therefore, here we will only report on results from case-control and log transformed antibody MFI analyses. Manhattan plots from these analyses are available in Supplement 1.

**GWAS**

A total of 46 GWASs were performed: 15 case-control analyses, and 31 logarithm transformed MFI analyses. Genome-wide significant loci are shown in Table 2. Genomic inflation factors (Supplement 2) were smaller than 1.04 for all GWASs except for mopp D MFI (log transformed) at 1.24. This analyses also yielded 22 separate loci using LD clumping, the largest number of genome-wide significant peaks. However, this was the analysis with the smallest sample size (276 individuals). Given the genomic inflation, an underlying population stratification and an elevated rate of false positive associations should be suspected.
Reassuringly, 14 analyses showed a locus at the MHC (Figure 1), with EBV (p=1.9x10^{-76} to 3.1x10^{-16}), JC virus (JCV) (p=1.7x10^{-47} to 4.5x10^{-24}), and Merkel Cell Virus (MCV) (p=7.3x10^{-37} to 1.0x10^{-18}) showing the strongest associations. Interestingly, the FUT2 gene on chromosome 19 was identified in both the case-control JCV seropositivity analysis and the BK virus VP1 log transformed MFI analysis. Otherwise, no other genetic loci were identified more than once, including within infectious agent families.

**HLA allele association analyses**

EBV antibody MFI (EA, EBNA-1, and ZEBRA) were associated with the most HLA alleles, with Lasso selecting 10 alleles for EBNA-1. The DQB1*02:01 and DQA1*03:01 alleles were found in 2 out of the three EBV antibody analyses (EA-D, EBNA-1, and ZEBRA) in Table 1. Interestingly, despite a strong MHC locus in the GWAS, Lasso did not select any alleles for VCA antibody MFI. This is most likely because the effect sizes and collinearity from any of the HLA alleles were likely too small to confidently select them in a multivariate analysis such as Lasso. Amongst the Polyomaviridae, both JCV and MCV replicated their GWAS finding, with DRB1*15:01 and DRB5*01:01 (also identified in EBNA-1) being selected in both antibody MFI analyses. Of note, these two HLA alleles are in almost complete linkage disequilibrium in European populations. Finally, varicella zoster virus antibody MFI was associated with three HLA alleles, one of which (DQB1*02:01) also being associated with two EBV antibodies (EA-D and EBNA). Results from the HLA allele association studies are summarized in Table 3.

**HLA amino acid residue association analyses**

Lasso selected amino acid residues from most selected alleles in the HLA allele association above, and also selected amino acid residues from proteins encoded by genes whose alleles were not previously selected. There were some differences, most significantly for the ZEBRA MFI analysis. This is likely to be due to the high rate of multicollinearity between amino-acids residues in encoded by MHC genes, which even Lasso was not able to entirely untangle. Full results can be found in Supplement 3.

**Review of previously reported associations**

Of the 88 associations obtained from either GWAS or HLA allele Lasso analysis, the NHGRI-EBI GWAS catalog found previously reported associations in 17 variants. In total, there were 73 previously reported associations, 57 in the HLA regions. Using the Phenoscanner and the 52 loci obtained from our GWASs, 24 had previously reported associations (total of 453 associations), of which 14 were in non-HLA regions (total of 271 associations). With both tools, associations included a wide range of diseases, including
autoimmune diseases, white blood cell counts, body-mass index, or respiratory functions (Supplement 4 and Supplement 5).

**Discussion**

Infectious diseases are a major contributor to the global burden of diseases and play a significant role in many NCDs. Given that multiple factors, both heritable and environmental, contribute to their transmission, their acquisition, and the human host response, studying their genetic determinants can be challenging. In this study, we have used measurements from the UKB’s serological results to perform the largest reported genetic association study of infectious diseases antibody measurements to date. While the HLA region was commonly associated with the host’s antibody-mediated immune responses, we have uncovered multiple biologically plausible genetic determinants of infectious diseases. For example, for EBV, we found loci at RASA3, MED12L, and IRF4 which are located on chromosomes 5, 17, and 15 respectively. For EBV, we found loci at RASA3, MED12L, and IRF4 which are located on chromosomes 5, 17, and 15 respectively. Further, while our results cannot prove causality, multiple previously reported epidemiological associations support our methodology. For example, allele DRB1*15:01 was associated with both JCV analysis and with the EBV ZEBRA analysis. DRB1*15:01 is the main genetic risk factor for multiple sclerosis, a disease previously associated with EBV. DRB1*15:01 was also selected by Lasso as a predictor of JCV serology, the cause of progressive multifocal leukoencephalopathy, another demyelinating disease which can also be triggered when giving immunosuppression to patients with multiple sclerosis. Similarly, FUT2 was associated with both BKV and JCV, two polyomaviridae with high genetic homology, and is therefore likely to be truly associated with their pathophysiology. FUT2 is a major determinant of the blood group secretor status and has been associated with multiple other viruses and NCDs, most notably kidney diseases. Uncovering these genetic determinants may allow for better informed drug and vaccine development, which typically is a long and expensive process. Similar genetics-informed prioritization of therapeutic targets for further drug and vaccine development has previously been used in multiple other diseases with success.

There are few published GWASs on human infectious diseases, and each has employed a different methodology to enroll patients, making comparisons difficult. Our study used serology data, but other have used prospective patient enrolment, or electronic medical record data, and the biggest GWAS on infectious diseases to date used self-reported history using questionnaires. A similar pre-print using the same data as ours is currently accessible and found similar significant associations. However, they performed a slightly different analysis that did not account for intrinsically limited serology testing specificity, or UKB HLA quality control recommendations. Nevertheless, as we found here, the HLA emerges as an important risk loci across most published GWASs.
Our study’s main strength lies in its careful methodology. Given the large number of statistical tests performed and inherent diagnostic limitations of serological tests, we made multiple analytic choices to ensure interpretable results, while lowering the risk of false associations. Most notably, we used Lasso regression to select HLA alleles and amino acid residues, and we used logarithmic transform to stabilize the variance of the antibody MFI analyses.

However, given that serological tests may have multiple possible interpretations our results should be interpreted with caution. That is because the differential diagnosis of a negative serological test includes having never been in contact with the infectious agent, the host not being able to mount an antibody mediated response, or antibodies not being a good proxy for either contact or immune response. Alternatively, a positive antibody titre may be explained by cross-reactivity with other antigens, especially if the antibody titers are low. Further, antibody levels are known to vary in time due to multiple host and environmental factors. To better assess antibody mediated response in hosts with a likely exposure, we have limited our quantitative analyses to individuals above the seropositivity threshold. Nevertheless, given that the population was randomly selected from the UKB cohort, we cannot rule out that unmeasured environmental or socio-economic confounders may have affected our results.

Our study has several limitations which can help guide future efforts in studying genetic determinants of infectious diseases. First, ideally future serological studies should be performed in individuals with clear history of exposure (or lack thereof) to the infectious agent. This would increase the serological test’s specificity and improve the chance of finding clinically significant genetic associations. Second, as the environment is a major non-heritable determinant of infectious diseases, it should be factored in the design of future genetic studies. Third, as individuals are possibly constantly re-exposed to infectious diseases, longitudinal follow-up and serial measurements of both IgM and IgG would also improve the chance of finding clinically important genetic variants. Lastly, our results are mostly hypothesis generating, and despite the fact that we many of the loci we found to be associated with infectious diseases have been associated with other diseases, this needs to be followed by in-vitro or in-vivo studies to establish true causal associations. This is especially true given that most of the associations we found were in the HLA, a region of the human genome that does not lend itself well to further in-silico study.

In summary, here we present the largest GWAS and HLA association analyses on infectious diseases and the resultant host antibody-mediated immune response to date. While, this work is hypothesis generating and should not be used to infer causality between infections and
NCDs without further research, with careful planning, we hope that future genetic studies will lead to further advances in our understanding of the interplay between, host, environment, and disease pathophysiology.

Data availability

Summary statistics from the GWAS will be made accessible through the NHGRI-EBI GWAS catalog.

Acknowledgement

This research has been conducted using the UK Biobank Resource under Application Number 27449.

Funding

The Richards research group is supported by the Canadian Institutes of Health Research (CIHR), the Lady Davis Institute of the Jewish General Hospital, the Canadian Foundation for Innovation, the NIH Foundation, Cancer Research UK and the Fonds de Recherche Québec Santé (FRQS). JBR is supported by a FRQS Clinical Research Scholarship. TwinsUK is funded by the Welcome Trust, Medical Research Council, European Union, the National Institute for Health Research (NIHR)-funded BioResource, Clinical Research Facility and Biomedical Research Centre based at Guy’s and St Thomas’ NHS Foundation Trust in partnership with King’s College London. These funding agencies had no role in the design, implementation or interpretation of this study.

Conflict of Interests

JBR has served as an advisor to GlaxoSmithKline.

Supplementary files description

Supplement 1: Manhattan Plots and QQ plots for seropositivity case-control studies, and quantitative log transformed MFI studies.

Supplement 2: Genomic Inflation Factors for case control and logarithmic transformed MFI GWASs.

Supplement 3: HLA amino acid residue analysis association analysis summary.
**Supplement 4:** Results from the NHGRI-EBI GWAS catalog search.

**Supplement 5:** Results from the Phenoscanner search.

**References**

1. O’Connor SM, Taylor CE, Hughes JM. Emerging infectious determinants of chronic diseases. *Emerg Infect Dis* 2006; 12(7): 1051-7.
2. de Martel C, Ferlay J, Franceschi S, et al. Global burden of cancers attributable to infections in 2008: a review and synthetic analysis. *The Lancet Oncology* 2012; 13(6): 607-15.
3. Broadley I, Pera A, Morrow G, Davies KA, Kern F. Expansions of Cytotoxic CD4+CD28− T Cells Drive Excess Cardiovascular Mortality in Rheumatoid Arthritis and Other Chronic Inflammatory Conditions and Are Triggered by CMV Infection. *Frontiers in Immunology* 2017; 8(195).
4. Readhead B, Haure-Mirande J-V, Funk CC, et al. Multiscale Analysis of Independent Alzheimer’s Cohorts Finds Disruption of Molecular, Genetic, and Clinical Networks by Human Herpesvirus. *Neuron* 2018; 99(1): 64-82.e7.
5. Vanheusden M, Broux B, Welten SPM, et al. Cytomegalovirus infection exacerbates autoimmune mediated neuroinflammation. *Sci Rep* 2017; 7(1): 663-.
6. Cusick MF, Libbey JE, Fujinami RS. Molecular mimicry as a mechanism of autoimmune disease. *Clin Rev Allergy Immunol* 2012; 42(1): 102-11.
7. Bonsack M, Mohan N, Öhlenschläger K, et al. Application of individual HLA-binding prediction thresholds increases the detection of target cell surface-presented HPV16 E6- and E7-derived epitopes as a basis for therapeutic HPV vaccine design. *The Journal of Immunology* 2020; 204(1 Supplement): 169.8.
8. Garrido F. HLA Class-I Expression and Cancer Immunotherapy. *Advances in experimental medicine and biology* 2019; 1151: 79-90.
9. Chowell D, Morris LGT, Grigg CM, et al. Patient HLA class I genotype influences cancer response to checkpoint blockade immunotherapy. *Science* 2018; 359(6375): 582-7.
10. Garrido F. HLA Class-I Expression and Cancer Immunotherapy. In: Garrido F, ed. MHC Class-I Loss and Cancer Immune Escape. Cham: Springer International Publishing; 2019: 79-90.
11. Kessler T, Vilne B, Schunkert H. The impact of genome-wide association studies on the pathophysiology and therapy of cardiovascular disease. *EMBO Mol Med* 2016; 8(7): 688-701.
12. Blackwell JM, Jamieson SE, Burgner D. HLA and infectious diseases. *Clinical microbiology reviews* 2009; 22(2): 370-85, Table of Contents.
13. de Bakker PIW, McVean G, Sabeti PC, et al. A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. *Nature genetics* 2006; 38(10): 1166-72.
14. Nelson CL, Pelak K, Podgoreanu MV, et al. A genome-wide association study of variants associated with acquisition of Staphylococcus aureus bacteremia in a healthcare setting. *BMC Infect Dis* 2014; 14: 83-.
15. Tian C, Hromatka BS, Kiefer AK, et al. Genome-wide association and HLA region fine-mapping studies identify susceptibility loci for multiple common infections. *Nat Commun* 2017; 8(1): 599-.
16. Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* 2018; 562(7726): 203-9.
17. Brenner N, Waterboer T. Serological measurement of infectious agents in UK Biobank: a pilot study in 10,000 samples. 2019. [https://biobank.ctsu.ox.ac.uk/crystal/crystal/docs/infldisease.pdf](https://biobank.ctsu.ox.ac.uk/crystal/crystal/docs/infldisease.pdf) (accessed 2019-11-21).
18. Tate J, Ward G. Interferences in immunoassay. *Clin Biochem Rev* 2004; 25(2): 105-20.
19. Morris JA, Kemp JP, Youlten SE, et al. An atlas of genetic influences on osteoporosis in humans and mice. Nature genetics 2019; 51(2): 258-66.
20. Persyn E, Redon R, Bellanger L, Dina C. The impact of a fine-scale population stratification on rare variant association test results. PLOS ONE 2018; 13(12): e0207677.
21. Jiang L, Zheng Z, Qi T, et al. A resource-efficient tool for mixed model association analysis of large-scale data. Nature genetics 2019; 51(12): 1749-55.
22. Xu C, Tachmazidou I, Walter K, et al. Estimating genome-wide significance for whole-genome sequencing studies. Genet Epidemiol 2014; 38(4): 281-90.
23. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007; 81(3): 559-75.
24. Purcell S. PLINK (v1.09b6.10) http://pngu.mgh.harvard.edu/purcell/plink/, 2019.
25. UKBiobank. Imputation of classical HLA types from UK Biobank genotype data. 2016. https://biobank.ctsu.ox.ac.uk/crystal/crystal/docs/HLA_imputation.pdf (accessed 2020-06-10).
26. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. New York: Springer; 2017.
27. Raychaudhuri S, Sandor C, Stahl EA, et al. Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. Nature genetics 2012; 44(3): 291-6.
28. Robinson J, Halliwell JA, Hayhurst JD, Flice P, Parham P, Marsh Steven GE. The IPD and IMGT/HLA database: allele variant databases. Nucleic Acids Research 2014; 43(D1): D423-D31.
29. Buniello A, MacArthur JAL, Cerezo M, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res 2019; 47(D1): D1005-d12.
30. Kamat MA, Blackshaw JA, Young R, et al. PhenoScanner V2: an expanded tool for searching human genotype-phenotype associations. Bioinformatics (Oxford, England) 2019; 35(22): 4851-3.
31. Staley JR, Blackshaw J, Kamat MA, et al. PhenoScanner: a database of human genotype-phenotype associations. Bioinformatics (Oxford, England) 2016; 32(20): 3207-9.
32. Dalianis T, Hirsch HH. Human polyomaviruses in disease and cancer. Virology 2013; 437(2): 63-72.
33. Quandt JA, Huh J, Baig M, et al. Myelin basic protein-specific TCR/HLA-DRB5*01:01 transgenic mice support the etiologic role of DRB5*01:01 in multiple sclerosis. Journal of immunology (Baltimore, Md : 1950) 2012; 189(6): 2897-908.
34. Fogdell A, Olerup O, Fredriksson S, Vrethem M, Hillert J. Linkage analysis of HLA class II genes in Swedish multiplex families with multiple sclerosis. Neurology 1997; 48(3): 758-62.
35. Ramasubramanyan S, Osborn K, Al-Mohammad R, et al. Epstein-Barr virus transcription factor Zta acts through distal regulatory elements to directly control cellular gene expression. Nucleic acids research 2015; 43(7): 3563-77.
36. Dresang LR, Vereide DT, Sugden B. Identifying sites bound by Epstein-Barr virus nuclear antigen 1 (EBNA1) in the human genome: defining a position-weighted matrix to predict sites bound by EBNA1 in viral genomes. Journal of virology 2009; 83(7): 2930-40.
37. Gao Y, Wang L, Lei Z, Li J, Forrest JC, Liang X. IRF4 promotes Epstein-Barr virus activation in Burkitt’s lymphoma cells. The Journal of general virology 2019; 100(5): 851-62.
38. Liu W, Kim GB, Krump NA, Zhou Y, Riley JL, You J. Selective reactivation of STING signaling to target Merkel cell carcinoma. Proceedings of the National Academy of Sciences 2020: 201919690.
39. DeCaprio JA. Merkel cell polyomavirus and Merkel cell carcinoma. Philos Trans R Soc Lond B Biol Sci 2017; 372(1732): 20160276.
40. Alcina A, Abad-Grau MD, Fedetz M, et al. Multiple sclerosis risk variant HLA-DRB1*1501 associates with high expression of DRB1 gene in different human populations. PloS one 2012; 7(1): e29819-e.
41. Guan Y, Jakimovski D, Ramanathan M, Weinstock-Guttman B, Zivadinov R. The role of Epstein-Barr virus in multiple sclerosis: from molecular pathophysiology to in vivo imaging. *Neural Regen Res* 2019; **14**(3): 373-86.
42. Saribas AS, Ozdemir A, Lam C, Safak M. JC virus-induced Progressive Multifocal Leukoencephalopathy. *Future Virol* 2010; **5**(3): 313-23.
43. Sadiq SA, Puccio LM, Brydon EWA. JCV detection in multiple sclerosis patients treated with natalizumab. *Journal of Neurology* 2010; **257**(6): 954-8.
44. Cubitt CL. Molecular Genetics of the BK Virus. In: Ahsan N, ed. Polyomaviruses and Human Diseases. New York, NY: Springer New York; 2006: 85-95.
45. Thorven M, Grahn A, Hedlund K-O, et al. A Homozygous Nonsense Mutation (428G→A) in the Human Secretor (<em>FUT2</em>) Gene Provides Resistance to Symptomatic Norovirus (GGII) Infections. *Journal of Virology* 2005; **79**(24): 15351-5.
46. Azad MB, Wade KH, Timpson NJ. FUT2 secretor genotype and susceptibility to infections and chronic conditions in the ALSPAC cohort. *Wellcome Open Res* 2018; **3**: 65-.
47. Wuttke M, Li Y, Li M, et al. A catalog of genetic loci associated with kidney function from analyses of a million individuals. *Nature genetics* 2019; **51**(6): 957-72.
48. Shapiro MD, Tavori H, Fazio S. PCSK9: From Basic Science Discoveries to Clinical Trials. *Circ Res* 2018; **122**(10): 1420-38.
49. Zheng R, Li Z, He F, et al. Genome-wide association study identifies two risk loci for tuberculosis in Han Chinese. *Nat Commn* 2018; **9**(1): 4072.
50. Butler-Laporte G, Harrold A, Forgetta V, Richards JB. Elevated body mass index is associated with an increased risk of infectious disease admissions and mortality: a mendelian randomization study. *Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases* 2020.
51. Sveinbjornsson G, Gudbjartsson DF, Halldorsson BV, et al. HLA class II sequence variants influence tuberculosis risk in populations of European ancestry. *Nature genetics* 2016; **48**(3): 318-22.
52. Ellinghaus D, Degenhardt F, Bujanda L, et al. Genomewide Association Study of Severe Covid-19 with Respiratory Failure. *New England Journal of Medicine* 2020.
53. Kachuri L, Francis SS, Morrison M, et al. The landscape of host genetic factors involved in immune response to common viral infections. *medRxiv* 2020: 2020.05.01.20088054.
54. Klutts JS, Ford BA, Perez NR, Gronowski AM. Evidence-Based Approach for Interpretation of Epstein-Barr Virus Serological Patterns. *Journal of Clinical Microbiology* 2009; **47**(10): 3204-10.
55. Larsen SA, Steiner BM, Rudolph AH. Laboratory diagnosis and interpretation of tests for syphilis. *Clinical microbiology reviews* 1995; **8**(1): 1-21.
56. Landais C, Fenollar F, Thuny F, Raoult D. From Acute Q Fever to Endocarditis: Serological Follow-Up Strategy. *Clinical Infectious Diseases* 2007; **44**(10): 1337-40.
57. Haworth S, Mitchell R, Corbin L, et al. Apparent latent structure within the UK Biobank sample has implications for epidemiological analysis. *Nat Commun* 2019; **10**(1): 333.
| Infectious agents groups | Infectious agents | Antigen targets | MFI seropositivity threshold used for quantitative GWAS | Number of seropositive samples (%) (for MFI quantitative GWAS) | Seropositivity definitions if multiple antigens available (for case-control GWAS) |
|--------------------------|------------------|----------------|---------------------------------------------------|---------------------------------------------------------------|----------------------------------------------------------------------------------|
| **Herpesviridae**        |                  |                |                                                   |                                                               |                                                                                  |
| Herpes Simplex Virus-1   |                  | mgG-1          | 150                                               | 6199 (69.0)                                                    | n/a                                                                              |
| Herpes Simplex Virus-2   |                  | mgG-2          | 150                                               | 1382 (15.4)                                                   | n/a                                                                              |
| Epstein-Barr Virus       |                  | VCA p18        | 250                                               | 8518 (94.8)                                                   | Positive for 2 or more antigens                                                  |
|                         |                  | EBNA-1         | 250                                               | 7972 (88.7)                                                   | Total: 8477 (94.4%)                                                              |
|                         |                  | ZEBRA          | 100                                               | 8191 (91.2)                                                   |                                                                                  |
|                         |                  | EA-D           | 100                                               | 7763 (86.4)                                                   |                                                                                  |
| Human Cytomegalovirus    |                  | pp150          | 100                                               | 5136 (57.2)                                                   | Positive for 2 or more antigens                                                  |
|                         |                  | pp52           | 150                                               | 5681 (63.2)                                                   | Total: 5045 (56.2%)                                                             |
|                         |                  | pp28           | 200                                               | 5087 (56.6)                                                   |                                                                                  |
| Human Herpesvirus-6      |                  | IE1A           | 100                                               | 6968 (77.6)                                                   | Three seropositivity definitions assessed separately:                            |
|                         |                  | IE1B           | 100                                               | 7119 (79.2)                                                   | - Positive for any antigen, total: 8171 (91.0%)                                   |
|                         |                  | p101k          | 100                                               | 1951 (21.7)                                                   | - Positive for IE1A, total: 6968 (77.6%)                                         |
|                         |                  |                |                                                   |                                                               | - Positive for IE1B, total: 7119 (79.2%)                                         |
| Human Herpesvirus-7      |                  | U14            | 100                                               | 8528 (94.9)                                                   |                                                                                  |
| Varicella Zoster Virus   |                  | Glycoproteins E and I | 100                                         | 7595 (84.5)                                                   | n/a                                                                              |
| **Polyomaviridae**       |                  |                |                                                   |                                                               |                                                                                  |
| Human Polyomavirus BKV   |                  | BK VP1         | 250                                               | 8555 (95.2)                                                   | n/a                                                                              |
| Human Polyomavirus JCV   |                  | JC VP1         | 250                                               | 5118 (57.0)                                                   | n/a                                                                              |
| Merkel Cell Polyomavirus |                  | MC VP1         | 250                                               | 5915 (65.8)                                                   | n/a                                                                              |
| **Bacteria**             |                  |                |                                                   |                                                               |                                                                                  |
| Chlamydia trachomatis    |                  | momp A         | 100                                               | 964 (10.7)                                                    | Seropositive if either:                                                          |
|                         |                  | momp D         | 100                                               | 1371 (15.3)                                                   | - Positive for pGP3                                                              |
|                         |                  | tarp-D F1      | 100                                               | 1635 (18.2)                                                   | - Negative for pGP3, but positive for 2 out of 5 remaining antigens (but momp D and momp A can only contribute once together) |
|                | tarp-D F2 | 2074 (23.1) | Total: 1784 (19.9%) |
|----------------|-----------|-------------|---------------------|
| PorB           | 80        | 273 (3.04)  |                     |
| pGP3           | 200       | 1784 (19.9) |                     |
|                |           |             |                     |
| Helicobacter pylori |          |             |                     |
| CagA           | 400       | 985 (11.0)† | Positive for 2 or more antigens, except for CagA†. |
| VacA           | 100       | 1571 (17.5) |                     |
| OMP            | 170       | 2640 (29.4) |                     |
| GroEL          | 80        | 2716 (30.2) |                     |
| Catalase       | 180       | 1558 (17.3) |                     |
| UreA           | 130       | 2251 (25.1) |                     |
| Parasite       |           |             |                     |
| Toxoplasma gondii |          |             |                     |
| p22            | 100       | 1308 (14.6) | Positive for either antigen |
| sag1           | 160       | 3919 (43.6) |                     |
|                |           |             | Total: 2449 (27.3%) |

**Table 1: Details of Infectious Agents Studied.** This table is adapted from the UKB documentation\(^1\): infectious agents selected for the GWAS analyses, the antigen targets used for antibody measurements, and the seropositivity definitions used for case-control status. All counts and ratios are restricted to the 8761 samples from white British individuals. †CagA was not available for all participants and was therefore not used for seropositivity calculation.
| Agent          | Analysis                  | Variant       | CHR | Beta | SE   | P-value | Effect Allele | Other Allele | Frequency | Overlapping or Nearest Gene |
|----------------|---------------------------|---------------|-----|------|------|---------|---------------|--------------|-----------|-----------------------------|
| BKV            | VP1 MFI (log)             | rs492602      | 19  | -0.09| 0.015| 4.3e-09| A             | G            | 0.479     | FUT2                                     |
| C. trachomatis | Seropositivity Case-Control | rs143335233  | 2   | -0.14| 0.023| 5.4e-09| C             | T            | 0.983     | NCK2                                     |
| PorB MFI (log) | Chr4:13265941             | rs74725117    | 4   | -1.53| 0.28 | 3.1e-08| CT            | G            | 0.983     | RNU6-962P                                |
|                | rs201129973               |               | 6   | -1.74| 0.31 | 3.1e-08| G             | A            | 0.981     | RP11-302L19.1                           |
|                | rs61957300                |               | 13  | -1.28| 0.23 | 3.6e-08| A             | C            | 0.975     | TPTE2P2                                  |
|                | rs140031044               |               | 18  | -2.05| 0.33 | 7.4e-10| A             | G            | 0.989     | CTD-2008L17.1                           |
| pGP3 MFI (log) | Chr7:66874490             |               | 7   | -0.32| 0.059| 4.2e-08| CTCTT         | C            | 0.916     | AC006480.1                               |
| CMV            | pp28 MFI (log)            | rs12698418    | 7   | 0.13 | 0.024| 4.4e-08| G             | A            | 0.248     | EN2                                      |
| EBV            | Seropositivity Case-Control | rs71437272   | 13  | 0.07 | 0.012| 2.1e-08| C             | T            | 0.980     | RASA3                                    |
|                | VCA MFI (log)             | rs9379862     | 6   | 0.10 | 0.018| 1.1e-08| T             | C            | 0.740     | BTN3A2                                   |
|                | rs9271536                 |               | 6   | 0.19 | 0.020| 7.3e-21| A             | T            | 0.180     | HLA-DQA1 (MHC)                           |
|                | ZEBRA MFI (log)           | rs34034915    | 6   | 0.10 | 0.017| 3.3e-09| T             | TG           | 0.660     | RP1-97D16.1                              |
|                | Chr6:32597087             |               | 6   | -0.29| 0.016| 3.0e-75| CA            | C            | 0.637     | HLA-DQA1 (MHC)                           |
| EBNA-1 MFI (log)| rs67886110               |               | 3   | 0.09 | 0.016| 2.1e-08| G             | T            | 0.599     | MED12L                                   |
|                | rs6927022                 |               | 6   | 0.30 | 0.016| 1.9e-76| A             | G            | 0.505     | HLA-DQA1 (MHC)                           |
| EA-D MFI (log) | rs2316515                 |               | 6   | 0.09 | 0.016| 4.2e-08| A             | G            | 0.388     | IRF4                                     |
|                | rs2395192                 |               | 6   | -0.13| 0.016| 3.1e-16| C             | T            | 0.440     | HLA-DRB9 (MHC)                           |
|                | rs73067509                |               | 7   | 0.14 | 0.025| 2.2e-08| C             | G            | 0.892     | AC004538.3                               |
| Virus   | MFI (log) | rsID   | Alleles | MAF  | p Value | Effect Size | Genes   |
|---------|-----------|--------|---------|------|---------|-------------|---------|
| HHV6    | IE1A MFI  | rs13079586 | 3       | -0.10 | 0.018   | 4.6e-08     | C       |
|         |           | rs2844606  | 6       | -0.10 | 0.017   | 1.2e-08     | A       |
|         |           | rs28752523 | 6       | -0.12 | 0.021   | 7.9e-09     | C       |
| HHV7    | U14 MFI   | rs144232229 | 3       | -1.0  | 0.17    | 1.9e-09     | G       |
|         | OMP MFI   | rs3104361  | 6       | -0.18 | 0.029   | 6.5e-10     | T       |
|         | UreA MFI  | rs71569678 | 6       | -0.34 | 0.061   | 3.0e-08     | A       |
| HSV2    | Seropositivity | rs538162817 | 3       | -0.14 | 0.025   | 2.9e-08     | T       |
|         |           | rs7503464  | 17      | 0.03  | 0.0054  | 1.2e-08     | G       |
|         | mgG-2 MFI | rs144232229 | 3       | -1.0  | 0.17    | 1.9e-09     | G       |
|         |           | Chr5:166282898 | 5   | 0.73  | 0.13    | 4.6e-08     | CT      |
| JCV     | Seropositivity | rs17843569 | 6       | -0.15 | 0.010   | 1.7e-47     | C       |
|         |           | rs2432132  | 19      | -0.06 | 0.0082  | 8.8e-15     | C       |
|         | VP1 MFI   | rs1610401  | 1       | 0.35  | 0.063   | 1.9e-08     | G       |
|         |           | rs374949924 | 6       | -0.26 | 0.026   | 4.5e-24     | G       |
| MCV     | Seropositivity | rs55792153 | 5       | 0.05  | 0.0082  | 3.6e-10     | A       |
|         |           | rs9269771  | 6       | -0.12 | 0.0096  | 7.3e-37     | T       |
|         | VP1 MFI   | rs7444313  | 5       | 0.16  | 0.021   | 6.5e-15     | G       |
|         |           | rs76148407 | 6       | 0.31  | 0.052   | 1.8e-09     | C       |
|         |           | rs114708114 | 6     | 0.37  | 0.066   | 2.1e-08     | C       |
| rs75040706 | 6 | 0.30 | 0.054 | 3.4e-08 | A | G | 0.970 | RP1-86C11.7 |
| rs28393149 | 6 | -0.26 | 0.029 | 1.0e-18 | C | G | 0.882 | HLA-DRB6 (MHC) |
| rs541989586 | 3 | -0.26 | 0.047 | 3.9e-08 | A | AT | 0.934 | SOX2-OT |
| rs148929820 | 5 | -0.44 | 0.072 | 8.5e-10 | G | A | 0.975 | RP11-510I6.1 |
| rs11881343 | 19 | -0.47 | 0.086 | 3.9e-08 | A | T | 0.982 | NOTCH3 |

**T. gondii**  
**sag1 MFI (log)**

| rs1766 | 6 | -0.04 | 0.0055 | 1.1e-11 | A | G | 0.530 | HLA-DQB1 (MHC) |
| rs13197633 | 6 | -0.20 | 0.025 | 1.4e-15 | G | A | 0.880 | TOB2P1 |
| rs34073492 | 6 | -0.19 | 0.025 | 5.8e-14 | C | T | 0.877 | RP11-457M11.5 |
| rs56401801 | 6 | -0.19 | 0.025 | 6.7e-14 | T | A | 0.883 | VN1R10P |
| rs13204572 | 6 | -0.16 | 0.026 | 1.1e-09 | G | C | 0.892 | HIST1H4D |
| rs1048381 | 6 | -0.22 | 0.021 | 6.3e-25 | G | A | 0.815 | HLA-DQA1 (MHC) |

**VZV**  
**Seropositivity**  
**Case-Control**

| rs13197633 | 6 | -0.20 | 0.025 | 1.4e-15 | G | A | 0.880 | TOB2P1 |
| rs34073492 | 6 | -0.19 | 0.025 | 5.8e-14 | C | T | 0.877 | RP11-457M11.5 |
| rs56401801 | 6 | -0.19 | 0.025 | 6.7e-14 | T | A | 0.883 | VN1R10P |
| rs13204572 | 6 | -0.16 | 0.026 | 1.1e-09 | G | C | 0.892 | HIST1H4D |
| rs1048381 | 6 | -0.22 | 0.021 | 6.3e-25 | G | A | 0.815 | HLA-DQA1 (MHC) |

**Table 2:** Lead significant variants per GWAS as determined by LD clumping. For variants within the major histocompatibility complex (MHC), only the variant with lowest p-value is reported. Where applicable, variant positions are given using the GRCh37/hg19 human genome assembly.
| Agent | Analysis                     | Selected Components | Univariate (95% CI) | Effect | Univariate P-value | Multivariate (95% CI) | Effect | Multivariate P-value |
|-------|------------------------------|---------------------|---------------------|--------|--------------------|-----------------------|--------|----------------------|
| EBV   | EA-D MFI (log)               | B*08:01             | -0.16 (-0.20, -0.11) | 6.17x10^{-12} | -0.065 (-0.13, -0.0011) | 0.05 |
|       |                              | DRB1*09:01          | 0.40 (0.27, 0.54)    | 8.67x10^{-7} | 0.24 (0.081, 0.40)    | 0.003 |
|       |                              | DQB1*02:01          | -0.17 (-0.21, -0.12) | 1.38x10^{-11} | -0.092 (-0.16, -0.028) | 0.005 |
|       |                              | DQA1*03:01          | 0.11 (0.070, 0.15)   | 2.96x10^{-4} | 0.072 (0.032, 0.11)   | 0.0004 |
|       |                              | DQB1*03:03          | 0.18 (0.11, 0.25)    | 7.3x10^{-7}  | 0.10 (0.019, 0.18)    | 0.02 |
|       |                              | Sex                | -0.23 (-0.27, -0.19) | <2x10^{-16} | -0.24 (-0.29, -0.20)  | <2x10^{-16} |
|       |                              | Age                | 0.007 (0.005, 0.01)  | 1.35x10^{-4} | 0.064 (0.042, 0.086)  | 1.04x10^{-8} |
|       |                              | PCI                | 0.08 (0.06, 0.11)    | 9.61x10^{-14} | 0.078 (0.056, 0.10)   | 4.16x10^{-12} |
| EBNA-1 MFI (log) | DRB4*01:03          | -0.17 (-0.20, -0.13) | <2x10^{-16} | -0.11 (-0.15, -0.067) | 3.75x10^{-7} |
|       | DRB3*02:02          | 0.21 (0.17, 0.26)   | <2x10^{-16} | 0.16 (0.11, 0.21)    | 3.88x10^{-10} |
|       | DRB1*07:01          | -0.22 (-0.27, -0.18) | <2x10^{-16} | -0.089 (-0.45, 0.27) | 0.63 |
|       | DRB1*12:01          | 0.53 (0.40, 0.67)   | 3.76x10^{-10} | 0.30 (0.16, 0.43)    | 1.57x10^{-5} |
|       | DRB1*15:01          | 0.26 (0.22, 0.31)   | <2x10^{-16} | 0.14 (0.061, 0.22)   | 0.0005 |
|       | DQB1*02:01          | -0.24 (-0.28, -0.19) | <2x10^{-16} | -0.25 (-0.30, -0.20) | <2x10^{-16} |
|       | DQA1*01:02          | 0.23 (0.19, 0.27)   | <2x10^{-16} | 0.058 (-0.016, 0.013) | 0.12 |
|       | DQA1*02:01          | -0.22 (-0.27, -0.18) | <2x10^{-16} | -0.091 (-0.45, 0.27) | 0.62 |
|       | DPM1*03:01          | -0.20 (-0.25, -0.14) | 1.99x10^{-14} | -0.19 (-0.25, -0.14) | 5.49x10^{-14} |
|       | DPB1*04:02          | 0.16 (0.11, 0.21)   | 4.96x10^{-10} | 0.11 (0.058, 0.16)   | 2.78x10^{-3} |
| ZEBRA MFI (log) | DRB4*01:01          | 0.28 (0.23, 0.34)   | <2x10^{-16} | 0.074 (-0.00029, 0.15) | 0.05 |
|       | DRB1*03:01          | -0.20 (-0.24, -0.16) | <2x10^{-16} | -0.081 (-0.12, -0.038) | 0.0002 |
|       | DRB1*04:04          | 0.45 (0.37, 0.53)   | <2x10^{-16} | 0.36 (0.27, 0.46)    | 3.19x10^{-24} |
|       | DQB1*03:02          | 0.24 (0.19, 0.28)   | <2x10^{-16} | 0.045 (-0.029, 0.12) | 0.23 |
|       | DQB1*04:02          | 0.36 (0.26, 0.46)   | 1.18x10^{-11} | 0.45 (0.35, 0.55)    | <2x10^{-16} |
|       | DQA1*02:01          | 0.27 (0.23, 0.31)   | <2x10^{-16} | 0.29 (0.23, 0.35)    | <2x10^{-16} |
|       | DQA1*03:01          | 0.17 (0.13, 0.20)   | <2x10^{-16} | 0.13 (0.075, 0.18)   | 1.29x10^{-6} |
|       | Sex                | -0.27 (-0.31, -0.23) | <2x10^{-16} | -0.29 (-0.33, -0.25) | <2x10^{-16} |
|       | Age                | 0.051 (0.029, 0.072) | 4.17x10^{-6} | 0.059 (0.039, 0.080) | 2.43x10^{-8} |
| JCV   | Seropositivity Case-Control* | DRB5*01:01         | 0.55 (0.50, 0.60)   | <2x10^{-16} | 1.02 (0.51, 1.55)    | 0.95 |
|       | DRB1*15:01         | 0.55 (0.50, 0.59)   | <2x10^{-16} | 0.63 (0.31, 1.26)    | 0.19 |
|       | DQB1*06:02         | 0.55 (0.21, 0.60)   | <2x10^{-16} | 0.94 (0.64, 1.39)    | 0.75 |
|       | DQA1*01:02         | 0.61 (0.57, 0.66)   | <2x10^{-16} | 0.89 (0.78, 1.02)    | 0.11 |
|       | Sex                | 1.25 (1.15, 1.36)   | 1.87x10^{-7} | 1.26 (1.16, 1.37)    | 8.06x10^{-8} |
| VPI MFI (log) | DRB5*01:01         | -0.29 (-0.35, -0.23) | <2x10^{-16} | -0.29 (-0.35, -0.23) | <2x10^{-16} |
| MCV   | VPI MFI (log)       | A*29:02             | -0.41 (-0.50, -0.31) | <2x10^{-16} | -0.41 (-0.50, -0.32) | <2x10^{-16} |
Table 3: HLA alleles association analysis summary. As explained in the text, the untransformed antibody MFI analyses are not shown. Only pathogens with Lasso selected alleles are reported. Age variable is on the standardized scale. Only variables selected by Lasso are shown. PC1: 1st principal component (effect is on the standardized scale). Multivariate effects stand for the effects observed when all alleles selected by Lasso (for each respective infection) are included in the same regression analysis. *Effects for the JCV seropositivity case-control analysis are reported as odds ratios.

| VZV | Glycoproteins E and I MFI (log) | DRB1*04:04 | 0.39 (-0.49, -0.29) | 1.9x10⁻¹⁴ | -0.41 (-0.51, -0.31) | 7.01x10⁻¹⁶ |
|-----|--------------------------------|------------|----------------------|------------|----------------------|------------|
|     |                                | DRB1*15:01 | -0.19 (-0.24, -0.13) | 1.67x10⁻¹¹ | -0.12 (-0.36, 0.11)  | 0.30       |
|     |                                | DQB1*05:01 | 0.21 (0.16, 0.27)    | 1.53x10⁻¹⁵ | 0.068 (-0.055, 0.19) | 0.28       |
|     |                                | DQB1*06:02 | -0.18 (-0.24, -0.13) | 2.62x10⁻¹¹ | -0.059 (-0.29, 0.18) | 0.62       |
|     |                                | DQA1*01:01 | 0.21 (0.16, 0.26)    | <2x10⁻¹⁶   | 0.10 (-0.016, 0.22)  | 0.09       |
| Sex |                                | A*01:01    | 0.17 (0.13, 0.20)    | <2x10⁻¹⁶   | 0.089 (0.039, 0.14)  | 0.0004     |
|     |                                | B*08:01    | 0.20 (0.16, 0.25)    | <2x10⁻¹⁶   | 0.040 (-0.033, 0.11) | 0.28       |
|     |                                | DQB1*02:01 | 0.21 (0.17, 0.25)    | <2x10⁻¹⁶   | 0.14 (0.080, 0.20)   | 8.67x10⁻⁶  |
|     |                                | Sex        | 0.21 (0.16, 0.25)    | <2x10⁻¹⁶   | 0.21 (0.17, 0.26)    | <2x10⁻¹⁶   |
Figure 1: QQ-Plot and Manhattan plots from selected GWAS from the seropositivity case-control and antibody log transformed MFI analyses. Each dot on a Manhattan plot (right) represents the p-value (y-axis, on logarithmic scale) associated with the association test at a genetic variant. Values above the dashed line are considered genome-wide significant. QQ-plots (left) show the observed p-values (y-axis) against the expected p-values (x-axis). Any deviance from the red line suggests that the effect seen is not explained only by chance alone. As can be seen, the MHC is a commonly identified locus. A: EBV EA-D MFI (log). B: EBV EBNA-1 MFI (log). C: EBV ZEBRA MFI (log). D: JCV seropositivity case-control.