After the Infection: A Survey of Pathogens and Non-communicable Human Disease

Michael Lape[1,2,3], Daniel Schnell[2], Sreeja Parameswaran[3,6], Kevin Ernst[3], Nathan Salomonis[1,2,4], Lisa J. Martin[4,6], Brett M. Harnett[1], Leah C. Kottyan[3,4,6*], Matthew T. Weirauch[2,3,4,5,6*]

1. Department of Biomedical Informatics, University of Cincinnati College of Medicine, Cincinnati, Ohio, USA.
2. Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA.
3. Center for Autoimmune Genomics and Etiology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA.
4. Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, Ohio, USA.
5. Division of Developmental Biology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA.
6. Division of Human Genetics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA.

*Co-correspondence: Leah.Kottyan@cchmc.org and Matthew.Weirauch@cchmc.org

Summary

There are many well-established relationships between pathogens and human disease, but far fewer when focusing on non-communicable diseases (NCDs). We leverage data from The UK Biobank and TriNetX to perform a systematic survey across 20 pathogens and 426 diseases, focused primarily on NCDs. To this end, we assess the association between disease status and infection history proxies. We identify 206 pathogen-disease pairs that replicate in both cohorts. We replicate many established relationships, including *Helicobacter pylori* with several gastroenterological diseases, and connections between Epstein-Barr virus with multiple sclerosis and lupus. Overall, our approach identified evidence of association for 15 of the pathogens and 96 distinct diseases, including a currently controversial link between human cytomegalovirus (CMV) and ulcerative colitis (UC). We validate this connection through two orthogonal analyses, revealing increased CMV gene expression in UC patients and enrichment for UC genetic risk signal near human genes that have altered expression upon CMV infection. Collectively, these results form a foundation for future investigations into mechanistic roles played by pathogens in disease.

Keywords: human disease, pathogens, viruses, epidemiology, UK Biobank, TriNetX, Electronic Health Records, biostatistics

Introduction

Humans are exposed to infectious agents throughout life. Many of the communicable diseases associated with specific infectious agents are well-characterized[1]. In acute infection, virus-driven mechanisms are clearly and predominantly seen as an agent of disease. For example, respiratory syncytial virus (RSV) causes an upper-respiratory disease that can be life-threatening in young infants and causes cold-like symptoms in adolescents and adults[2]. Likewise, varicella zoster virus (VZV) causes varicella (chickenpox) upon primary infection. This infection typically occurs during childhood and is well tolerated. However, if primary infection occurs in infants, adults, or the immunocompromised, the viral infection is less well contained, and the virally mediated pathology can be life-threatening. Even after primary infection, VZV lies dormant for decades, reactivating in a portion of adults with the lytic virus directly leading to zoster (shingles)[3].

The role of infectious agents in non-communicable diseases (NCDs) is much less well-explored, although several associations are well-known. For example, *Helicobacter pylori* infection is the strongest risk factor for gastric cancers[4]. Likewise, multiple studies have demonstrated an important role for both hepatitis B (HBV) and C viruses (HCV) in the development of cirrhosis and other chronic liver diseases[5,6]. Further, certain strains of human papillomavirus (HPV) are known to cause a large proportion of cervical cancers and a smaller proportion of several other cancers[7]. Each of these pathogens, as well as several others, have been classified as biological carcinogens by the International Agency for Research on Cancer (IARC)[8]. Indeed, it is estimated that up to

15% of all new cancer diagnoses are attributable to these and other infectious agents[9]. More recently, strong epidemiologic and mechanistic links have been identified between Epstein-Barr virus (EBV) infection and multiple sclerosis (MS)[10–12], following decades of suggestive epidemiological and molecular evidence[13,14]. It is likely that many unknown pathogen-disease connections remain to be discovered.

NCDs are often chronic diseases and have complex etiologies. Unlike acute viral infections, the virus itself might not contribute to the full etiology of virally associated NCDs. While many individuals might be exposed and mount an immune response to a virus, virally associated NCDs often occur in a small subset of individuals in the context of specific genetic factors and other environmental exposures[15].

Historically, connections between pathogens and diseases have been made one pair at a time. However, the recent establishment of large-scale national biobanks and the general shift to electronic health records (EHRs) enables the systematic analysis of many pathogen-disease pairs. Here, we leverage biobank data from The UK Biobank (UKB) and EHR data from TriNetX, LLC (TNX), resources that each contain both serologic and diagnostic records, enabling the systematic detection of associations between multiple pathogens and multiple diseases simultaneously. Using a discovery-replication approach, our analysis reveals 206 replicated pathogen-disease associations, including previously established pathogen-disease links and many novel or previously suggestive relationships. In particular, we identify strong evidence for the currently controversial connection between cytomegalovirus (CMV) infection and

ulcerative colitis (UC). Orthogonal analyses of this relationship reveal that: (1) patients with UC have elevated CMV mRNA levels in intestinal tissue samples compared to healthy controls; and (2) UC genetic risk loci are enriched near human genes that change expression upon CMV infection in two independent datasets. Collectively, these results implicate multiple pathogens in dozens of non-communicable human diseases, providing a unique and powerful resource for future studies of the mechanistic roles played by these pathogens in disease development.

Results

Pathogen-disease data collection in two large independent cohorts

We employed a discovery-replication experimental design using two newly available resources, The UK Biobank (UKB) and TriNetX (TNX). For our discovery cohort, we used the portion of UKB participants with systematically measured antibody titers (9,429 subjects). We sought sufficient statistical power to detect effect sizes observed for at least half of our “Tier 1” positive controls at a power level of 0.8, with an alpha of 0.05 (see Methods). This yielded minimum case and total sample (cases plus controls) requirements of 17 and 187, respectively. Within the set of UKB subjects with serologic data, ignoring ICD codes A00 - B99 covering “Certain infectious and parasitic diseases”, we found 408 NCDs and 18 communicable diseases (ICD10 codes between C00 and O99) for which we were powered (**Supplemental Dataset 1**). The primary focus of this study is on the 408 NCDs with a secondary assessment of communicable diseases that are associated with non-causative pathogens. We modeled the association between these 426 diseases and the titer levels of 45 different antibodies representing immune responses to 20 unique pathogens. Six infectious diseases (ICD codes between A00 - B99) were also included as part of a control set. Within the resulting 19,289 separate antibody-disease tests, we collapsed instances where there were multiple antibodies to the same pathogen by selecting the antibody with the most significant association. In total, this procedure resulted in 8,616 pathogen-disease tests. This discovery cohort (UKB) has the advantage of a carefully controlled experimental design with many covariates for detection of confounding effects, with the disadvantage of a relatively small number of subjects.

As our independent replication cohort, we used data extracted from The TriNetX Research Network. TriNetX, LLC (Cambridge, MA), is a global health research network that has integrated electronic health record data from more than 75 separate healthcare organizations, mostly from the United States. These data contain both the clinical diagnoses and the serologic test results required for our model for over 11 million individuals. A total of 209 separate binary (positive/negative) clinical laboratory tests targeting the 20 UKB pathogens were identified in the medical records provided by TNX. As with the discovery cohort, we restricted pathogen-disease tests to those for which we were statistically powered, ~75% of all 8,616 tested UKB pathogen-disease pairs. To minimize multiple testing burden, we only modeled those pairs that were significant in the discovery cohort (UKB). This replication cohort (TNX) has the disadvantage of a less well-controlled study design involving data from different clinical tests and sites, with the advantage of a very large number of subjects. The larger number of subjects enabled the use of an additional restriction requiring infection status to appear in a participant's medical record prior to disease diagnosis (see Methods).

Development of a statistical model with high sensitivity and specificity

We developed a workflow for the systematic detection and replication of pathogen-disease pairs within our discovery and replication cohorts (**Figure 1**). To this end, we used a logistic regression model to test for association between disease status and a proxy for a history of infection by a given pathogen: either continuous antibody titer values (UKB) or binary clinical laboratory test results (TNX). To mitigate potential

confounding, we adjusted each model for any of ten additional sociodemographic and health-related variables (**Supplemental Table 1**) that were found to be significantly associated with both disease status and the pathogen proxy in the discovery cohort (**Supplemental Table 2**). To prevent model overfitting, a backward elimination procedure was used to prune non-significant covariates prior to fitting the final model (see Methods).

The most commonly adjusted-for covariates were age, sex, and body mass index (BMI), included in 35.5%, 31.4%, and 24.4% of the UKB antibody-disease models, respectively (**Supplemental Figure 1**). As expected, age was most significant in models for diseases that are more common in the elderly such as “disorders of lipoprotein metabolism and other lipidaemias”, “other arthrosis”, and “other cataract”. Likewise, the strongest BMI associations were seen in models for metabolic syndrome diseases such as “obesity”, “essential primary hypertension” and “diabetes mellitus”. Roughly a quarter of models did not require adjustment for any covariates.

We confirmed the robustness of our analytical model in our UKB discovery cohort using a permutation-based approach. To this end, we calculated an empirical p-value for each antibody-disease pair by permuting disease status across individuals (see Methods). We observed exceptionally strong correlation (Pearson’s $r > 0.99$) between the nominal p-value obtained from our analytical model and the empirically derived permutation-based p-value (**Supplemental Figure 2**).

To reduce the multiple testing burden, we only tested pathogen-disease pairs in the replication cohort that were statistically significant in the discovery cohort. During replication, we refit the UKB model using the replication cohort data, adjusting for the same covariates when data were available. To reduce false negatives, a lenient per-disease Benjamini-Hochberg (BH) false discovery rate (FDR) threshold of 0.3 was applied to the discovery cohort results. Then, to reduce false positives, a more stringent per-disease FDR threshold of 0.01 was used in the replication stage. Only pairs with significant associations and effects (odds ratios) in the same direction across both cohorts were considered replicated pathogen-disease relationships.

We first assessed the sensitivity of our model on a set of positive controls. “Tier 1” positive controls were identified as infectious disease diagnoses paired with their causal pathogen, e.g., hepatitis C (HCV) paired with a diagnosis of “unspecified viral hepatitis” (ICD10: B19). As expected, our model found significant associations between the disease “unspecified viral hepatitis” and both hepatitis B (HBV) and HCV pathogens. It also identified an association with the BK virus (BKV) at the more lenient discovery cohort threshold. However, upon testing for replication, only the HBV and HCV results remained significant (**Figure 2**). To assess the specificity of the model, we used a set of “expected negatives”, which we identified as the complement of the Tier 1 positive controls, i.e., an infectious disease diagnoses with a pathogen that does not cause the disease, such as “unspecified viral hepatitis” with Epstein-Barr virus (EBV). We excluded HIV from the expected negative set due to its indirect involvement in many immune-mediated diseases.

Overall, our model identified significant association for all eight (100%) of the Tier 1 positive control pathogen-disease pairs in the UKB cohort, and all eight of these replicated in TNX. Conversely, the model identified only five (5.68%) of the expected negatives as significant in UKB, only one of which replicated in TNX, a diagnosis of "infectious mononucleosis" (ICD10: B27) and human herpesvirus 6 (HHV-6). Upon further investigation, this replicated "expected negative" pair represents a previously established relationship: HHV-6 infection can account for up to five percent of infectious mononucleosis (IM)-like syndrome diagnoses in adult patients[16]. Considering only the fully replicated pairs as predicted positives, and those pairs that were either identified as not significant in UKB or failed replication as predicted negatives, our model has a sensitivity of 1.0, a specificity of 0.80, and a precision of 0.89.

We next assessed a set of 83 pathogen-disease pairs with suggestive literature evidence ("Tier 2" positive controls) collected via a semi-automated literature search approach (see Methods). Sixteen (19.28%) of these Tier 2 pairs were significantly associated in UKB, and 11 of the 15 with available data (68.75%) replicated in TNX. Included in these are well-known associations such as *H. pylori* with several gastroenterological diseases such as "duodenal ulcer", "peptic ulcer, site unspecified", and "gastritis and duodenitis"[17–19]. We also validated connections between particular pathogens and hepatic diseases, such as "fibrosis and cirrhosis of liver" with HCV and "other diseases of liver" with both HBV and HCV[20]. Finally, our model replicated the well-established associations of EBV with multiple sclerosis (MS)[10,12,21] and with

systemic lupus erythematosus (SLE)[22,23]. Taken together, these results establish that our approach can capture both well-established and suggestive pathogen-disease relationships while maintaining a strong degree of specificity.

Identification of 206 replicated pathogen-disease relationships

Encouraged by the performance of our model on our positive and negative control sets, we next sought to identify novel pathogen-disease relationships. In total, of the 8,437 "unknown" (non-Tier 1 or 2) pathogen-disease tests that met our requirements in UKB, 569 were significant. 462 of these pairs had sufficient data in the TNX cohort to test for replication, and 195 of these were replicated (2.3% of the total 8,437 "unknown" pairs that were initially tested) (**Figure 3**). The 195 replicated pairs represent a diverse collection of diseases and pathogens, with 15 of the 20 tested pathogens connected to at least one of 96 distinct diseases, 89 of which are NCDs (**Figure 4** and **Supplemental Dataset 2**). Altogether, the 11 "Tier 2" and the 195 "unknown" pathogen-disease relationships equate to 206 replicated associations. While these relationships are correlations, due to the size of the TNX cohort we were able to restrict our study populations to only those who had a positive pathogenic test result prior to disease development. Thus these "temporal correlations" provide much stronger evidence for the pathogen playing a role in disease development.

Outside of our Tier 1 positive controls, the largest odds ratio (OR) obtained in the UKB is for systemic lupus erythematosus (SLE) with EBV (OR = 3.98), which also has a large TNX odds ratio of 4.96. We also replicated the strongly established connection between

EBV and multiple sclerosis (OR = 2.55 and 4.45). Overall, HCV and HBV have the most replicated associations (25) of any pathogens tested (**Supplemental Figure 3**). CMV has the most associations with an OR indicating risk, (21 of the 24 with OR > 1), whereas *Chlamydia trachomatis* has the largest number of protective relationships (15 of the 20 with OR < 1). “Other diseases of the urinary system” (ICD: N39) had the most replicated associations across all pathogens (10), all of which are predicted to increase the risk of disease. The 3-character ICD10 code N39 includes both the communicable diagnosis urinary tract infections (N39.0) as well as several non-communicable forms of incontinence (N39.3, N39.4).

To characterize the replicated pathogen-disease pairs more broadly, we examined our results at the ICD chapter level, which largely represent distinct body systems. *H. pylori* has the most replicated associations in a particular chapter: chapter 11 (K00 - K95), which contains diseases of the digestive system (**Supplemental Figure 4**). In addition to the three Tier 2 positive controls discussed above, *H. pylori* is predicted to increase the risk of four additional digestive system diseases and to decrease the risk of “irritable bowel syndrome”, “diaphragmatic hernia”, “other diseases of esophagus”, and “gastro-esophageal reflux disease” (GERD). The relationship between *H. pylori* and GERD continues to be debated, with one recent meta-analysis reporting that the eradication of *H. pylori* increases the risk of GERD, thus indicating a possible protective effect[24] and another recent systematic review rating the “evidence grade” for the association between *H. pylori* and GERD as low[24,25]. Outside of chapter 11, *H. pylori* also has a risk relationship with “iron-deficiency anemia”, an association that has been published

previously[26], and a protective relationship with asthma, which has also been previously reported[27,28]. EBV, HBV, and HCV all have replicated associations in over 10 different ICD chapters, reflecting the often-systemic effects of infections by these pathogens.

Orthogonal validations of the cytomegalovirus – ulcerative colitis relationship

We next sought orthogonal evidence supporting the 206 replicated associations identified by our approach. Virus-disease relationships are often reflected by higher expression levels of virus-encoded genes in patients compared to controls[29–33]. Likewise, many viruses manipulate host gene expression patterns[34,35], and the molecular processes of most complex diseases are now appreciated to be impacted by alterations to human gene expression levels[36,37]. We thus hypothesized that causative pathogen-disease relationships would be reflected in publicly available gene expression data. To test this hypothesis, we performed two complementary analyses. First, we examined viral gene expression levels in patients compared to controls. Second, we asked if genome-wide association study (GWAS) risk loci were enriched near human genes with altered expression levels following viral infection.

As a positive control, we first examined the well-established link between EBV and SLE. To this end, we identified six publicly available SLE case/control RNA-seq data sets performed in blood and B cell subsets (**Supplemental Dataset 3**). Collectively, these data contain 378 SLE cases and 74 control subjects. We used the VIRTUS software package[38] to identify and quantify viral read counts in these data. As expected, this analysis revealed significantly higher EBV transcript levels in SLE cases compared to

controls (p-value = 0.0049) (**Figure 5A**). We next considered ulcerative colitis (UC), a disease with suggestive, but still unclear roles for viral infection[39]. Our main analyses implicated both HIV and CMV in UC disease processes, with the replication cohort showing infection occurred prior to disease diagnosis. HIV has recently been linked to UC[40]. For CMV, a possible role is much less clear, perhaps due to the historical difficulty of differentiating between CMV colitis and inflammatory bowel diseases such as UC[39]. Although, the hypothesis that CMV may be causal of UC remains contentious[41], the ability of CMV to cause UC flare-ups is still heavily debated[42]. We thus examined RNA-seq data for a set of 669 UC cases and 59 controls obtained from seven studies using intestinal biopsies (**Supplemental Dataset 3**). Similar to EBV and SLE, we observe significantly higher levels of CMV transcripts relative to controls in these samples (p-value = 0.022) (**Figure 5B**), providing additional evidence for a role of CMV in UC disease processes.

As a second orthogonal analysis, we asked if GWAS disease risk loci are enriched near human genes with virus-induced expression level changes. To this end, we used our RELI algorithm[11] to relate GWAS risk loci and public RNA-seq experiments examining infected and uninfected cells. In brief, this procedure uses a permutation-based method to estimate the significance of the overlap between the genomic coordinates of GWAS genetic risk loci and 200 kilobase windows around the transcription start site for genes with virus-altered gene expression levels (see Methods).

As a positive control, we first examined EBV and SLE. As expected, our analyses revealed significant overlap between SLE risk loci and genes that are differentially expressed upon EBV infection, in two independent studies performed in B lymphocytes and Peripheral Blood Mononuclear Cells (**Figure 5C**, purple bars) (**Supplemental Dataset 4**). In contrast, when considering genes that did not change significantly upon infection (**Figure 5C**, grey bars), SLE risk loci are not enriched. These results are consistent with our previous observation that the genomic binding events of the EBNA2 regulatory protein, encoded by EBV, coincide with approximately half of all SLE risk loci[11]. Encouraged by these results, we next compared CMV-altered genes to UC genetic risk loci. Similar to the EBV-SLE results, we again observe highly significant overlap between CMV-altered genes and UC risk loci, and insignificant overlap for expressed but unchanged genes, in two different cell types (monocytes and dendritic cells) (**Figure 5D**, purple and grey bars, respectively) (**Supplemental Dataset 4**).

Collectively, these analyses provide compelling orthogonal evidence that the CMV-UC connection identified in both of our independent cohorts might represent a causative relationship.

Discussion

In this study, we sought a wider understanding of the role played by pathogens in what are traditionally considered non-communicable diseases. To this end, we developed a logistic regression model and applied the model to two large, independent biobank resources. Our model showed strong discriminatory performance on a set of positive and negative controls and replicated many additional well-documented pathogen-disease associations. We report evidence for over 200 new or previously tenuous pathogen-disease connections, including a role for CMV in UC, which was supported by two orthogonal genomics-based validations, and provide the corresponding data on a freely accessible and easily browsable web server.

A recent study took a similar approach to ours, with a specific focus on six neurodegenerative diseases[43]. Herein we report results from an analysis across the disease spectrum. In addition to the comparatively limited scope of the Levine *et al.* study, there are several key differences between the 2023 Levine *et al.* study and ours. The Levine *et al.* study used hospital diagnosis codes as a pathogen proxy, some of which link to multiple pathogens, such as viral encephalitis. In comparison, we were able to pinpoint specific pathogens due to our use of serology data. Further, by restricting to inpatient hospital databases, the Levine *et al.* study focused on patients with infections that were sufficiently severe to require hospitalization. In contrast, our analyses included systematically measured titers for select UKB participants[44], along with data from TNX, which pulls all available clinical laboratory test results for each patient, encompassing standard preventative screenings, outpatient diagnostic

workups, as well as panels ordered during hospital stays. This is likely one reason why the odds ratios we report are much more modest than those reported by Levine *et al.* Thus, although both studies are of great utility, the results of the two studies are not directly comparable.

Roughly 40% (81/206) of the replicated associations in our study have odds ratios of less than one, indicating a potentially protective pathogen-disease relationship. For example, while high-risk strains of HPV are known to cause over 70% of cervical cancer, our results also suggest that HPV-16 and -18 can be “protective” for diseases such as “seborrheic dermatitis” and “other dermatitis”. Indeed, viral infections that increase risk of one phenotype or disease can reduce risk for others[45–48]. More generally, the role of viral infection in shaping the human immune system and subsequent immune responses has been studied extensively. It is well appreciated that viral infection can rewire the chromatin of immune cells and shape subsequent responses of a person towards additional inflammatory insults[49,50]. For example, a viral response that results in an interferon-based immune response could be protective in the context of diseases driven by T cell helper-2 type inflammation[51]. Vaccination against a particular pathogen will protect against infection and thus the disease risks associated with that pathogen. However, further studies will be required to investigate if vaccination will confer the protective effects that we report in this study.

Our study was enabled by the release of the large UK Biobank and TriNetX datasets. The associations identified in our study depend upon a sufficient number of subjects

that have both accompanying diagnostic and serology data. As additional, larger datasets are released, it will be critical to validate the findings of this study and to use the additional statistical power to examine NCDs for which we were not powered. It will also be important to identify potentially causal etiological mechanisms driving these pathogen-disease associations, as recently exemplified by the EBV-MS field[11,12,49,52,53].

In summary, we present the largest systematic assessment to date of pathogens in the context of non-communicable human disease. Using complementary discovery and replication datasets, we identified 206 replicated pathogen-disease relationships, including additional orthogonal evidence strongly supporting a relationship between CMV infection and ulcerative colitis. We anticipate that this rich data resource will form the foundation for future characterization of the many currently unknown pathogen-disease relationships.

Acknowledgments

We thank Ivan Marazzi, Chris Benner, Brad Rosenberg, Emily Miraldi, Juan Fuxman-Bass, and Artem Babaian for insightful comments on the manuscript.

This research has been conducted using the UK Biobank Resource under Application Number 47377. We thank Matthew Skowronek, Eric Young, Jeff Warnick, and the rest of the engineering team at TriNetX for additional assistance working with the TriNetX data. TriNetX, LLC is compliant with the Health Insurance Portability and Accountability Act (HIPAA), the US federal law which protects the privacy and security of healthcare data, and any additional data privacy regulations applicable to the contributing HCO. We thank The University of Cincinnati Center for Clinical and Translational Science and Training (CCTST) for assistance with accessing TriNetX data. CCTST is supported by the National Center for Advancing Translational Sciences of the National Institutes of Health (NIH), under Award Number 2UL1TR001425-05A1. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

This research was funded by National Institutes of Health (NIH) R01 DK107502, R01 AI148276, U01 HG011172, U19 AI070235, and P30 AR070549 to L.C.K.; R01 HG010730, R01 GM055479, and U01 AI130830, to M.T.W.; R01 AR073228, R01 NS099068, and R01 AI024717 to M.T.W. and L.C.K.; R01 CA226802 to N.S Funding was also provided by CCHMC ARC Award 53632 to M.T.W. and L.C.K.

Author Contributions

Conceptualization: M.L., L.C.K., M.T.W.; Study design: M.L., L.C.K., M.T.W.; Statistical analysis design: M.L., D.S., L.M., M.T.W.; Computational analysis: M.L., S.P.; Writing: M.L., L.C.K., M.T.W.; Review and approve final manuscript: M.L., D.S., S.P., N.S., B.H., L.M., L.C.K., M.T.W.; Funding: N.S., L.C.K., M.T.W.

Figures

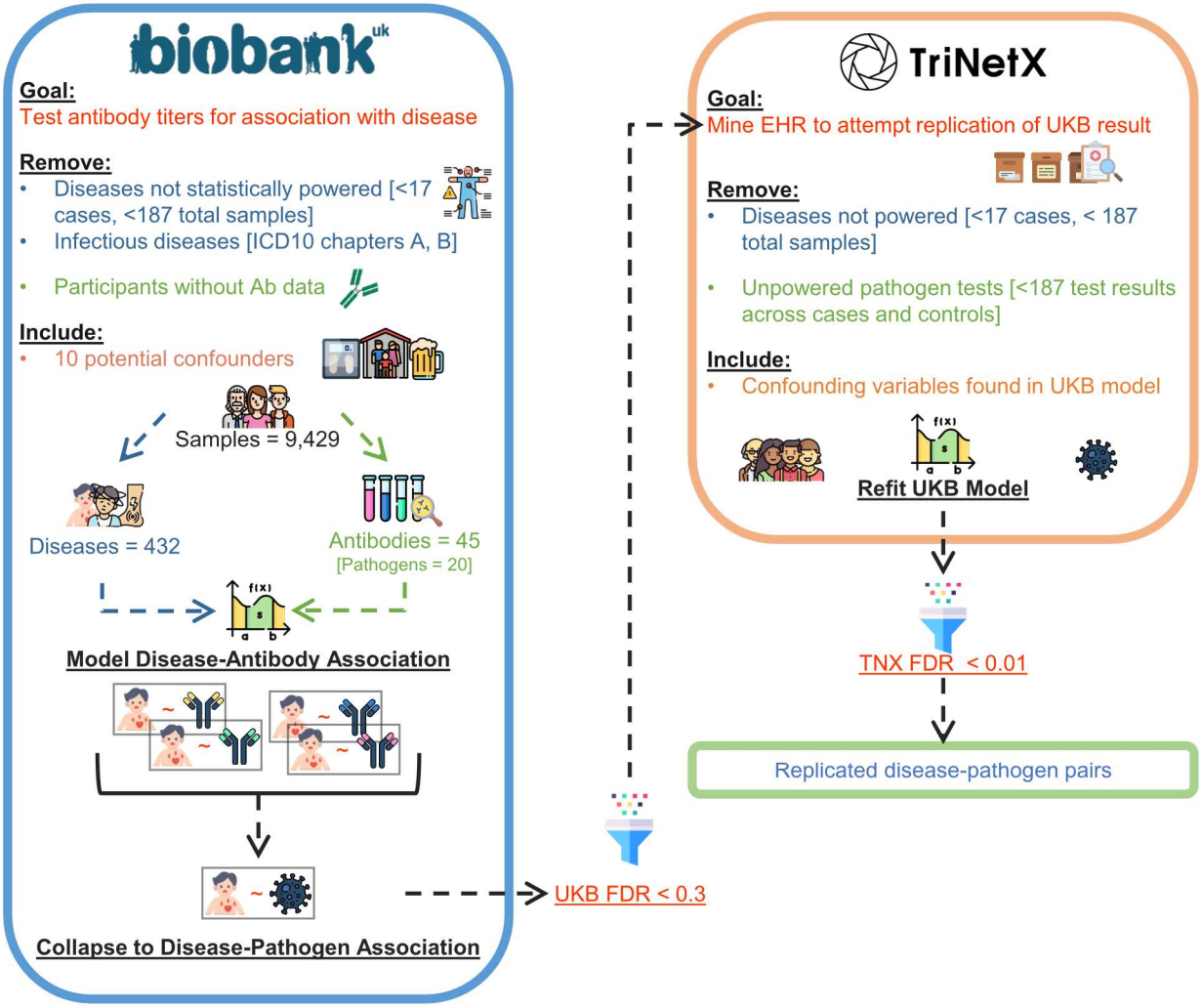

Figure 1 | Overview of study design.

Cartoon overview of the study design. The UK Biobank (UKB) was used as the discovery cohort (left). 426 diseases with sufficient sample counts in UKB (along with positive and negative controls) were tested for association with the 45 antibody titers representing immune responses to 20 unique pathogens across 9,429 UKB participants. Models were adjusted for any of ten additional health-related and sociodemographic variables that were determined to be confounding for a particular antibody-disease pair. Significant disease-pathogen pairs identified in UKB were tested

in the independent TriNetX (TNX) cohort (right). Pathogen-disease pairs that were significant in both the discovery and the replication cohort were considered replicated pairs.

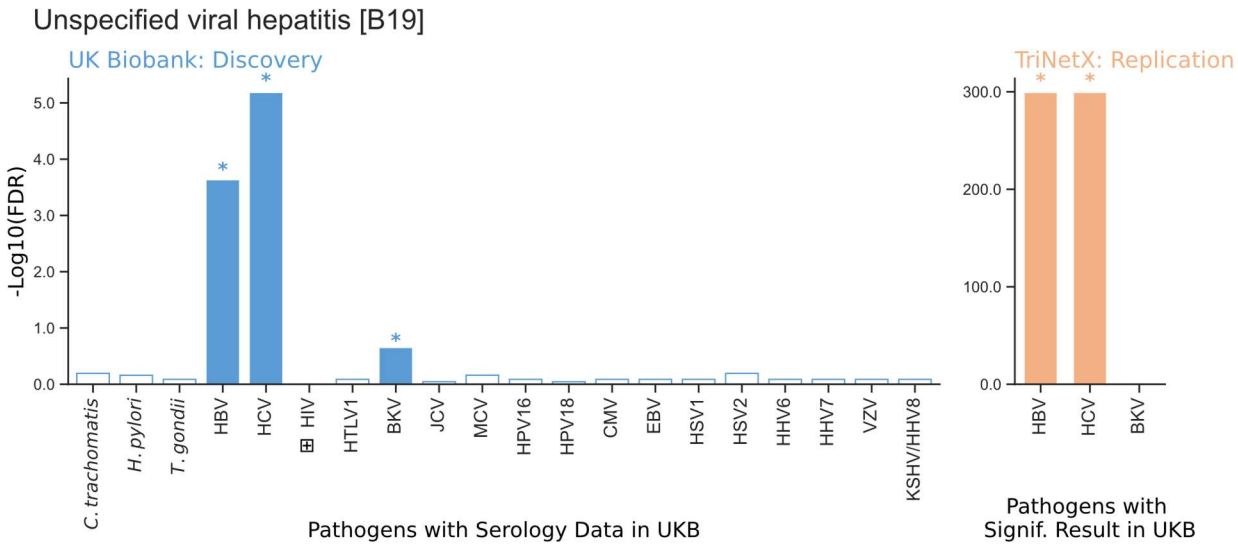

Figure 2 | Illustration of the two-step discovery-replication filtering process. Bar charts depicting an example of replicated pathogen-disease pairs. The left plot (blue) displays the significance of association (-Log10 p-value with a per-disease Benjamini-Hochberg false discovery rate (FDR) correction) of each pathogen with the disease of interest, here one of the Tier 1 positive control diseases, “unspecified viral hepatitis”. Significant associations are depicted as filled bars with colored asterisks above. Hepatitis B (HBV), hepatitis C (HCV), and BK Virus (BKV) are all significantly associated with “unspecified viral hepatitis” in the UKB cohort. The right plot (orange) shows the results of testing only the significant UKB results in the replication cohort, TriNetX. At the more stringent replication threshold, only HBV and HCV remain as significant, replicated pairs.

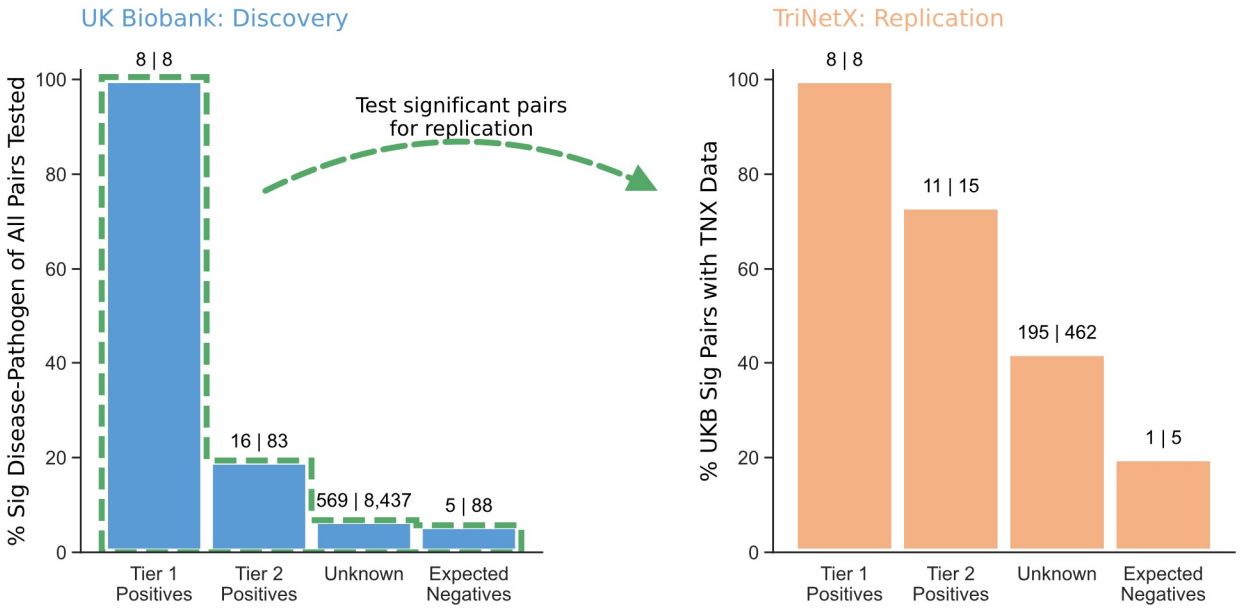

Figure 3 | Summary of pathogen-disease pairs identified across both cohorts.

Bar charts show the percent of pathogen-disease pairs in each subgroup ("Tier 1 Positives", "Tier 2 Positives", and "Unknown") that were significant in the discovery cohort (blue bars, left) and replication cohort (orange bars, right). The numbers used to calculate the percentages are indicated above each bar. All discovery cohort (left) significant pairs were assessed for replication in the independent replication cohort (right). Note that the total number of pairs tested for replication may not match the number of significant UKB pairs due to insufficient data available in the replication cohort for some pairs. Such cases are not considered replication successes or failures.

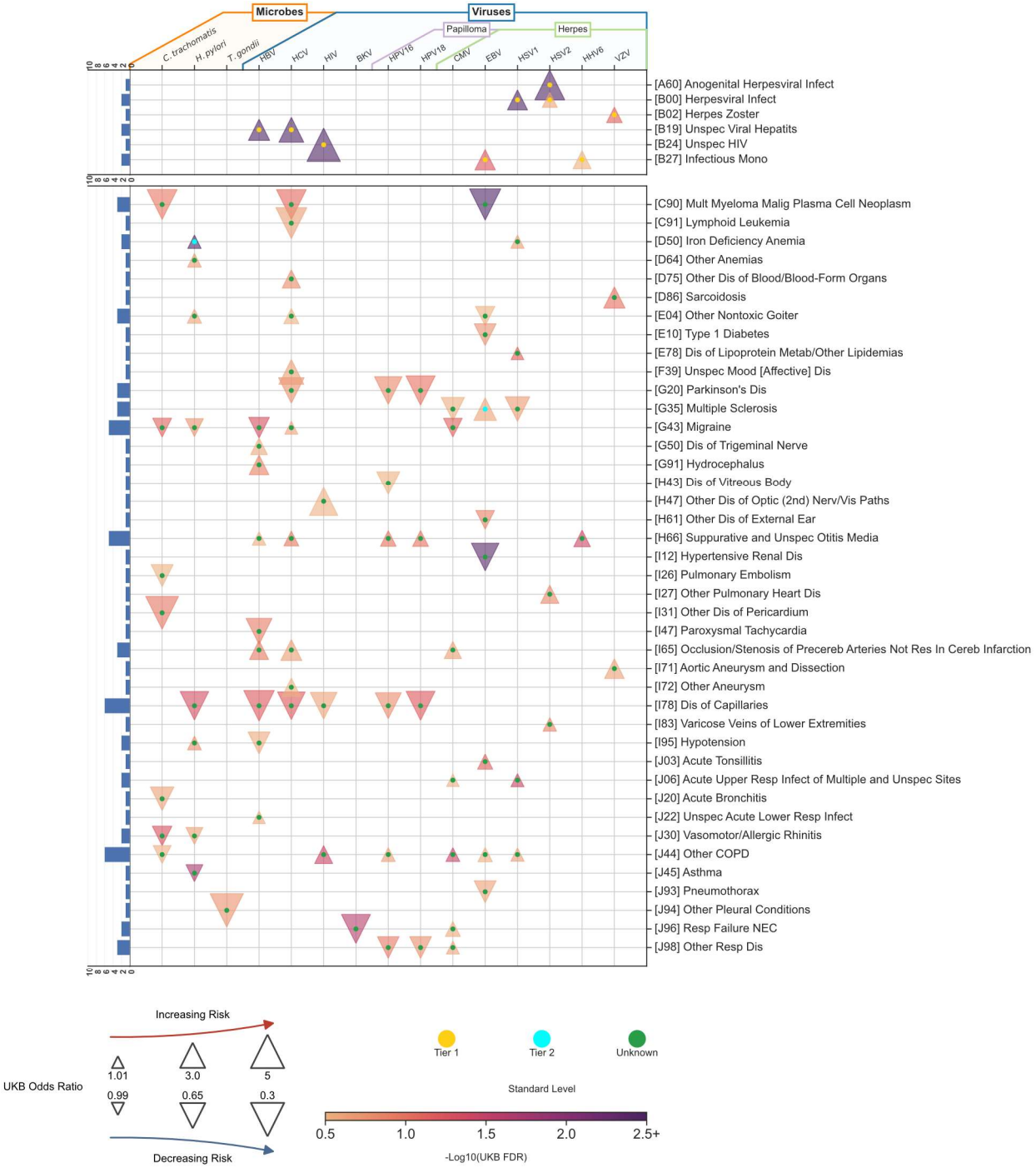

Figure 4A | Overview of all replicated pathogen-disease pairs (Tier 1 controls, C00 - J99)

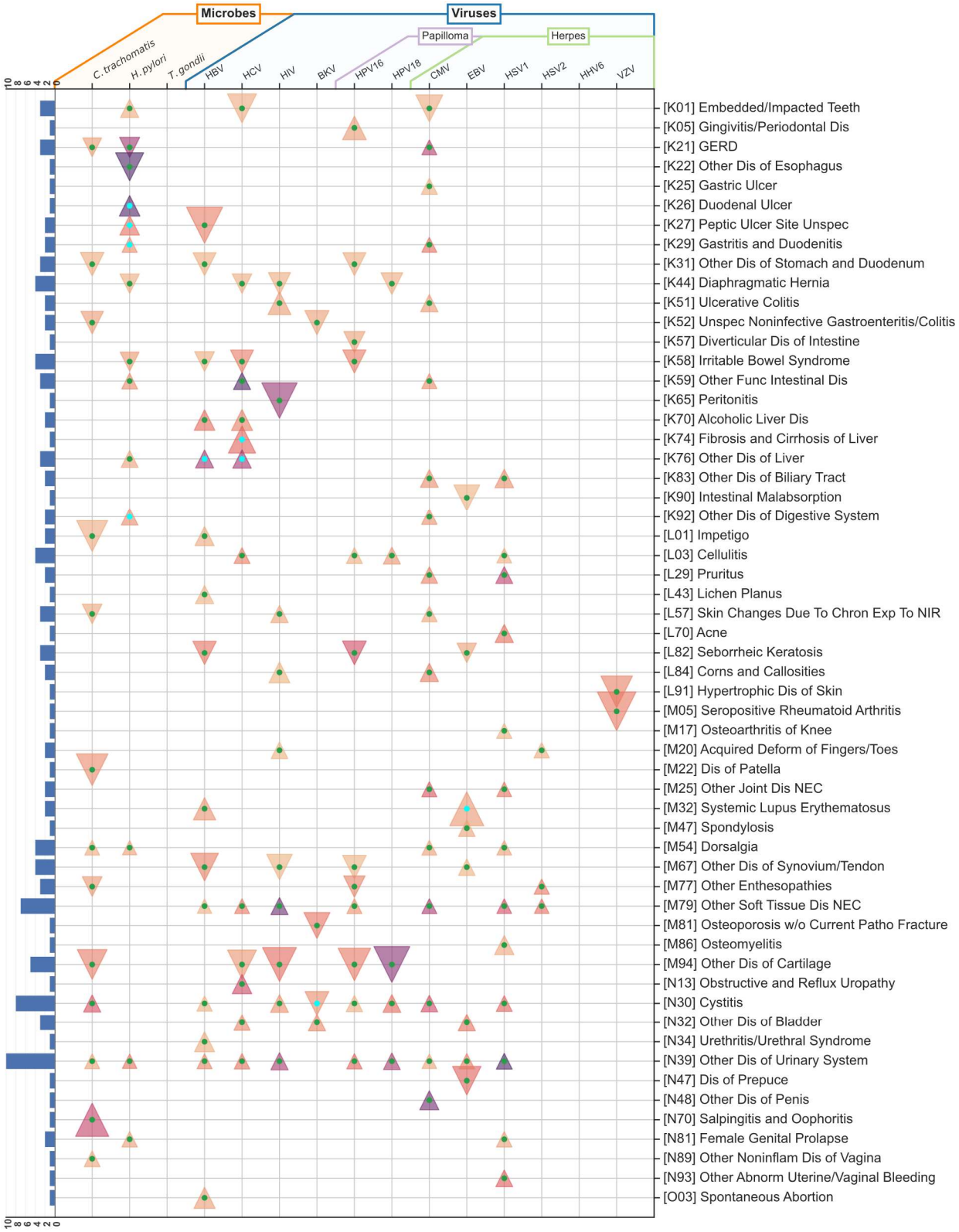

Figure 4B | Overview of all replicated pathogen-disease pairs (K00 - O99)

Figure 4 | Overview of all replicated pathogen-disease pairs.

Heatmap containing all pathogens and diseases with at least one replicated result. Pathogens are first grouped by type (microbe or virus), and then sub-grouped by family if more than one member is present. Diseases are ordered by ICD10 code, with the Tier 1 positive controls at the top. A histogram opposite to each disease indicates the total number of replicated associations between that disease and all pathogens investigated. Each replicated result for a pathogen-disease pair is represented by a triangle either pointing up (indicating an odds ratio greater than one) or down (indicating an odds ratio less than one). The size of each triangle represents the discovery cohort (UK Biobank or UKB) odds ratio and is capped at a maximum of 5 to account for the very large effect size between “unspecified hiv disease” and HIV (OR: 38.2). The color of each triangle represents the UKB negative log base-10 corrected p-value, capped at a maximum of 2.5 to enable better visual distinction in the region covering all but the eight most significant pathogen-disease pairs. Each triangle is marked with a central dot, the color of which indicates whether the pair is a Tier 1 (gold) or Tier 2 (cyan) positive control, or an unknown relationship (green).

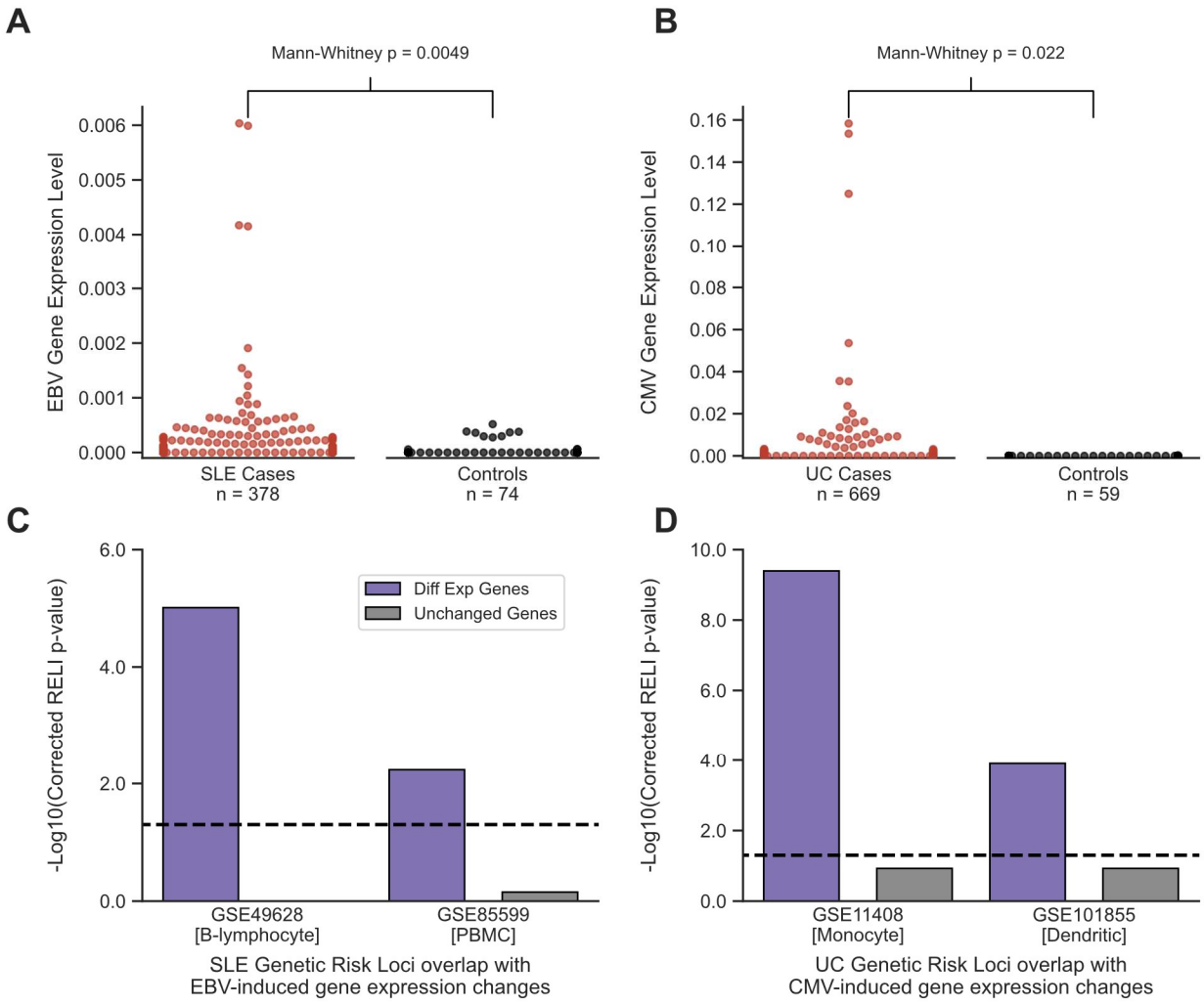

Figure 5 | Orthogonal validation of the EBV/SLE and CMV/UC associations.

Orthogonal validation of the EBV/SLE (positive control) and CMV/UC (new prediction) associations using virus gene expression levels (top) and enrichment of disease risk loci near virus-induced differentially expressed human genes (bottom). **A** and **B**. Swarm plots showing viral read counts normalized by the size of the viral genome and the total number of human mapped reads in that sample, providing the final 'Normalized Hit Rate' provided by the VIRTUS software package. Normalized hit rates were compared between cases and controls using a Mann-Whitney test with the p-values annotated on the plots: SLE vs controls (panel **A**) and UC vs controls (panel **B**). **C** and **D**. Bar plots

indicating the enrichment of SLE (left) or UC (right) genome-wide association study (GWAS) loci proximal to genes with altered expression (purple bars) or unaltered expression (grey bars) after infection by the viruses EBV and CMV, respectively. Gene Expression Omnibus (GEO) ID and cell type are provided below each plot. The black dashed line indicates statistical significance (p = 0.05)

Methods

UK BioBank cohort

The UK Biobank (UKB) is a prospective cohort study containing medical, sociodemographic, and genetic data for nearly 500,000 adults from across the United Kingdom[54]. In this study, diagnoses for all participants were extracted from two UKB fields, 'First Occurrences' [UKB Category 1712], which is a synthetic field generated by UKB analysts that collates diagnoses from primary care, hospital inpatient, death registry, and self-reported records, and 'Type of Cancer – ICD10' [UKB Field 40006], which contains cancer diagnoses pulled from national cancer registries for linked participants. The first occurrences data are limited to 3-character ICD10 codes, e.g., M32.9 is recorded as M32, so the cancer registry diagnoses were truncated to match. Analyses were limited to those diseases that had at least 17 cases and 187 total samples within the cohort (see power calculation below).

Data for 45 antibody titer levels representing immune responses to 20 unique pathogens [UKB Category 1307] were downloaded and Log10 transformed prior to analysis. Antibody titer measurements were performed using a multiplex serology approach based on the enzyme-linked immunosorbent assay (ELISA) concept as described by Mentzer *et al.*[44]. Similar to Mentzer *et al.,* a series of 10 additional health-related and sociodemographic variables that could potentially be associated with both disease status and antibody titers were collected to test for confounding during analyses. The continuous covariates age and BMI were scaled by a factor of 10, while all other covariates were either already categorical or were discretized (**Supplemental**

Table 1). The scikit-learn IterativeImputer method, which is based on the Multivariate Imputation by Chained Equations (MICE) algorithm, was used to impute missing covariate values (thirty-two participants were missing BMI values and eight were missing Townsend deprivation index values).

TriNetX cohort

TriNetX, LLC (TNX), is a private organization that has built a global medical research network that enables healthcare organizations to make their electronic health record (EHR) data more easily accessible to researchers in a de-identified manner enabling Real World Data analyses. To prevent possible outlier values due to results being reported in different units or encoding errors, continuous laboratory test results were excluded, thereby limiting our analysis to only binary tests where the results were either positive or negative. A query built by combining a complete list of Logical Observation Identifiers Names and Codes (LOINC) codes that corresponded to binary clinical laboratory tests for our pathogens of interest was run on 02-14-2023 across 73 healthcare organizations on the TNX Research Network, resulting in a list of just over 11 million unique patients. Diagnoses for all participants were collected and then cohorts were automatically generated for each pathogen-disease pair. The use of GNU Parallel[55] made the processing of the immense amount of TNX data much more tractable. Those with a result for a particular LOINC code but without the diagnosis of interest were considered controls, and only those with a test result (positive or negative) prior to the earliest diagnosis report in their medical record were considered cases. We removed those with the diagnosis appearing in the EHR before the laboratory test, as

the temporal relationship of infection prior to disease diagnosis could not be firmly established for them. We attempted to extract all potential confounding variables that were considered in the UKB analysis from the TNX data; however, data for only three of the ten were available. For situations in which a covariate that was included in the UKB model but was not available in the TNX data, such covariates were dropped from the TNX model prior to refitting. After finding few ICD10 B24 cases in TNX (a diagnosis used in UKB to indicate HIV infection, “Unspecified human immunodeficiency virus [HIV] disease”) it was determined that in the United States (the primary source of TNX data) the ICD10 code B20 is primarily used to indicate HIV infection. Thus, the TNX results for B20 were merged with the UKB B24 results, however, this is the only diagnosis for which this was done.

Establishment of the minimum number of cases and total samples for analysis

To ensure that the comparisons made had sufficient statistical power to identify associations if present, and minimize the multiple testing correction burden, we sought to determine what a minimum number of cases and controls would be. While our analytical strategy used a logistic regression model, to determine a preliminary power estimate we opted to consider the power to detect differences in antibody levels between those with and without disease for known antibody-disease pairs. This approach enabled us to calculate the effect sizes (differences between groups divided by standard deviation) and evaluate the power required to detect such differences. While there the logistic regression models (especially with covariates) are likely to have different power estimates, our goal was to come up with a pragmatic threshold for the

minimum number of cases which would have reasonable likelihood of success. We looked at effect sizes across a set of 14 known positive antibody-disease pairs in the UKB data, which collapsed to a set of eight pathogen-disease pairs. These pairs constitute our “Tier 1” positive controls and represent those infectious diseases that have been shown to be directly caused by a pathogen, such as “herpes zoster (HZ)” (ICD10: B02) diagnosis and varicella-zoster virus, the virus that causes HZ. Effect sizes at the antibody-disease level ranged from 0.1 to 3.9. After collapsing to the largest effect per pathogen-disease, effect sizes ranged from 0.2 to 3.9, with a median effect size of 0.72. Using the G*Power 3.1 software package[56], we calculated the number of cases required for 80% power with an alpha of 0.05, assuming 10 controls per case, yielding a minimum case number of 17.

We recognize that more controls may be available, but higher numbers of controls impact power only slightly (data not shown). Further, using the median effect size of known positives might underestimate power for some pathogen-disease associations; this conservative choice was made to balance the inclusion of diseases while ensuring sufficient statistical power.

Model development and application

We modeled the association between a given disease and a given pathogen using a logistic regression model with disease status as the binary outcome and the pathogen proxy value (continuous for UKB titers, categorical for TNX positive/negative) as the predictor (**Formula 1**). Since each cohort uses a different type of predictor, we note that

the odds ratio (OR) has a slightly different meaning between the two. In the UKB cohort, the OR represents the increase in odds of developing a disease per 10-fold increase in antibody titer level; in the TNX cohort, the OR represents the increase in odds of developing a particular disease in those subjects after testing positive for a given pathogen. All sex-specific diseases had controls limited to those participants at risk, e.g., cervical cancer (ICD10: C53) used only females without a cervical cancer diagnosis as controls. Further, for all pregnancy and childbirth-related diseases, only patients with a record of a healthy birth (ICD10 codes O80 - O84) were used as controls (patients with records of both a healthy birth and the diagnosis of interest were removed from the analysis).

Formula box

$$\mathrm{P(Dis\ status:\ Case)} = \frac{e^{\beta_0+\beta_1 X_1+\ \ldots}}{1+\ e^{\beta_0+\beta_1 X_1+\ \ldots}}$$

Formula 1: The logistic regression model employed in this study, examining the relationship between one disease and one pathogen proxy. The probability of having the disease of interest is modeled as a function of the pathogen proxy X_1, where X_1 is either a continuous titer value (UKB) or a binary pathogen test result (positive/negative) (TNX). A coefficient (β_n) for each variable is estimated during model fitting, with β_0 representing the intercept or bias, and β_1 the coefficient for the pathogen proxy and thus the pathogen itself. The model is adjusted for additional confounding covariates, as appropriate, by including additional terms, $\beta_2 X_2$, etc.

After adjusting our logistic regression model for all covariates found to be significantly associated with both disease status and pathogen proxy in separate univariate statistical tests (**Supplemental Table 2**), we ran a backward elimination procedure (stepAIC, MASS R library) to remove any non-significant covariates, to mitigate possible

overfitting. We attempted to replicate models that showed statistical significance in the UKB cohort using the TNX cohort, by simply refitting the same model.

In the rare situation where a categorical TNX pathogen test had five or fewer patients in one of the cells of the pathogen-disease contingency table we altered our model slightly. For example, for the diagnosis of “herpesviral [herpes simplex] infections” (ICD10: B00) and the pathogen test with LOINC code 93439-8, there is only one disease control with a positive test result for herpes simplex virus 1, the cause of this disease. This represents a very strong association that would be lost with a normal logistic regression model. Thus, in these situations we instead employed Firth’s bias-reduced logistic regression method[57] (R library logistf), a method which is appropriate in such situations.

To verify the robustness of the UKB model results, a permutation procedure was performed. Briefly, 10,000 permutations were run for each antibody-disease pair (45 total pairs per disease), where disease status was randomly shuffled amongst the participants while keeping the number of cases and controls as well as the model itself, i.e., covariates adjusted for, constant. Next, all permutation results for a particular disease were pooled into a larger per-disease null distribution now containing 450,000 permutation results. Empirical p-values for each antibody-disease pair were calculated by comparing the nominal p-value from our analytical model to the respective per-disease null distribution.

To control for multiple testing, we applied a per-disease Benjamini-Hochberg (BH) false discovery rate (FDR)[58] at the pathogen-disease level. Since we were using a discovery-replication model, we used a lenient FDR threshold of 0.3 for our discovery cohort to minimize possible false negatives, whereas upon replication a much more stringent FDR threshold of 0.01 was applied to minimize false positives.

Model assessment

To assess the model's performance, we calculated associations for a set of positive and negative control pairs. We used the Tier 1 controls as described above for the positive controls. Six infectious diseases are included in the Tier 1 controls, two of which can be caused by two different pathogens ("herpesviral [herpes simplex] infections" by herpes simplex virus 1 or 2 and "unspecified viral hepatitis" by HBV or HCV). The negative control set, termed "expected negatives", represents the complement of the Tier 1 controls, e.g., an HZ diagnosis paired with HBV instead of VZV. As an additional assessment, a second set of positive controls using only NCDs ("Tier 2") was collected using a literature mining approach. In brief, the log product frequency method[59] was employed to rank pathogen-disease pairs by the number of PubMed co-citations (disease and pathogen), normalized by the number of citations of each separately. Searches of PubMed were conducted using the R library easyPubMed v2.13 using default settings. The top 200 ranked results were manually reviewed, yielding a total of 83 "Tier 2" positive controls.

Orthogonal validation of CMV-UC association

We attempted to validate two of our replicated findings using two distinct 'omics-based methods.

First, to capture differences in viral gene expression levels between cases and controls for the diseases of interest, we used publicly available RNA-seq data from case/control cohorts and the bioinformatics tool VIRTUS v1.2.1[38] using default parameters except reducing the default "hit_cutoff" from 400 to 0 since we were only investigating a small number of pathogens. Briefly, VIRTUS is a pipeline that takes an input RNA-seq FASTQ file and aligns the reads to a reference human genome (hg38; GCF_000001405.39) using STAR[60]. It then attempts to align any unaligned reads to a second precompiled index file containing a user-specified set of viral genomes, here EBV (NC_007605.1) and CMV (NC_006273.2). The resulting number of mapped reads for each virus was first normalized by the pathogen's genome length, then normalized by the number of mapped human reads, providing the flexibility to compare results across both different pathogens as well as different samples and experiments. Finally, the non-parametric Mann-Whitney U test was used to compare the normalized read counts for a particular pathogen between cases and controls.

Second, to examine if GWAS loci for the diseases of interest were enriched near human genes that have altered expression levels upon infection by the pathogens of interest more so than unchanged genes, we used our RELI tool[11]. RELI estimates the statistical significance of overlap between a set of input genomic loci and a peak file, typically GWAS loci and ChIP-seq peaks. It does this through a permutation-based procedure by

first counting the number of overlaps between the input loci and peaks, then permuting the input loci around the genome and again counting the number of overlaps with the peaks, building a null distribution. Finally, to calculate significance, RELI compares the total number of overlaps seen in the input loci set to the null distribution.

We obtained GWAS data for ulcerative colitis (UC) and systemic lupus erythematosus (SLE) from the NHGRI-EBI GWAS catalog (v1.0.2-associations_e96_r2019-05-03)[61]. A genome-wide significance cutoff of $5x10^{-8}$ was used, and we considered only data from European populations due to the prevalence of GWAS data for this ancestry group. For each phenotype, independent loci were identified using LD-based pruning with PLINK[62] (window size 300,000 kb, SNP shift size 100,000 kb, and $r^2 < 0.2$). These independent loci were then expanded to incorporate variants in strong linkage disequilibrium (LD) ($r^2 > 0.8$) again using PLINK. We next downloaded lists of genes with expression changes in response to CMV or separately EBV infection from the VExD database (https://vexd.cchmc.org)[63] (**Supplemental Dataset 5)**. Differentially expressed genes (DEGs) are identified in VExD as those genes with an adjusted p-value < 0.05 and absolute fold change > 2, when comparing infected and uninfected cells of the same type within a single study. As a null model, for each study we also identified sets of genes that do not significantly change upon infection (adjusted p >= 0.05, and fold change < 1.2) and randomly selected the same number of genes to match the corresponding differentially expressed gene set. We then ran RELI using the GWAS risk loci for each disease as input against the genomic regions defined by a 200kb window centered on the transcription start stie of each input gene.

Supplemental Datasets

Supplemental Dataset 1 | Overview of diseases outside of ICD10 chapter 1 with an infectious component.

All diseases were tested at the 3-character ICD10 code level. This document contains a review of whether those diseases outside of ICD10 chapter 1 ("infectious diseases") are either infectious at the 3-character level, at a sub-code level, or not infectious at all. For example, J11 "influenza, virus not identified" is infectious at the 3-character level, whereas N39 "other disorders of urinary system" is not infectious at the 3-character level but does include the sub-code N39.0 "Urinary tract infection, site not specified", which is infectious. For several diseases it was not clear whether they were infectious at the 3-character level, indicated by question marks. These unclear diseases were counted as NCDs when summing the total number of NCDs, and communicable diseases examined in this analysis.

Supplemental Dataset 2 | Full pathogen-disease results across both UKB and TNX.

This document contains the results from our main analysis with association test results for all 8,616 pathogen-disease pair tests.

Supplemental Dataset 3 | NCBI GEO RNA-seq datasets used for VIRTUS analyses.

This document contains information on which NCBI GEO data sets were used for our VIRTUS analyses. The first sheet, 'SLE', contains information about the six lupus case/control RNA-seq data sets performed in blood and B cell subsets that were used

for the EBV-lupus VIRTUS analysis. The second sheet, 'UC', shows the same information for the seven ulcerative colitis case/control RNA-seq data sets examined in the CMV-UC VIRTUS analysis.

Supplemental Dataset 4 | RELI Results for EBV-SLE and CMV-UC.

This document shows the RELI results for both the set of differentially expressed genes upon either EBV or CMV infection as well as those for the random sampling of unaltered genes upon infection for all datasets tested. The sheet "EBV_SLE" has the results for the EBV-infected gene sets and risk loci for lupus. The sheet "CMV_UC" provides the RELI results for gene sets from CMV-infected cells and GWAS risk loci for ulcerative colitis.

Supplemental Dataset 5 | GWAS risk loci and RNA-seq data sets used for RELI analyses.

This document contains both the RNA-seq data sets ("Gene Sets" sheet) and the GWAS risk loci downloaded from The GWAS Catalog ("GWAS loci" sheet) that were used as inputs for our RELI analyses.

Code Availability

All code used in this project is available from the Weirauch Research Lab's GitHub page, https://github.com/WeirauchLab/pathogen_ncd

Data Availability Statement

The data used for the main analysis from The UK Biobank can be accessed via an application for Tier 2 UKB data. Further, the data from TriNetX (https://live.trinetx.com, accessed on 14 February 2023), can be requested directly from TriNetX. In both cases costs may be incurred, and a material transfer agreement or data sharing agreement is required.

References

1. Bennett, J.E., Dolin, R., and Blaser, M.J. eds. (2020). Mandell, Douglas, and Bennett's principles and practice of infectious diseases Ninth edition. (Elsevier).

2. Borchers, A.T., Chang, C., Gershwin, M.E., and Gershwin, L.J. (2013). Respiratory Syncytial Virus—A Comprehensive Review. Clin. Rev. Allergy Immunol. *45*, 331–379. 10.1007/s12016-013-8368-9.

3. Gershon, A.A., Breuer, J., Cohen, J.I., Cohrs, R.J., Gershon, M.D., Gilden, D., Grose, C., Hambleton, S., Kennedy, P.G.E., Oxman, M.N., et al. (2015). Varicella zoster virus infection. Nat. Rev. Dis. Primer *1*, 15016. 10.1038/nrdp.2015.16.

4. Hardbower, D.M., Peek, R.M., and Wilson, K.T. (2014). At the Bench: Helicobacter pylori, dysregulated host responses, DNA damage, and gastric cancer. J. Leukoc. Biol. *96*, 201–212. 10.1189/jlb.4BT0214-099R.

5. Lauer, G.M., and Walker, B.D. (2001). Hepatitis C Virus Infection. N. Engl. J. Med. *345*, 41–52. 10.1056/NEJM200107053450107.

6. Trépo, C., Chan, H.L.Y., and Lok, A. (2014). Hepatitis B virus infection. The Lancet *384*, 2053–2063. 10.1016/S0140-6736(14)60220-8.

7. zur Hausen, H. (2009). Papillomaviruses in the causation of human cancers — a brief historical account. Virology *384*, 260–265. 10.1016/j.virol.2008.11.046.

8. Bouvard, V., Baan, R., Straif, K., Grosse, Y., Secretan, B., Ghissassi, F.E., Benbrahim-Tallaa, L., Guha, N., Freeman, C., Galichet, L., et al. (2009). A review of human carcinogens—Part B: biological agents. Lancet Oncol. *10*, 321–322. 10.1016/S1470-2045(09)70096-8.

9. Plummer, M., de Martel, C., Vignat, J., Ferlay, J., Bray, F., and Franceschi, S. (2016). Global burden of cancers attributable to infections in 2012: a synthetic analysis. Lancet Glob. Health *4*, e609-616. 10.1016/S2214-109X(16)30143-7.

10. Bjornevik, K., Cortese, M., Healy, B.C., Kuhle, J., Mina, M.J., Leng, Y., Elledge, S.J., Niebuhr, D.W., Scher, A.I., Munger, K.L., et al. (2022). Longitudinal analysis reveals high prevalence of Epstein-Barr virus associated with multiple sclerosis. Science *375*, 296–301. 10.1126/science.abj8222.

11. Harley, J.B., Chen, X., Pujato, M., Miller, D., Maddox, A., Forney, C., Magnusen, A.F., Lynch, A., Chetal, K., Yukawa, M., et al. (2018). Transcription factors operate across disease loci, with EBNA2 implicated in autoimmunity. Nat. Genet. *50*, 699–707. 10.1038/s41588-018-0102-3.

12. Lanz, T.V., Brewer, R.C., Ho, P.P., Moon, J.-S., Jude, K.M., Fernandez, D., Fernandes, R.A., Gomez, A.M., Nadj, G.-S., Bartley, C.M., et al. (2022). Clonally

expanded B cells in multiple sclerosis bind EBV EBNA1 and GlialCAM. Nature *603*, 321–327. 10.1038/s41586-022-04432-7.

13. Thacker, E.L., Mirzaei, F., and Ascherio, A. (2006). Infectious mononucleosis and risk for multiple sclerosis: A meta-analysis. Ann. Neurol. *59*, 499–503. 10.1002/ana.20820.

14. Belbasis, L., Bellou, V., Evangelou, E., Ioannidis, J.P.A., and Tzoulaki, I. (2015). Environmental risk factors and multiple sclerosis: an umbrella review of systematic reviews and meta-analyses. Lancet Neurol. *14*, 263–273. 10.1016/S1474-4422(14)70267-4.

15. Virolainen, S.J., VonHandorf, A., Viel, K.C.M.F., Weirauch, M.T., and Kottyan, L.C. (2023). Gene-environment interactions and their impact on human health. Genes Immun. *24*, 1–11. 10.1038/s41435-022-00192-6.

16. Naito, T., Kudo, N., Inui, A., Matsumoto, N., Takeda, N., Isonuma, H., Dambara, T., and Hayashida, Y. (2006). Causes of Infectious Mononucleosis-like Syndrome in Adult Patients. Intern. Med. *45*, 833–834. 10.2169/internalmedicine.45.1725.

17. Ahmed, S., and Belayneh, Y.M. (2019). Helicobacter pylori And Duodenal Ulcer: Systematic Review Of Controversies In Causation. Clin. Exp. Gastroenterol. *12*, 441–447. 10.2147/CEG.S228203.

18. Suerbaum, S., and Michetti, P. (2002). Helicobacter pylori Infection. N. Engl. J. Med. *347*, 1175–1186. 10.1056/NEJMra020542.

19. Kandulski, A., Selgrad, M., and Malfertheiner, P. (2008). Helicobacter pylori infection: A clinical overview. Dig. Liver Dis. *40*, 619–626. 10.1016/j.dld.2008.02.026.

20. Smith, A., Baumgartner, K., and Bositis, C. (2019). Cirrhosis: Diagnosis and Management. Am. Fam. Physician *100*, 759–770.

21. Bar-Or, A., Pender, M.P., Khanna, R., Steinman, L., Hartung, H.-P., Maniar, T., Croze, E., Aftab, B.T., Giovannoni, G., and Joshi, M.A. (2020). Epstein–Barr Virus in Multiple Sclerosis: Theory and Emerging Immunotherapies. Trends Mol. Med. *26*, 296–310. 10.1016/j.molmed.2019.11.003.

22. James, J.A., Kaufman, K.M., Farris, A.D., Taylor-Albert, E., Lehman, T.J., and Harley, J.B. (1997). An increased prevalence of Epstein-Barr virus infection in young patients suggests a possible etiology for systemic lupus erythematosus. J. Clin. Invest. *100*, 3019–3026.

23. Li, Z.-X., Zeng, S., Wu, H.-X., and Zhou, Y. (2019). The risk of systemic lupus erythematosus associated with Epstein–Barr virus infection: a systematic review and meta-analysis. Clin. Exp. Med. *19*, 23–36. 10.1007/s10238-018-0535-0.

24. Mou, W.-L., Feng, M.-Y., and Hu, L.-H. (2020). Eradication of Helicobacter Pylori Infections and GERD: A systematic review and meta-analysis. Turk. J. Gastroenterol. *31*, 853–859. 10.5152/tjg.2020.19699.

25. Cheng, Y., Kou, F., Liu, J., Dai, Y., Li, X., and Li, J. (2021). Systematic assessment of environmental factors for gastroesophageal reflux disease: An umbrella review of systematic reviews and meta-analyses. Dig. Liver Dis. *53*, 566–573. 10.1016/j.dld.2020.11.022.

26. DuBois, S., and Kearney, D.J. (2005). Iron-Deficiency Anemia and Helicobacter pylori Infection: A Review of the Evidence. Off. J. Am. Coll. Gastroenterol. ACG *100*, 453.

27. Amedei, A., Codolo, G., Del Prete, G., de Bernard, M., and D'Elios, M.M. (2010). The effect of Helicobacter pylori on asthma and allergy. J. Asthma Allergy *3*, 139–147. 10.2147/JAA.S8971.

28. Zuo, Z.T., Ma, Y., Sun, Y., Bai, C.Q., Ling, C.H., and Yuan, F.L. (2021). The Protective Effects of Helicobacter pylori Infection on Allergic Asthma. Int. Arch. Allergy Immunol. *182*, 53–64. 10.1159/000508330.

29. Gross, A.J., Hochberg, D., Rand, W.M., and Thorley-Lawson, D.A. (2005). EBV and systemic lupus erythematosus: a new perspective. J. Immunol. Baltim. Md 1950 *174*, 6599–6607. 10.4049/jimmunol.174.11.6599.

30. Poole, B.D., Templeton, A.K., Guthridge, J.M., Brown, E.J., Harley, J.B., and James, J.A. (2009). Aberrant Epstein-Barr viral infection in systemic lupus erythematosus. Autoimmun. Rev. *8*, 337–342. 10.1016/j.autrev.2008.12.008.

31. Piroozmand, A., Haddad Kashani, H., and Zamani, B. (2017). Correlation between Epstein-Barr Virus Infection and Disease Activity of Systemic Lupus Erythematosus: a Cross-Sectional Study. Asian Pac. J. Cancer Prev. APJCP *18*, 523–527. 10.22034/APJCP.2017.18.2.523.

32. Moon, U.Y., Park, S.J., Oh, S.T., Kim, W.-U., Park, S.-H., Lee, S.-H., Cho, C.-S., Kim, H.-Y., Lee, W.-K., and Lee, S.K. (2004). Patients with systemic lupus erythematosus have abnormally elevated Epstein-Barr virus load in blood. Arthritis Res. Ther. *6*, R295-302. 10.1186/ar1181.

33. Agostini, S., Mancuso, R., Guerini, F.R., D'Alfonso, S., Agliardi, C., Hernis, A., Zanzottera, M., Barizzone, N., Leone, M.A., Caputo, D., et al. (2018). HLA alleles modulate EBV viral load in multiple sclerosis. J. Transl. Med. *16*, 80. 10.1186/s12967-018-1450-6.

34. Paschos, K., and Allday, M.J. (2010). Epigenetic reprogramming of host genes in viral and microbial pathogenesis. Trends Microbiol. *18*, 439–447. 10.1016/j.tim.2010.07.003.

35. Liu, X., Hong, T., Parameswaran, S., Ernst, K., Marazzi, I., Weirauch, M.T., and Fuxman Bass, J.I. (2020). Human Virus Transcriptional Regulators. Cell *182*, 24–37. 10.1016/j.cell.2020.06.023.

36. Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. Science *337*, 1190–1195. 10.1126/science.1222794.

37. Oh, E.S., and Petronis, A. (2021). Origins of human disease: the chrono-epigenetic perspective. Nat. Rev. Genet. *22*, 533–546. 10.1038/s41576-021-00348-6.

38. Yasumizu, Y., Hara, A., Sakaguchi, S., and Ohkura, N. (2021). VIRTUS: a pipeline for comprehensive virus analysis from conventional RNA-seq data. Bioinformatics *37*, 1465–1467. 10.1093/bioinformatics/btaa859.

39. Mourad, F.H., Hashash, J.G., Kariyawasam, V.C., and Leong, R.W. (2020). Ulcerative Colitis and Cytomegalovirus Infection: From A to Z. J. Crohns Colitis *14*, 1162–1171. 10.1093/ecco-jcc/jjaa036.

40. Elmahdi, R., Kochhar, G.S., Iversen, A.T., Allin, K.H., Dulai, P.S., Desai, A., and Jess, T. (2022). Development of Inflammatory Bowel Disease in HIV Patients: A Danish Cohort Study (1983–2018) With American Validation (1999–2018). Gastro Hep Adv. *1*, 1114–1121. 10.1016/j.gastha.2022.08.003.

41. Beswick, L., Ye, B., and van Langenberg, D.R. (2016). Toward an Algorithm for the Diagnosis and Management of CMV in Patients with Colitis. Inflamm. Bowel Dis. *22*, 2966–2976. 10.1097/MIB.0000000000000958.

42. Jentzer, A., Veyrard, P., Roblin, X., Saint-Sardos, P., Rochereau, N., Paul, S., Bourlet, T., Pozzetto, B., and Pillet, S. (2020). Cytomegalovirus and Inflammatory Bowel Diseases (IBD) with a Special Focus on the Link with Ulcerative Colitis (UC). Microorganisms *8*, 1078. 10.3390/microorganisms8071078.

43. Levine, K.S., Leonard, H.L., Blauwendraat, C., Iwaki, H., Johnson, N., Bandres-Ciga, S., Ferrucci, L., Faghri, F., Singleton, A.B., and Nalls, M.A. (2023). Virus exposure and neurodegenerative disease risk across national biobanks. Neuron *0*. 10.1016/j.neuron.2022.12.029.

44. Mentzer, A.J., Brenner, N., Allen, N., Littlejohns, T.J., Chong, A.Y., Cortes, A., Almond, R., Hill, M., Sheard, S., McVean, G., et al. (2022). Identification of host–pathogen-disease relationships using a scalable multiplex serology platform in UK Biobank. Nat. Commun. *13*, 1818. 10.1038/s41467-022-29307-3.

45. Bach, J.-F. (2002). The Effect of Infections on Susceptibility to Autoimmune and Allergic Diseases. N. Engl. J. Med. *347*, 911–920. 10.1056/NEJMra020100.

46. Cooper, P.J. (2009). Interactions between helminth parasites and allergy. Curr. Opin. Allergy Clin. Immunol. *9*, 29–37. 10.1097/ACI.0b013e32831f44a6.

47. Garn, H., and Renz, H. (2007). Epidemiological and immunological evidence for the hygiene hypothesis. Immunobiology *212*, 441–452. 10.1016/j.imbio.2007.03.006.

48. Kilpeläinen, M., Terho, E.O., Helenius, H., and Koskenvuo, M. (2000). Farm environment in childhood prevents the development of allergies. Clin. Exp. Allergy J. Br. Soc. Allergy Clin. Immunol. *30*, 201–208. 10.1046/j.1365-2222.2000.00800.x.

49. Hong, T., Parameswaran, S., Donmez, O.A., Miller, D., Forney, C., Lape, M., Saint Just Ribeiro, M., Liang, J., Edsall, L.E., Magnusen, A.F., et al. (2021). Epstein-Barr virus nuclear antigen 2 extensively rewires the human chromatin landscape at autoimmune risk loci. Genome Res. *31*, 2185–2198. 10.1101/gr.264705.120.

50. Wang, R., Lee, J.-H., Kim, J., Xiong, F., Hasani, L.A., Shi, Y., Simpson, E.N., Zhu, X., Chen, Y.-T., Shivshankar, P., et al. (2023). SARS-CoV-2 restructures host chromatin architecture. Nat. Microbiol. *8*, 679–694. 10.1038/s41564-023-01344-8.

51. Okada, H., Kuhn, C., Feillet, H., and Bach, J.-F. (2010). The “hygiene hypothesis” for autoimmune and allergic diseases: an update. Clin. Exp. Immunol. *160*, 1–9. 10.1111/j.1365-2249.2010.04139.x.

52. Thomas, O.G., Bronge, M., Tengvall, K., Akpinar, B., Nilsson, O.B., Holmgren, E., Hessa, T., Gafvelin, G., Khademi, M., Alfredsson, L., et al. (2023). Cross-reactive EBNA1 immunity targets alpha-crystallin B and is associated with multiple sclerosis. Sci. Adv. *9*, eadg3032. 10.1126/sciadv.adg3032.

53. Keane, J.T., Afrasiabi, A., Schibeci, S.D., Swaminathan, S., Parnell, G.P., and Booth, D.R. (2021). The interaction of Epstein-Barr virus encoded transcription factor EBNA2 with multiple sclerosis risk loci is dependent on the risk genotype. EBioMedicine *71*, 103572. 10.1016/j.ebiom.2021.103572.

54. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O’Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature *562*, 203. 10.1038/s41586-018-0579-z.

55. Tange, O. (2022). GNU Parallel 20220122 ('20 years’). 10.5281/zenodo.5893336.

56. Faul, F., Erdfelder, E., Buchner, A., and Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. Behav. Res. Methods *41*, 1149–1160. 10.3758/BRM.41.4.1149.

57. Firth, D. (1993). Bias reduction of maximum likelihood estimates. Biometrika *80*, 27–38. 10.1093/biomet/80.1.27.

58. Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. Ser. B Methodol. *57*, 289–300. 10.1111/j.2517-6161.1995.tb02031.x.

59. Hu, Y., Hines, L.M., Weng, H., Zuo, D., Rivera, M., Richardson, A., and LaBaer, J. (2003). Analysis of Genomic and Proteomic Data Using Advanced Literature Mining. J. Proteome Res. *2*, 405–412. 10.1021/pr0340227.

60. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinforma. Oxf. Engl. *29*, 15–21. 10.1093/bioinformatics/bts635.

61. Buniello, A., MacArthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., et al. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. *47*, D1005–D1012. 10.1093/nar/gky1120.

62. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., et al. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am. J. Hum. Genet. *81*, 559–575.

63. Dexheimer, P.J., Pujato, M., Roskin, K., and Weirauch, M.T. (2022). VExD: A curated resource for human gene expression alterations following viral infection. 2021.12.02.470974. 10.1101/2021.12.02.470974.