Supplementary Figures

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

$$nLPF = -\log_{10}\left(\frac{\text{Disease AND Pathogen nCite}}{\text{Disease nCite}} \times \frac{\text{Disease AND Pathogen nCite}}{\text{Pathogen nCite}}\right) \quad \textbf{(b)}$$

Figure S1 | Equations for logistic regression and negated log product frequency utilized.

The logistic regression model employed in this study **(a)** examines 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 antibody titer value in the case of UK Biobank data or a binary pathogen test result (positive/negative) for data obtained from TriNetX. A coefficient (β_n) is estimated for each variable during model fitting, with β_0 representing the intercept or bias, and β_1 representing 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. The negated form of the log product frequency (LPF) method (**b**) was used to rank order pathogen-disease pairs automatically to enrich the top 175 pathogen-disease pairs that were manually reviewed. This enrichment would lead to a larger set of high-evidence pairs that could be used as the "Tier 2" positive control set. The term "nCite" refers to the number of citations found when searching PubMed.

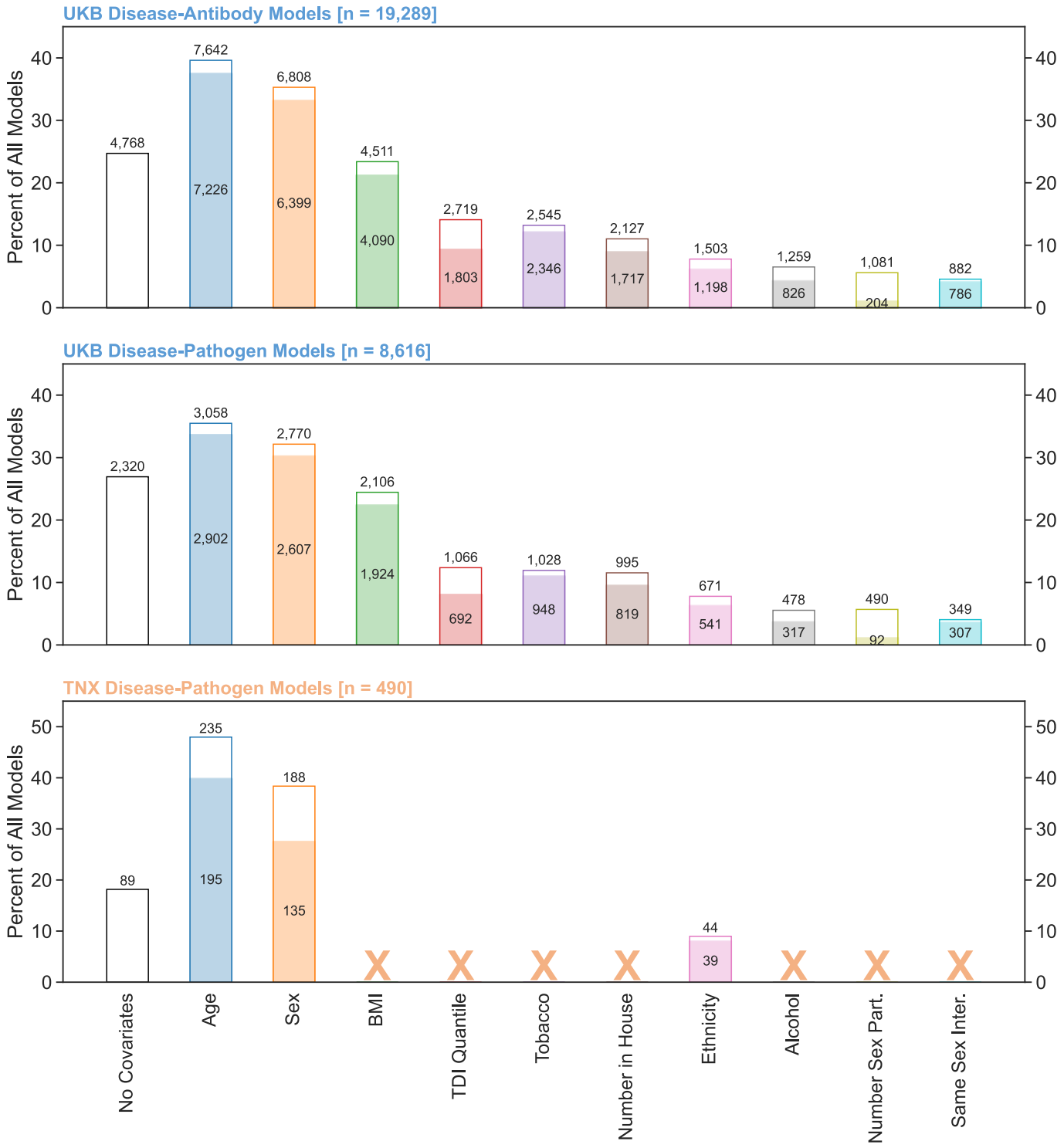

Figure S2 | Covariate use across models.

Bar charts summarizing covariate usage in antibody-disease and pathogen-disease tests across the discovery and replication cohorts. The height of each bar indicates the percentage of all models that were adjusted for the corresponding sociodemographic or health-related covariate. To be included in an antibody-disease model for adjustment, a covariate had to be significantly associated (unadjusted

$p < 0.05$) in univariate tests with both the disease status and the titer level separately (**Table S1**). The filled portion of each bar indicates the percentage of the final multivariate logistic regression models in which the covariate remained significant (unadjusted $p < 0.05$). The covariates that were unavailable in the TriNetX (TNX) cohort are marked by an orange “X”. Abbreviations: Part: Partners; Inter: Intercourse; TDI: Townsend deprivation index.

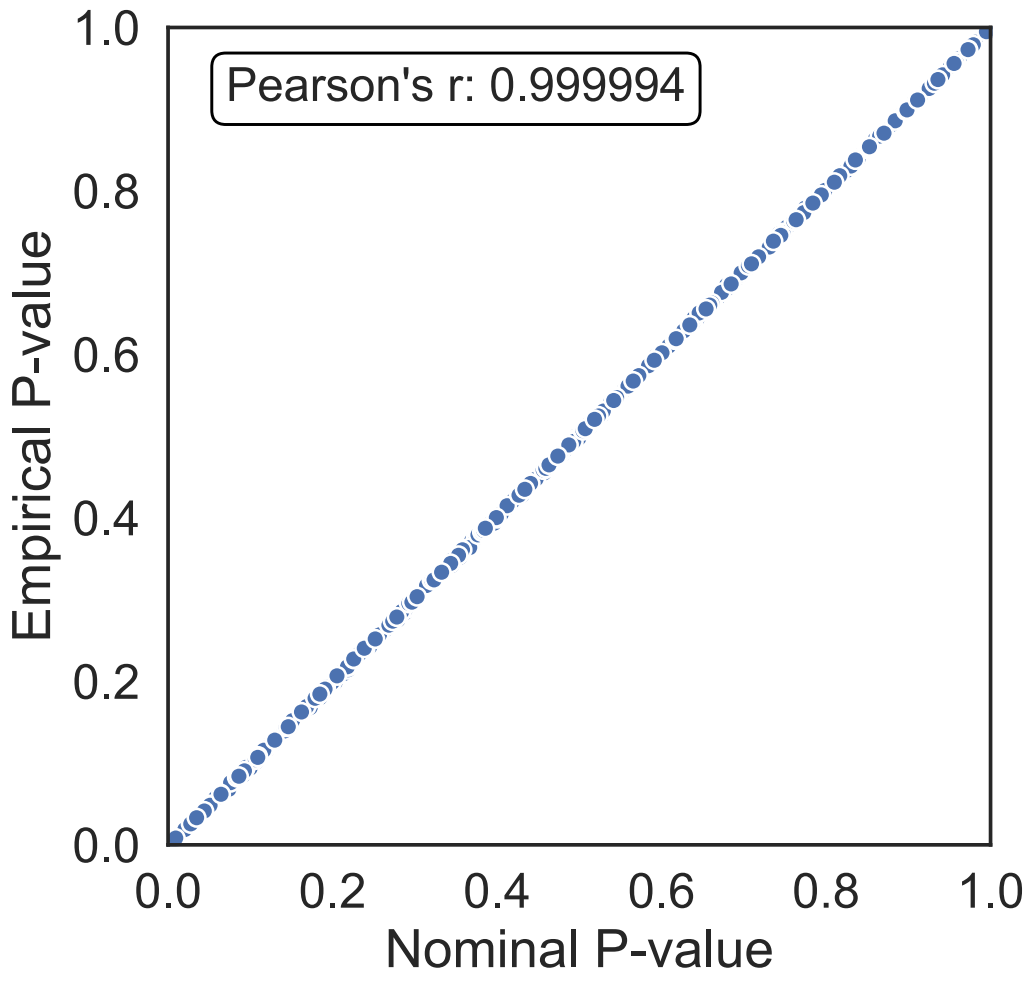

Figure S3 | Nominal versus empirical p-values across all UK Biobank antibody-disease models.

Scatter plot comparing the nominal and empirical p-values for all discovery cohort antibody-disease models. To calculate empirical p-values, 10,000 permutations of each antibody-disease model were performed. All permutations for a particular disease were combined into a per-disease null distribution, yielding 450,000 permutation results per disease (see Methods). The nominal p-value for a specific antibody-disease model was compared to the disease-specific null distribution to calculate the empirical p-value.

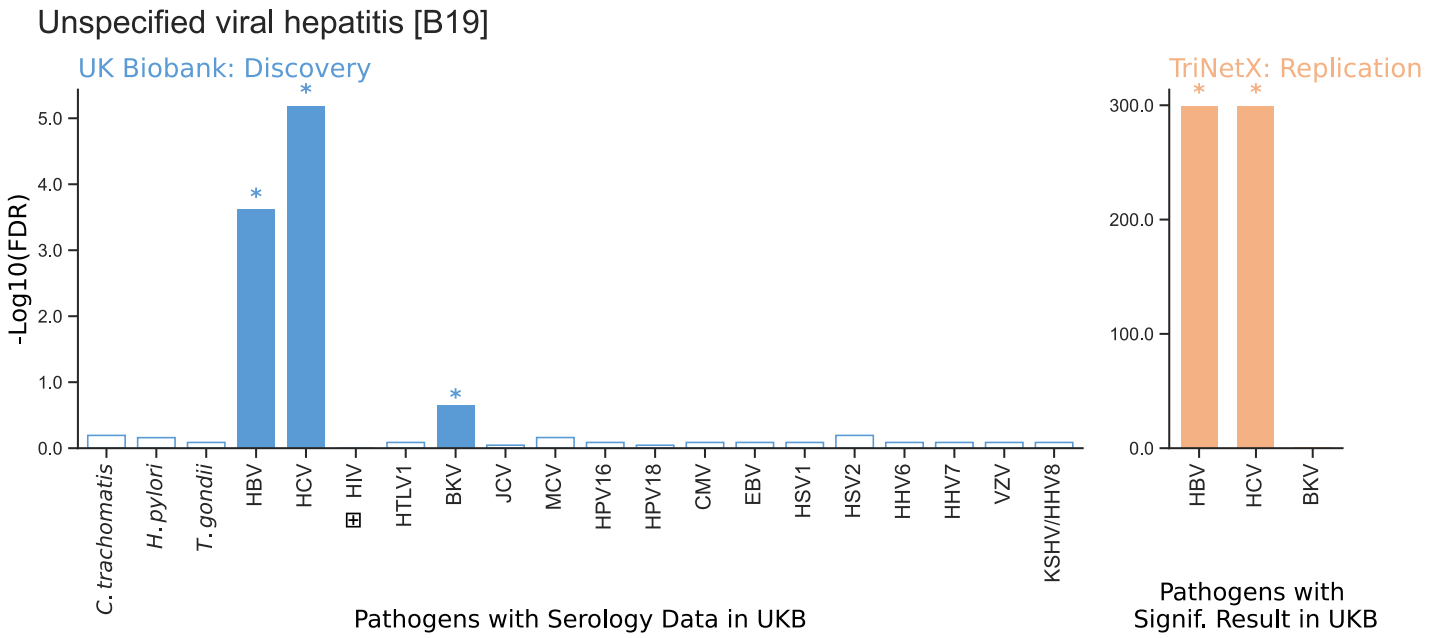

Figure S4 | Illustration of the two-step discovery-replication process.

Bar charts depicting our methods' results for a control disease. The left plot (blue) displays the significance of association (-Log10 transformed per-disease Benjamini-Hochberg false discovery rate (FDR)) for each pathogen with the disease of interest, here the control disease, “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 UK Biobank (UKB) cohort with our more lenient threshold of 0.3. The right plot (orange) shows the results of testing only the significant UKB results in the replication cohort, TriNetX. Only HBV and HCV remain significant at the more stringent replication threshold of per-disease FDR < 0.01, leaving the pairs HBV-unspecified viral hepatitis and HCV-unspecified viral hepatitis the only replicated results.

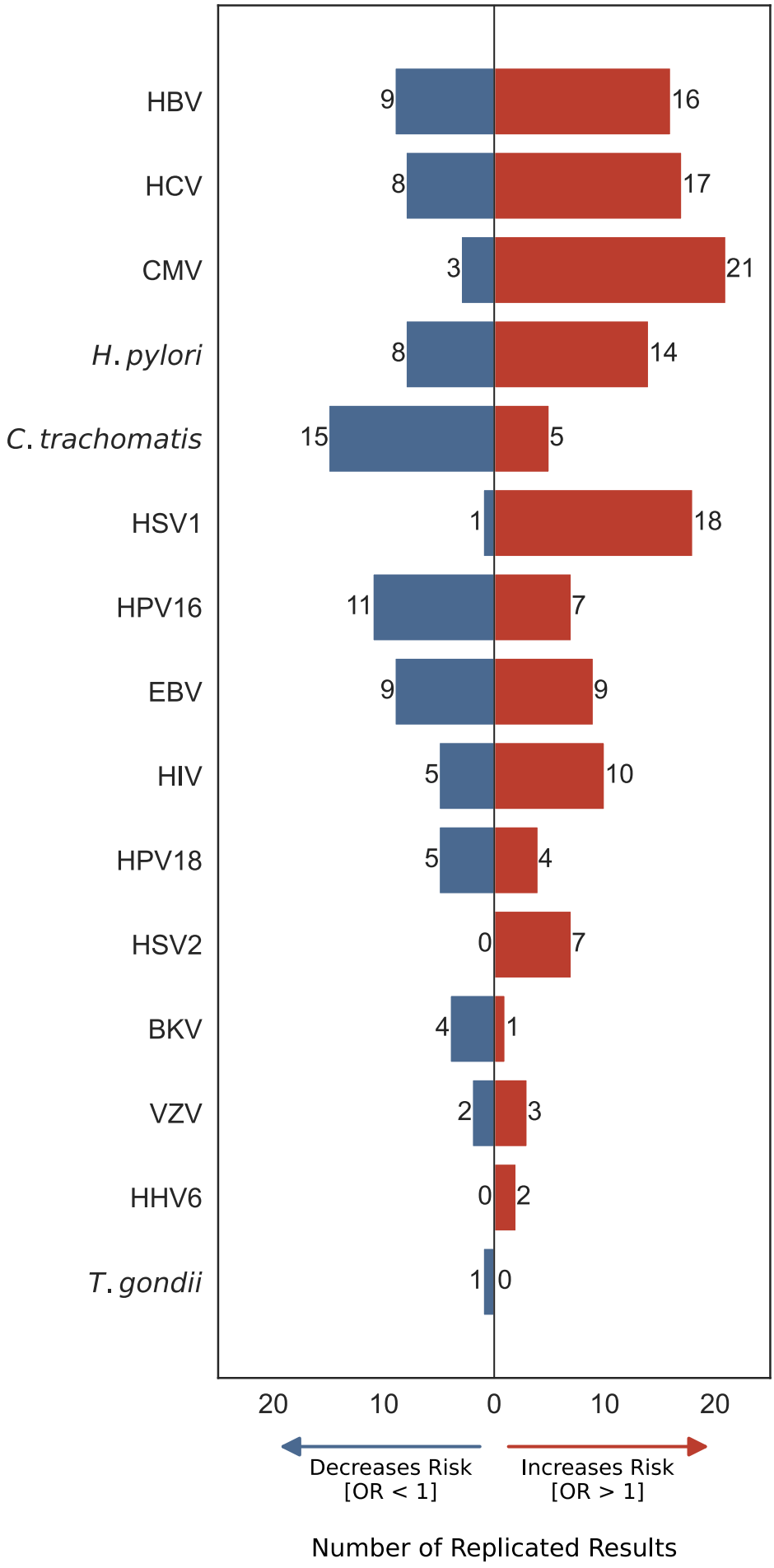
HBV
HCV
CMV
H. pylori
C. trachomatis
HSV1
HPV16
EBV
HIV
HPV18
HSV2
BKV
VZV
HHV6
T. gondii
9
16
8
17
3
21
8
14
15
5
1
18
11
7
9
9
5
10
5
4
0
7
4
1
2
3
0
2
1
0
20
10
0
10
20
Decreases Risk
[OR < 1]
Increases Risk
[OR > 1]
Number of Replicated Results

Figure S5 | Replicated results split by pathogen and effect direction.

A split bar chart showing the number of replicated “risk” and “protective” associations for each pathogen. The total number of replicated results for each pathogen is split by effect direction. Those with a UK Biobank (UKB) and TriNetX (TNX) odds ratio of less than one are represented by the blue bars to the left of the center line. While those with a UKB and TNX odds ratio greater than one are represented by the red bars to the right of the center line. Each bar is labeled with the total number of replicated results it represents.

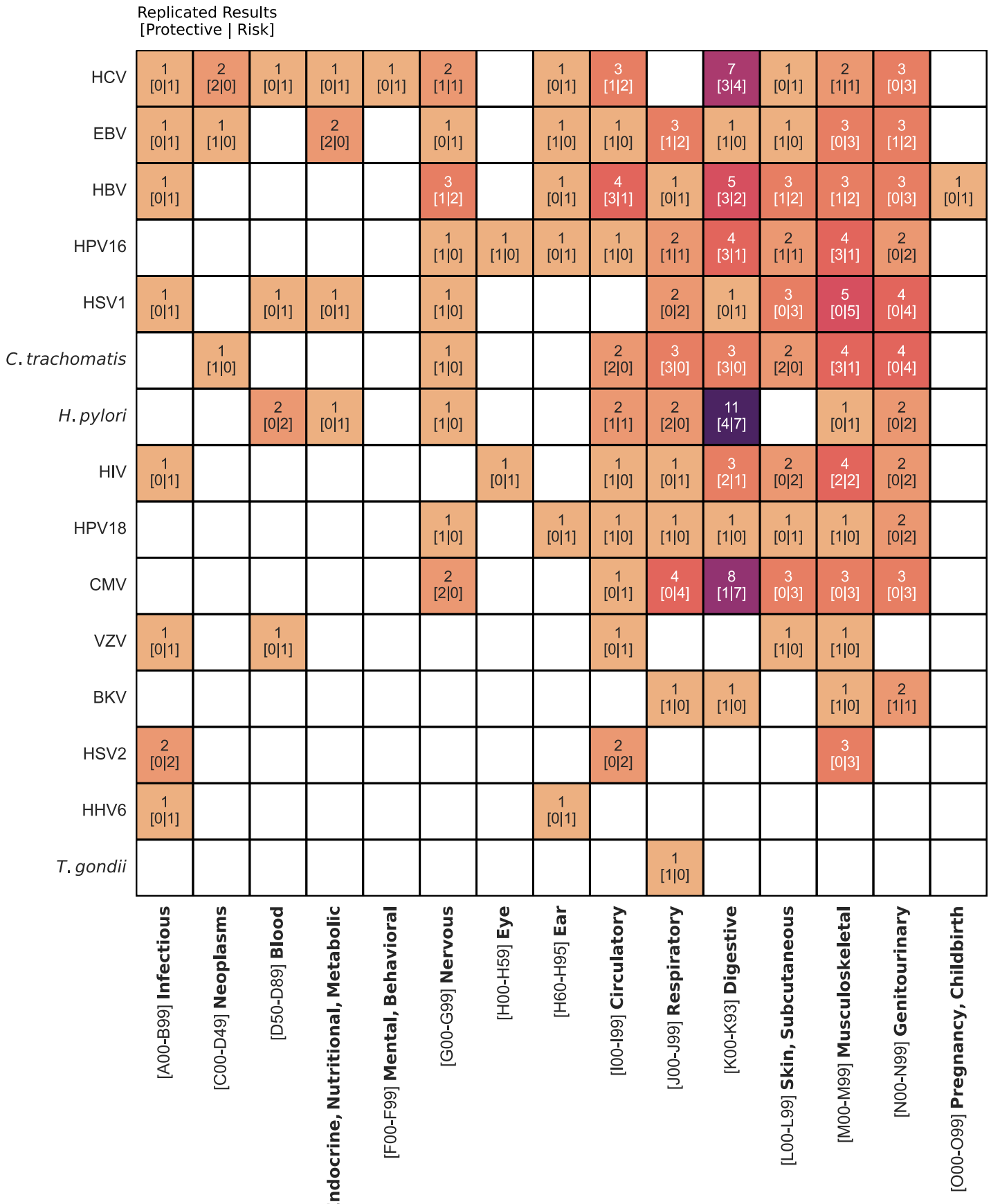
Replicated Results
[Protective | Risk]
HCV
EBV
HBV
HPV16
HSV1
C. trachomatis
H. pylori
HIV
HPV18
CMV
VZV
BKV
HSV2
HHV6
T. gondii
[A00-B99] Infectious
[C00-D49] Neoplasms
[D50-D89] Blood
[E00-E90] Endocrine, Nutritional, Metabolic
[F00-F99] Mental, Behavioral
[G00-G99] Nervous
[H00-H59] Eye
[H60-H95] Ear
[I00-I99] Circulatory
[J00-J99] Respiratory
[K00-K93] Digestive
[L00-L99] Skin, Subcutaneous
[M00-M99] Musculoskeletal
[N00-N99] Genitourinary
[O00-O99] Pregnancy, Childbirth

Figure S6 | Heatmap of replicated results at the ICD10 block level.

A heatmap showing the total number of replicated results for each pathogen that has a replicated result, across all diseases in an International Classification of Diseases 10th revision (ICD10) block, which generally includes only diseases of a particular body system. All cells are annotated with the total number of replicated results above, then in square brackets, the number of these replicated results with odds ratios less than one, followed by those with odds ratios greater than one. White squares with no annotation indicate that no replicated associations were found for that pathogen and any tested diseases in that ICD10 block.

Supplementary Tables

Covariate	Ab Titer Assoc Test	Disease Status Assoc Test
Sex	t-test	Chi-squared
Age	linear regression	t-test
BMI	linear regression	t-test
Race	ANOVA	Chi-squared
Townsend Deprivation Index	ANOVA	Chi-squared
Number in House	ANOVA	Chi-squared
Tobacco Use	ANOVA	Chi-squared
Alcohol Use	ANOVA	Chi-squared
Number of Sex Partners	ANOVA	Chi-squared
Same-sex Intercourse	ANOVA	Chi-squared

Table S1 | Univariate tests used to determine confounding.

This table shows the type of univariate test, all two-sided, performed between each covariate and antibody titer and each covariate and disease status to determine if the covariate was confounding. For a covariate to be considered a confounder, it had to be significantly associated (unadjusted $p < 0.05$) separately with both disease status and antibody titer.