The genetic architecture of human infectious diseases and pathogen-induced cellular phenotypes

Authors: Andrew T. Hale[1,2], Dan Zhou[2], Rebecca L. Sale[2], Lisa Bastarache[3], Liuyang Wang[4,5], Sandra S. Zinkel[6], Steven J. Schiff[7], Dennis C. Ko[4], and Eric R. Gamazon[2,8,9,10*]

[1]Vanderbilt University School of Medicine, Medical Scientist Training Program, Nashville, TN.

[2]Vanderbilt Genetics Institute & Division of Genetic Medicine, Vanderbilt University Medical Center, Nashville, TN.

[3]Department of Bioinformatics, Vanderbilt University School of Medicine, Nashville, TN.

[4]Department of Molecular Genetics and Microbiology, Duke University School of Medicine, Durham, NC.

[5]Division of Infectious Diseases, Department of Medicine, Duke University School of Medicine, Durham, NC.

[6]Division of Hematology and Oncology, Vanderbilt University Medical Center, Nashville, TN.

[7]Centers for Neural Engineering and Infectious Disease Dynamics, Departments of Neurosurgery, Engineering Science and Mechanics, and Physics, Penn State University. University Park, PA.

[8]Clare Hall, University of Cambridge, Cambridge, UK.

[9]MRC Epidemiology Unit, University of Cambridge, Cambridge, UK

[10]Lead Contact.

* Correspondence and requests for materials should be addressed to E.R.G. (eric.gamazon@vanderbilt.edu)

SUMMARY

Infectious diseases (ID) represent a significant proportion of morbidity and mortality across the world. Host genetic variation is likely to contribute to ID risk and downstream clinical outcomes, but there is a need for a genetics-anchored framework to decipher molecular mechanisms of disease risk, infer causal effect on potential complications, and identify instruments for drug target discovery. Here we perform transcriptome-wide association studies (TWAS) of 35 clinical ID traits in a cohort of 23,294 individuals, identifying 70 gene-level associations with 26 ID traits. Replication in two large-scale biobanks provides additional support for the identified associations. A phenome-scale scan of the 70 gene-level associations across hematologic, respiratory, cardiovascular, and neurologic traits proposes a molecular basis for known complications of the ID traits. Using Mendelian Randomization, we then provide causal support for the effect of the ID traits on adverse outcomes. The rich resource of genetic information linked to serologic tests and pathogen cultures from bronchoalveolar lavage, sputum, sinus/nasopharyngeal, tracheal, and blood samples (up to 7,699 positive pathogen cultures across 92 unique genera) and a large catalog of genome-wide associations of microbiome variation generated from phylogenetic analysis of 16S rRNA gene sequences are developed here into a platform to interrogate the genetic basis of compartment-specific infection and colonization. To accelerate insights into cellular mechanisms, we develop a TWAS repository of gene-level associations in a broad collection of human tissues with 79 pathogen-exposure induced cellular phenotypes as a discovery and replication platform. Cellular phenotypes of infection by 8 pathogens included pathogen invasion, intercellular spread, cytokine production, and pyroptosis. These rich datasets will facilitate mechanistic insights into the role of host genetic variation on ID risk and pathophysiology, with important implications for our molecular understanding of potentially severe phenotypic outcomes.

Keywords: PrediXcan; Human Genetics; Infectious Disease; Transcriptomics; TWAS; GWAS; Electronic Health Records; Hi-HOST; BioVU; UK Biobank; FinnGen; Functional Genomics; GTEx; PheWAS; Mendelian Randomization; Clinical Microbiology; 16S rRNA; Microbiome

HIGHLIGHTS

- Atlas of genome-wide association studies (GWAS) and transcriptome-wide association studies (TWAS) results for 35 clinical infectious disease (ID) phenotypes, with genome-wide and transcriptome-wide significant results for 13 and 26 clinical ID traits, respectively

- Phenome-scale scan of ID-associated genes across 197 hematologic, respiratory, cardiovascular, and neurologic traits, facilitating identification of genes associated with known complications of the ID traits

- Mendelian Randomization analysis, leveraging naturally occurring DNA sequence variation to perform “randomized controlled trials” to test the causal effect of ID traits on potential outcomes and complications

- A genomic resource of TWAS associations for 79 pathogen-induced cellular traits from High-throughput Human *in vitrO* Susceptibility Testing (Hi-HOST) across 44 tissues as a discovery and replication platform to enable *in silico* cellular microbiology and functional genomic experiments

INTRODUCTION

Genome-wide association studies (GWAS) and large-scale DNA biobanks with phenome-scale information are making it possible to identify the genetic basis of a wide range of complex traits in humans (Bycroft et al., 2018; Roden et al., 2008). A parallel development is the increasing availability of GWAS summary statistics, facilitating genetic analyses of entire disease classes and promising considerably improved resolution of genetic effects on human disease (Cotsapas et al., 2011; Gamazon et al., 2019). Recent analysis involving 558 well-powered GWAS results found that trait-associated loci cover ~50% of the genome, enriched in both coding and regulatory regions, and of these, ~90% are implicated in multiple traits (Watanabe et al., 2019). However, the breadth of clinical and biological information in these datasets will require new methodologies and additional high-dimensional data to advance our understanding of the genetic architecture of complex traits and relevant molecular mechanisms (Bulik-Sullivan et al., 2015; Gamazon et al., 2018; Shi et al., 2016). Approaches to understanding the functional consequences of implicated loci and genes are needed to determine causal pathways and potential mechanisms for pharmacological intervention.

The genetic basis of infectious disease (ID) risk and severity has been relatively understudied, and its implications for etiological understanding of human disease and drug target discovery may be investigated using phenome-scale information increasingly available in these biobanks. ID risk and pathogenesis is likely to be multifactorial, resulting from a complex interplay of host genetic variation, environmental exposure, and pathogen-specific molecular mechanisms. With few exceptions, the extent to which susceptibility to ID is correlated with host genetic variation remains poorly understood (de Bakker and Telenti, 2010). However, for at least some ID traits, including poliomyelitis, hepatitis, and *Helicobacter pylori* (Burgner et al., 2006; Herndon and Jennings, 1951; Hohler et al., 2002; Malaty et al., 1994), disease risk is heritable, based on twin studies. Although monogenic mechanisms of ID risk have been

demonstrated (Casanova, 2015a, b), the contribution of variants across the entire allele frequency spectrum to interindividual variability in ID risk remains largely unexplored.

Here we conduct genome-wide association studies (GWAS) and transcriptome-wide association studies (TWAS) of 35 ID traits. To implement the latter, we apply PrediXcan (Gamazon et al., 2018; Gamazon et al., 2015), which exploits the genetic component of gene expression to probe the molecular basis of disease risk. We combine information across a broad collection of tissues to determine gene-level associations using a multi-tissue approach, which displays markedly improved statistical power over a single-tissue approach (Barbeira et al., 2019; Gamazon et al., 2018; Gamazon et al., 2015). Notably, we identify 70 gene-level associations for 26 of 35 ID traits, i.e., heretofore referred to as ID-associated genes, and conduct replication using the corresponding traits in the UK Biobank and FinnGen consortia data (Bycroft et al., 2018; Locke et al., 2019). The rich resource of genetic information linked to clinical microbiology information (serology and culture data) across bacterial, fungal, and viral genera that we leverage provides a platform to interrogate the genetic basis of compartment-specific infection and colonization. Linking high-resolution taxonomic classification from 16S ribosomal RNA (rRNA) sequencing to host GWAS information has been used to investigate the contribution of host genetic variation to microbiome composition (Hughes et al., 2020). We therefore exploit host GWAS for 155 pathogens in the microbiome (Hughes et al., 2020) to gain further insights into identified genetic risk factors for an ID trait. To determine the phenotypic consequences of ID-associated genes, including adverse outcomes and complications, we perform a phenome-scale scan across hematologic, respiratory, cardiovascular, and neurologic traits. To extend these findings, we use a Mendelian Randomization framework (Lawlor et al., 2008) to conduct causal inference on the effect of a clinical ID trait on an adverse clinical outcome. To elucidate the cellular mechanisms through which host genetic variation influences disease risk, we generate an atlas of gene-level associations with 79 pathogen-induced cellular phenotypes determined by High-throughput Human *in vitrO* Susceptibility Testing (Hi-HOST)

(Wang et al., 2018) as a discovery and replication platform. The rich genomic resource we generate and the methodology we develop promise to accelerate discoveries on the molecular mechanisms of infection, improve our understanding of adverse outcomes and complications, and enable prioritization of new therapeutic targets.

RESULTS

A schematic diagram illustrating our study design and the reference resource we provide can be found in Figure 1. Here we analyzed 35 clinical ID traits, 79 pathogen-exposure-induced cellular traits, and 197 (cardiovascular, hematologic, neurologic, and respiratory) traits. We performed GWAS and TWAS (Gamazon et al., 2015; Gusev et al., 2016) to investigate the genetic basis of the ID traits and their potential adverse outcomes and complications. We exploited serology and culture data linked to genetic information and genome-wide associations with microbiome traits to investigate compartment-specific patterns of infection. We conducted causal inference within a Mendelian Randomization framework (Davey Smith and Hemani, 2014), exploiting genetic instruments for naturally “randomized controlled trials” to evaluate the causality of an observed association between a modifiable exposure or risk factor and a clinical phenotype. We generate a rich resource for understanding the genetic and molecular basis of infection and potential adverse effects and complications.

GWAS and TWAS of 35 infectious disease clinical phenotypes implicate broad range of molecular mechanisms

We sought to characterize the genetic determinants of 35 ID traits, including many which have never been investigated using a genome-wide approach. First, we performed GWAS of each of these phenotypes using a cohort of 23,294 patients of European ancestry with extensive EHR information from BioVU (Roden et al., 2008). We identified genome-wide significant associations ($p < 5x10^{-8}$) for 13 ID traits (Figure 2A and Supplementary Table 1). The

SNP rs17139584 on chromosome 7 was our most significant association ($p = 1.21 \times 10^{-36}$) across all traits, with bacterial pneumonia. A LocusZoom plot shows several additional genome-wide significant variants in the locus (Figure 2B), in low linkage disequilibrium ($r^2 < 0.20$) with the sentinel variant rs17139584, including variants in the *MET* gene and in *CFTR*. The *MET* gene acts as a receptor to *Listeria monocytogenes* internalin InlB, mediating entry into host cells; interestingly, listeriosis, a bacterial infection caused by this pathogen, can lead to pneumonia (García-Montero et al., 1995). Given the observed associations in the cystic fibrosis gene *CFTR* (~650 Kb downstream of *MET*), we also asked whether the rs17139584 association was driven by cystic fibrosis. Notably, the SNP remained nominally significant, though its significance was substantially reduced, after adjusting for cystic fibrosis status ($p = 0.007$; see Methods) or excluding the cystic fibrosis cases ($p = 0.02$). The LD profile of the genome-wide significant results in this locus (Figure 2B) is consistent with the involvement of multiple gene mechanisms (e.g., *MET* and *CFTR*) underlying bacterial pneumonia risk. The rs17139584 association replicated ($p = 5.3 \times 10^{-3}$) in the UK Biobank (Bycroft et al., 2018). Eighty percent to ninety percent of patients with cystic fibrosis suffer from respiratory failure due to chronic bacterial infection (with *Pseudomonas aeruginosa*) (Lyczak et al., 2002). Thus, future studies on the role of this locus in lung infection associated with cystic fibrosis may provide germline predictors of this complication; alternatively, the locus may confer susceptibility to lung inflammation, regardless of cystic fibrosis status. Collectively, our analysis shows strong support for allelic heterogeneity, with likely multiple independent variants in the locus contributing to interindividual variability in bacterial pneumonia susceptibility.

Additional examples of genome-wide significant associations with other ID traits were identified. For example, rs192146294 on chromosome 1 was significantly associated ($p = 1.23 \times 10^{-9}$) with *Staphylococcus* infection. In addition, 10 variants on chromosome 8 were significantly associated ($p < 1.17 \times 10^{-8}$) with Mycoses infection.

Next, to improve statistical power, we performed multi-tissue PrediXcan (Barbeira et al., 2019; Gamazon et al., 2018; Gamazon et al., 2015). We constructed an atlas of TWAS associations with these ID traits in separate European and African American ancestry cohorts (Supplementary Data File 1) as a resource to facilitate mechanistic studies. Notably, 70 genes reached experiment-wide or individual ID-trait significance for 26 of the 35 clinical ID traits (Figure 3A and Table 1). Sepsis, the clinical ID trait with the largest sample size in our data (Figure 3B; Phecode 994; number of cases 2,921; number of controls 22,874), was significantly associated ($p = 8.16 \times 10^{-7}$) with *IKZF5* after Bonferroni correction for the number of genes tested. The significant genes (Table 1) were independent of the sentinel variants from the GWAS (Supplementary Table 1), indicating that the gene-based test was identifying additional signals.

Our analysis identified previously implicated genes for the specific ID traits but also proposes novel genes and mechanisms. ID-associated genes include *NDUFA4* for intestinal infection, a component of the cytochrome oxidase and regulator of the electron transport chain (Balsa et al., 2012); *AKIRIN2* for candidiasis, an evolutionarily conserved regulator of inflammatory genes in mammalian innate immune cells (Tartey et al., 2015; Tartey et al., 2014); *ZNF577* for viral hepatitis C, a gene previously shown to be significantly hypermethylated in hepatitis C related hepatocellular carcinoma (Revill et al., 2013); and epithelial cell adhesion molecule (*EPCAM*) for tuberculosis, a known marker for differentiating malignant tuberculous pleurisy (Sun et al., 2014), among many others. These examples of ID-associated genes highlight the enormous range of molecular mechanisms that may contribute to susceptibility and complication phenotypes.

Replication of gene-level associations with infectious diseases in the UK Biobank and FinnGen

To bolster our genetic findings and show that our results were not driven by biobank-specific confounding, we performed replication analysis for a subset of ID traits available in the

independent UK Biobank and FinnGen consortia datasets (see Methods and Supplementary Data File 2). Notably, the genes associated with intestinal infection ($p < 0.05$,) in BioVU – the ID trait with the largest sample size in BioVU and with a replication dataset in the independent FinnGen biobank – showed a significantly greater level of enrichment for gene-level associations with the same trait in FinnGen compared to the remaining set of genes (Figure 3C). Thus, higher significance (i.e., lower p-value) was observed in FinnGen for the intestinal infection associated genes identified in BioVU, which included the top association *NDUFA4* (discovery $p = 1.83 \times 10^{-9}$, replication $p = 0.044$). These results illustrate the value of exploiting large-scale biobank resources for genetic studies of ID traits- despite well-known caveats (Ko and Urban, 2013; Power et al., 2017).

Replication of tissue-level PrediXcan results enables identification of robust tissue-specific associations, lending insights into putative molecular mechanisms. Here we provide tissue-level replication for ID traits available in the UK Biobank and FinnGen: bacterial pneumonia (lung, Phecode 480.1), influenza (lung, Phecode 481), viral pneumonia (lung, Phecode 480.2), meningitis (10 brain regions, Phecode 320), and encephalitis (10 brain regions, Phecode 323) (Supplementary Data File 2). For example, replication analysis in the UK Biobank (see Methods) for influenza (Figure 3D, Supplementary Data File 2) and bacterial pneumonia (Figure S1A, Supplementary Data File 2) associated genes in lung tissue revealed many robust tissue-level associations. We also replicated the top genes ($p < 0.05$) associated with influenza (Supplementary Data File 2) and viral pneumonia (Supplementary Data File 2) in FinnGen (see Methods), providing additional support for the role of these genes in pulmonary disease. We found substantial replication in the UK Biobank for individual brain tissue gene expression for meningitis in hippocampus, cerebellar hemisphere, and hypothalamus (Figure S1B-D, Supplementary Data File 2). Similar trends were observed for the remaining brain tissues (Supplementary Data File 2). In addition, we found substantial replication in the UK Biobank in caudate, cortex, and cerebellum tissue for encephalitis (Figure S1E-G, Supplementary Data File

2). Again, similar trends were observed across all brain regions (Supplementary Data File 2). Intriguingly, the cerebellar gene expression traits nominally associated with meningitis in BioVU improved the signal-to-noise ratio for identifying cerebellar associations in FinnGen (Figure 3E, Supplementary Data File 1-2). These data highlight reproducible genetically-determined mechanisms underlying ID risk.

Tissue expression profile of infectious disease associated genes suggests tissue-dependent mechanisms

The ID-associated genes tend to be less tissue-specific (i.e., more ubiquitously expressed) than the remaining genes (Figure S2A, Mann Whitney U test on the τ statistic, $p = 7.5 \times 10^{-4}$), possibly reflecting the multi-tissue PrediXcan approach we implemented, which prioritizes genes with multi-tissue support to improve statistical power, but also the genes' pleiotropic potential. We hypothesized that tissue expression profiling of ID-associated genes can provide additional insights into disease etiologies and mechanisms. For example, the intestinal infection associated gene *NDUFA4* is expressed in a broad set of tissues, including the alimentary canal, but displays relatively low expression in whole blood (Figure S2B). In addition, *TOR4A,* the most significant association with bacterial pneumonia (Table 1), is most abundantly expressed in lung, consistent with the tissue of pathology, but also in spleen (Figure S2C), whose rupture is a lethal complication of the disease (Domingo et al., 1996; Gerstein et al., 1967). These examples illustrate the diversity of tissue-dependent mechanisms that may contribute in complex and dynamic ways to interindividual variability in ID susceptibility and progression. We therefore provide a resource of single-tissue gene-level associations with the ID traits to facilitate molecular or clinical follow-up studies.

Genetic overlap reveals host gene expression programs and common pathways as targets for pathogenicity

We hypothesized that ID-associated genes implicate shared functions and pathways, which may reflect common targeted host transcriptional programs. Among the 70 gene-level associations with the 35 clinical ID traits, 40 proteins are post-translationally modified by phosphorylation (Supplementary Table 2), a significant enrichment (Benjamini-Hochberg adjusted $p < 0.10$ on DAVID annotations (Huang da et al., 2009)) relative to the rest of the genome, indicating that phosphoproteomic profiling can shed substantial light on activated host factors and perturbed signal transduction pathways during infection (Soderholm et al., 2016; Stahl et al., 2013). In addition, 16 proteins are acetylated, consistent with emerging evidence supporting this mechanism in the host antiviral response (Murray et al., 2018) (Supplementary Table 3). These data identify specific molecular mechanisms across ID traits with critical regulatory roles (e.g., protein modifications) in host response among the ID-associated genes.

We tested the hypothesis that distinct infectious agents exploit common pathways to find a compatible intracellular niche in the host, potentially implicating shared genetic risk factors. Notably, 64 of the 70 ID-associated genes (Table 1) were nominally associated ($p < 0.05$) with multiple ID traits (Supplementary Table 4). These genes warrant further functional study as broadly exploited mechanisms targeted by pathogens or as broadly critical to pathogen-elicited immune response. Gene Set Enrichment Analysis (GSEA) of these genes implicated a number of significant (FDR < 0.05) gene sets (Figure 4A), including those involved in actin-based processes and cytoskeletal protein binding, processes previously demonstrated to mediate host response to pathogen infection (Taylor et al., 2011). Since diverse bacterial and viral pathogens target host regulators that control the cytoskeleton (which plays a key role in the biology of infection) or modify actin in order to increase virulence, intracellular motility, or intercellular spread (Aktories and Barbieri, 2005; Yu et al., 2011; Zahm et al., 2013), these results reassuringly lend support to the involvement of the genes in infectious pathogenesis.

Notably, we identified an enrichment (FDR = 9.68×10^{-3}) for a highly conserved motif ("TCCCRNNRTGC"), within 4 kb of transcription start site (TSS) of multi-ID associated genes

(Figure 4A-B), that does not match any known transcription factor binding site (Xie et al., 2005) and may be pivotal for host-pathogen interaction for the diversity of infectious agents included in our study. In addition, we found that several of the multi-ID associated genes (with the sequence motif near the TSS) have been observed in host-pathogen protein complexes (by both coimmunoprecipitation and affinity chromatography approaches) for the specific pathogens responsible for the ID traits (Ammari et al., 2016). See Supplementary Data File 3 for complete list of host-pathogen interactions for these genes/proteins. One example is *CDK5*, a gene significantly associated with Gram-positive septicemia (Table 1) and nominally associated with multiple ID traits, including herpes simplex. CDK5 is activated by p35, whose cleaved form p25 results in subcellular relocalization of CDK5. The CDK5-p25 complex regulates inflammation (Na et al., 2015) (whose large-scale disruption is characteristic of septicemia) and induces cytoskeletal disruption in neurons (Patrick et al., 1999) (where the herpes virus is responsible for lifelong latent infection). The A and B chains of the CDK5-p25 complex (Figure 4C for structure diagram (Tarricone et al., 2001)) are required for cytoskeletal protein binding (CDK5), whereas the D and E chains (p25) are involved in actin regulation and kinase function, all molecular processes implicated in our pathway analysis. Intriguingly, blocking CDK5 can have a substantial impact on the outcome of inflammatory diseases including sepsis (Pfänder et al., 2019), enhancing the anti-inflammatory potential of immunosuppressive treatments, and has been shown to attenuate herpes virus replication (Man et al., 2019), suggesting that modulation of this complex is important for viral pathogenesis.

CDK5 is also altered by several other viruses, identified using unbiased mass spectrometry analysis (Davis et al., 2015) (Figure 4D), indicating a broadly exploited mechanism (across pathogens) that is consistent with the gene's multi-ID genetic associations in our TWAS data (Figure 4D). The CDK5-interaction proteins include: 1) M2_134A1 (matrix protein 2, influenza A virus), a component of the proton-selective ion channel required for viral genome release during cellular entry and is targeted by the anti-viral drug amantadine (Hay et al., 1985);

2) VE7_HPV16, a component of human papillomavirus (HPV) required for cellular transformation and trans-activation through disassembly of E2F1 transcription factor from RB1 leading to impaired production of type I interferons (Barnard et al., 2000; Chellappan et al., 1992; Phelps et al., 1988); 3) VE7_HPV31, which has been shown to engage histone deacetylases 1 and 2 to promote HPV31 genome maintenance (Longworth and Laimins, 2004); 4) VCYCL_HHV8P (cyclin homolog within the human herpesvirus 8 genome), which has been shown to control cell cycle through CDK6 and induce apoptosis through Bcl2 (Duro et al., 1999; Ojala et al., 1999; Ojala et al., 2000); and 5) F5HC81_HHV8, predicted to act as a viral cyclin homolog. Overall, these data underscore the evolutionary strategies that pathogens have evolved to promote infection, including the hijacking of the host transcriptional machinery and the biochemical alterations of the host proteome.

Serology and culture data reveal insights into clinical infection and pathogen colonization

We exploited extensive clinical microbiological laboratory analysis of blood (Figure 5A), bronchoalveolar lavage, sputum, sinus/nasopharyngeal, and tracheal cultures for bacterial and fungal pathogen genus identification (Figure S3A-D), as well as respiratory viral genus identification (Figure S3E) (see Methods) to evaluate phenotype resolution and algorithm. For example, we found that *Staphylococcus* infection (Phecode = 041.1) performed well in classifying *Staphylococcus aureus* infection based on blood culture data. The area under the Receiver Operating Characteristic (ROC) curve was 0.938 (Figure 5B) with standard error of 0.008 generated from bootstrapping (see Methods). The area under the curve (AUC) quantifies the probability that the Phecode classifier ranks a randomly chosen positive instance of *Staphylococcus aureus* infection in blood higher than a randomly chosen negative one. In comparison, the first principal component (PC) in our European ancestry samples showed AUC of 0.514 (Figure 5B) while sex and age performed even more poorly (AUC $\approx$ 0.50). We then tested a logistic model with the Phecode classifier, age, sex, and the first 5 PCs in the model.

The Phecode classifier was significantly associated ($p < 2.2x10^{-16}$) after conditioning on the remaining covariates. The fitted value from the joint model consisting of the remaining covariates showed AUC of 0.568 (Figure 5B). Collectively, culture data for improved resolution of clinical infection and pathogen colonization provide validation of our approach.

To expand these findings and further dissect the complex pathogen-colonization patterns in humans, we utilized host genome-wide associations with human gut microbiome variation for 155 pathogens (Hughes et al., 2020) identified through 16S rRNA sequencing. For example, among the top SNP associations with *Desulfovibrionaceae* ($p<0.0001$), a sulfate-reduced bacterium associated with intraabdominal infections and inflammatory bowel disease (Goldstein et al., 2003; Loubinoux et al., 2002), we observed an enrichment for SNPs associated with intestinal infection in BioVU (Figure 5C). These data provide a reference resource to elucidate how genetically determined microbiome variation influences ID trait susceptibility.

Phenome scan of clinical ID-associated genes identifies adverse outcomes and complications

Electronic Health Records (EHR) linked to genetic data may reveal insights into associated clinical sequalae (Bastarache et al., 2018; Denny et al., 2013; Unlu et al., 2020). To assess the phenomic impact of ID-associated genes (Table 1), we performed a phenome-scale scan across 197 hematologic, respiratory, cardiovascular, and neurologic traits available in BioVU (Figure 6A and Supplementary Data File 4). Correcting for total number of genes and phenotypes tested, we identified four gene-phenotype pairs reaching experiment-wide significance: 1) *WFDC12*, our most significant ($p = 4.23x10^{-6}$) association with meningitis and a known anti-bacterial gene (Hagiwara et al., 2003), is also associated with cerebral edema and compression of brain ($p = 1.35x10^{-6}$), a feared clinical complication of meningitis (Niemöller and Täuber, 1989); 2) *TM7SF3*, the most significant gene with Gram-negative sepsis ($p = 1.37x10^{-6}$), is also associated with acidosis ($p = 1.95x10^{-6}$), a known metabolic derangement associated

with severe sepsis (Suetrong and Walley, 2016), and a gene known to play a role in cell stress and the unfolded protein response (Isaac et al., 2017); 3) *TXLNB*, the most significant gene associated with viral warts and human papillomavirus infection ($p = 4.35x10^{-6}$), is also associated with abnormal involuntary movements, $p = 1.39x10^{-6}$; and 4) *RAD18*, the most significant gene associated with Streptococcus infection ($p = 2.01x10^{-6}$), is also associated with anemia in neoplastic disease ($p = 3.10x10^{-6}$). Thus, coupling genetic analysis to EHR data with their characteristic breadth of clinical traits offers the possibility of determining the phenotypic consequences of ID-associated genes, including known (in the case of *WFDC12* and *TM7SF3*) potentially adverse health outcomes and complications.

Mendelian Randomization provides causal support for the effect of infectious disease trait on identified adverse phenotypic outcomes/complications

Since our gene-level associations with clinical ID diagnoses implicated known adverse complications, we sought to explicitly evaluate the causal relation between the ID traits and the adverse outcomes/complications. We utilized the Mendelian Randomization paradigm (Lawlor et al., 2008) (Figure 6B), which exploits genetic instruments to make causal inferences in observational data, in effect, performing randomized controlled trials to evaluate the causal effect of “exposure” (i.e., ID trait) on “outcome” (e.g., the complication). Specifically, we conducted multiple-instrumental-variable causal inference using GWAS (Davey Smith and Hemani, 2014) and PrediXcan summary results. First, we used independent SNPs ($r^2 = 0.01$) that pass a certain threshold for significance with the ID trait ($p < 1.0x10^{-5}$) as genetic instruments. To control for horizontal pleiotropy and account for the presence of invalid genetic instruments, we utilized MR-Egger regression and weighted-median Mendelian Randomization (see Methods) (Bowden et al., 2015; Bowden et al., 2016).

Here, we performed Mendelian Randomization on the ID trait and a complication trait identified through the unbiased phenome scan. This analysis yields causal support for the effect

of 1) Gram-negative sepsis on acidosis (Figure 6C, weighted-median estimator $p = 2.0x10^{-7}$); and 2) meningitis on cerebral edema and compression of brain (Figure 5C, weighted-median estimator $p = 2.7x10^{-3}$). Our resource establishes a framework to elucidate the genetic component of an ID trait and its impact on the human disease phenome, enabling causal inference on the effect of an ID trait on potential complications.

TWAS of 79 pathogen-exposure induced cellular traits highlights cellular mechanisms and enables validation of ID gene-level associations

Elucidating how the genes influence infection-related cellular trait variation may provide a mechanistic link to ID susceptibility. We thus performed TWAS of 79 pathogen-induced cellular traits – including infectivity and replication, cytokine levels, and host cell death, among others (Wang et al., 2018) (Supplementary Data File 5). We identified 38 gene-level associations reaching trait-level significance ($p < 2.87x10^{-6}$, correcting for number of statistical tests; Figure 7A). In addition, we replicated SNP associations with the cellular traits using the genetic associations with the ID traits (Supplementary Data File 6), which map to cellular phenotypes (Supplementary Data File 7).

Integration of EHR data into Hi-HOST (Wang et al., 2018) may enable replication of gene-level associations with a clinical ID trait. Indeed, we observed a marked enrichment for genes associated with direct *Staphylococcus* toxin exposure cellular response in Hi-HOST among the human Gram-positive septicemia associated genes from BioVU (see Supplementary Data File 5 for genes with FDR < 0.05) (Figure 7B). In addition, integration of EHR data into Hi-HOST may improve the signal-to-noise ratio in Hi-HOST TWAS data. Indeed, the top 300 genes nominally associated ($p < 0.016$) with Staphylococcus infection (Phecode 041.1) in BioVU departed from null expectation for their associations with Staphylococcus toxin exposure in Hi-HOST compared to the full set of genes, which did not (Figure 7C), as perhaps expected due to

the modest sample size. Collectively, these results demonstrate that integrating the EHR-derived TWAS results into TWAS of the cellular trait can greatly improve identification of potentially relevant pathogenic mechanisms.

Phenome scan of TWAS findings from Hi-HOST

To identify potential adverse effects of direct pathogen exposure, we performed a phenome-scan across the 197 cardiovascular, hematologic, neurologic, and respiratory traits as described above. Our top gene-phenotype pairs include: 1) *FAM171B*, our most significant association with interleukin 13 (IL-13) levels is also associated with alveolar and parietoalveolar pneumonopathy ($p = 4.04 \times 10^{-5}$), a phenotype known to be modulated by IL-13 dependent signaling (Zheng et al., 2008); 2) *OSBPL10*, the most significant gene associated with cell death caused by *Salmonella enterica serovar Typhimurium*, is also associated with intracerebral hemorrhage ($p = 4.99 \times 10^{-5}$), a known complication of *S. Typhimurium* endocarditis (Gómez-Moreno et al., 2000). These data highlight the utility of joint genetic analysis of pathogen-exposure-induced phenotypes and clinical ID traits to gain insights into the molecular and cellular basis of complications and adverse outcomes. However, more definitive conclusions will require larger sample sizes and functional studies.

DISCUSSION

ID susceptibility is a complex interplay between host genetic variation and pathogen-exposure induced mechanisms. While GWAS has begun to identify population-specific loci conferring ID risk (Tian et al., 2017), the underlying function of identified variants, predominantly in non-coding regulatory regions, remains poorly understood. Molecular characterization of infectious processes has been, in general, agnostic to the genetic architecture of clinical infection. Although pathogen exposure is requisite to display clinical ID traits, the role of host genetic variation remains largely unexplored.

Our study provides a reference atlas of genetic variants and genetically-determined expression traits associated with 35 clinical ID traits from BioVU. We identified 70 gene-level associations, with replication for a subset of ID traits in the UK Biobank and FinnGen. To provide additional support to our findings, we leveraged a rich resource of genetic information linked to serologic tests and pathogen cultures from five clinical sample sites and exploited a large catalog of genome-wide associations of microbiome variation generated from 16S rRNA based taxonomic classification. A phenome scan across 197 hematologic, respiratory, cardiovascular, and neurologic traits proposes a molecular basis for the link between certain ID traits and outcomes. Using Mendelian Randomization, we determined the ID traits which, as exposure, show significant causal effect on outcomes. Finally, we developed a TWAS catalog of 79 pathogen-exposure induced cellular traits (Hi-HOST) in a broad collection of tissues, which provides a platform to interrogate mediating cellular and molecular mechanisms.

Genetic predisposition to ID onset and progression is likely to be complex (Casanova, 2015a). Monogenic mechanisms conferring ID risk have been proposed, but these mechanisms are unlikely to explain the broad contribution of host genetic influence on ID risk (Casanova, 2015b). Thus, a function-centric methodology is necessary to disentangle potentially causal pathways. Our approach builds on PrediXcan, which estimates the genetically-determined component of gene expression (Gamazon et al., 2015). The genetic component of gene expression can then be tested for association with the trait, enabling insights into potential pathogenic mechanisms (Gamazon et al., 2019) and novel therapeutic strategies (So et al., 2017).

Our study identified genes with diverse functions, including roles in mitochondrial bioenergetics (Balsa et al., 2012; El-Bacha and Da Poian, 2013), regulation of cell death (Labbé and Saleh, 2008), and of course links to host immune response (Brouwer et al., 2019; Liang et al., 2019; Pan et al., 2017; Saitoh et al., 2009; Sharfe et al., 1997; Tsuboi and Meerloo, 2007;

Walenna et al., 2018; Willis et al., 2009; Yu et al., 2017; Zhang et al., 2015). These diverse functions may therefore contribute to pleiotropic effects on clinical outcomes and complications.

In addition, we identified genes implicated in Mendelian diseases, for which susceptibility to infection is a predominant feature, including *WIPF1* (OMIM #614493; recurrent infections and reduced natural killer cell activity (Lanzi et al., 2012)), *IL2RA* (OMIM #606367; recurrent bacterial infections, recurrent viral infections, and recurrent fungal infections (Sharfe et al., 1997)), and *TBK1* (OMIM #617900; herpes simplex encephalitis (HSE), acute infection, and episodic HSE (Herman et al., 2012)). These examples show that the identified genes may also confer predisposition, with near-complete penetrance, to an infectious disease related trait displaying true Mendelian segregation.

Enrichment analysis of 64 of the 70 ID-associated genes with nominal support for associations with other clinical ID traits identified modulation of the actin cytoskeleton as a potential shared mechanism of host susceptibility to infection (Figure 4). While manipulation of the actin cytoskeleton by pathogens is hardly a new concept, our study identified specific host genetic variation in actin regulatory genes that is potentially causative of clinical ID manifestations. In addition to pathogen interaction with the cytoskeletal transport machinery, efficient exploitation of host gene expression program is crucial for successful invasion and colonization, and here we mapped several pathogenicity-relevant targets. Notably, we observed a significant enrichment for a highly conserved sequence motif, within 4 kb of a multi-ID-associated gene's TSS, that is not a known transcription factor binding site. The motif's presence near multi-ID associated genes suggests a broad regulatory role in host-pathogen interaction, involving the diversity of pathogens examined here, towards successful reprogramming of host gene expression. Furthermore, we identified a significant enrichment for phosphorylated host proteins, suggesting the value of global phosphoproteomic profiling, which has recently been used to prioritize pharmacological targets for the novel SARS-CoV-2 virus (Bouhaddou et al., 2020). These data provide several potential avenues by which host

susceptibility can be breached by a pathogen's requirement to maintain a niche through manipulation of host cellular machinery.

To obtain additional support for our gene-level associations, we leveraged two genomic resources with rich phenotypic information (UK Biobank (Bycroft et al., 2018) and FinnGen (Locke et al., 2019)). These data will prove increasingly useful to characterizing the genetic basis of the ID-associated adverse outcomes and complications. Despite the caveats for the use of EHR in genetic analyses of ID traits (Ko and Urban, 2013; Power et al., 2017), the growing availability of such independent datasets will facilitate identification of robust genetic associations. Perhaps more importantly, the breadth of clinical phenotypes in these EHR datasets should enable identification of associated adverse outcomes and complications for the ID-associated genes.

The primary challenges in conducting GWAS of ID traits include phenotype definition and case-control misclassification. Obstacles to accurate phenotype definition include the requirement of specialized laboratory testing to identify specific pathogens and administration of prophylactic therapeutics complicating identification of potentially causative pathogens. Seropositivity may result from the complex genetic properties of the pathogen and the particular mechanisms governing host-pathogen interaction. However, seropositivity may not indicate clinical manifestations of the disease. On the other hand, seronegativity may imply lack of exposure to the pathogen, the absence of infection even in the presence of exposure, or host resistance to infection. Anchoring the analysis to host genetic information (as in our use of genetically-determined expression) and replication of discovered associations may address some aspects of this challenge. Here we exploit an extensive resource of culture data (for identification of pathogens from clinical specimens) linked to whole-genome genetic information to provide additional support to our gene-level associations. One of the ubiquitous problems in diagnosis is that culture recovery is often for multiple organisms, or a contaminant not relevant to the actual pathogen. Similarly, molecular diagnostics of pathogen identification is often a

curation of multiple statistically relevant putative pathogens. The mapping of pathogen genome identification to transcriptional response (molecular seropositivity) is a valuable validation of a finding that a given pathogen is associated with a particular infectious syndrome, and our approach to identification of genetically determined expression changes may facilitate this mapping. Future studies may also implement more complex GWAS models, including incorporating the pathogen genome.

Our catalog of TWAS associations with microbiome composition may facilitate insights into molecular mechanisms of infectious disease risk and complications, inform studies of host-pathogen interactions, and improve anti-microbial pharmacologic strategies. Improved characterization of pathogen colonization and taxonomic classification at species and strain level through 16S rRNA sequencing-based approaches may lead to greater resolution of causative infectious processes. Disruption of pathogen equilibrium in the microbiome by environmental or genetic variation may determine susceptibility to human disease (Goodrich et al., 2016; Hall et al., 2017). However, critical challenges to understanding the patterns of host colonization include identification of rare pathogen populations as well as environmental pressures (i.e. medication use, dietary alterations, etc.) acutely altering the microbiome landscape (Kurilshikov et al., 2017). Thus, linking microbiome traits to host genetic variation promises to improve resolution of causative mechanisms for ID traits and potentially adverse outcomes.

Mendelian Randomization provides a framework to perform causal interference on the effect of the exposure on the outcome (Davey Smith and Hemani, 2014; Lawlor et al., 2008). We leveraged a summary statistics based approach to test the causal effect of an ID trait on potential adverse outcomes, using genetic instruments. Mendelian Randomization requires three assumptions: 1) the genetic instrument is associated with exposure (i.e., ID trait); 2) the genetic instrument is associated with the outcome (i.e., adverse outcome or complication) only through the exposure of interest; and 3) the genetic instrument is affecting the outcome

independent of other factors (i.e., confounders). Violations of these assumptions can have critical implications for the interpretation of the results. Thus, several approaches have been developed that are robust to these violations. In the case of ID traits, a methodology that distinguishes causality from comorbidity is critical. While many phenotypes are highly comorbid and suspected to have a causal relationship (e.g., smoking and depression/anxiety), Mendelian Randomization does not necessarily support the causal hypothesis (Taylor et al., 2014). Furthermore, since RCTs cannot be ethically conducted for ID traits and adverse outcomes, the methodology offers an approach for elucidating the role of an infection phenotype or pathogen exposure in disease causation using an observational study design. Here, we found strong causal support for the effect of certain clinical ID traits on potential adverse complications identified through a phenome scan of the ID-associated genes: 1) meningitis - cerebral edema and compression of brain; and 2) Gram-negative sepsis - acidosis. These data indicate that genetic risk factors for select adverse outcomes and complications exert their phenotypic effect through the relevant ID traits.

To enable investigations into mediating cellular and molecular traits for the ID-associated genes, we provide a functional genomics resource built on a high-throughput *in vitro* pathogen infection screen (Hi-HOST) (Wang et al., 2018). Integration of EHR data into Hi-HOST facilitates replication of gene-level associations with clinical ID traits and greatly improves the signal-to-noise ratio. This discovery and replication platform, encompassing human phenomics and cellular microbiology, provides a high-throughput approach to linking host cellular processes to clinical ID traits and adverse outcomes.

Although additional mechanistic studies are warranted, our study lays the foundation for anchoring targeted molecular studies in human genetic variation. Elucidation of host mechanisms exploited by pathogens requires multi-disciplinary approaches. Here, we show the broader role of host genetic variation, implicating diverse disease mechanisms. Our study generates a rich resource and a genetics-anchored methodology to facilitate investigations of

ID-associated clinical outcomes and complications, with important implications for the development of preventive strategies and more effective therapeutics. Causal inference on the clinical ID traits and potential complications promises to expand our understanding of the molecular basis for the link and, crucially, enable prediction and prevention of serious adverse events.

REFERENCES

(2015). Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science (New York, NY) *348*, 648-660.

Aktories, K., and Barbieri, J.T. (2005). Bacterial cytotoxins: targeting eukaryotic switches. Nat Rev Microbiol *3*, 397-410.

Ammari, M.G., Gresham, C.R., McCarthy, F.M., and Nanduri, B. (2016). HPIDB 2.0: a curated database for host-pathogen interactions. Database : the journal of biological databases and curation *2016*.

Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R. (2015). A global reference for human genetic variation. Nature *526*, 68-74.

Balsa, E., Marco, R., Perales-Clemente, E., Szklarczyk, R., Calvo, E., Landázuri, M.O., and Enríquez, J.A. (2012). NDUFA4 is a subunit of complex IV of the mammalian electron transport chain. Cell Metab *16*, 378-386.

Barbeira, A.N., Dickinson, S.P., Bonazzola, R., Zheng, J., Wheeler, H.E., Torres, J.M., Torstenson, E.S., Shah, K.P., Garcia, T., Edwards, T.L.*, et al.* (2018). Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. Nat Commun *9*, 1825.

Barbeira, A.N., Pividori, M., Zheng, J., Wheeler, H.E., Nicolae, D.L., and Im, H.K. (2019). Integrating predicted transcriptome from multiple tissues improves association detection. PLoS Genet *15*, e1007889.

Barnard, P., Payne, E., and McMillan, N.A. (2000). The human papillomavirus E7 protein is able to inhibit the antiviral and anti-growth functions of interferon-alpha. Virology *277*, 411-419.

Bastarache, L., Hughey, J.J., Hebbring, S., Marlo, J., Zhao, W., Ho, W.T., Van Driest, S.L., McGregor, T.L., Mosley, J.D., Wells, Q.S.*, et al.* (2018). Phenotype risk scores identify patients with unrecognized Mendelian disease patterns. Science (New York, NY) *359*, 1233-1239.

Battle, A., Brown, C.D., Engelhardt, B.E., and Montgomery, S.B. (2017). Genetic effects on gene expression across human tissues. Nature *550*, 204-213.

Bouhaddou, M., Memon, D., Meyer, B., White, K.M., Rezelj, V.V., Marrero, M.C., Polacco, B.J., Melnyk, J.E., Ulferts, S., Kaake, R.M.*, et al.* (2020). The Global Phosphorylation Landscape of SARS-CoV-2 Infection. Cell.

Bowden, J., Davey Smith, G., and Burgess, S. (2015). Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. Int J Epidemiol *44*, 512-525.

Bowden, J., Davey Smith, G., Haycock, P.C., and Burgess, S. (2016). Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. Genetic epidemiology *40*, 304-314.

Brouwer, W.P., Chan, H.L., Lampertico, P., Hou, J., Tangkijvanich, P., Reesink, H.W., Zhang, W., Mangia, A., Tanwandee, T., Montalto, G.*, et al.* (2019). Genome Wide Association Study Identifies Genetic Variants Associated With Early And Sustained Response To (Peg)Interferon In Chronic Hepatitis B Patients: The GIANT-B Study. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America.

Bulik-Sullivan, B., Finucane, H.K., Anttila, V., Gusev, A., Day, F.R., Loh, P.R., Duncan, L., Perry, J.R., Patterson, N., Robinson, E.B.*, et al.* (2015). An atlas of genetic correlations across human diseases and traits. Nat Genet *47*, 1236-1241.

Burgner, D., Jamieson, S.E., and Blackwell, J.M. (2006). Genetic susceptibility to infectious diseases: big is beautiful, but will bigger be even better? Lancet Infect Dis *6*, 653-663.

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*, 203-209.

Casanova, J.L. (2015a). Human genetic basis of interindividual variability in the course of infection. Proc Natl Acad Sci U S A *112*, E7118-7127.

Casanova, J.L. (2015b). Severe infectious diseases of childhood as monogenic inborn errors of immunity. Proc Natl Acad Sci U S A *112*, E7128-7137.

Chellappan, S., Kraus, V.B., Kroger, B., Munger, K., Howley, P.M., Phelps, W.C., and Nevins, J.R. (1992). Adenovirus E1A, simian virus 40 tumor antigen, and human papillomavirus E7 protein share the capacity to disrupt the interaction between transcription factor E2F and the retinoblastoma gene product. Proc Natl Acad Sci U S A *89*, 4549-4553.

Consortium, G. (2013). The Genotype-Tissue Expression (GTEx) project. Nat Genet *45*, 580-585.

Cotsapas, C., Voight, B.F., Rossin, E., Lage, K., Neale, B.M., Wallace, C., Abecasis, G.R., Barrett, J.C., Behrens, T., Cho, J.*, et al.* (2011). Pervasive sharing of genetic effects in autoimmune disease. PLoS Genet *7*, e1002254.

Datsenko, K.A., and Wanner, B.L. (2000). One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. Proc Natl Acad Sci U S A *97*, 6640-6645.

Davey Smith, G., and Hemani, G. (2014). Mendelian randomization: genetic anchors for causal inference in epidemiological studies. Human molecular genetics *23*, R89-98.

Davis, Z.H., Verschueren, E., Jang, G.M., Kleffman, K., Johnson, J.R., Park, J., Von Dollen, J., Maher, M.C., Johnson, T., Newton, W.*, et al.* (2015). Global mapping of herpesvirus-host protein complexes reveals a transcription strategy for late genes. Molecular cell *57*, 349-360.

de Bakker, P.I., and Telenti, A. (2010). Infectious diseases not immune to genome-wide association. Nat Genet *42*, 731-732.

Denny, J.C., Bastarache, L., Ritchie, M.D., Carroll, R.J., Zink, R., Mosley, J.D., Field, J.R., Pulley, J.M., Ramirez, A.H., Bowton, E.*, et al.* (2013). Systematic comparison of phenome-wide

association study of electronic medical record data and genome-wide association study data. Nature biotechnology *31*, 1102-1110.

Denny, J.C., Ritchie, M.D., Basford, M.A., Pulley, J.M., Bastarache, L., Brown-Gentry, K., Wang, D., Masys, D.R., Roden, D.M., and Crawford, D.C. (2010). PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics *26*, 1205-1210.

Derks, E.M., Zwinderman, A.H., and Gamazon, E.R. (2017). The Relation Between Inflation in Type-I and Type-II Error Rate and Population Divergence in Genome-Wide Association Analysis of Multi-Ethnic Populations. Behavior genetics *47*, 360-368.

Domingo, P., Rodriguez, P., Lopez-Contreras, J., Rebasa, P., Mota, S., and Matias-Guiu, X. (1996). Spontaneous rupture of the spleen associated with pneumonia. European journal of clinical microbiology & infectious diseases : official publication of the European Society of Clinical Microbiology *15*, 733-736.

Duro, D., Schulze, A., Vogt, B., Bartek, J., Mittnacht, S., and Jansen, D.r.P. (1999). Activation of cyclin A gene expression by the cyclin encoded by human herpesvirus-8. J Gen Virol *80 (Pt 3)*, 549-555.

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. Ann Statist *7*, 1-26.

El-Bacha, T., and Da Poian, A.T. (2013). Virus-induced changes in mitochondrial bioenergetics as potential targets for therapy. The international journal of biochemistry & cell biology *45*, 41-46.

Gamazon, E.R., Segre, A.V., van de Bunt, M., Wen, X., Xi, H.S., Hormozdiari, F., Ongen, H., Konkashbaev, A., Derks, E.M., Aguet, F.*, et al.* (2018). Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. Nat Genet *50*, 956-967.

Gamazon, E.R., Wheeler, H.E., Shah, K.P., Mozaffari, S.V., Aquino-Michaels, K., Carroll, R.J., Eyler, A.E., Denny, J.C., Nicolae, D.L., Cox, N.J.*, et al.* (2015). A gene-based association method for mapping traits using reference transcriptome data. Nat Genet *47*, 1091-1098.

Gamazon, E.R., Zwinderman, A.H., Cox, N.J., Denys, D., and Derks, E.M. (2019). Multi-tissue transcriptome analyses identify genetic mechanisms underlying neuropsychiatric traits. Nat Genet *51*, 933-940.

García-Montero, M., Rodríguez-García, J.L., Calvo, P., González, J.M., Fernández-Garrido, M., Loza, E., and Serrano, M. (1995). Pneumonia caused by Listeria monocytogenes. Respiration *62*, 107-109.

Gerstein, A.R., Riegel, N., and Dennis, M. (1967). Ruptured Spleen Simulating Pneumonia. JAMA *199*, 589-589.

Goldstein, E.J., Citron, D.M., Peraino, V.A., and Cross, S.A. (2003). Desulfovibrio desulfuricans bacteremia and review of human Desulfovibrio infections. Journal of clinical microbiology *41*, 2752-2754.

Gómez-Moreno, J., Moar, C., Román, F., Pérez-Maestu, R., and López de Letona, J.M. (2000). Salmonella endocarditis presenting as cerebral hemorrhage. Eur J Intern Med *11*, 96-97.

Goodrich, J.K., Davenport, E.R., Beaumont, M., Jackson, M.A., Knight, R., Ober, C., Spector, T.D., Bell, J.T., Clark, A.G., and Ley, R.E. (2016). Genetic Determinants of the Gut Microbiome in UK Twins. Cell Host Microbe *19*, 731-743.

Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B.W.J.H., Jansen, R., de Geus, E.J.C., Boomsma, D.I., Wright, F.A.*, et al.* (2016). Integrative approaches for large-scale transcriptome-wide association studies. Nature genetics *48*, 245-252.

Hagiwara, K., Kikuchi, T., Endo, Y., Huqun, Usui, K., Takahashi, M., Shibata, N., Kusakabe, T., Xin, H., Hoshi, S.*, et al.* (2003). Mouse SWAM1 and SWAM2 are antibacterial proteins composed of a single whey acidic protein motif. J Immunol *170*, 1973-1979.

Hall, A.B., Tolonen, A.C., and Xavier, R.J. (2017). Human genetic variation and the gut microbiome in disease. Nat Rev Genet *18*, 690-699.

Hay, A.J., Wolstenholme, A.J., Skehel, J.J., and Smith, M.H. (1985). The molecular basis of the specific anti-influenza action of amantadine. Embo j *4*, 3021-3024.

Herman, M., Ciancanelli, M., Ou, Y.H., Lorenzo, L., Klaudel-Dreszler, M., Pauwels, E., Sancho-Shimizu, V., Pérez de Diego, R., Abhyankar, A., Israelsson, E.*, et al.* (2012). Heterozygous TBK1 mutations impair TLR3 immunity and underlie herpes simplex encephalitis of childhood. The Journal of experimental medicine *209*, 1567-1582.

Herndon, C.N., and Jennings, R.G. (1951). A twin-family study of susceptibility to poliomyelitis. Am J Hum Genet *3*, 17-46.

Hohler, T., Reuss, E., Evers, N., Dietrich, E., Rittner, C., Freitag, C.M., Vollmar, J., Schneider, P.M., and Fimmers, R. (2002). Differential genetic determination of immune responsiveness to hepatitis B surface antigen and to hepatitis A virus: a vaccination study in twins. Lancet *360*, 991-995.

Huang da, W., Sherman, B.T., and Lempicki, R.A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc *4*, 44-57.

Hughes, D.A., Bacigalupe, R., Wang, J., Rühlemann, M.C., Tito, R.Y., Falony, G., Joossens, M., Vieira-Silva, S., Henckaerts, L., Rymenans, L.*, et al.* (2020). Genome-wide associations of human gut microbiome variation and implications for causal inference analyses. Nature microbiology *5*, 1079-1087.

Isaac, R., Goldstein, I., Furth, N., Zilber, N., Streim, S., Boura-Halfon, S., Elhanany, E., Rotter, V., Oren, M., and Zick, Y. (2017). TM7SF3, a novel p53-regulated homeostatic factor, attenuates cellular stress and the subsequent induction of the unfolded protein response. Cell Death Differ *24*, 132-143.

Jordan, D.M., Verbanck, M., and Do, R. (2019). HOPS: a quantitative score reveals pervasive horizontal pleiotropy in human genetic variation is driven by extreme polygenicity of human traits and diseases. Genome Biol *20*, 222.

Ko, D.C., Gamazon, E.R., Shukla, K.P., Pfuetzner, R.A., Whittington, D., Holden, T.D., Brittnacher, M.J., Fong, C., Radey, M., Ogohara, C.*, et al.* (2012). Functional genetic screen of human diversity reveals that a methionine salvage enzyme regulates inflammatory cell death. Proc Natl Acad Sci U S A *109*, E2343-2352.

Ko, D.C., Shukla, K.P., Fong, C., Wasnick, M., Brittnacher, M.J., Wurfel, M.M., Holden, T.D., O'Keefe, G.E., Van Yserloo, B., Akey, J.M.*, et al.* (2009). A genome-wide in vitro bacterial-infection screen reveals human variation in the host response associated with inflammatory disease. Am J Hum Genet *85*, 214-227.

Ko, D.C., and Urban, T.J. (2013). Understanding human variation in infectious disease susceptibility through clinical and cellular GWAS. PLoS Pathog *9*, e1003424.

Kurilshikov, A., Wijmenga, C., Fu, J., and Zhernakova, A. (2017). Host Genetics and Gut Microbiome: Challenges and Perspectives. Trends in immunology *38*, 633-647.

Labbé, K., and Saleh, M. (2008). Cell death in the host response to infection. Cell Death Differ *15*, 1339-1349.

Lanzi, G., Moratto, D., Vairo, D., Masneri, S., Delmonte, O., Paganini, T., Parolini, S., Tabellini, G., Mazza, C., Savoldi, G.*, et al.* (2012). A novel primary human immunodeficiency due to deficiency in the WASP-interacting protein WIP. The Journal of experimental medicine *209*, 29-34.

Lawlor, D.A., Harbord, R.M., Sterne, J.A., Timpson, N., and Davey Smith, G. (2008). Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. Stat Med *27*, 1133-1163.

Liang, X., Gupta, K., Quintero, J.R., Cernadas, M., Kobzik, L., Christou, H., Pier, G.B., Owen, C.A., and Cataltepe, S. (2019). Macrophage FABP4 is required for neutrophil recruitment and bacterial clearance in Pseudomonas aeruginosa pneumonia. Faseb j *33*, 3562-3574.

Locke, A.E., Steinberg, K.M., Chiang, C.W.K., Service, S.K., Havulinna, A.S., Stell, L., Pirinen, M., Abel, H.J., Chiang, C.C., Fulton, R.S.*, et al.* (2019). Exome sequencing of Finnish isolates enhances rare-variant association power. Nature *572*, 323-328.

Longworth, M.S., and Laimins, L.A. (2004). The binding of histone deacetylases and the integrity of zinc finger-like motifs of the E7 protein are essential for the life cycle of human papillomavirus type 31. J Virol *78*, 3533-3541.

Loubinoux, J., Bronowicki, J.P., Pereira, I.A., Mougenel, J.L., and Faou, A.E. (2002). Sulfate-reducing bacteria in human feces and their association with inflammatory bowel diseases. FEMS Microbiol Ecol *40*, 107-112.

Lyczak, J.B., Cannon, C.L., and Pier, G.B. (2002). Lung infections associated with cystic fibrosis. Clin Microbiol Rev *15*, 194-222.

Malaty, H.M., Engstrand, L., Pedersen, N.L., and Graham, D.Y. (1994). Helicobacter pylori infection: genetic and environmental influences. A study of twins. Annals of internal medicine *120*, 982-986.

Man, A., Slevin, M., Petcu, E., and Fraefel, C. (2019). The Cyclin-Dependent Kinase 5 Inhibitor Peptide Inhibits Herpes Simplex Virus Type 1 Replication. Scientific reports *9*, 1260.

Murray, L.A., Sheng, X., and Cristea, I.M. (2018). Orchestration of protein acetylation as a toggle for cellular defense and virus replication. Nat Commun *9*, 4967.

Na, Y.R., Jung, D., Gu, G.J., Jang, A.R., Suh, Y.-H., and Seok, S.H. (2015). The early synthesis of p35 and activation of CDK5 in LPS-stimulated macrophages suppresses interleukin-10 production. Science signaling *8*, ra121-ra121.

Niemöller, U.M., and Täuber, M.G. (1989). Brain edema and increased intracranial pressure in the pathophysiology of bacterial meningitis. European journal of clinical microbiology & infectious diseases : official publication of the European Society of Clinical Microbiology *8*, 109-117.

Odds, F.C., Brown, A.J., and Gow, N.A. (2004). Candida albicans genome sequence: a platform for genomics in the absence of genetics. Genome Biol *5*, 230.

Ojala, P.M., Tiainen, M., Salven, P., Veikkola, T., Castaños-Vélez, E., Sarid, R., Biberfeld, P., and Mäkelä, T.P. (1999). Kaposi's sarcoma-associated herpesvirus-encoded v-cyclin triggers apoptosis in cells with high levels of cyclin-dependent kinase 6. Cancer Res *59*, 4984-4989.

Ojala, P.M., Yamamoto, K., Castaños-Vélez, E., Biberfeld, P., Korsmeyer, S.J., and Mäkelä, T.P. (2000). The apoptotic v-cyclin-CDK6 complex phosphorylates and inactivates Bcl-2. Nature cell biology *2*, 819-825.

Pan, Y., Tian, T., Park, C.O., Lofftus, S.Y., Mei, S., Liu, X., Luo, C., O'Malley, J.T., Gehad, A., Teague, J.E*., et al.* (2017). Survival of tissue-resident memory T cells requires exogenous lipid uptake and metabolism. Nature *543*, 252-256.

Patrick, G.N., Zukerberg, L., Nikolic, M., de la Monte, S., Dikkes, P., and Tsai, L.H. (1999). Conversion of p35 to p25 deregulates Cdk5 activity and promotes neurodegeneration. Nature *402*, 615-622.

Pfänder, P., Fidan, M., Burret, U., Lipinski, L., and Vettorazzi, S. (2019). Cdk5 Deletion Enhances the Anti-inflammatory Potential of GC-Mediated GR Activation During Inflammation. Frontiers in Immunology *10*.

Phelps, W.C., Yee, C.L., Münger, K., and Howley, P.M. (1988). The human papillomavirus type 16 E7 gene encodes transactivation and transformation functions similar to those of adenovirus E1A. Cell *53*, 539-547.

Power, R.A., Parkhill, J., and de Oliveira, T. (2017). Microbial genome-wide association studies: lessons from human GWAS. Nature Reviews Genetics *18*, 41-50.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet *38*, 904-909.

Pujol, C., and Bliska, J.B. (2003). The ability to replicate in macrophages is conserved between Yersinia pestis and Yersinia pseudotuberculosis. Infect Immun *71*, 5892-5899.

Revill, K., Wang, T., Lachenmayer, A., Kojima, K., Harrington, A., Li, J., Hoshida, Y., Llovet, J.M., and Powers, S. (2013). Genome-wide methylation analysis and epigenetic unmasking identify tumor suppressor genes in hepatocellular carcinoma. Gastroenterology *145*, 1424-1435.e1421-1425.

Roden, D.M., Pulley, J.M., Basford, M.A., Bernard, G.R., Clayton, E.W., Balser, J.R., and Masys, D.R. (2008). Development of a large-scale de-identified DNA biobank to enable personalized medicine. Clinical pharmacology and therapeutics *84*, 362-369.

Saitoh, T., Fujita, N., Hayashi, T., Takahara, K., Satoh, T., Lee, H., Matsunaga, K., Kageyama, S., Omori, H., Noda, T.*, et al.* (2009). Atg9a controls dsDNA-driven dynamic translocation of STING and the innate immune response. Proc Natl Acad Sci U S A *106*, 20842-20846.

Saka, H.A., Thompson, J.W., Chen, Y.S., Kumar, Y., Dubois, L.G., Moseley, M.A., and Valdivia, R.H. (2011). Quantitative proteomics reveals metabolic and pathogenic properties of Chlamydia trachomatis developmental forms. Molecular microbiology *82*, 1185-1203.

Sharfe, N., Dadi, H.K., Shahar, M., and Roifman, C.M. (1997). Human immune disorder arising from mutation of the alpha chain of the interleukin-2 receptor. Proc Natl Acad Sci U S A *94*, 3168-3171.

Shi, H., Kichaev, G., and Pasaniuc, B. (2016). Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data. Am J Hum Genet *99*, 139-153.

So, H.C., Chau, C.K., Chiu, W.T., Ho, K.S., Lo, C.P., Yim, S.H., and Sham, P.C. (2017). Analysis of genome-wide association data highlights candidates for drug repositioning in psychiatry. Nat Neurosci *20*, 1342-1349.

Soderholm, S., Kainov, D.E., Ohman, T., Denisova, O.V., Schepens, B., Kulesskiy, E., Imanishi, S.Y., Corthals, G., Hintsanen, P., Aittokallio, T.*, et al.* (2016). Phosphoproteomics to Characterize Host Response During Influenza A Virus Infection of Human Macrophages. Molecular & cellular proteomics : MCP *15*, 3203-3219.

Stahl, J.A., Chavan, S.S., Sifford, J.M., MacLeod, V., Voth, D.E., Edmondson, R.D., and Forrest, J.C. (2013). Phosphoproteomic analyses reveal signaling pathways that facilitate lytic gammaherpesvirus replication. PLoS Pathog *9*, e1003583.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S.*, et al.* (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A *102*, 15545-15550.

Suetrong, B., and Walley, K.R. (2016). Lactic Acidosis in Sepsis: It's Not All Anaerobic: Implications for Diagnosis and Management. Chest *149*, 252-261.

Sun, W., Li, J., Jiang, H.G., Ge, L.P., and Wang, Y. (2014). Diagnostic value of MUC1 and EpCAM mRNA as tumor markers in differentiating benign from malignant pleural effusion. QJM : monthly journal of the Association of Physicians *107*, 1001-1007.

Tarricone, C., Dhavan, R., Peng, J., Areces, L.B., Tsai, L.H., and Musacchio, A. (2001). Structure and regulation of the CDK5-p25(nck5a) complex. Molecular cell *8*, 657-669.

Tartey, S., Matsushita, K., Imamura, T., Wakabayashi, A., Ori, D., Mino, T., and Takeuchi, O. (2015). Essential Function for the Nuclear Protein Akirin2 in B Cell Activation and Humoral Immune Responses. J Immunol *195*, 519-527.

Tartey, S., Matsushita, K., Vandenbon, A., Ori, D., Imamura, T., Mino, T., Standley, D.M., Hoffmann, J.A., Reichhart, J.M., Akira, S.*, et al.* (2014). Akirin2 is critical for inducing inflammatory genes by bridging IkappaB-zeta and the SWI/SNF complex. Embo j *33*, 2332-2348.

Taylor, A.E., Fluharty, M.E., Bjørngaard, J.H., Gabrielsen, M.E., Skorpen, F., Marioni, R.E., Campbell, A., Engmann, J., Mirza, S.S., Loukola, A.*, et al.* (2014). Investigating the possible causal association of smoking with depression and anxiety using Mendelian randomisation meta-analysis: the CARTA consortium. BMJ Open *4*, e006141.

Taylor, M.P., Koyuncu, O.O., and Enquist, L.W. (2011). Subversion of the actin cytoskeleton during viral infection. Nat Rev Microbiol *9*, 427-439.

Tian, C., Hromatka, B.S., Kiefer, A.K., Eriksson, N., Noble, S.M., Tung, J.Y., and Hinds, D.A. (2017). Genome-wide association and HLA region fine-mapping studies identify susceptibility loci for multiple common infections. Nat Commun *8*, 599.

Tsuboi, S., and Meerloo, J. (2007). Wiskott-Aldrich syndrome protein is a key regulator of the phagocytic cup formation in macrophages. The Journal of biological chemistry *282*, 34194-34203.

Unlu, G., Qi, X., Gamazon, E.R., Melville, D.B., Patel, N., Rushing, A.R., Hashem, M., Al-Faifi, A., Chen, R., Li, B.*, et al.* (2020). Phenome-based approach identifies RIC1-linked Mendelian syndrome through zebrafish models, biobank associations and clinical studies. Nat Med *26*, 98-109.

Walenna, N.F., Kurihara, Y., Chou, B., Ishii, K., Soejima, T., Itoh, R., Shimizu, A., Ichinohe, T., and Hiromatsu, K. (2018). Chlamydia pneumoniae exploits adipocyte lipid chaperone FABP4 to facilitate fat mobilization and intracellular growth in murine adipocytes. Biochem Biophys Res Commun *495*, 353-359.

Wang, L., Pittman, K.J., Barker, J.R., Salinas, R.E., Stanaway, I.B., Williams, G.D., Carroll, R.J., Balmat, T., Ingham, A., Gopalakrishnan, A.M.*, et al.* (2018). An Atlas of Genetic Variation Linking Pathogen-Induced Cellular Traits to Human Disease. Cell Host Microbe *24*, 308-323.e306.

Watanabe, K., Stringer, S., Frei, O., Umicevic Mirkov, M., de Leeuw, C., Polderman, T.J.C., van der Sluis, S., Andreassen, O.A., Neale, B.M., and Posthuma, D. (2019). A global overview of pleiotropy and genetic architecture in complex traits. Nat Genet.

Wei, W.-Q., Bastarache, L.A., Carroll, R.J., Marlo, J.E., Osterman, T.J., Gamazon, E.R., Cox, N.J., Roden, D.M., and Denny, J.C. (2017). Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. PloS one *12*, e0175508-e0175508.

Willis, K.L., Patel, S., Xiang, Y., and Shisler, J.L. (2009). The effect of the vaccinia K1 protein on the PKR-eIF2alpha pathway in RK13 and HeLa cells. Virology *394*, 73-81.

Wolfe, D., Dudek, S., Ritchie, M., and Pendergrass, S. (2013). Visualizing genomic information across chromosomes with PhenoGram. BioData Mining *6*, 18 - 18.

Xie, X., Lu, J., Kulbokas, E.J., Golub, T.R., Mootha, V., Lindblad-Toh, K., Lander, E.S., and Kellis, M. (2005). Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. Nature *434*, 338-345.

Yavorska, O.O., and Burgess, S. (2017). MendelianRandomization: an R package for performing Mendelian randomization analyses using summarized data. Int J Epidemiol *46*, 1734-1739.

Yu, B., Cheng, H.C., Brautigam, C.A., Tomchick, D.R., and Rosen, M.K. (2011). Mechanism of actin filament nucleation by the bacterial effector VopL. Nat Struct Mol Biol *18*, 1068-1074.

Yu, Z., Song, H., Jia, M., Zhang, J., Wang, W., Li, Q., Zhang, L., and Zhao, W. (2017). USP1-UAF1 deubiquitinase complex stabilizes TBK1 and enhances antiviral responses. The Journal of experimental medicine *214*, 3553-3563.

Zahm, J.A., Padrick, S.B., Chen, Z., Pak, C.W., Yunus, A.A., Henry, L., Tomchick, D.R., Chen, Z., and Rosen, M.K. (2013). The bacterial effector VopL organizes actin into filament-like structures. Cell *155*, 423-434.

Zhang, X., Bogunovic, D., Payelle-Brogard, B., Francois-Newton, V., Speer, S.D., Yuan, C., Volpi, S., Li, Z., Sanal, O., Mansouri, D.*, et al.* (2015). Human intracellular ISG15 prevents interferon-alpha/beta over-amplification and auto-inflammation. Nature *517*, 89-93.

Zheng, T., Liu, W., Oh, S.Y., Zhu, Z., Hu, B., Homer, R.J., Cohn, L., Grusby, M.J., and Elias, J.A. (2008). IL-13 receptor alpha2 selectively inhibits IL-13-induced responses in the murine lung. J Immunol *180*, 522-529.

AUTHOR CONTRIBUTIONS

Conceptualization, A.T.H. and E.R.G.; Methodology, A.T.H., D.Z., R.L.S., L.B., S.J.S., D.C.K., and E.R.G.; Investigation, A.T.H., D.Z., R.L.S., L.B., L.W., S.S.Z., S.J.S., D.C.K., and E.R.G. Writing – Original Draft, A.T.H. and E.R.G.; Writing – Review and Editing, A.T.H., D.Z., R.L.S., L.B., L.W., S.S.Z., S.J.S., D.C.K., and E.R.G, Funding Acquisition – A.T.H. and E.R.G., Supervision, E.R.G.

ACKNOWLEDGEMENTS

A.T.H. is supported by the National Institutes of Health (F30HL143826) and Vanderbilt University Medical Scientist Training Program (T32GM007347). E.R.G. is supported by the National Human Genome Research Institute of the National Institutes of Health under Award Numbers R35HG010718 and R01HG011138. E.R.G. and S.S.Z. are funded by the National Heart, Lung, & Blood Institute of the National Institutes of Health under Award Number R01HL133559. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. E.R.G. has also significantly benefitted from a Fellowship at Clare Hall, University of Cambridge (UK) and is grateful to the President and Fellows of the college for a stimulating intellectual home. Genomic data are also supported by individual investigator-led projects including U01-HG004798, R01-NS032830, RC2-GM092618, P50-GM115305, U01-HG006378, U19-HL065962, and R01-HD074711. Additional funding sources for BioVU are listed at https://victr.vanderbilt.edu/pub/biovu/. L.B. is supported by R01-LM010685. S.J.S. is supported by an NIH Director's Pioneer and Transformative Awards DP1-HD086071 and R01-AI145057. D.C.K. is supported by R01-AI118903, R21-AI144586, and R21-AI146520. D.C.K. and L.W. are supported by R21-AI133305.

DECLARATION OF INTERESTS

E.R.G. receives an honorarium from the journal *Circulation Research* of the American Heart Association, as a member of the Editorial Board. The other authors declare no competing interests.

FIGURE LEGENDS

Figure 1. Overview of ID atlas resource. List of ID traits tested with corresponding Phecode (phewascatalog.org) in parentheses.

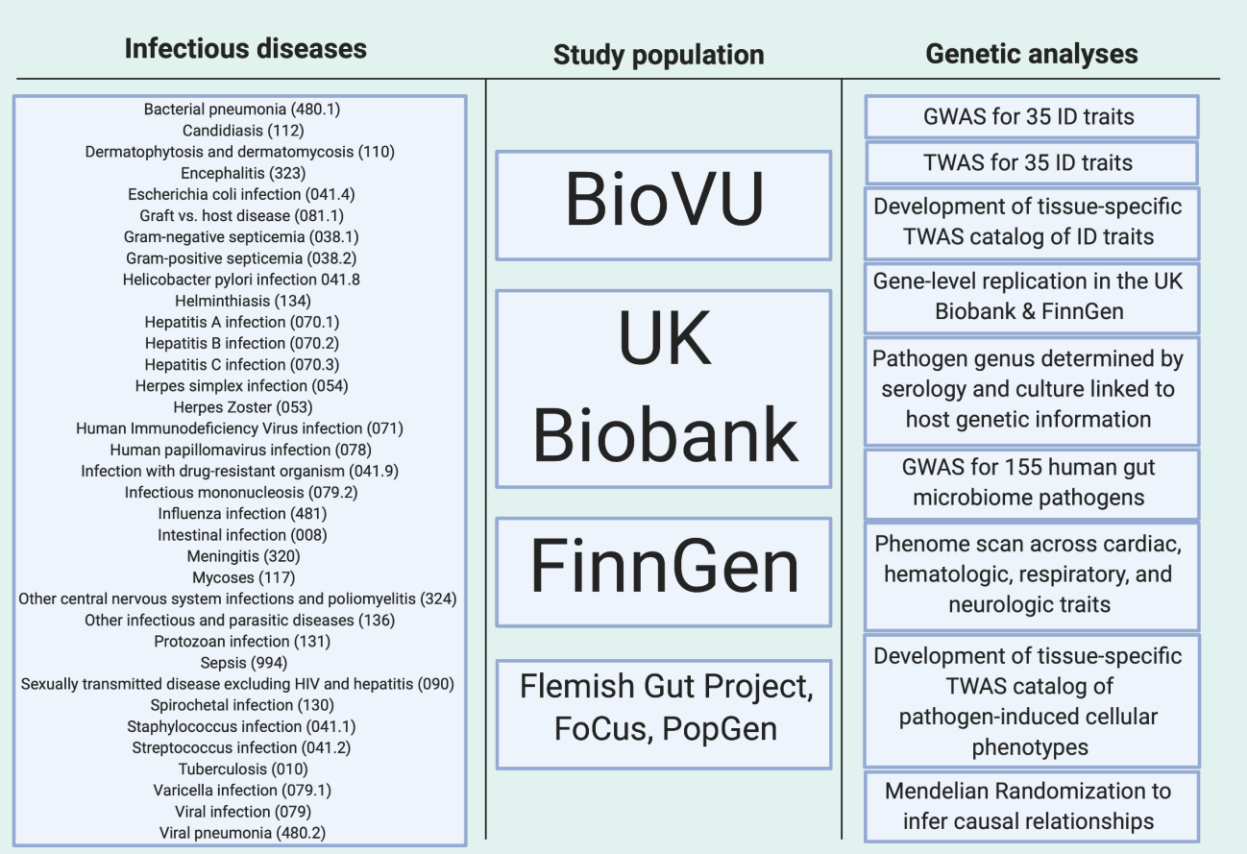


Figure 2. Genome-wide association study (GWAS) of ID traits. (A) Threshold for inclusion of SNP associations was set at 1.0×10^{-4}. Genome-wide significance for an ID trait was set at the conventional GWAS threshold $p = 5.0 \times 10^{-8}$, as indicated by the horizontal dotted line. The subset of 13 ID traits (among the full set tested) with variants that meet the traditional genome-wide significance threshold are included. The top variant association for each of the 13 traits is labeled. The most significant variant association is with bacterial pneumonia ($p < 1.0 \times 10^{-30}$). (B) LocusZoom plot at the sentinel variant, rs17139584, associated with bacterial pneumonia. Several variants in low LD ($r^2 < 0.20$) with the sentinel variant, including variants in the cystic fibrosis gene *CFTR* and in the *MET* gene >650 Kb upstream, are genome-wide significant for bacterial pneumonia. The sentinel variant remains statistically significant ($p = 0.007$) after adjusting for a diagnosis of cystic fibrosis.

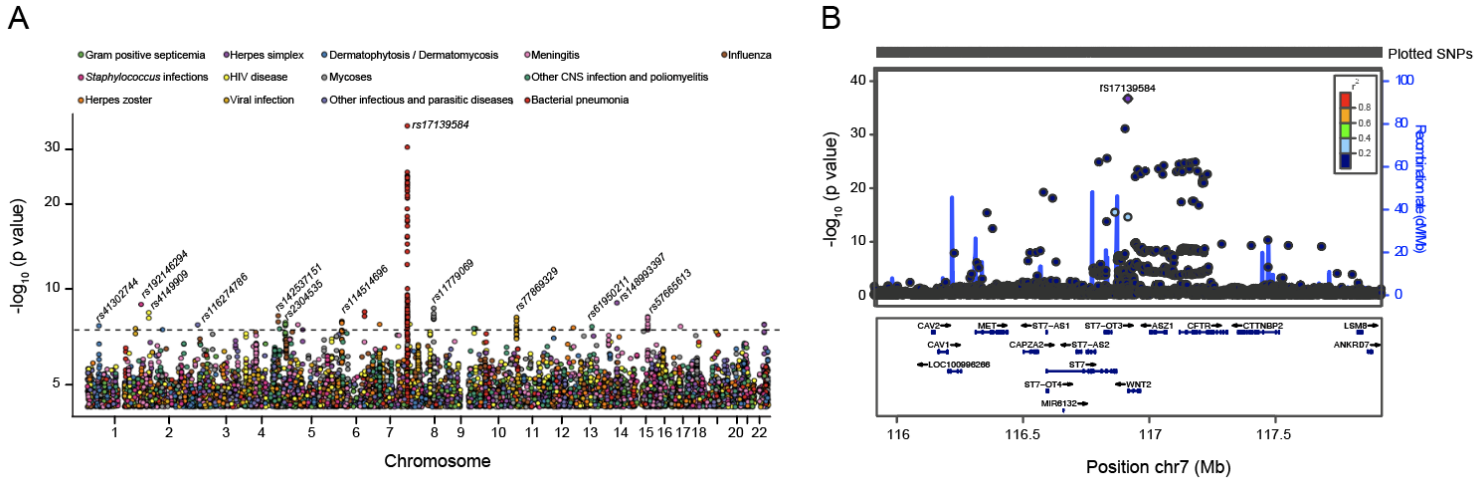
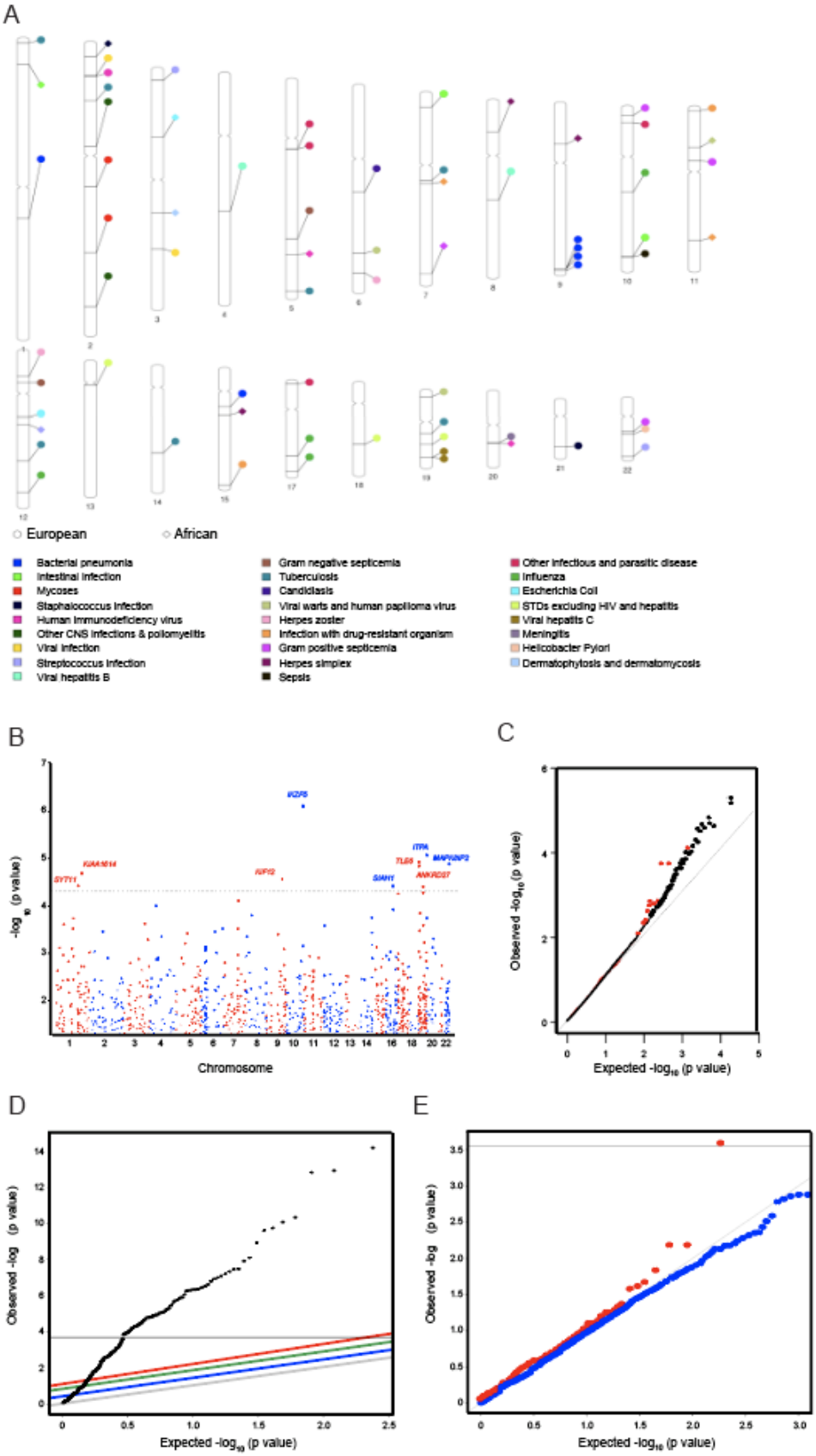


Figure 3. Transcriptome-wide association studies (TWAS) of 35 ID traits reveal novel ID-associated genes. The genetic component of gene expression for autosomal genes was individually tested for association with each of 35 ID traits (see Methods). (A) Experiment-wide or ID-specific significant genes are displayed on the ideogram using their chromosomal locations and color-coded using the associated ID traits. Most associations represent unique genes within the implicated loci, suggesting the genes are not tagging another causal gene. A locus on chromosome 9, by contrast, shows multiple associations with the same ID trait, which may indicate correlation of the expression traits with a single causal gene in the locus. (B) Manhattan plot shows the PrediXcan associations with sepsis (Phecode 994; number of cases 2,921; number of controls 22,874). Dashed line represents $p < 5x10^{-5}$. The gene *IKZF5* was significant ($p = 8.16x10^{-7}$) after Bonferroni correction for the number of genes tested. (C) Q-Q plot of FinnGen replication p-values for genes associated with intestinal infection ($p < 0.05$) in BioVU (red) compared to the remaining set of genes (black). The ID-associated genes tended to be more significant in the independent dataset than the remaining genes, as evidenced by the leftward shift in the Q-Q plot. (D) Q-Q plot of UK Biobank replication p-values for influenza in lung tissue demonstrating significant deviation from null expectation. The horizontal line corresponds to the Bonferroni threshold (adjusted $p < 0.05$). (E) Q-Q plot of the full set of FinnGen p-values from the association of gene expression in cerebellum with meningitis (blue) vs. the subset of FinnGen (replication) p-values for the top BioVU meningitis gene-level associations in cerebellum (red). The meningitis associated genes in BioVU dramatically improve signal-to-noise ratio, as shown by the leftward shift in the Q-Q plot. The horizontal line corresponds to the Bonferroni threshold (adjusted $p < 0.05$) based on the number of top BioVU genes tested.


A
European
African
Bacterial pneumonia
Intestinal infection
Mycoses
Staphylococcus infection
Human immunodeficiency virus
Other CNS infections & poliomyelitis
Viral infection
Streptococcus infection
Viral hepatitis B
Gram negative septicemia
Tuberculosis
Candidiasis
Viral warts and human papilloma virus
Herpes zoster
Infection with drug-resistant organism
Gram positive septicemia
Herpes simplex
Sepsis
Other infectious and parasitic disease
Influenza
Escherichia Coli
STDs excluding HIV and hepatitis
Viral hepatitis C
Meningitis
Helicobacter Pylori
Dermatophytosis and dermatomycosis
B
Chromosome
C
Observed -log10 (p value)
Expected -log10 (p value)
D
Observed -log (p value)
Expected -log10 (p value)
E
Observed -log (p value)
Expected -log10 (p value)

980

Figure 4. Enriched pathways across multiple ID traits and pathogen evolutionary strategies to promote infection. (A) Gene set enrichment analysis of ID associated genes having also nominal associations with additional ID traits. All gene sets satisfied false discovery rate < 0.05 for pathway enrichment and included known biological processes (e.g. protein complex formation, cytoskeletal protein binding, cell death, actin motility, etc.) relevant to the biology of infection. (B) Highly conserved motif "TCCCRNNRTGC", within 4 kb of the transcription start site (TSS) of ID-associated genes, is enriched among the multi-ID associated genes and does not match any known transcription factor binding site. Genes with this motif near the TSS include *AKIRIN2*, *CDK5*, *RAD50*, *PTCD3*, and *CERS3*. This suggests a strategy that the pathogens may broadly exploit to hijack the host transcriptional machinery. (C) CDK5 is an example of a multi-ID associated gene, significantly associated with Gram-positive septicemia and nominally associated with other IDs, including herpes simplex virus. CDK5 is activated by its regulatory subunit p35/p25. The CDK5-p25 complex regulates inflammation (whose large-scale disruption is characteristic of septicemia) and induces cytoskeletal disruption in neurons (where the herpes virus promotes lifelong latent infection). Structure of the CDK5-p25 complex (PDB: 1H4L, (Tarricone et al., 2001)) is shown here. The A and B chains are required for cytoskeletal protein binding (CDK5), whereas the D and E chains (p25) are involved in actin regulation and kinase function, all functions implicated in our pathway analysis. (D) Multi-ID associated genes identified by our study have also been observed in host-pathogen protein complexes (by coimmunoprecipitation, affinity chromatography, and two-hybrid approaches, among others) for the specific pathogens responsible for the ID traits. Interactions of pathogen proteins with CDK5 are shown here. M2_134A1 (UniProt: PO6821) is the matrix protein 2 component of the proton-selective ion channel required for influenza A viral genome release during cellular entry and is targeted by the anti-viral drug amantadine (Hay et al., 1985). VE7_HPV16 (UniProt: PO3129) is a component of human papillomavirus (HPV) required for cellular transformation and trans-activation through disassembly of E2F1 transcription factor from RB1 leading to impaired production of type I interferons (Barnard et al., 2000; Chellappan et al., 1992; Phelps et al., 1988). VE7_HPV31 (UnitProt: P17387) engages histone deacetylases 1 and 2 to promote HPV31 genome maintenance (Longworth and Laimins, 2004). VCYCL_HHV8P (UniProt: Q77Q36) is a cyclin homolog within the human herpesvirus 8 genome that has been shown to control cell cycle through CDK6 and induce apoptosis through Bcl2 (Duro et al., 1999; Ojala et al., 1999; Ojala et al., 2000). F5HC81_HHV8 (UniProt: F5HC81) is not well-characterized, but predicted to act as a viral cyclin homolog. This suggests a second strategy that the pathogens exploit, i.e., alteration of the host proteome, to promote infection.

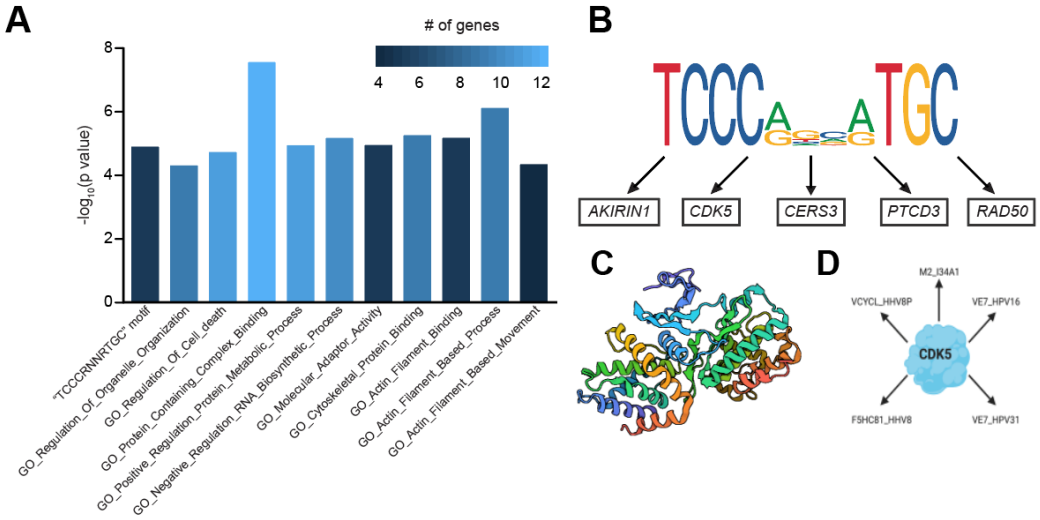
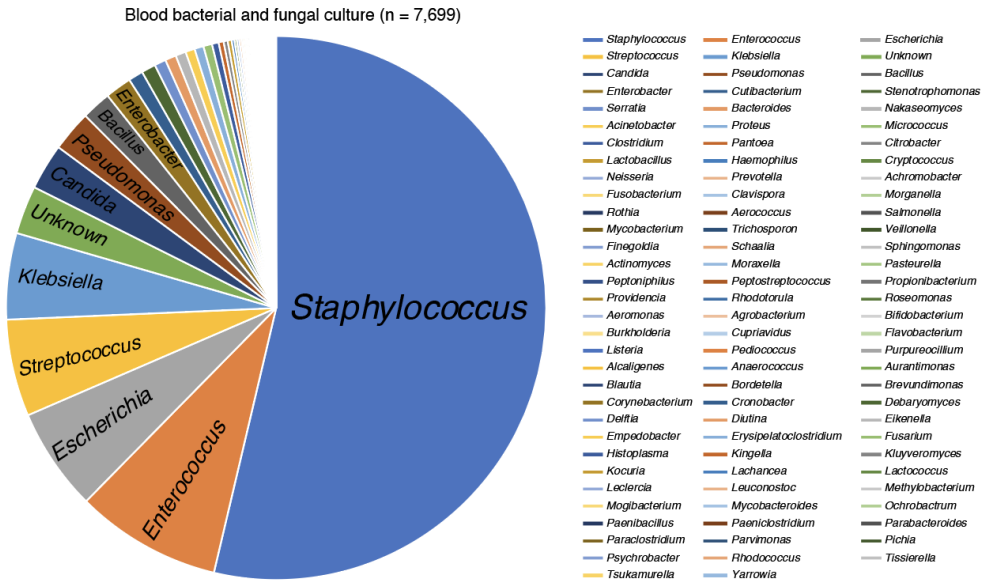


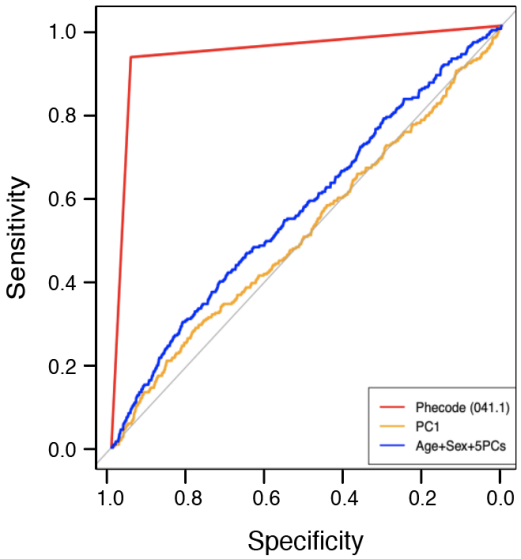
Figure 5. Pathogen genus identification from clinical blood cultures linked to whole-genome information reveals insights into host colonization and infection. (A) Bacterial and fungal pathogens identified from blood (n = 7,699 positive cultures across 94 genera) from 2,417 individuals. (B) Area under the receiver operating characteristic curve (AUC) showing that the clinical trait *Staphylococcus* infection (Phecode = 041.1) performs well in classifying *Staphylococcus aureus* infection based on blood culture data from (A), with AUC of 0.938 with standard error of 0.008. The first PC in the European ancestry samples and a model with age, sex, and the first 5 PCs, both with substantially lower performance (AUC of 0.514 and 0.568, respectively), are also shown. (C) Q-Q plot of top SNPs ($p < 0.0001$) associated with *Desulfovibrionaceae* colonization in the human gut microbiome showing the p-value distribution from the association with intestinal infection (Phecode 008) in BioVU. Note the enrichment as shown by the leftward shift relative to the diagonal line, with several genes satisfying the FDR<0.05 (red line) threshold.

1031

A


Blood bacterial and fungal culture (n = 7,699)
Staphylococcus
Enterococcus
Escherichia
Streptococcus
Klebsiella
Unknown
Candida
Pseudomonas
Bacillus
Enterobacter
Staphylococcus
Streptococcus
Candida
Enterobacter
Serratia
Acinetobacter
Clostridium
Lactobacillus
Neisseria
Fusobacterium
Rothia
Mycobacterium
Finegoldia
Actinomyces
Peptoniphilus
Providencia
Aeromonas
Burkholderia
Listeria
Alcaligenes
Blautia
Corynebacterium
Delftia
Empedobacter
Histoplasma
Kocuria
Leclercia
Mogibacterium
Paenibacillus
Paraclostridium
Psychrobacter
Tsukamurella
Enterococcus
Klebsiella
Pseudomonas
Cutibacterium
Bacteroides
Proteus
Pantoea
Haemophilus
Prevotella
Clavispora
Aerococcus
Trichosporon
Schaalia
Moraxella
Peptostreptococcus
Rhodotorula
Agrobacterium
Cupriavidus
Pediococcus
Anaerococcus
Bordetella
Cronobacter
Diutina
Erysipelatoclostridium
Kingella
Lachancea
Leuconostoc
Mycobacteroides
Paeniclostridium
Parvimonas
Rhodococcus
Yarrowia
Escherichia
Unknown
Bacillus
Stenotrophomonas
Nakaseomyces
Micrococcus
Citrobacter
Cryptococcus
Achromobacter
Morganella
Salmonella
Veillonella
Sphingomonas
Pasteurella
Propionibacterium
Roseomonas
Bifidobacterium
Flavobacterium
Purpureocillium
Aurantimonas
Brevundimonas
Debaryomyces
Eikenella
Fusarium
Kluyveromyces
Lactococcus
Methylobacterium
Ochrobactrum
Parabacteroides
Pichia
Tissierella

B


Sensitivity
Specificity
1.0
0.8
0.6
0.4
0.2
0.0
Phecode (041.1)
PC1
Age+Sex+5PCs

C

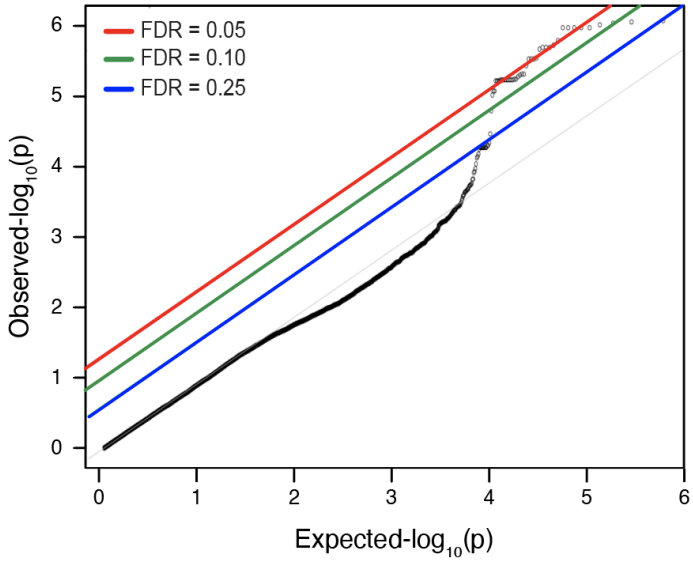

FDR = 0.05
FDR = 0.10
FDR = 0.25
Observed-log10(p)
Expected-log10(p)
0
1
2
3
4
5
6

Figure 6. Phenome-scale scan of 70 ID-associated genes across 197 cardiovascular, hematologic, neurologic, and respiratory phenotypes (cases > 200) in BioVU (phewascatalog.org) identifies genes association with both disease risk and corresponding known complications of the infection. (A) Each dot represents the association of an ID-associated gene with one of the 197 (hematologic, respiratory, cardiovascular, and neurologic) phenotypes. Horizontal red line indicates threshold for statistical significance correcting for number of phenotypes and ID-associated genes tested. We identify four gene-phenotype pairs reaching experiment-wide significance: 1) *WFDC12*, our most significant ($p = 4.23 \times 10^{-6}$) association with meningitis, is also associated with cerebral edema and compression of brain ($p = 1.35 \times 10^{-6}$), a feared clinical complication of meningitis (Niemöller and Täuber, 1989); 2) *TM7SF3*, the most significant gene with Gram-negative sepsis ($p = 1.37 \times 10^{-6}$), is also associated with acidosis ($p = 1.95 \times 10^{-6}$), a known metabolic derangement associated with severe sepsis (Suetrong and Walley, 2016); 3) *TXLNB*, the most significant gene associated with viral warts and human papillomavirus infection ($p = 4.35 \times 10^{-6}$), is also associated with abnormal involuntary movements, $p = 1.39 \times 10^{-6}$; and 4) *RAD18*, the most significant gene associated with Streptococcus infection ($p = 2.01 \times 10^{-6}$), is also associated with anemia in neoplastic disease ($p = 3.10 \times 10^{-6}$). (B) Mendelian randomization framework. P-value threshold used to define an instrumental variable was set at $p < 1.0 \times 10^{-5}$ and variants in linkage equilibrium ($r^2 = 0.01$) were used. (C) Mendelian Randomization provides strong support for causal exposure-outcome relationships for 1) meningitis and compression of brain (left, median-weighted estimator $p = 2.7 \times 10^{-3}$); and 2) gram-negative septicemia and acidosis (right, median-weighted estimator $p = 2.0 \times 10^{-7}$). Grey lines indicate 95% confidence interval.

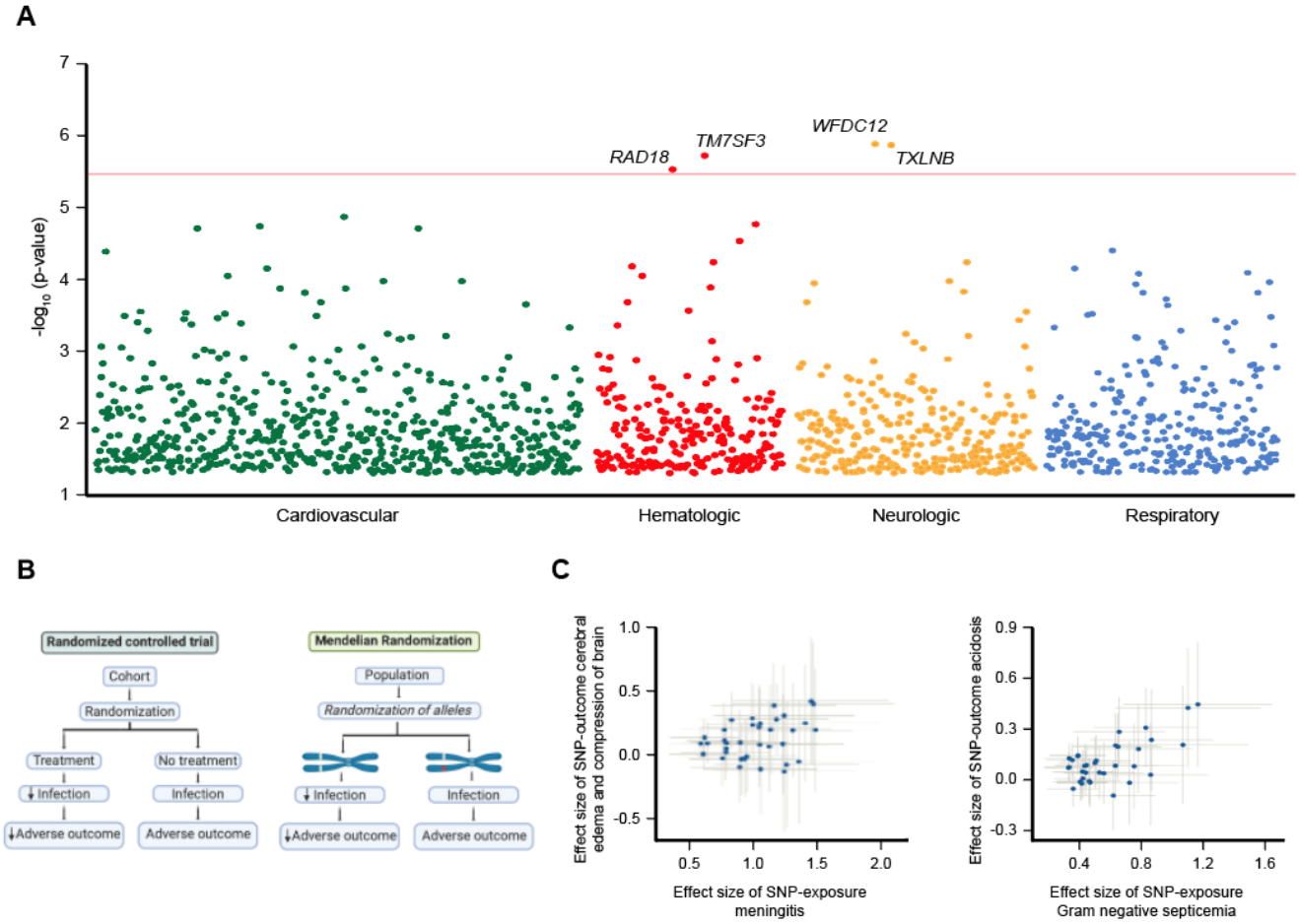


Figure 7. TWAS of 79 pathogen-exposure induced cellular traits improves identification of pathogen-induced cellular mechanisms. (A) Genes reaching significance in Hi-HOST after correction for the total number of genes and cellular phenotypes tested. (B) Integration of EHR data into Hi-HOST facilitates replication of gene-level associations with a clinical ID trait. Genes nominally associated ($p < 0.05$) with Gram-positive septicemia (Phecode 038.2) in BioVU show significant enrichment for *Staphylococcus* toxin exposure, a Hi-HOST phenotype. The Q-Q plot shows the distribution of TWAS p-values in the Hi-HOST data for the top genes in the BioVU data. False discovery rate (FDR) thresholds at 0.25 (blue), 0.10 (green), and 0.05 (red) are shown. (C) Integration of EHR data into Hi-HOST also improves the signal-to-noise ratio in Hi-HOST. For example, the top 300 genes nominally associated with *Staphylococcus* infection (Phecode 041.1) in BioVU ($p < 0.016$, red) depart from null expectation for their TWAS associations with *Staphylococcus* toxin exposure in Hi-HOST compared to the full set of genes (black).

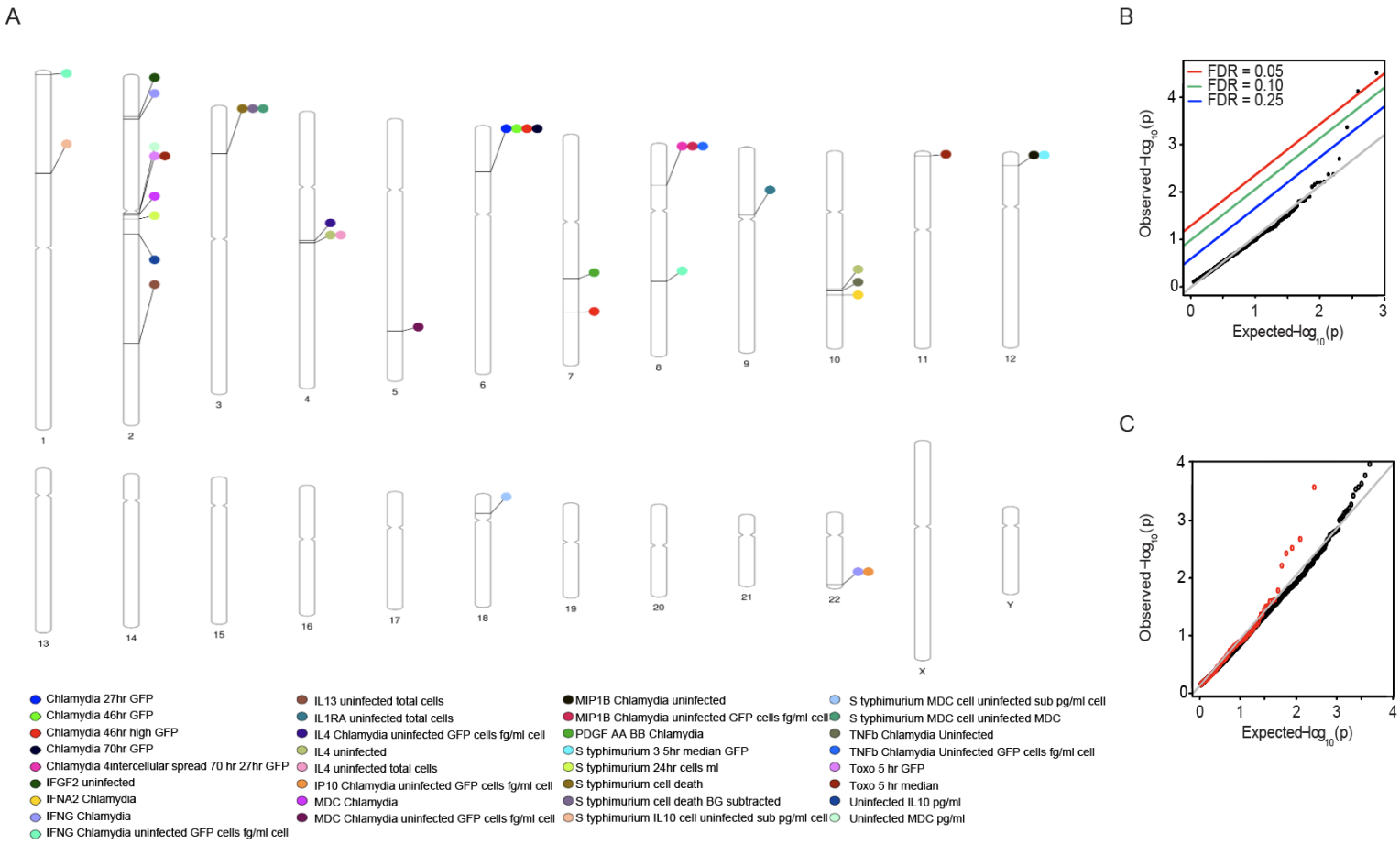


Table 1. Significant trait-specific gene-level associations with individual infectious disease phenotypes (for which number of cases > 100). Experiment-wide findings are noted in **bold**.

Gene	PheCode	Phenotype	Cases	Controls	Ancestry	Odds ratio	P value
IKZF5	994	Sepsis	2,921	22,874	European	0.91	$8.16x10^{-7}$
AKIRIN2	112	Candidiasis	2,284	21,426	European	0.91	$2.83x10^{-6}$
PSMG1	041.1	*Staphylococcus* infection	2,180	19,844	European	0.90	$3.13x10^{-6}$
AGTR1	079	Viral infection	1,811	20,904	European	1.12	$1.49x10^{-6}$
SLC35F6	079	Viral infection	1,811	20,904	European	0.89	$3.30x10^{-6}$
NDUFA4	**008**	**Intestinal infection**	**1,608**	**24,187**	**European**	**1.16**	**$1.83x10^{-9}$**
C10orf120	**008**	**Intestinal infection**	**1,608**	**24,187**	**European**	**1.13**	**$4.92x10^{-8}$**
RAD18	041.2	*Streptococcus* infection	1,262	19,844	European	1.16	$2.01x10^{-6}$
MAPK8IP2	041.2	*Streptococcus* infection	1,262	19,844	European	1.14	$3.81x10^{-6}$
AVIL	041.4	*Escherichia Coli*	1,231	19,844	European	1.16	$1.58x10^{-6}$
STAP2	078	Viral warts and human papilloma virus	1,152	20,904	European	1.15	$2.33x10^{-6}$
TXLNB	078	Viral warts and human papilloma virus	1,152	20,904	European	0.86	$4.35x10^{-6}$
SLCO1A2	053	Herpes zoster	989	20,904	European	0.93	$1.64x10^{-7}$
CLDN20	053	Herpes zoster	989	20,904	European	0.86	$4.54x10^{-6}$
IGF2	041.9	Infection with drug-resistant organism	893	19,844	European	0.83	$4.01x10^{-7}$
CERS3	041.9	Infection with drug-resistant organism	893	19,844	European	1.17	$4.18x10^{-6}$
TOR4A	**480.1**	**Bacterial pneumonia**	**862**	**18,054**	**European**	**0.84**	**$5.15x10^{-8}$**
FAM166A	480.1	Bacterial pneumonia	862	18,054	European	1.19	$1.10x10^{-7}$
C9orf173	480.1	Bacterial pneumonia	862	18,054	European	1.18	$4.48x10^{-7}$
PIP5K1A	480.1	Bacterial pneumonia	862	18,054	European	1.16	$7.00x10^{-7}$
NELFB	480.1	Bacterial pneumonia	862	18,054	European	0.86	$1.87x10^{-6}$
AVEN	480.1	Bacterial pneumonia	862	18,054	European	0.85	$3.31x10^{-6}$
TM7SF3	038.1	Gram negative septicemia	820	19,844	European	1.17	$1.37x10^{-6}$
RAD50	038.1	Gram negative septicemia	820	19,844	European	1.17	$4.50x10^{-6}$
ZNF577	070.3	Viral hepatitis C	808	20,904	European	0.84	$6.21x10^{-7}$
ZNF649	070.3	Viral hepatitis C	808	20,904	European	0.85	$1.85x10^{-6}$
SETD9	**136**	**Other infectious and parasitic diseases**	**746**	**24,770**	**European**	**0.83**	**$3.04x10^{-8}$**
AC022431.1	**136**	**Other infectious and parasitic diseases**	**746**	**24,770**	**European**	**1.20**	**$7.92x10^{-8}$**
MYO1C	136	Other infectious and parasitic diseases	746	24,770	European	1.10	$2.97x10^{-6}$
NUDT5	136	Other infectious and parasitic diseases	746	24,770	European	0.84	$3.52x10^{-6}$
MAATS1	110	Dermatophytosis and dermatomycosis	654	3,330	African	0.80	$4.82x10^{-6}$
PTPN4	**117**	**Mycoses**	**627**	**21,426**	**European**	**0.79**	**$1.56x10^{-7}$**
WIPF1	117	Mycoses	627	21,426	European	1.20	$2.72x10^{-6}$
ALX4	**038.2**	**Gram positive septicemia**	**613**	**19,844**	**European**	**1.25**	**$4.21x10^{-8}$**
C22orf31	038.2	Gram positive septicemia	613	19,844	European	0.81	$2.05x10^{-6}$
IL2RA	038.2	Gram positive septicemia	613	19,844	European	1.20	$3.88x10^{-6}$
VWA5B1	008	Intestinal infection	368	4,060	African	1.30	$3.85x10^{-6}$
ATP6V1C2	041.1	*Staphylococcus* infection	358	3,337	African	1.33	$1.51x10^{-6}$
WDR66	481	Influenza	272	18,054	European	0.71	$3.47x10^{-7}$
FAM20A	481	Influenza	272	18,054	European	077	$1.51x10^{-6}$
HKDC1	481	Influenza	272	18,054	European	1.34	$2.20x10^{-6}$
ASPSCR1	481	Influenza	272	18,054	European	1.35	$2.76x10^{-6}$
FAM208A	041.4	*Escherichia Coli*	243	3,337	African	1.42	$1.15x10^{-6}$
TBK1	041.2	*Streptococcus* infection	229	3,337	African	1.41	$4.70x10^{-6}$
DNAJC5G	071	Human immunodeficiency virus	196	20,904	European	1.08	$7.36x10^{-7}$
FABP4	070.2	Viral hepatitis B	166	20,904	European	0.70	$4.39x10^{-6}$
ANK2	070.2	Viral hepatitis B	166	20,904	European	0.77	$4.56x10^{-6}$
HIP1	041.9	Infection with drug-resistant organism	165	3,337	African	1.46	$6.74x10^{-7}$
C11orf53	041.9	Infection with drug-resistant organism	165	3,337	African	0.72	$4.98x10^{-6}$
EPCAM	**010**	**Tuberculosis**	**156**	**19,844**	**European**	**1.40**	**$5.04x10^{-8}$**
AL589739.1	**010**	**Tuberculosis**	**156**	**19,844**	**European**	**1.55**	**$9.46x10^{-8}$**
PROX2	010	Tuberculosis	156	19,844	European	1.40	$9.70x10^{-7}$
USP44	010	Tuberculosis	156	19,844	European	1.44	$1.07x10^{-6}$
GPRIN1	010	Tuberculosis	156	19,844	European	1.49	$2.55x10^{-6}$
NSUN5	010	Tuberculosis	156	19,844	European	0.75	$2.76x10^{-6}$
C19orf55/PROSER3	010	Tuberculosis	156	19,844	European	1.51	$3.30x10^{-6}$
DNAJC17	054	Herpes simplex	154	3,241	African	0.65	$2.75x10^{-7}$
CHMP5	054	Herpes simplex	154	3,241	African	1.36	$2.22x10^{-6}$

GNRH1	054	Herpes simplex	154	3,241	African	1.47	$4.64x10^{-6}$
TXNL1	152	Sexually transmitted infections excluding HIV and hepatitis	152	25,643	European	0.64	$3.92x10^{-7}$
LTBP4	152	Sexually transmitted infections excluding HIV and hepatitis	152	25,643	European	0.70	$2.13x10^{-6}$
CRYL1	152	Sexually transmitted infections excluding HIV and hepatitis	152	25,643	European	0.70	$2.23x10^{-6}$
LIMK2	041.8	Helicobacter Pylori	150	19,844	European	0.71	$2.88x10^{-6}$
WFDC12	320	Meningitis	144	25,170	European	1.16	$4.23x10^{-6}$
TNNC2	071	Human immunodeficiency virus	139	3,241	African	1.44	$2.49x10^{-6}$
PRELID2	071	Human immunodeficiency virus	139	3,241	African	1.47	$2.68x10^{-6}$
PTCD3	**324**	**Other CNS infections and poliomyelitis**	**136**	**25,170**	**European**	**0.89**	**$3.74x10^{-8}$**
ATG9A	324	Other CNS infections and poliomyelitis	136	25,170	European	0.76	$2.46x10^{-6}$
EIF3M	078	Viral warts and human papilloma virus	136	3,241	African	1.48	$3.22x10^{-6}$
CDK5	038.2	Gram positive septicemia	114	3,337	African	0.65	$3.64x10^{-6}$

STAR★METHODS

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Eric R. Gamazon (eric.gamazon@vanderbilt.edu).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

BioVU

BioVU, one of the largest DNA biobanks tied to an EHR database, is a subset of the synthetic derivative (SD), a deidentified electronic health record, consisting of individuals with whole-genome genetic information. Detailed information on the construction, utilization, ethics, and policies of the BioVU resource is described elsewhere (Roden et al., 2008). ID traits were defined based on a hierarchical grouping of International Classification of Diseases, Ninth Revision (ICD-9) codes into phenotype codes (Phecodes) representing clinical traits, as previously described (Denny et al., 2013; Denny et al., 2010). (See below for a description of pathogen culture and viral test data in the BioVU individuals, including genera detected from different types of cultures.) We used version 1.2 of the Phecode Map containing 1,965 Phecodes based on 20,203 ICD-9 codes, which substantially improves signal-to-noise and more accurately reflects the clinical trait. Phecodes may exclude related phenotypes (e.g., in the case of Gram negative septicemia (Phecode = 038.1), the range of Phecodes given by 010-041.99, involving bacterial infection) and, importantly, include the definition of the appropriate control group (Wei et al., 2017). Detailed description of Phecode trait maps can be found at phewascatalog.org. As an efficient and viable model for human genetics research, the Phecode system has been used to perform phenome-wide association studies (PheWAS) for validation of

known genetic associations and discovery of new genetic disorders (Denny et al., 2013; Unlu et al., 2020).

Pathogen culture and virology data linked to whole-genome genetic information

For individuals with whole-genome genetic information, we analyzed pathogen (bacterial, mycobacterial, and fungal) culture data derived from the following positive cultures for the indicated clinical samples: 1) blood (n = 7,699), 2) sputum (n = 2,478), 3) sinus/nasopharyngeal (n = 1,820), 4) bronchial-alveolar lavage (n = 1,265), and 5) tracheal sampling (n = 422). Furthermore, we analyzed a respiratory panel containing 28 viral strains from 2,890 individuals with whole-genome genetic information. Viral strains included the following: 1) Adenovirus, 2) Bocavirus, 3) Bordetella parapertussis, 4) Bordetella pertussis, 5) Chlamydia pneumoniae, 6) Coronavirus 229E, 7) Coronavirus HKU1, 8) Coronavirus NL63, 9) Coronavirus NOS, 10) Coronavirus OC43, 11) Enterovirus/Rhinovirus, 12) Human Metapneumovirus, 13) Influenza A, 14) Influenza A, H1, 15) Influenza A, H1N1, 16) Influenza A, H3, 17) Influenza B, 18) Mycoplasma pneumoniae, 19) Parainfluenza, 20) Parainfluenza 1, 21) Parainfluenza 2, 22) Parainfluenza 3, 23) Parainfluenza 4, 24) Respiratory syncytial virus (RSV), 25) RSV, A, 26) RSV, B, and 27) Rhinovirus. The pathogen information for each individual in our study included: 1) Total number of cultures; 2) Number of negative cultures (i.e., no pathogen growth); 3) Number of ambiguous cultures (i.e., normal upper respiratory bacteria or low level contamination); 4) Number of positive cultures (i.e., the number of cultures with growth consistent with clinical infection); 5) Genus or genera isolated (up to 96 unique genera per sample site), which ranged from zero to 10 per sample.

METHODS DETAILS

GWAS of ID traits

GWAS of the ID traits were performed on the 23,294 and 4,321 BioVU individuals of European and African ancestry, respectively. Quality control pre-processing and SNP-level imputation were conducted, as previously described (Unlu et al., 2020). Genomic ancestry was quantified using principal components analysis of the genotype data (Derks et al., 2017; Price et al., 2006). The association analysis was performed using age, gender, batch, and the first five principal components as covariates.

Conditional SNP-level analysis

We performed conditional analysis on the top GWAS association with the ID trait (in this case, bacterial pneumonia) to determine whether it was driven by a related covariate (in this case, cystic fibrosis status). We used logistic regression to model the conditional probability of the infectious disease:

$$\ln \frac{P(Y=1 \mid s)}{1-(P(Y=1 \mid s)} = \beta_0 + \beta_1 s + \beta_2 (CF)$$

where s is the genotype at the sentinel variant, Y is the disease (i.e., bacterial pneumonia) status, and CF is the covariate of interest (i.e., cystic fibrosis).

Transcriptome-wide association studies (TWAS) using PrediXcan

We performed multi-tissue PrediXcan (Barbeira et al., 2019; Gamazon et al., 2018; Gamazon et al., 2015) in the 23,294 BioVU subjects. Experiment-wide significance was determined using Bonferroni correction for the total number of genes tested (n = 9,868) across 35 phenotypes (i.e., $p < 1.4 \times 10^{-7}$). Trait-specific significance was determined using Bonferroni correction for the total number of genes tested (n = 9,868, $p < 5.07 \times 10^{-6}$). Genomic ancestry was quantified using principal components analysis (Derks et al., 2017; Price et al., 2006). TWAS results were visualized using PhenoGram (Wolfe et al., 2013).

GWAS and TWAS Replication in the UK Biobank and FinnGen consortia

Replication of GWAS and TWAS was performed in the UK Biobank (Bycroft et al., 2018) and FinnGen consortia (Locke et al., 2019). We used the UK Biobank (http://www.nealelab.is/uk-biobank) and the FinnGen (https://www.finngen.fi/en/access_results) summary results to generate the gene-level associations. GTEx v6p models were used to generate tissue-level results.

Classification of pathogen infection based on serology and culture data using several classifiers

Let X be a classifier (e.g., the Phecode or a logistic regression classifier) of serology and culture data based infection for a given pathogen, with probability density $\varphi_+(x)$ for positive instances and probability density $\varphi_-(x)$ for negative instances. The ROC curve plots the specificity (SP) and sensitivity (SN) at various thresholds:

$$SN(T) = \int_T^\infty \varphi_+(x)dx$$

$$SP(T) = 1 - \int_T^\infty \varphi_-(x)dx$$

The area Ω under the curve (AUC) is given by:

$$\Omega = \int_{-\infty}^{\infty} SN(T)SP'(T)dT = \int_{-\infty}^{\infty}\int_T^\infty \varphi_+(x)\,\varphi_-(T)dxdT = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} I(x > T)\,\varphi_+(x)\varphi_-(T)dxdT$$

where $I(A)$ is the indicator function, i.e., equal to one if $(x, T) \in A$ and zero otherwise. The last equals the probability that the classifier X ranks a randomly chosen positive instance (of culture data based infection) higher than a randomly chosen negative instance. We note that the expression for Ω suggests other metrics of interest, for example:

$$\Omega_c = \int_c^1 SN(SP)dSP$$

$$c = \underset{t}{\operatorname{argmin}} \sqrt{(1-t)^2 + (1-SN(t))^2}$$

Here $(c, SN(c)$ is the point on the ROC curve closest to true positive rate of 1 and false positive rate of 0. We estimated the sampling distribution of Ω (including standard error), using bootstrapping (n = 100) (Efron, 1979). We used the pROC package for visualization.

Leveraging GWAS of human microbiome traits to extend GWAS of ID traits

We leveraged genome-wide associations of microbiome traits, involving 155 pathogens derived from phylogenetic analysis of 16S rRNA gene sequences (Hughes et al., 2020).

Causal inference by Mendelian Randomization

To infer causality between the infectious diseases and potential complications, we performed Mendelian Randomization (MR, (Davey Smith and Hemani, 2014; Lawlor et al., 2008)) in 23,294 individuals of European ancestry in BioVU. To define instrumental variables (IVs), we clumped the exposure-associated SNPs with high linkage disequilibrium (LD) using Plink1.9 ($p < 1x10^{-5}$, $r^2 = 0.01$). Only biallelic non-palindromic variants were considered as IVs. Considering the pervasive horizontal pleiotropy in human genetic variation (Jordan et al., 2019), we applied summary statistics based MR-Egger regression (Bowden et al., 2015). MR-Egger regression generalizes the inverse-variance weighted method, where the intercept is assumed to be zero. We also used the weighted-median estimator (Bowden et al., 2016) to test the causal effect of the exposure trait on the outcome. We leveraged the R package 'MendelianRandomization'.

High-throughput Human in vitrO Susceptibility Testing (Hi-HOST)

We generated an atlas of TWAS associations with 79 pathogen-induced cellular traits – including infectivity and replication, cytokine levels, and host cell death (Wang et al., 2018)

using the Hi-HOST platform (Ko et al., 2012; Ko et al., 2009). A list of populations, pathogens and project description may be found at http://h2p2.oit.duke.edu/About/, and phenotype definitions and family-based GWAS of the Hi-HOST Phenome Project were previously described (Wang et al., 2018). Briefly, lymphoblastoid cell lines (LCLs) from the 1000 Genomes Consortium (Auton et al., 2015) were obtained from the Coriell Institute. The LCLs represented diverse populations, including ESN (Esan in Nigeria), GWD (Gambians in Western Divisions in the Gambia), IBS (Iberian Population in Spain), and KHV (Kinh in Ho Chi Minh City, Vietnam). LCLs were cultured in RPMI 1640 media containing 10% fetal bovine serum, 2 mM glutamine, 100 U/ml of penicillin-G, and 100 mg/ml streptomycin for 8 days prior to experimental use, as previously described (Wang et al., 2018). *Chlamydia trachomatis* infection of LCLs was performed using *C. trachomatis* LGV-L2 Rif^R pGFP::SW2 (Saka et al., 2011). *Salmonella* infection was performed using pMMB67GFP (Pujol and Bliska, 2003), and sifA deletion was constructed using lambda red and validated using PCR (Datsenko and Wanner, 2000; Ko et al., 2009). *Candida albicans* SC5314 infection was performed as previously described (Odds et al., 2004) and levels of fibroblast growth factor 2 were measured using enzyme linked immunosorbent assays. *Staphylococcus aureus* toxin (alpha-hemolysin) was obtained from Sigma and applied to LCLs at a concentration of 1 μg/ml for 23 hours. Cell death was measured using 7-AAD staining and flow cytometry. Additional experimental details can be found at http://h2p2.oit.duke.edu/About/.

We estimated the gene-level effect size on the Hi-HOST phenotypes, using GWAS summary statistics (Barbeira et al., 2018) in each of the 44 GTEx tissues (version 6p) (Battle et al., 2017). The gene expression prediction model was trained using GTEx as the reference dataset (https://zenodo.org/record/3572842/files/GTEx-V6p-HapMap-2016-09-08.tar.gz). The gene-level effect size was estimated using S-PrediXcan after allele harmonization (Barbeira et al., 2018). We also applied MultiXcan to improve the ability to identify potential target genes (Barbeira et al., 2019). In brief, MultiXcan regresses the cellular trait on the principal

components of the predicted expression data across all the available tissues. For each gene, MultiXcan yields a joint effect estimate across the 44 tissues. We applied the summary-statistic based version (S-MultiXcan) and followed the guides from the tool's webpage https://github.com/hakyimlab/MetaXcan.

DATA AND SOFTWARE AVAILABILITY

All code is available at the project's github page: https://github.com/gamazonlab/infectiousDiseaseResource. All trait-level GWAS, PrediXcan, and Hi-HOST TWAS results are available at www.phewascatalog.org.

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
BioVU	(Denny et al., 2010; Roden et al., 2008)	https://victr.vanderbilt.edu/pub/biovu/
PrediXcan genetic associations for 35 ID traits	This paper	phewascatalog.org
PrediXcan	(Gamazon et al., 2018; Gamazon et al., 2015)	https://github.com/hakyimlab/PrediXcan
MultiXcan	(Barbeira et al., 2019)	https://github.com/hakyimlab/multixcan-paper
GTEx	(2015; Battle et al., 2017; Consortium, 2013)	https://gtexportal.org/home/
Gene Set Enrichment Analysis (GSEA)	(Subramanian et al., 2005)	http://software.broadinstitute.org/gsea/index.jsp
Human gut microbiome GWAS	(Hughes et al., 2020)	https://data.bris.ac.uk/data/dataset/f12b2cf665593ac466149cc6da3ca867
Human gut microbiome TWAS	This paper	phewascatalog.org
Mendelian Randomization software package	(Yavorska and Burgess, 2017)	https://cran.r-project.org/web/packages/MendelianRandomization/MendelianRandomization.pdf
Hi-HOST GWAS	(Ko et al., 2012; Wang et al., 2018)	http://h2p2.oit.duke.edu/About/
Hi-HOST TWAS	This paper	phewascatalog.org