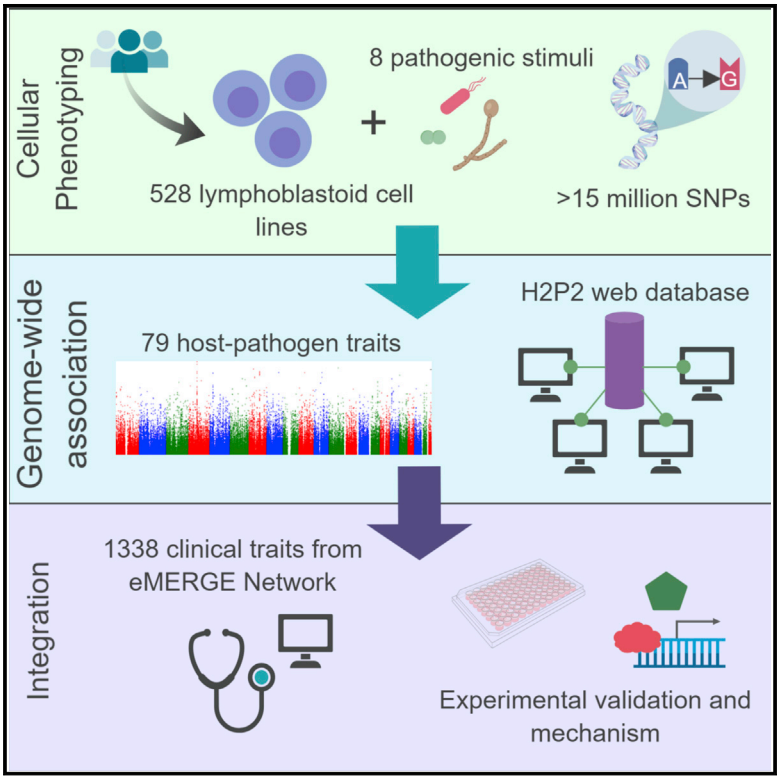
Resource

Cell Host & Microbe

# An Atlas of Genetic Variation Linking Pathogen-Induced Cellular Traits to Human Disease

## Graphical Abstract




## Authors

Liuyang Wang, Kelly J. Pittman, Jeffrey R. Barker, ..., Raphael H. Valdivia, David R. Crosslin, Dennis C. Ko

## Correspondence

dennis.ko@duke.edu

## In Brief

Approaches are needed for a deeper understanding of how human genetics impacts disease susceptibility. Wang et al. present a catalog of cellular genome-wide association studies comprising 79 phenotypes in response to 8 pathogens. Combining this with clinical association data and experimental validation revealed mechanisms and connections to disease.

## Highlights

- Heritable variation in 79 cellular responses to infection with 8 pathogens was assessed
- Phenotypes segregate in biologically meaningful clusters
- 17 significant genome-wide associations with infection phenotypes were identified
- Integration with clinical GWAS revealed SNPs associated with IBD and hepatitis

Wang et al., 2018, Cell Host & Microbe *24*, 308–323
August 8, 2018 
https://doi.org/10.1016/j.chom.2018.07.007

CellPress

CellPress

Cell Host & Microbe
Resource

# An Atlas of Genetic Variation Linking Pathogen-Induced Cellular Traits to Human Disease

Liuyang Wang,[1] Kelly J. Pittman,[1] Jeffrey R. Barker,[1] Raul E. Salinas,[1] Ian B. Stanaway,[2] Graham D. Williams,[1] Robert J. Carroll,[3] Tom Balmat,[4] Andy Ingham,[5] Anusha M. Gopalakrishnan,[1] Kyle D. Gibbs,[1] Alejandro L. Antonia,[1] The eMERGE Network, Joseph Heitman,[1,6] Soo Chan Lee,[7] Gail P. Jarvik,[8] Joshua C. Denny,[3] Stacy M. Horner,[1,6] Mark R. DeLong,[5] Raphael H. Valdivia,[1] David R. Crosslin,[2] and Dennis C. Ko[1,6,9,*]

[1]Department of Molecular Genetics and Microbiology, School of Medicine, Duke University, Durham, NC 27710, USA
[2]Department of Biomedical Informatics and Medical Education, School of Medicine, University of Washington, Seattle, WA 98195, USA
[3]Department of Biomedical Informatics, School of Medicine, Vanderbilt University, Nashville, TN 37212, USA
[4]Social Science Research Institute, Duke University, Durham, NC 27710, USA
[5]Duke Research Computing, Duke University, Durham, NC 27710, USA
[6]Division of Infectious Diseases, Department of Medicine, School of Medicine, Duke University, Durham, NC 27710, USA
[7]South Texas Center for Emerging Infectious Diseases (STCEID), Department of Biology, College of Sciences, the University of Texas at San Antonio, San Antonio, TX 78249, USA
[8]Department of Medicine, Division of Medical Genetics, School of Medicine, University of Washington, Seattle, WA 98195, USA
[9]Lead Contact
*Correspondence: dennis.ko@duke.edu
https://doi.org/10.1016/j.chom.2018.07.007

## SUMMARY

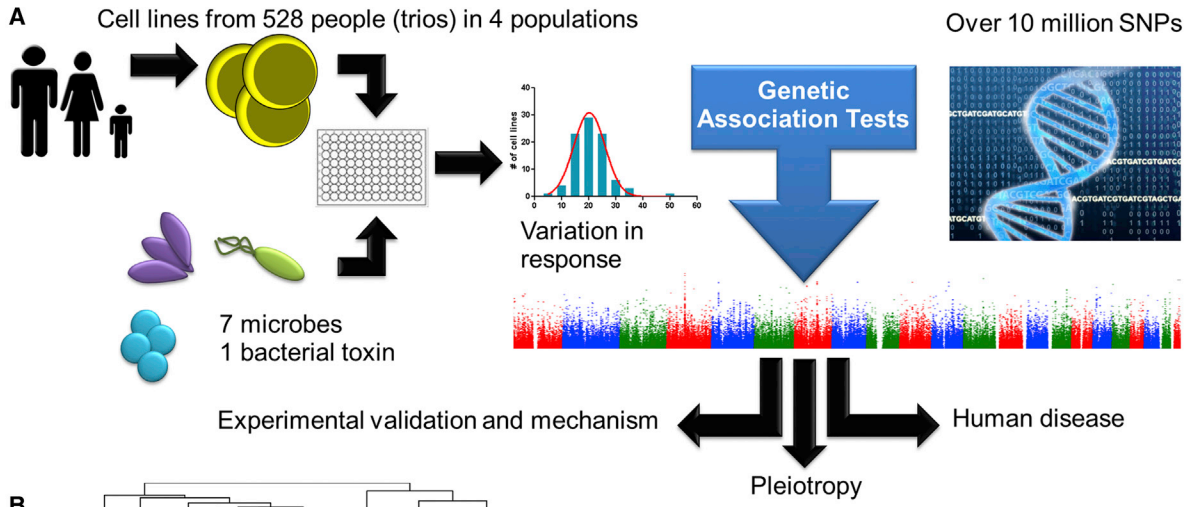
**Pathogens have been a strong driving force for natural selection. Therefore, understanding how human genetic differences impact infection-related cellular traits can mechanistically link genetic variation to disease susceptibility. Here we report the Hi-HOST Phenome Project (H2P2): a catalog of cellular genome-wide association studies (GWAS) comprising 79 infection-related phenotypes in response to 8 pathogens in 528 lymphoblastoid cell lines. Seventeen loci surpass genome-wide significance for infection-associated phenotypes ranging from pathogen replication to cytokine production. We combined H2P2 with clinical association data from patients to identify a SNP near *CXCL10* as a risk factor for inflammatory bowel disease. A SNP in the transcriptional repressor *ZBTB20* demonstrated pleiotropy, likely through suppression of multiple target genes, and was associated with viral hepatitis. These data are available on a web portal to facilitate interpreting human genome variation through the lens of cell biology and should serve as a rich resource for the research community.**

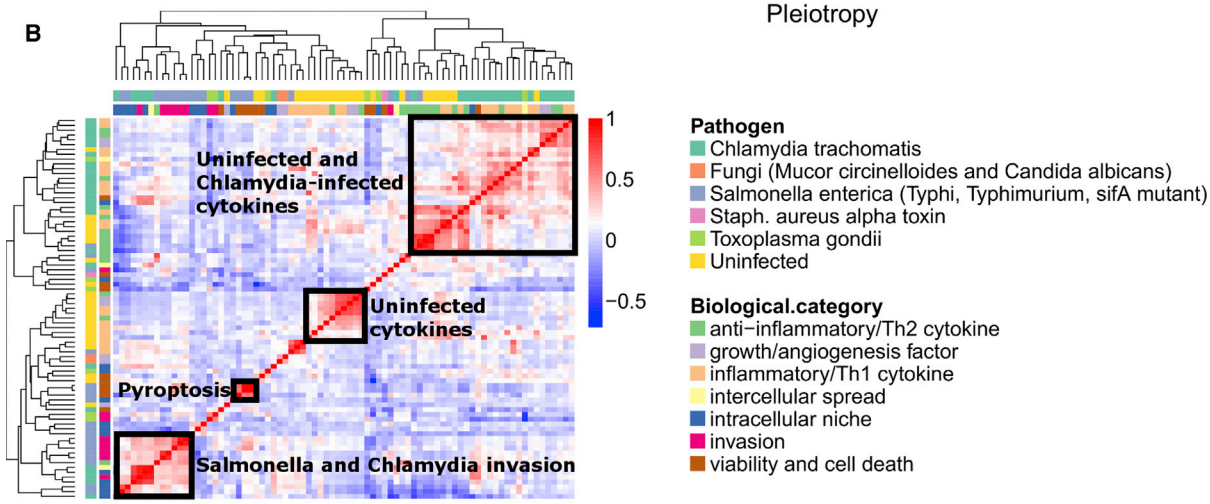
## INTRODUCTION

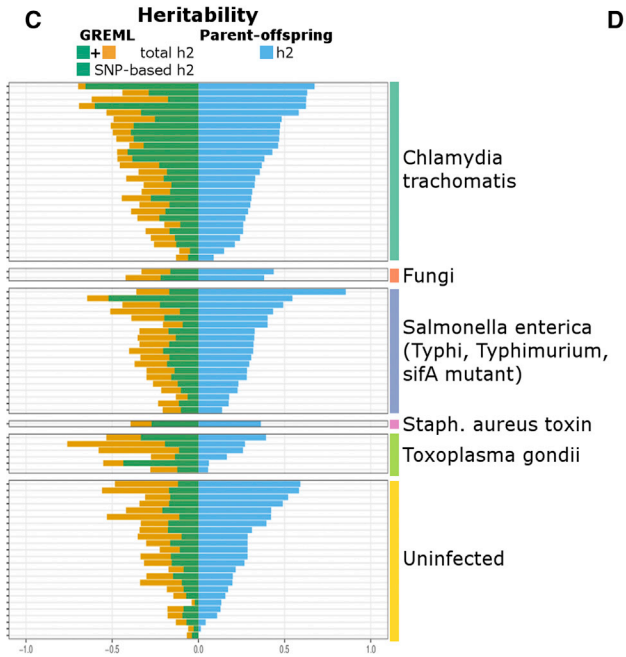
The human genome has been shaped by migration, drift, admixture, and natural selection (Cavalli-Sforza et al., 1994; Li et al., 2008). One of the strongest driving forces in natural selection has been pathogens (Fumagalli et al., 2011), as first exemplified with A.C. Allison's demonstration that sickle cell allele (rs334) confers resistance to malaria (Allison, 1954). Red blood cells from individuals with this allele are resistant to *Plasmodium* infection (Friedman, 1978). Similarly, human resistance to HIV infections afforded by the *CCR5Δ32* allele can also be seen at the level of individual T cells (Liu et al., 1996). Therefore, understanding how human genetic differences impact cellular traits can mechanistically link human genetic variation to disease susceptibility.

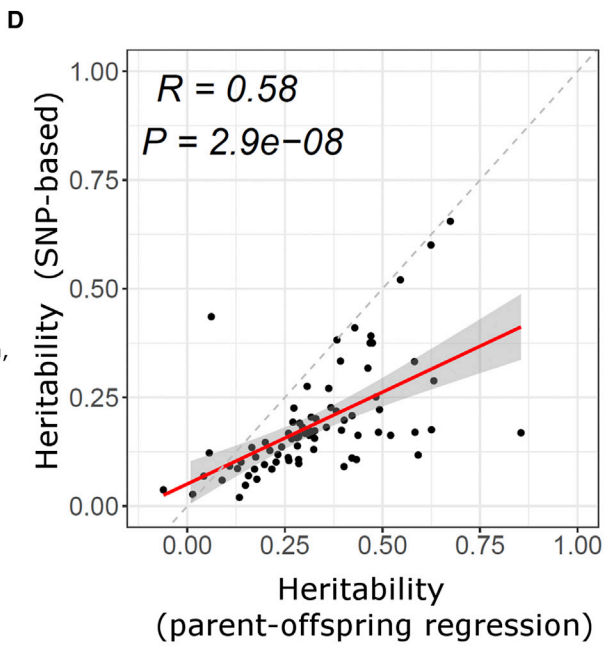
Key to delineating such causal links have been studies examining molecular traits, such as gene expression, in human populations. Expression quantitative trait loci (eQTL) studies in lymphoblastoid cell lines (LCLs) defined abundant associations between human SNPs and expression of nearby genes (Nica and Dermitzakis, 2013; Stranger et al., 2007). LCLs are Epstein-Barr virus (EBV)-transformed B cells that are transcriptionally similar to antigen-activated primary B cells (Cahir-McFarland et al., 2004). LCLs serve as a standardized resource for functional human genetic variation studies, as they have been densely genotyped (1000 Genomes Project Consortium et al., 2010; International HapMap Consortium, 2005). As eQTLs are often shared across tissues (e.g., 88% of *cis*-eQTLs are shared among LCLs, fibroblasts, and primary T cells [Flutre et al., 2013]), LCL eQTL studies have led to important insights not only in immunity-related diseases but also for disorders where B cells are not believed to be primary drivers of disease (Nica and Dermitzakis, 2013).

Using LCLs, we developed Hi-HOST (high-throughput human in vitro susceptibility testing) to identify human genetic differences in pathogen-induced cellular traits, serving as a cell biological link between eQTL studies and GWAS of disease (Ko et al., 2009). Hi-HOST uses live pathogens to examine variation in innate immune recognition, but also in pathogen-manipulated cell biological processes that can be quantified as phenotypes for genome-wide association. This work therefore builds on a long tradition of using cellular microbiology to elucidate basic




*(legend on next page)*

cell biology (Cossart et al., 1996) and expands that utility to interpret the human genome. Using Hi-HOST, we leveraged LCL responses to *Salmonella enterica* to demonstrate that genetic variation in the methionine salvage pathway regulates pyroptosis and human susceptibility to sepsis (Ko et al., 2012; Wang et al., 2017). Similarly, we recently reported that a genetic variant in *VAC14* is associated with both increased *S.* Typhi invasion into LCLs and risk of typhoid fever in a Vietnamese population (Alvarez et al., 2017).

Here, we present the Hi-HOST Phenome Project (H2P2) to explore the genetic basis of cellular outcomes in response to infectious agents. Using 7 microbes and 1 bacterial toxin, we carried out GWAS of 79 host-pathogen phenotypes that serve as cellular readouts for processes such as endocytosis, endosomal trafficking, signal transduction, cell death, and transcriptional regulation. We integrated H2P2 data with experimental validation and disease association data from the eMERGE Network PheWAS pipeline (Denny et al., 2013) to define functions for genes in disease and provide clues to pathophysiology.

## RESULTS

### Phenotypic Variation in H2P2 Traits Reveals Biologically Meaningful Clusters

We measured variation in cellular traits in 528 LCLs stimulated with 7 different microbes and 1 bacterial toxin (Figure 1A). LCLs were from 4 human populations and consisted of parent-offspring trios, allowing for heritability estimation and for GWAS analysis with protection from stratification through family-based methods (Purcell et al., 2005, 2007).

The microbes and toxin we selected affect billions of people. Non-typhoidal *Salmonella* infections caused 150 million diarrheal illnesses and 0.6 million cases of invasive enteric disease in 2010 (Kirk et al., 2015). Approximately 20 million cases of typhoid fever are caused by *S.* Typhi every year (Dougan and Baker, 2014). *Chlamydia trachomatis* causes 100 million cases of genital tract infection every year (Newman et al., 2015), and 1.3 million people are blind due to ocular infection (Burton and Mabey, 2009). *Staphylococcus aureus* is a common cause of skin and soft tissue infections, bacteremia, and infective endocarditis, and its alpha toxin, utilized in H2P2, is a key virulence determinant (Tong et al., 2015). *Candida albicans* is a frequent cause of genitourinary tract infection that can cause more severe disease in immunocompromised individuals (Yapar, 2014). *Mucor circinelloides* is another fungal species that causes severe infections in immunocompromised individuals (Mendoza et al., 2014). Finally, over 6 billion people have been infected with the protozoal pathogen *Toxoplasma gondii*, which can be fatal in infants and immunocompromised individuals (Furtado et al., 2011). Thus the microbes and toxin used in H2P2 are important causes of human disease.

The pathogens selected also exploit a wide range of host cellular processes to either kill the host cell or to create replicative niches within them. With H2P2, LCLs are used as a general cellular model for host-pathogen traits. For example, invasion into LCLs requires the same type III secretion system (T3SS) and secreted effectors that are utilized in invasion of epithelial cells (Alvarez et al., 2017). Traits were selected for screening based on pilot experiments and phenotype optimization in 7 LCLs measured by flow cytometry and a Luminex panel of 41 cytokines. *C. trachomatis*, *S. enterica* serovar Typhi, serovar Typhimurium (wild-type and Δ*sifA* mutant, which escapes from the pathogen-containing vacuole at a greater rate into the cytosol [Beuzon et al., 2000]), and *T. gondii* are intracellular pathogens that employ diverse lifestyles. These microbes were engineered to express GFP to allow quantitation of pathogen invasion, survival and replication, intercellular spread, and concurrent measurement of cell death by flow cytometry. Cell death was also measured as the readout for *S. aureus* alpha toxin. Microbes were also tested for induction or suppression of cytokines. *S.* Typhimurium and *C. trachomatis*-infected cells were screened for 3 and 17 cytokines, respectively. *M. circinelloides* and *C. albicans* were included for their ability to induce fibroblast growth factor 2 (Lee et al., 2015). Definitions, histograms, and a graphical breakdown for all phenotypes are provided (Table S1; Data S1; Figure S1). Importantly, 76 of 79 H2P2 phenotypes showed significant experimental repeatability based on measurements on LCLs from three different passages (Figure S2).

Hierarchical clustering of traits based on inter-individual variation confirmed robustness of our measurements (Figure 1B). Levels of three cytokines (CXCL10 [also known as IP-10], interleukin-10 [IL-10], and macrophage-derived cytokine [MDC]) measured in uninfected cells with two different methods (ELISA at 24 hr and Luminex at 70 hr) showed strong correlation ($R = 0.78$ for CXCL10, $R = 0.46$ for IL-10, $R = 0.90$ for MDC). The clustering of responses to *Salmonella* infection is consistent with previous findings: *S.* Typhimurium, Typhi, and Typhimurium Δ*sifA* cluster for the phenotype of invasion, as all utilize a similar T3SS for entry (Collazo and Galan, 1997) (correlation to *S.* Typhimurium, $R = 0.69$ for *S.* Typhi and $R = 0.90$ for *S.* Typhimurium Δ*sifA*). The correlation is weaker for intracellular survival and replication phenotypes (correlation to *S.* Typhimurium replication, $R = 0.36$ for *S.* Typhi and $R = 0.34$ for *S.* Typhimurium Δ*sifA*), reflecting the different replications niches for Δ*sifA* (host cell

**Figure 1. Inter-individual Variation in H2P2 Traits Revealed Clustering of Phenotypes and Heritable Variation in Cellular Responses to Infection**

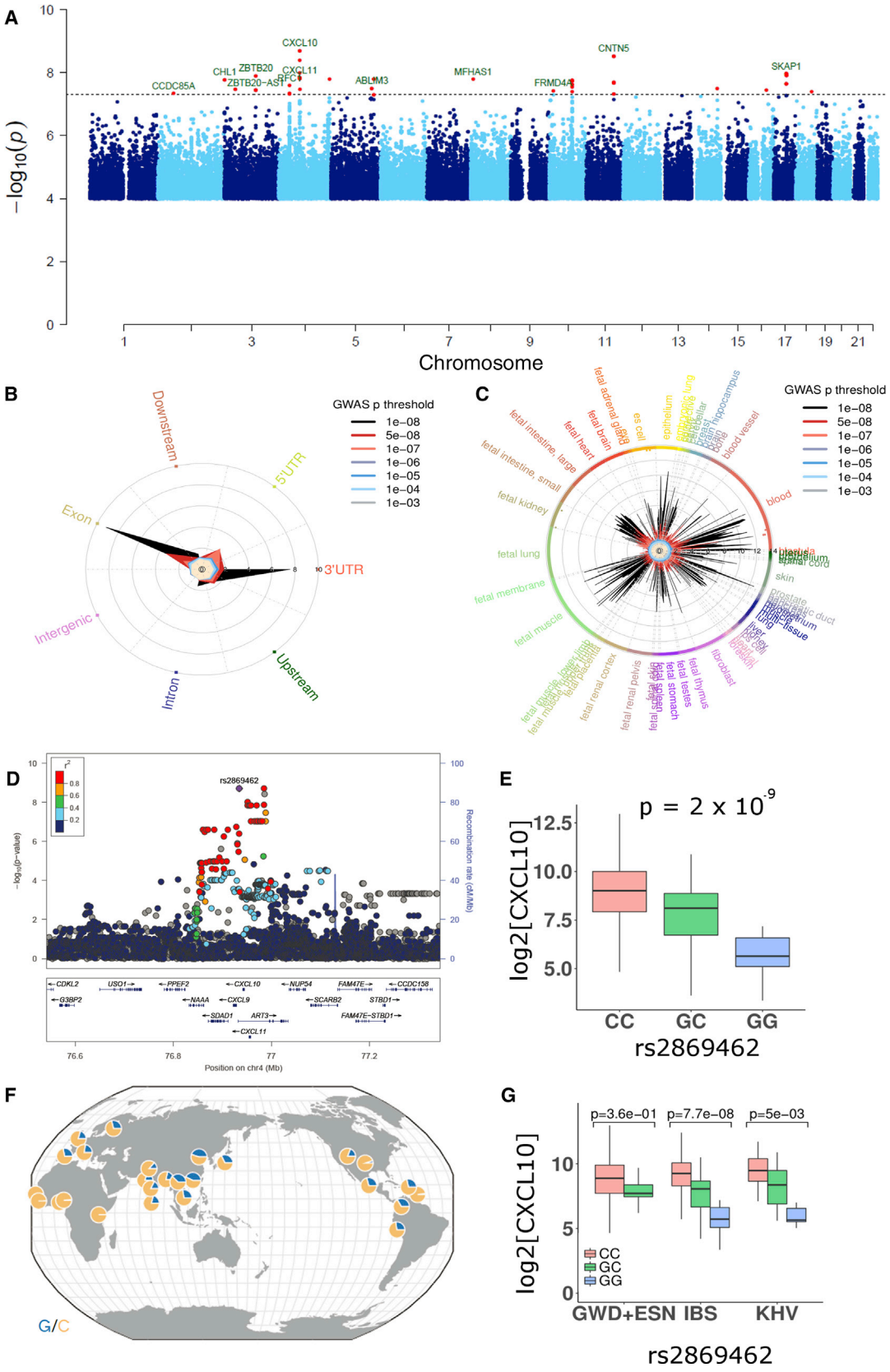
(A) Diagram of H2P2 workflow for connecting genetic variation to cell biology.

(B) Hierarchical clustering of H2P2 phenotypes revealed a map of trait similarity based on inter-individual phenotypic variation (Spearman correlation). Phenotypes are color-coded by stimuli (outer band) and biological category (inner band).

(C) Narrow-sense heritability ($h^2$) estimates for H2P2 phenotypes based on the Zaitlen method of GREML versus parent-offspring regression. The GREML method provided a SNP-based $h^2$ (green) as well as a total $h^2$ (yellow) (the sum of SNP-based $h^2$ [green] plus the non-SNP-based $h^2$).

(D) $h^2$ estimates from parent-offspring regression versus GREML SNP-based $h^2$ were well correlated for H2P2 traits. Gray shading indicates 95% confidence intervals.

See also Tables S1, S2, and S3; Figures S1–S3; Data S1 and S2.




*(legend on next page)*

cytoplasm) and wild-type Typhimurium (membrane bound vacuole) (Beuzon et al., 2000) or the use of a different repertoire of effectors by *S.* Typhi (Parkhill et al., 2001). In contrast, we observed almost no correlation between EBV copy number (from Mandage et al. (2017) for 284 LCLs also used in H2P2) and H2P2 traits (Figure S3; Table S2), indicating these phenotypes are not being driven by the LCL immortalization method. Thus, clustering based on phenotypic diversity verified reliability of measurements and confirmed biological relatedness established by much previous work.

## H2P2 Traits Are Heritable

To estimate contributions of genetic differences to phenotypic variance, we measured narrow-sense heritability ($h^2$) with two complementary methods: parent-offspring regression and SNP-based $h^2$. $h^2$ based on parent-offspring regression is estimated as the slope of the regression line for offspring phenotypes versus mid-parent phenotypes (Falconer and Mackay, 1996). We observed $h^2$ estimates from −0.06 to 0.85 (average $h^2$ = 0.33) (Figure 1C; Data S2; Table S3). The majority (64/79) of phenotypes showed significantly non-zero $h^2$ by this method ($p < 0.05$). In contrast, SNP-based $h^2$, as implemented in GCTA software (Yang et al., 2011) with the Zaitlen modification for related individuals (Zaitlen et al., 2013), calculates the proportion of variance that can be explained by all genotyped SNPs. With this method, we observed pedigree $h^2$ ranging from 0.04 to 0.76 (average $h^2$ = 0.36), and SNP-based $h^2$ ranging from 0.02 to 0.66 (average $h^2$ = 0.19) (Figure 1C; Table S3). The SNP-based $h^2$ estimates must be interpreted with caution, as small sample sizes resulted in large standard errors (Table S3).

We observed high correlation between $h^2$ estimated using the two different methods ($R = 0.58$; $p = 2.9 \times 10^{-8}$; Figure 1D). The strong correlation between the two estimates of $h^2$, based on distinct statistical frameworks, provides additional evidence that LCLs provide a robust system to identify human SNPs that contribute to the heritability of cellular phenotypes.

## H2P2 Reveals 17 Genome-wide Significant Associations and Enrichment for Genic SNPs and Regions of Active Chromatin

We performed family-based GWAS using dense genotyping information (15.5 million SNPs after imputation; see STAR Methods). Across 79 traits, 17 loci reached a genome-wide significance threshold of $p < 5 \times 10^{-8}$ (Figure 2A; Table 1). Phenotype permutation analysis demonstrated that to obtain an $\alpha$ of 0.1, a more stringent p value threshold of $p < 2.76 \times 10^{-8}$ is appropriate (Figure S4). At this threshold, 10 higher-confidence loci remain, although we emphasize that H2P2 should be viewed as a hypothesis-generating resource and that SNPs even at this more stringent threshold should undergo further validation.

Several of these SNPs demonstrated very different allele frequencies among populations (Table S4). If we had conducted this study only in European LCLs, we would not have detected five SNPs that are rare/absent in European populations. Likewise, if we had conducted this study only in African LCLs, we would not have detected four of the associated SNPs. This underscores the need for conducting GWAS in multiple populations for revealing a greater spectrum of genetic differences of functional significance.

H2P2 demonstrated enrichment of associated SNPs for functional genome annotations. We used GARFIELD to calculate and visualize fold enrichment of SNPs associated in H2P2 at variable p value thresholds with different genomic features (Iotchkova et al., 2016). In regard to SNP location, the greatest enrichment was observed for exonic SNPs (Figure 2B). The fold enrichment was highest at the most stringent p value threshold ($p < 1 \times 10^{-8}$) for H2P2 traits (9.2-fold enrichment; $p = 0.10$ by Fisher's exact test) and was statistically significant when using a $p < 5 \times 10^{-8}$ threshold for H2P2 traits (3.9-fold enrichment; $p = 0.047$). In contrast, there was depletion for intergenic SNPs at the $p < 5 \times 10^{-8}$ threshold (0.66-fold enrichment). H2P2-associated SNPs showed even greater enrichment for regions of active chromatin based on DNase hypersensitivity peaks from ENCODE (ENCODE Project Consortium, 2012) (Figure 2C). Consistent with H2P2 being conducted in LCLs, the second greatest enrichment was observed for DNAase hypersensitivity peaks measured in the LCL GM06990 out of 424 cell types measured (at $p < 5 \times 10^{-8}$, 5.7-fold enrichment; $p = 3.8 \times 10^{-3}$). Even stronger enrichment was noted at this threshold for human Th2 cells (7.3-fold enrichment; $p = 3.5 \times 10^{-4}$), and enrichment was observed for cells derived from most tissues (Figure 2C), consistent with LCLs being a relevant model for genetic analysis for multiple human cell types.

## A Large-Effect *cis*-Regulatory Variant Regulates CXCL10 Levels

The strongest association was observed for rs2869462 with levels of the chemokine CXCL10 (also known as IP-10) following

**Figure 2. GWA of H2P2 Revealed 17 Genome-wide Significant Loci Including a *cis*-Cytokine-QTL Near *CXCL10***

(A) Meta-Manhattan plot for 79 traits shows 17 peaks (red) with $p < 5 \times 10^{-8}$ (dotted line). –log(p values) were calculated using QFAM-parents in PLINK.

(B) GARFIELD enrichment plot of SNP location demonstrated enrichment of SNPs associated with H2P2 phenotypes in exons, and 5′ and 3′ UTRs. SNPs associated with H2P2 traits at various p value thresholds were plotted in the indicated colors and the height of the peak within each category indicates fold enrichment from 0 to 10.

(C) GARFIELD enrichment plot of DNase hypersensitivity peaks demonstrated enrichment of SNPs associated with H2P2 phenotypes in active chromatin regions in multiple cell/tissue types.

(D) Regional plot around the *CXCL10* gene demonstrated association of rs2869462 with CXCL10 levels from *C. trachomatis*-infected cells. SNPs are plotted by position on chromosome 4 and –log(p value) and color-coded by $r^2$ value to rs2869462 from 1000 Genomes European data.

(E) Genotypic medians, first and third quartiles (box), and maximum and minimum values (whiskers) for rs2869462 for CXCL10 levels from *C. trachomatis*-infected LCLs from all LCLs.

(F) Map of rs2869462 allele frequencies (C, orange; G, blue) from Geography of Genetic Variants Browser (Marcus and Novembre, 2017).

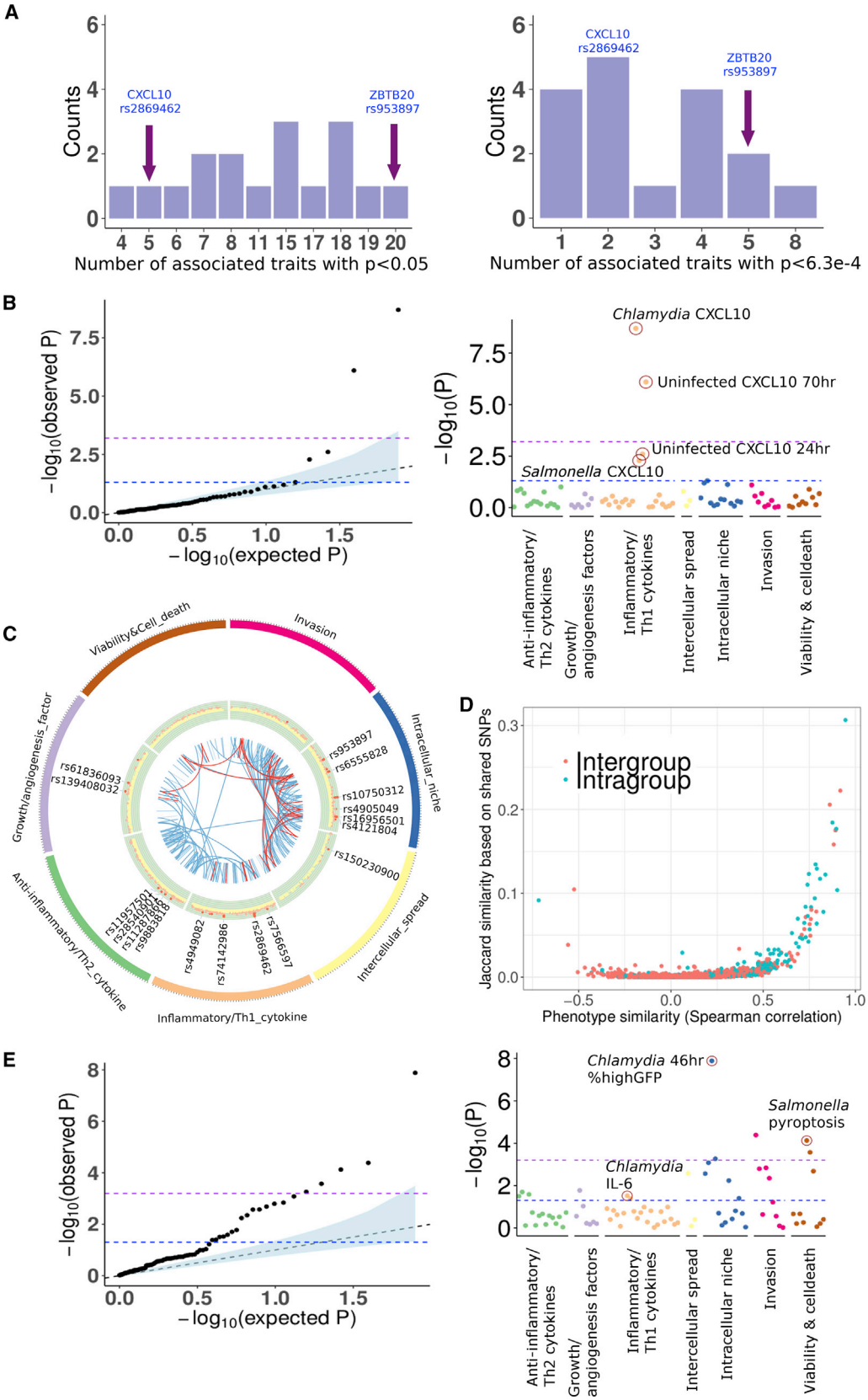
(G) Individual population genotypic median plots for rs2869462 for CXCL10 levels demonstrated C > G in all populations. p values were generated with QFAM-parents in PLINK.

See also Table S4; Figures S4 and S5.

**Table 1. Genome-wide Significant H2P2 SNPs**

| SNP ID | Chr | Position | Cellular Trait | p Value | Risk Allele | Gene[a] | Parent-Offspring $h^2$ for Trait (%) | SNP-Based $h^2$ for Trait (%) | Variance Explained by SNP (Parental LCLs Only) (%) |
|---|---|---|---|---|---|---|---|---|---|
| rs7566597 | 2 | 56289463 | MIP1B_Chlamydia | $4.67 \times 10^{-8}$ | G | CCDC85A | 58.2 | 33.3 | 4.4 |
| rs9883818 | 3 | 337848 | MDC_Uninfected_ELISA | $1.70 \times 10^{-8}$ | A | CHL1 | 28.6 | 9.8 | 1.2 |
| rs150230900 | 3 | 38822664 | Chlamydia_70hr_GFP | $3.50 \times 10^{-8}$ | – | near RP11-134J21.1 | 21.1 | 12.9 | 1.0 |
| rs953897 | 3 | 114349113 | Chlamydia_46hr_highGFP | $1.30 \times 10^{-8}$ | T | ZBTB20 | 33.0 | 20.1 | 2.4 |
| rs11287866 | 4 | 39317796 | IL4_Chlamydia | $2.60 \times 10^{-8}$ | TA | RFC1 | 36.8 | 22.7 | 1.6 |
| rs2869462 | 4 | 76013566 | IP10_Chlamydia | $2.00 \times 10^{-9}$ | C | ART3; near CXCL10 | 48.3 | 25.2 | 13.7 |
| rs28540901 | 4 | 184492384 | IL10_Uninfected_Luminex | $1.60 \times 10^{-8}$ | T | near IRF2 | 10.8 | 9.2 | 2.9 |
| rs11957501 | 5 | 149250737 | MDC_S_typhimurium | $3.25 \times 10^{-8}$ | C | ABLIM3 | 43.3 | 10.7 | 4.9 |
| rs6555828 | 5 | 157066729 | Chlamydia_70hr_median_GFP | $1.60 \times 10^{-8}$ | G | near HAVCR1 | 27.3 | 22.6 | 2.8 |
| rs139408032 | 8 | 8882324 | FGF2_Mucor | $1.65 \times 10^{-8}$ | A | MFHAS1 | 38.2 | 21.9 | 2.7 |
| rs61836093 | 10 | 14235152 | FGF2_Candida | $3.80 \times 10^{-8}$ | G | FRMD4A | 43.7 | 16.3 | 2.0 |
| rs74142986 | 10 | 81018235 | TNFb_Chlamydia | $1.85 \times 10^{-8}$ | T | – | 35.6 | 18.2 | 2.5 |
| rs10750312 | 11 | 99526242 | S_typhimurium_Intracellular_Replication_24_3_5hr_median_GFP | $3.00 \times 10^{-9}$ | G | CNTN5 | 22.7 | 10.2 | 3.2 |
| rs4905049 | 14 | 93121941 | Chlamydia_46hr_highGFP | $3.30 \times 10^{-8}$ | G | near ITPK1 | 33.0 | 20.1 | 0.3 |
| rs4949082 | 16 | 63851836 | IP10_S_typhimurium | $3.70 \times 10^{-8}$ | A | – | 40.1 | 9.1 | 1.2 |
| rs16956501 | 17 | 48419912 | Chlamydia_70hr_median_GFP | $1.10 \times 10^{-8}$ | C | SKAP1 | 27.3 | 22.6 | 0.2 |
| rs4121804 | 18 | 59790252 | Chlamydia_46hr_highGFP | $4.11 \times 10^{-8}$ | G | – | 33.0 | 20.1 | 0.6 |

A single SNP with the lowest p value is listed for each peak. SNPs described in the text are underlined.
[a]Gene the SNP is located in or "near" (within 20 kb).




(legend on next page)

*Chlamydia* infection ($p = 2 \times 10^{-9}$; Figures 2D and 2E). This SNP is located 7.5 kb 3′ of the *CXCL10* coding sequence (Figure 2D). CXCL10 mediates inflammation by coordinating T helper 1 recruitment and activating effector cells during infection and autoimmunity (Groom and Luster, 2011). The effect of this SNP is large: rs2869462 accounts for 13.7% of the variance in CXCL10 protein levels. While no dataset is available to replicate this association at the protein level, this SNP also demonstrated an association with *CXCL10* mRNA ($p = 7 \times 10^{-7}$) in an independent set of 465 uninfected LCLs (Lappalainen et al., 2013; none of the LCLs overlap the H2P2 LCLs) (Figure S5).

There were large differences in rs2869462 allele frequencies among the populations. The derived allele of rs2869462 (G) is present at the highest frequencies in Europe (28% in Iberian Population in Spain) and Asia (29% in Kinh in Ho Chi Minh City, Vietnam) and is substantially lower in Africa (0.8% in Esan in Nigeria, 0.8% in Gambians in Western Divisions in the Gambia; Figure 2F; Table S4). Strikingly, all populations demonstrated the same directionality of effect on CXCL10 levels (C > G) (Figure 2G).

## SNPs Lead to Pleiotropic Effects on Multiple Pathogen-Induced Traits

Different pathogens can target common signaling pathways to establish an intracellular niche or to modulate immune responses. Therefore, we examined whether genome-wide significant hits were associated with multiple traits at $p < 0.05$ or, with Bonferroni multiple test correction, $p < 6.3 \times 10^{-4}$. All genome-wide significant hits were associated with at least four H2P2 traits at $p < 0.05$ (Figure 3A). However, several phenotypes are closely related and therefore these cross-phenotype associations do not reflect true pleiotropy, multiple unrelated effects due to the same gene (Solovieff et al., 2013). For example, rs2869462 had five cross-phenotype associations but four of these traits are based on CXCL10 levels (Figure 3B). Beyond the existence of pleiotropy, the pattern of which traits shared genetic associations provided additional insight (Figure 3C). A circle plot of cross-phenotype associations showed most cross-phenotype associations in H2P2 connect invasion, establishment of an intracellular niche, and intercellular spread. Traits that had high phenotypic correlation (Figure 1B) were more likely to have cross-phenotypic associations as expected (Figure 3D).

The greatest number of associated traits at $p < 0.05$ occurred for rs953897, a SNP in the gene encoding the transcriptional repressor ZBTB20. Based on GTEx data, rs953897 is associated with *ZBTB20-AS1* transcript abundance ($p = 1 \times 10^{-9}$) (GTEx Consortium, 2015). While rs953897 is most strongly associated with high *C. trachomatis* burden at 46 hr ($p = 1.3 \times 10^{-8}$), this SNP was associated with 20 H2P2 traits ($p < 0.05$) and 5 traits using a multiple test-corrected threshold of $p < 6.3 \times 10^{-4}$. A Q-Q plot comparing p values for all traits for rs953897 confirmed the high degree of pleiotropy, demonstrating strong deviation from neutrality toward lower p values (Figure 3D). These associations even included other pathogens and biological processes, as the third most strongly associated trait was *S.* Typhimurium-induced pyroptosis ($p = 7.5 \times 10^{-5}$).

## ZBTB20 Affects the Outcome of *Chlamydia* and *Salmonella* Infections

The T allele of rs953897 was associated with both higher levels of *Chlamydia* replication (Figure 4A) and *Salmonella*-induced pyroptosis (Figure 4B). Reduction of *ZBTB20* expression by RNAi increased *Chlamydia* replication and pyroptosis, mimicking the effect of the T allele (Figures 4C and 4D). Depletion of ZBTB20 also indicated that some phenotypes that did not reach statistically significant associations with rs953897 after multiple-test correction were nonetheless mediated by ZBTB20. Specifically, expression of IL-6 after infection with *Chlamydia* ($p = 0.03$) was reduced after *ZBTB20* RNAi treatment, again mimicking the effect of the T allele (Figure 4E). Thus, the association data and functional validation point to a role for ZBTB20 in the regulation of multiple infection-related phenotypes.

ZBTB20 has been characterized as a transcriptional repressor during prenatal development in liver and brain (Mitchelmore et al., 2002; Xie et al., 2008). Rare protein-coding mutations in *ZBTB20* are responsible for Primrose syndrome, which has features as disparate as mental retardation, ossified external ears, and distal muscle wasting (Cordeddu et al., 2014). We hypothesized that one ZBTB20 target gene might regulate a pathway that impacted multiple biological processes related to pathogen immunity. An attractive target, IκB (*NFKBIA*), the canonical suppressor of nuclear factor κB (NF-κB) signaling, is subject to ZBTB20 transcriptional repression (Liu et al., 2013). Reduction of ZBTB20 expression in LCLs by RNAi caused a moderate increase in expression of IκB (Figure 4F). An increase of IκB should cause inhibition of NF-κB signaling, resulting in increased *Chlamydia* replication but decreased expression of pro-inflammatory cytokines including IL-6. Consistent with this prediction, depletion of IκB decreased *Chlamydia* replication and increased IL-6 production (Figures 4C and 4E). In contrast, depletion of IκB did not impact *Salmonella*-induced pyroptosis, indicating an IκB-independent mechanism (Figure 4D). These observations point to multiple roles for ZBTB20 in regulating cellular functions during infection, both through suppression of regulators of signaling pathways, such as NF-κB, but also through regulation of other unidentified targets.

**Figure 3. Cross-Phenotype Associations and Pleiotropy Are Abundant among Cellular Host-Pathogen Traits**

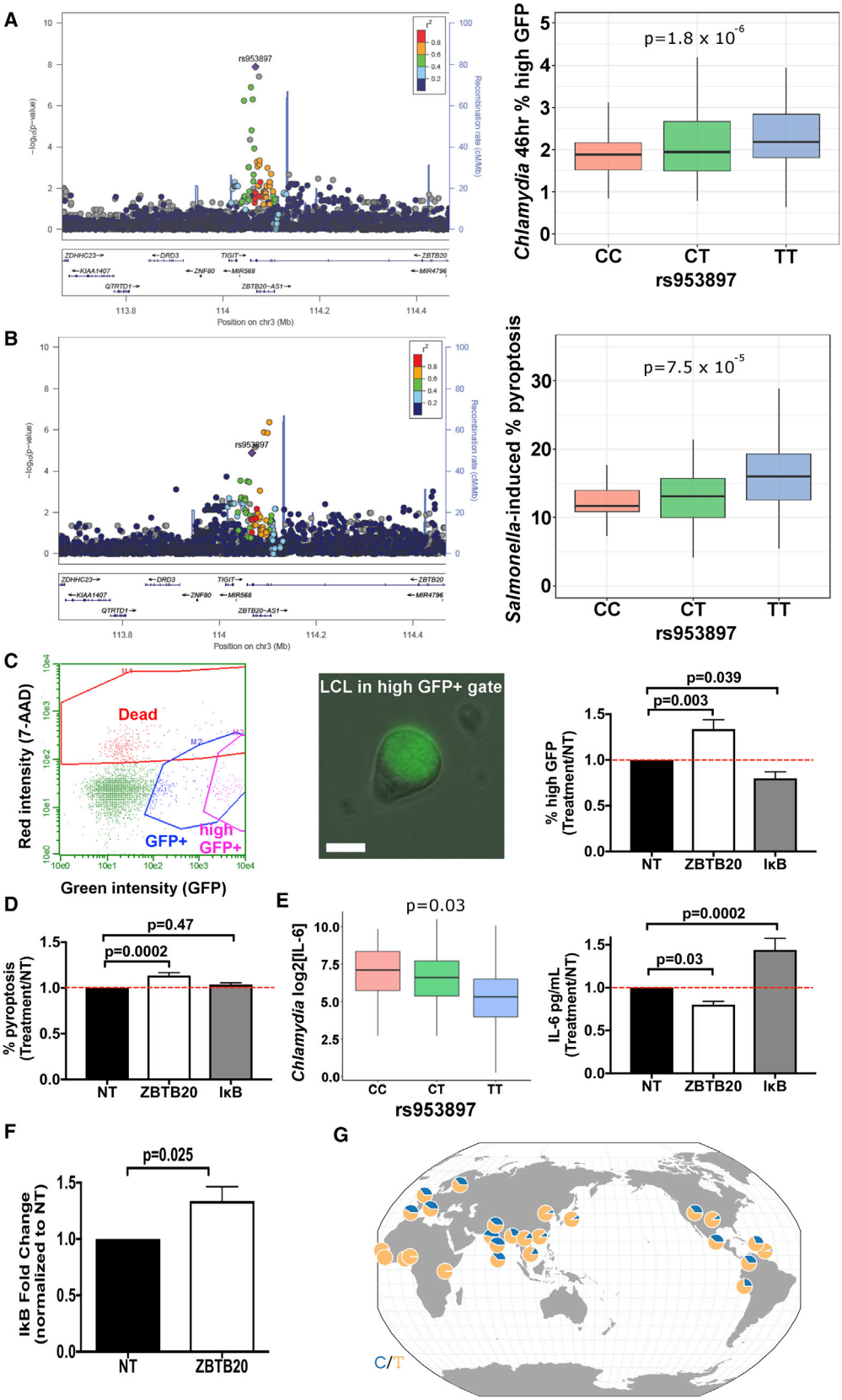
(A) Histograms of the number of cross-phenotype associations for the 17 H2P2 genome-wide significant hits. For the most strongly associated SNP in each GWAS peak, the number of traits with associations at $p < 0.05$ and $p < 6.33 \times 10^{-4}$ is shown.

(B) Q-Q plot and PheWAS plot for the association of rs2869462 with the 79 H2P2 phenotypes showed deviation from neutral expectation only for the four CXCL10 phenotypes. Gray shading indicates 95% confidence intervals.

(C) Circle plot of 79 phenotypes by category and lines connecting traits that share the same genome-wide significant hit at $p < 1 \times 10^{-5}$.

(D) Plot of pairwise trait phenotypic similarity (Spearman correlation) versus similarity of shared SNPs (Jaccard index). Traits that were more phenotypically similar have more shared SNPs with $p < 1 \times 10^{-3}$ for both traits.

(E) Q-Q plot and PheWAS plot for the association of rs953897 with the 79 H2P2 phenotypes showed deviation from neutral expectation for dozens of phenotypes, including traits in different biological categories in the PheWAS plot.




(legend on next page)

Interestingly, the geographic distribution of rs953897 was similar to rs2869462 (Figure 4G, compare with Figure 2F). For both SNPs, the derived allele (C for rs953897 and G for rs2869462) is more common in non-African populations (see Table S4). However, the effects of these alleles on inflammatory cytokine production are in opposite directions: the rs953897 C allele is associated with more IL-6, while rs2869462 G is associated with less CXCL10, suggesting complex evolutionary forces in shaping diversity of host-pathogen traits.

## SNPs Linked to CXCL10 Expression Are Associated with Inflammatory Bowel Disease

We determined if SNPs associated with cellular traits in H2P2 were associated with human disease. Consistent with the rs2869462 C allele being associated with increased CXCL10 and inflammation, we discovered that this allele is a previously unrecognized inflammatory bowel disease (IBD) risk allele. CXCL10 inhibitory antibodies have undergone phase II clinical trials for both subtypes of IBD, Crohn's disease (CD), and ulcerative colitis (UC) (Sandborn et al., 2016, 2017), based on evidence from animal models (Hyun et al., 2005) and of elevated levels of CXCL10 in patients (Ostvik et al., 2013). Examination of GWAS summary statistics from the IBD Genetics Consortium meta-analysis of 12,882 IBD cases and 21,770 controls (Liu et al., 2015) demonstrated that rs2869462 is associated with IBD ($p = 1.7 \times 10^{-4}$; odds ratio [OR] = 1.08), as well as with CD ($p = 1.9 \times 10^{-3}$; OR = 1.09) and UC subtypes (p = 0.016; OR = 1.06) separately. The direction of association is consistent with high levels of CXCL10 (the C allele) being associated with greater risk of IBD.

We conducted colocalization analysis to determine whether the CXCL10 protein level signal was the same as the IBD signal. We utilized COLOC, which uses a Bayesian framework to determine whether GWAS signals in the same region are likely due to the same causal variant (Giambartolomei et al., 2014). The posterior probability that both CXCL10 protein level and IBD share the same causal variant is as high as 0.80 for the region, with rs2869462 identified as the most likely causal SNP (Table S5). Comparison of regional plots of association indicate that the same linkage disequilibrium block is associated with both CXCL10 levels and IBD (Figure 5A).

We independently tested for this association using electronic medical record (EMR) data from the eMERGE Network (McCarty et al., 2011). The eMERGE dataset holds genotype-phenotype correlations of >80,000 individuals with phenotypes assigned based on ICD-9 patient billing codes (Denny et al., 2013). We tested for association with the codes for "inflammatory bowel disease and other gastroenteritis and colitis" and the more restrictive code for "ulcerative colitis." rs2869462 was associated with both phenotypes in the predicted direction (p = 0.003; OR = 1.12, C allele for IBD, p = 0.02; OR = 1.12, C allele for UC) (Figure 5B). Therefore by first identifying a SNP associated with CXCL10 levels in an LCL model of *C. trachomatis* infection, we have discovered and replicated an IBD risk allele.

## H2P2 SNPs Are Associated with Disease in PheWAS of EMR Traits

Next, we systematically generated hypotheses regarding the effects of the H2P2 genome-wide significant hits on human disease by employing a PheWAS (phenome-wide association study) approach, looking for associations across a large set (1,338) of clinical measurements and diseases cataloged in eMERGE (Denny et al., 2013). Five of 16 H2P2 genome-wide significant hits surpassed the multiple test corrected significance threshold with at least one EMR phenotype (Figure 6A; Table S6; 1 H2P2 SNP was not in eMERGE and had no good proxy).

The SNP with the greatest number of PheWAS associations was rs7566597. This SNP, associated with the H2P2 phenotype of *Chlamydia*-infected levels of the chemokine MIP-1β ([macrophage inflammatory protein-1beta] CCL4), was associated with five clinical traits. The most significant association ($p = 1.76 \times 10^{-6}$; p = 0.0024 after Bonferroni; OR = 1.40) was observed with otorrhea (Figure 6B). Otorrhea is ear drainage most commonly caused by an ear infection. MIP-1β is elevated in fluid from patients with middle ear infections (Kaur et al., 2015), mice with genetic predisposition to middle ear infections (Han et al., 2012), and primary middle ear epithelial cultures infected with influenza or *Streptococcus pneumoniae* (Tong et al., 2003). The directionality of this association (G allele associated with both higher MIP-1β levels and increased risk of otorrhea), and the fact that genotype is fixed prior to disease, lead to the hypothesis that the rs7566597 G allele causes higher levels of MIP-1β to increase risk of otorrhea.

eMERGE PheWAS also revealed that rs953897 in *ZBTB20* was associated with viral hepatitis (Figure 6C; $p = 3.17 \times 10^{-5}$;

**Figure 4. Genetic Variation Influencing *ZBTB20* Regulates Multiple Host-Pathogen Traits**

(A) Regional plot for *ZBTB20* demonstrated an association of rs953897 with *C. trachomatis* high GFP-infected cells at 46 hr ($p = 1.3 \times 10^{-8}$). Genotypic medians plot (with first and third quartiles [box] and maximum and minimum values [whiskers]) of rs953897 with high GFP-infected cells at 46 hr in IBS LCLs.

(B) Regional plot for ZBTB20 demonstrated an association of rs953897 with *S.* Typhimurium-induced pyroptosis ($p = 7.5 \times 10^{-5}$). Genotypic medians plot of rs953897 with *S.* Typhimurium-induced pyroptosis in all LCLs.

(C–F) LCL GM1761 was treated with NT, ZBTB20 (53% ± 9% knockdown), or IκB (83% ± 2% knockdown) Accell RNAi for 3 days prior to infection. Measurements were normalized to NT prior to statistical analysis. Means (±SEM) were plotted. (C) ZBTB20 and IκB regulate *C. trachomatis* replication. By 46hr, *C. trachomatis* replication resulted in a high GFP + population of cells with an enlarged GFP+ *Chlamydia*-containing vacuole. Scale bar, 10 μm. ZBTB20 knockdown resulted in a greater percentage of high GFP+ cells, similar to what was seen with the T allele, while IκB knockdown produced fewer. Mean (±SEM) percentage of high GFP+ cells in NT samples was 1.63% (±0.07%). (D) ZBTB20 regulates *Salmonella*-induced pyroptosis independent of IκB. ZBTB20 knockdown resulted in a greater percentage of pyroptotic cells, similar to what is seen with the T allele, while IκB knockdown showed no significant change. Percentage of pyroptotic cells in NT samples was 35.1% (±1.4%). (E) Both the T allele and ZBTB20 knockdown result in reduced IL-6. Genotypic median plot of rs953897 with *C. trachomatis*-induced IL-6 in all LCLs. ZBTB20 knockdown reduced IL-6 levels, while knockdown of IκB led to increased levels measured at 70 hr. IL-6 levels from NT LCLs were 186 pg/mL (±31.6 pg/mL). (F) ZBTB20 knockdown increased IκB mRNA (normalized by 18 s).

(G) Map of rs953897 allele frequencies (T, orange; C, blue) from Geography of Genetic Variants Browser (Marcus and Novembre, 2017). (C)–(F) were from 8 to 12 biological replicates from 2 to 4 experiments. p values for (C–E) were generated from one-way ANOVA analysis while (F) was calculated by an unpaired t test. p values in genotypic median plots (A, B, and E) were generated with QFAM-parents in PLINK.

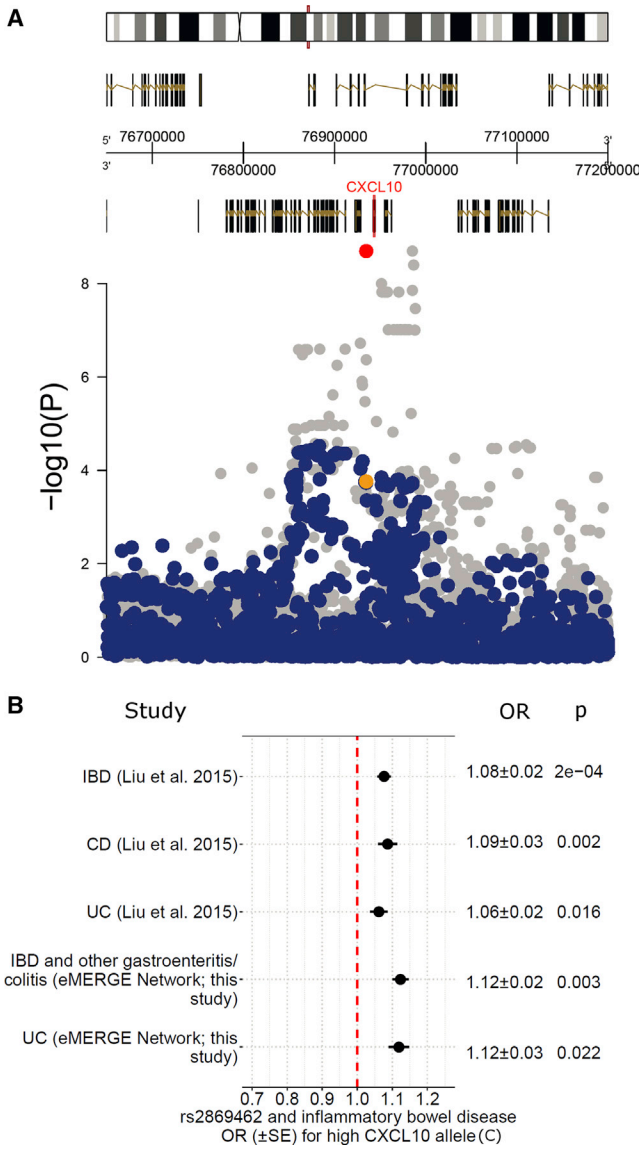



**Figure 5. SNPs Associated with High CXCL10 in H2P2 Are Associated with Increased Risk of IBD**

(A) Overlaid association plots of the CXCL10 region demonstrated colocalization of signals for *Chlamydia*-infected CXCL10 levels from H2P2 (gray) and IBD GWAS (blue) (Liu et al., 2015). rs2869462 is highlighted in red and yellow.
(B) OR plot for rs2869462 and IBD, CD, and UC based on data generated in Liu et al. (2015) and replication of the association with "IBD and other gastroenteritis and colitis" and UC from the eMERGE Network.
See also Table S5.

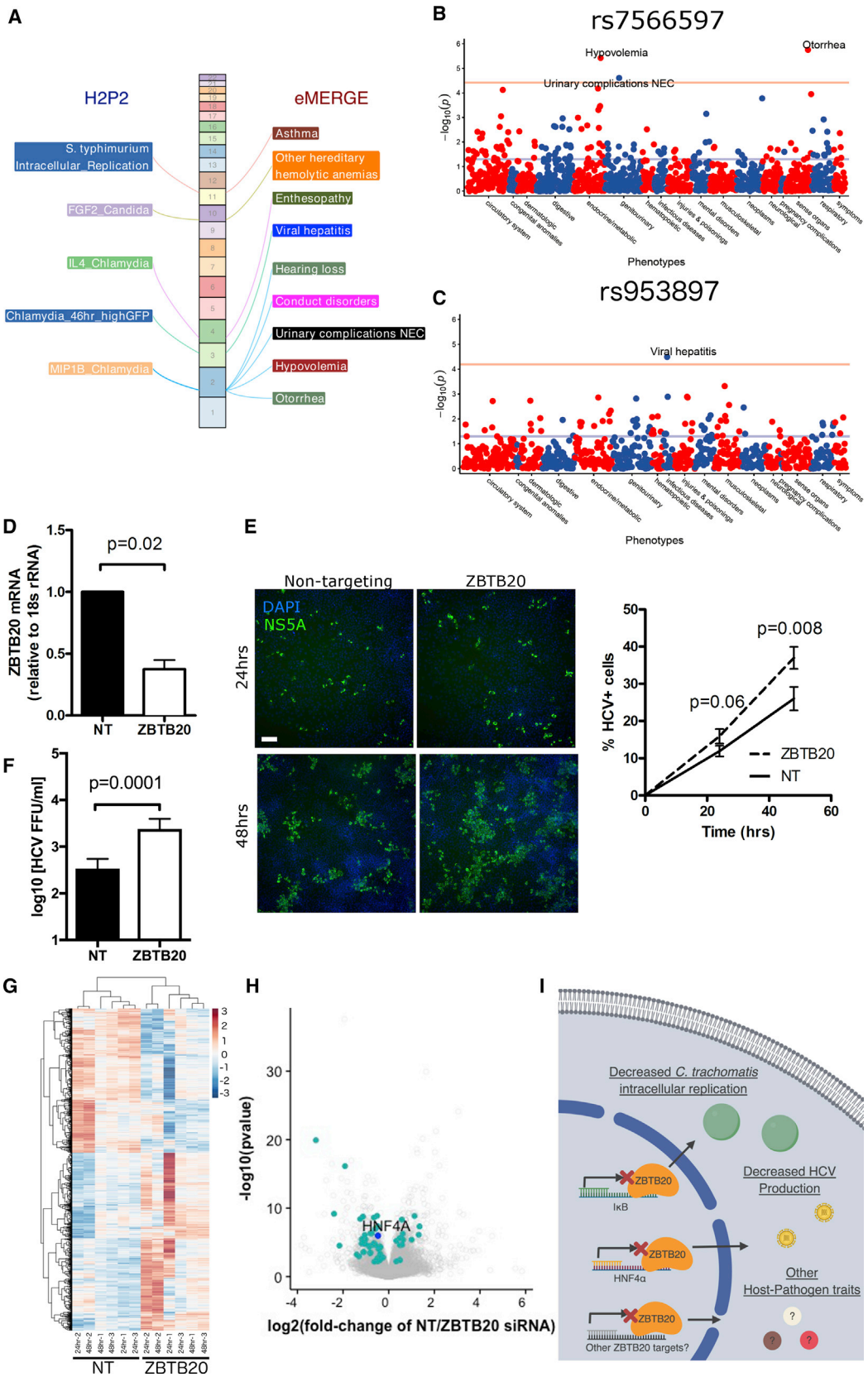
p = 0.025 after Bonferroni; OR = 1.20). To test this association experimentally, we performed RNAi against *ZBTB20* in Huh7 human hepatocytes. Depletion of *ZBTB20* mRNA (Figure 6D) increased the percentage of hepatitis C virus (HCV) infected cells over time (Figure 6E) and increased infectious virus production by 7-fold (Figure 6F). To determine what downstream targets of ZBTB20 repression might be mediating this increase, we performed RNA sequencing of uninfected Huh7 cells treated either with non-targeting (NT) or *ZBTB20* small interfering RNA (siRNA). A total of 1,123 genes were upregulated in *ZBTB20* siRNA-treated compared with NT (at false discovery rate [FDR]-corrected p = 0.05; Figure 6G; Table S7). Two of the top 10 gene sets enriched in upregulated genes were targets of the transcriptional activator hepatocyte nuclear factor 4 alpha (HNF4α) (Figure 6H; Table S7). HNF4α was 28% lower in NT compared with *ZBTB20* siRNA-treated (FDR-corrected $p = 6 \times 10^{-5}$), consistent with ZBTB20 suppression of HNF4α. Remarkably, depletion of HNF4α by a similar magnitude has been demonstrated to cause a 3-fold decrease in HCV production in hepatocytes (Li et al., 2014), similar to the 7-fold lower levels we observed with NT compared with *ZBTB20* siRNA (Figure 6F). Thus, genetic variation in *ZBTB20* has broad pleiotropic effects likely being mediated by suppression of different transcriptional targets, including HNF4α in hepatocytes (Figure 6I). This example demonstrates that combining H2P2 with PheWAS of clinical traits can lead to hypotheses that can be quickly tested in the most clinically relevant cell type.

## DISCUSSION

With H2P2, we have coupled the ability of pathogens to influence human genetic diversity to their use as cellular probes to elucidate mechanisms of disease. This cellular GWAS approach reveals: (1) molecules that could serve as possible biomarkers and therapeutic targets and (2) cellular models for validation and mechanistic dissection. We have developed an H2P2 web database to allow for exploration of this rich dataset (http://h2p2.oit.duke.edu).

Cellular GWAS studies have also been performed on levels of immune cells, cell surface proteins, and cytokines (Mikacenic et al., 2013; Orru et al., 2013; Roederer et al., 2015). However, Hi-HOST is unique in using live pathogens to induce complex cellular phenotypes, such as cell death and invasion, providing phenotypes intermediate between molecular phenotypes of gene/protein expression and human studies of disease. In addition, our use of parent-offspring trios allowed us to make estimates of $h^2$ through both parent-offspring regression and SNP-based methods that correlated quite strongly. Our estimates of $h^2$ are similar to other reports for immune-related traits. Orru et al. (2013) examined levels of 95 immune cell types and found mean $h^2$ of 41%. Our results are consistent with a strong genetic basic for variation in immune cell traits and host-pathogen interactions, although environment also has a large effect.

Integrating H2P2 with human genetic association data revealed how genetic variants impacting cellular traits also influenced human disease. While over a hundred IBD risk alleles have been identified (Liu et al., 2015), the fact rs2869462 was associated with levels of CXCL10 in H2P2 may make genotyping this SNP clinically actionable if coupled to anti-CXCL10 therapy. While anti-CXCL10 demonstrated some benefit in phase 2 trials, neither study met statistical significance (Sandborn et al., 2016, 2017). We hypothesize that rs2869462 genotype might be a predictive biomarker for identifying the genetic subtype of patients who show greatest benefit. This example, spanning molecular phenotype, human disease, and clinical utility serves as a template for how we envision the H2P2 web portal being used to make similar discoveries.




(legend on next page)

For *CXCL10*, *ZBTB20*, and other genes implicated by H2P2, there are numerous associations that do not reach genome-wide significance but are undoubtedly true-positives based on highly related phenotypes or experimental evidence. Indeed, our lab previously pursued non-genome-wide significant hits revealed by Hi-HOST, resulting in the identification of a metabolite biomarker for sepsis (Wang et al., 2017) and suggesting a potential therapeutic strategy for typhoid fever (Alvarez et al., 2017). However, to fully illuminate how genetic variation contributes to pathophysiology of disease will require the engagement of the research community with the H2P2 web portal and similar datasets. These users, already experts on particular genes and/or cellular pathways, would be well-equipped to then validate and discover the mechanisms underlying these associations. Future studies will expand the panel of stimuli and the cell types used in H2P2 to create a more complete picture of how cellular traits impact human health and disease, an important step toward a future of more personalized care.

Information about the eMERGE Network sites, leadership, and other details can be found at https://emerge.mc.vanderbilt.edu/.

## STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Cells
- METHOD DETAILS
  - LCL Screening
  - *C. trachomatis* Infection
  - *Salmonella* Infection
  - Fungal Infection
  - *S. aureus* Toxin Treatment
  - *T. gondii* Infection
  - LCL RNAi Experiments
  - HCV Infection Experiments
  - RNA-seq
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Phenotype Repeatability
  - Testing Effect of EBV Copy Number on H2P2 Cellular Phenotypes
  - Genotype and Imputation
  - Phenotype- and SNP-Based Heritability Analysis
  - Genome-wide Association Analysis
  - Permutation Analysis
  - Enrichment Analysis
  - PheWAS Analysis
  - Colocalization Analysis
  - Gene Expression Analysis on 1000 Genome RNA-seq Project
  - RNA-seq Analysis
  - Descriptive Statistics and Visualization
- DATA AND SOFTWARE AVAILABILITY

## SUPPLEMENTAL INFORMATION

Supplemental Information includes five figures, seven tables, and two data files and can be found with this article online at https://doi.org/10.1016/j.chom.2018.07.007.

## ACKNOWLEDGMENTS

L.W., K.J.P., R.E.S., K.D.G., A.L.A., and D.C.K. were supported by NIH grant R01AI118903. L.W., T.B., A.I., M.R.D., and D.C.K. were supported by NIH grant R21AI133305. L.W., K.J.P., J.R.B., R.E.S., A.M.G., R.H.V., and D.C.K. were supported by NIH grant U19AI084044. D.C.K. was supported by Duke University Whitehead Scholarship, Butler Pioneer Award, and Duke MGM Pilot Award. GDW and SMH were supported by NIH grants R01AI125416, R21AI124100 and Burroughs Wellcome Fund. I.B.S., R.J.C., J.C.D., G.P.J., and D.R.C. were supported by eMERGE Network (phase III), funded by the NHGRI through the following grants: U01HG8657 (Group Health Cooperative/U. of Washington); U01HG8685 (Brigham and Women's Hospital); U01HG8672 (Vanderbilt University Medical Center); U01HG8666 (Cincinnati Children's Hospital Medical Center); U01HG6379 (Mayo Clinic); U01HG8679 (Geisinger Clinic); U01HG8680 (Columbia University Health Sciences); U01HG8684 (Children's Hospital of Philadelphia); U01HG8673 (Northwestern University); U01HG8701 (Vanderbilt University Medical Center serving as the Coordinating Center); U01HG8676 (Partners Healthcare/Broad Institute); and U01HG8664 (Baylor College of Medicine). U01HG004438 (CIDR) and U01HG004424 (Broad Institute) served as eMERGE Genotyping Centers. S.C.L. was supported by NIH R03AI11917 and UTSA. J.H. was supported by NIH R37AI39115, R01AI50113, and an award from Astellas. T.B., A.I., and M.R.D. were supported by Duke Research Computing. We thank Yanlu Cao for performing FGF2 ELISAs, Luke Glover for help in plotting association data, Charles Rice for NS5A antibody, the Duke Immune Reconstitution & Biomarker Analysis Shared Resource in performing pilot Luminex measurements and the Duke Functional Genomics Shared Resource for use of Cellomics ArrayScan. DNA image from Figure 1 is from NHGRI Image Gallery. The content of this manuscript is solely the responsibility of the authors and

**Figure 6. PheWAS of H2P2 Hits Reveals Connections to Human Disease Including an Association of *ZBTB20* and Viral Hepatitis**

(A) Chromosome landscape of H2P2 genome-wide significant hits that demonstrate significant associations with eMERGE phenotypes. The five SNPs demonstrating significant associations are described in Table S6.

(B and C) eMERGE PheWAS plots of rs7566597 (B) and rs953897 (C). Red line indicates Bonferroni corrected p value = 0.05.

(D) qPCR demonstrated significant knockdown of ZBTB20 in Huh7 cells transfected with NT or ZBTB20 siRNA for 2 days.

(E) ZBTB20 suppresses hepatitis C virus (HCV) infection. The percentage of cells infected with HCV was assessed by immunofluorescence for HCV protein NS5A (green) and staining with DAPI (blue) for total cells. Scale bar, 50 μm. ZBTB20 depletion caused the percentage of HCV-infected cells to increase more rapidly.

(F) ZBTB20 suppresses HCV production. Supernatants collected at 72 hr post-infection were used in an HCV focus-forming assay to determine the concentration of productive HCV particles. ZBTB20 depletion caused a 7-fold increase in HCV focus-forming units (FFUs/mL). For (D)–(F), mean (±SEM) are shown and p values are from a paired t tests from three separate experiments.

(G) Heatmap of differentially expressed genes (FDR <= 0.05 and at least 2-fold change) for NT versus *ZBTB20* siRNA. Gene expression has been *Z* score normalized and samples (labeled by time point and experiment number) and genes are clustered by correlation distance with complete linkage.

(H) Volcano plot of $\log_2$(fold-change) versus $-\log_{10}$(p value) demonstrates upregulation of HNF4α targets (turquoise) with depletion of *ZBTB20*.

(I) Model of how ZBTB20 transcriptional repression affects multiple host-pathogen traits.

See also Tables S6 and S7.

does not necessarily represent the official views of the NIH or other funding sources.

## AUTHOR CONTRIBUTIONS

All authors critically reviewed the manuscript and contributed input to the final submission. D.C.K., L.W., K.J.P., A.L.A., I.B.S., G.D.W., T.B., S.M.H., R.H.V., and D.R.C. wrote the manuscript. D.C.K., L.W., K.J.P., A.L.A., J.H., S.C.L., S.M.H., R.H.V., and D.R.C. contributed to strategy and project planning. K.J.P., D.C.K., J.R.B., R.E.S., G.D.W., A.M.G., K.D.G., A.L.A., and S.C.L. carried out the experiments and analysis. L.W., I.B.S., R.J.C., The eMERGE Network, G.P.J., J.C.D., D.R.C., and D.C.K. planned and carried out computational analysis and provided datasets. T.B., A.I., L.W., M.R.D., and D.C.K. designed and implemented the H2P2 database and web portal.

## DECLARATION OF INTERESTS

Duke University has submitted a provisional patent application ("A Companion Diagnostic for IBD Therapy") on behalf of D.C.K., L.W., and A.L.A.

Received: March 5, 2018
Revised: May 28, 2018
Accepted: July 5, 2018
Published: August 8, 2018

## REFERENCES

1000 Genomes Project Consortium, Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E., and McVean, G.A. (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061–1073.

Aligeti, M., Roder, A., and Horner, S.M. (2015). Cooperation between the hepatitis C virus p7 and NS5B proteins enhances virion infectivity. J. Virol. *89*, 11523–11533.

Allison, A.C. (1954). Protection afforded by sickle-cell trait against subtertian malareal infection. Br. Med. J. *1*, 290–294.

Alvarez, M.I., Glover, L.C., Luo, P., Wang, L., Theusch, E., Oehlers, S.H., Walton, E.M., Tram, T.T.B., Kuang, Y.L., Rotter, J.I., et al. (2017). Human genetic variation in VAC14 regulates *Salmonella* invasion and typhoid fever through modulation of cholesterol. Proc. Natl. Acad. Sci. USA *114*, E7746–E7755.

Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics *31*, 166–169.

Bastidas, R.J., and Valdivia, R.H. (2016). Emancipating *Chlamydia*: advances in the genetic manipulation of a recalcitrant intracellular pathogen. Microbiol. Mol. Biol. Rev. *80*, 411–427.

Beuzon, C.R., Meresse, S., Unsworth, K.E., Ruiz-Albert, J., Garvis, S., Waterman, S.R., Ryder, T.A., Boucrot, E., and Holden, D.W. (2000). *Salmonella* maintains the integrity of its intracellular vacuole through the action of SifA. EMBO J. *19*, 3235–3249.

Burton, M.J., and Mabey, D.C. (2009). The global burden of trachoma: a review. PLoS Negl. Trop. Dis. *3*, e460.

Cahir-McFarland, E.D., Carter, K., Rosenwald, A., Giltnane, J.M., Henrickson, S.E., Staudt, L.M., and Kieff, E. (2004). Role of NF-kappa B in cell survival and transcription of latent membrane protein 1-expressing or Epstein-Barr virus latency III-infected cells. J. Virol. *78*, 4108–4119.

Carroll, R.J., Bastarache, L., and Denny, J.C. (2014). R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. Bioinformatics *30*, 2375–2376.

Cavalli-Sforza, L.L., Menozzi, P., and Piazza, A. (1994). The History and Geography of Human Genes (Princeton University Press).

Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience *4*, 7.

Collazo, C.M., and Galan, J.E. (1997). The invasion-associated type-III protein secretion system in *Salmonella* – a review. Gene *192*, 51–59.

Cordeddu, V., Redeker, B., Stellacci, E., Jongejan, A., Fragale, A., Bradley, T.E., Anselmi, M., Ciolfi, A., Cecchetti, S., Muto, V., et al. (2014). Mutations in ZBTB20 cause Primrose syndrome. Nat. Genet. *46*, 815–817.

Cossart, P., Boquet, P., Normark, S., and Rappuoli, R. (1996). Cellular microbiology emerging. Science *271*, 315–316.

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. USA *97*, 6640–6645.

Delaneau, O., Zagury, J.F., and Marchini, J. (2013). Improved whole-chromosome phasing for disease and population genetic studies. Nat. Methods *10*, 5–6.

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. Nat. Biotechnol. *31*, 1102–1110.

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

Dougan, G., and Baker, S. (2014). *Salmonella enterica* serovar Typhi and the pathogenesis of typhoid fever. Annu. Rev. Microbiol. *68*, 317–336.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Falconer, D.S., and Mackay, T.F.C. (1996). Introduction to Quantitative Genetics, Fourth Edition (Longmans Green).

Flutre, T., Wen, X., Pritchard, J., and Stephens, M. (2013). A statistical framework for joint eQTL analysis in multiple tissues. PLoS Genet. *9*, e1003486.

Friedman, M.J. (1978). Erythrocytic mechanism of sickle cell resistance to malaria. Proc. Natl. Acad. Sci. USA *75*, 1994–1997.

Fumagalli, M., Sironi, M., Pozzoli, U., Ferrer-Admetlla, A., Pattini, L., and Nielsen, R. (2011). Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. PLoS Genet. *7*, e1002355.

Furtado, J.M., Smith, J.R., Belfort, R., Jr., Gattey, D., and Winthrop, K.L. (2011). Toxoplasmosis: a global threat. J. Glob. Infect. Dis. *3*, 281–284.

Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. *5*, R80.

Giambartolomei, C., Vukcevic, D., Schadt, E.E., Franke, L., Hingorani, A.D., Wallace, C., and Plagnol, V. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet. *10*, e1004383.

Groom, J.R., and Luster, A.D. (2011). CXCR3 ligands: redundant, collaborative and antagonistic functions. Immunol. Cell Biol. *89*, 207–215.

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

Han, F., Yu, H., Li, P., Zhang, J., Tian, C., Li, H., and Zheng, Q.Y. (2012). Mutation in Phex gene predisposes BALB/c-Phex(Hyp-Duk)/Y mice to otitis media. PLoS One *7*, e43010.

Harrel, F.E., and Davis, C.E. (1982). A new distribution-free quantile estimator. Biometrika *69*, 635–640.

Howie, B.N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. *5*, e1000529.

Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., and Abecasis, G.R. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat. Genet. *44*, 955–959.

Hyun, J.G., Lee, G., Brown, J.B., Grimm, G.R., Tang, Y., Mittal, N., Dirisina, R., Zhang, Z., Fryer, J.P., Weinstock, J.V., et al. (2005). Anti-interferon-inducible chemokine, CXCL10, reduces colitis by impairing T helper-1 induction and recruitment in mice. Inflamm. Bowel Dis. *11*, 799–805.

International HapMap Consortium (2005). A haplotype map of the human genome. Nature *437*, 1299–1320.

Iotchkova, V., Ritchie, G.R.S., Geihs, M., Morganella, S., Min, J.L., Walter, K., Timpson, N.J., Consortium, U.K., Dunham, I., Birney, E., et al. (2016). GARFIELD–GWAS analysis of regulatory or functional information enrichment with LD correction. bioRxiv. https://doi.org/10.1101/085738.

Kanai, M., Tanaka, T., and Okada, Y. (2016). Empirical estimation of genome-wide significance thresholds based on the 1000 Genomes Project data set. J. Hum. Genet. *61*, 861–866.

Kaur, R., Casey, J., and Pichichero, M. (2015). Cytokine, chemokine, and Toll-like receptor expression in middle ear fluids of children with acute otitis media. Laryngoscope *125*, E39–E44.

Kersey, P.J., Staines, D.M., Lawson, D., Kulesha, E., Derwent, P., Humphrey, J.C., Hughes, D.S.T., Keenan, S., Kerhornou, A., Koscielny, G., et al. (2012). Ensembl genomes: an integrative resource for genome-scale data from non-vertebrate species. Nucleic Acids Res. *40*, D91–D97.

Kirk, M.D., Pires, S.M., Black, R.E., Caipo, M., Crump, J.A., Devleesschauwer, B., Dopfer, D., Fazil, A., Fischer-Walker, C.L., Hald, T., et al. (2015). World Health Organization estimates of the global and regional disease burden of 22 foodborne bacterial, protozoal, and viral diseases, 2010: a data synthesis. PLoS Med. *12*, e1001921.

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., 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. USA *109*, E2343–E2352.

Krueger, F. (2017). Trim Galore! Available at: http://www.bioinformatics.babraham.ac.uk/projects/trim_galore.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: an information aesthetic for comparative genomics. Genome Res. *19*, 1639–1645.

Lappalainen, T., Sammeth, M., Friedlander, M.R., t Hoen, P.A., Monlong, J., Rivas, M.A., Gonzalez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G., et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature *501*, 506–511.

Lee, S.C., Li, A., Calo, S., Inoue, M., Tonthat, N.K., Bain, J.M., Louw, J., Shinohara, M.L., Erwig, L.P., Schumacher, M.A., et al. (2015). Calcineurin orchestrates dimorphic transitions, antifungal drug responses and host-pathogen interactions of the pathogenic mucoralean fungus *Mucor circinelloides*. Mol. Microbiol. *97*, 844–865.

Li, J.Z., Absher, D.M., Tang, H., Southwick, A.M., Casto, A.M., Ramachandran, S., Cann, H.M., Barsh, G.S., Feldman, M., Cavalli-Sforza, L.L., et al. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. Science *319*, 1100–1104.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The sequence alignment/map format and SAMtools. Bioinformatics *25*, 2078–2079.

Li, C.H., Cervantes, M., Springer, D.J., Boekhout, T., Ruiz-Vazquez, R.M., Torres-Martinez, S.R., Heitman, J., and Lee, S.C. (2011). Sporangiospore size dimorphism is linked to virulence of *Mucor circinelloides*. PLoS Pathog. *7*, e1002086.

Li, X., Jiang, H., Qu, L., Yao, W., Cai, H., Chen, L., and Peng, T. (2014). Hepatocyte nuclear factor 4alpha and downstream secreted phospholipase A2 GXIIB regulate production of infectious hepatitis C virus. J. Virol. *88*, 612–627.

Liu, R., Paxton, W.A., Choe, S., Ceradini, D., Martin, S.R., Horuk, R., MacDonald, M.E., Stuhlmann, H., Koup, R.A., and Landau, N.R. (1996). Homozygous defect in HIV-1 coreceptor accounts for resistance of some multiply-exposed individuals to HIV-1 infection. Cell *86*, 367–377.

Liu, X., Zhang, P., Bao, Y., Han, Y., Wang, Y., Zhang, Q., Zhan, Z., Meng, J., Li, Y., Li, N., et al. (2013). Zinc finger protein ZBTB20 promotes Toll-like receptor-triggered innate immune responses by repressing IkappaBalpha gene transcription. Proc. Natl. Acad. Sci. USA *110*, 11097–11102.

Liu, J.Z., van Sommeren, S., Huang, H., Ng, S.C., Alberts, R., Takahashi, A., Ripke, S., Lee, J.C., Jostins, L., Shah, T., et al. (2015). Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nat. Genet. *47*, 979–986.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.

Mandage, R., Telford, M., Rodriguez, J.A., Farre, X., Layouni, H., Marigorta, U.M., Cundiff, C., Heredia-Genestar, J.M., Navarro, A., and Santpere, G. (2017). Genetic factors affecting EBV copy number in lymphoblastoid cell lines derived from the 1000 Genome Project samples. PLoS One *12*, e0179446.

Marcus, J.H., and Novembre, J. (2017). Visualizing the geography of genetic variants. Bioinformatics *33*, 594–595.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. *17*, 10–12.

McCarty, C.A., Chisholm, R.L., Chute, C.G., Kullo, I.J., Jarvik, G.P., Larson, E.B., Li, R., Masys, D.R., Ritchie, M.D., Roden, D.M., et al. (2011). The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC Med. Genomics *4*, 13.

Mendoza, L., Vilela, R., Voelz, K., Ibrahim, A.S., Voigt, K., and Lee, S.C. (2014). Human fungal pathogens of mucorales and entomophthorales. Cold Spring Harb. Perspect. Med. *5*, https://doi.org/10.1101/cshperspect.a019562.

Mikacenic, C., Reiner, A.P., Holden, T.D., Nickerson, D.A., and Wurfel, M.M. (2013). Variation in the TLR10/TLR1/TLR6 locus is the major genetic determinant of interindividual difference in TLR1/2-mediated responses. Genes Immun. *14*, 52–57.

Mitchelmore, C., Kjaerulff, K.M., Pedersen, H.C., Nielsen, J.V., Rasmussen, T.E., Fisker, M.F., Finsen, B., Pedersen, K.M., and Jensen, N.A. (2002). Characterization of two novel nuclear BTB/POZ domain zinc finger isoforms. Association with differentiation of hippocampal neurons, cerebellar granule cells, and macroglia. J. Biol. Chem. *277*, 7598–7609.

Mootha, V.K., Lindgren, C.M., Eriksson, K.-F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E., et al. (2003). PGC-1[alpha]-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat. Genet. *34*, 267–273.

Newman, L., Rowley, J., Vander Hoorn, S., Wijesooriya, N.S., Unemo, M., Low, N., Stevens, G., Gottlieb, S., Kiarie, J., and Temmerman, M. (2015). Global estimates of the prevalence and incidence of four curable sexually transmitted infections in 2012 based on systematic review and global reporting. PLoS One *10*, e0143304.

Nica, A.C., and Dermitzakis, E.T. (2013). Expression quantitative trait loci: present and future. Philos. Trans. R. Soc. Lond. B Biol. Sci. *368*, 20120362.

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.

Orru, V., Steri, M., Sole, G., Sidore, C., Virdis, F., Dei, M., Lai, S., Zoledziewska, M., Busonero, F., Mulas, A., et al. (2013). Genetic variants regulating immune cell levels in health and disease. Cell *155*, 242–256.

Ostvik, A.E., Granlund, A.V., Bugge, M., Nilsen, N.J., Torp, S.H., Waldum, H.L., Damas, J.K., Espevik, T., and Sandvik, A.K. (2013). Enhanced expression of CXCL10 in inflammatory bowel disease: potential role of mucosal Toll-like receptor 3 stimulation. Inflamm. Bowel Dis. *19*, 265–274.

Parkhill, J., Dougan, G., James, K.D., Thomson, N.R., Pickard, D., Wain, J., Churcher, C., Mungall, K.L., Bentley, S.D., Holden, M.T., et al. (2001). Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18. Nature *413*, 848–852.

Pruim, R.J., Welch, R.P., Sanna, S., Teslovich, T.M., Chines, P.S., Gliedt, T.P., Boehnke, M., Abecasis, G.R., and Willer, C.J. (2010). LocusZoom: regional

visualization of genome-wide association scan results. Bioinformatics *26*, 2336–2337.

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.

Purcell, S., Sham, P., and Daly, M.J. (2005). Parental phenotypes in family-based association analysis. Am. J. Hum. Genet. *76*, 249–259.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.

Roederer, M., Quaye, L., Mangino, M., Beddall, M.H., Mahnke, Y., Chattopadhyay, P., Tosi, I., Napolitano, L., Terranova Barberio, M., Menni, C., et al. (2015). The genetic architecture of the human immune system: a bioresource for autoimmunity and disease pathogenesis. Cell *161*, 387–403.

Roncero, M.I.G. (1984). Enrichment method for the isolation of auxotrophic mutants of Mucor using the polyene antibiotic N-glycosyl-polifungin. Carlsberg Res. Commun. *49*, 685.

Saeij, J.P., Boyle, J.P., Grigg, M.E., Arrizabalaga, G., and Boothroyd, J.C. (2005). Bioluminescence imaging of *Toxoplasma gondii* infection in living mice reveals dramatic differences between strains. Infect. Immun. *73*, 695–702.

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. Mol. Microbiol. *82*, 1185–1203.

Sandborn, W.J., Colombel, J.F., Ghosh, S., Sands, B.E., Dryden, G., Hebuterne, X., Leong, R.W., Bressler, B., Ullman, T., Lakatos, P.L., et al. (2016). Eldelumab [anti-IP-10] induction therapy for ulcerative colitis: a randomised, placebo-controlled, phase 2b study. J. Crohns Colitis *10*, 418–428.

Sandborn, W.J., Rutgeerts, P., Colombel, J.F., Ghosh, S., Petryka, R., Sands, B.E., Mitra, P., and Luo, A. (2017). Eldelumab [anti-interferon-gamma-inducible protein-10 antibody] induction therapy for active Crohn's disease: a randomised, double-blind, placebo-controlled phase IIa study. J. Crohns Colitis *11*, 811–819.

Solovieff, N., Cotsapas, C., Lee, P.H., Purcell, S.M., and Smoller, J.W. (2013). Pleiotropy in complex traits: challenges and strategies. Nat. Rev. Genet. *14*, 483–495.

Stranger, B.E., Nica, A.C., Forrest, M.S., Dimas, A., Bird, C.P., Beazley, C., Ingle, C.E., Dunning, M., Flicek, P., Koller, D., et al. (2007). Population genomics of human gene expression. Nat. Genet. *39*, 1217–1224.

Team, R.C. (2016). R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing). http://www.R-project.org/.

Tong, H.H., Long, J.P., Shannon, P.A., and DeMaria, T.F. (2003). Expression of cytokine and chemokine genes by human middle ear epithelial cells induced by influenza A virus and *Streptococcus pneumoniae* opacity variants. Infect. Immun. *71*, 4289–4296.

Tong, S.Y., Davis, J.S., Eichenberger, E., Holland, T.L., and Fowler, V.G., Jr. (2015). *Staphylococcus aureus* infections: epidemiology, pathophysiology, clinical manifestations, and management. Clin. Microbiol. Rev. *28*, 603–661.

Visscher, P.M., Hill, W.G., and Wray, N.R. (2008). Heritability in the genomics era – concepts and misconceptions. Nat. Rev. Genet. *9*, 255–266.

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. *38*, e164.

Wang, L., Ko, E.R., Gilchrist, J.J., Pittman, K.J., Rautanen, A., Pirinen, M., Thompson, J.W., Dubois, L.G., Langley, R.G., Jaslow, S.L., et al. (2017). Human genetic and metabolite variation reveal methylthioadenosine is a prognostic biomarker and inflammatory regulator in sepsis. Sci. Adv. *3*, e1602096.

Xie, Z., Zhang, H., Tsai, W., Zhang, Y., Du, Y., Zhong, J., Szpirer, C., Zhu, M., Cao, X., Barton, M.C., et al. (2008). Zinc finger protein ZBTB20 is a key repressor of alpha-fetoprotein gene transcription in liver. Proc. Natl. Acad. Sci. USA *105*, 10859–10864.

Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011). GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. *88*, 76–82.

Yapar, N. (2014). Epidemiology and risk factors for invasive candidiasis. Ther. Clin. Risk Manag. *10*, 95–105.

Zaitlen, N., Kraft, P., Patterson, N., Pasaniuc, B., Bhatia, G., Pollack, S., and Price, A.L. (2013). Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. PLoS Genet. *9*, e1003520.

## STAR★METHODS

### KEY RESOURCES TABLE

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---|---|---|
| Antibodies | | |
| HCV NS5A antibody | Charles Rice | N/A |
| Bacterial and Virus Strains | | |
| *S. enterica* Typhi (strain Ty2) +pMMB67GFP | Dennis Ko; Alvarez et al., 2017 | DCK33 |
| *S. enterica* Typhimurium (strain 14028s) +pMMB67GFP | Dennis Ko; Ko et al., 2009 | DCK22 |
| *S. enterica* Typhimurium (strain 14028S *ΔsifA*) +pMMB67GFP | This paper | DCK32 |
| *Chlamydia trachomatis* LGV-L2, $Rif^R$ + pGFP::SW2 | Raphael Valdivia; Bastidas and Valdivia, 2016 | N/A |
| *Mucor circinelloides* f. *lusitanicus* R7B (*leuA*-) | Soo Chan Lee; Roncero, 1984 | N/A |
| *Candida albicans* SC5314 | Joseph Heitman; Odds et al., 2004 | N/A |
| *Toxoplasma gondii* strain RHgfpluc | John Boothroyd; Saeij et al., 2005 | N/A |
| Hepatitis C virus | Stacy Horner | genotype 2A JFH1 HCV |
| Chemicals, Peptides, and Recombinant Proteins | | |
| alpha-hemolysin (Staph. aureus alpha toxin) | Sigma | H9395-.5MG |
| 7-AAD | Enzo Life Sciences | BML-AP400-0001 |
| Critical Commercial Assays | | |
| Human FGF basic DuoSet | R&D Systems | DY233 |
| Human IL-10 DuoSet | R&D Systems | DY217B |
| Human CXCL10 DuoSet | R&D Systems | DY266 |
| Human MDC Duoset | R&D Systems | DY336 |
| Human IL-6 Duoset | R&D Systems | DY206 |
| Milliplex MAP Human Cytokine Custom 17-plex panel | Millipore EMD | HCYTOMAG |
| Deposited Data | | |
| Illumina HumanOmni 2.5M array | 1000 Genomes | ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/hd_genotype_chip/ |
| H2P2 database and web portal | This paper | http://h2p2.oit.duke.edu |
| RNA-seq Dataset: NT vs. ZBTB20 in Huh7 | This paper | GEO Accession# GSE116172 |
| Experimental Models: Cell Lines | | |
| 1000 Genomes LCLs | Coriell | GWD, ESN, IBS, KHV |
| Oligonucleotides | | |
| Accell siRNA for non-targeting #1, *ZBTB20*, *NFKBIA* | Dharmacon | D-001910-01-50; E-020529-00-0010; E-004765-00-0010 |
| siGenome siRNA for *ZBTB20* | Dharmacon | M-020529-01-0005 |
| Taqman human gene expression assays for *ZBTB20*, *NFKBIA, 18S RNA* | ThermoFisher | Hs00210321_m1; Hs00355671_g1; Hs03928990_g1 |
| Software and Algorithms | | |
| PLINK v1.9 | Chang et al., 2015 | https://www.cog-genomics.org/plink/1.9/ |
| LocusZoom | Pruim et al., 2010 | http://locuszoom.sph.umich.edu/ |
| IMPUTE v2.3.2 | Howie et al., 2009 | http://mathgen.stats.ox.ac.uk/impute/impute_v2.html |

(*Continued on next page*)

*Continued*

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---|---|---|
| SHAPEIT v2.r790 | Delaneau et al., 2013 | https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html |
| SAMtools v1.5 | Li et al., 2009 | http://samtools.sourceforge.net/ |
| GCTA v1.26 | Yang et al., 2011 | http://cnsgenomics.com/software/gcta |
| GARFIELD | Iotchkova et al., 2016 | http://www.ebi.ac.uk/birney-srv/GARFIELD |
| CIRCOS v0.69 | Krzywinski et al., 2009 | http://circos.ca/software/download/circos/ |
| ANNOVAR v2016FEB01 | Wang et al., 2010 | http://annovar.openbioinformatics.org/ |
| GGV Browser | Marcus and Novembre, 2017 | http://popgen.uchicago.edu/ggv/ |
| TrimGalore v0.4.5 | Krueger, 2017 | https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/ |
| Cutadapt v1.9.1 | Martin, 2011 | https://cutadapt.readthedocs.io/ |
| R v3.3.2 | Team, 2016 | https://cran.r-project.org/ |
| ggplot2 | The R Foundation | https://cran.r-project.org/web/packages/ggplot2/index.html |
| data.table | The R Foundation | https://cran.r-project.org/web/packages/data.table/index.html |
| dplyr | The R Foundation | https://cran.r-project.org/web/packages/dplyr/index.html |
| biomaRt | Bioconductor | https://bioconductor.org/packages/release/bioc/html/biomaRt.html |
| stringr | The R Foundation | https://cran.r-project.org/web/packages/stringr/index.html |
| GenomicRanges | Bioconductor | https://bioconductor.org/packages/release/bioc/html/GenomicRanges.html |
| coloc v2.3.1 | The R Foundation | https://github.com/chr1swallace/coloc |
| STAR v2.6 | Dobin et al., 2013 | https://github.com/alexdobin/STAR |
| HTSeq v0.10.0 | Anders et al., 2015 | https://htseq.readthedocs.io/ |
| DESeq2 v1.20.0 | Love et al., 2014 | https://bioconductor.org/packages/DESeq2/ |
| PheWAS | The R Foundation | https://phewascatalog.org/ |

## CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for reagents should be directed to and will be fulfilled by the Lead Contact, Dennis Ko (dennis.ko@duke.edu).

## EXPERIMENTAL MODEL AND SUBJECT DETAILS

### Cells

1000 Genomes LCLs (528; all trios) from 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) populations were purchased from the Coriell Institute. LCLs were maintained at 37°C in a 5% $CO_2$ atmosphere and were grown in RPMI 1640 media (Invitrogen) supplemented with 10% fetal bovine serum (FBS), 2 mM glutamine, 100 U/mL penicillin-G, and 100 mg/mL streptomycin.

Human hepatoma Huh7 cells were grown in DMEM (Mediatech) supplemented with 10% fetal bovine serum (HyClone), 2.5 mM HEPES, and 1× non-essential amino acids (complete, cDMEM; Thermo Fisher Scientific). The identity of the Huh7 cell line was verified using the Promega GenePrint STR kit (DNA Analysis Facility, Duke University), and cells were verified as mycoplasma free by the LookOut Mycoplasma PCR detection kit (Sigma). Infectious stocks of a cell culture-adapted strain of genotype 2A JFH1 HCV were generated and titrated by focus-forming assay (FFA), as described (Aligeti et al., 2015).

## METHOD DETAILS

### LCL Screening

LCLs were received from Coriell and cultured for 8 days prior to assays. LCLs were counted with a Guava Easycyte Plus flow cytometer (Millipore). LCLs were washed once with RPMI 1% FBS and then plated out in RPMI 10% FBS at 200,000 cells/200 μL for

Salmonellae, 100,000 cells/100 μL for fungi, and 40,000 cells/100 μL for *S. aureus* alpha toxin, *C. trachomatis*, and *T. gondii*. Cells were passaged at 150,000/mL in 20 mL total volume for three days.

### *C. trachomatis* Infection

*C. trachomatis* LGV-L2 $Rif^R$ pGFP::SW2 was grown and purified as previously described (Saka et al., 2011). *C. trachomatis* was added at MOI 5 in 100 μL assay media, mixed by multichannel pipetting, and centrifuged onto cells at 3000 RPM for 30 min at 4°C. At 27, 46, and 70 hr, cells were mixed and 25 μL was taken for flow cytometry measurement (4000 cells). 25 μL of supernatant at 70 hr was measured by Luminex assay for 17 cytokines.

### *Salmonella* Infection

Salmonellae were tagged with an inducible GFP plasmid (pMMB67GFP from Pujol and Bliska, 2003). *sifA* deletion mutant was constructed with lambda red (Datsenko and Wanner, 2000) and verified by PCR. Assaying LCLs for Salmonellae infection was conducted as previously described (Ko et al., 2009). Overnight bacterial cultures were subcultured with a 1:33 dilution and grown for 2 hr 40 min at 37°C. Invasion was conducted for 1 hr at a multiplicity of infection (MOI) of 10 for *S.* Typhi and MOI 30 for *S.* Typhimurium, followed by addition of gentamicin (50 μg/mL) for 1 hr, and then culture was split into two separate cultures of 60 μL of cells with 140 μL of media to dilute gentamicin (15 μg/mL) and allow for collection at two timepoints. IPTG (1.4 mM) was added to turn on GFP expression for 75 min prior to 3.5 hr and 24 hr timepoints. For the 3.5 hr timepoint, 150 μL of cells were stained with 7-AAD (7-aminoactinomycin D; Enzo Life Sciences) and green and red fluorescence of 7000 cells was measured on a Guava Easycyte Plus flow cytometer (Millipore). For the 24 hr timepoint, cells were spun down and 2 aliquots of 55 μL of supernatant was removed and stored at –80°C for subsequent IL10 (25 μL), CXCL10 (25 μL), and MDC (4 μL) ELISAs (R&D Systems). 55 μL of cells were stained with 7-AAD and measured by flow cytometry.

### Fungal Infection

The *Mucor circinelloides* f. *lusitanicus* R7B (*leuA*-) (Roncero, 1984) strain and *Candida albicans* SC5314 (Odds et al., 2004) strain were used to induce FGF-2 from the LCLs. Leucine autotropism ($leuA^-$) was found not to impact virulence (Li et al., 2011). To prepare *Mucor* spores, potato dextrose agar (PDA, 4 g potato starch, 20 g dextrose, and 15 g agar per liter) was inoculated and incubated for 4 days at 26°C in the light. To collect spores, sterile deionized distilled water was added onto the plates and spores were released by gently scraping the colonies with a cell spreader. Spores were counted by using a hemocytometer. To prepare *C. albicans* yeast, yeast dextrose broth (10 g yeast extract, 20 g peptone, 20 g glucose per liter) was inoculated and incubated at 30°C by shaking at 250 rpm overnight. The yeast cells were quantified by using a hemocytometer. To co-culture with LCLs, all fungal cells were washed with sterile PBS twice. Fungi were added at MOI 1 in 10 μL and incubated for 24 hr. Culture supernatant was collected and stored at –80°C for later FGF-2 ELISA analysis (R&D Systems).

### *S. aureus* Toxin Treatment

LCLs were treated with alpha-hemolysin (Sigma) at 1 μg/mL for 23 hr. Cells were mixed and cell death quantified by 7-AAD staining (concentration) and flow cytometry.

### *T. gondii* Infection

*T. gondii* strain RHgfpluc was grown on confluent human foreskin fibroblast cells. The infected cells were then scraped and transferred to a 50 mL polystyrene tube and centrifuged at 500 x g for 10 min at 4°C. Pellet was resuspended in 3 mL of PBS and the suspension was aspirated three times using a 20 g needle attached to a 10 mL syringe. 30 mL of PBS was added and centrifuged at 500 x g for 10 min at 4°C. Supernatant was removed and pellet resuspended in 5 mL of PBS. Concentration of a 1:200 dilution was determined by flow cytometry and added at MOI 2 to cells. At 5, 30, and 48 hr infection, cells were mixed, and 25 μL taken for measuring 4000 cells by flow cytometry.

Note that *T. gondii* parasites were prepared separately for each infection assay from human fibroblasts that were passaged continuously for H2P2 screening. Therefore, there was more inter-experiment variation in the pathogen compared to, for example, *C. trachomatis* (which was prepared as a single batch for the entire screen and frozen into single-use aliquots). Additionally, *T. gondii* was incorporated into H2P2 after the screen had already commenced, and only 335 LCLs were assayed with this pathogen. For these reasons, measured inter-individual variation for the *Toxoplasma* traits was less reliable, and there was less power to detect genetic associations.

### LCL RNAi Experiments

LCLs (2 x $10^5$ cells) were treated for three days in 500 μL of Accell media (Dharmacon) with either non-targeting Accell siRNA #1 or an Accell SmartPool directed against human *ZBTB20* or *NFKBIA* (1 μM total siRNA; Dharmacon). Prior to infection, cells were plated at 1 x $10^5$ in 100 μl RPMI complete media (without antibiotics) in 96-well plates. Infections were conducted as described above.

### HCV Infection Experiments

Huh7 cells were seeded in 12-well plates at a density of 1 x $10^5$ cells per well in cDMEM and transfected the next day using 9 μL RNAiMAX (Thermo) and 3 μL of the indicated 10 μM siRNA (siGenome Smartpool [Dharmacon]), in Optimem (Thermo). Four hours

post-transfection, the transfection mixture was removed and 1 mL fresh cDMEM was added. HCV infections were performed at an MOI of 0.3 for 24, 48, or 72 hr. For each condition, duplicate wells were either infected or mock-infected, RNA was harvested from one well and the other utilized for visualization of infected cells. Supernatant was collected from both wells for virus titration.

HCV focus forming assay. Serial dilutions of supernatants collected from non-targeting or ZBTB20-targeting siRNA treated cells collected 72 hpi were used to infect naive Huh7.5 cells in triplicate wells of a 48-well plate. At 48 hpi, cells were fixed, permeabilized, and immunostained with HCV NS5A antibody (1:500; gift of Charles Rice, Rockefeller University). Following binding of horseradish peroxidase (HRP)-conjugated secondary antibody (1:500; Jackson ImmunoResearch), infected foci were visualized with the VIP Peroxidase Substrate Kit (Vector Laboratories) and counted at 40× magnification.

Visualization of HCV infected cells. 48 hr post-transfection, Huh7 cells treated with non-targeting or ZBTB20-targeting siRNA were infected with HCV (MOI 0.3) or mock-infected. 24 or 48 hr post-infection, cells were fixed in 4% paraformaldehyde in PBS, permeabilized with 0.2% Triton X-100 in PBS, and blocked with 3% BSA in PBS, then immunostained with HCV NS5A antibody (1:1000), washed 3x in PBS-Tween, then visualized with Alexa Fluor 488 Donkey anti-Mouse secondary antibody (1:1000, Thermo). Cell nuclei were stained with DAPI in the first of three PBS-Tween washes following the addition of secondary antibody. Two-color images were collected with the Cellomics ArrayScan VTI HCS (Thermo), at 20x magnification in the Duke Functional Genomics Shared Resource. 10 fields of per-well, per-condition were acquired and the percentage of identified nuclei with detectable NS5A staining was quantified using VHSview software (Thermo).

### RNA-seq

RNA was obtained from three independent experiments from Huh7 cells transfected with siGENOME siRNA for 48 hr (described above) and then mock-infected and incubated for an additional 24 or 48 hr. Stranded mRNA-seq libraries were generated and run on an Illumina NovaSeq 6000 instrument with 50 bp paired-end reads by the Duke Sequencing and Genomic Technology Shared Resource.

## QUANTIFICATION AND STATISTICAL ANALYSIS

### Phenotype Repeatability

Repeatability of each cellular trait was calculated from 3 independent experiments. The inter- and within-individual component of variance was calculated by fitting with one-way ANOVA. The estimated within-individual component of variance gave the repeatability coefficient.

### Testing Effect of EBV Copy Number on H2P2 Cellular Phenotypes

EBV relative copy numbers were retrieved for 1753 LCLs (Mandage et al., 2017), of which 284 cell lines overlapped with H2P2 samples (73 ESN; 105 GWD; 106 IBS). Prior to analysis, H2P2 phenotypes were averaged from three experiment replicates and then transformed to Z-scores within each experimental batch. Correlation between H2P2 cellular traits and EBV loads was tested using linear regression with population as a covariate.

### Genotype and Imputation

Genotypes for 1000 Genome LCLs (1000 Genomes Project Consortium et al., 2010) were from Illumina HumanOmni 2.5M array (905,788 SNPs; see details in Key Resources Table). Genome-wide imputation of autosomal genotypes with 1000 genome Phase 3 haplotype as reference panel were performed through two steps, a pre-phasing step using SHAPEIT2 (Delaneau et al., 2013) and an imputation step using IMPUTE2 (Howie et al., 2012). Imputed genotype was further filtered by imputation accuracy score (IMPUTE's INFO) $< 0.9$ and minor allele frequency $< 0.01$. A total of 339 samples overlap with 1000 genome Phase 3 individuals. We merged direct sequenced genotypes from 1000 genome Phase 3 project into our imputed genotypes. We obtained 15,581,278 SNPs (8,817,925 SNPs have minor allele frequency $\geq 0.05$). The human genome reference assembly (GRCh37/hg19) was used for all analysis.

### Phenotype- and SNP-Based Heritability Analysis

Two different methods were applied to estimate heritability. The parent-offspring (PO) regression method estimated additive heritability exclusively using phenotypic values. Linear regression of child against average of parents was performed, and the slope was used as a heritability estimator. Batch was incorporated as a covariate. Genotype-based heritability was estimated using the GCTA GREML method (Yang et al., 2011). Autosomal SNPs with minor allele frequency filtering of 0.05 were used to create a genetic relationship matrix (GRM). Zaitlen and colleagues developed a method, "big K/small K," which estimates heritability by jointly using closely related and unrelated individuals (Zaitlen et al., 2013). Following Zaitlen's method, variance explained by genome-wide SNPs ($\sigma_g^2$) was then estimated for each cellular trait. The Zaitlen modification provides joint estimates of 1) $h^2$ based on pedigree relatedness and 2) $h^2$ based on inferred relatedness from genome-wide SNPs.

While the standard error for SNP-based $h^2$ estimates were quite large, we nonetheless observed very strong correlation between these estimates and the parent-offspring $h^2$ estimates (see Figure 1D). We estimated $h^2$ based on the analysis of LCLs from all populations in H2P2 to increase the precision of our estimates by including more individuals. Although $h^2$ is a population-specific

parameter, $h^2$ are often quite similar across different populations and even species (Visscher et al., 2008). Nonetheless, we include estimates and standard errors for the combined analysis as well as individual populations in Table S3.

### Genome-wide Association Analysis

Genome-wide association analysis was conducted with PLINK v1.9 (Chang et al., 2015). Analysis was carried out using the QFAM-parents approach with adaptive permutation and a maximum of $10^9$ permutations. The QFAM procedures implemented in PLINK use linear regression to test for association while employing permutation of within- and between-family components separately to control for family structure (Purcell et al., 2005).

### Permutation Analysis

Simulation were carried out to estimate the genome-wide significance thresholds for H2P2 traits. For each cellular traits, we permutated the phenotype and ran family based association analysis using "qfam-parents" in PLINK 1.9. Following Kanai et al., 2016, we calibrated the empirical genome-wide significance threshold using the minimum p value from each simulation. We calculated the 90th percentile of the empirical distribution of $-\log_{10}P_{min}$ using the Harrel-Davis estimator (Harrel and Davis, 1982) at $\alpha = 0.1$, and the 95% confidence intervals of the quantile were obtained using 10000 bootstraps.

### Enrichment Analysis

Enrichment analyses were carried out using GARFIELD (Iotchkova et al., 2016). GARFIELD has predefined a total of 1005 features from ENCODE and the NIH Roadmap project, and applies generalized linear regression models while accounting for the effects of linkage disequilibrium (LD), minor allele frequency, and local gene density. The GWAS summary statistics were used to quantify fold-enrichment against predefined annotation features at different GWAS p value thresholds.

### PheWAS Analysis

Testing for association of H2P2 genome-wide hits with clinical phenotypes was performed with the eMERGE biobank dataset of 83,717 individuals from 12 contributing medical centers (McCarty et al., 2011) with ICD-9 derived PheWAS codes (Denny et al., 2013). A merged set of unified variant genotypes across 78 batches of samples with different genotype platforms (e.g. various Illumina and Affymetrix arrays) was produced by imputation using the Michigan Imputation Server (MIS) with the HRC1.1 haplotype reference set. Sixteen variants were selected for PheWAS based on association in H2P2 and their inclusion in the imputed eMERGE biobank dataset. The PheWAS codes were defined by query of the ICD-9 electronic medical record datasets of the contributing medical centers. Two types of PheWAS code phenotypes were used in the association to ascertain more chronic versus singleton diagnoses: the minimum code count of one (mcc1) ICD-9 code to define an individual as a PheWAS code case, and minimum code count of two (mcc2) instances to define an individual as a chronically represented case. In the mcc2 cases, individuals were excluded from analysis if they only had one instance of the ICD-9 code. If there were less than 500 cases we did not include the ICD-9 derived PheWAS code in analysis because it would likely be underpowered and impact the multiple testing correction. We also did not include medical centers that had low ascertainment of the ICD-9 by excluding medical centers which had less than 10 cases. This resulted in 1,338 for mcc1 and 788 for mcc2 phenotypes being included in the analysis. We used the PLINK1.9 identity by descent genome file to find the set of unrelated individuals to bring forward for analysis. PheWAS association was implemented in the R *glm()* logistic regression of the case-control data and plotting was carried out using the PheWAS R package (Carroll et al., 2014). The covariates of gender and the PLINK1.9 computed 1 and 2 principal components from the pruned (>5% minor allele frequency, genotype, and sample missingness > 0.1 and LD r-square<0.7.) genome-wide imputation variant genotypes were included in the regressions. The p.adjust() R function with Bonferroni methods was used to adjust p values of the tested PheWAS codes within a particular SNVs sets of tests for multiple comparisons. Bonferroni of less than 0.05 was used as a significance threshold.

### Colocalization Analysis

Colocalization analysis was performed using R "coloc" v2.3.1 package (available at http://cran.r-project.org/web/packages/coloc). This software applies a Bayesian framework to estimate the posterior probability of genomic variants affecting both cellular trait and disease based on pre-computed GWAS p values, odds ratios, and minor allele frequencies. We ran colocalization on a 400-kb region centered on the focus SNP rs2869462 using default COLOC parameters ($P1=P2=1x10^{-4}$; $P12=1x10^{-5}$). Summary statistics of IBD GWAS (Liu et al., 2015) were obtained from www.ibdgenetics.org.

### Gene Expression Analysis on 1000 Genome RNA-seq Project

Gene expression data of 465 individuals (Lappalainen et al., 2013) were obtained from the EBI website (https://www.ebi.ac.uk/Tools/geuvadis-das/). The rs2869462 genotype data were downloaded from the 1000 genome project (1000 Genomes Project Consortium et al., 2010). Effects of rs2869462 on *CXCL10* gene expression were tested by linear regression on both combined populations and individual population.

### RNA-seq Analysis

RNA-seq data were processed using the TrimGalore toolkit v0.4.5 (Krueger, 2017) which employs Cutadapt v1.9.1 (Martin, 2011) to trim low quality bases and Illumina sequencing adapters from the 3′ end of the reads. Only reads that were 20 nt or longer after trimming were kept for further analysis. Reads were mapped to the GRCh37v75 version of the human genome and transcriptome (Kersey et al., 2012) using the STAR v2.6 (Dobin et al., 2013). Reads were kept for subsequent analysis if they mapped to a single genomic location. Gene counts were compiled using the HTSeq v0.10.0 (Anders et al., 2015). Only genes that had at least 10 reads in any given library were used in subsequent analysis. Normalization and differential expression was carried out using the DESeq2 v1.20.0 (Love et al., 2014) Bioconductor (Gentleman et al., 2004) package with the R statistical programming environment (Team, 2016). Using batch and time point as cofactors in the model, we identified differentially expressed genes between the *ZBTB20* and the NT siRNA conditions. The false discovery rate was calculated to control for multiple hypothesis testing. Gene set enrichment analysis (Mootha et al., 2003) was performed to identify gene ontology terms and pathways associated with altered gene expression for each of the comparisons performed.

### Descriptive Statistics and Visualization

Descriptive statistics were performed with GraphPad Prism 6 (GraphPad Software, US) and with R (Team, 2016). QQ plots were plotted using quantile-quantile function in R. Regional Manhattan plot were made using LocusZoom (Pruim et al., 2010). Circos v0.69 was used to visualize the shared SNPs among different groups. The size of each study or number of replicates, along with the statistical tests performed can be found in Figure Legends. All numerical data are presented as the mean ± SEM (standard error of mean).

## DATA AND SOFTWARE AVAILABILITY

All H2P2 data are available for browsing and download at http://h2p2.oit.duke.edu.

The H2P2 application server is running Red Hat Enterprise Linux Server v7.4, Apache v2.4.6, Shiny Server v1.5.3.838, R v3.4.1, and Microsoft ODBC Driver 13 for SQL Server. Implemented R packages include shiny v1.0.3, RODBC v1.3-15, ggplot2 v2.2.1, d3heatmap v0.6.11, and DT v0.2. The H2P2 database server is running MS Windows Server 2016 and MS SQL Server 2016. Large volume tables and indexes (2.5 billion GWAS observations and 8.5 billion genotype observations) were partitioned for improved query performance. Parallelized query implementation was also used to improve performance.

RNA-seq data are available in GEO: GSE116172.