COMPUTATIONAL MEDICINE

Disease Risk Factors Identified Through Shared Genetic Architecture and Electronic Medical Records

Li Li,[1,2]* David J. Ruau,[1,2]* Chirag J. Patel,[1,2,3] Susan C. Weber,[4] Rong Chen,[1,5] Nicholas P. Tatonetti,[6] Joel T. Dudley,[7] Atul J. Butte[1,2†]

Genome-wide association studies have identified genetic variants for thousands of diseases and traits. We evaluated the relationships between specific risk factors (for example, blood cholesterol level) and diseases on the basis of their shared genetic architecture in a comprehensive human disease–single-nucleotide polymorphism association database (VARIMED), analyzing the findings from 8962 published association studies. Similarity between traits and diseases was statistically evaluated on the basis of their association with shared gene variants. We identified 120 disease-trait pairs that were statistically similar, and of these, we tested and validated five previously unknown disease-trait associations by searching electronic medical records (EMRs) from three independent medical centers for evidence of the trait appearing in patients within 1 year of first diagnosis of the disease. We validated that the mean corpuscular volume is elevated before diagnosis of acute lymphoblastic leukemia; both have associated variants in the gene *IKZF1*. Platelet count is decreased before diagnosis of alcohol dependence; both are associated with variants in the gene *C12orf51*. Alkaline phosphatase level is elevated in patients with venous thromboembolism; both share variants in *ABO*. Similarly, we found that prostate-specific antigen and serum magnesium levels were altered before the diagnosis of lung cancer and gastric cancer, respectively. Disease-trait associations identify traits that could serve as future prognostics, if validated through EMR and subsequent prospective trials.

INTRODUCTION

Genome-wide association studies (GWAS) and candidate gene approaches have identified genetic variants for thousands of traits (*1*–*3*). Studied traits included clinical measurements (for example, cholesterol levels), social behavior (for example, smoking), patient characteristics (for example, weight), and disease susceptibility. At the same time, the number of GWAS performed to study diseases has rapidly increased since 2007, and their findings provide opportunities to investigate the potential impact of common genetic variants on complex diseases (*4*, *5*). It has already been noted that seemingly different diseases and conditions that share associated single-nucleotide polymorphisms (SNPs) may have common biological mechanisms (*6*, *7*).

With so many successful GWAS already completed on nondisease traits (referred hereafter as traits), we hypothesized that diseases and traits could be similarly related to each other through shared genetic variation. Preliminary work by us (*8*) and others (*5*) suggests that traits could indeed share variants with diseases. There could be high value in such a disease-trait association for medicine if the trait is easily or cheaply measured, or is already commonly measured in health care setting, and if the trait can be identified before the disease.

We hypothesized that traits could serve as potential new prognostic markers or risk factors for disease susceptibility, if those traits significantly shared genetic associations with diseases. We theorized that if variant-associated genes found in a GWAS of traits significantly matched gene variants found associated with a disease, those traits might be predictive for diseases, especially if that trait was one already measured in a clinical care settings, already captured in an electronic medical record (EMR).

RESULTS

Genes associated with diseases and traits

This study reports a method for predicting new markers for disease from genetic associations found for thousands of diseases and traits from GWAS. We started with findings from VARIMED (VARiants Informing MEDicine) (*9*–*13*), a manually curated database of disease-SNP associations, containing more than 100 features of association studies from 8962 human genetics papers covering 2376 diseases and traits. VARIMED has been used to interpret the genome sequences of patients and other individuals (*9*, *14*). We identified a list of disease-trait pairs based on shared genetic architecture.

Figure 1 shows our overall experimental design. From VARIMED, we identified significant associations between 801 unique genes and 69 diseases (median = 10 per disease), and between 796 unique genes and 85 traits (median = 10 per trait). In each case, there were at least three significant genes per disease or trait, and the P value was $<1 \times 10^{-8}$ at the genome-wide significance level from individual GWAS (table S1, A and B). The three diseases with the most associated genes were rheumatoid arthritis (122 genes), membranous nephropathy (88 genes), and myocardial infarction (73 genes). The top 3 traits with the most associated genes were height (120 genes), blood cholesterol level (50 genes),

[1]Division of Systems Medicine, Department of Pediatrics, Stanford University School of Medicine, 1265 Welch Road, Stanford, CA 94305, USA. [2]Lucile Packard Children's Hospital, Palo Alto, CA 94305, USA. [3]Stanford Prevention Research Center, Department of Medicine, Stanford University School of Medicine, Stanford, CA 94305, USA. [4]Stanford Center for Clinical Informatics, Stanford University School of Medicine, Stanford, CA 94305, USA. [5]Personalis Inc., 1350 Willow Road, Suite 202, Menlo Park, CA 94025, USA. [6]Department of Biomedical Informatics, Columbia Initiative for Systems Biology, and Department of Medicine, Columbia University, 622 West 168th Street, VC5, New York, NY 10027, USA. [7]Department of Genetics and Genomics Sciences, Institute for Genomics and Multiscale Biology, Mount Sinai School of Medicine, One Gustave L. Levy Place, Box 1498, New York, NY 10029, USA.
*These authors contributed equally to this work.
†Corresponding author. E-mail: abutte@stanford.edu

and blood protein C levels (49 genes). We plotted the distributions of the gene counts as a density map by kernel density estimation (fig. S1A). We found no significant difference between the distribution of gene-disease associations and gene-trait associations via the Kolmogorov-Smirnov test ($P = 0.16$). We concluded that the number of genes associated with either traits or diseases was unbiased and comparable.

Disease and trait associations identified by shared variant-associated gene

We searched for pairs of diseases and traits that shared variants in common genes. To evaluate the significance of the association, we assigned an information content measure to each gene on the basis of how frequently a gene was associated across diseases and traits using term frequency–inverse document frequency (TF-IDF), and then controlled for multiple hypothesis testing by random shuffling 1000 times. We identified 120 disease-trait pairs significant at $q \leq 0.01$ based on the pairwise cosine distance calculation (see Materials and Methods). Among the 120 pairs, 96 (80%) pairs linked a disease and trait that were originally published in different GWAS or candidate gene studies (table S2). Forty-five unique diseases and 50 unique traits were identified out of the 120 significant disease-trait pairs. To evaluate the accuracy of our predictions, we manually reviewed the biomedical literature to see if we could corroborate these 120 predicted associations. Ninety-four pairs were known, published associations between diseases and traits. Twenty-six pairs were previously undescribed, without previous evidence in the literature (table S2). We plotted the distribution of the PubMed counts for shared genes for disease-trait pairs. We found no significant difference between the distribution of the number of published human genetic papers in genes shared in known and newly discovered disease-trait pairs via the Kolmogorov-Smirnov test ($P = 0.51$) (fig. S1B).

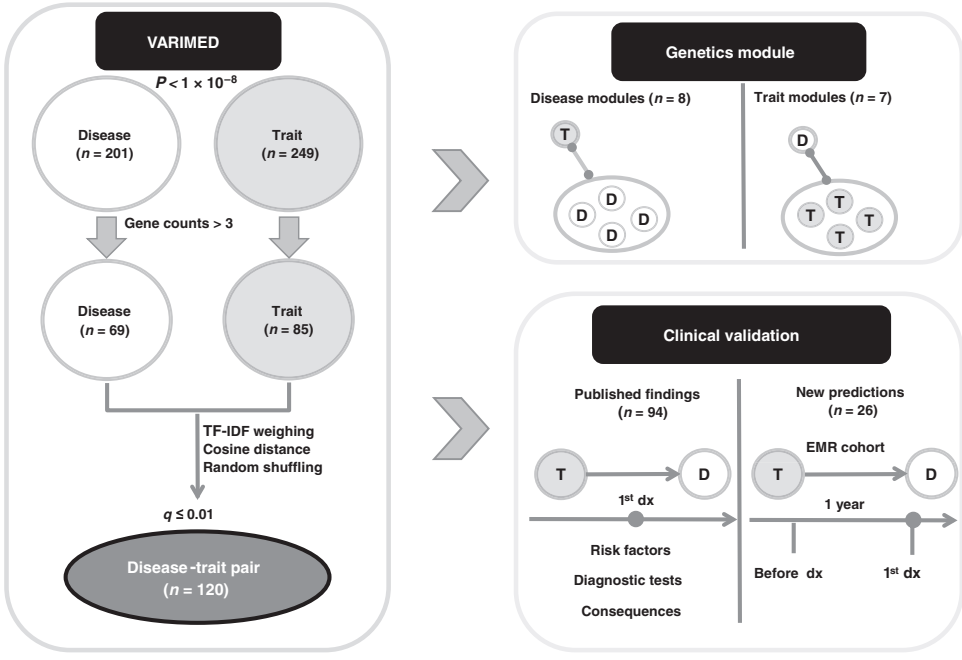

Fig. 1. Diagram for identifying significant disease-trait genetic associations.

Genetic commonality between diseases and traits

We generated a comprehensive network for visualizing all 120 disease-trait pairs (Fig. 2 and table S2). Diseases (blue circles) and traits (orange triangles) were connected to each other by edges when there was a significant association at $q \leq 0.01$. If multiple diseases or traits were connected to the similar traits or diseases, these were grouped into super sets (termed "modules"), simplifying the visualization of this complex network. Eight major disease modules (blue circles) were revealed in the network, which represent groups of diseases sharing a significant genetic association to a particular trait or a group of traits.

Four modules presented known classifications based on the physiological system affected by the disorder. For instance, solid organ cancer (Fig. 2, module D1) was connected with prostate-specific antigen (PSA) levels because this trait and these diseases were significantly associated through *TERT*. The skin cancer module (Fig. 2, module D2) was connected with pigmentary characteristics, as a trait, through *SLC45A2* or *MC1R*. The autoimmune disorder module (D6) was connected with antibody titer levels through association with major histocompatibility complex (MHC) class I/II or MHC class–related molecules. Finally, type 2 diabetes–related syndromes (Fig. 2, module D3) were connected with proinsulin levels. Most of these connections were through *ARAP1*, *MADD*, or *TCF7L2* (table S2).

The remaining four disease modules (Fig. 2, D4, D5, D7, and D8) exhibited multiple-to-multiple relationships underlying unexpected shared genetic commonality. One module (Fig. 2, module D4) connected esophageal cancer and alcohol dependence with cholesterol levels through *ALDH2*, *BRAP*, and *C12orf51*, whereas another (Fig. 2, module D5) connected Kawasaki disease and chronic obstructive pulmonary disease (COPD) with smoking through *RAB4B*.

We identified seven trait modules (Fig. 2, T1 to T7, orange circles). Three modules had known associations: pigmentary characteristics (Fig. 2, T1) with skin cancer (D2) through *MC1R* or *SLC45A2*, and a subset (freckles and eye colors) with chronic lymphocytic leukemia through *IRF4*. Coagulation factor activity tests (Fig. 2, T4) were connected with venous thromboembolism (VTE). Three were related through *ABO* (table S2). Lipid panel (Fig. 2, T5) was connected through *APOC1*, *APOE*, *PVRL2*, and *TOMM40* to Alzheimer's disease, through *CELSR2*, *LDLR*, *PSRC1*, and *ZNF259* to coronary artery disease (CAD), and through *ZNF259* to metabolic syndrome.

Detecting traits known to be associated with diseases

Ninety-four of the 120 significant disease-trait associations were known findings supported by published studies (table S2); these disease-trait associations could be classified into one of three types based on the temporal relationship between the trait and disease pathogenesis: (i) risk factors, for which traits manifest before disease onset and may cause the disease; (ii) diagnostic tests, for which traits manifest contemporaneously with disease onset; and (iii) consequences or complications, for

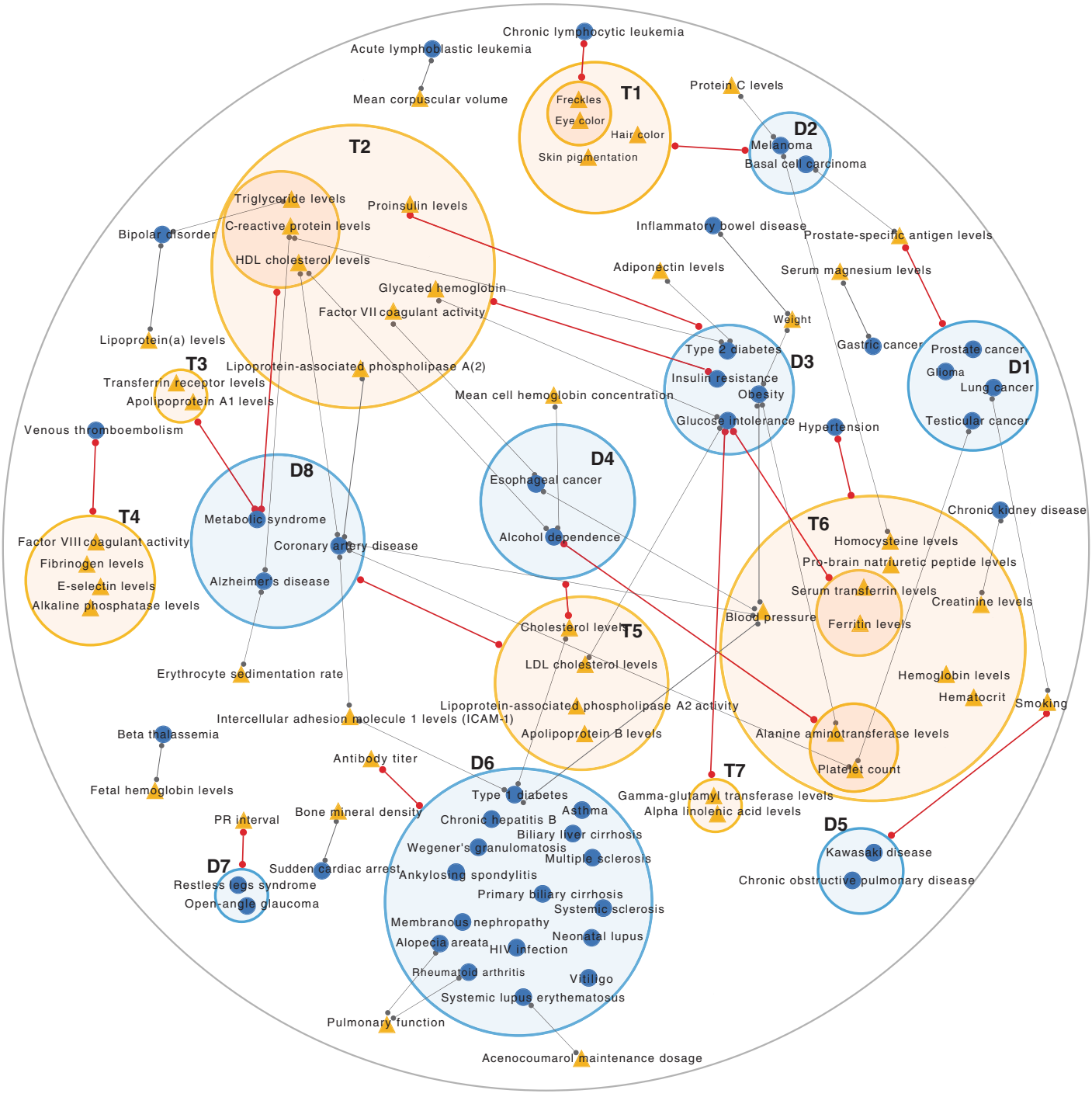

Fig. 2. Disease-trait network of 120 significant pairs. The network consists of the 120 significant disease-trait pairs with $q \leq 0.01$. Diseases (blue circles) and traits (orange circles) are connected by gray lines (single connection between trait and disease) or red lines (one to a group of diseases or traits). T1 to T7 indicate trait modules (light orange circles) connected to a disease or disease module by red lines. D1 to D8 indicate disease modules (light blue circles) connected to a trait or trait module by red lines. This network was visualized by Cytoscape 2.6.0 (*48*) and the CyOog (*49*) plug-in.

which traits manifest after the disease diagnosis (Fig. 3 and table S2). We manually categorized each known finding into one of these three categories on the basis of original clinical studies (table S2). Thirty-nine pairs were classified as risk factors, 27 pairs were described as diagnostic tests in current clinical practice, and 28 pairs were defined as consequences or complications.

One of the 39 known pairs from the risk factors category (Fig. 3) linked smoking and COPD ($q < 0.001$). Three genes containing variants

were shared between smoking and COPD: *AGPHD1*, *CHRNA3*, and *RAB4B* (Fig. 2 and table S2). The COPD patients in all six GWAS were former or current smokers (*15*–*20*). Smoking is the primary risk factor for COPD (*21*–*23*), and little is known about the nature of the inflammatory response leading to the pathogenesis of COPD (*21*). Therefore, of the six genetic variants previously discovered and published to be associated with COPD, these three might have been indirectly influenced by smoking (concept illustrated in Fig. 3) and might actually reflect variants related to smoking (that is, propensity to addiction, noncessation, and variable action of nicotine).

Existing diagnostic tests were also reidentified through our approach. In one GWAS, 21 genes were associated with antibody titer levels after inoculation with hepatitis B vaccine (*24*). However, this study did not include patients with autoimmune diseases. We found that antibody titer levels, as a trait, were significantly associated with 16 autoimmune diseases. Antinuclear antibody and autoantibody tests can serve as diagnostic tests in autoimmune disorders and diseases (table S2 and Fig. 2). Although the GWAS (*24*) did not explicitly enroll participants with these autoimmune diseases, our method inferred known relationships between clinical measurements, such as autoantibody tests, and autoimmune diseases on the basis of their shared genetic architecture (Fig. 3).

Last, among the 28 known pairs reflecting comorbidity or consequence (table S2), alcohol dependence syndrome (ADS) was associated with three traits: cholesterol levels through shared variants in *ALDH2*, *BRAP*, and *C12orf51*; alanine aminotransferase (ALT) levels through shared variants in *C12orf51*; and high-density lipoprotein cholesterol (HDL-C) levels through shared variants in *C12orf51* and *OAS3*. In this case, we speculate that the three genes found associated with cholesterol levels reported by Kato *et al.* (*25*) and two genes for ALT and HDL-C reported by Kim *et al.* (*26*) were discovered in cohorts containing individuals who might have been influenced by alcohol, although these authors did not control for any alcohol effect in their GWAS investigations on these genes (*25*–*27*). In addition, high HDL-C has been previously observed with triple frequency in individuals with ADS (*28*). Further, a high cholesterol content diet has been found in patients with ADS (*29*). ALT levels are associated with increased daily alcohol intake in individuals with ADS (*30*).

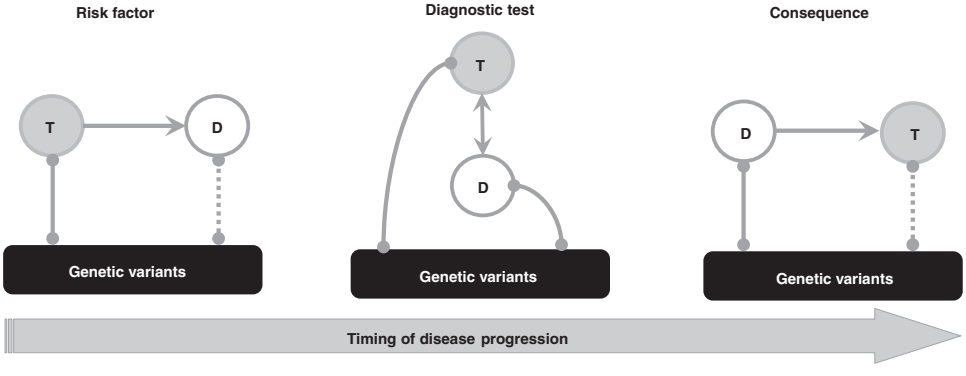

Fig. 3. Three ways traits and diseases can temporally interrelate. Traits (that is, risk factors) can manifest before disease, at the same time as disease diagnosis, or represent consequences occurring after diagnosis. Genetic variants were either directly observed in traits and diseases (solid edges) or indirectly observed or potentially influenced by a preceding trait or disease (dotted edges). Arrow direction indicates the timing of the interrelation.

Clinical validation of previously undescribed disease-trait pairs with EMR

To evaluate our new associations between traits and diseases, we obtained EMR data, because they represented a patient cohort independent from our curated GWAS studies. We obtained deidentified EMR data from three independent clinical centers: Stanford Hospital and Clinics (SHC) (*31*), Mount Sinai Medical Center (MSMC), and Columbia University Medical Center (CUMC). Among 26 new disease-trait pairs, we studied 5 that could be validated solely by electronic means based on clinical data available in the three centers. In addition, we tested a positive control disease-trait pair, and two nonrelated disease-trait pairs as negative controls.

Our first new pair was that mean corpuscular volume (MCV) and acute lymphoblastic leukemia (ALL) were both associated with *IKZF1* ($q = 0.001$; table S2). To validate this finding, we selected as cases individuals at SHC and MSMC who had an MCV measurement within 1 year before a recorded diagnosis of ALL, where that recorded diagnosis was the first such diagnosis for each individual within our EMR. There were 640 and 307 cases of ALL at SHC and MSMC, respectively [mean age, 49 ± 18 (range, 18 to 91) at SHC and 48 ± 19 (range, 18 to 102) at MSMC; 45% female at both centers]. We selected as controls those individuals at SHC and MSMC with at least one MCV measurement and no diagnosis of ALL, yielding 254,624 and 367,292 control patients at SHC and MSMC, respectively. Patients with an abnormal MCV were significantly more likely to get a first recorded diagnosis of ALL within 1 year compared to patients with normal MCV [odds ratio (OR), 3.31 (95% CI, 2.84 to 3.87), with $P = 3.79 \times 10^{-57}$ at SHC; OR, 2.4 (95% CI, 1.91 to 3), with $P = 9.16 \times 10^{-15}$ at MSMC; Table 1]. Besides the increase in cases, the MCV values themselves were significantly higher in cases compared to controls ($P = 1.32 \times 10^{-48}$ and 3.36×10^{-11} for SHC and MSMC, respectively; Fig. 4A).

Our second new finding was that serum magnesium (MGN) level was associated with gastric cancer (GCA) through *MUC1*, *THBS3*, and *TRIM46* ($q < 0.001$; table S2). We validated this finding by selecting the 305 and 499 individuals at CUMC and MSMC, respectively, who had an MGN measurement within 1 year before our first EMR recorded diagnosis of GCA, where that recorded diagnosis was the first such diagnosis for each individual within our EMR [mean age, 51 ± 19 (range, 18 to 90) at CUMC and 66 ± 15 (range, 18 to 99) at MSMC; 41 and 52% female in CUMC and MSMC, respectively]. We selected 204,575 and 119,585 patients as controls at CUMC and MSMC, respectively, who had at least one MGN measurement and no diagnosis of GCA. We found that patients with an abnormal MGN level were significantly more likely to develop GCA within 1 year compared to patients with normal MCV [OR, 1.59 (95% CI, 1.26 to 2.01), with $P = 1.04 \times 10^{-4}$ at CUMC; OR, 1.54 (95% CI, 1.29 to 1.84), with $P = 1.45 \times 10^{-6}$ at MSMC; Table 1]. In addition, the MGN measurement values were significantly higher in those diagnosed with GCA within 1 year before our first diagnosis compared to all other MGN measurements ($P = 4.81 \times 10^{-10}$ and 9.48×10^{-5} for CUMC and MSMC, respectively; Fig. 4B).

Our third validation related PSA level (PSA) to lung cancer (LCA) through *CLPTM1L* and *TERT* (q = 0.001; table S2). Cases were those 114 and 126 males at SHC and MSMC, respectively, who had a PSA measurement within 1 year before our first recorded diagnosis of LCA [mean age, 60 ± 12 (range, 21 to 101) at SHC and 69 ± 10 (range, 46 to 99) at MSMC]. Control individuals at SHC and MSMC had at least one PSA measurement and no diagnosis of LCA. Patients with an abnormally high PSA were significantly more likely to develop LCA within 1 year compared to patients with normal PSA [OR, 2.08 (95% CI, 1.36 to 3.18), with $P = 5 \times 10^{-4}$ at SHC; OR, 2.33 (95% CI, 1.58 to 3.44), with $P = 1.87 \times 10^{-5}$ at MSMC; Table 1]. Just as with the previous findings, the PSA values were significantly higher in those diagnosed with LCA within 1 year before our first diagnosis compared to all other PSA measurements (P = 0.002 and 0.028 for SHC and MSMC, respectively; Fig. 4C).

We similarly validated our fourth finding, alkaline phosphatase (ALP) level related to VTE through *ABO* and *TERT* (q = 0.008; table S2), finding that patients at CUMC and MSMC with an abnormal ALP were significantly more likely to develop VTE within 1 year compared to patients with normal ALP [OR, 1.91 (95% CI, 1.81 to 2.01), with $P = 1.67 \times 10^{-133}$ at MSMC; OR, 1.30 (95% CI, 1.16 to 1.45), with $P = 3.97 \times 10^{-6}$ at CUMC; Table 1]. Like the previous findings, the ALP values themselves were significantly higher in those diagnosed with VTE within 1 year before our first diagnosis compared to all other ALP measurements ($P = 4.48 \times 10^{-252}$ and 7.33×10^{-55} for CUMC and MSMC, respectively; Fig. 4D).

The fifth and final validation was to test the relation between PLT counts and ADS, linked through *C12orf51* (q = 0.007; table S2). Patients were selected at all three centers if they had a PLT measurement within 1 year of a recorded diagnosis of ADS, where that recorded diagnosis was the first such diagnosis for each individual within our EMR. These cases were compared to individuals with at least one PLT measurement and no diagnosis of ADS. Patients with abnormal PLT were significantly more likely to be newly assigned a diagnosis of ADS within 1 year compared to patients with normal PLT [OR, 2.12 (95% CI, 1.92 to 2.35), with $P = 1.24 \times 10^{-52}$ at SHC; OR, 1.84 (95% CI, 1.74 to 1.95), with $P = 1.42 \times 10^{-109}$ at MSMC; OR, 1.25 (95% CI, 1.09 to 1.45), with P = 0.0016 at CUMC; Table 1]. PLT values were consistently lower in ADS patients versus controls within 1 year before our first ADS diagnosis ($P = 4.37 \times 10^{-32}$ at SHC, $P = 2.47 \times 10^{-43}$ at MSMC, and $P = 2.67 \times 10^{-6}$ at CUMC; Fig. 4E).

To evaluate whether the significance of our five validated disease-trait pairs was confounded by age and gender, we adjusted age and gender variables in a logistic regression model for each of these five tests. We discovered that significant associations still persisted for MCV and ALL [adjusted OR, 3.5 (95% CI, 3.02 to 4.14), with $P < 2 \times 10^{-16}$ at SHC; adjusted OR, 2.49 (95% CI, 1.99 to 3.13), with $P = 2.73 \times 10^{-15}$ at MSMC], MGN and GCA [adjusted OR, 1.44 (95% CI, 1.21 to 1.72), with $P = 5.03 \times 10^{-5}$ at MSMC; adjusted OR, 1.63 (95% CI, 1.29 to 2.07), with $P = 4.02 \times 10^{-5}$ at CUMC], ALP and VTE [adjusted OR, 1.80 (95% CI, 1.71 to 1.90), with $P < 2 \times 10^{-16}$ at MSMC; adjusted OR, 1.3 (95% CI, 1.17 to 1.46), with $P = 2.84 \times 10^{-6}$ at CUMC], and PLT and ADS [adjusted OR, 1.95 (95% CI, 1.76 to 2.16), with $P < 2 \times 10^{-16}$ at SHC; adjusted OR, 1.78 (95% CI, 1.69 to 1.89), with $P < 2 \times 10^{-16}$ at MSMC; adjusted OR, 1.25 (95% CI, 1.08 to 1.44), with P = 0.0025 at CUMC]. Only PSA and LCA did not reach significance after age matching [adjusted OR, 1.48 (95% CI, 0.99 to 2.23), with P = 0.058 at MSMC; adjusted OR, 1.3 (95% CI, 0.83 to 2.03), with P = 0.25 at SHC], which may be due to insufficient sample size or a possible confounding in the underlying original association with PSA and prostate cancer (PCA).

Table 1. Summary of clinical validation through EMR from three independent medical centers. CI, confidence interval.

Finding	Disease-trait pair	Center	Total *N*	Cases	Controls	Gender	Laboratory values*	OR (95%CI)	$P^{\dagger}$	$P^{\ddagger}$
New	ALL-MCV	SHC	255,264	640	254,624	Both	High + low	3.31 (2.84–3.87)	3.79×10^{-57}	1.32×10^{-48}
	ALL-MCV	MSMC	367,599	307	367,292	Both	High + low	2.40 (1.91–3.00)	9.16×10^{-15}	3.36×10^{-11}
	GCA-MGN	MSMC	120,084	499	119,585	Both	High + low	1.54 (1.29–1.84)	1.45×10^{-6}	9.48×10^{-5}
	GCA-MGN	CUMC	204,880	305	204,575	Both	High + low	1.59 (1.26–2.01)	1.04×10^{-4}	4.81×10^{-10}
	LCA-PSA	SHC	19,203	114	19,089	Male	High	2.08 (1.36–3.18)	5.0×10^{-4}	2.0×10^{-3}
	LCA-PSA	MSMC	25,326	126	25,200	Male	High	2.33 (1.58–3.44)	1.87×10^{-5}	0.028
	VTE-ALP	MSMC	256,876	6,470	250,406	Both	High + low	1.91 (1.81–2.01)	1.67×10^{-133}	4.48×10^{-252}
	VTE-ALP	CUMC	406,845	1,554	405,291	Both	High + low	1.30 (1.16–1.45)	3.97×10^{-6}	7.33×10^{-55}
	ADS-PLT	SHC	249,091	1,635	247,456	Both	High + low	2.12 (1.92–2.35)	1.24×10^{-52}	4.37×10^{-32}
	ADS-PLT	MSMC	360,628	5,445	355,183	Both	High + low	1.84 (1.74–1.95)	1.42×10^{-109}	2.47×10^{-43}
	ADS-PLT	CUMC	610,169	965	609,204	Both	High + low	1.25 (1.09–1.45)	0.0016	2.67×10^{-6}
Positive	PCA-PSA	SHC	17,481	595	16,886	Male	High	10.96 (9.25–12.98)	4.43×10^{-248}	1.02×10^{-83}
	PCA-PSA	MSMC	24,219	1,231	22,988	Male	High	7.51 (6.67–8.46)	2.0×10^{-316}	7.01×10^{-69}
	PCA-PSA	CUMC	51,952	4,253	47,699	Male	High	9.45 (8.83–10.11)	1.02×10^{-300}	6.02×10^{-308}
Negative	ALL-PSA	SHC	19,268	17	19,251	Male	High	1.00 (0.12–8.13)	1	0.1
	GCA-PSA	SHC	19,300	31	19,269	Male	High	0.65 (0.20–2.13)	0.47	0.5

*High, high versus normal laboratory value; High + low, high and low versus normal laboratory values. †χ^2 test. ‡Wilcoxon rank-sum test.

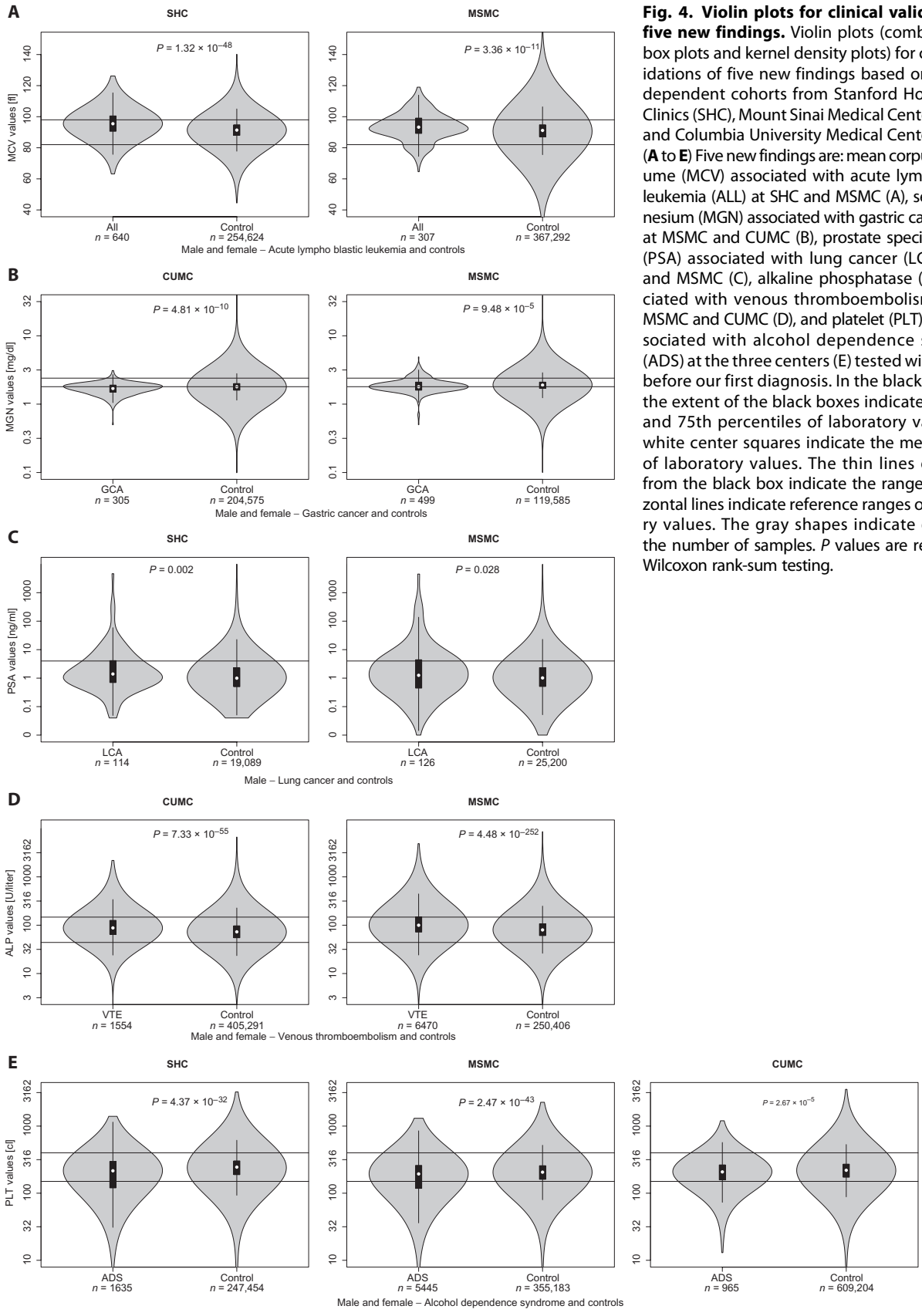

Fig. 4. Violin plots for clinical validations of five new findings. Violin plots (combination of box plots and kernel density plots) for clinical validations of five new findings based on three independent cohorts from Stanford Hospital and Clinics (SHC), Mount Sinai Medical Center (MSMC), and Columbia University Medical Center (CUMC). (**A** to **E**) Five new findings are: mean corpuscular volume (MCV) associated with acute lymphoblastic leukemia (ALL) at SHC and MSMC (A), serum magnesium (MGN) associated with gastric cancer (GCA) at MSMC and CUMC (B), prostate specific antigen (PSA) associated with lung cancer (LCA) at SHC and MSMC (C), alkaline phosphatase (ALP) associated with venous thromboembolism (VTE) at MSMC and CUMC (D), and platelet (PLT) counts associated with alcohol dependence syndrome (ADS) at the three centers (E) tested within 1 year before our first diagnosis. In the black box plots, the extent of the black boxes indicates the 25th and 75th percentiles of laboratory values, and white center squares indicate the median value of laboratory values. The thin lines extending from the black box indicate the range. The horizontal lines indicate reference ranges of laboratory values. The gray shapes indicate density of the number of samples. *P* values are reported by Wilcoxon rank-sum testing.

To evaluate our data resource in validating our findings, we selected one well-known association as a positive control (PSA levels and PCA) from all three centers. We obtained 595, 1231, and 4253 PCA male patient samples with PSA results [mean age, 70 ± 10 (range, 44 to 96) at SHC; mean age, 70 ± 11 (range, 34 to 98) at MSMC; and mean age, 58 ± 13 (range, 18 to 90) at CUMC] and 16,886, 22,988, and 47,699 control patients from SHC, MSMC, and CUMC, respectively. As expected, patients with abnormally high PSA were associated with PCA within 1 year before the first PCA diagnosis [OR, 10.96 (95% CI, 9.25 to 12.98), with $P = 4.43 \times 10^{-248}$ at SHC; OR, 7.51 (95% CI, 6.67 to 8.46), with $P = 2 \times 10^{-316}$ at MSMC; OR, 9.45 (95% CI, 8.83 to 10.11), with $P = 1.02 \times 10^{-300}$ at CUMC; Table 1]. Additionally, PSA values were higher in PCA patients compared to controls within 1 year before diagnosis ($P = 1.02 \times 10^{-83}$ at SHC, $P = 7.01 \times 10^{-69}$ at MSMC, and $P = 6.02 \times 10^{-308}$ at CUMC; fig. S2A).

We also tested two nonrelated associations as negative controls (PSA and ALL or GCA) using data from SHC. For the two negative control experiments, we performed the same tests, and we did not observe an association between laboratory values and disease (fig. S2, B and C, and Table 1).

DISCUSSION

We have developed a systematic approach for identifying genetic associations between traits and disease susceptibilities through shared genetic architecture. The goal was to identify traits as potential disease prognostic markers or risk factors. We identified 120 disease-trait pairs for traits associated with diseases; 80% of the pairs linked a disease and trait that had been published in distinct GWAS. Ninety-four had previous evidence in the literature, whereas 26 disease-trait pairs were newly described. We showed that these predicted relationships can be tested using medical center EMRs, when sufficient numbers of patients have data with assessments of both the trait and disease. We validated the relationships for five previously unreported findings—MCV to ALL, MGN to GCA, ALP to VTE, PSA to LCA, and PLT to ADS—using independent clinical EMR data from three independent academic medical centers.

The network representation for the significant 120 disease-trait pairs enabled us to highlight the complex genetic relationships between diseases and traits. The network revealed interconnections within and across eight disease modules and seven trait modules. Diseases and traits with shared genetic architecture can point to new markers and, potentially, therapeutic intervention and monitoring strategies. We noted that the traits and diseases associated with the most genes did not have more connections than diseases or traits with fewer gene associations, suggesting an accurate prioritizing strategy.

The strength of our strategy is that this approach can connect diseases and traits across the nosology or taxonomy of diseases. Another strength is that it provides a tractable framework that enables initial steps toward the development or redefinition of human disease nomenclatures informed by genetic variation. This gives the method potential utility in clinical care.

We found interesting relationships even with this known set of 94 relations beyond behavioral risk factors and diseases themselves. Examples include shared architecture for smoking and COPD, as well as ALT levels and alcohol dependence. For instance, because COPD commonly results from smoking, variants that have been discovered and associated with COPD could be influenced by smoking; the true genetic variants for COPD might only be unmasked if the smoking variable is controlled for in COPD GWAS. Similarly, the association of the four genetic variants with ALT, cholesterol, and HDL-C could be biased by the effect of alcohol. The GWAS to identify concrete genetic variants for these three clinical measurements should be performed in patients, ensuring that alcohol dependence is not a confounder. Thus, our study indicates that some findings from GWAS may have been influenced by or resulted from subject behaviors.

In addition, although we focus on disease-trait association in this study, a disease could be the potential confounder to another disease as well. For instance, ADS is a risk factor to HDL-C, which is a known risk factor to CAD (*32*), and *C12orf51* was shared among them; therefore, *C12orf51* variants associated with CAD could be confounded by ADS. Similarly, metabolite levels, such as MGN levels, are distorted in severe gastrointestinal disorders, and these disorders might actually be the causal factor for patients with subsequent diagnosis of another disease. We suggest that known and newly discovered risk factors should be considered in future GWAS design to properly identify variants more independent of behavioral or environmental influence (*33*, *34*). Lack of full consideration of behavioral risk factors and their interaction with the genome may be one explanation of the small effect sizes or ORs (1.1 to 1.5) in published GWAS (*35*), although this is speculation.

Causal relationships between risk factors and disease are difficult to determine. However, investigators can now use genetic information to ascertain causality between risk factors and disease in an observational study (for example, HDL-C and cardiovascular disease) by using Mendelian randomization (*36*–*38*). Mendelian randomization is a method of using measured variation in genes of known function to examine the causal effect of a modifiable (nongenetic) exposure on disease in nonexperimental studies in epidemiology. If a trait exists on the causal pathway for disease, carriers of genetic variants associated with abnormal levels of the trait would be expected to be at different risk for disease. For example, Voight and colleagues have cast doubt on whether higher level of HDL-C is connected with a lower risk for myocardial infarction (*39*). The method described here provides a way of predicting relationships between traits and diseases, complementing Mendelian randomization. Predictions arising from similarity in genetic architecture such as the ones we have reported here may be tested in subsequent studies by using Mendelian randomization.

Another strategy to test predicted disease-trait associations is to use information from EMR, a resource that can provide patient phenotypic and physiological measurements, in the context of the clinical care setting, even before the diagnosis of disease (*40*, *41*). We used this approach here to validate five of our newly described disease-trait pairs. Our results show that these five clinical measurements can be risk factors for their paired diseases. This method could be expanded to cover larger and smaller units of time, or more distant time frames, as well as to take age into account.

Nevertheless, associations between complex traits and diseases discovered via genetic similarity and subsequent EMR-based retrospective validation cannot fully distinguish the causal relationships between traits and diseases. GWAS inherently capture only common variants, and consequently, certain associations between diseases and traits could be missing in our approach. In a tertiary care hospital setting, it is not always clear when and where the first diagnosis of disease took place by just looking at EMR data. We do not always know if a patient had been

diagnosed elsewhere or how long the patient has had a disease before their first observed diagnosis at each medical center. (The median onset age was correlated with known average ages of onset of each disease, suggesting that most of these patients did not receive care for any significant period of time elsewhere before presenting to a hospital setting.) *ICD-9 (International Classification of Diseases, 9th Revision)* codes also may not be clear enough for specific phenotype identification. That being said, we speculate that the codes we used for cancers are more likely to be accurately assigned than those for obesity and less severe disorders. Although methods for phenotyping from the eMERGE (*42*) project could have been deployed to reduce misclassification, the phenotypes we studied here were not yet listed in PheKB (*42*).

Laboratory values and measurements can be influenced by other related diseases or conditions and comorbidities. We did not control for these effects because there is no well-documented list of potential confounders for every laboratory measurement; however, we assumed that cases and controls were matched by a common set of characteristics. Additionally, it has been shown that hospitalized patients make poor control subjects, a phenomenon described as the Berkson bias, where a noncausal association exists between exposure and disease because of the condition that the subject has to come to the hospital to be involved in the study (*43*). Each individual relationship described through shared genetic architecture should be further tested in prospective epidemiology studies.

Here, we had also desired to evaluate the rest of the predicted disease-trait pairs. For instance, PSA was associated with testicular cancer (TCA), through *CLPTM1L* and *TERT* ($q < 0.001$; table S2). However, because the disease incidences were low at all three centers (only 22 at SHC, 33 at MSMC, and 65 patients at CUMC had PSA laboratory values measured before the first diagnosis for TCA), we did not have sufficient power to perform such analysis. Another finding was bone mineral density related to sudden cardiac arrest through the *ESR1* gene. Validation of findings such as these may be possible by using public health and longitudinal study data. Future studies to validate disease-trait pairs may require linking the EMR of multiple centers to gain the necessary numbers of patients needed.

In conclusion, investigation of traits that share genetic architecture with a disease and validating them through EMR is a powerful way to identify risk factors and prognostics. These associated traits show that risk factors need to be better considered or controlled in GWAS design to identify independent variants without the confounding of behavioral, environmental, or informative disease pathophysiology. Whether these traits can serve as diagnostic markers for complex diseases will depend on prospective trials.

MATERIALS AND METHODS

Extracting diseases and traits from VARIMED

As of this writing, VARIMED is a database of SNPs and diseases obtained from the manual review of 8962 human genetics papers including GWAS and candidate gene studies, with 87,553 SNPs mapped to 8913 genes and 1119 diseases and 1256 traits. We considered only diseases and traits whose genetic variants had genome-wide significance ($P < 1 \times 10^{-8}$) (*44*). Using this filter, we identified 201 diseases and 249 traits with at least one variant that mapped to a genic region. All genetic variants were then systematically mapped to genes with the most recent National Center for Biotechnology Information Entrez Gene identifiers through Entrez dbSNP using AILUN (*45*). SNPs in intergenic regions could not be associated with specific genes and were not considered. Next, to capture only highly relevant associations for enrichment, we kept only diseases and traits associated with at least three genes, yielding 69 diseases and 85 traits associated with 1439 genes. Distributions for the number of genes associated with diseases and traits were evaluated with Kolmogorov-Smirnov test (fig. S1A).

TF-IDF weighting scheme for shared genetic architecture between diseases and traits

For each gene associated with a disease or trait, we computed the gene popularity using the TF-IDF weighing method (*46*) to down-weight the ubiquitous genes that are associated with many diseases. For instance, *LPL* is associated with seven diseases/traits, whereas *CR1* is associated only with two diseases/traits (table S2). The detailed TF-IDF (*46*) calculation procedure for all 5865 combinations of disease-trait pairs (69 × 85) with 8913 genes is described as follows. First, we calculated a TF using $\mathrm{TF}_{(i,j)} = \frac{n_{i,j}}{\sum_k n_{k,j}}$, where $n_{i,j}$ is the number of occurrences of gene i in a particular disease or trait j. $\sum_k n_{k,j}$ indicates the total number of occurrences of all genes in a particular disease or trait j. The value of $\mathrm{TF}_{(i,j)}$ indicates the level of occurrence frequency of gene i in disease or trait j. Next, we calculated IDF using $\mathrm{IDF}_{(i)} = \log_{10}(\frac{D}{D_i})$. Here, D is the total number of diseases and traits, and D_i is the number of disease and trait containing gene i. A larger $\mathrm{IDF}_{(i)}$ implies a lower popularity of gene i among the diseases or traits, translating into more weight because it might only be shared between these two phenotypes among 8913 genes. Last, we calculated a TF-IDF score using $\mathrm{TF-IDF}_{(i,j)} = \mathrm{TF}_{(i,j)} \times \mathrm{IDF}_{(i)}$ for each gene within individual disease or trait by taking into account the popularity of the gene.

Assessing significance of disease-trait distance via the false discovery rate (*q* value)

We then calculated the false discovery rate (q value) to control for multiple-hypothesis testing and assess significance of similarity between diseases and traits. A q value (*47*) is an estimate of the rate of false positives incurred at a given significance threshold. Disease-trait similarity was estimated using the cosine distance between $\mathrm{TF-IDF}_{(i,j)}$ scores for all disease-trait combinations (equation as follows, where D and T are disease or trait and i is the gene shared between them).

$$\text{Cosine similarity}\,(D,T) = \frac{D \cdot T}{\|D\|\|T\|} = \frac{\sum_{i=1}^{n} D_i \times T_i}{\sqrt{\sum_{i=1}^{n} (D_i)^2} \times \sqrt{\sum_{i=1}^{n} (T_i)^2}}$$

Next, to evaluate the significance of a disease-trait distance score, we randomly shuffled the genes across all the traits and recomputed the disease-trait distance. We repeated the randomization procedure 1000 times to estimate the null distribution of the cosine distance for each pair. The q values were calculated as the ratio of the expected number of false positives over the total number of hypotheses tested (*47*). A q value of ≤0.01 was chosen as a significant association level between disease-trait pairs. Distributions for the number of PubMed counts reported for shared genes in known versus new discovered disease-trait pairs were evaluated with Kolmogorov-Smirnov test (fig. S1B).

Network visualization of the significant disease-trait pairs

We visualized a network representation of the disease-trait pairs identified as significant. We used Cytoscape 2.6.0 (*48*) and the CyOog (*49*) plug-in to represent and visualize the modular nature of the network, using all default settings. Diseases connected to the same trait were grouped into a super set (termed modules), as were traits connected to the same diseases. Each edge indicates a minimum significant association with $q \leq 0.01$; edge formation was not based on Cytoscape or CyOog.

Using EMR from three independent medical center database systems

We used adult patient EMR data from three medical centers after 1 January 2005 as independent cohorts to validate our findings. We identified case groups with the first diagnoses of target diseases using *ICD-9* diagnosis codes: 204.0 for acute lymphoid leukemia (ALL), 303 for ADS, 151 for GCA, 186 for TCA, 162 for LCA, 453 for VTE, and 185 for PCA. The control group for each analysis was taken from the adult patients without the diagnosis of target disease. Reference ranges for laboratory tests were based on MedlinePlus from the National Library of Medicine. They were as follows: serum/plasma PLT count, 150 to 400 K/μl; serum/plasma MGN, 1.8 to 2.4 mg/dl; MCV, 82 to 98 fl; ALP, 44 to 147 IU/liter; and PSA, <4 ng/ml.

Validation of newly described disease-trait pairs with EMR data

Use of EMR data was approved by individual's Institutional Review Board. To perform χ^2 tests, laboratory values were discretized. Values outside the reference range were defined as being in the "abnormal range." Values less than the low reference was defined as "low range," and those greater than the high reference were "high range." For a given test, we compared the maximum and minimum laboratory values to the reference range if multiple tests had been performed on a patient during the analysis time frame, which was defined as 1 year before disease diagnosis. Patients were defined as normal if laboratory results were within reference ranges, and abnormal if they were high or low range. Patients were excluded if multiple laboratory values were both high and low range.

We performed Wilcoxon rank-sum test by evaluating the actual laboratory values and χ^2 tests by calculating the ORs for abnormal ranges versus normal reference range between case and control groups. We report the ORs along with 95th percentile confidence intervals and *P* value. We compared the percentage of abnormal results for case and control patients 1 year before our first diagnosis code of the target disease in case patients, and in control patients who were cared for at SHC, MSMC, and CUMC and without diagnosis of target disease. This allowed us to investigate whether changes in laboratory values could be risk factors for predicting case incidence. In addition, logistic regression using generalized linear model function was also performed by adjusting age and gender variables in each prediction model, and the adjusted OR was also reported.

All statistics were computed by SAS 9.2 (SAS Institute) and R 2.15.1 (*50*).

SUPPLEMENTARY MATERIALS

www.sciencetranslationalmedicine.org/cgi/content/full/6/234/234ra57/DC1

Table S1A. Number of genes with variants associated with the 85 traits.

Table S1B. Number of genes with variants associated with the 69 diseases.

Table S2. One hundred twenty disease-trait pairs with shared common genes, *q* values derived from random sampling methods, and original GWAS studies from VARIMED.

Fig. S1A. Gene density for traits and diseases.

Fig. S1B. Published human genetic study density for the shared genes in new discovered and known disease-trait pairs.

Fig. S2. Violin plots for one positive control and two negative controls.

References (*51*–*127*)

REFERENCES AND NOTES

1. Wellcome Trust Case Control Consortium, Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
2. A. D. Johnson, C. J. O'Donnell, An open access database of genome-wide association results. *BMC Med. Genet.* **10**, 6 (2009).
3. U. P. Steinbrecher, M. Lougheed, Scavenger receptor-independent stimulation of cholesterol esterification in macrophages by low density lipoprotein extracted from human aortic intima. *Arterioscler. Thromb.* **12**, 608–625 (1992).
4. L. A. Hindorff, P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins, T. A. Manolio, Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 9362–9367 (2009).
5. H. Li, Y. Lee, J. L. Chen, E. Rebman, J. Li, Y. A. Lussier, Complex-disease networks of trait-associated single-nucleotide polymorphisms (SNPs) unveiled by information theory. *J. Am. Med. Inform. Assoc.* **19**, 295–305 (2012).
6. M. Sirota, M. A. Schaub, S. Batzoglou, W. H. Robinson, A. J. Butte, Autoimmune disease classification by inverse association with SNP alleles. *PLOS Genet.* **5**, e1000792 (2009).
7. K. I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, A. L. Barabási, The human disease network. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 8685–8690 (2007).
8. L. Li, D. Ruau, R. Chen, S. Weber, A. Butte, Systematic identification of risk factors for Alzheimer's disease through shared genetic architecture and electronic medical records. *Pac. Symp. Biocomput.* **18**, 224–235 (2013).
9. E. A. Ashley, A. J. Butte, M. T. Wheeler, R. Chen, T. E. Klein, F. E. Dewey, J. T. Dudley, K. E. Ormond, A. Pavlovic, A. A. Morgan, D. Pushkarev, N. F. Neff, L. Hudgins, L. Gong, L. M. Hodges, D. S. Berlin, C. F. Thorn, K. Sangkuhl, J. M. Hebert, M. Woon, H. Sagreiya, R. Whaley, J. W. Knowles, M. F. Chou, J. V. Thakuria, A. M. Rosenbaum, A. W. Zaranek, G. M. Church, H. T. Greely, S. R. Quake, R. B. Altman, Clinical assessment incorporating a personal genome. *Lancet* **375**, 1525–1535 (2010).
10. R. Chen, E. Corona, M. Sikora, J. T. Dudley, A. A. Morgan, A. Moreno-Estrada, G. B. Nilsen, D. Ruau, S. E. Lincoln, C. D. Bustamante, A. J. Butte, Type 2 diabetes risk alleles demonstrate extreme directional differentiation among human populations, compared to other diseases. *PLOS Genet.* **8**, e1002621 (2012).
11. R. Chen, E. V. Davydov, M. Sirota, A. J. Butte, Non-synonymous and synonymous coding SNPs show similar likelihood and effect size of human disease association. *PLOS One* **5**, e13574 (2010).
12. C. J. Patel, R. Chen, A. J. Butte, Data-driven integration of epidemiological and toxicological data to select candidate interacting genes and environmental factors in association with disease. *Bioinformatics* **28**, i121–i126 (2012).
13. S. Suthram, J. T. Dudley, A. P. Chiang, R. Chen, T. J. Hastie, A. J. Butte, Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets. *PLOS Comput. Biol.* **6**, e1000662 (2010).
14. F. E. Dewey, R. Chen, S. P. Cordero, K. E. Ormond, C. Caleshu, K. J. Karczewski, M. Whirl-Carrillo, M. T. Wheeler, J. T. Dudley, J. K. Byrnes, O. E. Cornejo, J. W. Knowles, M. Woon, K. Sangkuhl, L. Gong, C. F. Thorn, J. M. Hebert, E. Capriotti, S. P. David, A. Pavlovic, A. West, J. V. Thakuria, M. P. Ball, A. W. Zaranek, H. L. Rehm, G. M. Church, J. S. West, C. D. Bustamante, M. Snyder, R. B. Altman, T. E. Klein, A. J. Butte, E. A. Ashley, Phased whole-genome genetic risk in a family quartet using a major allele reference sequence. *PLOS Genet.* **7**, e1002280 (2011).
15. M. H. Cho, P. J. Castaldi, E. S. Wan, M. Siedlinski, C. P. Hersh, D. L. Demeo, B. E. Himes, J. S. Sylvia, B. J. Klanderman, J. P. Ziniti, C. Lange, A. A. Litonjua, D. Sparrow, E. A. Regan, B. J. Make, J. E. Hokanson, T. Murray, J. B. Hetmanski, S. G. Pillai, X. Kong, W. H. Anderson, R. Tal-Singer, D. A. Lomas, H. O. Coxson, L. D. Edwards, W. MacNee, J. Vestbo, J. C. Yates, A. Agusti, P. M. Calverley, B. Celli, C. Crim, S. Rennard, E. Wouters, P. Bakke, A. Gulsvik, J. D. Crapo, T. H. Beaty, E. K. Silverman, A genome-wide association study of COPD identifies a susceptibility locus on chromosome 19q13. *Hum. Mol. Genet.* **21**, 947–957 (2012).
16. S. G. Pillai, X. Kong, L. D. Edwards, M. H. Cho, W. H. Anderson, H. O. Coxson, D. A. Lomas, E. K. Silverman; ECLIPSE and ICGN Investigators, Loci identified by genome-wide association studies influence different disease-related phenotypes in chronic obstructive pulmonary disease. *Am. J. Respir. Crit. Care Med.* **182**, 1498–1505 (2010).
17. J. Wang, M. R. Spitz, C. I. Amos, A. V. Wilkinson, X. Wu, S. Shete, Mediating effects of smoking and chronic obstructive pulmonary disease on the relation between the CHRNA5-A3 genetic locus and lung cancer risk. *Cancer* **116**, 3458–3462 (2010).
18. M. H. Cho, N. Boutaoui, B. J. Klanderman, J. S. Sylvia, J. P. Ziniti, C. P. Hersh, D. L. DeMeo, G. M. Hunninghake, A. A. Litonjua, D. Sparrow, C. Lange, S. Won, J. R. Murphy, T. H. Beaty,

E. A. Regan, B. J. Make, J. E. Hokanson, J. D. Crapo, X. Kong, W. H. Anderson, R. Tal-Singer, D. A. Lomas, P. Bakke, A. Gulsvik, S. G. Pillai, E. K. Silverman, Variants in *FAM13A* are associated with chronic obstructive pulmonary disease. *Nat. Genet.* **42**, 200–202 (2010).
19. D. Lambrechts, I. Buysschaert, P. Zanen, J. Coolen, N. Lays, H. Cuppens, H. J. Groen, W. Dewever, R. J. van Klaveren, J. Verschakelen, C. Wijmenga, D. S. Postma, M. Decramer, W. Janssens, The 15q24/25 susceptibility variant for lung cancer and chronic obstructive pulmonary disease is associated with emphysema. *Am. J. Respir. Crit. Care Med.* **181**, 486–493 (2010).
20. S. G. Pillai, D. Ge, G. Zhu, X. Kong, K. V. Shianna, A. C. Need, S. Feng, C. P. Hersh, P. Bakke, A. Gulsvik, A. Ruppert, K. C. Lødrup Carlsen, A. Roses, W. Anderson, S. I. Rennard, D. A. Lomas, E. K. Silverman, D. B. Goldstein; ICGN Investigators, A genome-wide association study in chronic obstructive pulmonary disease (COPD): Identification of two major susceptibility loci. *PLOS Genet.* **5**, e1000421 (2009).
21. J. Vestbo, S. S. Hurd, A. G. Agustí, P. W. Jones, C. Vogelmeier, A. Anzueto, P. J. Barnes, L. M. Fabbri, F. J. Martinez, M. Nishimura, R. A. Stockley, D. D. Sin, R. Rodriguez-Roisin, Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: GOLD executive summary. *Am. J. Respir. Crit. Care Med.* **187**, 347–365 (2013).
22. Centers for Disease Control and Prevention, *The Health Consequences of Smoking: A Report of the Surgeon General* (Centers for Disease Control and Prevention, Atlanta, GA, 2004).
23. R. A. Pauwels, A. S. Buist, P. M. Calverley, C. R. Jenkins, S. S. Hurd; GOLD Scientific Committee, Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease. NHLBI/WHO Global Initiative for Chronic Obstructive Lung Disease (GOLD) Workshop summary. *Am. J. Respir. Crit. Care Med.* **163**, 1256–1276 (2001).
24. E. Png, A. Thalamuthu, R. T. Ong, H. Snippe, G. J. Boland, M. Seielstad, A genome-wide association study of hepatitis B vaccine response in an Indonesian population reveals multiple independent risk variants in the HLA region. *Hum. Mol. Genet.* **20**, 3893–3898 (2011).
25. N. Kato, F. Takeuchi, Y. Tabara, T. N. Kelly, M. J. Go, X. Sim, W. T. Tay, C. H. Chen, Y. Zhang, K. Yamamoto, T. Katsuya, M. Yokota, Y. J. Kim, R. T. Ong, T. Nabika, D. Gu, L. C. Chang, Y. Kokubo, W. Huang, K. Ohnaka, Y. Yamori, E. Nakashima, C. E. Jaquish, J. Y. Lee, M. Seielstad, M. Isono, J. E. Hixson, Y. T. Chen, T. Miki, X. Zhou, T. Sugiyama, J. P. Jeon, J. J. Liu, R. Takayanagi, S. S. Kim, T. Aung, Y. J. Sung, X. Zhang, T. Y. Wong, B. G. Han, S. Kobayashi, T. Ogihara, D. Zhu, N. Iwai, J. Y. Wu, Y. Y. Teo, E. S. Tai, Y. S. Cho, J. He, Meta-analysis of genome-wide association studies identifies common variants associated with blood pressure variation in east Asians. *Nat. Genet.* **43**, 531–538 (2011).
26. Y. J. Kim, M. J. Go, C. Hu, C. B. Hong, Y. K. Kim, J. Y. Lee, J. Y. Hwang, J. H. Oh, D. J. Kim, N. H. Kim, S. Kim, E. J. Hong, J. H. Kim, H. Min, Y. Kim, R. Zhang, W. Jia, Y. Okada, A. Takahashi, M. Kubo, T. Tanaka, N. Kamatani, K. Matsuda; MAGIC consortium, T. Park, B. Oh, K. Kimm, D. Kang, C. Shin, N. H. Cho, H. L. Kim, B. G. Han, Y. S. Cho, Large-scale genome-wide association studies in East Asians identify new genetic loci influencing metabolic traits. *Nat. Genet.* **43**, 990–995 (2011).
27. Y. Kamatani, K. Matsuda, Y. Okada, M. Kubo, N. Hosono, Y. Daigo, Y. Nakamura, N. Kamatani, Genome-wide association study of hematological and biochemical traits in a Japanese population. *Nat. Genet.* **42**, 210–215 (2010).
28. K. G. Kahl, W. Greggersen, U. Schweiger, J. Cordes, C. U. Correll, J. Ristow, J. Burow, C. Findel, A. Stoll, C. Balijepalli, L. Göres, C. Lösch, T. Hillemacher, S. Bleich, S. Moebus, Prevalence of the metabolic syndrome in men and women with alcohol dependence: Results from a cross-sectional study during behavioural treatment in a controlled environment. *Addiction* **105**, 1921–1927 (2010).
29. G. A. Gross, Drug and alcohol abuse and cholesterol levels. *J. Am. Osteopath. Assoc.* **94**, 55–56, 61–62 (1994).
30. A. Imhof, M. Froehlich, H. Brenner, H. Boeing, M. B. Pepys, W. Koenig, Effect of alcohol consumption on systemic markers of inflammation. *Lancet* **357**, 763–767 (2001).
31. H. J. Lowe, T. A. Ferris, P. M. Hernandez, S. C. Weber, STRIDE—An integrated standards-based translational research informatics platform. *AMIA Annu. Symp. Proc.* **2009**, 391–395 (2009).
32. P. W. Wilson, R. B. D'Agostino, D. Levy, A. M. Belanger, H. Silbershatz, W. B. Kannel, Prediction of coronary heart disease using risk factor categories. *Circulation* **97**, 1837–1847 (1998).
33. E. Zeggini, M. N. Weedon, C. M. Lindgren, T. M. Frayling, K. S. Elliott, H. Lango, N. J. Timpson, J. R. Perry, N. W. Rayner, R. M. Freathy, J. C. Barrett, B. Shields, A. P. Morris, S. Ellard, C. J. Groves, L. W. Harries, J. L. Marchini, K. R. Owen, B. Knight, L. R. Cardon, M. Walker, G. A. Hitman, A. D. Morris, A. S. Doney; Wellcome Trust Case Control Consortium (WTCCC), M. I. McCarthy, A. T. Hattersley, Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* **316**, 1336–1341 (2007).
34. T. M. Frayling, N. J. Timpson, M. N. Weedon, E. Zeggini, R. M. Freathy, C. M. Lindgren, J. R. Perry, K. S. Elliott, H. Lango, N. W. Rayner, B. Shields, L. W. Harries, J. C. Barrett, S. Ellard, C. J. Groves, B. Knight, A. M. Patch, A. R. Ness, S. Ebrahim, D. A. Lawlor, S. M. Ring, Y. Ben-Shlomo, M. R. Jarvelin, U. Sovio, A. J. Bennett, D. Melzer, L. Ferrucci, R. J. Loos, I. Barroso, N. J. Wareham, F. Karpe, K. R. Owen, L. R. Cardon, M. Walker, G. A. Hitman, C. N. Palmer, A. S. Doney, A. D. Morris, G. D. Smith, A. T. Hattersley, M. I. McCarthy, A common variant in the *FTO* gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* **316**, 889–894 (2007).
35. T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. Mackay, S. A. McCarroll, P. M. Visscher, Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
36. S. C. Harrison, M. V. Holmes, S. E. Humphries, Mendelian randomisation, lipids, and cardiovascular disease. *Lancet* **380**, 543–545 (2012).
37. N. A. Sheehan, V. Didelez, P. R. Burton, M. D. Tobin, Mendelian randomisation and causal inference in observational epidemiology. *PLOS Med.* **5**, e177 (2008).
38. G. D. Smith, S. Ebrahim, 'Mendelian randomization': Can genetic epidemiology contribute to understanding environmental determinants of disease? *Int. J. Epidemiol.* **32**, 1–22 (2003).
39. B. F. Voight, G. M. Peloso, M. Orho-Melander, R. Frikke-Schmidt, M. Barbalic, M. K. Jensen, G. Hindy, H. Hólm, E. L. Ding, T. Johnson, H. Schunkert, N. J. Samani, R. Clarke, J. C. Hopewell, J. F. Thompson, M. Li, G. Thorleifsson, C. Newton-Cheh, K. Musunuru, J. P. Pirruccello, D. Saleheen, L. Chen, A. Stewart, A. Schillert, U. Thorsteinsdottir, G. Thorgeirsson, S. Anand, J. C. Engert, T. Morgan, J. Spertus, M. Stoll, K. Berger, N. Martinelli, D. Girelli, P. P. McKeown, C. C. Patterson, S. E. Epstein, J. Devaney, M. S. Burnett, V. Mooser, S. Ripatti, I. Surakka, M. S. Nieminen, J. Sinisalo, M. L. Lokki, M. Perola, A. Havulinna, U. de Faire, B. Gigante, E. Ingelsson, T. Zeller, P. Wild, P. I. de Bakker, O. H. Klungel, A. H. Maitland-van der Zee, B. J. Peters, A. de Boer, D. E. Grobbee, P. W. Kamphuisen, V. H. Deneer, C. C. Elbers, N. C. Onland-Moret, M. H. Hofker, C. Wijmenga, W. M. Verschuren, J. M. Boer, Y. T. van der Schouw, A. Rasheed, P. Frossard, S. Demissie, C. Willer, R. Do, J. M. Ordovas, G. R. Abecasis, M. Boehnke, K. L. Mohlke, M. J. Daly, C. Guiducci, N. P. Burtt, A. Surti, E. Gonzalez, S. Purcell, S. Gabriel, J. Marrugat, J. Peden, J. Erdmann, P. Diemert, C. Willenborg, I. R. König, M. Fischer, C. Hengstenberg, A. Ziegler, I. Buysschaert, D. Lambrechts, F. Van de Werf, K. A. Fox, N. E. El Mokhtari, D. Rubin, J. Schrezenmeir, S. Schreiber, A. Schäfer, J. Danesh, S. Blankenberg, R. Roberts, R. McPherson, H. Watkins, A. S. Hall, K. Overvad, E. Rimm, E. Boerwinkle, A. Tybjaerg-Hansen, L. A. Cupples, M. P. Reilly, O. Melander, P. M. Mannucci, D. Ardissino, D. Siscovick, R. Elosua, K. Stefansson, C. J. O'Donnell, V. Salomaa, D. J. Rader, L. Peltonen, S. M. Schwartz, D. Altshuler, S. Kathiresan, Plasma HDL cholesterol and risk of myocardial infarction: A mendelian randomisation study. *Lancet* **380**, 572–580 (2012).
40. J. C. Denny, D. C. Crawford, M. D. Ritchie, S. J. Bielinski, M. A. Basford, Y. Bradford, H. S. Chai, L. Bastarache, R. Zuvich, P. Peissig, D. Carrell, A. H. Ramirez, J. Pathak, R. A. Wilke, L. Rasmussen, X. Wang, J. A. Pacheco, A. N. Kho, M. G. Hayes, N. Weston, M. Matsumoto, P. A. Kopp, K. M. Newton, G. P. Jarvik, R. Li, T. A. Manolio, I. J. Kullo, C. G. Chute, R. L. Chisholm, E. B. Larson, C. A. McCarty, D. R. Masys, D. M. Roden, M. de Andrade, Variants near *FOXE1* are associated with hypothyroidism and other thyroid conditions: Using electronic medical records for genome- and phenome-wide studies. *Am. J. Hum. Genet.* **89**, 529–542 (2011).
41. K. D. Mandl, I. S. Kohane, Escaping the EHR trap—The future of health IT. *N. Engl. J. Med.* **366**, 2240–2242 (2012).
42. A. N. Kho, J. A. Pacheco, P. L. Peissig, L. Rasmussen, K. M. Newton, N. Weston, P. K. Crane, J. Pathak, C. G. Chute, S. J. Bielinski, I. J. Kullo, R. Li, T. A. Manolio, R. L. Chisholm, J. C. Denny, Electronic medical records for genetic research: Results of the eMERGE consortium. *Sci. Transl. Med.* **3**, 79re1 (2011).
43. D. Westreich, Berkson's bias, selection bias, and missing data. *Epidemiology* **23**, 159–164 (2012).
44. J. Ott, Y. Kamatani, M. Lathrop, Family-based designs for genome-wide association studies. *Nat. Rev. Genet.* **12**, 465–474 (2011).
45. R. Chen, L. Li, A. J. Butte, AILUN: Reannotating gene expression data automatically. *Nat. Methods* **4**, 879 (2007).
46. H. C. Wu, R. W. P. Luk, K. F. Wong, K. L. Kwok, Interpreting TF-IDF term weights as making relevance decisions. *ACM Trans. Inform. Syst.* **26**, Article No. 13 (2008).
47. J. D. Storey, R. Tibshirani, Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 9440–9445 (2003).
48. P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, T. Ideker, Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
49. L. Royer, M. Reimann, B. Andreopoulos, M. Schroeder, Unraveling protein networks with power graph analysis. *PLOS Comput. Biol.* **4**, e1000108 (2008).
50. R. Ihaka, R. Gentleman, R: A language for data analysis and graphics. *J. Comput. Graphic. Stat.* **5**, 299–314 (1996).
51. D. J. Tobin, N. Orentreich, D. A. Fenton, J. C. Bystryn, Antibodies to hair follicles in alopecia areata. *J. Invest. Dermatol.* **102**, 721–724 (1994).
52. P. Caramelli, R. Nitrini, R. Maranhao, A. C. Lourenco, M. C. Damasceno, C. Vinagre, B. Caramelli, Increased apolipoprotein B serum concentration in Alzheimer's disease. *Acta Neurol. Scand.* **100**, 61–63 (1999).
53. S. E. O'Bryant, S. C. Waring, V. Hobson, J. R. Hall, C. B. Moore, T. Bottiglieri, P. Massman, R. Diaz-Arrastia, Decreased C-reactive protein levels in Alzheimer disease. *J. Geriatr. Psychiatry Neurol.* **23**, 49–53 (2010).

54. T. Matsuzaki, K. Sasaki, J. Hata, Y. Hirakawa, K. Fujimi, T. Ninomiya, S. O. Suzuki, S. Kanba, Y. Kiyohara, T. Iwaki, Association of Alzheimer disease pathology with abnormal lipid metabolism: The Hisayama Study. *Neurology* **77**, 1068–1075 (2011).
55. G. T. Lesser, V. Haroutunian, D. P. Purohit, M. Schnaider Beeri, J. Schmeidler, L. Honkanen, R. Neufeld, L. S. Libow, Serum lipids are related to Alzheimer's pathology in nursing home residents. *Dement. Geriatr. Cogn. Disord.* **27**, 42–49 (2009).
56. M. van Oijen, I. M. van der Meer, A. Hofman, J. C. Witteman, P. J. Koudstaal, M. M. Breteler, Lipoprotein-associated phospholipase A2 is associated with risk of dementia. *Ann. Neurol.* **59**, 139–144 (2006).
57. J. H. Ringrose, HLA-B27 associated spondyloarthropathy, an autoimmune disease based on crossreactivity between bacteria and HLA-B27? *Ann. Rheum. Dis.* **58**, 598–610 (1999).
58. D. H. Bryant, M. W. Burns, L. Lazarus, The correlation between skin tests, bronchial provocation tests and the serum level of IgE specific for common allergens in patients with asthma. *Clin. Allergy* **5**, 145–157 (1975).
59. V. P. Chinem, H. A. Miot, Prevalence of actinic skin lesions in patients with basal cell carcinoma of the head: A case-control study. *Rev. Assoc. Med. Bras.* **58**, 188–196 (2012).
60. R. Zanetti, S. Rosso, C. Martinez, A. Nieto, A. Miranda, M. Mercier, D. I. Loria, A. Østerlind, R. Greinert, C. Navarro, G. Fabbrocini, C. Barbera, H. Sancho-Garnier, L. Gafà, A. Chiarugi, R. Mossotti, Comparison of risk patterns in carcinoma and melanoma of the skin in men: A multi-centre case–case–control study. *Br. J. Cancer* **94**, 743–751 (2006).
61. A. Kikuchi, H. Shimizu, T. Nishikawa, Clinical histopathological characteristics of basal cell carcinoma in Japanese patients. *Arch. Dermatol.* **132**, 320–324 (1996).
62. K. M. Musallam, V. G. Sankaran, M. D. Cappellini, L. Duca, D. G. Nathan, A. T. Taher, Fetal hemoglobin levels and morbidity in untransfused patients with β-thalassemia intermedia. *Blood* **119**, 364–367 (2012).
63. M. M. Kaplan, M. E. Gershwin, Primary biliary cirrhosis. *N. Engl. J. Med.* **353**, 1261–1273 (2005).
64. E. Emanuele, M. V. Carlin, A. D'Angelo, E. Peros, F. Barale, D. Geroldi, P. Politi, Elevated plasma levels of lipoprotein(a) in psychiatric patients: A possible contribution to increased vascular risk. *Eur. Psychiatry* **21**, 129–133 (2006).
65. M. Sagud, A. Mihaljevic-Peles, N. Pivac, M. Jakovljevic, D. Muck-Seler, Lipid levels in female patients with affective disorders. *Psychiatry Res.* **168**, 218–221 (2009).
66. V. M. Villarejos, J. Serra, K. A. Visoná, C. E. Eduarte, Antibodies to single stranded DNA: A diagnostic aid in chronic hepatitis B virus infections. *J. Med. Virol.* **4**, 97–101 (1979).
67. M. Benn, Apolipoprotein B levels, *APOB* alleles, and risk of ischemic cardiovascular disease in the general population, a review. *Atherosclerosis* **206**, 17–30 (2009).
68. G. W. Burggraf, J. O. Parker, Prognosis in coronary artery disease. Angiographic, hemodynamic, and clinical factors. *Circulation* **51**, 146–156 (1975).
69. M. P. Reilly, M. L. Wolfe, J. Dykhouse, K. Reddy, A. R. Localio, D. J. Rader, Intercellular adhesion molecule 1 (ICAM-1) gene variant is associated with coronary artery calcification independent of soluble ICAM-1 levels. *J. Investig. Med.* **52**, 515–522 (2004).
70. C. J. Packard, D. S. O'Reilly, M. J. Caslake, A. D. McMahon, I. Ford, J. Cooney, C. H. Macphee, K. E. Suckling, M. Krishna, F. E. Wilkinson, A. Rumley, G. D. Lowe, Lipoprotein-associated phospholipase A2 as an independent predictor of coronary heart disease. West of Scotland Coronary Prevention Study Group. *N. Engl. J. Med.* **343**, 1148–1155 (2000).
71. K. Vikenes, M. Farstad, J. E. Nordrehaug, Serotonin is associated with coronary artery disease and cardiac events. *Circulation* **100**, 483–489 (1999).
72. A. Sako, J. Kitayama, S. Kaisaki, H. Nagawa, Hyperlipidemia is a risk factor for lymphatic metastasis in superficial esophageal carcinoma. *Cancer Lett.* **208**, 43–49 (2004).
73. M. A. Martínez-García, M. Luque-Ramírez, J. L. San-Millán, H. F. Escobar-Morreale, Body iron stores and glucose intolerance in premenopausal women: Role of hyperandrogenism, insulin resistance, and genomic variants related to inflammation, oxidative stress, and iron metabolism. *Diabetes Care* **32**, 1525–1530 (2009).
74. S. Haghighi, M. Amini, Z. Pournaghshband, P. Amini, S. Hovsepian, Relationship between gamma-glutamyl transferase and glucose intolerance in first degree relatives of type 2 diabetics patients. *J. Res. Med. Sci.* **16**, 123–129 (2011).
75. R. R. Little, J. D. England, H. M. Wiedmeyer, E. M. McKenzie, D. J. Pettitt, W. C. Knowler, D. E. Goldstein, Relationship of glycosylated hemoglobin to oral glucose tolerance. Implications for diabetes screening. *Diabetes* **37**, 60–64 (1988).
76. J. D. McGarry, Disordered metabolism in diabetes: Have we underemphasized the fat component? *J. Cell. Biochem.* **55**, 29–38 (1994).
77. F. Fumeron, F. Péan, F. Driss, B. Balkau, J. Tichet, M. Marre, B. Grandchamp; Insulin Resistance Syndrome (DESIR) Study Group, Ferritin and transferrin are both predictive of the onset of hyperglycemia in men and women over 3 years: The data from an epidemiological study on the Insulin Resistance Syndrome (DESIR) study. *Diabetes Care* **29**, 2090–2094 (2006).
78. Y. Urwijitaroon, S. Barusrux, A. Romphruk, C. Puapairoj, P. Thongkrajai, Anti-HIV antibody titer: An alternative supplementary test for diagnosis of HIV-1 infection. *Asian Pac. J. Allergy Immunol.* **15**, 193–198 (1997).
79. A. V. Chobanian, G. L. Bakris, H. R. Black, W. C. Cushman, L. A. Green, J. L. Izzo Jr., D. W. Jones, B. J. Materson, S. Oparil, J. T. Wright Jr., E. J. Roccella; Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure. National Heart, Lung, and Blood Institute; National High Blood Pressure Education Program Coordinating Committee, Seventh report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure. *Hypertension* **42**, 1206–1252 (2003).
80. J. Coresh, G. L. Wei, G. McQuillan, F. L. Brancati, A. S. Levey, C. Jones, M. J. Klag, Prevalence of high blood pressure and elevated serum creatinine level in the United States: Findings from the third National Health and Nutrition Examination Survey (1988-1994). *Arch. Intern. Med.* **161**, 1207–1216 (2001).
81. M. K. Kim, K. H. Baek, K. H. Song, M. I. Kang, J. H. Choi, J. C. Bae, C. Y. Park, W. Y. Lee, K. W. Oh, Increased serum ferritin predicts the development of hypertension among middle-aged men. *Am. J. Hypertension* **25**, 492–497 (2012).
82. M. Cirillo, M. Laurenzi, M. Trevisan, J. Stamler, Hematocrit, blood pressure, and hypertension. The Gubbio Population Study. *Hypertension* **20**, 319–326 (1992).
83. G. F. Strippoli, J. C. Craig, C. Manno, F. P. Schena, Hemoglobin targets for the anemia of chronic kidney disease: A meta-analysis of randomized, controlled trials. *J. Am. Soc. Nephrol.* **15**, 3154–3165 (2004).
84. K. Sutton-Tyrrell, A. Bostom, J. Selhub, C. Zeigler-Johnson, High homocysteine levels are independently related to isolated systolic hypertension in older adults. *Circulation* **96**, 1745–1749 (1997).
85. R. Rahim, K. Nahar, I. A. Khan, Platelet count in 100 cases of pregnancy induced hypertension. *Mymensingh Med. J.* **19**, 5–9 (2010).
86. S. M. Dogan, M. Aydin, M. Gursurer, A. Dursun, G. Mungan, T. Onuk, N-terminal probrain natriuretic peptide predicts altered circadian variation in essential hypertension. *Coron. Artery Dis.* **18**, 347–352 (2007).
87. D. J. Buchan, Diagnosis and management of inflammatory bowel disease. *Can. Fam. Physician* **22**, 47–51 (1976).
88. A. D. Pradhan, J. E. Manson, N. Rifai, J. E. Buring, P. M. Ridker, C-reactive protein, interleukin 6, and risk of developing type 2 diabetes mellitus. *JAMA* **286**, 327–334 (2001).
89. M. W. Mansfield, D. M. Heywood, P. J. Grant, Circulating levels of factor VII, fibrinogen, and von Willebrand factor and features of insulin resistance in first-degree relatives of patients with NIDDM. *Circulation* **94**, 2171–2176 (1996).
90. A. Borai, C. Livingstone, F. Abdelaal, A. Bawazeer, V. Keti, G. Ferns, The relationship between glycosylated haemoglobin (HbA1c) and measures of insulin resistance across a range of glucose tolerance. *Scand. J. Clin. Lab. Invest.* **71**, 168–172 (2011).
91. A. Laws, G. M. Reaven, Evidence for an independent relationship between insulin resistance and fasting plasma HDL-cholesterol, triglyceride and insulin concentrations. *J. Intern. Med.* **231**, 25–30 (1992).
92. T. L. Nelson, M. L. Biggs, J. R. Kizer, M. Cushman, J. E. Hokanson, C. D. Furberg, K. J. Mukamal, Lipoprotein-associated phospholipase A_2 (Lp-PLA_2) and future risk of type 2 diabetes: Results from the Cardiovascular Health Study. *J. Clin. Endocrinol. Metab.* **97**, 1695–1701 (2012).
93. A. J. Hanley, R. D'Agostino Jr., L. E. Wagenknecht, M. F. Saad, P. J. Savage, R. Bergman, S. M. Haffner; Insulin Resistance Atherosclerosis Study, Increased proinsulin levels and decreased acute insulin response independently predict the incidence of type 2 diabetes in the insulin resistance atherosclerosis study. *Diabetes* **51**, 1263–1270 (2002).
94. J. Bazelmans, P. J. Nestel, C. Nolan, Insulin-induced glucose utilization influences triglyceride metabolism. *Clin. Sci.* **64**, 511–516 (1983).
95. J. C. Fink, R. A. Burdick, S. J. Kurth, S. A. Blahut, N. C. Armistead, M. S. Turner, L. M. Shickle, P. D. Light, Significance of serum creatinine values in new end-stage renal disease patients. *Am. J. Kidney Dis.* **34**, 694–701 (1999).
96. S. A. Khuder, Effect of cigarette smoking on major histological types of lung cancer: A meta-analysis. *Lung Cancer* **31**, 139–148 (2001).
97. E. A. Holly, D. A. Aston, R. D. Cress, D. K. Ahn, J. J. Kristiansen, Cutaneous melanoma in women. II. Phenotypic characteristics and other host-related factors. *Am. J. Epidemiol.* **141**, 934–942 (1995).
98. M. A. Tucker, Melanoma epidemiology. *Hematol. Oncol. Clin. North Am.* **23**, 383–395 (2009).
99. M. B. Veierød, E. Weiderpass, M. Thörn, J. Hansson, E. Lund, B. Armstrong, H. O. Adami, A prospective study of pigmentation, sun exposure, and risk of cutaneous malignant melanoma in women. *J. Natl. Cancer Inst.* **95**, 1530–1538 (2003).
100. W. Yumura, S. Suganuma, K. Nitta, Y. Sano, K. Uchida, H. Nihei, Prolonged membranous lupus nephritis with change of anti-ssDNA antibody titer and repeated renal relapse. *Clin. Exp. Nephrol.* **8**, 363–368 (2004).
101. J. Sierra-Johnson, V. K. Somers, F. H. Kuniyoshi, C. A. Garza, W. L. Isley, A. S. Gami, F. Lopez-Jimenez, Comparison of apolipoprotein-B/apolipoprotein-AI in subjects with versus without the metabolic syndrome. *Am. J. Cardiol.* **98**, 1369–1373 (2006).
102. M. Fröhlich, A. Imhof, G. Berg, W. L. Hutchinson, M. B. Pepys, H. Boeing, R. Muche, H. Brenner, W. Koenig, Association between C-reactive protein and features of the metabolic syndrome: A population-based study. *Diabetes Care* **23**, 1835–1839 (2000).
103. T. Gombet, B. Longo-Mbenza, B. Ellenga-Mbolla, M. S. Ikama, E. Mokondjimobe, G. Kimbally-Kaky, J. L. Nkoua, Aging, female sex, migration, elevated HDL-C, and inflammation are associated

with prevalence of metabolic syndrome among African bank employees. *Int. J. Gen. Med.* **5**, 495–503 (2012).
104. H. P. Gong, Y. M. Du, L. N. Zhong, Z. Q. Dong, X. Wang, Y. J. Mao, Q. H. Lu, Plasma lipoprotein-associated phospholipase A2 in patients with metabolic syndrome and carotid atherosclerosis. *Lipids Health Dis.* **10**, 13 (2011).
105. E. Burns, G. P. Mulley, Practical problems with eye-drops among elderly ophthalmology outpatients. *Age Ageing* **21**, 168–170 (1992).
106. P. H. Lalive, T. Menge, C. Delarasse, B. Della Gaspera, D. Pham-Dinh, P. Villoslada, H. C. von Büdingen, C. P. Genain, Antibodies to native myelin oligodendrocyte glycoprotein are serologic markers of early inflammation in multiple sclerosis. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 2280–2285 (2006).
107. A. Brucato, R. Cimaz, R. Caporali, V. Ramoni, J. Buyon, Pregnancy outcomes in patients with autoimmune diseases and anti-Ro/SSA antibodies. *Clin. Rev. Allergy Immunol.* **40**, 27–41 (2011).
108. T. S. Burgert, S. E. Taksali, J. Dziura, T. R. Goodman, C. W. Yeckel, X. Papademetris, R. T. Constable, R. Weiss, W. V. Tamborlane, M. Savoye, A. A. Seyal, S. Caprio, Alanine aminotransferase levels and fatty liver in childhood obesity: Associations with insulin resistance, adiponectin, and visceral fat. *J. Clin. Endocrinol. Metab.* **91**, 4287–4294 (2006).
109. I. L. Mertens, L. F. Van Gaal, Overweight, obesity, and blood pressure: The effects of modest weight reduction. *Obes. Res.* **8**, 270–278 (2000).
110. T. Kamoda, H. Saitoh, M. Inudoh, K. Miyazaki, A. Matsui, The serum levels of proinsulin and their relationship with IGFBP-1 in obese children. *Diabetes Obes. Metab.* **8**, 192–196 (2006).
111. M. Nakamura, Y. Shimizu-Yoshida, Y. Takii, A. Komori, T. Yokoyama, T. Ueki, M. Daikoku, K. Yano, T. Matsumoto, K. Migita, H. Yatsuhashi, M. Ito, N. Masaki, H. Adachi, Y. Watanabe, Y. Nakamura, T. Saoshiro, T. Sodeyama, M. Koga, S. Shimoda, H. Ishibashi, Antibody titer to gp210-C terminal peptide as a clinical parameter for monitoring primary biliary cirrhosis. *J. Hepatol.* **42**, 386–392 (2005).
112. D. S. Smith, P. A. Humphrey, W. J. Catalona, The early detection of prostate carcinoma with prostate specific antigen: The Washington University experience. *Cancer* **80**, 1852–1856 (1997).
113. K. Egerer, E. Feist, G. R. Burmester, The serological diagnosis of rheumatoid arthritis: Antibodies to citrullinated antigens. *Dtsch. Arztebl. Int.* **106**, 159–163 (2009).
114. L. S. Avnon, F. Manzur, A. Bolotin, D. Heimer, D. Flusser, D. Buskila, S. Sukenik, M. Abu-Shakra, Pulmonary functions testing in patients with rheumatoid arthritis. *Isr. Med. Assoc. J.* **11**, 83–87 (2009).
115. J. M. Gill, A. M. Quisel, P. V. Rocca, D. T. Walters, Diagnosis of systemic lupus erythematosus. *Am. Fam. Physician* **68**, 2179–2186 (2003).
116. Y. Morita, Y. Muro, K. Sugiura, Y. Tomita, Anti-cyclic citrullinated peptide antibody in systemic sclerosis. *Clin. Exp. Rheumatol.* **26**, 542–547 (2008).
117. F. Vaziri-Sani, S. Oak, J. Radtke, K. Lernmark, K. Lynch, C. D. Agardh, C. M. Cilio, A. L. Lethagen, E. Ortqvist, M. Landin-Olsson, C. Törn, C. S. Hampe, ZnT8 autoantibody titers in type 1 diabetes patients decline rapidly after clinical onset. *Autoimmunity* **43**, 598–606 (2010).
118. M. Rönnback, J. Fagerudd, C. Forsblom, K. Pettersson-Fernholm, A. Reunanen, P. H. Groop; Finnish Diabetic Nephropathy (FinnDiane) Study Group, Altered age-related blood pressure pattern in type 1 diabetes. *Circulation* **110**, 1076–1082 (2004).
119. H. Gylling, J. A. Tuominen, V. A. Koivisto, T. A. Miettinen, Cholesterol metabolism in type 1 diabetes. *Diabetes* **53**, 2217–2222 (2004).
120. J. Ma, A. Mollsten, M. Prazny, H. Falhammar, K. Brismar, G. Dahlquist, S. Efendic, H. F. Gu, Genetic influences of the intercellular adhesion molecule 1 (ICAM-1) gene polymorphisms in development of type 1 diabetes and diabetic nephropathy. *Diabetic Med.* **23**, 1093–1099 (2006).
121. S. Li, H. J. Shin, E. L. Ding, R. M. van Dam, Adiponectin levels and risk of type 2 diabetes: A systematic review and meta-analysis. *JAMA* **302**, 179–188 (2009).
122. A. D. Pradhan, J. E. Manson, J. B. Meigs, N. Rifai, J. E. Buring, S. Liu, P. M. Ridker, Insulin, proinsulin, proinsulin:insulin ratio, and the risk of developing type 2 diabetes mellitus in women. *Am. J. Med.* **114**, 438–444 (2003).
123. J. A. Chirinos, G. A. Heresi, H. Velasquez, W. Jy, J. J. Jimenez, E. Ahn, L. L. Horstman, A. O. Soriano, J. P. Zambrano, Y. S. Ahn, Elevation of endothelial microparticles, platelets, and leukocyte activation in patients with venous thromboembolism. *J. Am. Coll. Cardiol.* **45**, 1467–1471 (2005).
124. S. Shrivastava, P. M. Ridker, R. J. Glynn, S. Z. Goldhaber, S. Moll, H. Bounameaux, K. A. Bauer, C. M. Kessler, M. Cushman, D-dimer, factor VIII coagulant activity, low-intensity warfarin and the risk of recurrent venous thromboembolism. *J. Thromb. Haemost.* **4**, 1208–1214 (2006).
125. P. V. Jenkins, O. Rawley, O. P. Smith, J. S. O'Donnell, Elevated factor VIII levels and risk of venous thrombosis. *Bri. J. Haematol.* **157**, 653–663 (2012).
126. Y. K. Park, N. S. Kim, S. K. Hann, S. Im, Identification of autoantibody to melanocytes and characterization of vitiligo antigen in vitiligo patients. *J. Dermatol. Sci.* **11**, 111–120 (1996).
127. S. N. Wong, V. Shah, M. J. Dillon, Antineutrophil cytoplasmic antibodies in Wegener's granulomatosis. *Arch. Dis. Child.* **79**, 246–250 (1998).

Acknowledgments: We are grateful to M. Snyder (Department of Genetics, Stanford School of Medicine) and J. Caballero (Department of Anesthesia, Stanford School of Medicine) for reviewing this manuscript and providing helpful comments and suggestions. We thank A. Skrenchuk at Stanford for Linux cluster computing and database support. **Funding:** This study was supported in part by the Lucile Packard Foundation for Children's Health. The STRIDE project was supported by the National Center for Research Resources and the National Center for Advancing Translational Sciences, NIH, through grant UL1 RR025744. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. **Author contributions:** L.L. and A.J.B. conceived and designed the study. L.L. performed the experiments. L.L. and D.J.R. analyzed the data. L.L., D.J.R., C.J.P., S.C.W., R.C., J.T.D., N.P.T., and A.J.B. contributed reagents/materials/analysis tools. L.L., D.J.R., C.J.P., and A.J.B. wrote the paper. **Competing interests:** A.J.B. is a founder and consultant of Personalis Inc., a genetic testing company. R.C. is an employee of Personalis Inc. The other authors declare that they have no competing interests.

Submitted 30 July 2013
Accepted 11 March 2014
Published 30 April 2014
10.1126/scitranslmed.3007191

Citation: L. Li, D. J. Ruau, C. J. Patel, S. C. Weber, R. Chen, N. P. Tatonetti, J. T. Dudley, A. J. Butte, Disease risk factors identified through shared genetic architecture and electronic medical records. *Sci. Transl. Med.* **6**, 234ra57 (2014).

Science Translational Medicine AAAS

Disease Risk Factors Identified Through Shared Genetic Architecture and Electronic Medical Records

Li Li, David J. Ruau, Chirag J. Patel, Susan C. Weber, Rong Chen, Nicholas P. Tatonetti, Joel T. Dudley and Atul J. Butte (April 30, 2014)

Science Translational Medicine **6** (234), 234ra57. [doi: 10.1126/scitranslmed.3007191]

Editor's Summary

Medicine by Association

As data get bigger, the challenge is to extract human-sized conclusions that we can comprehend and use. Li and colleagues have done exactly this by exploiting VARIMED, a hand-curated database of single-nucleotide polymorphisms (SNP) associated with diseases or clinical parameters such as cholesterol level and smoking status, extracted from the literature.

By finding pairs of diseases and these nondisease clinical parameters (which they call traits) that are associated with the same SNP variants, they construct hypotheses that the traits could be prognostic markers or risk factors for the disease. Ninety-four of the 120 pairs they identified were known and published in the literature; 26 pairs were previously undescribed. The known associations tended to fall into groups: solid organ cancer with prostate-specific antigen (PSA) and autoimmune disorders with major histocompatibility complex (MHC)–related molecules, for example. The authors were able to validate several of the newly associated traits and diseases by extracting data from electronic medical records from three clinical centers: They found that patients with abnormal mean corpuscular volume were more than three times more likely to receive a diagnosis of acute lymphoblastic leukemia within a year than those with normal values. Similarly, abnormal magnesium levels predicted a greater risk of developing gastric cancer within a year, and abnormally high PSA levels predicted a doubling in the odds of receiving a lung cancer diagnosis within a year.

This all in silico discovery and validation of potential risk factors for disease present an important hypothesis-generating tool for medicine. Prospective clinical trials will test whether these clinical traits can serve as informative diagnostic and prognostic markers.

Science Translational Medicine (print ISSN 1946-6234; online ISSN 1946-6242) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue, NW, Washington, DC 20005.