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

Li Li,1,2a David J. Ruau,1,2a Chirag J. Patel,1,2,3 Susan C. Weber,4 Rong Chen,1,5 Nicholas P. Tatonetti,5 Joel T. Dudley,7 Atul J. Butte1,2,†

INTRODUCTION

Genome-wide association studies (GWAS) and candidate gene approaches 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.

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 × 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),
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 of the 120 significant disease-trait associations 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).

**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 MCIR. 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: pigmented characteristics (Fig. 2, T1) with skin cancer (D2) through MCIR 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. 1. Diagram for identifying significant disease-trait genetic associations.**

<table>
<thead>
<tr>
<th>Disease modules (n = 8)</th>
<th>Trait modules (n = 7)</th>
</tr>
</thead>
<tbody>
<tr>
<td>T D D D D D D D</td>
<td>T T T T T T T</td>
</tr>
<tr>
<td>Published findings (n = 94)</td>
<td>New predictions (n = 26)</td>
</tr>
<tr>
<td>1st dx Before dx 1st dx</td>
<td></td>
</tr>
<tr>
<td>Risk factors Diagnostic tests Consequences</td>
<td></td>
</tr>
</tbody>
</table>

\( q \leq 0.01 \)
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: AGP HD1, CHR N A3, 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 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).

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 \( 3.36 \times 10^{-11} \) 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 \((q < 0.001; \text{table S2})\). 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 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 \( P = 1.04 \times 10^{-48} \) 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 \( 9.48 \times 10^{-12} \) for CUMC and MSMC, respectively; Fig. 4B).

---

**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.
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 \[\text{mean age, } 60 \pm 12 \text{(range, 21 to 101) at SHC and } 69 \pm 10 \text{(range, 46 to 99) at MSMC.}\] 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 \[\text{OR, } 2.08 \text{ (95\% CI, 1.36 to 3.18), with } P = 5 \times 10^{-5}\text{ at SHC; OR, 2.33 \text{ (95\% CI, 1.58 to 3.44), with } P = 1.87 \times 10^{-5}\text{ at MSMC; Table 1.}\] 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 \( (q = 0.008; \text{table S2})\), 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 \[\text{OR, 1.91 \text{ (95\% CI, 1.36 to 3.18), with } P = 1.67 \times 10^{-133}\text{ at MSMC; OR, 1.30 \text{ (95\% CI, 1.16 to 1.45), with } P = 3.97 \times 10^{-6}\text{ at CUMC; Table 1.}\] 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 \(7.33 \times 10^{-55}\) for CUMC and MSMC, respectively; Fig. 4D).

The fifth and final validation was to test the relation between PLT and TERT \( (q = 0.007; \text{table S2})\), finding that patients at CUMC and MSMC with an abnormal TERT were significantly more likely to have abnormal PLT \[\text{adjusted OR, 1.30 \text{ (95\% CI, 1.16 to 1.45), with } P = 2.73 \times 10^{-15}\text{ at MSMC}, MGN and GCA [adjusted OR, 1.44 \text{ (95\% CI, 1.21 to 1.72)}, with } P = 5.03 \times 10^{-3}\text{ at MSMC; adjusted OR, 1.63 \text{ (95\% CI, 1.29 to 2.07)}, with } P = 4.02 \times 10^{-5}\text{ at CUMC}, ALP and VTE [adjusted OR, 1.80 \text{ (95\% CI, 1.71 to 1.90)}, with } P < 2 \times 10^{-16}\text{ at MSMC; adjusted OR, 1.3 \text{ (95\% CI, 1.17 to 1.46), with } P = 2.84 \times 10^{-7}\text{ at CUMC}, and PLT and ADS \text{[adjusted OR, 1.95 \text{ (95\% CI, 1.76 to 2.16)}, with } P < 2 \times 10^{-16}\text{ at SHC; adjusted OR, 1.78 \text{ (95\% CI, 1.69 to 1.89)}, with } P < 2 \times 10^{-16}\text{ at MSMC; adjusted OR, 1.25 \text{ (95\% CI, 1.08 to 1.44), with } P = 0.0025\text{ at CUMC.}\]

Only PSA and LCA did not reach significance after age matching \[\text{[adjusted OR, 1.48 \text{ (95\% CI, 0.99 to 2.23)}, with } P = 0.058\text{ at MSMC; adjusted OR, 1.3 \text{ (95\% CI, 0.83 to 2.03), with } P = 0.25\text{ at SHC, which may be due to insufficient sample size or a possible confounding in the underlying original association with PSA and prostate cancer (PCA).}\]

Table 1. Summary of clinical validation through EMR from three independent medical centers.

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

*High, high versus normal laboratory value; High + low, high and low versus normal laboratory values. †χ² 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 \( (P = 1.02 \times 10^{-43} \text{ at SHC, } P = 7.01 \times 10^{-39} \text{ at MSMC, and } P = 6.02 \times 10^{-308} \text{ at CUMC; fig. S2A}).\)

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 founders 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 ESRI 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 ALLIN (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/trait, whereas CRI1 is associated only with two diseases/trait (table S2). The detailed TF-IDF (46) calculation procedure for all 5865 combinations of disease-trait pairs \((69 \times 85)\) with 8913 genes is described as follows. First, we calculated a TF using \( \text{TF}_{i,j} = \frac{n_{i,j}}{\sum \text{n}_{i,j}} \), where \( n_{i,j} \) is the number of occurrences of gene \( i \) in a particular disease or trait \( j \). \( \sum \text{n}_{i,j} \) indicates the total number of occurrences of all genes in a particular disease or trait \( j \). The value of \( \text{TF}_{i,j} \) indicates the level of occurrence frequency of gene \( i \) in disease or trait \( j \). Next, we calculated IDF using \( \text{IDF}_{i} = \log_{10} \frac{\text{n}_{\text{genes}}}{\text{D}} \). Here, \( D \) is the total number of diseases and traits, and \( D \) is the number of disease and trait containing gene \( i \). A larger \( \text{IDF}_{i} \) 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 \( \text{TF-IDF}_{i,j} = \text{TF}_{i,j} \times \text{IDF}_{i} \) 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 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).

\[
\text{Cosine similarity} (D,T) = \frac{D \cdot T}{\|D\| \|T\|} = \frac{\sum_{j=1}^{n} D_i \times T_i}{\sqrt{\sum_{j=1}^{n} (D_j)^2} \times \sqrt{\sum_{j=1}^{n} (T_j)^2}}
\]

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 \( \leq 0.01 \) 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 AIDS, 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 $\chi^2$ tests, laboratory values were discretized. Values outside the reference range were defined as being in the “abnormal range.” Values less than the low reference were 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 $\chi^2$ 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.

**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).


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

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