Title

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

Journal
Science Translational Medicine, 6(234)

ISSN
1946-6234

Authors
Li, Li
Ruau, David J
Patel, Chirag J
et al.

Publication Date
2014-04-30

DOI
10.1126/scitranslmed.3007191

Published in final edited form as:
Sci Transl Med. 2014 April 30; 6(234): 234ra57. doi:10.1126/scitranslmed.3007191.

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], and **Atul J. Butte**[1,2,*]

[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, 1265 Welch Road, Stanford, CA 94305, USA

[4]Stanford Center for Clinical Informatics, Stanford University School of Medicine, 1265 Welch Road, 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, & Department of Medicine, Columbia University, 622 West 168th St. 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

Abstract

Genome-Wide Association Studies (GWAS) have identified genetic variants for thousands of diseases and traits. In this study, 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-SNP association database (VARIMED), analyzing the findings from 8,962 published association studies. Similarity between traits and diseases was statistically evaluated based on 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 (EMR) from 3 independent medical centers for evidence of the trait appearing in patients within one year of first diagnosis of the disease. We validated that 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

[*]Correspondence to: Atul J. Butte, M.D., Ph.D., Department of Pediatrics, Stanford University School of Medicine, 1265 Welch Road MSOB X163, Stanford, CA, 94305-5415, abutte@stanford.edu, Phone: +1-650-723-3465, Fax: +1-650-723-7070.
[†]Authors contributed equally

Author Contributions: Conceived and designed the study: LL AJB. Performed the experiments: LL. Analyzed the data: LL DJR. Contributed reagents/material/analysis tools: LL DJR CJP SCW RC JTD NPT AJB. Wrote the paper: LL DJR CJP AJB

Competing interests: Dr. Butte is a founder and consultant of Personalis, Inc., a genetic testing company. Rong Chen is an employee of Personalis, Inc. The rest of authors declare that they have no competing interests.

C12orf51. Alkaline phosphatase level is elevated in patients with venous thromboembolism; both share variants in *ABO*. Similarly, we found prostate specific antigen and serum magnesium levels were altered before the diagnosis of lung cancer and gastric cancer, respectively. Disease-trait associations identifies traits that can potentially serve a prognostic function clinically; validating disease-trait associations through EMR can whether these candidates are risk factors for complex diseases.

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 (e.g., cholesterol levels), social behavior (e.g., smoking), patient characteristics (e.g., 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 non-disease 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 over 100 features of association studies from 8,962 human genetics papers covering 2,376 diseases and traits. VARIMED has been 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/disease), and between

796 unique genes and 85 traits (median = 10/trait). In each case, there were at least 3 significant genes per disease or trait, and the p-value was $< 1\times10^{-8}$ at the genome-wide significance level from individual GWAS (Table S1A and S1B). 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 (Figure 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 were unbiased and comparable.

Diseases and traits 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 one thousand times. We identified 120 disease-trait pairs significant at a q-value ≤ 0.01 based on the pair-wise cosine distance calculation (see 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 prior 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) (Figure S1B).

Genetic commonality between diseases and traits

We generated a comprehensive network for visualizing all 120 disease-trait pairs (Figure 2, Table S2). Diseases (blue circles) and traits (orange triangles) were connected to each other by edges when there was a significant association at q ≤ 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 (Figure 2, module D1) was connected with prostate-specific antigen levels (PSA), as this trait and these diseases were significantly associated through *TERT*. The skin cancer module (Figure 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 MHC class I/II or MHC class related molecules. Finally, type 2 diabetes related syndromes (Figure 2, module D3) were connected with proinsulin levels. Most of these connections were through *ARAP1, MADD*, or *TCF7L2* (Table S2).

The remaining 4 disease modules (Figure 2, D4-5, D7-8) exhibited multiple-to-multiple relationships underlying unexpected shared genetic commonality. One module (Figure 2, module D4) connected esophageal cancer and alcohol dependence with cholesterol levels through *ALDH2, BRAP*, and *C12orf51*, while another (Figure 2, module D5) connected Kawasaki disease and chronic obstructive pulmonary disease with smoking through *RAB4B*.

We identified seven trait modules (Figure 2, T1-7, orange circles). Three modules had known associations: pigmentary characteristics (Figure 2, T1) with skin cancer (D2) through *MC1R or SLC45A2*, and a subset (freckles and eye colors) with chronic lymphocytic leukemia (CLL) through *IRF4*. Coagulation factor activity tests (Figure 2, T4) were connected with venous thromboembolism. Three were related through *ABO* (Table S2). Lipid panel (Figure 2, T5) was connected through *APOC1, APOE, PVRL2* and *TOMM40* to Alzheimer's disease, through *CELSR2, LDLR, PSRC1*, and *ZNF259* to coronary artery disease, and through *ZNF259* to metabolic syndrome.

Detecting traits known to be associated with diseases

Ninety-four out of the 120 significant disease-trait associations were known findings supported by published studies (Table S2), these disease-trait association could be classified into one of three types, based on the temporal relationship between the trait and disease pathogenesis: 1) risk factors, for which traits manifest prior to disease onset and may cause the disease, 2) diagnostic tests, for which traits manifest contemporaneously with disease onset, and 3) consequences or complications, for which traits manifest after the disease diagnosis (Figure 3, Table S2). We manually categorized each known finding into one of these 3 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 (Figure 3) linked smoking and chronic obstructive pulmonary disease (COPD; q<0.001). Three genes containing variants were shared between smoking and COPD: *AGPHD1, CHRNA3*, and *RAB4B* (Figure 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 Figure 3), and might actually reflect variants related to smoking (i.e. propensity to addiction, non-cessation, 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 Figure 2). Even though the GWAS (24) did not explicitly enroll participants with these autoimmune diseases, our method inferred known relationships between clinical measurements, such as auto-antibody tests, and autoimmune diseases on the basis of their shared genetic architecture (Figure 3).

Last, among the 28 known pairs reflecting comorbidity or consequence (Table S2), alcohol dependence syndrome (ADS) was associated with 3 traits: cholesterol levels through shared variants in *ALDH2*, *BRAP*, and *C12orf51*; alanine aminotransferase levels (ALT) through shared variants in *C12orf51*; and HDL cholesterol levels (HDL-C) 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 that might have been influenced by alcohol, while 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 increase daily alcohol intake in individuals with ADS (30).

Clinical validation of previously undescribed disease-trait pairs with EMR

To evaluate our new associations between traits and diseases, we obtained EMR data, as it represented a patient cohort independent from our curated GWAS studies. We obtained deidentified EMR data from 3 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 five 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 non-related 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 one 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 [18-91] at SHC and 48 ± 19 [18-102] 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 one year, compared to patients with normal MCV (Odds Ratio (OR): 3.31 [2.84-3.87] with $p = 3.79\times10^{-57}$ at SHC; OR: 2.4 [1.91-3] with $p = 9.16\times10^{-15}$ at MSMC, Table 1). Besides the increase in cases, the MCV values themselves were significantly higher in cases compared to controls ($p = 1.32\times10^{-48}$ and 3.36×10^{-11} for SHC and MSMC respectively, Figure 4A).

Our second new finding was that serum magnesium level (MGN) 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 one 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 [18-90] at CUMC and 66 ± 15 [18-99] at MSMC; 41% and 52% female in CUMC and MSMC). 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 one year, compared to patients with normal MCV (OR: 1.59 [1.26-2.01] with $p = 1.04\times10^{-4}$ at CUMC; OR: 1.54 [1.29-1.84] with $p = 1.45\times10^{-6}$ 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 ($p = 4.81\times10^{-10}$ and 9.48×10^{-5} for CUMC and MSMC respectively, Figure 4B).

Our third validation related prostate specific antigen 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 one year before our first recorded diagnosis of LCA (mean age 60 ± 12 [21-101] at SHC and 69 ± 10 [46-99] at MSMC). Controls individuals at SHC and MSMC had at least one PSA measurement and no diagnosis of LCA. Patients with an abnormal high PSA were significantly more likely to develop LCA within one year compared to patients with normal PSA (OR: 2.08 [1.36-3.18] with $p = 5\times10^{-4}$ at SHC; OR: 2.33 [1.58-3.44] with $p = 1.87\times10^{-5}$ 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, Figure 4C).

We similarly validated our fourth finding, alkaline phosphatase level (ALP) related to venous thromboembolism (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 one year compared to patients with normal ALP (OR: 1.91 [1.81-2.01] with $p = 1.67\times10^{-133}$ at MSMC; OR: 1.30 [1.16-1.45] with $p = 3.97\times10^{-6}$ 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 ($p = 4.48\times10^{-252}$ and 7.33×10^{-55} for CUMC and MSMC respectively, Figure 4D).

The fifth and final validation was to test the relation between platelet counts (PLT) and alcohol dependence syndrome (ADS), linked through *C12orf51* (q=0.007; Table S2). Patients were selected at all three centers if they had a PLT measurement within one 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 one year compared to patients with normal PLT (OR: 2.12 [1.92-2.35] with $p = 1.24\times10^{-52}$ at SHC; OR: 1.84 [1.74-1.95] with $p=1.42\times10^{-109}$ at MSMC; OR: 1.25 [1.09-1.45] with p=0.0016 at CUMC, Table 1). PLT values were consistently lower in ADS patients versus controls within one year before our first ADS diagnosis ($p = 4.37\times10^{-32}$ at SHC, $p = 2.47\times10^{-43}$ at MSMC, and $p = 2.67\times10^{-6}$ at CUMC, Figure 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 5 tests. We discovered that significant associations still persisted for MCV and ALL (adjusted OR 3.5: [3.02-4.14] with $p < 2\times10^{-16}$ at SHC; adjusted OR: 2.49 [1.99-3.13] with $p = 2.73\times10^{-15}$ at MSMC), MGN and GCA (adjusted OR: 1.44 [1.21-1.72] with $p = 5.03\times10^{-5}$ at MSMC; adjusted OR: 1.63 [1.29-2.07] with $p = 4.02\times10^{-5}$ at CUMC), ALP and VTE (adjusted OR: 1.80 [1.71-1.90] with $p < 2\times10^{-16}$ at MSMC; adjusted OR: 1.3 [1.17-1.46] with $p = 2.84\times10^{-6}$ at CUMC), and PLT and ADS (adjusted OR 1.95 [1.76-2.16] with $p < 2\times10^{-16}$ at SHC, adjusted OR 1.78 [1.69-1.89] with $p < 2\times10^{-16}$ at MSMC; adjusted OR 1.25 [1.08-1.44] with p =0.0025 at CUMC). Only PSA and LCA did not reach significance after age matching (adjusted OR: 1.48 [0.99-2.23] with p = 0.058 at MSMC, adjusted OR: 1.3 [0.83-2.03] with p = 0.25 at SHC), which may due to insufficient sample size or a possible confounding in the underlying original association with PSA and prostate cancer.

To evaluate our data resource in validating our findings, we selected one well-known association as a positive control (PSA levels and prostate cancer [PCA]) from all three centers. We obtained 595, 1,231, and 4,253 PCA male patient samples with PSA results (mean age 70 ± 10 [44-96] at SHC; 70 ± 11 [34-98] at MSMC; and mean age 58 ± 13 [18-90] 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 one year before the first PCA diagnosis (OR: 10.96 [9.25-12.98] with $p = 4.43\times10^{-248}$ at SHC; OR: 7.51 [6.67-8.46] with $p = 2\times10^{-316}$ at MSMC; OR: 9.45 [8.83-10.11] with $p = 1.02\times10^{-300}$ at CUMC Table 1). Additionally, PSA values were higher in PCA patients compared to controls within one year before diagnosis ($p = 1.02\times10^{-83}$ at SHC, $p = 7.01\times10^{-69}$ at MSMC, and $p = 6.02\times10^{-308}$ at CUMC, Figure S2A).

We also tested two non-related 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 lab values and disease (Figure S2B and 2C, 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 prior evidence in the literature, while 26 disease-trait pairs were newly described. We showed that these predicted relationships can be tested medical-center electronic medical records, when sufficient numbers of patients have data with assessments of both the trait and disease. We validated the relationships for 5 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 3 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 towards 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 chronic obstructive pulmonary disease (COPD), as well as alanine aminotransferase levels and alcohol dependence. For instance, as 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 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, alcohol dependence syndrome (ADS) is a risk factor to HDL-C, which is a known risk factor to coronary artery disease (CAD) (32) and *C12orf51* was shared among them; therefore, *C12orf51* variants associated with CAD could be confounded by ADS. Similarly, metabolite levels, such as magnesium 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 odds ratios (1.1-1.5) 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 (e.g., HDL-C and cardiovascular disease) by using Mendelian randomization (36) (37) (38). Mendelian randomization is a method of using measured variation in genes of known function to examine the causal effect of a modifiable (non-genetic) exposure on disease in non-experimental 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 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 prior to their first observed diagnosis at each medical center. (The median onset age were correlated with known average ages of onset of each disease, suggesting the majority of these patients did not receive care for any significant period of time elsewhere before presenting to a hospital setting.) ICD-9 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 codes 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 co-morbidities. We did not control for these effects, as there is no well-documented list of potential confounders for every laboratory measurement; however, we assumed 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 non-causal 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.

In this study, we had also desired to evaluate the rest of the predicted disease-trait pairs. For instance, prostate specific antigen (PSA) was associated with testicular cancer (TCA), through *CLPTM1L* and *TERT* (q<0.001; Table S2). However, as the disease incidences were low at all three center (only 22 at SHC, 33 at MSMC, and 65 patients at CUMC had PSA labs measured prior to first diagnosis for TCA), we did not have sufficient power to perform such analysis. Another finding was bone mineral density (BMD) related to sudden cardiac arrest (SCA) through a gene *ESR1*. 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, prognostics, and diagnostic markers for complex diseases, although risk factors need to be better considered or controlled in GWAS design to identify independent variants without the confounding of behavioral, environmental or informative of disease pathophysiology.

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 8,962 human genetics papers including GWAS and candidate gene studies, with 87,553 SNPs mapped to 8,913 genes and 1,119 diseases and 1,256 traits. We considered only diseases and traits whose genetic variants had genome-wide significance ($p < 1\times10^{-8}$) (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 NCBI 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 1,439 genes. Distributions for the number of genes associated with diseases and traits were evaluated with Kolmogorov-Smirnov test (Figure 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 Term Frequency–Inverse Document Frequency (TF-IDF) weighing method (46) to down-weight the ubiquitous genes which are associated with many diseases. For instance, *LPL* is associated with 7 diseases/traits while *CR1* is associated only with 2 disease/traits (Table S2). The detailed Term Frequency-Inverse Document Frequency (TF-IDF) (46) calculation procedure for all 5,865 combinations of disease-trait pairs (69 × 85) with 8,913 genes is described as follows. First, we calculated a *term frequency* (TF) using $tf_{(i,j)}=\frac{n_{i,j}}{\sum_k n_{k,j}}$, where $n_{i,j}$ is the number of occurrences of gene i in a particular disease or trait j. $\Sigma_k n_{k,j}$

indicates the total number of occurrences of all genes in a particular disease or trait *j*. The value of $tf_{(i,\ j)}$ indicates the level of occurrence frequency of gene *i* in disease or trait *j*. Next, we calculated *inverse document frequency* (IDF) using $idf_{(i)}=\log_{10}(\frac{D}{D_i})$. 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 $idf_{(i)}$ implies a lower popularity of gene *i* among the diseases or traits, translating into more weight as it might only be shared between this two phenotypes among 8,913 genes. Last, we calculated a TF-IDF score using $tf-idf_{(i,j)}=tf_{(i,j)}\times idf_{(i)}$ 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 $tf\text{-}idf_{(i,\ j)}$ 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).

$$\mathrm{cos}ine-similarity\left(D,T\right)=\frac{D\bullet T}{\|D\|\|T\|}=\frac{\sum_{i=1}^{n}D_i\times T_i}{\sqrt{\sum_{i=1}^{n}\left(D_i\right)^2}\times\sqrt{\sum_{i=1}^{n}\left(T_i\right)^2}}$$

Next, to evaluate the significance of a disease-trait distance score, we randomly shuffled the genes across all the traits and re-computed the disease-trait distance. We repeated the randomization procedure 1,000 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 ≤ 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 vs. new discovered disease-trait pairs were evaluated with Kolmogorov-Smirnov test (Figure 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) plugin 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 ≤ 0.01; edge formation was not based on Cytoscape or CyOog.

Utilizing EMR from three independent medical center database systems

We used adult patient EMR data from three medical centers after 1/1/2005 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 alcohol dependence syndrome (ADS), 151 for gastric cancer (GCA), 186 for testicular cancer (TCA), 162 for lung cancer (LCA), 453 for venous thromboembolism (VTE), and 185 for prostate cancer (PCA). The control group for each analysis was taken from the adult patients without the diagnosis of target disease. Reference ranges for lab tests were based on

MedlinePlus from the National Library of Medicine. They were as follows: serum/plasma platelet count (PLT): 150-400 k/uL, serum/plasma magnesium (MGN): 1.8-2.4 mg/dL, mean corpuscular volume (MCV): 82-98 fL, Alkaline phosphatase (ALP): 44-147 IU/L, and prostate specific antigen (PSA): < 4 ng/mL.

Validation of newly describe disease-trait pairs with EMR data

Use of EMR data was approved by individual's Institutional Review Board. To perform chi-square tests, lab values were discretized. Values outside the reference range were defined as being in the “abnormal range”. Values less than 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 lab values to the reference range if multiple tests had been performed on a patient during the analysis time frame, which was defined as one year before disease diagnosis. Patients were defined as normal if lab results were within reference ranges, and abnormal if they were high or low range. Patients were excluded if multiple lab values were both high range and low range.

We performed Wilcoxon sum-rank test by evaluating the actual lab values and chi-square tests by calculating the odds ratios for abnormal ranges versus normal reference range between case and control groups. We report the odds ratios along with 95th percentile confidence intervals and p-value. We compared the percentage of abnormal results for case and control patients one year prior to our first diagnosis code of the target disease in case patients, and in control patients who were cared for at SCH, MSMC, and CUMC and without diagnosis of target disease. This allowed us to investigate whether changes in lab 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 Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We are grateful of Dr. Michael Snyder in the Genetics Department and Dr. Jorge Caballero in the Department of Anesthesia, Stanford School of Medicine, for reviewing this manuscript and providing helpful comments and suggestions. We thank Alex 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, National Institutes of Health, 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.

References and Notes

1. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. Jun 7.2007 447:661. [PubMed: 17554300]

2. Johnson AD, O'Donnell CJ. An open access database of genome-wide association results. BMC medical genetics. 2009; 10:6. [PubMed: 19161620]

3. Steinbrecher UP, Lougheed M. Scavenger receptor-independent stimulation of cholesterol esterification in macrophages by low density lipoprotein extracted from human aortic intima. Arteriosclerosis and thrombosis : a journal of vascular biology / American Heart Association. May. 1992 12:608. [PubMed: 1576122]

4. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings of the National Academy of Sciences of the United States of America. Jun 9.2009 106:9362. [PubMed: 19474294]

5. Li H, Lee Y, Chen JL, Rebman E, Li J, Lussier YA. Complex-disease networks of trait-associated single-nucleotide polymorphisms (SNPs) unveiled by information theory. Journal of the American Medical Informatics Association : JAMIA. Mar-Apr;2012 19:295. [PubMed: 22278381]

6. Sirota M, Schaub MA, Batzoglou S, Robinson WH, Butte AJ. Autoimmune disease classification by inverse association with SNP alleles. PLoS genetics. Dec.2009 5:e1000792. [PubMed: 20041220]

7. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabasi AL. The human disease network. Proceedings of the National Academy of Sciences of the United States of America. May 22.2007 104:8685. [PubMed: 17502601]

8. Li L, Ruau D, Chen R, Weber S, Butte A. SYSTEMATIC IDENTIFICATION OF RISK FACTORS FOR ALZHEIMER'S DISEASE THROUGH SHARED GENETIC ARCHITECTURE AND ELECTRONIC MEDICAL RECORDS. Pac Symp Biocomput. 2013; 18:224. [PubMed: 23424127]

9. Ashley EA, Butte AJ, Wheeler MT, Chen R, Klein TE, Dewey FE, Dudley JT, Ormond KE, Pavlovic A, Morgan AA, Pushkarev D, Neff NF, Hudgins L, Gong L, Hodges LM, Berlin DS, Thorn CF, Sangkuhl K, Hebert JM, Woon M, Sagreiya H, Whaley R, Knowles JW, Chou MF, Thakuria JV, Rosenbaum AM, Zaranek AW, Church GM, Greely HT, Quake SR, Altman RB. Clinical assessment incorporating a personal genome. Lancet. May 1.2010 375:1525. [PubMed: 20435227]

10. Chen R, Corona E, Sikora M, Dudley JT, Morgan AA, Moreno-Estrada A, Nilsen GB, Ruau D, Lincoln SE, Bustamante CD, Butte AJ. Type 2 diabetes risk alleles demonstrate extreme directional differentiation among human populations, compared to other diseases. PLoS genetics. Apr.2012 8:e1002621. [PubMed: 22511877]

11. Chen R, Davydov EV, Sirota M, Butte AJ. Non-synonymous and synonymous coding SNPs show similar likelihood and effect size of human disease association. PloS one. 2010; 5:e13574. [PubMed: 21042586]

12. Patel CJ, Chen R, Butte AJ. Data-driven integration of epidemiological and toxicological data to select candidate interacting genes and environmental factors in association with disease. Bioinformatics. Jun 15.2012 28:i121. [PubMed: 22689751]

13. Suthram S, Dudley JT, Chiang AP, Chen R, Hastie TJ, Butte AJ. Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets. PLoS computational biology. Feb.2010 6:e1000662. [PubMed: 20140234]

14. Dewey FE, Chen R, Cordero SP, Ormond KE, Caleshu C, Karczewski KJ, Whirl-Carrillo M, Wheeler MT, Dudley JT, Byrnes JK, Cornejo OE, Knowles JW, Woon M, Sangkuhl K, Gong L, Thorn CF, Hebert JM, Capriotti E, David SP, Pavlovic A, West A, Thakuria JV, Ball MP, Zaranek AW, Rehm HL, Church GM, West JS, Bustamante CD, Snyder M, Altman RB, Klein TE, Butte AJ, Ashley EA. Phased whole-genome genetic risk in a family quartet using a major allele reference sequence. PLoS genetics. Sep.2011 7:e1002280. [PubMed: 21935354]

15. Cho MH, Castaldi PJ, Wan ES, Siedlinski M, Hersh CP, Demeo DL, Himes BE, Sylvia JS, Klanderman BJ, Ziniti JP, Lange C, Litonjua AA, Sparrow D, Regan EA, Make BJ, Hokanson JE, Murray T, Hetmanski JB, Pillai SG, Kong X, Anderson WH, Tal-Singer R, Lomas DA, Coxson HO, Edwards LD, MacNee W, Vestbo J, Yates JC, Agusti A, Calverley PM, Celli B, Crim C, Rennard S, Wouters E, Bakke P, Gulsvik A, Crapo JD, Beaty TH, Silverman EK. A genome-wide association study of COPD identifies a susceptibility locus on chromosome 19q13. Human molecular genetics. Feb 15.2012 21:947. [PubMed: 22080838]

16. Pillai SG, Kong X, Edwards LD, Cho MH, Anderson WH, Coxson HO, Lomas DA, Silverman EK. Loci identified by genome-wide association studies influence different disease-related

phenotypes in chronic obstructive pulmonary disease. American journal of respiratory and critical care medicine. Dec 15.2010 182:1498. [PubMed: 20656943]

17. Wang J, Spitz MR, Amos CI, Wilkinson AV, Wu X, Shete S. Mediating effects of smoking and chronic obstructive pulmonary disease on the relation between the CHRNA5-A3 genetic locus and lung cancer risk. Cancer. Jul 15.2010 116:3458. [PubMed: 20564069]

18. Cho MH, Boutaoui N, Klanderman BJ, Sylvia JS, Ziniti JP, Hersh CP, DeMeo DL, Hunninghake GM, Litonjua AA, Sparrow D, Lange C, Won S, Murphy JR, Beaty TH, Regan EA, Make BJ, Hokanson JE, Crapo JD, Kong X, Anderson WH, Tal-Singer R, Lomas DA, Bakke P, Gulsvik A, Pillai SG, Silverman EK. Variants in FAM13A are associated with chronic obstructive pulmonary disease. Nature genetics. Mar.2010 42:200. [PubMed: 20173748]

19. Lambrechts D, Buysschaert I, Zanen P, Coolen J, Lays N, Cuppens H, Groen HJ, Dewever W, van Klaveren RJ, Verschakelen J, Wijmenga C, Postma DS, Decramer M, Janssens W. The 15q24/25 susceptibility variant for lung cancer and chronic obstructive pulmonary disease is associated with emphysema. American journal of respiratory and critical care medicine. Mar 1.2010 181:486. [PubMed: 20007924]

20. Pillai SG, Ge D, Zhu G, Kong X, Shianna KV, Need AC, Feng S, Hersh CP, Bakke P, Gulsvik A, Ruppert A, Lodrup Carlsen KC, Roses A, Anderson W, Rennard SI, Lomas DA, Silverman EK, Goldstein DB. A genome-wide association study in chronic obstructive pulmonary disease (COPD): identification of two major susceptibility loci. PLoS genetics. Mar.2009 5:e1000421. [PubMed: 19300482]

21. Vestbo J, Hurd SS, Agusti AG, Jones PW, Vogelmeier C, Anzueto A, Barnes PJ, Fabbri LM, Martinez FJ, Nishimura M, Stockley RA, Sin DD, Rodriguez-Roisin R. Global Strategy for the Diagnosis, Management and Prevention of Chronic Obstructive Pulmonary Disease, GOLD Executive Summary. American journal of respiratory and critical care medicine. Aug 9.2012

22. The Health Consequences of Smoking: A Report of the Surgeon General. Atlanta (GA): 2004.

23. Pauwels RA, Buist AS, Calverley PM, Jenkins CR, Hurd SS. 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. American journal of respiratory and critical care medicine. Apr.2001 163:1256. [PubMed: 11316667]

24. Png E, Thalamuthu A, Ong RT, Snippe H, Boland GJ, Seielstad M. A genome-wide association study of hepatitis B vaccine response in an Indonesian population reveals multiple independent risk variants in the HLA region. Human molecular genetics. Oct 1.2011 20:3893. [PubMed: 21764829]

25. Kato N, Takeuchi F, Tabara Y, Kelly TN, Go MJ, Sim X, Tay WT, Chen CH, Zhang Y, Yamamoto K, Katsuya T, Yokota M, Kim YJ, Ong RT, Nabika T, Gu D, Chang LC, Kokubo Y, Huang W, Ohnaka K, Yamori Y, Nakashima E, Jaquish CE, Lee JY, Seielstad M, Isono M, Hixson JE, Chen YT, Miki T, Zhou X, Sugiyama T, Jeon JP, Liu JJ, Takayanagi R, Kim SS, Aung T, Sung YJ, Zhang X, Wong TY, Han BG, Kobayashi S, Ogihara T, Zhu D, Iwai N, Wu JY, Teo YY, Tai ES, Cho YS, He J. Meta-analysis of genome-wide association studies identifies common variants associated with blood pressure variation in east Asians. Nature genetics. Jun.2011 43:531. [PubMed: 21572416]

26. Kim YJ, Go MJ, Hu C, Hong CB, Kim YK, Lee JY, Hwang JY, Oh JH, Kim DJ, Kim NH, Kim S, Hong EJ, Kim JH, Min H, Kim Y, Zhang R, Jia W, Okada Y, Takahashi A, Kubo M, Tanaka T, Kamatani N, Matsuda K, Park T, Oh B, Kimm K, Kang D, Shin C, Cho NH, Kim HL, Han BG, Cho YS. Large-scale genome-wide association studies in East Asians identify new genetic loci influencing metabolic traits. Nature genetics. Oct.2011 43:990. [PubMed: 21909109]

27. Kamatani Y, Matsuda K, Okada Y, Kubo M, Hosono N, Daigo Y, Nakamura Y, Kamatani N. Genome-wide association study of hematological and biochemical traits in a Japanese population. Nature genetics. Mar.2010 42:210. [PubMed: 20139978]

28. Kahl KG, Greggersen W, Schweiger U, Cordes J, Correll CU, Ristow J, Burow J, Findel C, Stoll A, Balijepalli C, Gores L, Losch C, Hillemacher T, Bleich S, Moebus S. 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. Nov.2010 105:1921. [PubMed: 20735365]

29. Gross GA. Drug and alcohol abuse and cholesterol levels. The Journal of the American Osteopathic Association. Jan.1994 94:55. [PubMed: 8169159]

30. Imhof A, Froehlich M, Brenner H, Boeing H, Pepys MB, Koenig W. Effect of alcohol consumption on systemic markers of inflammation. Lancet. Mar 10.2001 357:763. [PubMed: 11253971]

31. Lowe HJ, Ferris TA, Hernandez PM, Weber SC. STRIDE--An integrated standards-based translational research informatics platform. AMIA … Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2009; 2009:391. [PubMed: 20351886]

32. Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. Circulation. May 12.1998 97:1837. [PubMed: 9603539]

33. Zeggini E, Weedon MN, Lindgren CM, Frayling TM, Elliott KS, Lango H, Timpson NJ, Perry JR, Rayner NW, Freathy RM, Barrett JC, Shields B, Morris AP, Ellard S, Groves CJ, Harries LW, Marchini JL, Owen KR, Knight B, Cardon LR, Walker M, Hitman GA, Morris AD, Doney AS, McCarthy MI, Hattersley AT. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. Science. Jun 1.2007 316:1336. [PubMed: 17463249]

34. Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, Lindgren CM, Perry JR, Elliott KS, Lango H, Rayner NW, Shields B, Harries LW, Barrett JC, Ellard S, Groves CJ, Knight B, Patch AM, Ness AR, Ebrahim S, Lawlor DA, Ring SM, Ben-Shlomo Y, Jarvelin MR, Sovio U, Bennett AJ, Melzer D, Ferrucci L, Loos RJ, Barroso I, Wareham NJ, Karpe F, Owen KR, Cardon LR, Walker M, Hitman GA, Palmer CN, Doney AS, Morris AD, Smith GD, Hattersley AT, McCarthy MI. A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. Science. May 11.2007 316:889. [PubMed: 17434869]

35. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM. Finding the missing heritability of complex diseases. Nature. Oct 8.2009 461:747. [PubMed: 19812666]

36. Harrison SC, Holmes MV, Humphries SE. Mendelian randomisation, lipids, and cardiovascular disease. Lancet. Aug 11.2012 380:543. [PubMed: 22607824]

37. Sheehan NA, Didelez V, Burton PR, Tobin MD. Mendelian randomisation and causal inference in observational epidemiology. PLoS medicine. Aug 26.2008 5:e177. [PubMed: 18752343]

38. Smith GD, Ebrahim S. ‘Mendelian randomization’: can genetic epidemiology contribute to understanding environmental determinants of disease? International journal of epidemiology. Feb. 2003 32:1. [PubMed: 12689998]

39. Voight BF, Peloso GM, Orho-Melander M, Frikke-Schmidt R, Barbalic M, Jensen MK, Hindy G, Holm H, Ding EL, Johnson T, Schunkert H, Samani NJ, Clarke R, Hopewell JC, Thompson JF, Li M, Thorleifsson G, Newton-Cheh C, Musunuru K, Pirruccello JP, Saleheen D, Chen L, Stewart A, Schillert A, Thorsteinsdottir U, Thorgeirsson G, Anand S, Engert JC, Morgan T, Spertus J, Stoll M, Berger K, Martinelli N, Girelli D, McKeown PP, Patterson CC, Epstein SE, Devaney J, Burnett MS, Mooser V, Ripatti S, Surakka I, Nieminen MS, Sinisalo J, Lokki ML, Perola M, Havulinna A, de Faire U, Gigante B, Ingelsson E, Zeller T, Wild P, de Bakker PI, Klungel OH, Maitland-van der Zee AH, Peters BJ, de Boer A, Grobbee DE, Kamphuisen PW, Deneer VH, Elbers CC, Onland-Moret NC, Hofker MH, Wijmenga C, Verschuren WM, Boer JM, van der Schouw YT, Rasheed A, Frossard P, Demissie S, Willer C, Do R, Ordovas JM, Abecasis GR, Boehnke M, Mohlke KL, Daly MJ, Guiducci C, Burtt NP, Surti A, Gonzalez E, Purcell S, Gabriel S, Marrugat J, Peden J, Erdmann J, Diemert P, Willenborg C, Konig IR, Fischer M, Hengstenberg C, Ziegler A, Buysschaert I, Lambrechts D, Van de Werf F, Fox KA, El Mokhtari NE, Rubin D, Schrezenmeir J, Schreiber S, Schafer A, Danesh J, Blankenberg S, Roberts R, McPherson R, Watkins H, Hall AS, Overvad K, Rimm E, Boerwinkle E, Tybjaerg-Hansen A, Cupples LA, Reilly MP, Melander O, Mannucci PM, Ardissino D, Siscovick D, Elosua R, Stefansson K, O'Donnell CJ, Salomaa V, Rader DJ, Peltonen L, Schwartz SM, Altshuler D, Kathiresan S. Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study. Lancet. Aug 11.2012 380:572. [PubMed: 22607825]

40. Denny JC, Crawford DC, Ritchie MD, Bielinski SJ, Basford MA, Bradford Y, Chai HS, Bastarache L, Zuvich R, Peissig P, Carrell D, Ramirez AH, Pathak J, Wilke RA, Rasmussen L, Wang X, Pacheco JA, Kho AN, Hayes MG, Weston N, Matsumoto M, Kopp PA, Newton KM, Jarvik GP, Li R, Manolio TA, Kullo IJ, Chute CG, Chisholm RL, Larson EB, McCarty CA, Masys DR, Roden DM, de Andrade M. Variants near FOXE1 are associated with hypothyroidism and other thyroid conditions: using electronic medical records for genome- and phenome-wide studies. American journal of human genetics. Oct 7.2011 89:529. [PubMed: 21981779]

41. Mandl KD, Kohane IS. Escaping the EHR trap--the future of health IT. The New England journal of medicine. Jun 14.2012 366:2240. [PubMed: 22693995]

42. Kho AN, Pacheco JA, Peissig PL, Rasmussen L, Newton KM, Weston N, Crane PK, Pathak J, Chute CG, Bielinski SJ, Kullo IJ, Li R, Manolio TA, Chisholm RL, Denny JC. Electronic medical records for genetic research: results of the eMERGE consortium. Science translational medicine. Apr 20.2011 3:79re1.

43. Westreich D. Berkson's bias, selection bias, and missing data. Epidemiology. Jan.2012 23:159. [PubMed: 22081062]

44. Ott J, Kamatani Y, Lathrop M. Family-based designs for genome-wide association studies. Nature reviews Genetics. Jul.2011 12:465.

45. Chen R, Li L, Butte AJ. AILUN: reannotating gene expression data automatically. Nature methods. Nov.2007 4:879. [PubMed: 17971777]

46. Wu HC, Luk RWP, Wong KF, Kwok KL. Interpreting TF-IDF term weights as making relevance decisions. Acm T Inform Syst. 2008; 26

47. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences of the United States of America. Aug 5.2003 100:9440. [PubMed: 12883005]

48. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research. Nov.2003 13:2498. [PubMed: 14597658]

49. Royer L, Reimann M, Andreopoulos B, Schroeder M. Unraveling protein networks with power graph analysis. PLoS computational biology. 2008; 4:e1000108. [PubMed: 18617988]

50. Ihaka R, Gentleman R. R: A Language for Data Analysis and Graphics. Journal of computational and graphical statistics. 1996; 5:299.

51. Tobin DJ, Orentreich N, Fenton DA, Bystryn JC. Antibodies to hair follicles in alopecia areata. The Journal of investigative dermatology. May.1994 102:721. [PubMed: 8176253]

52. Caramelli P, Nitrini R, Maranhao R, Lourenco AC, Damasceno MC, Vinagre C, Caramelli B. Increased apolipoprotein B serum concentration in Alzheimer's disease. Acta neurologica Scandinavica. Jul.1999 100:61. [PubMed: 10416513]

53. O'Bryant SE, Waring SC, Hobson V, Hall JR, Moore CB, Bottiglieri T, Massman P, Diaz-Arrastia R. Decreased C-reactive protein levels in Alzheimer disease. Journal of geriatric psychiatry and neurology. Mar.2010 23:49. [PubMed: 19933496]

54. Matsuzaki T, Sasaki K, Hata J, Hirakawa Y, Fujimi K, Ninomiya T, Suzuki SO, Kanba S, Kiyohara Y, Iwaki T. Association of Alzheimer disease pathology with abnormal lipid metabolism: the Hisayama Study. Neurology. Sep 13.2011 77:1068. [PubMed: 21911734]

55. Lesser GT, Haroutunian V, Purohit DP, Schnaider Beeri M, Schmeidler J, Honkanen L, Neufeld R, Libow LS. Serum lipids are related to Alzheimer's pathology in nursing home residents. Dementia and geriatric cognitive disorders. 2009; 27:42. [PubMed: 19129700]

56. van Oijen M, van der Meer IM, Hofman A, Witteman JC, Koudstaal PJ, Breteler MM. Lipoprotein-associated phospholipase A2 is associated with risk of dementia. Annals of neurology. Jan.2006 59:139. [PubMed: 16278861]

57. Ringrose JH. HLA-B27 associated spondyloarthropathy, an autoimmune disease based on crossreactivity between bacteria and HLA-B27? Annals of the rheumatic diseases. Oct.1999 58:598. [PubMed: 10491358]

58. Bryant DH, Burns MW, Lazarus L. The correlation between skin tests, bronchial provocation tests and the serum level of IgE specific for common allergens in patients with asthma. Clinical allergy. Jun.1975 5:145. [PubMed: 1139766]

59. Chinem VP, Miot HA. Prevalence of actinic skin lesions in patients with basal cell carcinoma of the head: a case-control study. Rev Assoc Med Bras. Apr.2012 58:188. [PubMed: 22569613]

60. Zanetti R, Rosso S, Martinez C, Nieto A, Miranda A, Mercier M, Loria DI, Osterlind A, Greinert R, Navarro C, Fabbrocini G, Barbera C, Sancho-Garnier H, Gafa L, Chiarugi A, Mossotti R. Comparison of risk patterns in carcinoma and melanoma of the skin in men: a multi-centre case-case-control study. British journal of cancer. Mar 13.2006 94:743. [PubMed: 16495934]

61. Kikuchi A, Shimizu H, Nishikawa T. Clinical histopathological characteristics of basal cell carcinoma in Japanese patients. Archives of dermatology. Mar.1996 132:320. [PubMed: 8607638]

62. Musallam KM, Sankaran VG, Cappellini MD, Duca L, Nathan DG, Taher AT. Fetal hemoglobin levels and morbidity in untransfused patients with beta-thalassemia intermedia. Blood. Jan 12.2012 119:364. [PubMed: 22096240]

63. Kaplan MM, Gershwin ME. Primary biliary cirrhosis. The New England journal of medicine. Sep 22.2005 353:1261. [PubMed: 16177252]

64. Emanuele E, Carlin MV, D'Angelo A, Peros E, Barale F, Geroldi D, Politi P. Elevated plasma levels of lipoprotein(a) in psychiatric patients: a possible contribution to increased vascular risk. European psychiatry : the journal of the Association of European Psychiatrists. Mar.2006 21:129. [PubMed: 16516110]

65. Sagud M, Mihaljevic-Peles A, Pivac N, Jakovljevic M, Muck-Seler D. Lipid levels in female patients with affective disorders. Psychiatry research. Aug 15.2009 168:218. [PubMed: 19560828]

66. Villarejos VM, Serra J, Visona KA, Eduarte CE. Antibodies to single stranded DNA: a diagnostic aid in chronic hepatitis B virus infections. Journal of medical virology. 1979; 4:97. [PubMed: 314970]

67. Benn M. Apolipoprotein B levels, APOB alleles, and risk of ischemic cardiovascular disease in the general population, a review. Atherosclerosis. Sep.2009 206:17. [PubMed: 19200547]

68. Burggraf GW, Parker JO. Prognosis in coronary artery disease. Angiographic, hemodynamic, and clinical factors. Circulation. Jan.1975 51:146. [PubMed: 1109313]

69. Reilly MP, Wolfe ML, Dykhouse J, Reddy K, Localio AR, Rader DJ. Intercellular adhesion molecule 1 (ICAM-1) gene variant is associated with coronary artery calcification independent of soluble ICAM-1 levels. Journal of investigative medicine : the official publication of the American Federation for Clinical Research. Dec.2004 52:515. [PubMed: 15682683]

70. Packard CJ, O'Reilly DS, Caslake MJ, McMahon AD, Ford I, Cooney J, Macphee CH, Suckling KE, Krishna M, Wilkinson FE, Rumley A, Lowe GD. Lipoprotein-associated phospholipase A2 as an independent predictor of coronary heart disease. West of Scotland Coronary Prevention Study Group. The New England journal of medicine. Oct 19.2000 343:1148. [PubMed: 11036120]

71. Vikenes K, Farstad M, Nordrehaug JE. Serotonin is associated with coronary artery disease and cardiac events. Circulation. Aug 3.1999 100:483. [PubMed: 10430761]

72. Sako A, Kitayama J, Kaisaki S, Nagawa H. Hyperlipidemia is a risk factor for lymphatic metastasis in superficial esophageal carcinoma. Cancer letters. May 10.2004 208:43. [PubMed: 15105044]

73. Martinez-Garcia MA, Luque-Ramirez M, San-Millan JL, Escobar-Morreale HF. 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. Aug.2009 32:1525. [PubMed: 19401444]

74. Haghighi S, Amini M, Pournaghshband Z, Amini P, Hovsepian S. Relationship between gamma-glutamyl transferase and glucose intolerance in first degree relatives of type 2 diabetics patients. Journal of research in medical sciences : the official journal of Isfahan University of Medical Sciences. Feb.2011 16:123. [PubMed: 22091220]

75. Little RR, England JD, Wiedmeyer HM, McKenzie EM, Pettitt DJ, Knowler WC, Goldstein DE. Relationship of glycosylated hemoglobin to oral glucose tolerance. Implications for diabetes screening. Diabetes. Jan.1988 37:60. [PubMed: 3335278]

76. McGarry JD. Disordered metabolism in diabetes: have we underemphasized the fat component&. Journal of cellular biochemistry. 1994; 55(Suppl):29. [PubMed: 7929616]

77. Fumeron F, Pean F, Driss F, Balkau B, Tichet J, Marre M, Grandchamp B. 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. Sep. 2006 29:2090. [PubMed: 16936158]

78. Urwijitaroon Y, Barusrux S, Romphruk A, Puapairoj C, Thongkrajai P. Anti-HIV antibody titer: an alternative supplementary test for diagnosis of HIV-1 infection. Asian Pacific journal of allergy and immunology / launched by the Allergy and Immunology Society of Thailand. Dec.1997 15:193. [PubMed: 9579612]

79. Chobanian AV, Bakris GL, Black HR, Cushman WC, Green LA, Izzo JL Jr, Jones DW, Materson BJ, Oparil S, Wright JT Jr, Roccella EJ. Seventh report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure. Hypertension. Dec. 2003 42:1206. [PubMed: 14656957]

80. Coresh J, Wei GL, McQuillan G, Brancati FL, Levey AS, Jones C, Klag MJ. 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). Archives of internal medicine. May 14.2001 161:1207. [PubMed: 11343443]

81. Kim MK, Baek KH, Song KH, Kang MI, Choi JH, Bae JC, Park CY, Lee WY, Oh KW. Increased serum ferritin predicts the development of hypertension among middle-aged men. American journal of hypertension. Apr.2012 25:492. [PubMed: 22278211]

82. Cirillo M, Laurenzi M, Trevisan M, Stamler J. Hematocrit, blood pressure, and hypertension. The Gubbio Population Study. Hypertension. Sep.1992 20:319. [PubMed: 1516951]

83. Strippoli GF, Craig JC, Manno C, Schena FP. Hemoglobin targets for the anemia of chronic kidney disease: a meta-analysis of randomized, controlled trials. Journal of the American Society of Nephrology : JASN. Dec.2004 15:3154. [PubMed: 15579519]

84. Sutton-Tyrrell K, Bostom A, Selhub J, Zeigler-Johnson C. High homocysteine levels are independently related to isolated systolic hypertension in older adults. Circulation. Sep 16.1997 96:1745. [PubMed: 9323056]

85. Rahim R, Nahar K, Khan IA. Platelet count in 100 cases of pregnancy induced hypertension. Mymensingh medical journal : MMJ. Jan.2010 19:5. [PubMed: 20046164]

86. Dogan SM, Aydin M, Gursurer M, Dursun A, Mungan G, Onuk T. N-terminal probrain natriuretic peptide predicts altered circadian variation in essential hypertension. Coronary artery disease. Aug. 2007 18:347. [PubMed: 17627183]

87. Buchan DJ. Diagnosis and management of inflammatory bowel disease. Canadian family physician Medecin de famille canadien. Aug.1976 22:47. [PubMed: 21308011]

88. Pradhan AD, Manson JE, Rifai N, Buring JE, Ridker PM. C-reactive protein, interleukin 6, and risk of developing type 2 diabetes mellitus. JAMA : the journal of the American Medical Association. Jul 18.2001 286:327.

89. Mansfield MW, Heywood DM, Grant PJ. Circulating levels of factor VII, fibrinogen, and von Willebrand factor and features of insulin resistance in first-degree relatives of patients with NIDDM. Circulation. Nov 1.1996 94:2171. [PubMed: 8901668]

90. Borai A, Livingstone C, Abdelaal F, Bawazeer A, Keti V, Ferns G. The relationship between glycosylated haemoglobin (HbA1c) and measures of insulin resistance across a range of glucose tolerance. Scandinavian journal of clinical and laboratory investigation. Apr.2011 71:168. [PubMed: 21348785]

91. Laws A, Reaven GM. Evidence for an independent relationship between insulin resistance and fasting plasma HDL-cholesterol, triglyceride and insulin concentrations. Journal of internal medicine. Jan.1992 231:25. [PubMed: 1732395]

92. Nelson TL, Biggs ML, Kizer JR, Cushman M, Hokanson JE, Furberg CD, Mukamal KJ. Lipoprotein-associated phospholipase A2 (Lp-PLA2) and future risk of type 2 diabetes: results from the Cardiovascular Health Study. The Journal of clinical endocrinology and metabolism. May.2012 97:1695. [PubMed: 22399516]

93. Hanley AJ, D'Agostino R Jr, Wagenknecht LE, Saad MF, Savage PJ, Bergman R, Haffner SM. Increased proinsulin levels and decreased acute insulin response independently predict the incidence of type 2 diabetes in the insulin resistance atherosclerosis study. Diabetes. Apr.2002 51:1263. [PubMed: 11916954]

94. Bazelmans J, Nestel PJ, Nolan C. Insulin-induced glucose utilization influences triglyceride metabolism. Clin Sci (Lond). May.1983 64:511. [PubMed: 6339152]

95. Fink JC, Burdick RA, Kurth SJ, Blahut SA, Armistead NC, Turner MS, Shickle LM, Light PD. Significance of serum creatinine values in new end-stage renal disease patients. American journal of kidney diseases : the official journal of the National Kidney Foundation. Oct.1999 34:694. [PubMed: 10516351]

96. Khuder SA. Effect of cigarette smoking on major histological types of lung cancer: a meta-analysis. Lung Cancer. Feb-Mar;2001 31:139. [PubMed: 11165392]

97. Holly EA, Aston DA, Cress RD, Ahn DK, Kristiansen JJ. Cutaneous melanoma in women. II. Phenotypic characteristics and other host-related factors. American journal of epidemiology. May 15.1995 141:934. [PubMed: 7741123]

98. Tucker MA. Melanoma epidemiology. Hematology/oncology clinics of North America. Jun.2009 23:383. [PubMed: 19464592]

99. Veierod MB, Weiderpass E, Thorn M, Hansson J, Lund E, Armstrong B, Adami HO. A prospective study of pigmentation, sun exposure, and risk of cutaneous malignant melanoma in women. Journal of the National Cancer Institute. Oct 15.2003 95:1530. [PubMed: 14559875]

100. Yumura W, Suganuma S, Nitta K, Sano Y, Uchida K, Nihei H. Prolonged membranous lupus nephritis with change of anti-ssDNA antibody titer and repeated renal relapse. Clinical and experimental nephrology. Dec.2004 8:363. [PubMed: 15619038]

101. Sierra-Johnson J, Somers VK, Kuniyoshi FH, Garza CA, Isley WL, Gami AS, Lopez-Jimenez F. Comparison of apolipoprotein-B/apolipoprotein-AI in subjects with versus without the metabolic syndrome. The American journal of cardiology. Nov 15.2006 98:1369. [PubMed: 17134631]

102. Frohlich M, Imhof A, Berg G, Hutchinson WL, Pepys MB, Boeing H, Muche R, Brenner H, Koenig W. Association between C-reactive protein and features of the metabolic syndrome: a population-based study. Diabetes care. Dec.2000 23:1835. [PubMed: 11128362]

103. Gombet T, Longo-Mbenza B, Ellenga-Mbolla B, Ikama MS, Mokondjimobe E, Kimbally-Kaky G, Nkoua JL. Aging, female sex, migration, elevated HDL-C, and inflammation are associated with prevalence of metabolic syndrome among African bank employees. International journal of general medicine. 2012; 5:495. [PubMed: 22807636]

104. Gong HP, Du YM, Zhong LN, Dong ZQ, Wang X, Mao YJ, Lu QH. Plasma lipoprotein-associated phospholipase A2 in patients with metabolic syndrome and carotid atherosclerosis. Lipids in health and disease. 2011; 10:13. [PubMed: 21247435]

105. Burns E, Mulley GP. Practical problems with eye-drops among elderly ophthalmology outpatients. Age and ageing. May.1992 21:168. [PubMed: 1615776]

106. Lalive PH, Menge T, Delarasse C, Della Gaspera B, Pham-Dinh D, Villoslada P, von Budingen HC, Genain CP. Antibodies to native myelin oligodendrocyte glycoprotein are serologic markers of early inflammation in multiple sclerosis. Proceedings of the National Academy of Sciences of the United States of America. Feb 14.2006 103:2280. [PubMed: 16461459]

107. Brucato A, Cimaz R, Caporali R, Ramoni V, Buyon J. Pregnancy outcomes in patients with autoimmune diseases and anti-Ro/SSA antibodies. Clinical reviews in allergy & immunology. Feb.2011 40:27. [PubMed: 20012231]

108. Burgert TS, Taksali SE, Dziura J, Goodman TR, Yeckel CW, Papademetris X, Constable RT, Weiss R, Tamborlane WV, Savoye M, Seyal AA, Caprio S. Alanine aminotransferase levels and fatty liver in childhood obesity: associations with insulin resistance, adiponectin, and visceral fat. The Journal of clinical endocrinology and metabolism. Nov.2006 91:4287. [PubMed: 16912127]

109. Mertens IL, Van Gaal LF. Overweight, obesity, and blood pressure: the effects of modest weight reduction. Obesity research. May.2000 8:270. [PubMed: 10832771]

110. Kamoda T, Saitoh H, Inudoh M, Miyazaki K, Matsui A. The serum levels of proinsulin and their relationship with IGFBP-1 in obese children. Diabetes, obesity & metabolism. Mar.2006 8:192.

111. Nakamura M, Shimizu-Yoshida Y, Takii Y, Komori A, Yokoyama T, Ueki T, Daikoku M, Yano K, Matsumoto T, Migita K, Yatsuhashi H, Ito M, Masaki N, Adachi H, Watanabe Y, Nakamura Y, Saoshiro T, Sodeyama T, Koga M, Shimoda S, Ishibashi H. Antibody titer to gp210-C terminal peptide as a clinical parameter for monitoring primary biliary cirrhosis. Journal of hepatology. Mar.2005 42:386. [PubMed: 15710222]

112. Smith DS, Humphrey PA, Catalona WJ. The early detection of prostate carcinoma with prostate specific antigen: the Washington University experience. Cancer. Nov 1.1997 80:1852. [PubMed: 9351559]

113. Egerer K, Feist E, Burmester GR. The serological diagnosis of rheumatoid arthritis: antibodies to citrullinated antigens. Deutsches Arzteblatt international. Mar.2009 106:159. [PubMed: 19578391]

114. Avnon LS, Manzur F, Bolotin A, Heimer D, Flusser D, Buskila D, Sukenik S, Abu-Shakra M. Pulmonary functions testing in patients with rheumatoid arthritis. The Israel Medical Association journal : IMAJ. Feb.2009 11:83. [PubMed: 19432035]

115. Gill JM, Quisel AM, Rocca PV, Walters DT. Diagnosis of systemic lupus erythematosus. American family physician. Dec 1.2003 68:2179. [PubMed: 14677663]

116. Morita Y, Muro Y, Sugiura K, Tomita Y. Anti-cyclic citrullinated peptide antibody in systemic sclerosis. Clinical and experimental rheumatology. Jul-Aug;2008 26:542. [PubMed: 18799082]

117. Vaziri-Sani F, Oak S, Radtke J, Lernmark K, Lynch K, Agardh CD, Cilio CM, Lethagen AL, Ortqvist E, Landin-Olsson M, Torn C, Hampe CS. ZnT8 autoantibody titers in type 1 diabetes patients decline rapidly after clinical onset. Autoimmunity. Dec.2010 43:598. [PubMed: 20298127]

118. Ronnback M, Fagerudd J, Forsblom C, Pettersson-Fernholm K, Reunanen A, Groop PH. Altered age-related blood pressure pattern in type 1 diabetes. Circulation. Aug 31.2004 110:1076. [PubMed: 15326070]

119. Gylling H, Tuominen JA, Koivisto VA, Miettinen TA. Cholesterol metabolism in type 1 diabetes. Diabetes. Sep.2004 53:2217. [PubMed: 15331530]

120. Ma J, Mollsten A, Prazny M, Falhammar H, Brismar K, Dahlquist G, Efendic S, Gu HF. Genetic influences of the intercellular adhesion molecule 1 (ICAM-1) gene polymorphisms in development of Type 1 diabetes and diabetic nephropathy. Diabetic medicine : a journal of the British Diabetic Association. Oct.2006 23:1093. [PubMed: 16978373]

121. Li S, Shin HJ, Ding EL, van Dam RM. Adiponectin levels and risk of type 2 diabetes: a systematic review and meta-analysis. JAMA : the journal of the American Medical Association. Jul 8.2009 302:179.

122. Pradhan AD, Manson JE, Meigs JB, Rifai N, Buring JE, Liu S, Ridker PM. Insulin, proinsulin, proinsulin:insulin ratio, and the risk of developing type 2 diabetes mellitus in women. The American journal of medicine. Apr 15.2003 114:438. [PubMed: 12727576]

123. Chirinos JA, Heresi GA, Velasquez H, Jy W, Jimenez JJ, Ahn E, Horstman LL, Soriano AO, Zambrano JP, Ahn YS. Elevation of endothelial microparticles, platelets, and leukocyte activation in patients with venous thromboembolism. Journal of the American College of Cardiology. May 3.2005 45:1467. [PubMed: 15862420]

124. Shrivastava S, Ridker PM, Glynn RJ, Goldhaber SZ, Moll S, Bounameaux H, Bauer KA, Kessler CM, Cushman M. D-dimer, factor VIII coagulant activity, low-intensity warfarin and the risk of recurrent venous thromboembolism. Journal of thrombosis and haemostasis : JTH. Jun.2006 4:1208. [PubMed: 16706961]

125. Jenkins PV, Rawley O, Smith OP, O'Donnell JS. Elevated factor VIII levels and risk of venous thrombosis. British journal of haematology. Jun.2012 157:653. [PubMed: 22530883]

126. Park YK, Kim NS, Hann SK, Im S. Identification of autoantibody to melanocytes and characterization of vitiligo antigen in vitiligo patients. Journal of dermatological science. Feb. 1996 11:111. [PubMed: 8869031]

127. Wong SN, Shah V, Dillon MJ. Antineutrophil cytoplasmic antibodies in Wegener's granulomatosis. Archives of disease in childhood. Sep.1998 79:246. [PubMed: 9875021]

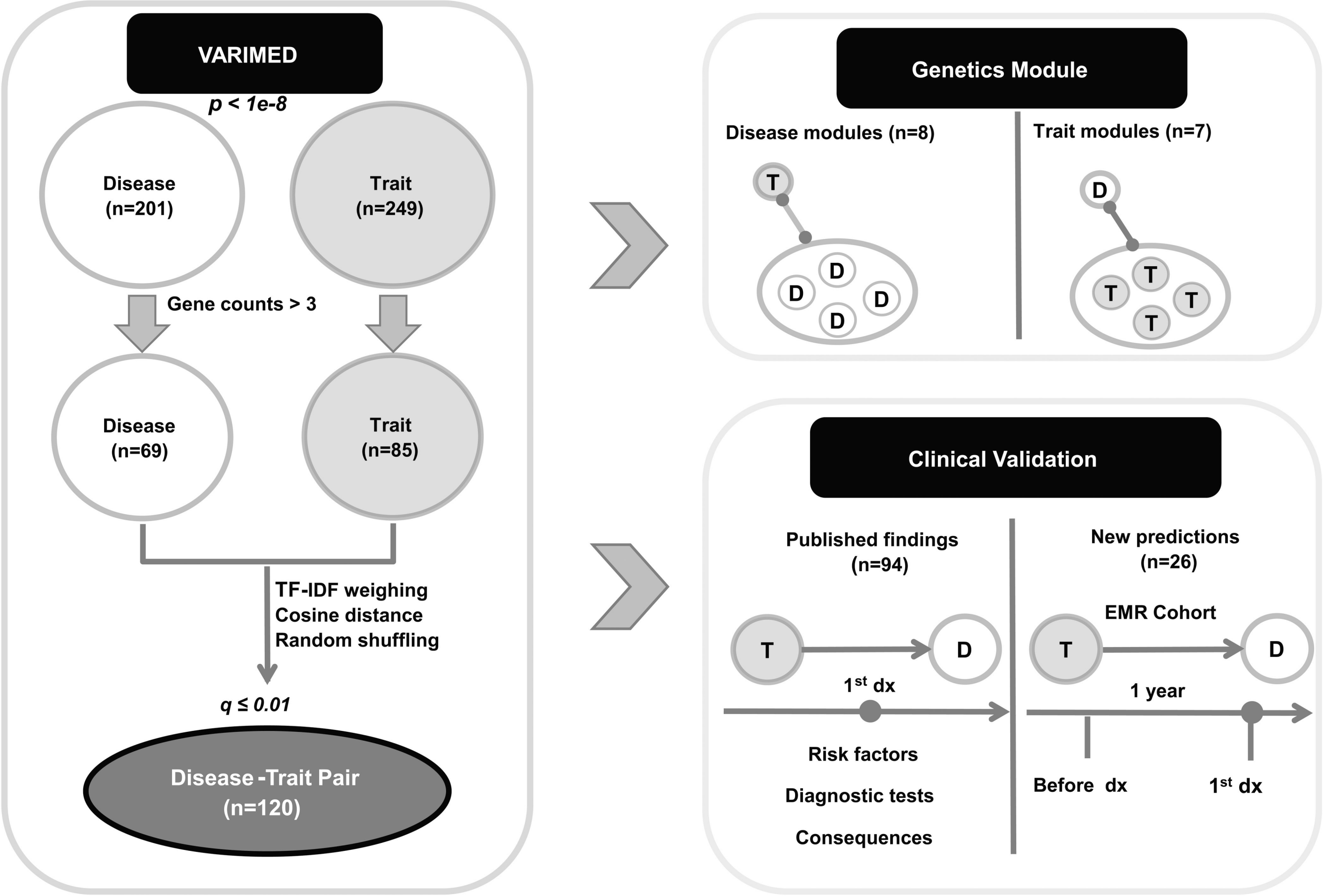

Figure 1.

Diagram for identifying significant disease-trait genetic associations.

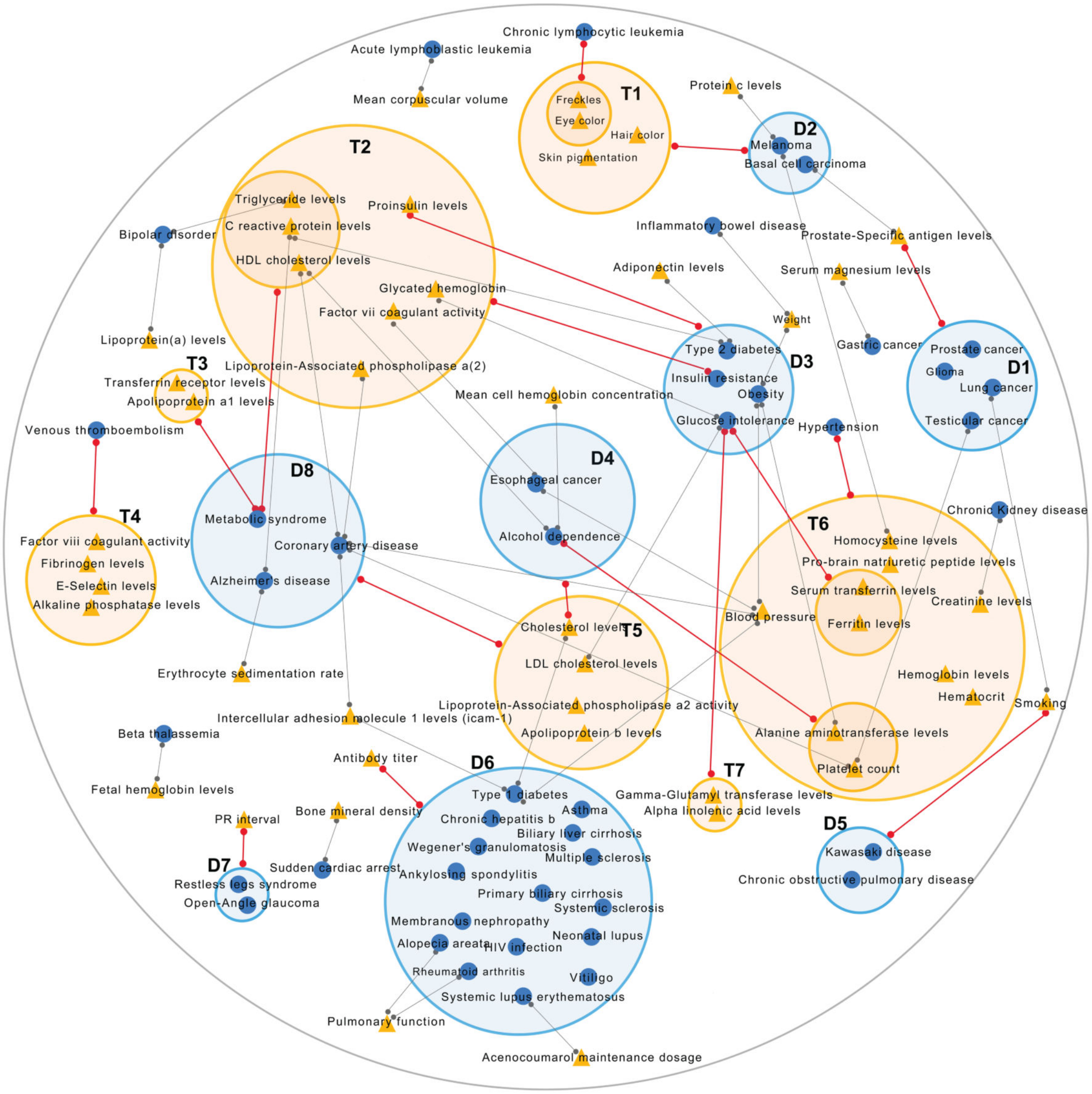

Figure 2. Disease-trait network of 120 significant pairs

The network consists of the 120 significant disease-trait pairs with q ≤ 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-T7 indicate trait modules (light orange circles) connected to a disease or disease module by red lines. D1-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) plugin.

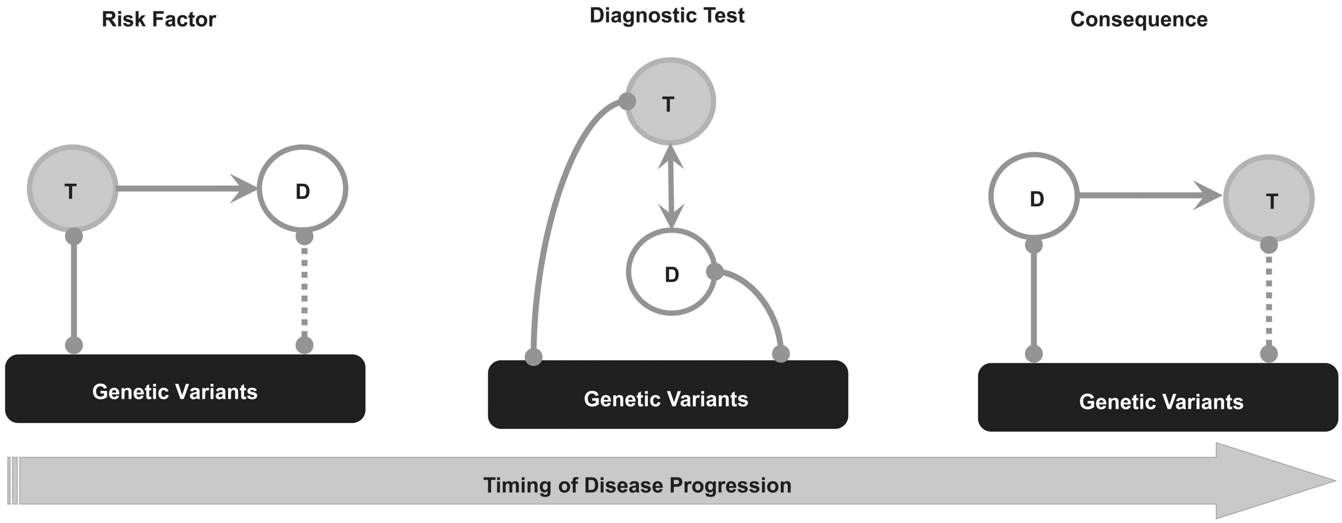

Figure 3. Three ways traits and diseases can temporally interrelate

Traits (i.e. risk factors) can manifest prior to 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.

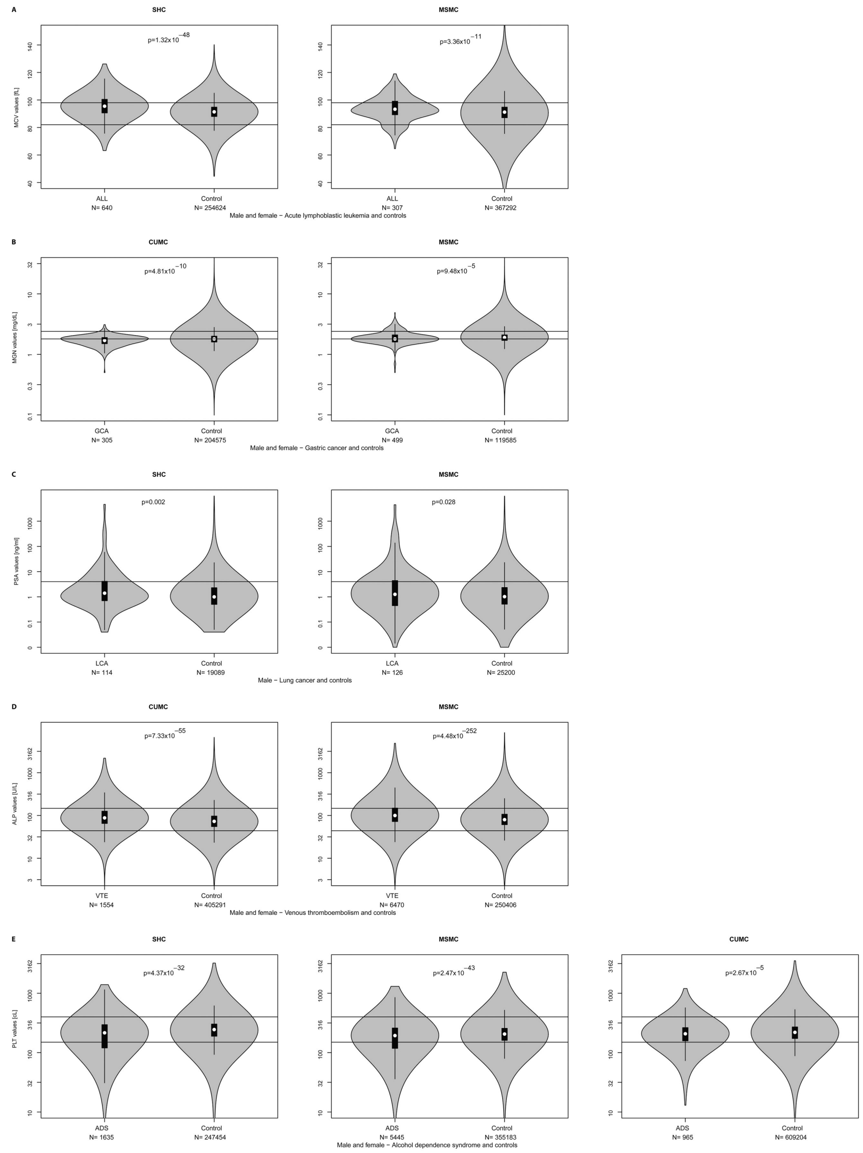

Figure 4. Violin plots for clinical validations of five new findings

Violin plots (combination of boxplots and kernel density plots) for clinical validations of 5 new findings based on three independent cohorts from SHC, MSMC, and CUMC. Five new findings are MCV associated with ALL at SHC and MSMC (**4A**), MGN associated with GCA at MSMC and CUMC (**4B**), PSA associated with LCA at SHC and MSMC (**4C**), ALP associated with VTE at MSMC and CUMC (**4D**), and PLT counts associated with ADS at three centers (**4E**) tested within one year lab tested before our first diagnosis. In the black box plots, bold black lines boundaries indicate the 25th, 75th percentiles of lab values, and white center squares indicate the median value of lab values. The horizontal lines indicate reference ranges of lab values. The grey shapes indicate density of the number of samples. P-values are reported by Wilcoxon Sum Rank testing.

Table 1

Summary of clinical validation through EMR from 3 independent medical centers

Finding	Disease-trait pair	Center	Total N	Cases	Controls	Gender	Lab Values*	OR (95%CI)	P-value ††	P-value †
New	ALL-MCV	SHC	255264	640	254624	Both	High+Low	3.31 (2.84-3.87)	3.79×10^{-57}	1.32×10^{-48}
	ALL-MCV	MSMC	367599	307	367292	Both	High+Low	2.40 (1.91-3.00)	9.16×10^{-15}	3.36×10^{-11}
	GCA-MGN	MSMC	120084	499	119585	Both	High+Low	1.54 (1.29-1.84)	1.45×10^{-6}	9.48×10^{-5}
	GCA-MGN	CUMC	204880	305	204575	Both	High+Low	1.59 (1.26-2.01)	1.04×10^{-4}	4.81×10^{-10}
	LCA-PSA	SHC	19203	114	19089	Male	High	2.08 (1.36-3.18)	5.0×10^{-4}	2.0×10^{-3}
	LCA-PSA	MSMC	25326	126	25200	Male	High	2.33 (1.58-3.44)	1.87×10^{-5}	0.028
	VTE-ALP	MSMC	256876	6470	250406	Both	High+Low	1.91 (1.81-2.01)	1.67×10^{-133}	4.48×10^{-252}
	VTE-ALP	CUMC	406845	1554	405291	Both	High+Low	1.30 (1.16-1.45)	3.97×10^{-6}	7.33×10^{-55}
	ADS-PLT	SHC	249091	1635	247456	Both	High+Low	2.12 (1.92-2.35)	1.24×10^{-52}	4.37×10^{-32}
	ADS-PLT	MSMC	360628	5445	355183	Both	High+Low	1.84 (1.74-1.95)	1.42×10^{-109}	2.47×10^{-43}
	ADS-PLT	CUMC	610169	965	609204	Both	High+Low	1.25 (1.09-1.45)	0.0016	2.67×10^{-6}
Positive	PCA-PSA	SHC	17481	595	16886	Male	High	10.96 (9.25-12.98)	4.43×10^{-248}	1.02×10^{-83}
	PCA-PSA	MSMC	24219	1231	22988	Male	High	7.51 (6.67-8.46)	2.0×10^{-316}	7.01×10^{-69}
	PCA-PSA	CUMC	51952	4253	47699	Male	High	9.45 (8.83-10.11)	1.02×10^{-300}	6.02×10^{-308}
Negative	ALL-PSA	SHC	19268	17	19251	Male	High	1.00 (0.12-8.13)	1	0.1
	GCA-PSA	SHC	19300	31	19269	Male	High	0.65 (0.20-2.13)	0.47	0.5

*: High=High vs. normal lab value; High+low=high and low vs. normal lab values

††: Chi-square test

†: Wilcoxon sum-rank test