Pleiotropic Variability Score: A Genome Interpretation Metric to Quantify Phenomic Associations of Genomic Variants

Khader Shameer[1, 2*], Benjamin S. Glicksberg[1, 2^], Marcus A. Badgeley[1, 2^], Kipp W. Johnson[1, 2], and Joel T. Dudley[1,2,3*]

1) Department of Genetics and Genomics, Icahn Institute of Genomics and Multiscale Biology, New York, NY 10029

2) Harris Center for Precision Wellness, Mount Sinai Health System, New York, NY 10029, NY

3) Population Health Science and Policy, Mount Sinai Health System, New York, NY 10029, NY.

*Corresponding author: joel.dudley@mssm.edu; drshameerkhader@gmail.com

^Equal contributions

Keywords: phenomics, genomics, genetic variant, variant interpretation, translational bioinformatics

Abstract

A more complete understanding of phenomic space is critical for elucidating genome-phenome relationships and for assessing disease risk from genome sequencing. To incorporate knowledge of phenomic variability into variant interpretation, we developed a new genome interpretation metric called, Pleiotropic Variability Score (PVS). PVS uses ontologies of human diseases and medical phenotypes, namely human phenotype ontology (HPO) and disease ontology (DO), to compute the similarities of disease and clinical phenotypes associated with a genetic variant based on semantic reasoning algorithms. We tested 78 unique semantic similarity methods and integrated six robust metrics to define the pleiotropy score of SNPs. We computed PVS for 12, 541 SNPs (10, 021 SNPs mapped to DO phenotype and 8, 569 SNPs mapped to HPO phenotypes) using a repertoire of 382 HPO and 317 DO unique phenotype terms compiled from genotype-phenotype catalog. We validated the utility of PVS by computing pleiotropy using an electronic health record linked genomic database (Bio*ME*, n=11,210) and generated "allele-specific pleiotropy". Further we demonstrate the application of PVS in personalized medicine using "personalized pleiotropy score" reports for individuals with genomic data that could potentially aid in variant interpretation. We further developed a software framework to incorporate PVS into VCF files and to consolidate pleiotropy assessment as part of genome interpretation pipelines. As the genome-phenome catalogs are growing, PVS will be a useful metric to assess genetic variation to find SNPs with highly pleiotropic effects. Additionally, variants with varying degree of pleiotropy can be prioritized for explorative studies to understand specific roles of SNPs and pleiotropic hubs in mediating novel phenotypes and drug development.

INTRODUCTION

Phenomics is the systematic study of the physical and biochemical traits of organisms due to variation in genetic, metabolic, environmental or other perturbations[1, 2]. Phenomics is an emerging component of multiscale biology which seeks to improve the understanding of the pathophysiology of human disease to advance translational science[3] and aid drug discovery[4]. Evaluation of biological or clinical phenotypes (phenome) will help in ascertaining the genotype-phenotype relationships mediating complex diseases and genetic disorders[5]. The translational bioinformatics community provides tools and methods to capture, mine, and analyze phenomics data[6-8]. With the recent growth in human genome-phenome databases, pleiotropy can be compiled from by systematic integration of phenotypes associated with a genetic loci[9, 10]. Phenomic inference tools (PheGenI[11], PhenoGen[12]), ontologies[13] and databases (PhenomicDB[14]) for capturing and integrating biological and clinical phenomic data are expanding. Recent efforts to unify human phenotype knowledge into ontologies have fuelled quantitative estimates of human phenotypes to predict disease modalities[15, 16]. Electronic medical record (EMR) linked genome-wide and phenome-wide association studies (GWAS and PheWAS) expanded our understanding of how a single genotype could influence multiple phenotypes. Variant-specific, trait-specific, or systematic PheWAS has uncovered novel genotype-phenotype associations that further show the network architecture of human diseases at the genome-phenome level. For example, genetic variants affecting platelet traits were associated with various conditions including inflammatory spondylopathies, hemorrhoids, acute bronchospasm, and myocardial infarction[17-19].

Multiple studies have shown that rare, orphan, complex, or genetic diseases have shared genomic and phenomic components[20-23]. Thus, "pleiotropy"[24] the phenomenon of a genetic variant, gene or specific genomic region within an linkage disequilibrium block influencing multiple phenotypic traits plays an important role in the evolution of traits and complex phenotypes[25-29]. Studying the network

properties of human genes in the context of pleiotropic effects reveals the role of one gene in different disease pathogeneses[30, 31]. For example, 8q24[32] is associated with prostate cancer, breast cancer, ovarian cancer, colorectal cancer and bladder cancer whereas 9p21[33, 34] is associated multiple cardiovascular diseases. While these regions affect distinct clinical manifestations – they are mediated by an underlying common mechanism of pathophysiology (e.g. cancers and disease of cardiovascular systems) and specific genomic regions showing hypermutation (kataegis) trends[35].

Pleiotropy on a SNP level is often more apparent. For example, rs3184504, a trypsin to arginine variant on an adapter protein SH2B3, is associated with more than 12 phenotypes including hypothyroidism, blood pressure, type 1 diabetes, coronary heart disease, arthritis, myocardial infarction, peripheral arterial disease and several medical traits like eosinophil counts[19, 36]. On the other end of the pleiotropy spectrum, an intronic SNP within *BTBD9,* rs9357271, is only associated with restless leg syndrome[37]. Here the SNPs have several phenomic associations, but a quantification of pleiotropy is not possible without the understanding of individual diseases and their etiology by manual abstraction of disease classes using manual curation. In a typical genome-interpretation workflow, the variants with multiple phenotypes get prioritized irrespective of how similar or correlated the phenotypes are. Precisely interpreting phenomic similarity of a given variant could help to more accurately quantify the pleiotropy associated with a genetic variant.

Genomic variant interpretation methods help to understand the contributions of SNPs or structural variants to the clinical manifestation of a complex diseases [38, 39]. Existing metrics for variant interpretations are based on gene or protein sequence information, sequence conservation, evolutionary divergence, and perturbation impact of variant on the protein structure[38]. These measures provide possible links between the clinical phenotype and variant data generated from whole genome sequencing (WGS) or whole exome sequencing (WES) [40, 41]. Current methods do not estimate the similarity or diversity of the phenotype set associated with a given

variant. Furthermore, current guidelines for genome interpretation or software frameworks do not consider phenomic associations of a genetic variant for prioritizing variants[39, 42, 43]. Despite the growth in translational biology and large-scale PheWAS studies[18, 44-47], phenomics based metrics are sparse in variant interpretation[39, 42].

In this report, we introduce a new metric called Pleiotropic Variability Score (PVS) to quantify the diversity of phenotypes associated with genetic variants. The method is developed using annotations compiled from a database of genotypes and phenotypes and two biomedical ontologies. A wide range of ontology based similarity assessment measures have been developed for information extraction from ontologies and structured data in the fields of biocuration[48], natural language processing (NLP)[49], artificial intelligence[50], and Semantic Web mining[51]. Such semantic similarity scores were used as the algorithmic bases for analyzing massive biological and clinical datasets including functional similarity of gene ontology (GO) terms, comparative phenomics of plant species, gene prioritization, and disease network reconstruction[52-55].

We have validated PVS as a metric using an EMR-linked genomics database (Bio*ME* biobank). We packaged the PVS system into a software framework (PVSAnnotate) to introduce PVS into VCF files as part of genome-annotation pipelines and to generate reports for genomic interpretations. Collectively our method, software, and the associated database for quantifying pleiotropy of genetic variants could provide a quantitative phenomics score as a component of genome interpretation pipelines.

METHODS

We designed a scoring method to quantify pleiotropy of phenotypes associated with a variant. We used GWASdb v2 to define canonical genotype-phenotype relationships and two ontologies to relate phenotypes via computed phenomic similarities. We developed the scoring method using a systematic survey of various semantic similarity measures and identify the critical combination of metrics that

can effectively compute pleiotropy. Briefly, pleiotropy can be defined as the inverse of the semantic similarity of two biological entities (for example genes, or SNPs) in a mathematical expression. This simple estimation gets complex due to the availability of multiple semantic similarity methods, and often a gene or an SNP may have more than two functions associated with GWAS, PheWAS, or function test experimental assays. Here we describe the ontologies, semantic similarity scores, and estimation used to develop a new method for quantifying phenomic associations of genetic variants called Pleiotropic Variability Score (PVS). While the technique could be applied to any biological entities (genes, proteins, metabolites, transcripts, microbiome, etc.), we have only tested the approach's utility as a genomic interpretation metric. Analyses and validation of the utility beyond SNPs are out of scope for this work.

Genotype-phenotype annotation databases:

Multiple databases catalog genotype-phenotype relationships of common or rare genetic variants[56-60]. While many of these databases provide phenotype information, to the best of our knowledge only GWASdb[61] and GWASdb v2[62] curate phenotype codes using standard ontologies. Other databases map phenotypes defined in the original publication of genotype-phenotype discovery or replication.

Ontologies for pleiotropy assessment:

An ontology is defined as a compendium of knowledge about a domain (e.g., medicine, biology) represented using a controlled, standardized vocabulary for describing both concepts and the relationships between concepts. A well-defined ontology can aid in the logical abstraction of knowledge about a field of study (e.g. gene ontology to summarize biological function of a gene or enrichment analyses of gene sets). Biofoundry lists 136 active biological, healthcare, and clinical ontologies. Out of these, 35 ontologies incorporate phenotype concepts. Some of these ontologies are species specific (e.g. Drosophila Phenotype Ontology, Fission Yeast Phenotype Ontology, Human Phenotype Ontology), while others are organism agnostic and describe primary attributes (e.g. Measurement method ontology). We

used human phenotype ontology (HPO) and human disease ontology (DO) to compute phenomic similarities as a pleiotropy measure of genetic variants. Ontological structure and various metrics of HPO and DO are provided in Table 1.

Human Phenotype Ontology (HPO): HPO is a structured and controlled vocabulary that relates the phenotypic features encountered in human hereditary and non-hereditary disease [63, 64]. HPO has been used in comparative disease analysis tools designed for clinical phenotyping and diagnosis (e.g. Phenomizer[65]). Clinical modifier, Mode of inheritance, Mortality/Aging, and Phenotypic abnormality represent the base classes of HPO. HPO has more than 11,000 terms and over 115,000 annotations of hereditary diseases and provides annotations to approximately 4,000 common diseases. We have used the data-version/format-version: 1.2 with date stamp releases/2015-02-14 for the analyses described in this study.

Human Disease Ontology (DO): DO is a hierarchical, controlled vocabulary for classifying human disease [66, 67]. Disease by infectious agent, disease of anatomical entity, disease of cellular proliferation, disease of mental health, disease of metabolism, genetic disease, physical disorder, and syndrome represent the base classes of DO. DO has been used as the knowledgebase for disease enrichment analyses[68]. We have used the data-version: releases/2015-02-12 with date stamp 11/02/2015 for the analyses described in this study.

Computing phenomic similarity of clinical traits and human diseases:
The pleiotropic score of a variant is derived using semantic similarity metrics computed using HPO and DO. Semantic similarity metric compute an objective distance between terms in a knowledgebase or ontology, where the distance is based on the meaning or semantic content as opposed to similarity which can be extracted from their syntactical representations. Multiple methods have been implemented to derive semantic similarity scores that use different features (topological similarity, natural language processing, text mining, or statistical

similarity) of knowledge corpus[69]. Semantic similarity can be derived using a diverse set of enumeration methods. Some of the commonly used metrics in biological and clinical domain include the methods proposed by Resnik[70], Lin[71], and Wang et al.[72]. Semantic similarity scores are typically computed on a pair-wise basis or set-wise; furthermore, these scores can be summarized as min, max or best metric average (BMA). These metrics have been part of software frameworks to calculate gene-function or gene-disease similarities and similarities of terms in ontologies[52, 65, 73, 74]. We used an ontology agnostic software framework (Semantic Measures Library and Toolkit, http://www.semantic-measures-library.org/) to compute phenomic similarities as part of PVS algorithm and derive the best-representative scores from the library of 128 methods. To avoid known methodical bias observed in phenotype similarity associated with diseases[75] and to address the coverage of phenotype ontologies[76], we performed a data-driven comparative analyses of various semantic similarity measurements to identify the most robust metrics to derive pleiotropic similarities across phenotype ontologies.

Phenomic similarity of genetic variants using Resnik method: The semantic similarity of a variant with two phenotypes (t_1 and t_2) can be computed using Resnik method[70] by deriving the similarity between terms in the context of an ontology (HPO or DO) using information content (IC) of their most informative common ancestor (MICA) of two terms (t_1 and t_2) using the following equation:

$$SS_{RESNIK}(t_1, t_2) = IC(t_{MICA})$$

Phenomic similarity of genetic variants using Stojanovic method: The Stojanovic method is a graph theory based method that uses a shortest path algorithm. The similarity between two terms is computed by considering their relative location in a common hierarchy, specifically the shortest path to reach a common term. The Stojanovic method does not need to traverse through the entire ontology to derive the similarity but the computation will terminate upon finding a common parent term using shortest path.

Group wise integration of pairwise scores: Pairwise similarity metrics like Resnik and Stojanovic can be summarized as (SS_{Max}), minimum (SS_{Min}), mean (SS_{Mean}), and best-match average (SS_{BMA}). If a variant is associated with two phenotypes, similarity is computed in pair-wise manner; for more than two phenotypes, a group wise integration is used. Each of the group wise semantic similarity computing methods is briefly discussed below and a more detailed account can be found in the literature.

Maximum similarity of pairwise scores: The SS_{Max} metric computes the maximum semantic similarity score over all pair wise similarities (S_{ij}) using the equation:

$$SS_{Max}(t_1,t_2) = \underset{1<=i<=m,1<=j<=n}{Min} Sij$$

Minimum similarity of phenotypes: The SS_{Min} represents the minimum semantic similarity score of all pairwise scores (S_{ij}) using the equation:

$$SS_{Min}(t_1,t_2) = \underset{1<=i<=m,1<=j<=n}{Min} Sij$$

Average phenomic similarities: The SS_{MEAN} computes the average semantic similarity of all pairs of term sets in respective ontologies using the equation:

$$SS_{Mean}(t_1,t_2) = \frac{1}{m*n}\sum_{i=1}^{m}\sum_{j=1}^{n} Sij$$

Best-match average of phenomic similarity: The SS_{BMA} method derives a best-match average from the scores using the following equation. BMA is obtained from the similarity matrix S with by computing the average of all maximum similarities on each row and column.

$$SS_{BMA}(t_1,t_2) = \frac{\sum_{i=1}^{m} SS_{\underline{\underset{1<=j<=n}{Max}}} Sij + \sum_{j=1}^{n} SS_{\underline{\underset{1<=i<=n}{Max}}} Sij}{m*n}$$

We generated 172 combinatorial group similarity metrics by implementing each possible dependency combination as in Fig SXX1. We tested several metrics from each fundamentally different basis of evaluating the ontology in combinatory fashion, including 1) single node evaluation of intrinsic information content measures (3 methods:), 2) similarity of pairs of 2 nodes (3 methods based on node set:, 3 based on node information content: , and 2 based on edges:), 3) group similarity of measures of pairs (5 methods:), and 4) group similarity of measures of many individual nodes (2 methods based on connectivity: , 2 methods based on information content:). Semantic similarity of pair of diseases were computed using the Java-based Semantic Measures Library and Toolkit[77] .

Pleiotropic Variability Score:

Multiple combinations of semantic similarity metrics and approximation yields an extensive library of scores. Integrating these scores into a single parameter is ideal to incorporate it in genome analyses and interpretation workflow. To understand how each of these metric influence scoring, we compiled data on various metrics across both ontologies (See Figure S1-S5). After assessing 128 semantic similarity method combinations, we selected the top six methods based on consensus correlation and scale of assessment (i.e. local similarity vs. global similarity of two terms or concepts) to define pleiotropy. After computing the semantic similarity using these methods, we defined PVS as the inverse of the semantic similarity.

Pleiotropy Variability Score using HPO was derived using the following equation:

$$PVS_{HPO}(V(t_1,t_2,...t_n)) = \left\langle SS_{Max,1}, SS_{Max,2}, SS_{Min,1}, SS_{Min,2}, SS_{Mean,1}, SS_{Mean,2} \right\rangle^{-1}$$

Pleiotropy using DO was derived using the following equation

$$PVS_{DO}(V(t_1,t_2,...t_n)) = \left\langle SS_{Max,1}, SS_{Max,2}, SS_{Min,1}, SS_{Min,2}, SS_{Mean,1}, SS_{Mean,2} \right\rangle^{-1}$$

Where top2MinMetrics were MIN_Stojanovic and MIN_Sim, top2MaxMetrics were MAX_Resnik and MAX_Sim, and top2MeanMetrics were AVERAGE and BMA.

Finally, we average the score across both ontologies for variants with both phenotypes to offer a complete spectrum of disease and phenotypes represented in the equation using the equations:

$$PVS = \left(\frac{PVS_{HPO}(V(t_1, t_2, \ldots t_n)) + PVS_{DO}(V(t_1, t_2, \ldots t_n)}{2} \right)$$

Different semantic similarity metrics are characterizing set similarity of terms with different scopes of consideration. To obtain a better coverage across all similarity scores that reflect the full range of similarity scopes we use an average of the top pair of metrics (Max, Min and Mean) from each of the fundamentally different scoring approaches. By incorporating multiple methods, PVS incorporates multiple layers of information in a single score. For example: *Min* metrics provides the intrinsic and extrinsic similarity of terms using global similarity trends in ontology. *Max* based metric provides information content-based similarity using local similarity. *Mean* or *BMA* based method capture shared all the best matching semantic similarity values for each term. To conclude PVS provides a single metric by wrapping the entire compendium of scoring methods to capture phenomic similarity to quantify pleiotropy.

Biological implications Comparison with other variant interpretation scores and methods:

Variants were mapped to genes and stratified by PVS score levels (L1, L2, L3, L4) and tested for biological and functional enrichment with Reactome pathways and Gene Ontology (molecular function) annotations using Enrichr (https://maayanlab.cloud/Enrichr/). We compared PVS with various genome interpretation metrics including FATHMM[78], SIFT[79], PolyPhen-2[80], CADD[81], GERP++[82], LRT[82], MetaLR, MetaSVM[83], MutationAssessor[84, 85], MutationTaster[86], PROVEAN[87], PhastCons[88], PhyloP[82], SiPhy[89], CONDEL[90], and VEST3[91]. Variant annotations were obtained from dbNSFP[92] and Ensemble Variant Effect Predictor (VEP)[93] and all analyses were based on the impact on the canonical transcript. We have also compared the PVS of variants by SNP location, CNV annotation to

understand the biological implications of PVS. For both analyses SNP position was mapped to SNP type and CNVs using SNP-Nexus pipeline.

Clinical validation:

Canonical annotations from biomedical genomics database can be used for developing theoretical models and prediction methods. To test the validity of PVS in genome interpretation, we tested the incidence of pleiotropic variants using an EMR-linked genomic biorepository. Our goal is to determine the disease comorbidity profile within pleiotropic allele carriers in a patient population. We estimated the prevalence of pleiotropic hub carriers using our validation cohort using disease phenotype mapping and SNP based overlap of PVS scores.

PVSAnnotate:

We developed a software pipeline called PVSAnnotate (Figure 3) to incorporate the computation of PVS as part of genome interpretation. VCF data handling routines in PVSAnnotate is developed using R, and the software workflow is a shell script-based implementation amenable to integration with existing genome analyzers and interpretation packages. The pipeline uses SMLTK to compute semantic similarity scores for variants in VCF to derive the allele-specific pleiotropic score and will output a formatted VCF incorporating PVS.

RESULTS:

We designed, developed and validated a new objective scoring method for genome interpretation that will rank SNPs based on known phenotype associations. We used PVS algorithm and annotated ~12K variants in human genome. A new schema for quantifying pleiotropy is introduced and examined the unique biological functions across pleiotropic levels (L1 to L4). Our results suggests that pleiotropic levels have implicit biological meaning and comparable to scores like SIFT. PVSAnnotate, a software framework based on PVS is used to annotate personalized pleiotropy of a reference genome and population scale, genomic data from EHR-linked biobank

We envisage that PVS will be part of genome interpretation pipelines and contribute to the understanding of molecular basis of disease comorbidities and network architecture of pleiotropy in human disease physiology. We demonstrate possible applications of this method by annotating a sample genome and estimating pleiotropy from an EMR-linked genomic bank using the PVSAnnotate software framework as well.

Summary of genotype-phenotype dataset:

We compiled and analyzed a total of 12, 541 SNPs with rs* identifiers and 699 phenotypes to derive canonical PVS. 72.9% of the SNPs were Transitions (changes from A <-> G and C <-> T) and 27.1% were Transversions (changes from A <-> C, A <-> T, G <-> C or G <-> T). Transitions are expected to occur twice as frequently as transitions; our dataset is indicative and thus present a good representative sample. Scores were generated with all genetic variants reported in the reference database (union of GWAS Catalog + GWASdb v2 (version gwasdb_20140908_snp_trait) + GRASP) with more than one DO (*n*=317 unique phenotypes; 88.536±22.277) or HPO (*n*=382 individual phenotypes; 72.48±13.628) associated phenotypes.
Within the DO database, type 2 diabetes mellitus, coronary artery disease, myocardial infarction, cardiovascular system disease, and bipolar disorder were the most represented diseases, comprising 3.5% of all terms annotations (*n*=8,939 terms in the background). For HPO, phenotypic abnormality, Type 2 diabetes mellitus, bipolar affective disorder, myocardial infarction, and schizophrenia were most represented phenotypes, comprising 3.45% of all HPO terms (*n*=11,065 terms in the background).

Deriving the PVS metric to quantify pleiotropy:

We tested several metrics (Figure S1) from each categories in combinatorial fashion, including 1) single node evaluation of intrinsic information content measures (3 methods:), 2) similarity of pairs of 2 nodes (3 methods based on node set: 3 based on node information content: and 2 based on edges:), 3) group similarity of measures of pairs (5 methods:), and 4) group similarity of measures of many

individual nodes (2 methods based on connectivity: 2 methods based on information content). We list brief descriptions of individual methods in Table 2 and outline each combination of methods assessed to find best-representative metrics to derive PVS in Supplementary Data (PVS_SS_Metrics_Tested.xlsx). We tested a total of 172 semantic similarity measurements and compiled the pleiotropy score using 6 metrics. Finally, we detail correlation consensus of 172 metrics assessed used to derive final set of semantic measures from different phenotype lengths (See Supplementary Data PVS-Phenotype_Len_Correlations.xlsx). Correlation plots of methods were examined, and we picked methods that agree across both ontologies using R^2 as a parameter. Finally, we picked a total of 6 methods that provides depth, breadth and reduce the bias to top terms. It should be noted that semantic similarity algorithms are sensitive to specificity of information and severity of impact increases with more general phenotypes (Figure S2-S5).

Quantifying pleiotropy using PVS:

We provide a flowchart summarizing the PVS algorithm in Figure 1 along with the semantic similarity scores computed using phenotypes associated with 12, 646 variants (10,1114/8,569 SNPs). Additionally, we detail the PVS derived from the semantic scores of the metrics computed using HPO, DO, and the final integrated score within the same figure. We calculated quartiles of the PVS (Kolmogorov-Smirnov, P<0.01) and organized the scores into four different levels of pleiotropy as follows. We hope the levels of pleiotropy using a simple L1-to-L4 system would further help to establish the use of PVS as an easy to interpret metric as part of genome interpretation workflows.

- L1 (no or low pleiotropy; 0.351±0.001), a variant with a PVS score x for a variant V is considered L1 when $Vx\{\in L1 | 0.264 \leq PVS < 0.436\}$; An example for an L1 PVS SNP would be rs1002552 with a PVS=0.308; PVS-Class=L1;

associated phenotypes used for estimating pleiotropic variability are acute lymphocytic leukemia and acute leukemia.

- L2 (low or medium pleiotropy; 0.48±0.001) when $Vx\{\in L2 | 0.436 \leq PVS < 0.52\}$; An example for an L2 PVS SNP would be rs1004446 with a PVS=0.51; PVS-Class=L2; associated phenotypes used for estimating pleiotropic variability are type 1 diabetes mellitus and multiple sclerosis.
- L3 (medium or high pleiotropy; 0.539±0) when $Vx\{\in L3 | 0.52 \leq PVS < 0.56\}$; An example for an L4 PVS SNP would be rs1046747 with a PVS=0.543; PVS-Class=L3; associated phenotypes used for estimating pleiotropic variability are asthma, leukemia, lymphoma, lymphopenia, cardiovascular system disease
- L4 (high or super pleiotropy; 2.351±0.005) when $Vx\{\in L4 | 0.56 \leq PVS < 0.673\}$; An example for an L4 PVS SNP would be rs964184 with a PVS=0.565; PVS-Class=L4; with more than 45 disease, phenotypes, or traits associations.

Determining disease prevalence for patients with highly pleiotropic variants: We sought to validate the PVS metric in a clinical cohort by first identifying any pleiotropic variant carriers and then assessing prevalence of associated disease phenotypes within these carriers. Specifically, we determined for each hub, the amount of individuals that had any combination of diseases associated with the hub. For each rsID-allele combination we identified total number of patients with the risk allele (numwithallele), number of patients heterozygous for the risk allele (numheterozygous), number of patients with a homozygous alternative genotype (numhomoalt). We then compared the disease frequency data including total number of diseases associated with variant (diseases) followed by number of patients with all but one disease in decrement order (d, d1, d2…d10) between

groups (See Figures 2, 3 and Supplementary Data File: PVS-BioME-Validation_Results.xlsx).

Comparing PVS with canonical annotations of genetic variants:

We mapped12,541 variants to 4,053 genes, 15,383 transcripts, and 1,874 regulatory features. These SNPs affect 33 different biotypes including regions (e.g. enhancers), transcription factor binding sites, and non-sense mediated decay. We outline a summary of variation consequences for the subset of coding variants along with the variant annotations compiled using Ensembl VEP and dbNSFP databases (See Figures 4, 5, 6 Supplementary Data: PVS-SNPs-Annotation.xlsx).

Allele specific pleiotropy of rs867186-G: The variant rs867186 (1KG-MAF: 0.0001) is a coding variant encoded in protein c receptor (*PROCR*) gene. *PROCR* (synonyms *CCCA*, *EPCR,* or *CCD41*) is a serine protease and involved in blood coagulation pathways, malaria, and cancer[96]. Variants in this genes have been associated with myocardial infarction, venous thromboembolism, late fetal loss during pregnancy and play a key role in inflammatory pathways [97, 98]. The variant is associated with seven phenotypes that include traits (protein C levels, coagulation factor levels, anticoagulant levels, hemostatic factors and hematological phenotypes, and D-dimer levels) and diseases (coronary heart diseases and coronary artery disease) across three populations (EUR, AFR, and SAN). These phenotypes were mapped to four terms in ontologies (coronary artery disease, arteriosclerotic cardiovascular disease, blood coagulation disease, and vein disease). Out of 1985 patients with the rs867186-G allele (5.7% homozygous patients) only 115 patients had homozygous alleles and 478 patients had one or more of the associated diseases (24.0%).

Allele specific pleiotropy of rs780094-T: The rs780094-T variant is a coding SNP on glucokinase (hexokinase 4) regulator (*GCKR*) gene and inhibits glucokinase in the liver and pancreas [99, 100]. This variant is associated with multiple clinical traits including LDL cholesterol, triglycerides, and phospholipid levels, as well as and

disease phenotypes such as calcium metabolism disease, coronary artery disease, atherosclerosis, and gout across five different populations (EUR, AFR, SAN, ASN, and AMR). Within our population, out of the 5,442 patients with the risk allele (20% homozygous patients), 25.6% individuals had one of the four diseases. Collectively these examples show the prevalence of genotypes with multiple phenotypes is lower than expected. This trend could be driven by the age of the individuals, ethnic origin, individualized susceptibility profiles, environment, and the influence of modifiable risk attributes. Replicating our initial findings using a large genotype-phenotype database will provides an estimate of pleiotropy on a population-scale and could help in identifying trends in adaptive evolution and antagonistic pleiotropic trends.

Biological Implications and Comparison with variant interpretations metrics: To understand the biological role of the genes harboring SNP, we have conducted enrichment analyses using genes mapped to the different levels. Some of the key Reactome pathways enriched across PVS levels includes platelet homeostasis (L1); axon guidance (L2); cGMP effects (L3) and channel activity (L4). Similarly, Molecular function terms enriched across PVS levels includes olfactory receptor activity (L1); calcium activated cation channel activity (L2); phosphatidylinositol kinase activity (L3); and aldo-keto reductase (NADP) activity (L4).

By mapping the distribution of PVS by SNP types and CNV types (Figure 6), interesting biological insights can be gained (Kruskal-Wallis, $p < 0.001$; unequal medians; Figure 6). For example, SNP types mapped "canonical splice site" have a mean PVS of 0.564. This is an interesting finding as splicing events are considered as an important mechanism by which genes like G-protein coupled follitropin receptor, SR proteins and TMPRSS2/ERG fusion transcripts have demonstrated pleiotropic properties. Further analyses and function test experiments may yield novel biological insights on how low or higher order pleiotropy computed by PVS may influence biological mechanisms.

We have compared PVS with several genetic variant interpretation metrics including SIFT, PolyPhen2, etc. to illustrate its utility as a genomic interpretation tool. PVS and SIFT have a negative correlation (Correlation test, $p < 0.001$, n=481 SNPs; Figure S6). From a biological perspective, this correlation is potentially indicative of higher the pleiotropy; lower the damage estimated by SIFT approach. We did not observe a correlation for PolyPhen2 scores and PVS (Correlation test, $p = 0.669$, n=557). Further analyses using expanded set of SNP data and stratified analyses by levels may yield correlations.

PVSAnnotate – A software framework for quantifying pleiotropy of genomes

The PVSAnnotate software pipeline and all related source code and data are available in the public domain. Source code and data to implement software pipeline (PVSAnnotate; See Figure 3) to annotate genomic data using pleiotropic variability score is available from the URL: https://bitbucket.org/dudleylab/pvs. The potential users should note that our knowledge of genotype-phenotype associations is rapidly expanding, hence confirming data provenance is critical to ensure the reproductivity of these scores.

Annotating a genome using PVSAnnotate:

The PVSAnnotate software pipeline can import a standard Variant Calling Format (VCF file) and will annotate all variants with PVS scores. The variants were segregated into individuals and mapped to the genotype-phenotype table and associations will be gathered. The pipeline can also produce personal PVS report and personalized risk network as by integrating with various popular bioinformatics tools including Cytoscape (See Figures 2 and 3). To illustrate the application of PVSAnnotate, we have annotated variants from whole genome profiling of a patient (NA12878, genome in a bottle sample) and provide the output (See Supplementary Data File: PVSAnnotate-VCF-Output.xlsx). For this example, we obtained the individual genome in VCF format with SNPs and genotype from ftp://ussd-ftp.illumina.com/hg19/8.0.1/NA12878. We mapped the SNPs to PVS

catalog to EBI-GWAS catalog to check the risk alleles (https://www.genome.gov/admin/gwascatalog.txt) by assuming an EA ancestry. Risk alleles were then mapped to phenotypes using canonical annotations in PVSAnnotate with GWASdb v2(http://jjwanglab.org/gwasdb) as a bridge. There is no phenotype data was available for this individual at the time of our analyses in the public domain other than a pharmacogenomic variant information (http://www.nist.gov/mml/bbd/ppgenomeinabottle2.cfm). Our annotations provide a pleiotropic assessment of various variants.

Discussion:

Pleiotropy – a fundamental biological phenomenon discovered in the '50s where a single genomic feature affects multiple phenotypes has growing relevance in the current era of precision medicine and high-throughput biology. With the availability of cost-effective phenotyping algorithms, affordable genome profiling, and digital health tools, such as EMR-linked biobanks, more genotype-phenotype associations have been discovered in the last decade than any time in the history of biology and medicine. We can now leverage genome-scale pleiotropy for various applications, including genome interpretations and complex disease risk estimations. We designed the PVS system to make use of genotype-phenotype annotations and ontologies for systematic and reproducible computation of phenotype pleiotropy of genetic variants.

PVS encompasses the pleiotropic variability of a variant by computing the similarity or diversity of all phenotypes associated with a genetic variant. Using a multitude of semantic similarity metrics, PVS can be utilized in multiple ways, such as finding novel variants from a WGS or WES study, or to assess if a gene or variant is part of a known phenomic hub. A genetic variant with high PVS can help to reveal the biological mechanisms underlying all associated phenotypes. PVS is a metric that illustrates the integration of massive amount of heterogeneous genomic and phenomic data could aid in developing better metrics for genome interpretations. Pleiotropy, along with variations such as antagonistic pleiotropy, offer options to

understand the pervasiveness of certain disease pairs and absence of others driven by genetic variants. PVS could also be used as a metric to find potential antagonistic pleiotropic trends, where the associated phenotypes do not co-manifest, on a population scale. Additionally, using PVS to systematically analyze antagonistic pleiotropy by identifying inverse comorbidity relationships in the patient population could help to identify new drug targets (pleiotropic hubs in human phenome) that drive disease pathogenesis mechanisms and could aid in developing precision therapies. Phenomic pleiotropy could also help and guide to understand molecular underpinnings of drug development strategies including drug repositioning. For example, an L1 PVS SNP rs11360 is associated with stroke; where as an L4 PVS rs877098 is associated with alcohol dependence[94] and multiple complex diseases (bipolar disorder, coronary artery disease, Crohn's disease, rheumatoid arthritis, type 1 diabetes and type 2 diabetes)[95]. Exploring these genomic targets may yield potential mechanistic insights and putative drug targets but the first could be target for a specific disease; whereas the second could be a pleiotropic target that could modulate multiple diseases.

In our dataset 10,021 SNPs have >= 2 DO associations and 8,569 SNPs have >=2 HP associations. Top 3 phenotypes with SNP associations from DO were type 2 diabetes mellitus (*n*=1471), coronary artery disease (*n*=1288) and myocardial infarction (*n*=1231). Top 3 phenotypes with SNP associations from HPO were phenotypic abnormality (*n*=1144), type 2 diabetes mellitus (*n*=958), and bipolar affective disorder (*n*=705). These examples indicates that phenotype-based data also have inherent differences in genome-phenome association and hence there is a need to bring evidence from different annotations and ontologies. In this work, we decided on a combination of semantic similarity metrics as the most appropriate method for estimating pleiotropy to overcome score inflation biases due to overrepresentation of phenotypes within a single class. Specifically, we used products of two metrics for each approximation (*Max, Min and Average*), deriving PVS from Max similarity would compute the similarity represented by the absolute terms in a given ontology. Incorporating all three approximations into PVS creates a more robust metric,

which overcomes limitations of each approximation individually. For example, if we consider two genetic variants (VarA and VarB): VarA is associated with two different phenotypes (infection and cancer) and VarB is associated with the pair of clinically different but biologically related conditions (Parkinson's disease and Alzheimer's disease). During similarity computation *Min* based approximation rank both of these variants with high pleiotropy and lead to scoring of any two terms from same or different part of ontology as pleiotropy. *Max* would rank only the extreme example as pleiotropic, but do not account for clinically interesting differential diagnoses (Alzheimer's disease and Parkinson's Disease). The *Average* method assigns the variant associated with infection and cancer as highly pleiotropic (infection and cancer) but the other variant will get an estimate on the medium scale of pleiotropy.

Previous attempts to classify or score pleiotropic effects[101-104] including seven types of pleiotropy, Fisher's geometric model, and degree of pleiotropy rely on manual assignment of phenotypes or traits to a gene, which is difficult to scale and keep up-to-date. PVS applies a more pragmatic approach by using algorithms to consider the entire, current phenotype knowledgebase. Such approaches are becoming increasingly complex with the rapid growth of low-cost, high-depth genomics projects[9, 27, 28, 105, 106]. PVS offers unbiased estimation of pleiotropy using semantic similarity as a surrogate measure for deriving pleiotropy using multiple ontologies. Compared to a previous method that relies on counts of phenotypes and a degree of surprise scoring[107], our method utilizes a systematic approach to quantify pleiotropy by considering the inherent relationships between clinical traits and disease phenotypes in a cohesive and automated fashion.

The automated clinical-grade genome interpretation offers many challenges to clinicians and translational bioinformatics communities. Efficient and precise analysis and interpretation of genomic data are critical for its use and for application in risk estimation of complex diseases, diagnoses of rare disease, pharmacogenomics, and tumor profiling for personalized, cancer-specific drug

recommendation. As pleiotropy is a crucial mechanism that drives multiple diseases and phenotypes pathogeneses, direct quantification of pleiotropy is a critical annotation layer that should be incorporated as part of genome interpretation workflows. PVS and PVSAnnotate collectively fill this gap by providing a method and software for incorporating pleiotropic variability as part of genome analyses.

Limitations and future work:

In this work, we designed, developed, and implemented a new computational approach to quantify pleiotropy of phenotypes associated with genetic variants. Our work has several limitations. In the current implementation of PVS, we calculate the semantic similarity-based estimates on two ontologies. Using ontology merging methods and incorporating additional ontologies (e.g., Experimental Factor Ontology) could improve the scoring system and will be incorporated into future versions of the PVS. Additionally, some of the genotype-phenotype databases (GWAS Central, dbGAP) provide mapping to Medical Subject Headings (MeSH). Computing semantic similarity score on the entire MeSH tree will be computationally expensive but could be useful. The broad nature of the MeSH terms could squeeze the phenotype similarity scoring to extremes when deriving *Min, Max*, or *Average*, but the addition and merging of additional ontologies would help in keeping the PVSAnnotate pipeline up to date, as well as making it more robust.

Incorporating other biological data types, like expressed quantitative trait loci (eQTL) data[108] , gene-tissue annotations[109], and gene-gene interactions[109] for refining genotype-phenotype annotations could further improve the scoring and quantification of pleiotropy. PVSAnnotate only performs the computation of the terms that are compiled in the database and this mapping is performed using *rs* identifiers, which are not stable. As such, future version of PVSAnnotate will use one or more identification system or mapping strategy to connect variants to phenotypes. Ontologies are often merged, revised, or deprecated, which could affect stability of ontological terms and could influence the scoring system. It should also be noted that different classification schemas have been proposed to describe

pleiotropy in biology (for example, Hodgkin proposed a seven-type classification schema that includes artefactual, secondary, adoptive, parsimonious, opportunistic, combinatorial, and unifying. In the current implementation of PVS do not account these differences.

Conclusion:

Current generation of genomic variant scoring algorithms and interpretation tools often use genomic conservation, evolutionary substitution and other cross-species alignment driven approaches to rank and prioritize genomic variants. These scoring systems have been used as utility to rank variants for clinical use of genomic variants. Such scores do not account the growing evidence from genome-wide association studies and phenome-wide association studies. In this work, we are introducing a new genomic variant interpretation metric PVS, developed using a rigorous quantitative assessment framework that provides a score to depict pleiotropy by accounting all phenotypes associated with a genomic variant.

We designed, developed, and validated an objective scoring method that will rank and classify variants based on their associated phenotypes. This metric, PVS, can quantify the SNP-phenotype associations through calculating semantic similarity from HPO and DO ontologies. PVS describes the pleiotropic nature of the variant of interest complementing existing genome-interpretation scores. Standard genetic variant filtering applications can use PVSAnnotate to filter variants according to pleiotropy. For example, SNPs with higher PVS (highly influential across the phenome) study phenotypic correlations or SNPs with lower PVS (low influence; novel) for explorative research. We have extrapolated the concept of PVS from SNPs to genes and defined pleiotropic hubs. From a clinical perspective, PVS helps interpret allelic heterogeneities and helps assess genetic variant associations with the same disease class (e.g., prostate cancer and neuroendocrine prostate cancer) or several distinct diseases (e.g., myocardial infarction and cancer). We envisage that PVS will further contribute to understanding the molecular basis of disease

comorbidities and network architecture of pleiotropy in human physiology.

Availability

Software and data discussed in this manuscript to compute pleiotropic variability score is available from the URL: https://bitbucket.org/dudleylab/pvs

Supplementary Data

Supplementary Data include six figures and 6 Excel files and can be found with this article online.

ACKNOWLEDGEMENTS

Authors would like to thank Harris Center For Precision Wellness and Icahn Institute for Genomics and Multiscale Biology (http://icahn.mssm.edu/departments-and-institutes/genomics), of Mount Sinai Health System for infrastructural support. Authors would like to acknowledge Corey Xu, Ben Readhead, Brian A. Kidd, Douglas Ruderfer, Pei Wang and Rong Chen for discussion.

FUNDING

This work was funded by the following grants from National Institutes of Health: National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK, R01-DK098242-03); National Cancer Institute (NCI, U54-CA189201-02) to JTD and National Center for Advancing Translational Sciences (NCATS, UL1TR000067) Clinical and Translational Science Award (CTSA) grant to JTD and KS. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

AUTHOR CONTRIBUTIONS

KS, MAB and BSG wrote the manuscript, KS, MAB and BSG curated the data and performed the analyses, validation and developed the software pipelines. KS, MAB, BR, BAK, PW, RC contributed ideas and participated in derivation of the scoring

method. JTD contributed to the overall planning of the project, scoring method and the manuscript. All authors read and approved the final manuscript.

COMPETING AND FINANCIAL INTERESTS

KS: Received salary, consulting fees or honoraria from Philips Healthcare, Kencor Health, OccamzRazor, Alphabet, McKinsey& Company, BCG, LEK. Employee of AstraZeneca at the time of publication; MAB: none; BSG: None declared; KWJ: None declared. JTD has received consulting fees or honoraria from Janssen Pharmaceuticals, GSK, AstraZeneca, and Hoffman-La Roche. JTD holds equity in NuMedii Inc, Ayasdi, Inc. and Ontomics, Inc. No writing assistance was utilized in the production of this manuscript.

Tables:

Table 1: Summary of Human Phenotype Ontology and Human Disease Ontology used to derive PVS

Feature	**HPO**[1]	**DO**[2]
Average number of child terms	3	3
Maximum depth of ontology	15	12
Maximum number of child terms in ontology	37	81
Terms (or classes) in ontology	11592	9189
Terms with a single child	979	539
Terms with more than 25 children	14	18
Terms with no children	3236	5600
Correlation between terms and PVS	No*	Yes**
Ontology-specific PVS	0.495±0.006	0.498±0.002
Phenotype count per SNP	3.228±0.079	2.801±0.03
SNPs with phenotypes used to derive PVS	8569	10021
Unique phenotypes used to derive PVS	382	317

All ontological metric data compiled from BioPortal (accessed on 12/01/2015).
[1]Human Phenotype Ontology https://bioportal.bioontology.org/ontologies/HP
[2]Human Disease Ontology https://bioportal.bioontology.org/ontologies/DOID
*Correlation test, P=0.232; **Correlation test, P< 0.001

Table 2: Different semantic similarity measures assessed to identify best-representative methods to derive pleiotropy

Scope	Approach	Methods
Information content	Intrinsic	sanchez, zhou seco
Information content	Extrinsic	resnik
Pairwise measures	IC-based	resnik,lin, schilcker, jc
Groupwise measures	Direct	to, nto, ui, lp, lee, ali_and_deane and gic
Groupwise measures	Indirect	min, max, avg, avgnorm, bma, bmm, maxnorm

Adapted from http://semantic-measures-library.org/sml/index.php?q=sml-semantic-measures

Full list of methods evaluated to derive PVS is provided in Supplementary File X.

Figures:

Figure 1: Graphical outline of PVS methodology

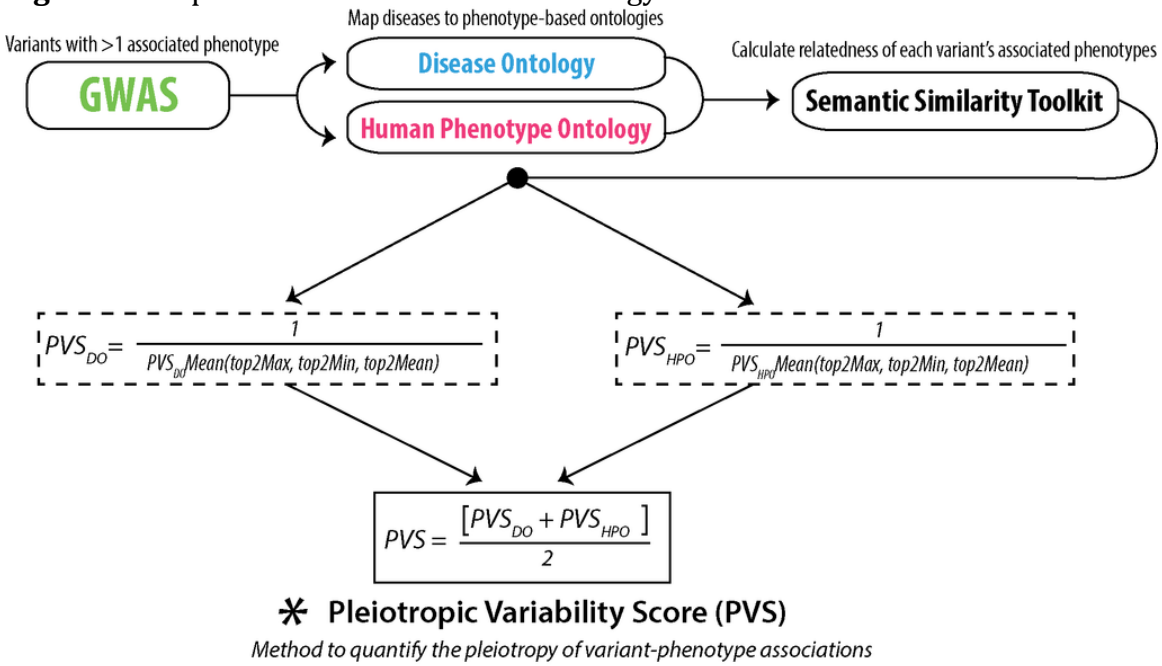
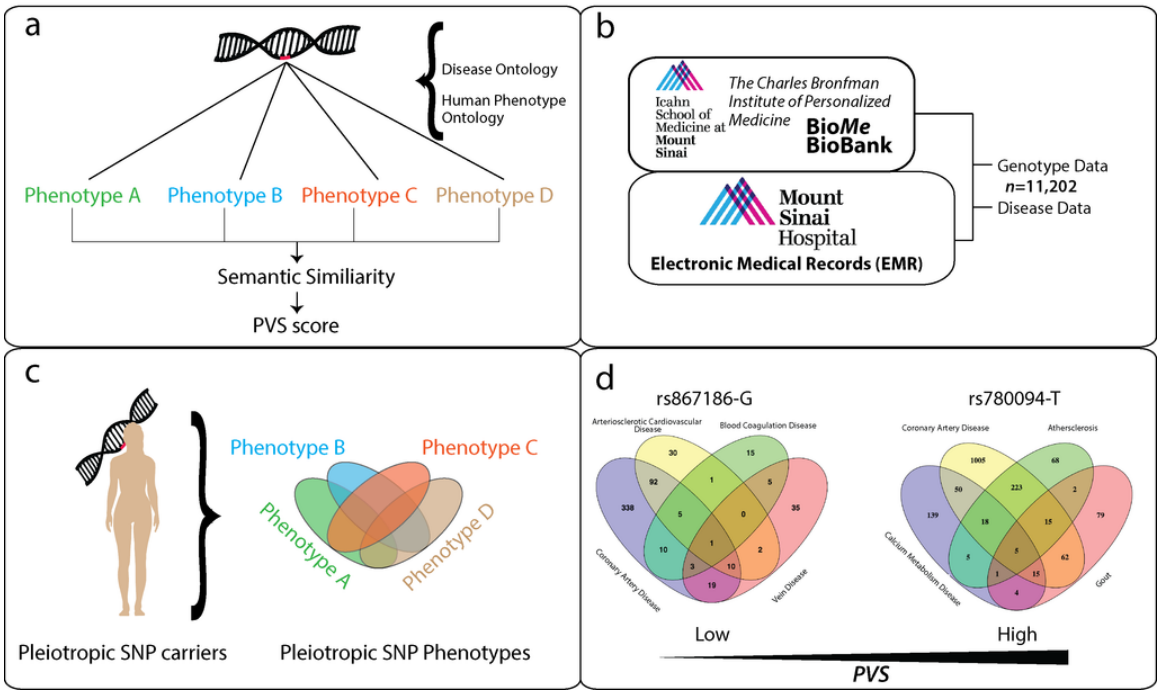


Figure 2: Visual graphic of validation approach used for Pleiotropic Variability Score using an electronic medical record (EMR) linked biobank (BioME)



a) Computation of pleiotropic variability score (PVS) b) Validation cohort c) Overlapping diseases associated with pleiotropy using EMR d) representative examples using allele specific pleiotropy (ASP) rs867186-G and rs780094-T

Figure 3: Flow-chart of software pipeline (PVSAnnotate) to incorporate pleiotropic variability as part of genome annotation and interpretation and visual graphic of a personal PVS report

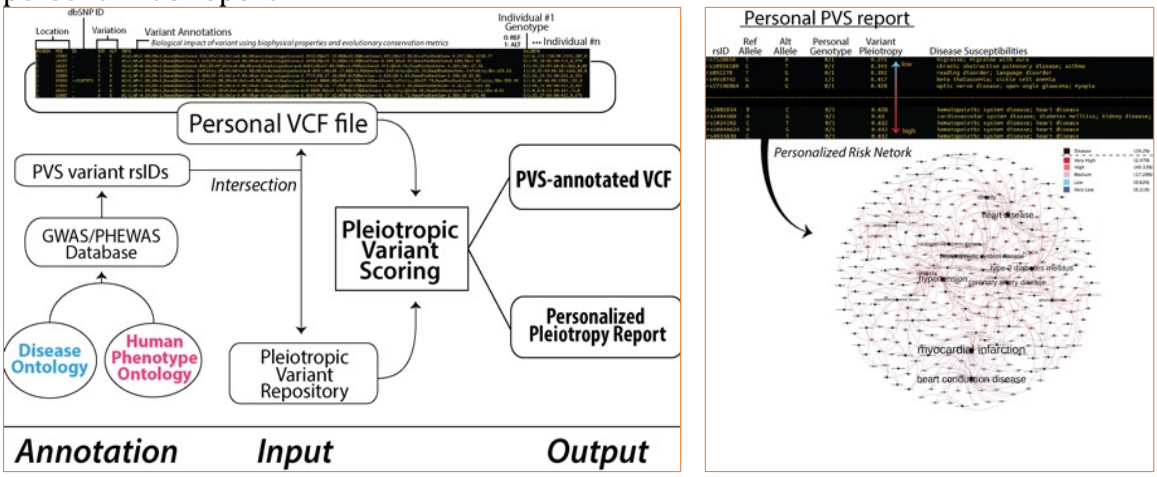


Figure 4: Summary of variants and individual scores used in PVS. Variant consequences a) all and b) coding sequence), c) SNP-phenotype coverage and semantic similarity scores using six metrics using e) DO and f) HPO ontologies

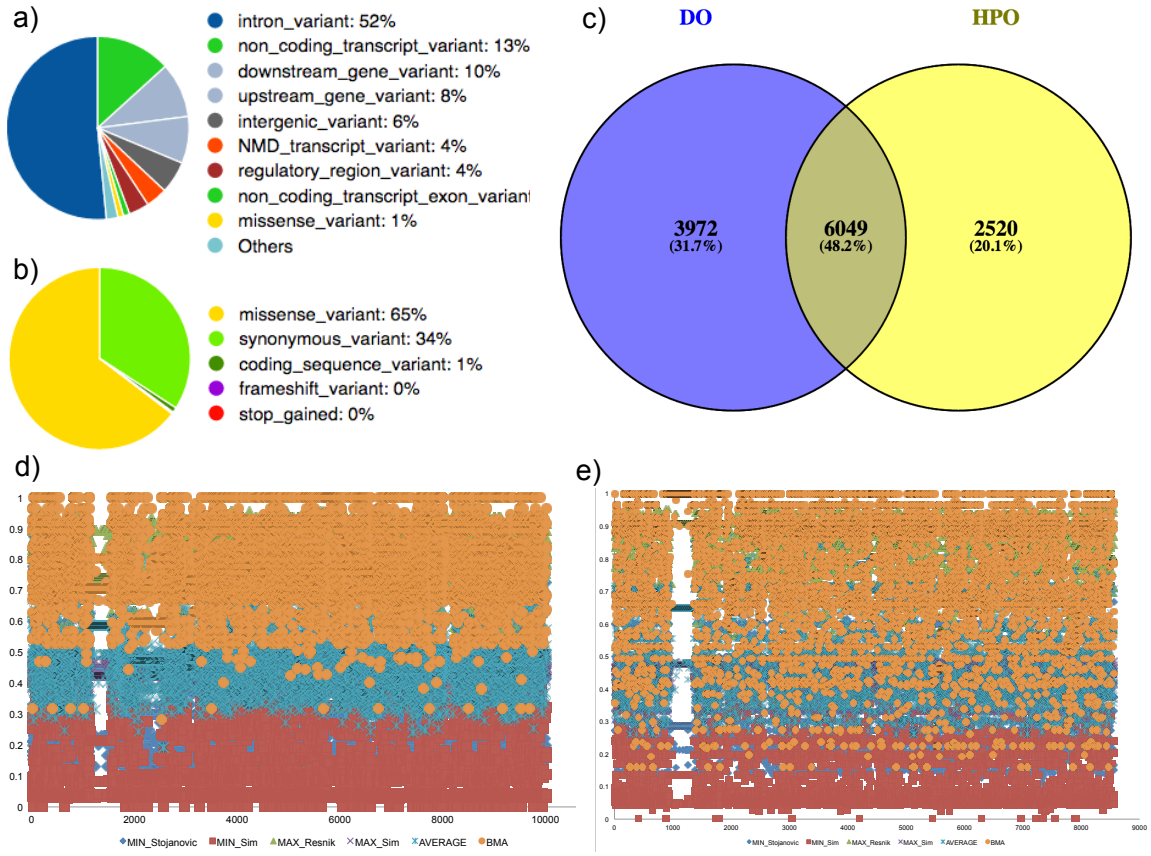


Figure 5: Pleiotropic variability scores of genetic variants a) distribution of scores estimated b) Functional enrichment of genes mapped from the variants (Reactome pathways and Ontology: Molecular Functions) across different levels of pleiotropy

a)

L1 when

$$Vx\{\in L1 | 0.264 \le PVS < 0.436\}$$

L2 when

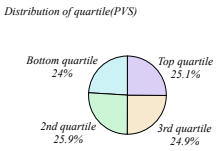
$$Vx\{\in L2 | 0.436 \le PVS < 0.52\}$$

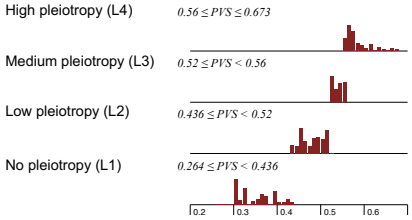
L3 when

$$Vx\{\in L3 | 0.52 \le PVS < 0.56\}$$

L4 when

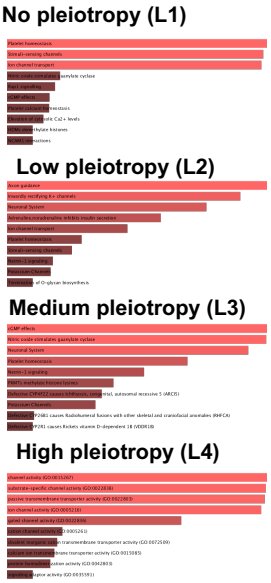
$$Vx\{\in L4 | 0.56 \le PVS < 0.673\}$$





Features	L4 Pleiotropy	L3 Pleiotropy	L2 Pleiotropy	L1 Pleiotropy
Variants	3166	3129	3905	2432
Genes	1492	1603	1405	1261
Transcripts	6003	6220	4902	4656
Regulatory regions	494	546	534	361

b)



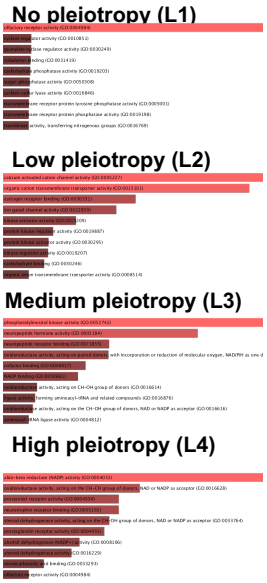
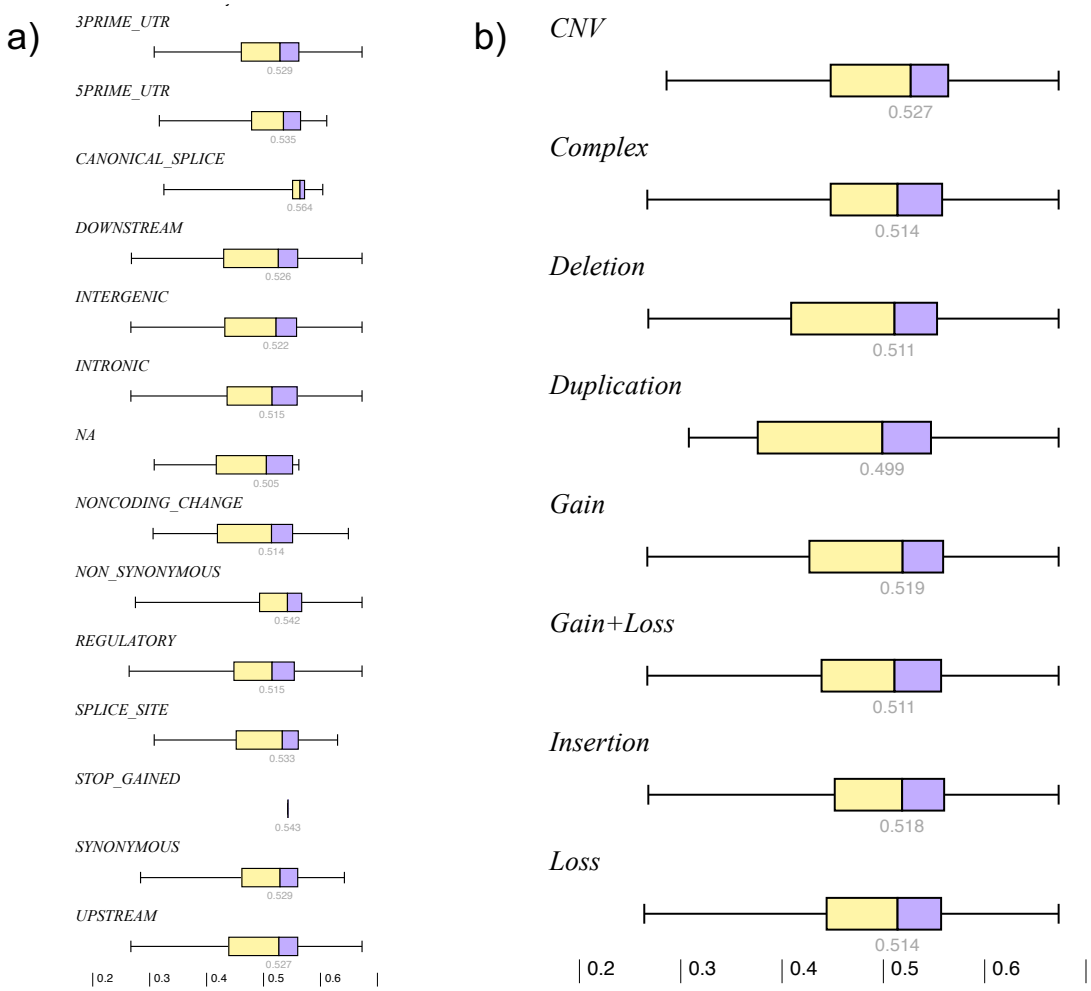


Figure 6: Distribution of PVS across SNP types and copy number variations



Reference:

1. Houle, D., Govindaraju, D.R. & Omholt, S. Phenomics: the next challenge. *Nat Rev Genet* **11**, 855-866 (2010).
2. Bilder, R.M. et al. Phenomics: the systematic study of phenotypes on a genome-wide scale. *Neuroscience* **164**, 30-42 (2009).
3. Deans, A.R. et al. Finding our way through phenotypes. *PLoS Biol* **13**, e1002033 (2015).
4. Groth, P., Leser, U. & Weiss, B. Phenotype mining for functional genomics and gene discovery. *Methods Mol Biol* **760**, 159-173 (2011).
5. Lanktree, M.B., Hassell, R.G., Lahiry, P. & Hegele, R.A. Phenomics: expanding the role of clinical evaluation in genomic studies. *J Investig Med* **58**, 700-706 (2010).
6. Toga, A.W. et al. Big Biomedical data as the key resource for discovery science. *J Am Med Inform Assoc* (2015).
7. Ritchie, M.D., Holzinger, E.R., Li, R., Pendergrass, S.A. & Kim, D. Methods of integrating data to uncover genotype-phenotype interactions. *Nat Rev Genet* **16**, 85-97 (2015).
8. Oellrich, A. et al. The digital revolution in phenotyping. *Brief Bioinform* (2015).
9. Solovieff, N., Cotsapas, C., Lee, P.H., Purcell, S.M. & Smoller, J.W. Pleiotropy in complex traits: challenges and strategies. *Nat Rev Genet* **14**, 483-495 (2013).
10. Wang, Z., Liao, B.Y. & Zhang, J. Genomic patterns of pleiotropy and the evolution of complexity. *Proc Natl Acad Sci U S A* **107**, 18034-18039 (2010).
11. Ramos, E.M. et al. Phenotype-Genotype Integrator (PheGenI): synthesizing genome-wide association study (GWAS) data with existing genomic resources. *Eur J Hum Genet* **22**, 144-147 (2014).
12. Bhave, S.V. et al. The PhenoGen informatics website: tools for analyses of complex traits. *BMC Genet* **8**, 59 (2007).
13. Gkoutos, G.V., Schofield, P.N. & Hoehndorf, R. Computational tools for comparative phenomics: the role and promise of ontologies. *Mamm Genome* **23**, 669-679 (2012).
14. Kahraman, A. et al. PhenomicDB: a multi-species genotype/phenotype database for comparative phenomics. *Bioinformatics* **21**, 418-420 (2005).
15. Petrovski, S. & Goldstein, D.B. Phenomics and the interpretation of personal genomes. *Sci Transl Med* **6**, 254fs235 (2014).
16. Baynam, G. et al. Phenotyping: targeting genotype's rich cousin for diagnosis. *J Paediatr Child Health* **51**, 381-386 (2015).
17. Denny, J.C. et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* **26**, 1205-1210 (2010).
18. Pendergrass, S.A. et al. Phenome-wide association study (PheWAS) for detection of pleiotropy within the Population Architecture using Genomics and Epidemiology (PAGE) Network. *PLoS Genet* **9**, e1003087 (2013).

19. Shameer, K. et al. A genome- and phenome-wide association study to identify genetic variants influencing platelet count and volume and their pleiotropic effects. *Hum Genet* **133**, 95-109 (2014).
20. Joy, T. & Hegele, R.A. Genetics of metabolic syndrome: is there a role for phenomics? *Curr Atheroscler Rep* **10**, 201-208 (2008).
21. Goh, K.I. et al. The human disease network. *Proc Natl Acad Sci U S A* **104**, 8685-8690 (2007).
22. Blair, D.R. et al. A nondegenerate code of deleterious variants in Mendelian loci contributes to complex disease risk. *Cell* **155**, 70-80 (2013).
23. Hidalgo, C.A., Blumm, N., Barabasi, A.L. & Christakis, N.A. A dynamic network approach for the study of human phenotypes. *PLoS Comput Biol* **5**, e1000353 (2009).
24. Williams, G.C. Pleiotropy, Natural Selection, and the Evolution of Senescence. *Evolution* **11**, 398-411 (1957).
25. Stearns, F.W. One hundred years of pleiotropy: a retrospective. *Genetics* **186**, 767-773 (2010).
26. Hill, W.G. & Zhang, X.S. Assessing pleiotropy and its evolutionary consequences: pleiotropy is not necessarily limited, nor need it hinder the evolution of complexity. *Nat Rev Genet* **13**, 296; author reply 296 (2012).
27. Hill, W.G. & Zhang, X.S. On the pleiotropic structure of the genotype-phenotype map and the evolvability of complex organisms. *Genetics* **190**, 1131-1137 (2012).
28. Wagner, G.P. & Zhang, J. The pleiotropic structure of the genotype-phenotype map: the evolvability of complex organisms. *Nat Rev Genet* **12**, 204-213 (2011).
29. Tyler, A.L., Crawford, D.C. & Pendergrass, S.A. Detecting and Characterizing Pleiotropy: New Methods for Uncovering the Connection Between the Complexity of Genomic Architecture and Multiple phenotypes. *Pac Symp Biocomput*, 183-187 (2014).
30. Chavali, S., Barrenas, F., Kanduri, K. & Benson, M. Network properties of human disease genes with pleiotropic effects. *BMC Syst Biol* **4**, 78 (2010).
31. Feldman, I., Rzhetsky, A. & Vitkup, D. Network properties of genes harboring inherited disease mutations. *Proc Natl Acad Sci U S A* **105**, 4323-4328 (2008).
32. Sur, I., Tuupanen, S., Whitington, T., Aaltonen, L.A. & Taipale, J. Lessons from functional analysis of genome-wide association studies. *Cancer Res* **73**, 4180-4184 (2013).
33. Harismendy, O. et al. 9p21 DNA variants associated with coronary artery disease impair interferon-gamma signalling response. *Nature* **470**, 264-268 (2011).
34. Visel, A. et al. Targeted deletion of the 9p21 non-coding coronary artery disease risk interval in mice. *Nature* **464**, 409-412 (2010).
35. Nik-Zainal, S. et al. Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979-993 (2012).
36. Kullo, I.J. et al. The ATXN2-SH2B3 locus is associated with peripheral arterial disease: an electronic medical record-based genome-wide association study. *Front Genet* **5**, 166 (2014).

37. Yang, Q. et al. Association studies of variants in MEIS1, BTBD9, and MAP2K5/SKOR1 with restless legs syndrome in a US population. *Sleep Med* **12**, 800-804 (2011).
38. Shameer, K., Tripathi, L.P., Kalari, K.R., Dudley, J.T. & Sowdhamini, R. Interpreting functional effects of coding variants: challenges in proteome-scale prediction, annotation and assessment. *Brief Bioinform* (2015).
39. Richards, S. et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* **17**, 405-424 (2015).
40. Dewey, F.E. et al. Clinical interpretation and implications of whole-genome sequencing. *JAMA* **311**, 1035-1045 (2014).
41. Ashley, E.A. et al. Clinical assessment incorporating a personal genome. *Lancet* **375**, 1525-1535 (2010).
42. MacArthur, D.G. et al. Guidelines for investigating causality of sequence variants in human disease. *Nature* **508**, 469-476 (2014).
43. Dewey, F.E. et al. Sequence to Medical Phenotypes: A Framework for Interpretation of Human Whole Genome DNA Sequence Data. *PLoS Genet* **11**, e1005496 (2015).
44. Pendergrass, S.A. et al. The use of phenome-wide association studies (PheWAS) for exploration of novel genotype-phenotype relationships and pleiotropy discovery. *Genet Epidemiol* **35**, 410-422 (2011).
45. Ye, Z. et al. Phenome-wide association studies (PheWASs) for functional variants. *Eur J Hum Genet* **23**, 523-529 (2015).
46. Denny, J.C. et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat Biotechnol* **31**, 1102-1110 (2013).
47. Robinson, P.N. & Webber, C. Phenotype ontologies and cross-species analysis for translational research. *PLoS Genet* **10**, e1004268 (2014).
48. Howe, D. et al. Big data: The future of biocuration. *Nature* **455**, 47-50 (2008).
49. Sfakianaki, P. et al. Semantic biomedical resource discovery: a Natural Language Processing framework. *BMC Med Inform Decis Mak* **15**, 77 (2015).
50. Martinez-Costa, C., Cornet, R., Karlsson, D., Schulz, S. & Kalra, D. Semantic enrichment of clinical models towards semantic interoperability. The heart failure summary use case. *J Am Med Inform Assoc* **22**, 565-576 (2015).
51. Machado, C.M., Rebholz-Schuhmann, D., Freitas, A.T. & Couto, F.M. The semantic web in translational medicine: current applications and future directions. *Brief Bioinform* **16**, 89-103 (2015).
52. Deng, Y., Gao, L., Wang, B. & Guo, X. HPOSim: an R package for phenotypic similarity measure and enrichment analysis based on the human phenotype ontology. *PLoS One* **10**, e0115692 (2015).
53. Hoehndorf, R., Schofield, P.N. & Gkoutos, G.V. Analysis of the human diseasome using phenotype similarity between common, genetic, and infectious diseases. *Sci Rep* **5**, 10888 (2015).
54. Oellrich, A. et al. An ontology approach to comparative phenomics in plants. *Plant Methods* **11**, 10 (2015).

55. Vissers, L.E. & Veltman, J.A. Standardized phenotyping enhances Mendelian disease gene identification. *Nat Genet* **47**, 1222-1224 (2015).
56. Welter, D. et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* **42**, D1001-1006 (2014).
57. Leslie, R., O'Donnell, C.J. & Johnson, A.D. GRASP: analysis of genotype-phenotype results from 1390 genome-wide association studies and corresponding open access database. *Bioinformatics* **30**, i185-194 (2014).
58. Eicher, J.D. et al. GRASP v2.0: an update on the Genome-Wide Repository of Associations between SNPs and phenotypes. *Nucleic Acids Res* **43**, D799-804 (2015).
59. Johnson, A.D. & O'Donnell, C.J. An open access database of genome-wide association results. *BMC Med Genet* **10**, 6 (2009).
60. Thorisson, G.A. et al. HGVbaseG2P: a central genetic association database. *Nucleic Acids Res* **37**, D797-802 (2009).
61. Li, M.J. et al. GWASdb: a database for human genetic variants identified by genome-wide association studies. *Nucleic Acids Res* **40**, D1047-1054 (2012).
62. Li, M.J. et al. GWASdb v2: an update database for human genetic variants identified by genome-wide association studies. *Nucleic Acids Res* (2015).
63. Robinson, P.N. et al. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *Am J Hum Genet* **83**, 610-615 (2008).
64. Kohler, S. et al. The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Res* **42**, D966-974 (2014).
65. Kohler, S. et al. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *Am J Hum Genet* **85**, 457-464 (2009).
66. Du, P. et al. From disease ontology to disease-ontology lite: statistical methods to adapt a general-purpose ontology for the test of gene-ontology associations. *Bioinformatics* **25**, i63-68 (2009).
67. Schriml, L.M. et al. Disease Ontology: a backbone for disease semantic integration. *Nucleic Acids Res* **40**, D940-946 (2012).
68. Osborne, J.D. et al. Annotating the human genome with Disease Ontology. *BMC Genomics* **10 Suppl 1**, S6 (2009).
69. McInnes, B.T. & Pedersen, T. Evaluating measures of semantic similarity and relatedness to disambiguate terms in biomedical text. *J Biomed Inform* **46**, 1116-1124 (2013).
70. Resnik, P. in Proceedings of the 14th international joint conference on Artificial intelligence - Volume 1 448-453 (Morgan Kaufmann Publishers Inc., Montreal, Quebec, Canada; 1995).
71. Lin, D. in Proceedings of the Fifteenth International Conference on Machine Learning 296-304 (Morgan Kaufmann Publishers Inc., 1998).
72. Wang, J.Z., Du, Z., Payattakool, R., Yu, P.S. & Chen, C.F. A new method to measure the semantic similarity of GO terms. *Bioinformatics* **23**, 1274-1281 (2007).
73. Yu, G. et al. GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics* **26**, 976-978 (2010).

74. Li, J. et al. DOSim: an R package for similarity between diseases based on Disease Ontology. *BMC Bioinformatics* **12**, 266 (2011).
75. Jing, W., Xianxiao, Z., Jing, Z. & Zheng, G. in Bioinformatics and Biomedical Engineering (iCBBE), 2010 4th International Conference on 1-4 (2010).
76. Winnenburg, R. & Bodenreider, O. Coverage of Phenotypes in Standard Terminologies.
77. Harispe, S., Ranwez, S., Janaqi, S. & Montmain, J. The semantic measures library and toolkit: fast computation of semantic similarity and relatedness using biomedical ontologies. *Bioinformatics* **30**, 740-742 (2014).
78. Shihab, H.A. et al. Ranking non-synonymous single nucleotide polymorphisms based on disease concepts. *Hum Genomics* **8**, 11 (2014).
79. Ng, P.C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* **31**, 3812-3814 (2003).
80. Adzhubei, I., Jordan, D.M. & Sunyaev, S.R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet* **Chapter 7**, Unit7 20 (2013).
81. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310-315 (2014).
82. Cooper, G.M. et al. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* **15**, 901-913 (2005).
83. Dong, C. et al. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum Mol Genet* **24**, 2125-2137 (2015).
84. Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res* **39**, e118 (2011).
85. Reva, B., Antipin, Y. & Sander, C. Determinants of protein function revealed by combinatorial entropy optimization. *Genome Biol* **8**, R232 (2007).
86. Schwarz, J.M., Cooper, D.N., Schuelke, M. & Seelow, D. MutationTaster2: mutation prediction for the deep-sequencing age. *Nat Methods* **11**, 361-362 (2014).
87. Choi, Y. & Chan, A.P. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics* **31**, 2745-2747 (2015).
88. Siepel, A. et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**, 1034-1050 (2005).
89. Garber, M. et al. Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics* **25**, i54-62 (2009).
90. Gonzalez-Perez, A. & Lopez-Bigas, N. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *Am J Hum Genet* **88**, 440-449 (2011).
91. Carter, H., Douville, C., Stenson, P.D., Cooper, D.N. & Karchin, R. Identifying Mendelian disease genes with the variant effect scoring tool. *BMC Genomics* **14 Suppl 3**, S3 (2013).
92. Liu, X., Jian, X. & Boerwinkle, E. dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. *Hum Mutat* **34**, E2393-2402 (2013).

93. McLaren, W. et al. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**, 2069-2070 (2010).
94. Kendler, K.S. et al. Genomewide association analysis of symptoms of alcohol dependence in the molecular genetics of schizophrenia (MGS2) control sample. *Alcohol Clin Exp Res* **35**, 963-975 (2011).
95. Wellcome Trust Case Control, C. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661-678 (2007).
96. Turner, L. et al. Severe malaria is associated with parasite binding to endothelial protein C receptor. *Nature* **498**, 502-505 (2013).
97. Van de Wouwer, M., Collen, D. & Conway, E.M. Thrombomodulin-protein C-EPCR system: integrated to regulate coagulation and inflammation. *Arterioscler Thromb Vasc Biol* **24**, 1374-1383 (2004).
98. Ruf, W. & Schaffner, F. Role of the protein C receptor in cancer progression. *Thromb Res* **133 Suppl 2**, S85-89 (2014).
99. Choi, J.M., Seo, M.H., Kyeong, H.H., Kim, E. & Kim, H.S. Molecular basis for the role of glucokinase regulatory protein as the allosteric switch for glucokinase. *Proc Natl Acad Sci U S A* **110**, 10171-10176 (2013).
100. Global Lipids Genetics, C. et al. Discovery and refinement of loci associated with lipid levels. *Nat Genet* **45**, 1274-1283 (2013).
101. Hodgkin, J. Seven types of pleiotropy. *Int J Dev Biol* **42**, 501-505 (1998).
102. Knott, S.A. & Haley, C.S. Multitrait least squares for quantitative trait loci detection. *Genetics* **156**, 899-911 (2000).
103. Orr, H.A. Adaptation and the cost of complexity. *Evolution* **54**, 13-20 (2000).
104. Poon, A. & Otto, S.P. Compensating for our load of mutations: freezing the meltdown of small populations. *Evolution* **54**, 1467-1479 (2000).
105. Tyler, A.L., Crawford, D.C. & Pendergrass, S.A. The detection and characterization of pleiotropy: discovery, progress, and promise. *Brief Bioinform* **17**, 13-22 (2016).
106. Rozet, J.M. & Gerard, X. Understanding disease pleiotropy: From puzzle to solution. *Sci Transl Med* **7**, 291fs224 (2015).
107. Sivakumaran, S. et al. Abundant pleiotropy in human complex diseases and traits. *Am J Hum Genet* **89**, 607-618 (2011).
108. Mackay, T.F., Stone, E.A. & Ayroles, J.F. The genetics of quantitative traits: challenges and prospects. *Nat Rev Genet* **10**, 565-577 (2009).
109. Santos, A. et al. Comprehensive comparison of large-scale tissue expression datasets. *PeerJ* **3**, e1054 (2015).