Methodology

For reprint orders, please contact: reprints@futuremedicine.com

Pharmacogenomics

Gene-wise variant burden and genomic characterization of nearly every gene

Yoomi Park[1], Heewon Seo[1,2], Brian Y Ryu[1] & Ju Han Kim*[,1,3]

[1]Seoul National University Biomedical Informatics (SNUBI), Division of Biomedical Informatics, Seoul National University College of Medicine, Seoul, 03080, Korea

[2]Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario, M5G 2M9, Canada

[3]Center for Precision Medicine, Seoul National University Hospital, Seoul, 03080, Korea

*Author for correspondence: Tel.: +82 2 740 8320; Fax: +82 2 747 8928; juhan@snu.ac.kr

Aim: Current gene-level prioritization methods aim to provide information for further prioritization of 'disease-causing' mutations. Since, they are inherently biased toward disease genes, methods specific to pharmacogenetic (PGx) genes are required. **Methods:** We proposed a gene-wise variant burden (GVB) method that integrates *in silico* deleteriousness scores of the multitude of variants of a given gene at a personal-genome level. **Results:** GVB in its simplest form outperformed the two state-of-the-art methods with regard to predicting pharmacogenes and complex disease genes but not for rare Mendelian disease genes. GVB* adjusted by paralog counts robustly performed well in most of the pharmacogenetic sub-categories. Seven molecular genetic features well characterized the unique genomic properties of PGx, complex, and Mendelian disease genes. **Conclusion:** Altogether, GVB is an individual-specific genescore, especially advantageous for PGx studies.

First draft submitted: 23 March 2020; Accepted for publication: 5 June 2020; Published online: 23 July 2020

Keywords: common complex disease • gene prioritization • genetic variant burden • molecular genetic feature • pharmacogenetic • rare monogenic disease

With the advent of next-generation sequencing, sequence-based gene-level approaches have emerged as useful tools for prioritizing genes and/or variants. The residual variation intolerance score (RVIS) [1] assigns a score to a gene as computed by ranking human genes of a population in terms of genic intolerance against functional genetic variations. The gene damage index (GDI) [2] involves directly computing the mutational load of each human gene. However, the current gene-level approaches focus only on investigating putative functional genes associated with rare monogenic diseases with Mendelian traits. Conversely, genes and variants associated with variable drug responses and common complex diseases are severely under-researched and comprehensive genomic characterizations of pharmacogenetic (PGx), common complex disease and neutral (or nondisease) genes remain to be elucidated.

Mendelian diseases generally express their life-threatening disease phenotypes at early stages of life whereas common complex diseases tend to express their complex or indistinct phenotypes after several decades with a significant contribution from environmental factors [3]. PGx variants, by definition, exhibit no phenotype unless exposed to a specific drug. If a gene expresses its distinct phenotype without specific (drug) exposure, it would properly be classified as a disease gene rather than a pharmacogene. In particular, in genetic terms, Mendelian diseases are characterized by very few rare variants of high penetrance in a single gene under strong selective pressure. In comparison, variants associated with common complex diseases are relatively frequent and tend to be spread across multiple genes in heterogeneous forms, suggesting weak negative selection against such variants in complex disease genes [4,5]. PGx variants occur in different contexts, generally in haplotypes consisting of one or more common and rare variants exhibiting extensive interethnic variability compared with that of disease variants [6,7]. Notably, however, even the proper use of medical terminologies to describe genetic variants is challenged. In ClinVar [8], medical terms such as 'pathogenic', 'likely pathogenic' and 'benign' have been commonly used for annotating rare Mendelian disease variants. Alternatively, it is currently recommended to annotate PGx variants with 'metabolism:

rapid, intermediate and poor', 'efficacy: resistant, responsive and sensitive', or 'risk' [9]. Nevertheless, despite the rather distinct genomic characteristics, these different genetic categories are not differentially considered by the current gene-level approaches.

RVIS [1] and GDI [2] constitute leading gene-level methods that compute a population-level score for each gene. Since the scores depend on the genomic architecture of a specific population, they poorly considered issues of interindividual and interethnic genetic variability, which has been considered as a strong predictor for explaining unexpected responses to certain drugs [10]. Development of a gene-level method independent of or unbiased toward the genomic architecture of a specific population constitutes an unmet need. Another well-recognized concern is the knowledge-driven bias of obtaining information regarding candidate gene functions and phenotypes such as biological pathways and gene ontologies [11]. One approach to alleviate this bias is to catalog distinct patterns of various genomic properties for the purpose of discerning genes from different genetic categories and systematically adopt the patterns in the subsequent analysis steps in an unbiased manner for all genes including understudied genes with unknown functions.

Furthermore, genetic intolerance-based approach is unlikely to suit all purposes. It is pretty likely that a method designed for disease-gene identification would not be successful at effectively discovering pharmacogenes or common complex disease genes. Gene-wise variant burden (GVB), an integrated gene-level measure of the cumulative impact of the multitude of deleterious variants on a given gene in a population data-free manner, has been successfully applied to address numerous PGx problems [10,12–14], but the utility of GVB in a variety of genetic backgrounds has not yet been performed. Thus, in the present study, we systematically evaluated the comprehensive genetic categories of PGx, complex disease, Mendelian disease, nondisease and lethal genes using GVB along with the gene-level methods by means of comprehensive genomic characterizations incorporating seven molecular genetic features: number of paralogs, number of singletons, per-person mutability, protein–protein interaction (PPI) degree, coding sequence (CDS) length, McDonald–Kreitman neutrality index (NI) and protein complexity.

Materials & methods

Variant annotation methods for GVB

The variants were annotated using the Phase III data available from the 1000 Genomes Project (http://www.1000genomes.org/) [15]. Protein-coding gene regions were determined using ANNOVAR (http://annovar.openbioinformatics.org/) [16]. Variant deleteriousness was predicted using seven *in silico* bioinformatics prediction methods: SIFT (http://sift.jcvi.org/) [17–19], combined annotation-dependent depletion (CADD) [20], PolyPhen2 HIVD, PolyPhen2 HVAR [21], PhyloP [22], MutationTaster [23] and GERP++ [24].

Collecting gene lists for comprehensive genetic categories

To evaluate how well the gene-level scoring methods predicted genes across multiple genetic conditions, we organized known gene lists in three genetic categories: drug-related phenotypes, complex diseases and rare Mendelian diseases. For the PGx category, we extracted five pharmacological gene sets from absorption, distribution, metabolism, and excretion (ADME) (http://www.pharmaadme.org) and the Pharmacogenomics Knowledgebase (PharmGKB; 22 April 2020) [25]; ADME core, ADME extended, ADME all, PharmGKB VIP and PharmGKB. To ensure that the evaluation was comprehensive, we classified the genes according the following pharmacological characteristics; Pharmacological effect (i.e., toxicity, metabolism/PK, dosage and efficacy) in PharmGKB, pharmacodynamic (PD) and pharmacokinetic (PK) phenotypes (i.e., target, enzyme, transporter and carrier) in the DrugBank (http://www.drugbank.ca/; v5.15) [26], and enzyme families (i.e., CYPs and UGTs) (http://www.genenames.org/cgi-bin/genefamilies).

Overall, 16 lists for common complex diseases were extracted from the Genetic Association Database (GAD; 23 July 2011) [27] using the disease class filter with the following criteria: aging, cancer, cardiovascular, chemdependency, developmental, hematological, immune, infection, metabolic, neurological, normalvariation, pharmacogenomic, psych, renal, reproduction and vision. Entries annotated with 'Y' were selected, which indicates a significant association between a gene and a disease. Unclear phenotypes (unknown and others categories) and diseases with few genes (mitochondrial category) were excluded from the 19 GAD disease classes.

For rare Mendelian disease categories, six gene lists from Online Mendelian Inheritance in Man (OMIM) [28] (i.e., recessive, haploinsufficiency, dominant-negative, *de novo*, OMIM all and nondisease) were adopted based on the report of [1]). Five nonviable gene lists were obtained from Bartha *et al.* (i.e., *in vivo* essential, *in vitro* essential and mice essential) and Petrovski *et al.* (i.e., mouse genome informatics (MGI) lethality and MGI seizure) [1,29]. The

performances of GVB, RVIS and GDI were evaluated using the pROC package of R software [30]. The predictive performance was measured by calculating the proportion of genes with a score less than a threshold among known PGx or disease gene set (true positives) against true negatives for all possible thresholds on the ranked score lists.; the gene lists are summarized in Supplementary Table 1.

Gene-specific molecular genetic features

Genes were annotated with data on seven biologically meaningful molecular genetic features. We extracted the number of paralogs, CDS length, NI and D value for 17,040 human genes from the study of Itan *et al.* [2]. According to Itan *et al.*'s work, CDS length and number of paralogs were extracted via the Ensembl Biomart (version 75). NI $[(P_N/P_S)/(D_N/D_S)]$ was estimated by comparing the numbers of nonsilent and silent substitutions (D_N and D_S) with the numbers of nonsilent and silent polymorphisms (P_N and P_S) [31]. Protein complexity was estimated using Clark's distance ($D = \sqrt{\Sigma_{i=1}^{d}(\frac{|P_i - Q_i|}{P_i + Q_i})^2}$, where P_i is the number of residues of amino acid i in protein P with n amino acids, and Q_i is the number of residues of amino acid i in the average human protein Q) by calculating the squared root of half of the divergence [32]. Information regarding the human PPI degree was obtained using the recent IntAct database (22 April 2020) for 12,128 genes [33]. The number of singletons and per-person mutability for each human gene were calculated from the 1000 Genomes Project Phase III. Coding variants observed only once in the 1000 Genomes Project data were considered as singletons. We defined per-person mutability for a gene as $M_i = (\Sigma_{j=1}^{n} V_{ij})/n$, where V_{ij} is the total number of SIFT-annotated variants within gene i of the j^{th} individual.

Calculation of GVB

To quantify the cumulative genetic effect for all coding variants of a gene, we calculated the GVB score as described previously [10,12–14]. We obtained GVB scores ranging from 0 to 1 for each of the 17,502 human genes for the 2504 personal genome sequences of the 1000 Genomes Project. A gene is predicted to be more deleterious as the GVB score approaches 0; GVB scores are listed in Supplementary Table 2. Based on a recently published GVB study to find the optimal threshold value for tolerable variation [34], variants with SIFT score >0.7 were assumed to be neutral and ignored (Equation 1).

$$G_i = \{v \mid v \text{ with a SIFT score less than } 0.7\} \quad \text{(Eq. 1)}$$

Considering allelic dosage effects, homozygous variants were considered to be more damaging than a variant in a heterozygous state (Equation 2).

$${}_{adj}v_j = \begin{cases} (SIFT\ score)^{0.5}, & if\ v_j \in G_i\ and\ heterozygote \\ SIFT\ score, & if\ v_j \in G_i\ and\ homozygote \end{cases} \quad \text{(Eq. 2)}$$

The GVB score for each gene G_i with n deleterious variants was calculated as the geometric mean of SIFT score v (Supplementary Figure 1, Equation 3).

$$GVB(G_i) = \begin{cases} 1, & if\ n(G_i) = 0 \\ (\Pi_{j=1\ adj}^{n} v_j)^{\frac{1}{n}}, & if\ n(G_i) > 0 \end{cases} \quad \text{(Eq. 3)}$$

The GVB score is 1 if count n of variant j with scores >0.7 in a gene equals 0, indicating that the gene does not harbor any potentially damaging variant. We replaced zero with near-zero (10^{-8}) values to handle a multiply-by-zero problem, when the SIFT score of the variant was zero or zero equivalent.

Adjustment of GVB by the seven molecular genetic features

Using the seven genetic parameters as weights, we systematically generated four variations of the GVB score to comprehensively evaluate its predictive power under different genetic conditions by simply dividing or multiplying each of the seven biologically meaningful molecular genetic features (Supplementary Table 3). A pseudo value of 10^{-8} was used when the parameter value was zero. Since SIFT showed the best overall predictive performance, the results are drawn mainly using the SIFT-based GVB score. We applied the adjustment scheme independently for the genes for which each of the genetic molecular features is available (Supplementary Figure 2).

Table 1. Characteristics of gene-level prioritization methods.

Feature	GVB	RVIS	GDI
Score assignment for a gene	Per person score	Population-wise score	Population-wise score
Population dependency	Independent	ESP6500 (Exome Sequencing Project v. 6500)	The 1000 Genomes Project, Phase III
Variables for variants	SIFT, CADD, PolyPhen HIVD, PolyPhen HVAR, PhyloP, MutationTaster and GERP *in silico* prediction scores	Mutation frequency	Allele frequency and CADD
Value range	[0, 1]	[-8.29, 29.75]	[0, 42.91]
Value normalization	Yes	No	No
Value type	Absolute	Relative	Relative

CADD: Combined annotation-dependent depletion; GDI: Gene damage index; GERP: Genomic evolutionary rate profiling; GVB: Gene-wise variant burden; RVIS: Residual variation intolerance score.

Results

Comparison of the genomic characteristics revealed by GVB, RVIS & GDI

We defined the GVB as the overall impact of multiple deleterious variants on a gene (Supplementary Figure 1). GVB assigns a score for each gene for each individual using *in silico* prediction scores (e.g., SIFT [17–19], CADD [20], PolyPhen2 HIVD, PolyPhen2 HVAR [21], PhyloP [22], MutationTaster [23] and GERP++ [24]), while RVIS and GDI methods use relative frequencies of alleles (i.e., variability) in a given population and/or a variant deleteriousness score (i.e., CADD [20]). Whereas scores of population-based design are sensitive to the genomic architecture of different populations, GVB is not only independent of but also a useful tool for investigating the genomic architecture of a population (Table 1). Population-specific GVB scores and their distributions for all genes can easily be created by aggregating the individually assigned GVB scores.

Given the simple mathematical GVB equations (Equations 1–3), it is straightforward for GVB to assimilate genetically significant molecular features as parameters (Supplementary Figure 2). We defined GVB* as a modification of GVB to include the seven molecular genetic features: number of paralogs, number of singletons, per-person mutability, PPI degree, CDS length, McDonald–Kreitman NI and protein complexity.

Comprehensive evaluation of GVB across multiple genetic categories

Figure 1 exhibits the distributions of the areas under the receiver operating characteristic (ROC) curves (AUCs) as the measures of the discriminant powers of GVB (and GVB*), RVIS and GDI for PGx, common complex disease, and rare Mendelian disease gene categories and subcategories. Overall, GVB outperformed RVIS and GDI in PGx and complex disease gene categories (Figure 1, top and middle panels). GVB* augmented by the number of paralogs and CDS lengths further improved the predictive performances of GVB across almost all genetic subcategories (Supplementary Table 4).

Both GVB and GVB* outperformed for determining all five PGx categories (i.e., ADME core, ADME extended, ADME all, PharmGKB VIP and PharmGKB) in addition to all four pharmacological effect categories (i.e., toxicity, metabolism/PK, efficacy and dosage). Notably, performances of both GVB and GVB* were higher for core than peripheral PGx categories, in other words, ADME core > ADME extended (AUC_{GVB}: 0.71 > 0.58; AUC_{GVB*}: 0.76 > 0.68) and PharmGKB VIP > PharmGKB (AUC_{GVB}: 0.62 > 0.55; AUC_{GVB*}: 0.72 > 0.58) (Figure 1, top panel and Figure 2A). Both GVB and GVB* demonstrated markedly high AUCs for the PK carrier, enzyme and transporter but not for the PD target category. Moreover, both GVB and GVB* demonstrated higher performances for core than peripheral PGx categories, in other words, Target PGx > target, transporter PGx > transporter, enzyme (UGT) > enzyme (CYP) > enzyme PGx > enzyme, and carrier PGx > carrier (Figure 1 top panel and Figure 2B).

Under the common complex disease model provided by the GAD (Figure 1, middle panel), each gene-level method demonstrated its own merits in different subsets. GVB consistently outperformed RVIS and GDI with regard to the GAD Pharmacogenomic subcategory (AUC_{GVB*} = 0.71, AUC_{GVB} = 0.69). Overall, GVB and GVB* exhibited higher AUCs for the majority of the GAD common complex disease subcategories, excepting GAD Developmental, Psych and Chemdependency subcategories, which do not belong to common complex diseases.

Under the traditional rare Mendelian disease model, both RVIS and GDI outperformed GVB in OMIM Haploinsufficiency, *de novo*, and dominant negative categories. Overall, RVIS performed best in predicting monogenic

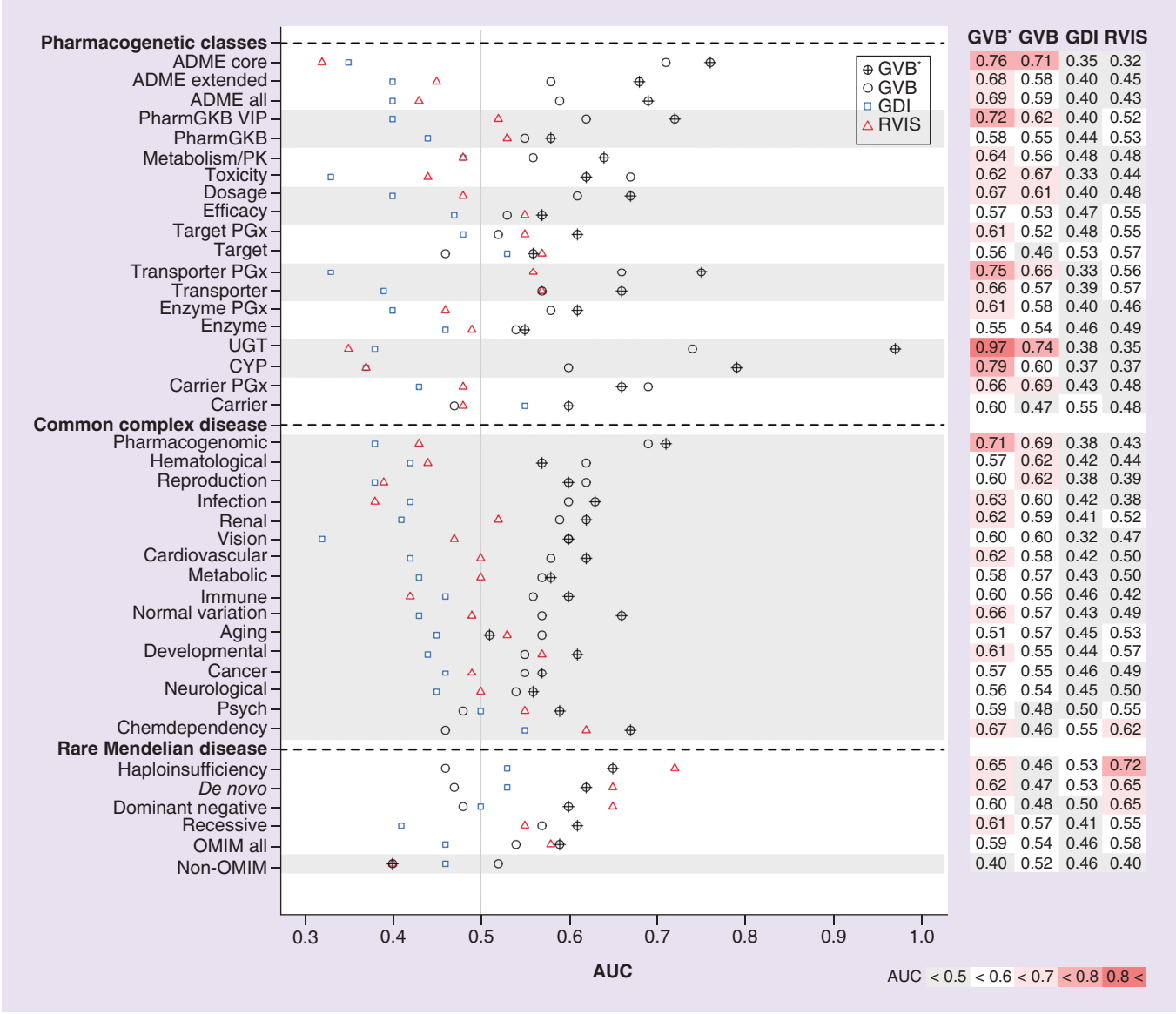

Figure 1. Performance comparison of gene-wise variant burden, Residual Variation Intolerance Score and Gene Damage Index for determining pharmacogenetic, complex disease and Mendelian disease genes. The AUC values for each subcategory are represented as colors ranging from gray (low AUCs) to red (high AUCs) according to their intensity scale. GVB*: GVB adjusted by the number of paralogs and CDS length.

ADME: Absorption, distribution, metabolism and excretion; AUC: Area under the receiver operating characteristic curve; CDS: Coding sequence; GAD: Genetic association database; GDI: Gene damage index; GVB: Gene-wise variant burden; OMIM: Online Mendelian Inheritance in Man; PGx: Pharmacogenetic; PharmGKB: Pharmacogenomics Knowledgebase; PK: Pharmacokinetic; RVIS: Residual variation intolerance score.

disease genes. Notably, GVB outperformed RVIS and GDI in the OMIM recessive subcategory (AUC_{GVB} = 0.57), for which both RVIS (AUC_{RVIS} = 0.55) and GDI (AUC_{GDI} = 0.41) functioned poorly, in part owing to the population-independent design. The rank orders of the prediction AUCs of RVIS and GDI decreased as follows: OMIM haploinsufficiency > *de novo* > dominant negative > recessive, which corresponds to the order of phenotypic severity (Figures 1 & 2).

AUCs for all of the PGx, GAD common-disease and OMIM rare-disease categories showed an inverse correlation between GVB and RVIS (Spearman's r = -0.82, -0.50 and 0.81, respectively) along with GVB and GDI (Spearman's r = -0.70, -0.80 and -0.97, respectively). This tendency could be explained by the different approaches of the methods. Whereas RVIS and GDI use population datasets to determine highly intolerant genes with severe

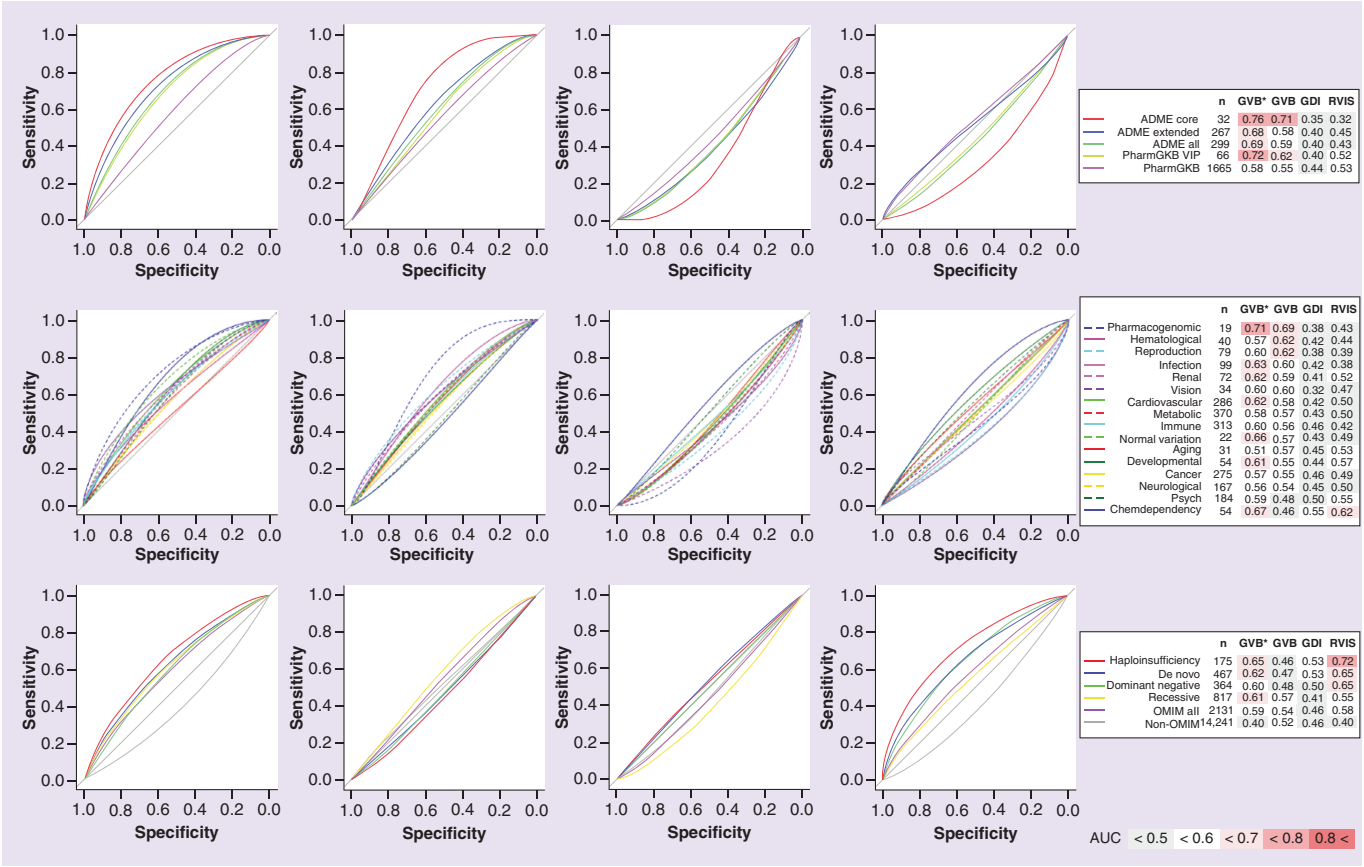

Figure 2. Performance comparison of gene-wise variant burden, Residual Variation Intolerance Score and Gene Damage Index for determining pharmacogenetic, complex disease and Mendelian disease gene categories and subcategories. (A) Comparison of GVB, GVB*, RVIS and GDI by ROC curves for five pharmacological, 16 complex disease and six Mendelian disease gene categories and **(B)** four pharmacological-effect, four pharmacodynamic- and kinetic-phenotype and two enzyme-family gene subcategories. GVB*: GVB adjusted by the number of paralogs and CDS length.
AUC: Area under the curve; CDS: Coding sequence; GDI: Gene damage index; GVB: Gene-wise variant burden; ROC: Receiver operating characteristic; RVIS: Residual variation intolerance score.

evolutionary conservation, GVB attempts to balance variant deleteriousness (by means of applying *in silico* variant prediction methods) and interindividual variabilities (by means of bottom-up aggregation of individualized scores) across the population, as appropriate.

Notably, GVB* adjusted by the number of paralogs and CDS lengths consistently improved the performances of GVB not only for PGx and GAD common-disease gene categories but also for OMIM rare-disease categories (Figures 1 & 2). Moreover, GVB* reversed the negative correlation of GVB with RVIS and GDI for the OMIM category to be positive (GVB* vs RVIS: Spearman's r = 0.78 and Kendall's τ = 0.69; GVB* vs GDI: Spearman's r = 0.62 and Kendall's τ = 0.50). The same genetic molecular features for GVB*; in other words, the number of paralogs and CDS lengths, markedly improved both GVBs albeit in the opposite directions for pharmaco- and complex disease genes and rare monogenic-disease genes. The results presented here are drawn mainly using the SIFT-based GVB that showed the best overall predictive performance (Supplementary Table 4), and consistent and successful evaluations for improved performances of boosting the seven molecular genetic features across seven different *in silico* variant prediction methods were demonstrated (Supplementary Figure 2).

Further evaluation of PGx subcategories including pharmacological effect, PK/PD and enzyme families demonstrated that GVB outperformed and GVB* further outperformed both RVIS and GDI (Figure 2B). Whereas Toxicity showed the highest AUC (AUC_{GVB} = 0.67), the AUCs most benefitting by GVB* adjustment were those for metabolism/PK (AUC_{GVB*} = 0.64 and AUC_{GVB} = 0.56) and Dosage (AUC_{GVB*} = 0.67 and AUC_{GVB} = 0.61), likely owing to the higher number of paralogs of PK enzymes than those of PD toxicity genes. When the enzyme

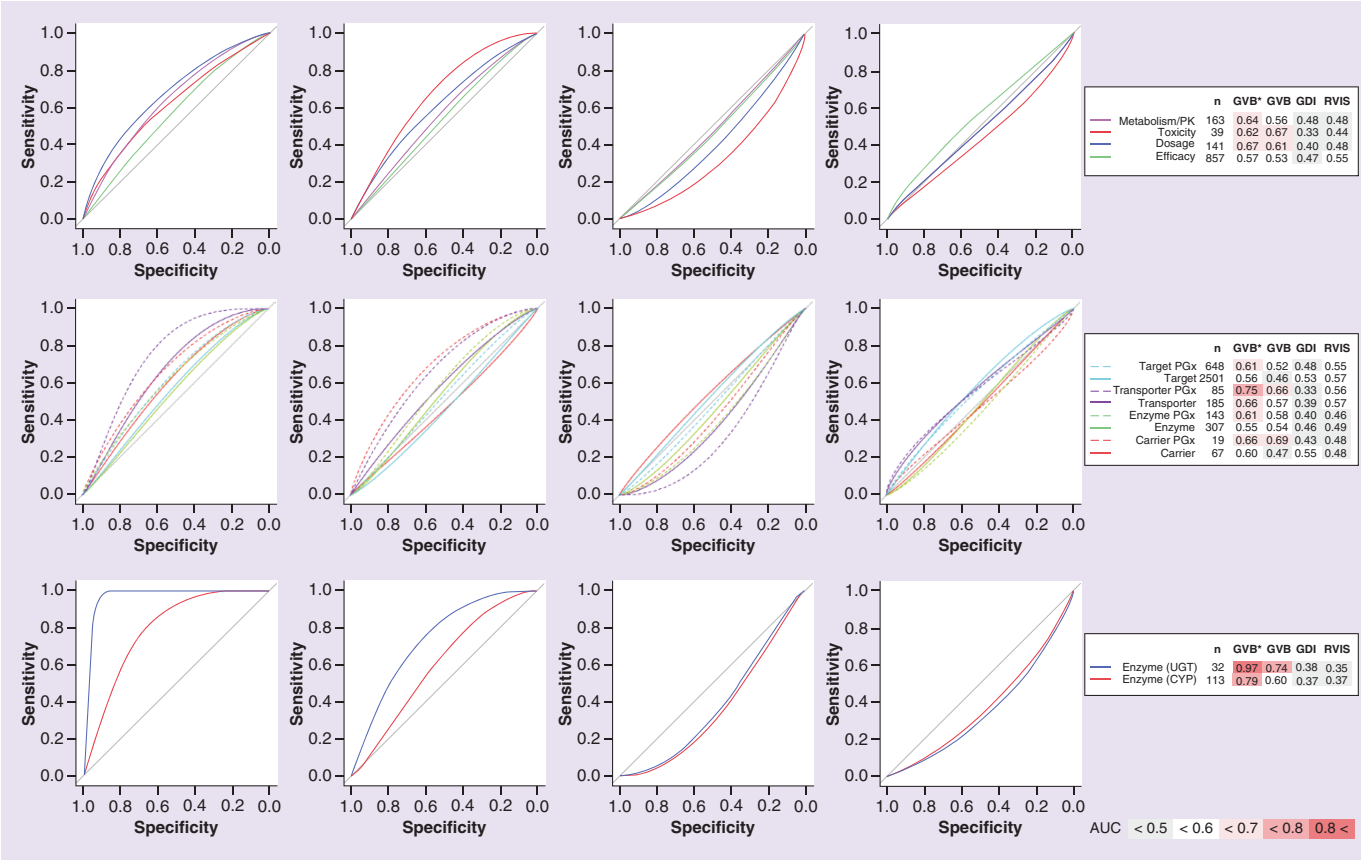

Figure 2. Performance comparison of gene-wise variant burden, Residual Variation Intolerance Score and Gene Damage Index for determining pharmacogenetic, complex disease and Mendelian disease gene categories and subcategories (cont.). (A) Comparison of GVB, GVB*, RVIS and GDI by ROC curves for five pharmacological, 16 complex disease and six Mendelian disease gene categories and **(B)** four pharmacological-effect, four pharmacodynamic- and kinetic-phenotype and two enzyme-family gene subcategories. GVB*: GVB adjusted by the number of paralogs and CDS length.
AUC: Area under the curve; CDS: Coding sequence; GDI: Gene damage index; GVB: Gene-wise variant burden; ROC: Receiver operating characteristic; RVIS: Residual variation intolerance score.

genes were categorized into gene families, the best performance was achieved among all categories by the UGT families (AUC_{GVB*} = 0.97, AUC_{GVB} = 0.74, AUC_{GDI} = 0.38 and AUC_{RVIS} = 0.35).

Genomic characterization of PGx, complex disease & Mendelian disease genes

The number of paralogs of a gene constitutes a useful indicator of essentiality based on phyletic characteristics [35]. Genes that belong to a paralogous family are assumed to exhibit high functional redundancy in order to cope with the possibility of their disruption, whereas genes without paralogs are more likely to be essential [36,37]. Figure 3 demonstrates that pharmacogenes (n = 3552) tend to be more paralogous than other genetic categories ($p < 0.001$). Mendelian disease (n = 2052) and nonviable genes (n = 1118) were less paralogous than the complex disease genes (n = 1046), likely resulting from strong selective pressure. Notably, the nondisease gene category (n = 14,316) shows fewer paralogs. When extracting genes with the most (25%) and the least intolerant quartile RVIS values, the consistent result that genes in both intolerant (5.31 ± 5.8) and tolerant (5.91 ± 8.1) quartiles tend to be more paralogous than the whole genes (4.94 ± 6.6) is shown (Supplementary Figure 3).

Singletons are defined as ultra-rare variants observed only once in a population, which corresponds in the present study to the 2504 subjects in the 1000 Genomes Project [15]. The accumulation of rare, recently evolved and potentially damaging variants within a gene itself can be important evidence of evolution, just as gene length is recognized as one of the most important evidence in evolution [38,39]. Mendelian disease genes tend to harbor a larger number of singletons, representing a variety of evolutionarily recent rare variants. Common complex disease genes are more likely to be comprised of fewer rare variants albeit more common variants that are sufficiently

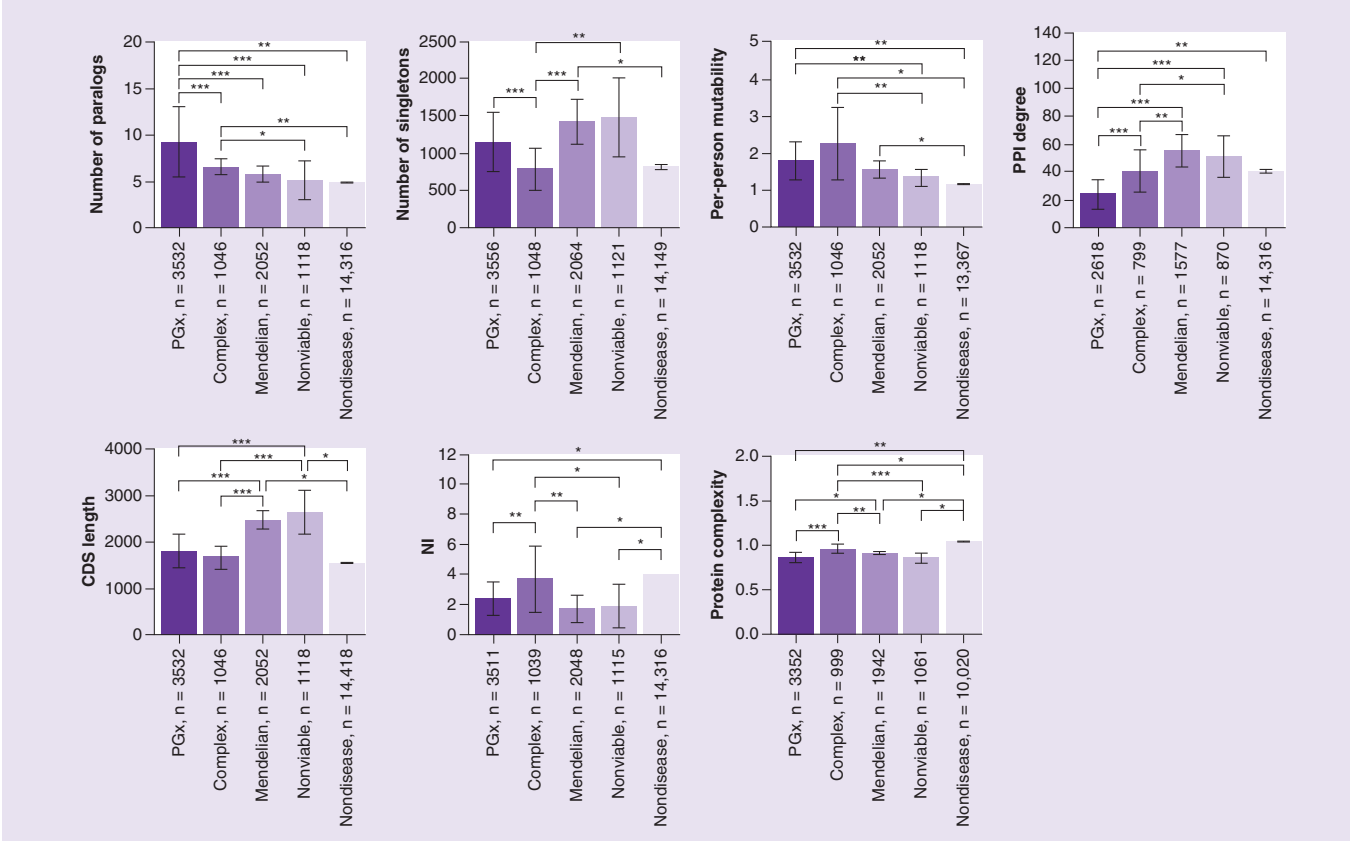

Figure 3. Genomic characterization of pharmacogenetic, complex disease, Mendelian disease, nondisease and nonviable genes with the seven genetic molecular features. The distributions for the average genetic features of the subcategories are displayed.
*p < 0.1; **p < 0.05; ***p < 0.01 by the Wilcoxon test.
CDS: Coding sequence; NI: Neutrality index; PGx: Pharmacogenetic; PPI: Protein–protein interaction.

old to be shared/tolerated across populations. PGx genes harbor more singletons than complex disease genes but fewer than Mendelian disease genes (Figure 3). Local adaptive selection is a plausible explanation for PGx genes carrying common functional variants. Positively selected variants can lead to an increased frequency of linked deleterious variants by genetic hitchhiking, resulting in higher interethnic heterogeneity that helps to avoid adverse drug reactions in certain populations [6,7].

We defined the 'per-person mutability' of a gene as the average number of protein-coding variants of a gene to indicate the endurable degree of genetic variability for an individual human genome. Complex disease genes showed the highest per-person mutability, allowing a gene to harbor multiple variants associated with complex diseases and/or phenotypes. Pharmacogenes exhibited moderate and Mendelian disease genes the lowest levels of per-person mutability (Figure 3).

NI estimates the selective pressure on human genes, where NI <1 (an excess of nonsynonymous divergence) indicates a signal of purifying selection and NI >1 (an excess of nonsynonymous polymorphism) indicates a signal of positive selection [2,40,41]. The average NI values for complex disease genes were higher than those of Mendelian disease genes (Figure 3), which may imply that increased levels of deleterious variants have long been accumulated in complex disease genes although their CDS lengths are shorter than those of Mendelian disease and nonviable genes. Figure 3 demonstrates that PGx genes have also accumulated more deleterious variants (a signal of (locally adaptive) purifying selection) than Mendelian disease and nonviable genes. GVB successfully captured these complex genomic characteristics of high NI and per-person mutability for PGx along with complex disease genes.

Mendelian disease and nonviable genes tended to have greater PPI degrees and CDS lengths (Figure 3), which are indicators of gene essentiality and disease-associated mutations [42–44]. Common variants accumulate in disordered regions whereas pathogenic variants are significantly depleted in disordered regions [2,45,46]. Disordered amino acid

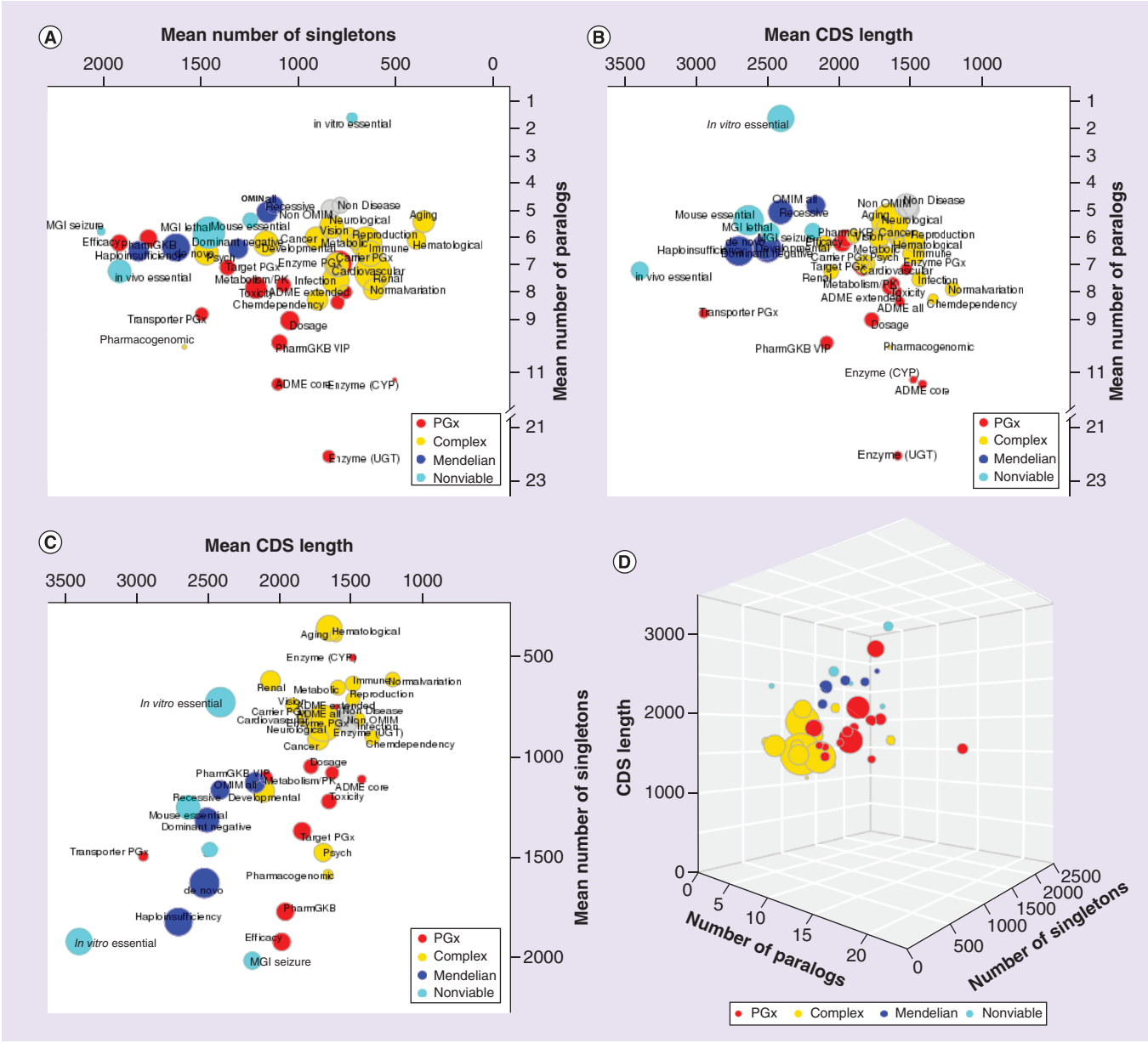

Figure 4. Genomic landscape of genetic molecular features among pharmacogenetic, complex disease and Mendelian disease gene categories. Genomic landscape of genetic molecular features differentiating **(A)** pharmaco- versus common complex disease genes, **(B)** pharmaco- versus rare Mendelian disease genes and **(C)** common complex versus rare Mendelian disease genes. Mean numbers of singletons (as evidence for component of genetic variability) and paralogs (as evidence for degree of selective pressure) and mean CDS lengths are represented by the horizontal and vertical axes and PPI degree and protein complexity by circle size. **(D)** Genomic landscape of phenotypic subgroups is visualized using four genetic features including per-person mutability.
CDS: Coding sequence; PGx: Pharmacogenetic; PPI: Protein–protein interaction.

composition was estimated using Clark's distance [2,32,47]. Our results showed that complex disease genes had higher D values than Mendelian disease genes.

Figure 4 exhibits the genomic landscape of molecular features across the different genetic categories. Through introduction of the concept of per-person mutability, clear separation of the three geno-phenotypic categories of genes is demonstrated (Figure 4D). Mendelian disease genes were characterized by rare frequency, slow evolution and highly functional, thereby being highly intolerant to variations in a personal genome. Common complex disease genes were characterized by the highest per-person mutability with fewer singletons, allowing the multiple co-existence of many neutral variants that were tolerated. Pharmacogenes were characterized by a moderate number of

singletons, moderate per-person mutability and the largest number of paralogs. The present study thus demonstrated the manner in which we understand and identify meaningful molecular genetic features and apply them to develop robust gene-level methods to differentiate PGx, common complex disease and rare Mendelian disease genes in the context of comprehensive genomic characterizations.

Correlation of GVB-derived interindividual variability with the number of paralogs

According to the prevailing hypothesis that genes with a larger number of paralogs are under looser evolutionary constraints [48], we assumed that the genetic variability among individuals increased with the number of paralogs through the accumulation of mutations. Consistent with this, we found that the coefficient of variation for the 2504 individuals of the 1000 Genomes Project positively correlated with the number of paralogs (Kendall's $\tau = 0.085$, $p < 0.0001$, Supplementary Figure 5). This suggested that GVB might be used as a quantitative measure for the effect of an allelic function that is significantly correlated with the evolutionary potential of a genetic variation.

Discussion

Proper population-based gene-level scores for all genes for particular ethnic or demographic groups are not yet readily available. This difficulty may prohibit the wide utility of the previous population-based gene-level approaches. In contrast to the leading gene-level methods including RVIS and GDI, GVB constitutes an unbiased method to evaluate the genomic architectures of populations, as this method assigns a gene-level score for each gene without using population data.

Particularly in the fields of pharmacogenetics and pharmacogenomics, the associated interindividual and interethnic genetic variability tends to be much larger than that of Mendelian disease genetics. Genetic variability thus constitutes a main factor contributing to the variabilities of responses to many drugs among different individuals and ethnic groups, which can lead to severe adverse drug reactions and poor efficacies. Consistent with this, interethnic variability, a phenomenon long been considered to be inherently unpredictable, was reported to be a strong predictor of pharmaceutical market withdrawals [10]. Conversely, when bringing drugs to market, all known risk factors are highly regulated through the clinical trial I to III phases for comprehensive evaluations, with the exception of the interindividual genetic variability that is well captured by GVB. We therefore recently evaluated GVB in direct comparison to the standard clinical pharmacogenetics of thiopurine metabolism by *TPMT* and *NUDT15* genes in pediatric patients with acute lymphoblastic leukemia. Patients at high risk of 6-mercaptopurine-induced adverse reactions were identified using GVB by integrating a number of clinically well-established gene effects. Notably, we successfully demonstrated that the GVB method is comparable or even superior to the best established pharmacogenetics of thiopurine therapy based on traditional diplotype-based methods [13].

In the present study, we found that GVB complemented RVIS and GDI, the leading gene-level methods focusing on Mendelian genetics. This finding reflects the distinct genomic characteristics of rare Mendelian- and common complex disease genes [4,42,49–51]. The genetic background of complex disease genes may be more similar to that of pharmacogenes than Mendelian disease genes, as chronic diseases are influenced by a complex combination of genetic effects, environmental stimuli and lifestyle factors [52], similar to the phenotypic consequences in drug metabolism that occur when externally triggered by a stimulus, such as a drug. The disease burdens of common complex diseases are obviously much greater than those of rare Mendelian diseases. Nevertheless, further fine tuning of GVB for specific common complex disease categories considering the underlying genetic architecture of each common disease category is necessary to fully exploit this methodology.

For the nondisease gene set, unexpected distributions of genetic features, such as the lowest number of paralogs and per-person mutability or a relatively high degree of PPI were shown. These results are consistent with previous studies reporting that the nondisease category contains genes inducing nonviable phenotypes with functional disruptions, which do not present disease phenotypes by definition [53]. Nonviable genes showed a pattern of genomic characteristics that was very similar to that of Mendelian disease genes (Figures 3 & 4). However, nonviable genes had fewer paralogs, more singletons and lower per-person mutability than Mendelian disease genes (Figure 3), suggesting that they are under stronger selective pressure with a higher variability of rare variants that have not survived to an extent allowing frequent occurrence in individual genomes. Overall, nonviable genes appeared to exhibit a slight bias toward extreme values compared with those of Mendelian disease genes (Figures 3 & 4).

Comprehensive genomic characterization of molecular features displayed clear separation of the three geno-phenotypic categories of genes. Common complex disease genes exhibited smaller number of singletons (x-axis) and

the highest protein complexity (dot size) with modest number of paralogs (y-axis). Pharmacogenes showed the most paralogs with moderate protein complexity and singletons (Figure 4A). The smaller number of singletons reflects that complex disease genes accumulate environmentally well-adapted neutral variants in disordered regions (higher protein complexity) under stronger selective pressure compared with pharmacogenes. Figure 4B demonstrates that Mendelian disease genes exhibit long CDS lengths (x-axis), with high PPI (dot size), and few paralogs (y-axis). Pharmacogenes have the shortest average CDS length, low PPI and the most paralogs. CDS length has been reported as important evidence of evolution, such that longer genes evolve slowly as they cost more to duplicate and this renders them more deleterious upon overexpression [38,39]. Figure 4C demonstrates that, as expected, complex disease genes exhibit low PPI (dot size), short CDS length (x-axis) and the lowest number of singletons (y-axis).

Genomic characteristics for the genetic categories of GAD psych and GAD developmental were similar to those of Mendelian disease genes, harboring higher number of singletons and longer CDS lengths (Supplementary Figure 4). This finding is consistent with the previous studies reporting that the high heritability of developmental disorders and major psychiatric conditions resemble the pattern expected from Mendelian-like genes [54] although the complex natures of those diseases (e.g., the strength and specificity of association and nongenetic inheritance) do not necessarily obey the rules of simple Mendelian inheritance [55,56]. Similarly, PGx PD Targets and Efficacy categories also exhibited a relatively small number of paralogs and a large number of singletons, similarly to Mendelian disease genes. This result is consistent with the previous studies that target genes of successful inhibitors are intolerant to the functional variants as extreme as known haploinsufficient genes [57,58]. The GAD pharmacogenomic subcategory showed a similar pattern with the Pharmacogene category, as expected. GAD chemdependency, represented by alcohol or substance abuse/addiction, also showed a similar pattern with that of Pharmacogenes.

The present study mainly described the results using the SIFT algorithm, which yielded the most robust predictions (Supplementary Table 4). When using the optimal threshold value to determine tolerable variation (SIFT >0.7), a reliable and high level of performance of SIFT was consistently replicated [34]. GVBs composed of other *in silico* variant prediction methods other than the SIFT algorithm performed well with unique profiles of advantages and limitations (i.e., CADD, PolyPhen2 HIVD, PolyPhen2 HVAR, PhyloP, MutationTaster, and GERP++; Supplementary Figure 2).

It has long been assumed that genes from different genetic backgrounds inherit different genomic characteristics. To the authors' best knowledge, the present study is the first to systematically explore this historical intuition in the context of PGx, common complex disease and rare Mendelian disease together with nonviable and nondisease gene categories and subcategories. GVB also provides a genetic condition-specific scheme for molecular genetic-feature optimizations. However, caution is needed in interpreting the effect of aggregate variables. Future work should apply much more sophisticated schemes based on various types of genetic features under a variety of genetic scenarios. Further investigation of large-scale genomic DNA changes such as copy number variations or subdivided regions such as protein domains and exons is necessary to improve and facilitate proper clinical interpretations

Summary points

- Current gene-level approaches including the residual variation intolerance score (RVIS) and the gene damage index (GDI) use a population dataset to find pathogenic genes by assigning a score for each gene for a given population. Thus, they provide population-dependent and collective evaluation scores that are not individualized.
- Gene-wise variant burden (GVB) assigns a score for each gene for each individual in a population-independent and individual-specific manner, which is necessary for clinical applications.
- GVB outperformed RVIS and GDI in predicting pharmacogenetic (PGx) genes and common complex disease genes, while RVIS and GDI outperformed for rare Mendelian disease gene predictions.
- Evolutionary conservation is different between variants within genes associated with congenital diseases that are under purifying selection and those which are not congenital-disease associated, such as most pharmacogenes and most common complex disease genes in their later lives. Different strategies are needed for pharmacogenicity versus early and late pathogenicity investigations.
- GVB performance was further improved by molecular genetic features including numbers of paralogs and singletons, per-person mutability, protein interaction degrees, gene length, selective pressure, and protein complexity.
- The GVB score provides fully individualized gene-level PGx scores to be applied to predict subgroups vulnerable to severe ADRs for certain drugs.

of the findings [59]. GVB considers allelic dosage effects, and the potential dosage effects potentially created by duplications and deletions can further be considered.

GVB obtained by integrating variant deleteriousness scores *in silico* of the multitude of variants of a given gene better predicted PGx and complex disease genes than the leading gene-level prioritization methods, RVIS and GDI. GVB adjusted for genetic molecular features further improved genomic characterizations and complemented RVIS and GDI for prioritizing 'disease-causing' rare mutations. Moreover, GVB not only successfully differentiated but also revealed the genomic landscape of the comprehensive genetic categories and subcategories of PGx, complex disease, rare disease, neutral (nondisease) and lethal genes. The insights presented here should aid personal–genome interpretations in the context of pharmacogenetics and common- and rare-disease phenotypes in the era of personal genomics.

Supplementary data

To view the supplementary data that accompany this paper please visit the journal website at: www.futuremedicine.com/doi/suppl/10.2217/pgs-2020-0039

Author contributions

All authors designed the model and the framework. Y Park and H Seo analyzed the data and carried out the implementation. Y Park and JH Kim wrote the manuscript. BY Ryu edited the manuscript. JH Kim conceived the study and was in charge of overall direction and planning. All authors read and approved the final manuscript.

Financial & competing interests disclosure

This research was supported by grants from the Ministry of Food and Drug Safety in 2019 (no. 16183MFDS541) and the Korean Health Technology R&D Project by the Ministry of Health and Welfare in the Republic of Korea (no. HI18C2386). The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

References

1. Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet* 9(8), e1003709 (2013).
2. Itan Y, Shang L, Boisson B *et al.* The human gene damage index as a gene-level approach to prioritizing exome variants. *Proc. Natl Acad. Sci. USA* 112(44), 13615–13620 (2015).
3. Stessman HA, Bernier R, Eichler EE. A genotype-first approach to defining the subtypes of a complex disease. *Cell* 156(5), 872–877 (2014).
4. Blekhman R, Man O, Herrmann L *et al.* Natural selection on genes that underlie human disease susceptibility. *Curr. Biol.* 18(12), 883–889 (2008).
5. Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L. Natural selection has driven population differentiation in modern humans. *Nat. Genet.* 40(3), 340–345 (2008).
6. Hovelson DH, Xue Z, Zawistowski M *et al.* Characterization of ADME gene variation in 21 populations by exome sequencing. *Pharmacogenet. Genomics* 27(3), 89–100 (2017).
7. Li J, Zhang L, Zhou H, Stoneking M, Tang K. Global patterns of genetic diversity and signals of natural selection for human ADME genes. *Hum. Mol. Genet.* 20(3), 528–540 (2011).
8. Landrum MJ, Lee JM, Riley GR *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* 42(Database issue), D980–D985 (2014).
9. Richards S, Aziz N, Bale 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(5), 405–424 (2015).
10. Lee KH, Baik SY, Lee SY, Park CH, Park PJ, Kim JH. Genome sequence variability predicts drug precautions and withdrawals from the market. *PLoS ONE* 11(9), e0162135 (2016).
11. Haynes WA, Tomczak A, Khatri P. Gene annotation bias impedes biomedical research. *Sci. Rep.* 8(1), 1362 (2018).
12. Seo H, Kwon EJ, You YA *et al.* Deleterious genetic variants in ciliopathy genes increase risk of ritodrine-induced cardiac and pulmonary side effects. *BMC Med. Genom.* 11(1), 4 (2018).
13. Park Y, Kim H, Choi JY *et al.* Star allele-based haplotyping versus gene-wise variant burden scoring for predicting 6-mercaptopurine intolerance in pediatric acute lymphoblastic leukemia patients. *Frontiers Pharmacol* 10, 654 (2019).

14. Lee KH, Kim SH, Kim CH *et al.* Identifying genetic variants underlying medication-induced osteonecrosis of the jaw in cancer and osteoporosis: a case control study. *J. Transl. Med.* 17(1), 381 (2019).
15. 1000 Genomes Project Consortium; Auton A, Brooks LD *et al.* A global reference for human genetic variation. *Nature* 526(7571), 68–74 (2015).
16. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 38(16), e164 (2010).
17. Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. *Genome Res.* 11(5), 863–874 (2001).
18. Ng PC, Henikoff S. Accounting for human polymorphisms predicted to affect protein function. *Genome Res.* 12(3), 436–446 (2002).
19. Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* 31(13), 3812–3814 (2003).
20 . Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* 46(3), 310–315 (2014).
21. Adzhubei IA, Schmidt S, Peshkin L *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* 7(4), 248–249 (2010).
22. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* 20(1), 110–121 (2010).
23. Schwarz JM, Rödelsperger C, Schuelke M, Seelow D. Mutation taster evaluates disease-causing potential of sequence alterations. *Nat. Methods* 7(8), 575–576 (2010).
24. Cooper GM, Stone EA, Asimenos G *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* 15(7), 901–913 (2005).
25. Gong L, Owen RP, Gor W, Altman RB, Klein TE. PharmGKB: an integrated resource of pharmacogenomic data and knowledge. *Curr. Prot. Bioinformatics* 14(7), (2008).
26. Wishart DS, Knox C, Guo AC *et al.* DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* 36(Database issue), D901–906 (2008).
27. Becker KG, Barnes KC, Bright TJ, Wang SA. The genetic association database. *Nat. Genet.* 36(5), 431–432 (2004).
28. Hamosh A, Scott AF, Amberger J, Bocchini C, Valle D, Mckusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* 30(1), 52–55 (2002).
29. Bartha I, Di Iulio J, Venter JC, Telenti A. Human gene essentiality. *Nat. Rev. Genet.* 19(1), 51–62 (2018).
30. Robin X, Turck N, Hainard A *et al.* pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinform.* 12, 77 (2011).
31. Stoletzki N, Eyre-Walker A. Estimation of the neutrality index. *Mol. Biol. Evol.* 28(1), 63–70 (2011).
32. Cha S. Comprehensive survey on distance/similarity measures between probability density functions. *Int. J. Math Model Methods Appl. Sci.* 1(4), 300–307 (2007).
33. Hermjakob H, Montecchi-Palazzi L, Lewington C *et al.* IntAct: an open source molecular interaction database. *Nucleic Acids Res.* 32(Database issue), D452–455 (2004).
34. Park J, Lee SY, Baik SY *et al.* Gene-wise burden of coding variants correlates to noncoding pharmacogenetic risk variants. *Int. J. Mol. Sci.* 21(9), (2020).
35. Doyle MA, Gasser RB, Woodcroft BJ, Hall RS, Ralph SA. Drug target prediction and prioritization: using orthology to predict essentiality in parasite genomes. *BMC Genom.* 11, 222 (2010).
36. Gu X. Evolution of duplicate genes versus genetic robustness against null mutations. *Trends Genet.* 19(7), 354–356 (2003).
37. Conant GC, Wagner A. Duplicate genes and robustness to transient gene knock-downs in *Caenorhabditis elegans*. *Proc. Biol. Sci.* 271(1534), 89–96 (2004).
38. O'Toole AN, Hurst LD, Mclysaght A. Faster evolving primate genes are more likely to duplicate. *Mol. Biol. Evol.* 35(1), 107–118 (2018).
39. Ma L, Pang CN, Li SS, Wilkins MR. Proteins deleterious on overexpression are associated with high intrinsic disorder, specific interaction domains, and low abundance. *J. Proteome Res.* 9(3), 1218–1225 (2010).
40. McDonald JH, Kreitman M. Adaptive protein evolution at the ADH locus in *Drosophila*. *Nature* 351(6328), 652 (1991).
41. Osada N, Mano S, Gojobori J. Quantifying dominance and deleterious effect on human disease genes. *Proc. Natl Acad. Sci USA* 106(3), 841–846 (2009).
42. Jin W, Qin P, Lou H, Jin L, Xu S. A systematic characterization of genes underlying both complex and Mendelian diseases. *Hum. Mol. Genet.* 21(7), 1611–1624 (2012).
43. Eyre-Walker YC, Eyre-Walker A. The role of mutation rate variation and genetic diversity in the architecture of human disease. *PLoS ONE* 9(2), e90166 (2014).

44. López-Bigas N, Ouzounis CA. Genome-wide identification of genes likely to be involved in human genetic disease. *Nucleic Acids Res.* 32(10), 3108–3114 (2004).
45. Lu HC, Chung SS, Fornili A, Fraternali F. Anatomy of protein disorder, flexibility and disease-related mutations. *Front. Mol. Biosci.* 2, 47 (2015).
46. Walter J, Charon J, Hu Y *et al.* Comparative analysis of mutational robustness of the intrinsically disordered viral protein VPg and of its interactor eIF4E. *PLoS ONE* 14(2), e0211725 (2019).
47. Wootton JC, Federhen S. Statistics of local complexity in amino acid sequences and sequence databases. *Comput. Chem.* 17(2), 149–163 (1993).
48. Shakhnovich BE, Koonin EV. Origins and impact of constraints in evolution of gene families. *Genome Res.* 16(12), 1529–1536 (2006).
49. Thomas PD, Kejariwal A. Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: evolutionary evidence for differences in molecular effects. *Proc. Natl Acad. Sci. USA* 101(43), 15398–15403 (2004).
50. Cooper DN, Mort M. Do inherited disease genes have distinguishing functional characteristics? *Genet. Test Mol. Biomarkers* 14(3), 289–291 (2010).
51. Podder S, Ghosh TC. Exploring the differences in evolutionary rates between monogenic and polygenic disease genes in human. *Mol. Biol. Evol.* 27(4), 934–941 (2010).
52. Craig J. Complex diseases: research and applications. *Nat. Educat.* 1, 184 (2008).
53. Chen WH, Zhao XM, van Noort V, Bork P. Human monogenic disease genes have frequently functionally redundant paralogs. *PLoS Comput. Biol.* 9(5), e1003073 (2013).
54. Gandal MJ, Leppa V, Won H, Parikshak NN, Geschwind DH. The road to precision psychiatry: translating genetics into disease mechanisms. *Nat. Neurosci.* 19(11), 1397–1407 (2016).
55. Toth M. Mechanisms of non-genetic inheritance and psychiatric disorders. *Neuropsychopharmacology* 40(1), 129–140 (2015).
56. Kendler KS. "A gene for. . . ": the nature of gene action in psychiatric disorders. *Am. J. Psychiatry* 162(7), 1243–1252 (2005).
57. Minikel EV, Karczewski KJ, Martin HC *et al.* Evaluating potential drug targets through human loss-of-function genetic variation. *BioRxiv* 530881 (2019).
58. Fuselli S. Beyond drugs: the evolution of genes involved in human response to medications. *Proc. Biol. Sci.* 286(1913), 20191716 (2019).
59. Gussow AB, Petrovski S, Wang Q, Allen AS, Goldstein DB. The intolerance to functional genetic variation of protein domains predicts the localization of pathogenic mutations within genes. *Genome Biol.* 17, 9 (2016).