The human gene damage index as a gene-level approach to prioritizing exome variants

Yuval Itan[a,1], Lei Shang[a], Bertrand Boisson[a], Etienne Patin[b,c], Alexandre Bolze[a], Marcela Moncada-Vélez[a,d], Eric Scott[e], Michael J. Ciancanelli[a], Fabien G. Lafaille[a], Janet G. Markle[a], Ruben Martinez-Barricarte[a], Sarah Jill de Jong[a], Xiao-Fei Kong[a], Patrick Nitschke[f], Aziz Belkadi[g,h], Jacinta Bustamante[a,g,h,i], Anne Puel[g,h], Stéphanie Boisson-Dupuis[a,g,h], Peter D. Stenson[j], Joseph G. Gleeson[k,l,m], David N. Cooper[j], Lluis Quintana-Murci[b,c,2], Jean-Michel Claverie[n,2], Shen-Ying Zhang[a,g,h,3], Laurent Abel[a,g,h,3], and Jean-Laurent Casanova[a,g,h,m,o,1,3]

[a]St. Giles Laboratory of Human Genetics of Infectious Diseases, Rockefeller Branch, The Rockefeller University, New York, NY 10065; [b]Human Evolutionary Genetics Unit, Institut Pasteur, 75015 Paris, France; [c]Centre National de la Recherche Scientifique, CNRS URA 3012, Paris, France; [d]Group of Primary Immunodeficiencies, Faculty of Medicine, University of Antioquia UdeA, Medellín, Colombia; [e]Neurogenetics Laboratory, Department of Neurosciences, University of California, San Diego, CA 92093-0662; [f]Bioinformatics Platform, University Paris Descartes, 75015 Paris, France; [g]Laboratory of Human Genetics of Infectious Diseases, Necker Branch, INSERM U.1163, Necker Hospital for Sick Children, 75015 Paris, France; [h]Paris Descartes University, Imagine Institute, 75015 Paris, France; [i]Center for Study of Primary Immunodeficiencies, Necker Hospital for Sick Children, Paris, France; [j]Institute of Medical Genetics, Cardiff University, Cardiff CF14 4XN, United Kingdom; [k]Laboratory of Pediatric Brain Disease, The Rockefeller University, New York, NY 10065; [l]New York Genome Center, New York, NY 10013; [m]Howard Hughes Medical Institute, New York, NY 10065; [n]APHM & Structural and Genomic Information Laboratory, UMR7256, CNRS Aix-Marseille University, 13288 Marseille, France; and [o]Pediatric Hematology-Immunology Unit, Necker Hospital for Sick Children, 75015 Paris, France

Contributed by Jean-Laurent Casanova, September 22, 2015 (sent for review June 26, 2015); reviewed by Jay Shendure and David B. Goldstein

The protein-coding exome of a patient with a monogenic disease contains about 20,000 variants, only one or two of which are disease causing. We found that 58% of rare variants in the protein-coding exome of the general population are located in only 2% of the genes. Prompted by this observation, we aimed to develop a gene-level approach for predicting whether a given human protein-coding gene is likely to harbor disease-causing mutations. To this end, we derived the gene damage index (GDI): a genome-wide, gene-level metric of the mutational damage that has accumulated in the general population. We found that the GDI was correlated with selective evolutionary pressure, protein complexity, coding sequence length, and the number of paralogs. We compared GDI with the leading gene-level approaches, genic intolerance, and de novo excess, and demonstrated that GDI performed best for the detection of false positives (i.e., removing exome variants in genes irrelevant to disease), whereas genic intolerance and de novo excess performed better for the detection of true positives (i.e., assessing de novo mutations in genes likely to be disease causing). The GDI server, data, and software are freely available to noncommercial users from lab.rockefeller.edu/casanova/GDI.

mutational damage | gene-level | gene prioritization | variant prioritization | next generation sequencing

Germ-line mutations can contribute to the long-term adaptation of humans, but at the expense of causing a large number of genetic diseases (1). The advent of next-generation sequencing (NGS)-based approaches, including whole-exome sequencing (WES), whole-genome sequencing (WGS), and RNA-Seq, has facilitated the large-scale detection of gene variants at both the individual and population levels (2–6). In patients suffering from a monogenic disease, at most two variants are disease causing [true positives (TP)], and the other 20,000 or so protein-coding exome variants are false positives (FP; type I error). Several variant-level metrics predicting the biochemical impact of DNA mutations (7–9) can be used to prioritize candidate variants for a phenotype of interest (10, 11). Gene-level metrics aim to prioritize the genes themselves, providing information that can be used for the further prioritization of variants. There are currently fewer gene-level than variant-level computational methods. They provide complementary information, as it is best to predict the impact of a variant by also taking into account population genetics data for its locus. Current gene-level methods include genic intolerance, as measured by the residual variation intolerance score (RVIS) (12) and de novo excess (DNE) (13). These metrics are particularly useful for determining whether a given gene (and, by inference, its variants) is a plausible candidate for involvement in a particular genetic disease (i.e., for the selection of a short list of candidate genes and variants, which include the TPs). However, owing to the large number and diversity of variants, the selection of a single candidate gene from the NGS data for a given patient with a specific disease remains challenging.

We reasoned that genes frequently mutated in healthy populations would be unlikely to cause inherited and rare diseases, but would probably make a disproportionate contribution to the variant calls observed in any given patient. Conversely, mutations in genes that are never or only rarely mutated under normal circumstances are more likely to be disease-causing. Leading gene-level strategies are based on selective pressure (12) and de novo mutation rate estimates (13). These methods are tailored to detect genes likely to harbor TPs. However, these methods do not directly calculate quantitatively the mutational load for human genes in the general (i.e., "healthy") population or the frequencies of mutant alleles. These methods may, therefore, not be optimal for filtering out highly mutated genes, which are likely

Significance

The protein-coding exome of a patient with a monogenic disease contains about 20,000 variations, of which only one or two are disease causing. When attempting to select disease-causing candidate mutation(s), a challenge is to filter out as many false-positive (FP) variants as possible. In this study, we describe the gene damage index (GDI), a metric for the nonsynonymous mutational load in each protein-coding gene in the general population. We show that the GDI is an efficient gene-level method for filtering out FP variants in genes that are highly damaged in the general population.

Author contributions: Y.I., B.B., A. Bolze, J.G.M., R.M.-B., S.B.-D., D.N.C., L.Q.-M., J.-M.C., S.-Y.Z., L.A., and J.-L.C. designed research; Y.I., L.S., E.P., M.M.-V., E.S., M.J.C., F.G.L., S.J.d. J., X.-F.K., J.B., A.P., P.D.S., J.G.G., and S.-Y.Z. performed research; Y.I., B.B., E.P., M.M.-V., E.S., J.G.M., S.J.d.J., P.N., J.B., S.B.-D., P.D.S., J.G.G., D.N.C., L.Q.-M., J.-M.C., S.-Y.Z., and L.A. contributed new reagents/analytic tools; Y.I., L.S., E.P., E.S., P.N., A. Belkadi, and P.D.S. analyzed data; and Y.I., A. Bolze, M.J.C., R.M.-B., X.-F.K., L.Q.-M., J.-M.C., S.-Y.Z., L.A., and J.-L.C. wrote the paper.

Reviewers: J.S., University of Washington; D.B.G., Columbia University.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

[1]To whom correspondence may be addressed. Email: yitan@rockefeller.edu or casanova@rockefeller.edu.

[2]L.Q.-M. and J.-M.C. contributed equally to this work.

[3]S.-Y.Z., L.A., and J.-L.C. contributed equally to this work.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1518646112/-/DCSupplemental.

to harbor many FPs. Moreover, there has been no formal comparison of the power of these gene-level methods and their combinations for maximizing the discovery of FPs and TPs by NGS. We therefore aimed to generate a robust metric of the cumulative mutational damage to each human protein-coding gene, to make it easier to distinguish the FP variants harbored by highly damaged genes (e.g., under relaxed constraint or positive selection) from potential candidate genes and variants, including the TPs. By damaged genes, we refer to genes displaying many nonsynonymous mutations, which are not necessarily damaging biochemically or evolutionarily. We developed the gene damage index (GDI), which defines, in silico, the mutational damage accumulated by each protein-coding human gene in the general population, and reflecting the combined influences of drifts and selections. We then tested this approach with the WES data for 84 patients in our in-house database, each of these patients having a known primary immunodeficiency (PID). Finally, we used receiver operating characteristic (ROC) curves for formal comparisons of performance between GDI and the existing gene-level RVIS and DNE approaches, and to assess the power of the gene-level methods for detecting enrichment in de novo mutations in cases versus controls. We also tested whether these methods could act in synergy to filter out FPs and select TPs.

Results

Genes Found to Be Highly Mutated in Patients with Monogenic Diseases. We hypothesized that the genes most frequently mutated in the general population would also probably be the genes most frequently mutated in patients with monogenic diseases. They would therefore be unlikely to cause disease, despite contributing a large proportion of the variants harbored by patients. As highly mutated genes form a large proportion of the total number of variants present in patients, they would also be expected to account for a large proportion of the FP variants in patients. We therefore set out to identify a set of highly mutated FP genes from the WES data derived from 84 patients with monogenic PIDs for whom the true disease-causing mutation had already been experimentally validated and reported (1). We used standard variant filtering methods for rare diseases, retaining only those with a minor allele frequency (MAF) <0.01 in the 1,000 Genomes Project, together with missense, nonsense, and frameshift variants (including start-loss and stop-loss), and in-frame indels and splice variants for which sequencing quality was high (see *Materials and Methods* for further details) (4, 14). We then determined the frequency of the remaining 44,668 variants among the 11,190 genes harboring them. By an analysis of outliers based on modified *Z*-score, we identified 496 genes carrying significantly larger numbers of variants than expected (≥10 per gene). Despite accounting for only 2.42% of all human protein-coding genes, these 496 genes harbored 58.32% of all of the rare variants found in patients. An approach that efficiently filters out highly damaged human genes should therefore efficiently eliminate a large proportion of the FP variants in patients. We would expect to obtain similar results with patients suffering from more common monogenic disorders, because the number of disease-causing mutations would remain negligible with respect to the total number of variants in these patients. Fig. S1 shows the patients' WES genes scaled by the number of variants they harbor.

The Human GDI. Following up on this observation, we defined the GDI as the cumulative mutational damage to a given human gene in the general population (the 1,000 Genomes Project). We chose to use the combined annotation dependent depletion (CADD) score as the variant-level damage prediction metric (9), because (*i*) it has been shown to be the best method for distinguishing between deleterious and benign variants and (*ii*) unlike methods such as PolyPhen-2 and SIFT, which predict only missense variants (7, 8), it can be used to assess the impact of most types of variant (9). We showed an inverse relationship between MAF and the CADD damage prediction value: rare variants tended to have higher CADD scores than common variants from the 1,000 Genomes Project (Fig. S2*A*). We then compared the performance of four heuristic gene-level GDI models, with the purpose of maximizing the differentiation of FPs from TPs (see the section on ROC curves and the Materials and Methods for further details and equations): (*i*) "raw" GDI, calculated for each human gene by first multiplying each variant's CADD score by the corresponding variant's number of alleles in the 1,000 Genomes Project (a total of 610,160 missense/nonsense/frameshift/in-frame indels/splice variants, with a MAF < 0.5, from a total of 20,243,313 alleles), then summing up all (CADD × allele count) products for one gene; (*ii*) the "CADD-normalized" gene-level model of accumulated mutational damage, calculated as in *i*, with each CADD score divided by the expected (median) CADD score of a variant with a similar allele frequency (Fig. S2*A*); (*iii*) the observed/expected GDI-normalized GDI model, in which the observed GDI was calculated as in *i* and then divided by the expected raw GDI, calculated as in *i* and using the expected CADD as in *ii*; and (*iv*) the "gene size-normalized" GDI model, calculated as in ref. 1 and divided by the length of the coding sequence (CDS) of the canonical transcript of the gene. We found that the GDI models that performed best under a general, autosomal dominant (AD), or autosomal recessive (AR) mode of inheritance were *i* and *ii*, with model *ii* outperforming *i* in all cases, as shown by calculations of the area under the curve (AUC; Fig. S2 *B–D*). We therefore used model *ii*, representing accumulated mutational damage normalized by dividing by the expected CADD score for this study. The GDI scores and their Phred-scaled GDI scores for 19,558 human protein genes are summarized in Dataset S1, Tab S1.

Definition of the Most and Least Damaged Human Genes. We hypothesized that mutations in the genes most damaged in healthy individuals are unlikely to be responsible for monogenic diseases, whereas mutations in the least damaged genes are more likely to be associated with the most severe monogenic disorders (or, alternatively, would be embryo-lethal). We therefore characterized the functional attributes of the human genes with the highest and lowest GDI values. By calculating the outliers from the gene damage data for all 19,558 protein-coding human genes, we defined the 751 human genes with the highest GDI values, on the basis of modified *Z*-score outliers for GDI. The 977 genes with the lowest GDI values were defined as those not displaying any nonsynonymous variation in healthy individuals from the 1,000 Genomes Project (although some will probably have nonsynonymous variants in larger databases such as ExAC (exac.broadinstitute.org), or when different annotation software was used. Owing to the gamma distribution of the data, there are no trivial statistical outliers at the lower end of the range (Fig. S3).

Characterization of the Most and Least Damaged Genes. We found that biological proximity, as predicted by the human gene connectome, was greater among high-GDI genes and among low-GDI genes than for randomly selected human genes ($P < 1.0 \times 10^{-5}$ for both sets) (15–17). This biological proximity suggests that high-GDI genes are functionally related to each other, as are low-GDI genes. We further performed biological ontology and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway functional enrichment analyses for the human genes with the highest and lowest GDI values (18–20). We found that the list of genes with the highest GDI values was strongly enriched in sensory perception genes ($P = 1.00 \times 10^{-23}$), including, in particular, the genes of the olfactory receptor superfamily (~400 protein-coding genes), which has been shown to be under positive selection constraints in the human lineage (21, 22). The list of genes with the lowest GDI values was enriched in ribosome, chemokine signaling, proteasome, and spliceosome genes, all of which are highly conserved in species lineages predating the emergence of vertebrates (23–26) (Dataset S1, Tab S2 and Figs. S4 and S5). The GDI can therefore be considered to be a surrogate indicator of the relative biological indispensability (low GDI) or redundancy (high GDI) of a given human gene.

Correlation of GDI with Selective Pressure and Number of Paralogs. We found that genes with a low GDI tended to be under purifying selection stronger than the median selective pressure acting on human genes, as ascertained by the estimated McDonald–

Kreitman neutrality index (NI; $P = 1.79 \times 10^{-36}$), whereas genes with a high GDI tended to be under less purifying selective pressure than the median human gene ($P = 3.14 \times 10^{-74}$). These results are plausible, because CADD score (used here as the basic metric for GDI calculation) is strongly dependent on evolutionary conservation (9). We also demonstrated a strong association between gene damage and the number of paralogs (27, 28): genes with high GDI had significantly larger numbers of paralogs than human genes generally ($P = 3.96 \times 10^{-11}$), whereas genes with low GDI had significantly smaller numbers of paralogs than human genes generally ($P = 2.98 \times 10^{-21}$) (21, 29). These results lend support to Ohno's neo/pseudofunctionalization hypothesis, according to which, following a gene duplication event, one copy is freed from evolutionary constraints and can therefore evolve a new function (or alternatively become a pseudogene) through the accumulation of mutations (30–32).

Correlation of GDI with Protein Complexity and Coding Sequence Length. We then investigated the possible association between the complexity of protein amino acid composition D and GDI. Low D values are associated with proteins with an amino acid composition similar to that of the average human protein, whereas high D values suggest a biased amino acid composition in the protein concerned, generally a disordered low-complexity protein. We found that genes with a low GDI had a significantly higher than normal D (i.e., low complexity, biased amino acid composition with respect to the median composition of human proteins; $P = 2.80 \times 10^{-54}$), whereas genes with a high GDI had low D values (i.e., a relatively unbiased amino acid composition; $P = 0.04$). These results are consistent with previous studies reporting a correlation between positive selection and protein disorder/complexity bias (33, 34). As expected, we found that genes with a low GDI had significantly shorter CDS than the median value for all genes ($P = 4.19 \times 10^{-152}$), whereas genes with a high GDI had significantly longer CDS ($P = 5.25 \times 10^{-35}$). Overall, these correlations reveal the existence of significant associations between GDI and various molecular properties of human genes and proteins.

Genes Carrying FP Variants Have Higher Than Normal GDI. We next assessed whether the GDI was an appropriate tool for filtering out genes with abundant FP variants. We first compared the GDI values of known Mendelian disease-causing genes (TP) defined by OMIM (35), for which the mode of inheritance was unambiguous, with those of genes harboring FP variants from the WES for the 84 PID patients described above. We performed three types of comparison with account for the different modes of inheritance and the corresponding candidate variants in patients: (*i*) all 1,217 Mendelian disease-causing genes vs. the genes harboring all 44,668 FP variants in patients; (*ii*) 375 AD Mendelian disease-causing genes vs. genes harboring 42,863 heterozygous FP variants in patients; and (*iii*) 585 AR Mendelian disease-causing genes vs. genes harboring 1,805 homozygous FP variants in the patients. We used one gene instance for each gene containing one or more variants in a patient (for example, a gene harboring 27 variants in 25 patients under a specific model had 25 instances of its GDI value in the specific analysis). We did not perform comparisons with X-linked genes, due to overall poor coverage in the WES data for the patients. We found that, in each of the three comparisons, the GDI values for genes with FP variants were significantly higher than those for the corresponding TP OMIM genes, both when comparing TPs with all FPs (Fig. 1; $P < 1.0 \times 10^{-200}$, $P = 2.08 \times 10^{-101}$, and $P = 4.68 \times 10^{-197}$ for tests *i*, *ii*, and *iii*, respectively) and then comparing TPs with FPs for the 58.32% of rare FP variants harbored by the 2.42% most mutated genes (Fig. S6; $P < 1.0 \times 10^{-200}$, $P = 7.44 \times 10^{-92}$, and $P = 3.75 \times 10^{-155}$ for the above tests *i*, *ii*, and *iii*, respectively). We did not exclude OMIM genes that were not validated as disease causing in the specific patients from the FP sets in any of the tests. We performed bootstrap simulations for comparisons of TPs and FPs (Fig. 2 and Fig. S7), with random sampling from the TP set to assess its validity as a predictor. We confirmed that the GDI of genes containing FPs in the patients was much higher than the GDI of disease-causing genes, for both highly mutated genes and for all genes. These results suggested that the GDI might be useful for filtering out a large proportion of the variants in genes that are unlikely to be disease causing.

Performance Assessment by ROC Curve Analysis. We assessed the performance of GDI for differentiating between disease-causing genes (TP, see above) and non–disease-causing genes (FPs, genes harboring variants detected in the patients but not responsible for disease; see above), by formally comparing the GDI with the raw scores obtained by two state-of-the-art gene-level approaches: genic intolerance (RVIS) (12) and DNE (13). Briefly, the RVIS approach ranks human genes in terms of the strength and consistency of the purifying selection acting against functional variation of the gene, whereas DNE estimates the rate of de novo mutation on a per-gene basis, globally and per gene set. With the aim of maximizing performance, we also tested the four possible combinations, GDI+RVIS, GDI+DNE, GDI+RVIS+DNE, and RVIS+DNE (see *Materials and Methods* for details regarding the integration of the methods into single scores). Using ROC curves, we demonstrated that GDI had the best performance of the three standalone methods under a general model and under models of AD or AR inheritance, for comparisons both of TPs with FPs for the 58.32% FP variants present in the 2.42% most mutated genes, and of TPs with all FPs (sensitivity and specificity, respectively; Fig. 3 and Fig. S8; see Dataset S1, Tab S3 for all AUC values). GDI+RVIS had the best performance of the four combinations of methods for all modes of inheritance. Of the six conditions tested, GDI+RVIS also outperformed GDI as a standalone method for the set of AR variants in frequently mutated genes. This analysis suggested that GDI and RVIS captured different sets of complementary information from the population genetics data.

Performance Assessment by Hot Zone Analyses. We then tested the performance of GDI and the combinations presented above to estimate enrichment in de novo mutations hypothesized as damaging

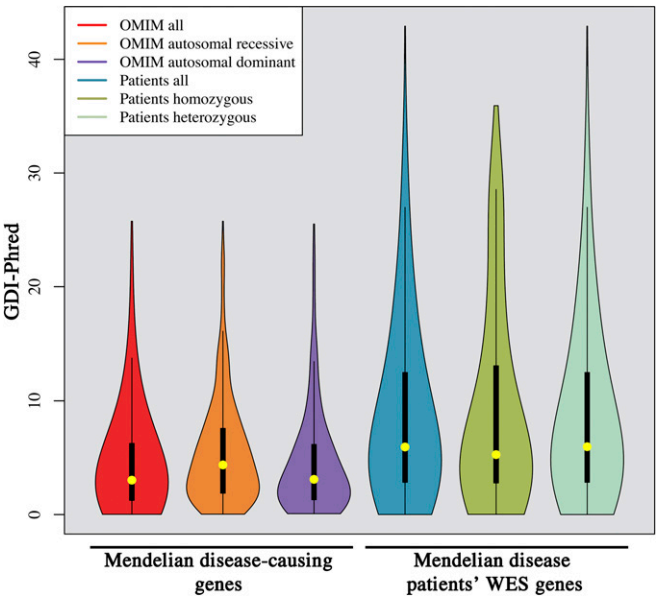

Fig. 1. GDI scores of disease-causing genes and of all of the patients' genes with variants. (*Left*) Violin plots of GDI values for true Mendelian disease-causing genes (all, AR, and AD). (*Right*) Violin plots of the corresponding GDI values (all, homozygous, and heterozygous) for the observed WES variant data from patients: missense/nonsense/frameshift/in-frame indels/splice variants in the genes with a MAF < 0.01 and a high sequencing quality, in 84 PID patients with known disease-causing mutations, after removal of the disease-causing mutation, for all of the patients' genes.

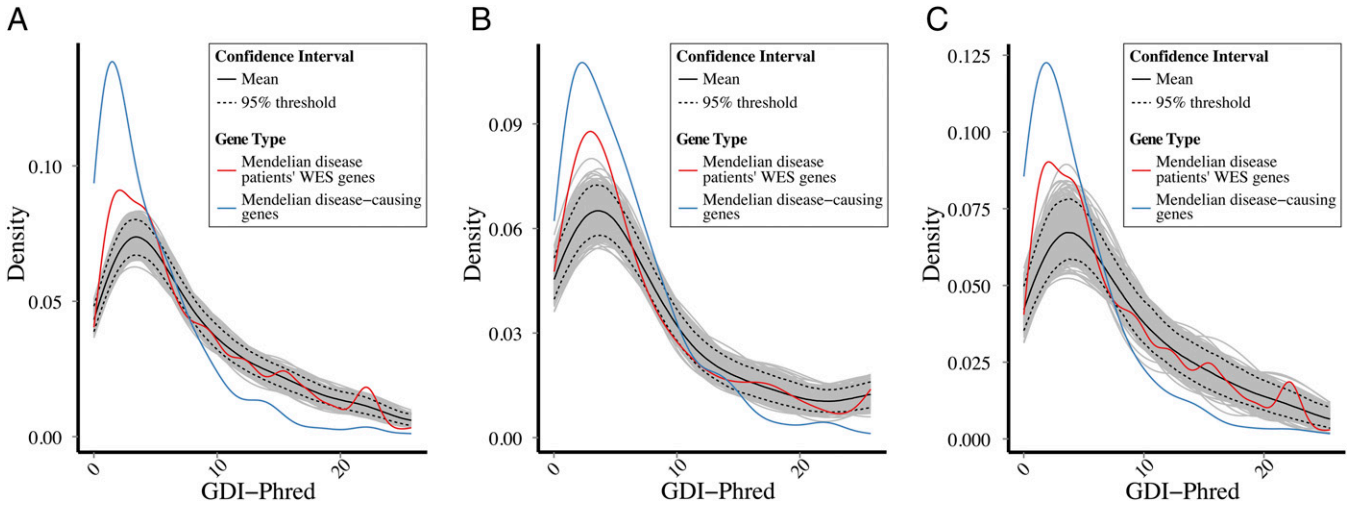

Fig. 2. Bootstrapping of GDI values: Mendelian genes vs. variants from patients. Bootstrapping simulation plots of the Mendelian disease-causing genes, together with the observed GDI values for WES rare variant data (MAF < 0.01) from patients, demonstrating a difference in the densities of the observed and expected sets. (*A*) All Mendelian disease-causing genes and all of the patients' variants. (*B*) Autosomal recessive disease-causing genes and homozygous variants from patients. (*C*) Autosomal dominant disease-causing genes and heterozygous variants from the patients.

in cases vs. controls by the hot zone approach (12, 36, 37). A combination of fixed gene-level and variant-level cutoffs was used to estimate a variant as a candidate to be damaging (see *Materials and Methods* for further details). In this test, RVIS and DNE outperformed GDI as standalone methods (all combined with a similar PolyPhen-2 cutoff, $P = 2.28 \times 10^{-07}$, $P = 5.41 \times 10^{-10}$, and $P = 5.48 \times 10^{-05}$, respectively). Interestingly, the highest overall performance ($P = 2.75 \times 10^{-10}$) was achieved with the combination of GDI+RVIS+DNE (Dataset S1, Tab S4). This analysis further suggests that the three methods capture different sets of complementary information from the available population genetic data. Altogether, these results suggest that, among the standalone methods, GDI is preferable for WES FP detection and filtration, whereas RVIS and DNE are better for TP detection of enrichment in de novo mutations. Moreover, combinations of these methods can optimize performance for TP (and potentially FP) detection.

Use of the GDI for Filtering Out False-Positive Variants in PID NGS Data. We then assessed the utility of the GDI for filtering out FP variants in the WES data for patients in a specific disease group. We used the PID disease group described above as a case study (38–40). We first estimated the GDI cutoff above which a gene is unlikely to be disease causing (PID causing in this case). For this purpose, we first summarized all currently known 229 PID genes (39, 40) and estimated their GDI scores (Dataset S1, Tab S5). We then adopted the standard assumption of experimental biologists that the maximum tolerable false-negative (FN) rate is 5% (i.e., 5% probability that the true disease-causing gene would be filtered out if the specific cutoff were applied). We therefore estimated the 95% CI for the GDI scores of AD and AR PID genes. The upper limit of this confidence interval was defined as the GDI cutoff above which a gene was considered to be a FP (i.e., having a GDI too high for it to be PID causing), with an estimated FN rate that should be <5%. We were able to filter out 60.62% of heterozygous variants under a model of AD inheritance (i.e., using the cutoff based on all AD PID genes), and 53.30% of homozygous variants under an AR model of inheritance (i.e., using the cutoff based on all AR PID genes) in the

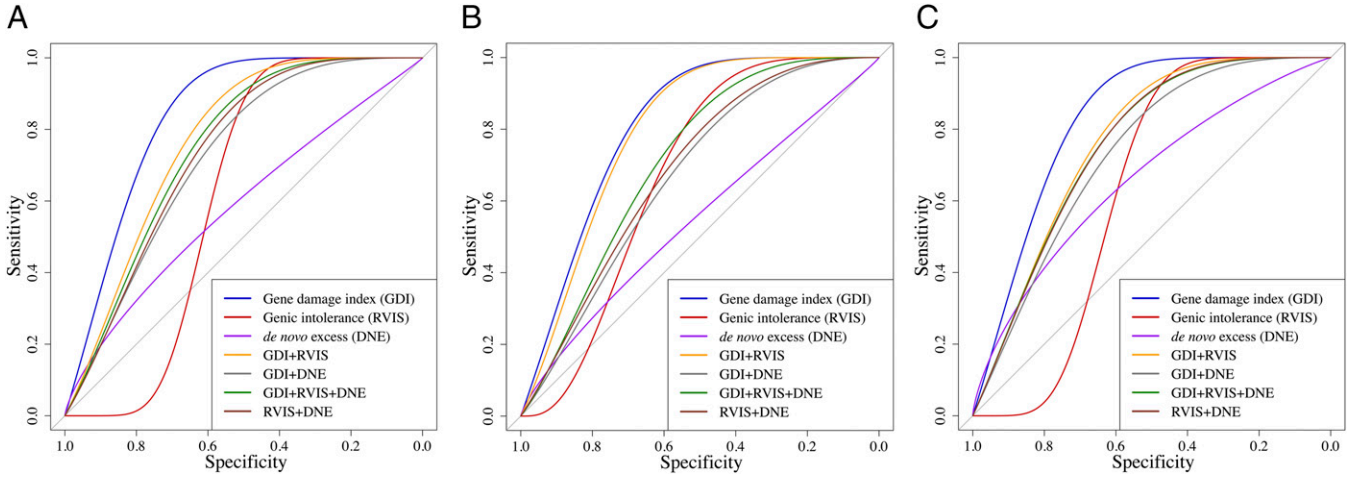

Fig. 3. Comparing GDI with state-of-the-art gene-level methods: Mendelian genes vs. rare variants in patients. ROC curve comparisons between GDI and two state-of-the-art gene-level methods (genic intolerance and de novo excess) and combinations of these methods. (*A*) All Mendelian disease-causing genes and all patients' variants. (*B*) Autosomal recessive disease-causing genes and homozygous variants from patients. (*C*) Autosomal dominant disease-causing genes and heterozygous variants from patients.

469 most highly mutated genes in patients. The GDI was therefore highly effective for detecting FP variants of highly mutated genes.

GDI Cutoffs for All, Mendelian, PID, and Cancer Disease-Causing Genes. Following the same principle as above, we then proposed GDI cutoffs and estimated FP prediction rates for various diseases (Dataset S1, Tab S6), including a general hypothetical cutoff generated from all 3,490 Human Gene Mutation Database (HGMD) (41) genes with strong experimental evidence for disease causality, 1,207 Mendelian disease-causing genes from OMIM (all, 375 distinctively AD, and 585 distinctively AR) (35), 229 PID genes (all, 42 distinctively AD, and 168 distinctively AR) (39), and 498 cancer genes (involving both germ-line and somatic mutations) extracted from the COSMIC project (all, 120 distinctively recessive, and 360 distinctively dominant) (42). For each human gene, we determined, under the FN <5% model, a low/medium/high damage prediction for the all/Mendelian/PID/cancer disease groups and the different modes of inheritance (Dataset S1, Tab S1). Finally, we suggest that the variant- and gene-level approaches could be used in synergy to create a phenotypic impact gradient (Fig. 4, also demonstrated in Fig. S9 for PID TP/FP differentiation with the hot zone approach using both CADD and GDI) (12), in which benign variants of highly damaged genes have the lowest predicted impact, and putatively damaging variants of genes with low levels of damage are predicted to have the highest impact.

Discussion

We describe here a genome-wide, population-based metric for mutational damage in all known human protein-coding genes (Dataset S1, Tab S1). We identified and characterized the most and least damaged human genes and calculated an associated GDI with various molecular genetic properties (Dataset S1, Tab S7). We demonstrated that genes highly damaged in the general population are unlikely to cause monogenic disorders. We suggest that the GDI is currently the best performing method (at least as a standalone method) for detecting FPs in patients' NGS data, whereas RVIS and DNE are better at detecting TPs. The combination of these methods, particularly for the selection of de novo mutations (TPs), appears to be synergistic. The three methods thus appear to capture different and complementary sets of population genetic information. We calculated the power of the GDI for identifying the abundant FP alleles unlikely to be responsible for PIDs in the WES data of patients. We propose GDI cutoff values for different disease groups under a general model or models of AD or AR inheritance. See lab.rockefeller.edu/casanova/GDI for programs and an easy-to-use web server providing GDI and selective pressure predictions for sets of genes.

One advantage of GDI in particular, and of gene-level metrics in general, over the more commonly used variant-level metrics is that GDI information is available for all human genes. By contrast, damage predictions are not always available for variant-level metrics, even with the CADD score, particularly for large insertions/deletions and copy number variation. Furthermore, although variant-level methods are probably best used for predicting a high impact of disruptive mutations, GDI is better suited to the prediction of low impact for variants in highly mutated genes. The two approaches are complementary. Another important and often neglected issue in the selection of an in silico approach is the FN rate. This rate should be considered carefully, and we suggest that 5% is a plausible FN rate for determining the cutoffs of GDI (and other metrics). However, the GDI cutoff could be tailored by the use of different FN rates (such as 1% or 10%), according to the nature of the study. Further studies of mutational damage to human genes should include population-specific analyses, as the GDI probably varies with ethnic background and the demographic history of the population (43). It will also be interesting to extend the GDI to the different isoforms of protein-coding genes (44–46) and noncoding RNAs (47) and to take regulatory variants into account (48). Finally, a consideration of copy number variation would also refine the calculation of the GDI (49). The rigorous study of mutational damage across human genes, at the genome-wide and population levels, together with other genome-wide approaches (11, 15, 50, 51), should facilitate studies of human genetics, particularly for monogenic disorders.

Materials and Methods

A detailed description of the methods applied can be found in *SI Materials and Methods*. Briefly, we first annotated all alleles with a global MAF <0.5 in the 1,000 Genomes Project (52, 53). We then attributed a predicted damage impact score, C, to each variant with CADD (9). We extracted WES data for 84 patients suffering from PIDs from our in-house database and analyzed in a bioinformatics pipeline described in our previous WES studies (54). For each human gene, we calculated GDI with four different models, and each was Phred homogenized. We compared the performance of GDI with that of state-of-the-art gene-level methods (including all possible combinations between the methods) for detecting FP variants in patients' WES data with ROC curves and AUCs, using TP sets of known disease-causing genes from OMIM and corresponding FP sets of rare variants from the WES data for 84 PID patients as above (55, 56). We then estimated the power of the above methods (as standalone methods and in combination) to detect TPs in patients (36), identifying a hot zone for each patient and control set separately for each method and combination. We estimated the selective pressure acting on

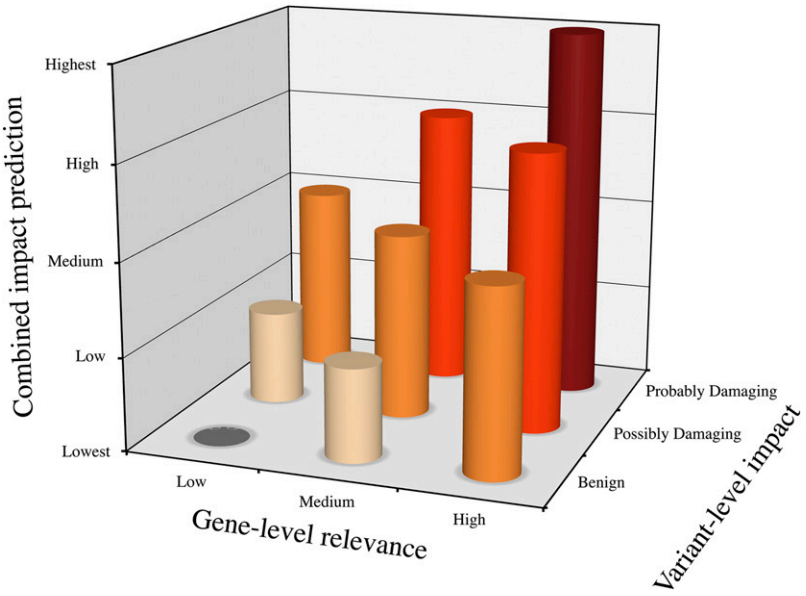

Fig. 4. Phenotypic impact predicted by a combination of variant-level and gene-level approaches. The hypothesis of combined variant- and gene-level metrics: a benign variant (variant-level) of a highly damaged gene (gene-level) would be expected to have the lowest phenotypic impact, whereas a damaging variant of a gene displaying low levels of damage would be expected to have the greatest phenotypic impact.

each human gene by the neutrality index (NI) (57–59) at the population level (1,000 Genomes Project). We identified the outliers of the GDI metric (following a gamma distribution) with a modified Z-score (60). We determined the statistical significance of various correlations between GDI and other gene properties and of GDI signature differences between FPs and TPs by Mood's median test. We used the sets of all known Mendelian PID genes (for which information about AD and AR inheritance was available) to estimate gene-level GDI upper cutoff values, whereas we used HGMD (41) to extract all known human disease-causing genes. We determined the GDI cutoff as the upper limit of the 95% CI for the known disease-causing genes of the disease group (61). We also performed bootstrapping simulations (1,000 iterations each) by Gaussian kernel density random sampling (61, 62). We calculated proteins complexity by first extracting the amino acid sequences corresponding to the proteins (27, 28) and then estimating the relative amino acid composition complexity using Clark's distance (63, 64).

ACKNOWLEDGMENTS. We thank Mark G. Thomas for providing insight into molecular evolution, Luis Barreiro for advice about estimating selective pressure, Martin Kircher for assistance with the CADD scoring algorithm, Avner Schlessinger for protein composition bias expertise, Slavé Petrovski for sharing de novo mutation data and gene-level metrics advice, and Yelena Nemirovskaya, Eric Anderson, and Mark Woollett for administrative support. This research was supported in part by March of Dimes Grant 1-FY12-440; National Institute of Allergy and Infectious Diseases Grants 5R37AI095983, 5R01AI088364, 5U01AI088685, and P01AI061093; the Rockefeller University; INSERM; University Paris Descartes; and the St. Giles Foundation. Y.I. was supported in part by Grant UL1 TR000043 from the National Center for Advancing Translational Sciences, National Institutes of Health Clinical and Translational Science Award program; F.G.L. by the New York Stem Cell Foundation; J.G.M. by the Canadian Institutes of Health Research; R.M.-B. by the European Molecular Biology Organization, S.J.d.J. by the German Research Foundation; and X.-F.K. by the Stony-Wold Herbert Fund. D.N.C. and P.D.S. received funding from Qiagen through a licensing agreement with Cardiff University. The laboratory of L.Q.-M. has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007–2013)/ERC Grant Agreement 281297.

1. Casanova JL, Conley ME, Seligman SJ, Abel L, Notarangelo LD (2014) Guidelines for genetic studies in single patients: Lessons from primary immunodeficiencies. *J Exp Med* 211(11):2137–2149.
2. Bamshad MJ, et al. (2011) Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet* 12(11):745–755.
3. Gilissen C, Hoischen A, Brunner HG, Veltman JA (2012) Disease gene identification strategies for exome sequencing. *Eur J Hum Genet* 20(5):490–497.
4. Goldstein DB, et al. (2013) Sequencing studies in human genetics: Design and interpretation. *Nat Rev Genet* 14(7):460–470.
5. Metzker ML (2010) Sequencing technologies: The next generation. *Nat Rev Genet* 11(1):31–46.
6. Belkadi A, et al. (2015) Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proc Natl Acad Sci USA* 112(17):5473–5478.
7. Kumar P, Henikoff S, Ng PC (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* 4(7):1073–1081.
8. Adzhubei IA, et al. (2010) A method and server for predicting damaging missense mutations. *Nat Methods* 7(4):248–249.
9. Kircher M, et al. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 46(3):310–315.
10. Miosge LA, et al. (2015) Comparison of predicted and actual consequences of missense mutations. *Proc Natl Acad Sci USA* 112(37):E5189–E5198.
11. Itan Y, Casanova JL (2015) Can the impact of human genetic variations be predicted? *Proc Natl Acad Sci USA* 112(37):11426–11427.
12. Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB (2013) Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet* 9(8):e1003709.
13. Samocha KE, et al. (2014) A framework for the interpretation of de novo mutation in human disease. *Nat Genet* 46(9):944–950.
14. McKenna A, et al. (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20(9):1297–1303.
15. Itan Y, et al. (2013) The human gene connectome as a map of short cuts for morbid allele discovery. *Proc Natl Acad Sci USA* 110(14):5558–5563.
16. Hagberg AA, Schult DA, Swart PJ (2008) Exploring network structure, dynamics, and function using NetworkX. *Proceedings of the 7th Python in Science Conference (SciPy2008)*, eds Varoquaux G, Vaught T, Millman J (SciPy2008, Pasadena, CA), pp 11–15.
17. Itan Y, et al. (2014) HGCS: An online tool for prioritizing disease-causing gene variants by biological distance. *BMC Genomics* 15:256.
18. Huang W, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 4(1):44–57.
19. Ashburner M, et al.; The Gene Ontology Consortium (2000) Gene ontology: Tool for the unification of biology. *Nat Genet* 25(1):25–29.
20. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res* 40(Database issue):D109–D114.
21. Gilad Y, Lancet D (2003) Population differences in the human functional olfactory repertoire. *Mol Biol Evol* 20(3):307–314.
22. Gilad Y, Man O, Glusman G (2005) A comparison of the human and chimpanzee olfactory receptor gene repertoires. *Genome Res* 15(2):224–230.
23. Ast G (2004) How did alternative splicing evolve? *Nat Rev Genet* 5(10):773–782.
24. Glickman MH, Ciechanover A (2002) The ubiquitin-proteasome proteolytic pathway: Destruction for the sake of construction. *Physiol Rev* 82(2):373–428.
25. Smith E, Morowitz HJ (2004) Universality in intermediary metabolism. *Proc Natl Acad Sci USA* 101(36):13168–13173.
26. Tomer R, Denes AS, Tessmar-Raible K, Arendt D (2010) Profiling by image registration reveals common origin of annelid mushroom bodies and vertebrate pallium. *Cell* 142(5):800–809.
27. Flicek P, et al. (2013) Ensembl 2013. *Nucleic Acids Res* 41(Database issue):D48–D55.
28. Haider S, et al. (2009) BioMart Central Portal–unified access to biological data. *Nucleic Acids Res* 37(Web Server issue):W23–W27.
29. Manry J, et al. (2011) Evolutionary genetic dissection of human interferons. *J Exp Med* 208(13):2747–2759.
30. Ohno S (1970) *Evolution by Gene Duplication* (Springer-Verlag, New York).
31. Koonin EV (2005) Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet* 39:309–338.
32. Itan Y, Bryson K, Thomas MG (2010) Detecting gene duplications in the human lineage. *Ann Hum Genet* 74(6):555–565.
33. Brown CJ, Johnson AK, Dunker AK, Daughdrill GW (2011) Evolution and disorder. *Curr Opin Struct Biol* 21(3):441–446.
34. Fuxreiter M, Tompa P, Simon I (2007) Local structural disorder imparts plasticity on linear motifs. *Bioinformatics* 23(8):950–956.
35. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* 33(Database issue):D514–D517.
36. Zhu X, et al. (January 15, 2015) Whole-exome sequencing in undiagnosed genetic diseases: interpreting 119 trios. *Genet Med*, 10.1038/gim.2014.191.
37. Zhu X, Need AC, Petrovski S, Goldstein DB (2014) One gene, many neuropsychiatric disorders: Lessons from Mendelian diseases. *Nat Neurosci* 17(6):773–781.
38. Conley ME, Casanova JL (2014) Discovery of single-gene inborn errors of immunity by next generation sequencing. *Curr Opin Immunol* 30:17–23.
39. Al-Herz W, et al. (2014) Primary immunodeficiency diseases: An update on the classification from the international union of immunological societies expert committee for primary immunodeficiency. *Front Immunol* 5:162.
40. Itan Y, Casanova JL (2015) Novel primary immunodeficiency candidate genes predicted by the human gene connectome. *Front Immunol* 6:142.
41. Stenson PD, et al. (2014) The Human Gene Mutation Database: Building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum Genet* 133(1):1–9.
42. Forbes SA, et al. (2015) COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res* 43(Database issue):D805–811.
43. Hussin JG, et al. (2015) Recombination affects accumulation of damaging and disease-associated mutations in human populations. *Nat Genet* 47(4):400–404.
44. Esteller M (2011) Non-coding RNAs in human disease. *Nat Rev Genet* 12(12):861–874.
45. Zhang C, et al. (2010) Integrative modeling defines the Nova splicing-regulatory network and its combinatorial controls. *Science* 329(5990):439–443.
46. Siddle KJ, et al. (2014) A genomic portrait of the genetic architecture and regulatory impact of microRNA expression in response to infection. *Genome Res* 24(5):850–859.
47. Consortium EP; ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489(7414):57–74.
48. Martin AR, et al. (2014) Transcriptome sequencing from diverse human populations reveals differentiated regulatory architecture. *PLoS Genet* 10(8):e1004549.
49. Stankiewicz P, Lupski JR (2010) Structural variation in the human genome and its role in disease. *Annu Rev Med* 61:437–455.
50. Snyder MW, Adey A, Kitzman JO, Shendure J (2015) Haplotype-resolved genome sequencing: Experimental methods and applications. *Nat Rev Genet* 16(6):344–358.
51. Heinzen EL, Neale BM, Traynelis SF, Allen AS, Goldstein DB (2015) The genetics of neuropsychiatric diseases: Looking in and beyond the exome. *Annu Rev Neurosci* 38:47–68.
52. Abecasis GR, et al.; 1000 Genomes Project Consortium (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature* 491(7422):56–65.
53. Wang K, Li M, Hakonarson H (2010) ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 38(16):e164.
54. Bolze A, et al. (2013) Ribosomal protein SA haploinsufficiency in humans with isolated congenital asplenia. *Science* 340(6135):976–978.
55. Robin X, et al. (2011) pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 12:77.
56. Lloyd CJ, Yong Z (1999) Kernel estimators of the ROC curve are better than empirical. *Stat Probab Lett* 44(3):221–228.
57. Rand DM, Kann LM (1996) Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from Drosophila, mice, and humans. *Mol Biol Evol* 13(6):735–748.
58. Cock PJ, et al. (2009) Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25(11):1422–1423.
59. Stoletzki N, Eyre-Walker A (2011) Estimation of the neutrality index. *Mol Biol Evol* 28(1):63–70.
60. Iglewicz B, Hoaglin D (1993) *How to Detect and Handle Outliers* (ASQC Quality Press, Milwaukee), p 87.
61. Oliphant TE (2007) Python for scientific computing. *Comput Sci Eng* 9(3):10–20.
62. Wickham H (2009) *ggplot2: Elegant Graphics for Data Analysis* (Springer, New York).
63. Cha S (2007) Comprehensive survey on distance/similarity measures between probability density functions. *Int J Math Models Methods Appl Sci* 1(4):300–307.
64. Wootton JC, Federhen S (1993) Statistics of local complexity in amino acid sequences and sequence databases. *Comput Chem* 17(2):149–163.