Genetics inMedicine | SPECIAL ARTICLE

© American College of Medical Genetics and Genomics

A literature review at genome scale: improving clinical variant assessment

Christopher A. Cassa, PhD[1,2], Daniel M. Jordan, PhD[3], Ivan Adzhubei, PhD[1,2] and Shamil Sunyaev, PhD[1,2]

Purpose: Over 150,000 variants have been reported to cause Mendelian disease in the medical literature. It is still difficult to leverage this knowledge base in clinical practice, as many reports lack strong statistical evidence or may include false associations. Clinical laboratories assess whether these variants (along with newly observed variants that are adjacent to these published ones) underlie clinical disorders.

Methods: We investigated whether citation data—including journal impact factor and the number of cited variants (NCV) in each gene with published disease associations—can be used to improve variant assessment.

Results: Surprisingly, we found that impact factor is not predictive of pathogenicity, but the NCV score for each gene can provide statistical support for prediction of pathogenicity. When this gene-level citation metric is combined with variant-level evolutionary conservation and structural features, classification accuracy reaches 89.5%. Further, variants identified in clinical exome sequencing cases have higher NCVs than do simulated rare variants from the Exome Aggregation Consortium database within the same set of genes and functional consequences ($P < 2.22 \times 10^{-16}$).

Conclusion: Aggregate citation data can complement existing variant-based predictive algorithms, and can boost their performance without the need to access and review large numbers of papers. The NCV is a slow-growing metric of scientific knowledge about each gene's association with disease.

Genet Med advance online publication 1 February 2018

Key Words: citation data; clinical interpretation; genomic medicine; impact factor; variant assessment

INTRODUCTION

Vast numbers of genetic variants have been reported in the medical and scientific literature as causing disease,[1] but it is still difficult to leverage this knowledge base clinically.[2] Variants that are identified in well-characterized genes such as *BRCA2* or *LDLR* would naturally warrant further review and potential clinical surveillance.[3] However, there are now approximately 5,000 genes with some clinical significance, many of which lack strong statistical evidence of pathogenicity, which we hereafter colloquially refer to as "disease-associated." For the majority of these genes only a few publications have described their clinical significance.[2] Patients undergoing clinical genomic sequencing often carry novel variants in these less-characterized disease-associated genes,[4] leading to challenges in diagnosis and complications in medical management.[5]

This dearth of information makes it difficult to translate the presence of a variant into an estimate of disease risk, particularly in carrier testing for asymptomatic individuals or in patients with multiple phenotypes where the range of associated disease genes is unclear.[6–8] When there are no reports describing an identified variant, clinicians must rely primarily on computational predictive techniques and published gene-level evidence to predict its functional effect.[2]

Even in more established disease genes, the majority of cases contain variants that are observed in only a single family, and laboratories must routinely prioritize and classify these variants of unknown significance.[9,10] Further, when specific variants have been previously described in publications, the relevant knowledge base may be inadequate for clinical interpretation. Because many of the variants in publication databases such as the Human Genome Mutation Database (HGMD) were originally identified in small, symptomatic populations without matched control groups, their associations can suffer from incorrect estimates of significance or effect size and a nontrivial fraction is likely to be spurious.[11,12] The result is a mixture of well-established associations with unverified or even false-positive findings.

For these reasons, a central challenge in clinical genomics is to prioritize variants that are identified during sequencing. Here, we analyze new features related to publication support at the gene level for their accuracy in estimating functional impact and potential to improve variant assessment.

METHODS

Citation-based predictive features

While estimates of effect size or significance may be missing or unreliable in case reports, it is difficult to ignore the clinical significance of many existing associations.[13–15] To address the

[1]Division of Genetics, Brigham and Women's Hospital, Boston, Massachusetts, USA; [2]Department of Medicine, Harvard Medical School, Boston, Massachusetts, USA; [3]Department of Genetic and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, New York, USA. Correspondence: Christopher A. Cassa (cassa@alum.mit.edu)

Submitted 21 June 2017; accepted 8 November 2017; advance online publication 1 February 2018. doi:10.1038/gim.2017.230

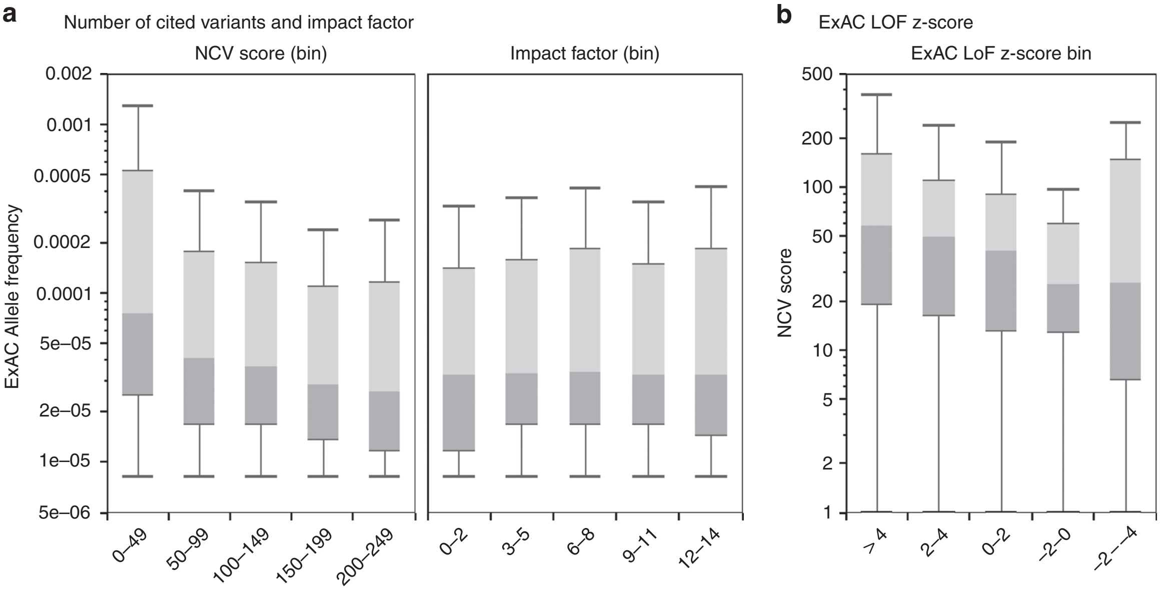

Figure 1 Correlation between the NCV score and measures of selective constraint. (**a**) Box plots of the allele frequency (Exome Aggregation Consortium, $N = 60{,}706$) of a set of putative disease variants (HGMD 2012.1), grouped by the number of cited variants (NCV) in each gene and the journal impact factor for each citation. We observed an inverse correlation between the NCV score and allele frequency, while the impact factor of the journal in which a variant is cited was not correlated with allele frequency. (**b**) We found that genes under stronger selective constraint (those with positive *z*-scores reflect genes that have fewer loss-of-function variants than expected based on a mutational model) have higher NCV scores than genes that are already associated with disease but are under lower selective constraint ($P = 2.36 \times 10^{-3}$).

uncertainty in individual reports, we extracted information from publications—in aggregate across suspected disease genes—to assess the relative strength of each gene's association with disease. The goal was to identify a metric that would have broad utility in separating highly suspicious genes from those that are more weakly associated with disease in large gene panels and clinical exome interpretation.

We evaluated two citation-based features in variant-level assessment:

(1) Number of cited variants (NCV) score. Using a large set of published disease associations (HGMD 2012.1),[1] we calculated the number of distinct variant sites in each gene that are described as causal. In addition, we separately analyzed the density of variants with citations, using the canonical protein-coding length (**Supplementary Materials** online).

(2) Impact factor (IF) score. Each variant in HGMD is associated with a single publication, for which we identified the journal IF; Thomson Reuters).[16] The IF is defined as the average number of citations over the previous 2 years for all papers in a journal.[17]

RESULTS

To determine whether these features are predictive of pathogenicity, we applied each score to a large set of putative disease-associated variants. For each variant in HGMD 2012.1 (restricted to "DM" variant class; importantly, this version was not explicitly filtered by variant frequency), we calculated the allele frequency from a large population cohort (Exome Aggregation Consortium (ExAC), $N = 60{,}706$ individuals). We then compared the NCV and IF scores for each variant with its observed allele frequency. In this set of disease-associated variants, we expected that variants with higher population frequencies would be less likely to be highly penetrant and pathogenic and those with lower population frequencies would be more likely to have stronger association with disease.[18]

Surprisingly, journal IF was not correlated with variant allele frequency, indicating that variants reported in high-impact journals, on average, are not more likely to be pathogenic (**Figure 1a**, **Supplementary Materials**). However, we found a significant inverse correlation between NCV and variant allele frequency in putative disease variants. This suggests that variants in genes with lower NCV scores are less likely to be highly penetrant pathogenic variants (**Figure 1a**). This trend was observed when the analysis was restricted to rare variants (allele frequency <0.01) as well as in a broader set of variants (allele frequency <0.05, **Supplementary Materials**).

At the gene level, we also found that the NCV score correlated well with an independently ascertained measure of genic intolerance to variation.[19] Disease-associated genes that are under stronger selective constraint (as estimated by the depletion of loss-of-function variants in comparison with what was expected) have higher NCV scores than disease-associated genes that are under neutral or relaxed constraint (**Figure 1b**, Mann-Whitney U *P*-value 2.36×10^{-3}).

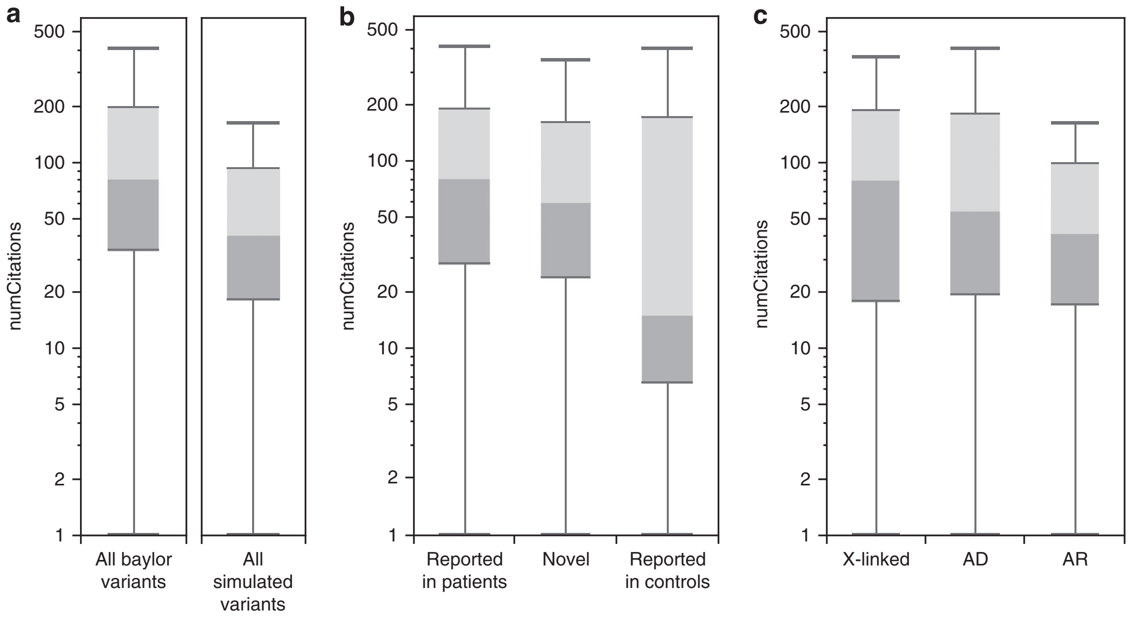

Figure 2 Number of cited variants (NCV) score applied to clinical exome-sequencing cases from Baylor Genetics and from simulated individuals. (**a**) Variants ascertained from clinical exome-sequencing cases that resulted in a molecular diagnosis of Mendelian disorder had significantly higher NCV scores than those simulated from individuals in the Exome Aggregation Consortium database (Mann-Whitney U $P < 2.22 \times 10^{-16}$). (**b**) Variants identified as causal in clinical sequence interpretation that had been previously reported in patients had higher NCV scores, followed by novel variants in known genes and by variants also seen in control populations. (**c**) By mode of inheritance, cases with X-linked inheritance had higher NCV scores than those associated with autosomal dominant (AD) and autosomal recessive (AR) inheritances.

Predicting variant pathogenicity in clinical exome sequencing cases

Next, we measured the performance of the NCV score in discriminating variants described as causal in clinical exome sequencing cases (Baylor Genetics).[9] We compared the distribution of observed NCV values in Baylor cases with those in a set of realistic simulated "candidate" variants. The variants were drawn randomly in proportion to their ExAC allele frequencies,[19] and were restricted to variants with frequencies below 0.1% and specific functional effects in the same set of disease genes (**Supplementary Materials**). We found that causal Mendelian variants from this clinical exome-sequencing program have significantly higher NCV scores than simulated variants (**Figure 2a**; $P < 2.22 \times 10^{-16}$). This demonstrates that NCV can potentially improve standard filtering and prioritization by allele frequency in clinical exome-sequencing cases.

As expected, variants identified in clinical sequencing that had previously been reported as causal in clinic patients had significantly higher NCV scores than those previously reported in control populations ($P = 1.26 \times 10^{-16}$) (**Figure 2b**). We also found that X-linked recessive variants from Mendelian cases had higher NCV scores than autosomal dominant ($P = 3.51 \times 10^{-4}$) and autosomal recessive cases ($P = 7.77 \times 10^{-4}$; **Figure 2c**). This may be due to differences in ability to identify recessive disorders, differences in the genetic heterogeneity of recessive disorders, or other sources of bias.

Integrating the NCV score with variant-level predictions of functional impact

Next, we sought to determine whether the gene-level NCV score provides information that complements predictions made using structural and evolutionary features at the variant level. We first used the NCV score to classify the pathogenicity of individual variants, a common requirement in clinical interpretation. We applied the NCV score to each variant in HumVar, a gold-standard test data set that contains missense variants that have already been classified as either benign or pathogenic, derived from UniProtKB-Swiss-Prot,[20] which is often used in classifier training. We made extensive efforts to prevent bias and overfitting in our predictions[21] (**Supplementary Materials** and **Supplementary Table 1**).

We found that the NCV score in each gene has predictive value and can classify HumVar variant pathogenicity with an area under the curve (AUC) of 0.847 (naive Bayes classifier in fivefold cross-validation) (**Figure 3**, green). We also found that the density of cited variants in a gene (as defined by the NCV divided by the canonical protein coding length) also improves pathogenicity predictions (**Supplementary Materials**). These gene-level predictions of pathogenicity are provided alongside NCV values in **Supplementary Table 2** and online at http://genetics.bwh.harvard.edu/genescores/.

We then supplemented this gene-level citation feature with individual variant-level predictions made by PolyPhen-2, a variant classifier that uses evolutionary conservation and structural features. We integrated these scores using the

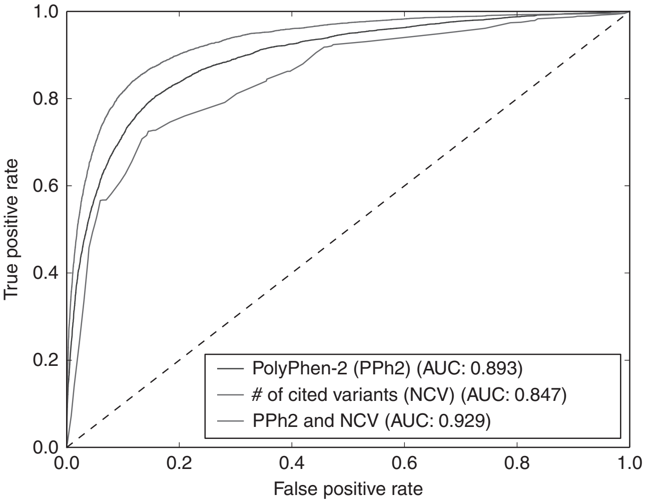

Figure 3 Receiver operating characteristic curves from naive Bayes classifiers trained on the same gold-standard data set from UniProtKB (HumVar). The number of cited variants (NCV) score is generated using a fully disparate set of citation data from Human Genome Mutation Database variants that are not recorded in UniProtKB. A classifier trained using PolyPhen-2 features (blue) is compared with one that uses only the NCV score (green). Combining the NCV score with PolyPhen-2 features improves prediction accuracy substantially (red). AUC, area under the curve. A full color version of this figure is available at the *Genetics in Medicine* journal online.

probability of pathogenicity output by PolyPhen-2 as a prior for our own naive Bayes NCV gene score, producing a composite score (**Supplementary Table 1**). The existing performance of PolyPhen-2 on the same data set (without NCV) has an AUC of 0.893 (**Figure 3**, blue). When the gene-level NCV score is combined with PolyPhen-2 features, the AUC on the same data set rises to 0.929 (**Figure 3**, red), a substantial improvement in classification performance.

We found that the accuracy of this approach is not generated by a small set of well-known disease genes. We removed all variants appearing in any of the 56 genes included in the American College of Medical Genetics and Genomics guidelines[3]—accounting for 1,999 of the 28,016 variants (7.14%)—and observed only a small reduction in AUC in our test data set (0.8586 to 0.8575). We also found only a small effect when removing genes with outlier NCV scores (**Supplementary Materials**).

Lower NCV scores observed for dosage-sensitive, severe-disease genes

Interestingly, we found the opposite effect in genes under the very strongest selection, where the loss of even one copy results in a severe phenotype. In a set of high-confidence haploinsufficient autosomal disease genes ($N = 127$),[22] we found that genes with the strongest phenotypic severity, earliest age of onset, and highest fraction of de novo variants are associated with lower NCV scores (**Figures 4a–c**). This is consistent with previous work that demonstrated that genes under extremely strong selection (e.g., genes with increased embryonic lethality in mice, those with cell essentiality, and those severely depleted of variation in human populations) have fewer publications in PubMed than genes that are associated with weaker selective effects.[23]

DISCUSSION

In this study, we demonstrated that aggregate citation data can be used at both the gene and variant levels to improve pathogenicity prediction based on novel variants. We have shown that these data can be combined with existing variant-level classifications and have potential clinical utility in classifying variants in rare-disease cases along with other indicators of pathogenicity. This helps address a substantial issue in clinical sequence interpretation: the presence of several variants under consideration in each case.

Ultimately, when a variant of unknown significance is encountered, the steps required to determine which variant is causal are costly and include an in-depth literature review of each gene, familial segregation, and functional validation studies.[6] The NCV score can be used, along with variant-level scores, to prioritize variants that should be considered first for these deeper studies. Although clinical laboratories have developed efficient variant-assessment processes, it is difficult to assess the accuracy or clinical relevance of all publications mentioning a gene or to understand the relative importance of published literature on one gene relative to that on all other genes. For this reason, it is useful to employ automated approaches to assist with prioritization so that resource-intensive processes can initially be restricted to a small set of genes.

If a gene has many variant sites associated with disease and few benign sites, it is more likely that subsequently identified variants in that gene will be associated with disease. The NCV score is a gene-based feature that captures the number of sites in a gene that are reported to be disease-associated versus those that are neutral. We gain predictive value using the genic distribution of both neutral and pathogenic variants in each gene in HumVar, and use these data to provide a posterior probability based on the NCV score for each gene.[24,25] While the NCV is a gene-level score, a high NCV score only increases our confidence that the gene itself has a well-documented role in disease, but there may be many variants within well-known disease genes may not be causal. This is precisely where information from functional predictions at the variant level yield information complementary to the NCV score and can provide utility in variant assessment.

For example, if a variant with a possibly damaging PolyPhen-2 score is observed in *BRCA2*, it is likely to rise to a probably damaging score level because of the preponderance of variants in *BRCA2* that are pathogenic (66.8%) and the lack of neutral or benign variants in the gene. In this case, a variant identified in *BRCA2* would have its PolyPhen-2 variant-level score multiplied by 1.335 (scores in **Supplementary Table 2**), which would alter its predictive power for pathogenicity. For other genes, the score change may be smaller, especially if the knowledge base for a gene is less comprehensive.

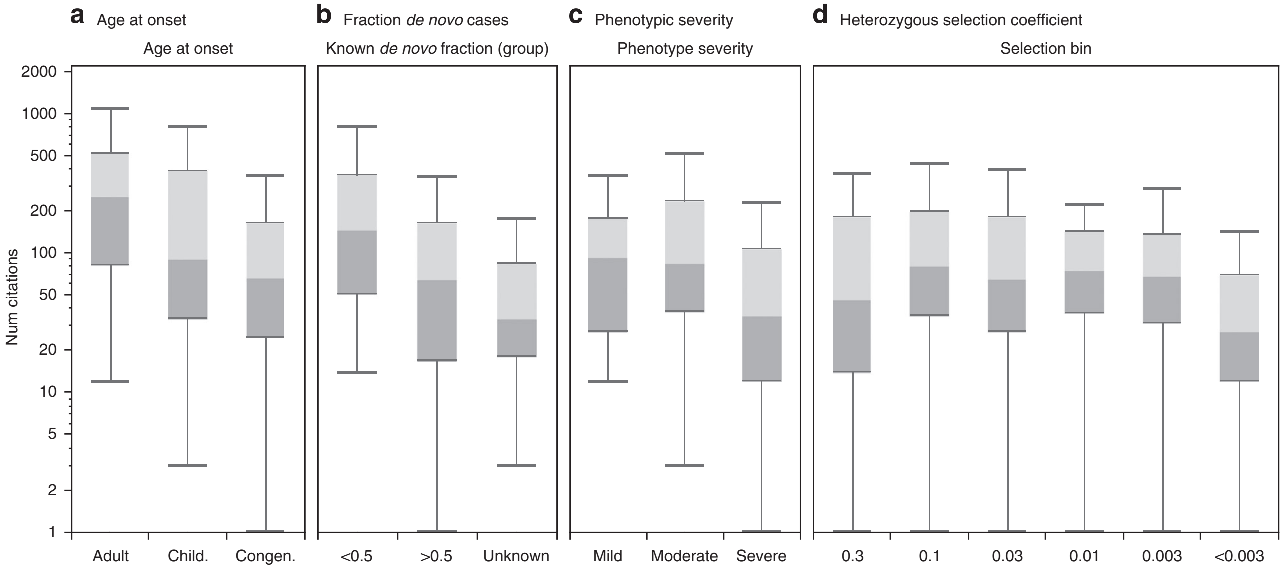

Figure 4 Number of cited variants (NCV) score applied to haploinsufficient disease genes (ClinGen Dosage Sensitivity Project). For 127 autosomal genes, we annotated each disease-associated gene with its NCV score, for each disease category and classification. Higher NCV scores were associated with (**a**) earlier age of onset (Mann–Whitney U $P = 0.0139$), (**b**) a larger fraction of de novo variants ($P = 0.00345$), and (**c**) increased phenotypic severity ($P = 0.00494$). Box plots range from 25th- to 75th-percentile values, and whiskers include 1.5 times the interquartile range. (**d**) Genes under the very strongest selection (s_het > 0.3) have lower NCV scores than genes under moderate selection, which may be due to the fact that many of these variants cause such severe effects that they may be less common in clinical cases.

In this case, the NCV score provides an assessment of gene morbidity and the variant classification provides information about the impact of specific variants, differentiating between those that are likely to disrupt a specific functional region of consequence or disrupt protein stability and those that have little or no effect on protein function.[26] This is consistent with previous results showing that different genes have measurably different thresholds of pathogenicity, and the accuracy of prediction methods can be improved by applying different thresholds or priors to different genes.[27]

Interestingly, we found that IF, on average, is not predictive of pathogenicity, and it serves as a negative control in this study (**Figure 1**). It may be the case that journal IF is simply too statistically noisy to predict pathogenicity, or that individual article citation data would correlate more closely than the journal IF. The NCV score is less susceptible to noise, as it is a slow-growing metric that reflects a broad perspective about the knowledge base for each gene, and outperforms other citation-based features.

The many factors that may potentially influence the NCV score are difficult to model without extensive data. Genetic tests that are commonly indicated (e.g., for cancer, hearing loss, cardiomyopathy) based on disease prevalence will lead to increased testing and observation of variants in known genes, potentially overweighting the NCV score. Conversely, genes associated with autosomal recessive disorders, as well as disorders driven largely by gain-of-function mutations, will both have lower NCV scores.

While longer genes have a larger target size, we made our comparisons with features that are unrelated to gene length (e.g., allele frequency, loss-of-function depletion, and clinical categories.) We specifically chose to feature NCV rather than a proportion of variant sites normalized by gene length owing to two concerns: (i) gene length potentially serves as a confounder, as longer genes are known to have enriched expression in the brain,[28] and (ii) genes that have relatively small gain-of-function regions associated with disease[29] may have diluted scores. We have also provided length-normalized results (**Supplementary Materials**), which provide similar predictive accuracy.

There are several potential limitations of this study. First, we did not attempt to remove review articles or identify articles that refute previous claims, due to review complexity. We also recognize the feedback potential for variants that are classified as pathogenic merely because previous variants in that gene have been classified as such. These issues remain difficult to resolve in computational predictions.[21]

While many HGMD variants may not be high-penetrance Mendelian variants,[30,31] the literature reports many high-quality gene–disease associations.[15,32] This approach extracts aggregate value from the large number of previously reported disease associations whose uncertain significance would make clinical classification difficult.

SUPPLEMENTARY MATERIAL

Supplementary material is linked to the online version of the paper at http://www.nature.com/gim

ACKNOWLEDGMENTS

This research was supported by NHGRI grant HG007229 (C.A.C.) and NIGMS grant GM078598 (I.A., D.M.J., and S.S.). We thank HGMD for permission to publish these aggregated data, and

Anne O'Donnell-Luria for access to curated data related to disorders in the ClinGen Dosage Sensitivity Project.

DISCLOSURE

The authors declare no conflict of interest.

REFERENCES

1. Stenson PD, Ball E, Howells K, Phillips A, Mort M, Cooper DN. Human Gene Mutation Database: towards a comprehensive central mutation database. *J Med Genet* 2008;45:124–126.
2. Walsh R, Thomson K, Ware JS, et al. Reassessment of Mendelian gene pathogenicity using 7,855 cardiomyopathy cases and 60,706 reference samples. *Genet Med*. 2017;19:192–203.
3. Green RC, Berg JS, Grody WW, et al. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet Med* 2013;15:565–574.
4. Biesecker LG, Green RC. Diagnostic clinical genome and exome sequencing. *N Engl J Med* 2014;370:2418–2425.
5. MacArthur DG, Manolio TA, Dimmock DP, et al. Guidelines for investigating causality of sequence variants in human disease. *Nature* 2014;508:469–476.
6. Kohane IS, Masys DR, Altman RB. The incidentalome: a threat to genomic medicine. *JAMA* 2006;296:212–215.
7. Ashley EA, Butte AJ, Wheeler MT, et al. Clinical assessment incorporating a personal genome. *Lancet* 2010;375:1525–1535.
8. Brunham LR, Hayden MR. Medicine. Whole-genome sequencing: the new standard of care? *Science* 2012;336:1112–1113.
9. Yang Y, Muzny DM, Xia F, et al. Molecular findings among patients referred for clinical whole-exome sequencing. *JAMA* 2014;312:1870–9.
10. Lee H, Deignan JL, Dorrani N, et al. Clinical exome Ssequencing for genetic identification of rare Mendelian disorders. *JAMA*. 2014;312:1880–7.
11. National Human Genome Research Institute. 2014. Genome-Wide Association Studies. http://www.genome.gov/20019523 Accessed 20 December, 2017.
12. Genomes Project C, Abecasis GR, Altshuler D, et al. A map of human genome variation from population-scale sequencing. *Nature* 2010;467: 1061–1073.
13. Dorschner MO, Amendola LM, Turner EH, et al. Actionable, pathogenic incidental findings in 1,000 participants' exomes. *Am J Hum Genet*. 2013;93:631–640.
14. MacArthur DG, Balasubramanian S, Frankish A, et al. A systematic survey of loss-of-function variants in human protein-coding genes. *Science* 2012;335:823–828.
15. Cassa CA, Savage SK, Taylor PL, Green RC, McGuire AL, Mandl KD. Disclosing pathogenic genetic variants to research participants: quantifying an emerging ethical responsibility. *Genome Res* 2012;22:421–428.
16. Thompson Reuters. The Thompson Reuters Impact Factor. 2014. http://wokinfo.com/essays/impact-factor/ Accessed 20 December, 2017.
17. Hirsch JE. An index to quantify an individual's scientific research output. *Proc Natl Acad Sci USA*. 2005;102:16569–16572.
18. Kryukov G V, Pennacchio LA, Sunyaev SR. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am J Hum Genet*. 2007;80:727–739.
19. Lek M, Karczewski KJ, Minikel E V., et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 2016;536:285–291.
20. Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. *Nat Methods* 2010;7:248–249.
21. Grimm DG, Azencott C-A, Aicheler F, et al. The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. *Hum Mutat* 2015;36:513–523.
22. Rehm HL, Berg JS, Brooks LD, et al. ClinGen—The Clinical Genome Resource. *N Engl J Med* 2015;372:2235–2242.
23. Cassa CA, Weghorn D, Balick DJ, et al. Estimating the selective effect of heterozygous protein truncating variants from human exome data.. *bioRxiv* 2016; 075523.
24. Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet* 2013;9:e1003709.
25. Samocha KE, Robinson EB, Sanders SJ, et al. A framework for the interpretation of de novo mutation in human disease. *Nat Genet* 2014; 46:944–950.
26. Baux D, Blanchet C, Hamel C, et al. Enrichment of LOVD-USHbases with 152 USH2A genotypes defines an extensive mutational spectrum and highlights missense hotspots. *Hum Mutat* 2014;35:1179–1186.
27. Itan Y, Shang L, Boisson B, et al. The mutation significance cutoff: gene-level thresholds for variant predictions. *Nat Methods* 2016;13: 109–110.
28. Sibley CR, Emmett W, Blazquez L, et al. Recursive splicing in long vertebrate genes. *Nature* 2015;521:371–375.
29. Herman DS, Lam L, Taylor MRG, et al. Truncations of titin causing dilated cardiomyopathy. *N Engl J Med* 2012;366:619–628.
30. Cassa CA, Tong MY, Jordan DM. Large numbers of genetic variants considered to be pathogenic are common in asymptomatic individuals. *Hum Mutat* 2013;34:1216–20.
31. Xue Y, Chen Y, Ayub Q, et al. Deleterious- and disease-allele prevalence in healthy individuals: insights from current predictions, mutation databases, and population-scale resequencing. *Am J Hum Genet*. 2012; 91:1022–1032.
32. Fabsitz RR, McGuire A, Sharp RR, et al. Ethical and practical guidelines for reporting genetic research results to study participants: updated guidelines from a National Heart, Lung, and Blood Institute working group. *Circ Cardiovasc Genet*. 2010;3:574–580.