Received: 15 September 2021 | Revised: 18 January 2022 | Accepted: 23 February 2022

DOI: 10.1002/humu.24362

INFORMATICS

Human Mutation HGVS HUMAN GENOME VARIATION SOCIETY WILEY

An expanded phenotype centric benchmark of variant prioritisation tools

Denise Anderson | Timo Lassmann

Telethon Kids Institute Precision Health Computational Biology, The University of Western Australia, Subiaco, Western Australia, Australia

Correspondence
Timo Lassmann, Telethon Kids Institute, Northern Entrance, Perth Children's Hospital, 15 Hospital Avenue, Nedlands, Western Australia, 6009, Australia.
Email: Timo.Lassmann@telethonkids.org.au

Funding information
McCusker Charitable Foundation; Stan Perron Charitable Foundation; Feilman Foundation

Abstract

Identifying the causal variant for diagnosis of genetic diseases is challenging when using next-generation sequencing approaches and variant prioritization tools can assist in this task. These tools provide in silico predictions of variant pathogenicity, however they are agnostic to the disease under study. We previously performed a disease-specific benchmark of 24 such tools to assess how they perform in different disease contexts. We found that the tools themselves show large differences in performance, but more importantly that the best tools for variant prioritization are dependent on the disease phenotypes being considered. Here we expand the assessment to 37 tools and refine our assessment by separating performance for nonsynonymous single nucleotide variants (nsSNVs) and missense variants (i.e., excluding nonsense variants). We found differences in performance for missense variants compared to nsSNVs and recommend three tools that stand out in terms of their performance (BayesDel, CADD, and ClinPred).

KEYWORDS
dbNSFP, disease, human phenotype ontology, phenotype, variant prioritization

Next-generation sequencing for clinical diagnosis of genetic diseases is routinely used, however, filtering and interpreting the tens of thousands (whole exome sequencing) or millions (whole genome sequencing) of variants identified by these approaches remains challenging (Caspar et al., 2018). Variant prioritization tools assist in this task by predicting the likely pathogenicity of variants in silico, thereby enabling ranking and filtering of variants. We previously performed a benchmark study of 24 variant prioritization tools and reported that performance differs depending on the disease phenotype and recommended use of five top performing tools (Anderson & Lassmann, 2018). Here we present an update to our benchmark that incorporates additional variant prioritization tools added to the latest version of dbNSFP (Liu et al., 2020), increasing the number of assessed tools to 37. Furthermore, we refined our assessment by considering the performance of tools for nonsynonymous single nucleotide variants (nsSNVs) and missense variants (i.e., excluding nonsense variants) separately. In total, for missense variants we tested 37 tools across 4890 disease phenotypes and for nsSNVs we tested 22 tools across 5723 disease phenotypes.

Performance of the variant prioritization tools was assessed through creation of disease specific benchmark datasets. To create these datasets we (1) used terms for human phenotypic abnormalities from the Human Phenotype Ontology (HPO) resource (Köhler et al., 2014), (2) obtained the genes associated with each HPO term from the disease to gene mapping tool Phenolyzer (Yang, Robinson, & Wang, 2015) and (3) obtained the pathogenic variants residing in these genes from ClinVar (Landrum et al., 2016). For each HPO term, performance of tools was based on how well they could discriminate pathogenic variants from a set of benign variants (Niroula & Vihinen, 2019) based on the area under the precision-recall curve (auPRC) which is suitable for inherently unbalanced data (i.e., the ratio of pathogenic to benign variants is small). We also assessed each tool based on the proportion of ClinVar pathogenic variants contained in the top 25 variants after ranking by predicted pathogenicity (PP25).

We categorized the variant prioritization tools into those that predict pathogenicity based primarily on (1) conservation scores derived from sequence alignments, (2) machine learning classifiers incorporating a diverse set of functional genomic features and (3) ensemble methods that incorporate pathogenicity scores from a number of variant prioritization tools. We used 16 conservation scores (bStatistic, FATHMM, GERP++, LIST-S2, LRT, MutationAsssessor, phastCons17way-primate, phastCons30way-mammalian, phastCons100way-vertebrate, phyloP17 way-primate, phyloP30way-mammalian, phyloP100way-vertebrate, PROVEAN, SIFT, SIFT4G and SiPhy), 15 machine learning scores (CADD, DANN, DEOGEN2, Eigen, Eigen-PC, fathmm-MKL, fathmm-XF, fitCons-i6, GenoCanyon, MPC, MutationTaster, PolyPhen2-HDIV, PolyPhen2-HVAR, PrimateAI, and VEST4) and 6 ensemble scores (BayesDel with allele frequency, BayesDel without allele frequency, ClinPred, MetaLR, MetaSVM, and REVEL). Though our focus is on classification of nsSNVs, a small number of these tools (BayesDel, CADD, MutationTaster2, PROVEAN, and SIFT) also classify insertion/deletion variants (InDels) which may be relevant for the disease under study.

Not all variant prioritization tools predict pathogenicity of nonsense variants, hence we evaluated performance for nsSNVs and missense variants separately. For missense variants, the top performing tools based on the auPRC included all of the ensemble scores (BayesDel_addAF, BayesDel_noAF, ClinPred, MetaLR, MetaSVM, and REVEL) and two machine learning scores (DEOGEN2 and VEST4) (Figure 1a and Table S1). The types of tools that perform well is more mixed when considering the PP25, with three conservation scores (LRT, phastCons100way, and SIFT), two machine learning scores (MutationTaster and Polyphen2-HDIV) and one ensemble score (BayesDel_addAF) being the best performers (Figure S1a and Table S2). For nsSNVs, the top performing tools based on the auPRC included both ensemble scores (BayesDel_addAF and BayesDel_noAF) and four of the machine learning scores (CADD, Eigen, Eigen-PC, and VEST4) (Figure 1b and Table S3). Of note, CADD, Eigen and Eigen-PC were overall weak performers when prioritizing missense variants but were excellent at prioritizing nsSNVs. Again, for PP25, performance is mixed with three conservation scores (LRT, phastCons30way, and phastCons100-way), two machine learning scores (CADD and MutationTaster) and two ensemble scores (BayesDel_addAF and BayesDel_noAF) showing very strong performance (Figure S1b and Table S4). The 50 HPO terms with the most variable performance across the tools for the auPRC and PP25 are shown for both missense variants and nsSNVs in Figures S2 through S5.

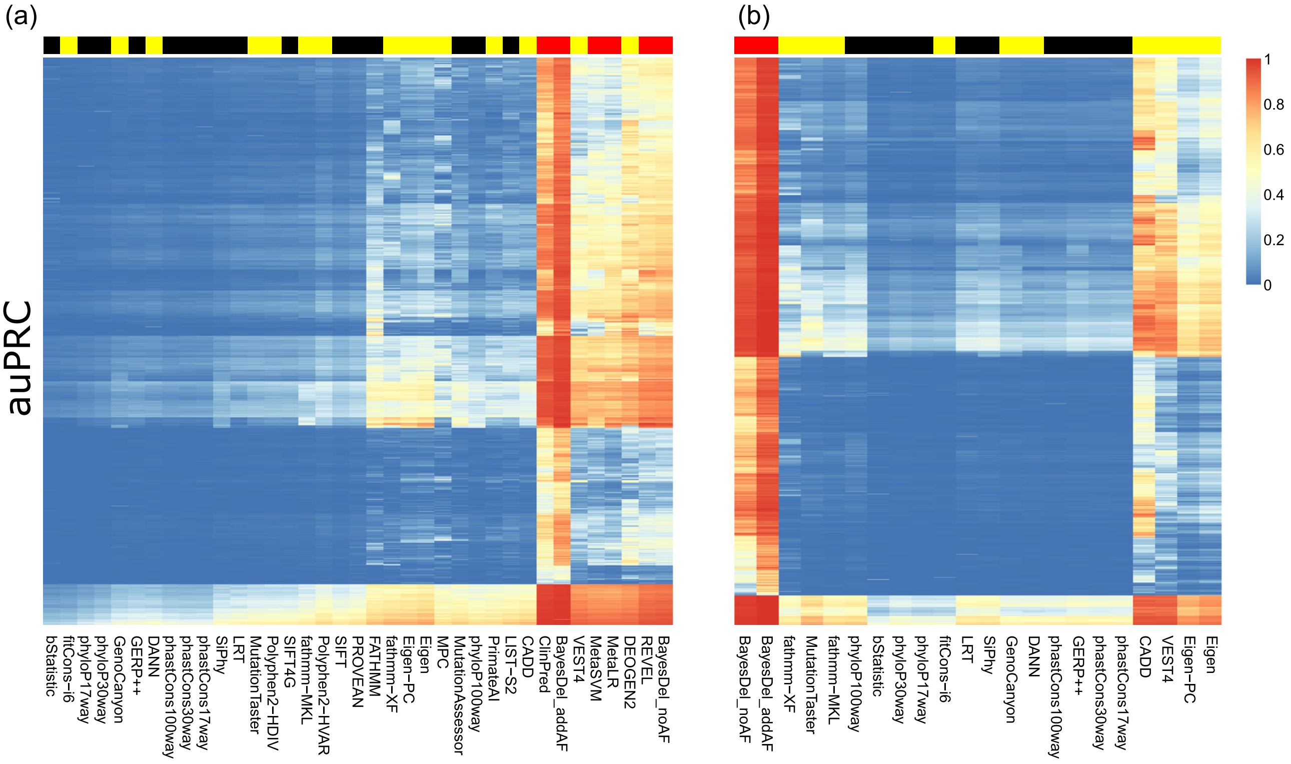

FIGURE 1 Heatmaps showing performance (auPRC) of variant prioritization tools for missense variants (a) and nsSNVs (b). Color coding of columns is based on the method used to predict pathogenicity, where black = conservation scores, yellow = machine learning scores and red = ensemble scores. Hierarchical cluster analysis with Euclidean distance and complete agglomeration was used to cluster both the tools and the HPO terms. HPO, Human Phenotype Ontology

Next, we examined performance of the top performing tools across different disease contexts. We limited this to the auPRC as there was more variability in performance across the HPO terms in comparison to the PP25 where performance was strong for most terms. We included four top level HPO terms and their descendant terms. The four top level terms were Abnormality of metabolism/homeostasis (HP:0001939), Abnormality of the immune system (HP:0002715), Abnormality of the nervous system (HP:0000707) and Neoplasm (HP:0002664).

BayesDel_addAF was clearly the strongest performer across the four top level HPO terms for missense variants, with median auPRC values ranging from 0.8 to 0.94 (Figure 2). ClinPred was the second best performer for missense variants (median auPRC range: 0.65–0.82), however, the interquartile range (IQR) was wider than that seen for BayesDel_addAF. Again, for nsSNVs, BayesDel_addAF was the best performer and auPRCs were higher and IQRs smaller than those seen for missense variants (median auPPRC range: 0.93–0.99). For both BayesDel_addAF and BayesDel_noAF, performance was stronger for nsSNVs compared to missense variants across all four top level terms. This is in contrast to VEST4, the only other tool with scores for both missense variants and nsSNVs, where performance was similar for Abnormality of metabolism/homeostasis and Abnormality of the immune system but improved for Abnormality of the nervous system and

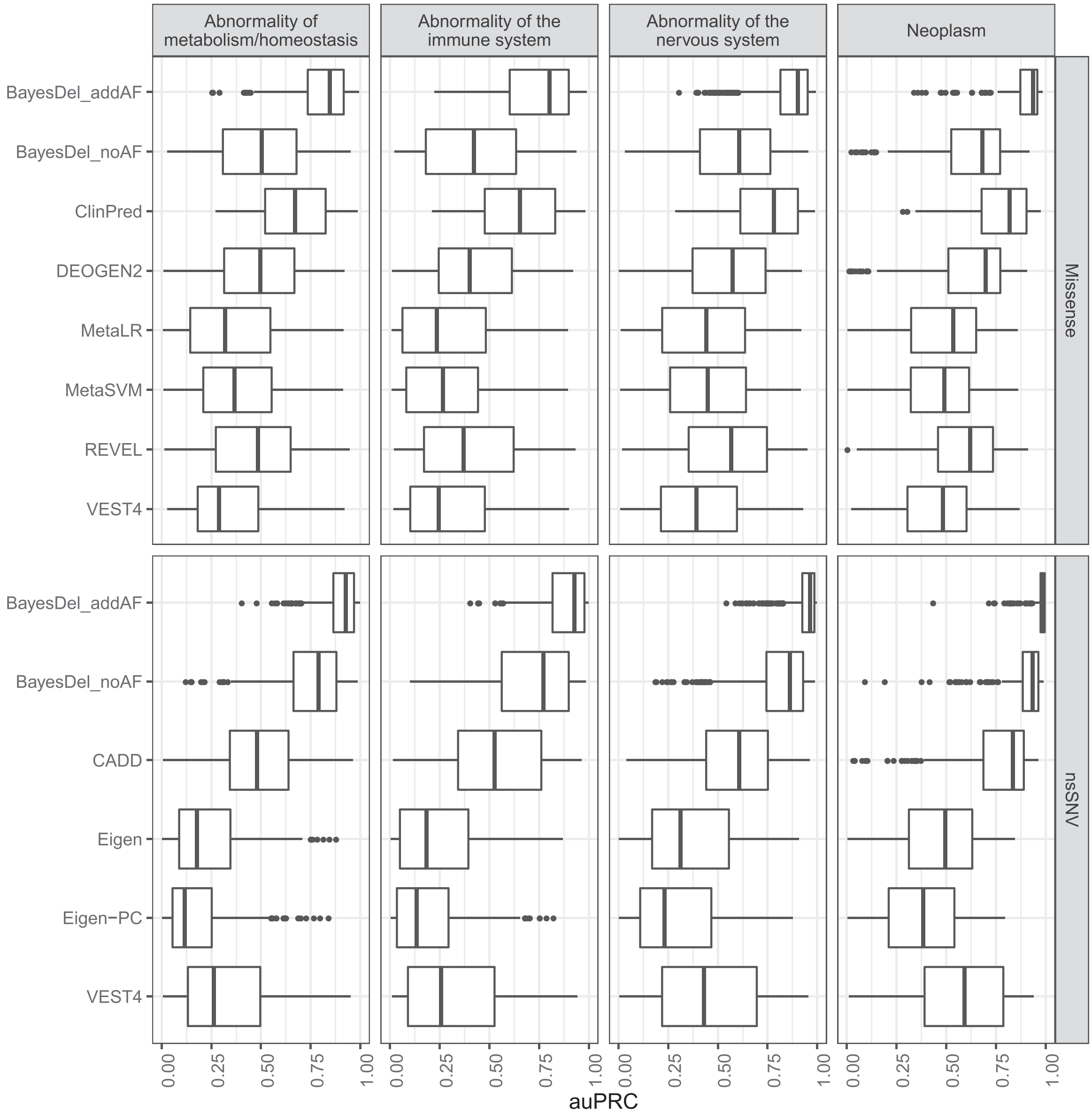

FIGURE 2 Boxplots showing auPRC of the top performing variant prioritization tools across selected top level HPO phenotypic abnormality terms and all their descendant terms for missense variants and nsSNVs. Abnormality of metabolism/homeostasis includes 340 terms for missense variants and 443 for nsSNVs. Abnormality of the immune system includes 242 terms for missense variants and 288 for nsSNVs. Abnormality of the nervous system includes 806 terms for missense variants and 898 for nsSNVs. Neoplasm includes 227 terms for missense variants and 291 for nsSNVs. auPRC, area under the precision-recall curve; nsSNV, nonsynonymous single nucleotide variant

Neoplasm. Though CADD was not a top performer for missense variants, it did exhibit strong performance for nsSNVs for terms associated with Neoplasm (median auPRC = 0.83) and moderate performance for the other three top level HPO terms (median auPRC range: 0.48–0.61).

When comparing the four top level HPO terms, strongest performance was seen across all tools for HPO terms associated with Neoplasm for both missense variants (median auPRC range: 0.48–0.94) and nsSNVs (median auPRC range: 0.38–0.99). For missense variants, all tools showed weakest performance for terms associated with Abnormality of the immune system (median auPRC range: 0.23–0.80). For nsSNVs, weakest performance was seen for terms associated with Abnormality of metabolism/homeostasis (median auPRC range: 0.12–0.93) and Abnormality of the immune system (median auPRC range: 0.13–0.93). In summary we found BayesDel_addAF to be the best performing tool for both missense variants and nsSNVs. Additionally, all tools exhibited stronger performance when prioritizing missense variants and nsSNVs for HPO terms associated with Neoplasm versus terms associated with abnormalities of metabolism/homeostasis, the immune system and the nervous system.

In summary, we found that the best performing variant prioritization tools differ depending on whether they are being used to prioritize missense variants or nsSNVs. Prioritization of missense variants is a more challenging task when compared to nonsense variants as nonsense variants usually affect protein function due to truncation. Whilst missense variants can also cause loss of protein function, the occurrence of this is rarer (around 20%) than that seen for nonsense variants (Kryukov et al., 2007).

The top performing tool in terms of auPRC for both missense variants and nsSNVs was BayesDel_addAF, with strongest performance seen for prioritization of nsSNVs. We also recommend ClinPred, the second best performer for missense variants as it showed consistent performance across a range of disease phenotypes. Whilst CADD was an overall weak performer for prioritizing missense variants, its overall performance for prioritizing nsSNVs was much improved. Hence, we also recommend CADD as a tool for prioritization of nsSNVs.

When considering performance based on PP25, BayesDel_addAF was again a top performer, consistently ranking ClinVar pathogenic variants within the top 25 ranked variants for both missense variants and nsSNVs across most HPO terms. However, in contrast to auPRC, strong performance was seen for conservation scores for both missense variants (LRT, phastCons100way and SIFT) and nsSNVs (LRT, phastCons30way, and phastCons100way). Similarly to the auPRC, CADD was also a strong performer for nsSNVs but not for missense variants.

Performance of the variant prioritization tools differs, even amongst the top performers, across the four top level HPO terms. Strongest performance for both missense variants and nsSNVs was seen for disease phenotypes associated with Neoplasm (HP:0002664). This is likely due to cancer being a more common disease that is better studied than rare diseases associated with Abnormality of metabolism/homeostasis (HP:0001939), Abnormality of the immune system (HP:0002715) and Abnormality of the nervous system (HP:0000707). This means pathogenic variants related to cancer will be overrepresented when compared to rarer diseases and hence also be overrepresented in training datasets of machine learning and ensemble methods. Furthermore, this points to the importance of developing tools that prioritize variants in a disease aware manner rather than the agnostic approach of the tools assessed here (Masica & Karchin, 2016).

In line with estimates of auPRC from our previous benchmark study (Anderson & Lassmann, 2018), we find that machine learning scores and ensemble scores show far superior performance than conservation scores when prioritizing variants across disease phenotypes. However, we do note that the training datasets used by machine learning and ensemble methods overlap in terms of the variants being assessed in this benchmark. This will result in more optimistic auPRC values for these methods in comparison to conservation methods. BayesDel and ClinPred in particular were trained on ClinVar pathogenic variants and given that our benchmark includes the same variants this will be contributing to their strong performance. Therefore, we cannot comment on whether the performance generalizes to yet unseen variants. Regardless of this, machine learning and ensemble methods can be expected to be superior to conservation methods as the pathogenicity of a variant can be predicted based on data that does not directly relate to conservation. Our benchmark is pragmatic in the sense that we focus on how these tools perform when used “out of the box” for the task of prioritizing variants. Though we do not recommend conservation measures based on the auPRC, some did perform well based on the PP25. In particular, LRT and phastCons100way were both strong performers for missense variants and nsSNVs.

In summary, we recommend use of BayesDel_addAF for prioritization of missense variants and nsSNVs. Given that in silico prediction tools have not reached the level of robustness required for clinical diagnostics (Richards et al., 2015; Strande et al., 2018), we further recommend use of ClinPred and CADD alongside BayesDel_addAF when prioritizing missense variants and nsSNVs respectively. BayesDel_addAF is also recommended for those who wish to examine a small number of top ranked missense variants or nsSNVs and for this task we further recommend simultaneous ranking with either LRT or phastCons100way. Of the five top performers we previously recommended (FATHMM, M-CAP, MetaLR, MetaSVM, and VEST3) (Anderson & Lassmann, 2018), MetaLR, MetaSVM and VEST4 (updated from VEST3) were amongst the top performing tools but their performance has been surpassed by new tools included in the current benchmark. The task of prioritizing variants remains a challenge, however the tools recommended here should prove useful for reducing the number of variants for follow up and ultimately contribute to disease diagnosis.

2 | METHODS

We previously described in detail our automated pipeline to integrate phenotypes with annotated variants (Anderson & Lassmann, 2018). Therefore, we only briefly describe each component and focus on describing updates to the benchmark.

2.1 | Human phenotype ontology

We used package ontologyIndex (Greene et al., 2017) within R 3.6.3 (R Core Team, 2020) to read in and process the HPO (Köhler et al., 2014) (HPO) obo file which was downloaded from http://purl.obolibrary.org/obo/hp.obo on the 28th of January 2021. We retrieved all 15,290 descendant terms of the Phenotypic abnormality (HP:0000118) term using the get_descendants() function.

2.2 | Linking disease phenotypes to genes using phenolyzer

Phenolyzer (Yang et al., 2015) was used to generate gene lists for the 15,290 HPO terms obtained above (File S1). We used the command line version available at https://github.com/WGLab/phenolyzer with default settings (i.e., options -p -ph -logistic -addon DB_DISGENET_GENE_DISEASE_SCORE,DB_GAD_GENE_DISEASE_SCORE -addon_weight 0.25).

2.3 | Linking candidate genes to causative variants using dbNSFP annotations

The database for nonsynonymous SNPs' functional predictions (dbNSFP) contains annotation for 84,013,490 potential nsSNVs and splicing-site SNVs in the human genome (Liu et al., 2011; Liu et al., 2020). We used dbNSFP version 4.1a (release 16 June, 2020) which is based on Gencode release 29/Ensembl version 94 (Cunningham et al., 2019; Frankish et al., 2019). We selected all variants occurring in the gene lists returned by Phenolyzer. We restricted our analysis to ClinVar (Landrum et al., 2016) "pathogenic" variants that were associated with a single gene. In total we obtained 35,167 pathogenic variants linked to genes associated with disease phenotypes (File S2). Of these, 16,411 were nonsense variants and 18,756 were missense variants.

2.4 | Benign variants

We used a set of 63,197 common (allele frequency ≥1% and <25%) missense variants obtained from the Exome Aggregation Consortium (ExAC) database (Niroula & Vihinen, 2019). These variants were downloaded from VariBench (Sasidharan Nair & Vihinen, 2013) (http://structure.bmc.lu.se/VariBench/ExAC_AAS_20171214.xlsx) and annotated with dbNSFP. We removed variants with ClinVar annotation other than "benign" and variants associated with more than one gene. We further filtered the variants to those 29,173 that had scores across all variant prioritization tools and used these in the benchmark analysis (File S3).

2.5 | Performance evaluation

For each HPO term, we evaluated the performance of variant prioritization tools by assessing their ability to separate ClinVar pathogenic variants from benign variants. These assessments were performed separately for nsSNVs and missense variants (i.e., excluding nonsense variants) as not all tools score nonsense variants. We required each HPO term to be associated with at least 25 pathogenic variants and to have complete scores across all tools. In total, for missense variants we tested 37 tools across 4890 HPO terms and for nsSNVs we tested 22 tools across 5723 HPO terms.

We assessed the following 22 variant prioritization tools that score nsSNVs: BayesDel (with and without allele frequency) (Feng, 2017), bStatistic (McVicker et al., 2009), CADD (Kircher et al., 2014; Rentzsch et al., 2019), DANN (Quang et al., 2015), Eigen (Ionita-Laza et al., 2016), Eigen-PC (Ionita-Laza et al., 2016), fathmm-MKL (Shihab et al., 2015), fathmm-XF (Rogers et al., 2018), fitCons-i6 (Gulko et al., 2015), GenoCanyon (Lu et al., 2015), GERP++ (Davydov et al., 2010), LRT (Chun & Fay, 2009), MutationTaster (Schwarz et al., 2014), phastCons (17way_primate, 30way_mammalian, 100way_vertebrate) (Siepel et al., 2005), phyloP (17way_primate, 30way_mammalian, 100way_vertebrate) (Siepel et al., 2006), SiPhy (Garber et al., 2009) and VEST4 (Carter et al., 2013). Additionally, we assessed a further 15 tools that only score missense variants: ClinPred (Alirezaie et al., 2018), DEOGEN2 (Raimondi et al., 2017), FATHMM (Shihab et al., 2013), LIST-S2 (Malhis et al., 2020), MetaLR (Dong et al., 2015), MetaSVM (Dong et al., 2015), MPC (Samocha et al., 2017), MutationAssessor (Reva et al., 2011), Polyphen2 (HDIV and HVAR) (Adzhubei et al., 2010), PrimateAI (Sundaram et al., 2018), PROVEAN (Choi et al., 2012), REVEL (Ioannidis et al., 2016), SIFT (Sim et al., 2012) and SIFT4G (Vaser et al., 2016). Further detail on the aforementioned tools is available in Table S1 of the dbNSFP v4 publication (Liu et al., 2020). We used the dbNSFP converted rank scores for each tool. We did not assess LINSIGHT (Huang et al., 2017) as this tool is focussed on prioritization of noncoding variants. We also omitted M-CAP (Jagadeesh et al., 2016), MutPred (Pejaver et al., 2020) and MVP (Qi et al., 2021) as these tools were missing scores for a substantial proportion of the benign variants.

We used R package PRROC (Keilwagen et al., 2014) to calculate the area under the precision recall curve (auPRC) based on the interpolation of Davis and Goadrich (Davis & Goadrich, 2006). We also constructed another performance measure called PP25 that calculates the proportion of ClinVar pathogenic variants in the top 25 ranked variants. Whilst the auPRC quantifies how well each tool can separate pathogenic variants from the whole set of benign variants, PP25 focuses on how well each tool does in ranking pathogenic variants amongst the top 25 most pathogenic. Heatmaps of performance (auPRC) were produced using the R NMF package (Gaujoux & Seoighe, 2010).

ACKNOWLEDGMENTS

This work was supported by the McCusker Charitable Foundation and the Stan Perron Foundation. Timo Lassmann is supported by a fellowship from the Feilman Foundation. Open access publishing facilitated by The University of Western Australia, as part of the Wiley - The University of Western Australia agreement via the Council of Australian University Librarians.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

AUTHOR CONTRIBUTIONS

Denise Anderson: performed analysis, interpreted results and drafted the manuscript. **Timo Lassmann**: conceived the study, interpreted results and drafted the manuscript.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available in Files S1, S2, and S3. Code used to generate results for this study is available as File S4.

ORCID

Denise Anderson http://orcid.org/0000-0003-0643-4136

REFERENCES

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., & Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature Methods*, *7*(4), 248–249. https://doi.org/10.1038/nmeth0410-248

Alirezaie, N., Kernohan, K. D., Hartley, T., Majewski, J., & Hocking, T. D. (2018). ClinPred: Prediction tool to identify disease-relevant nonsynonymous single-nucleotide variants. *American Journal of Human Genetics*, *103*(4), 474–483. https://doi.org/10.1016/j.ajhg.2018.08.005

Anderson, D., & Lassmann, T. (2018). A phenotype centric benchmark of variant prioritisation tools. *NPJ Genomic Medicine*, *3*, 5. https://doi.org/10.1038/s41525-018-0044-9

Carter, H., Douville, C., Stenson, P. D., Cooper, D. N., & Karchin, R. (2013). Identifying Mendelian disease genes with the variant effect scoring tool. *BMC Genomics*, *14*(Suppl 3), S3. https://doi.org/10.1186/1471-2164-14-S3-S3

Caspar, S. M., Dubacher, N., Kopps, A. M., Meienberg, J., Henggeler, C., & Matyas, G. (2018). Clinical sequencing: From raw data to diagnosis with lifetime value. *Clinical Genetics*, *93*(3), 508–519. https://doi.org/10.1111/cge.13190

Choi, Y., Sims, G. E., Murphy, S., Miller, J. R., & Chan, A. P. (2012). Predicting the functional effect of amino acid substitutions and indels. *PLoS One*, *7*(10), e46688. https://doi.org/10.1371/journal.pone.0046688

Chun, S., & Fay, J. C. (2009). Identification of deleterious mutations within three human genomes. *Genome Research*, *19*(9), 1553–1561. https://doi.org/10.1101/gr.092619.109

Cunningham, F., Achuthan, P., Akanni, W., Allen, J., Amode, M. R., Armean, I. M., Bennett, R., Bhai, J., Billis, K., Boddu, S., Cummins, C., Davidson, C., Dodiya, K. J., Gall, A., Girón, C. G., Gil, L., Grego, T., Haggerty, L., Haskell, E., … Flicek, P. (2019). Ensembl 2019. *Nucleic Acids Research*, *47*(D1), D745–D751. https://doi.org/10.1093/nar/gky1113

Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and ROC curves. *Proceedings of the 23rd International Conference on Machine Learning*, 233–240.

Davydov, E. V., Goode, D. L., Sirota, M., Cooper, G. M., Sidow, A., & Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Computational Biology*, *6*(12), e1001025. https://doi.org/10.1371/journal.pcbi.1001025

Dong, C., Wei, P., Jian, X., Gibbs, R., Boerwinkle, E., Wang, K., & Liu, X. (2015). Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Human Molecular Genetics*, *24*(8), 2125–2137. https://doi.org/10.1093/hmg/ddu733

Feng, B. J. (2017). PERCH: A unified framework for disease gene prioritization. *Human Mutation*, *38*(3), 243–251. https://doi.org/10.1002/humu.23158

Frankish, A., Diekhans, M., Ferreira, A. M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J. M., Sisu, C., Wright, J., Armstrong, J., Barnes, I., Berry, A., Bignell, A., Carbonell Sala, S., Chrast, J., Cunningham, F., Di Domenico, T., Donaldson, S., Fiddes, I. T., … Flicek, P. (2019). GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research*, *47*(D1), D766–D773. https://doi.org/10.1093/nar/gky955

Garber, M., Guttman, M., Clamp, M., Zody, M. C., Friedman, N., & Xie, X. (2009). Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics*, *25*(12), i54–i62. https://doi.org/10.1093/bioinformatics/btp190

Gaujoux, R., & Seoighe, C. (2010). A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics*, *11*, 367. https://doi.org/10.1186/1471-2105-11-367

Greene, D., Richardson, S., & Turro, E. (2017). ontologyX: A suite of R packages for working with ontological data. *Bioinformatics*, *33*(7), 1104–1106. https://doi.org/10.1093/bioinformatics/btw763

Gulko, B., Hubisz, M. J., Gronau, I., & Siepel, A. (2015). A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nature Genetics*, *47*(3), 276–283. https://doi.org/10.1038/ng.3196

Huang, Y. F., Gulko, B., & Siepel, A. (2017). Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nature Genetics*, *49*(4), 618–624. https://doi.org/10.1038/ng.3810

Ioannidis, N. M., Rothstein, J. H., Pejaver, V., Middha, S., McDonnell, S. K., Baheti, S., Musolf, A., Li, Q., Holzinger, E., Karyadi, D., Cannon-Albright, L. A., Teerlink, C. C., Stanford, J. L., Isaacs, W. B., Xu, J., Cooney, K. A., Lange, E. M., Schleutker, J., Carpten, J. D., … Sieh, W. (2016). REVEL: An ensemble method for predicting the pathogenicity of rare missense variants. *American Journal of Human Genetics*, *99*(4), 877–885. https://doi.org/10.1016/j.ajhg.2016.08.016

Ionita-Laza, I., McCallum, K., Xu, B., & Buxbaum, J. D. (2016). A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nature Genetics*, *48*(2), 214–220. https://doi.org/10.1038/ng.3477

Jagadeesh, K. A., Wenger, A. M., Berger, M. J., Guturu, H., Stenson, P. D., Cooper, D. N., Bernstein, J. A., & Bejerano, G. (2016). M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nature Genetics*, *48*(12), 1581–1586. https://doi.org/10.1038/ng.3703

Keilwagen, J., Grosse, I., & Grau, J. (2014). Area under precision-recall curves for weighted and unweighted data. *PLoS One*, *9*(3), e92209. https://doi.org/10.1371/journal.pone.0092209

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, *46*(3), 310–315. https://doi.org/10.1038/ng.2892

Köhler, S., Doelken, S. C., Mungall, C. J., Bauer, S., Firth, H. V., Bailleul-Forestier, I., Black, G. C., Brown, D. L., Brudno, M., Campbell, J., FitzPatrick, D. R., Eppig, J. T., Jackson, A. P., Freson, K., Girdea, M., Helbig, I., Hurst, J. A., Jähn, J., Jackson, L. G., … Robinson, P. N. (2014). The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Research*, *42*(Database issue), D966–D974. https://doi.org/10.1093/nar/gkt1026

Kryukov, G. V., Pennacchio, L. A., & Sunyaev, S. R. (2007). Most rare missense alleles are deleterious in humans: Implications for complex disease and association studies. *American Journal of Human Genetics*, *80*(4), 727–739. https://doi.org/10.1086/513473

Landrum, M. J., Lee, J. M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Hoover, J., Jang, W., Katz, K., Ovetsky, M., Riley, G., Sethi, A., Tully, R., Villamarin-Salomon, R., Rubinstein, W., & Maglott, D. R. (2016). ClinVar: Public archive of interpretations of clinically relevant variants. *Nucleic Acids Research*, *44*(D1), D862–D868. https://doi.org/10.1093/nar/gkv1222

Liu, X., Jian, X., & Boerwinkle, E. (2011). dbNSFP: A lightweight database of human nonsynonymous SNPs and their functional predictions. *Human Mutation*, *32*(8), 894–899. https://doi.org/10.1002/humu.21517

Liu, X., Li, C., Mou, C., Dong, Y., & Tu, Y. (2020). dbNSFP v4: A comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Genome Medicine*, *12*(1), 103. https://doi.org/10.1186/s13073-020-00803-9

Lu, Q., Hu, Y., Sun, J., Cheng, Y., Cheung, K. H., & Zhao, H. (2015). A statistical framework to predict functional non-coding regions in the human genome through integrated analysis of annotation data. *Scientific Reports*, *5*, 10576. https://doi.org/10.1038/srep10576

Malhis, N., Jacobson, M., Jones, S. J. M., & Gsponer, J. (2020). LIST-S2: Taxonomy based sorting of deleterious missense mutations across species. *Nucleic Acids Research*, *48*(W1), W154–W161. https://doi.org/10.1093/nar/gkaa288

Masica, D. L., & Karchin, R. (2016). Towards increasing the clinical relevance of in silico methods to predict pathogenic missense variants. *PLoS Computational Biology*, *12*(5), e1004725. https://doi.org/10.1371/journal.pcbi.1004725

McVicker, G., Gordon, D., Davis, C., & Green, P. (2009). Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genetics*, *5*(5), e1000471. https://doi.org/10.1371/journal.pgen.1000471

Niroula, A., & Vihinen, M. (2019). How good are pathogenicity predictors in detecting benign variants? *PLoS Computational Biology*, *15*(2), e1006481. https://doi.org/10.1371/journal.pcbi.1006481

Pejaver, V., Urresti, J., Lugo-Martinez, J., Pagel, K. A., Lin, G. N., Nam, H. J., Mort, M., Cooper, D. N., Sebat, J., Iakoucheva, L. M., Mooney, S. D., & Radivojac, P. (2020). Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. *Nature Communications*, *11*(1), 5918. https://doi.org/10.1038/s41467-020-19669-x

Qi, H., Zhang, H., Zhao, Y., Chen, C., Long, J. J., Chung, W. K., Guan, Y., & Shen, Y. (2021). MVP predicts the pathogenicity of missense variants by deep learning. *Nature Communications*, *12*(1), 510. https://doi.org/10.1038/s41467-020-20847-0

Quang, D., Chen, Y., & Xie, X. (2015). DANN: A deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics*, *31*(5), 761–763. https://doi.org/10.1093/bioinformatics/btu703

R Core Team. (2020). R: A Language and Environment for Statistical Computing, *R Foundation for Statistical Computing*.

Raimondi, D., Tanyalcin, I., Ferté, J., Gazzo, A., Orlando, G., Lenaerts, T., Rooman, M., & Vranken, W. (2017). DEOGEN2: Prediction and interactive visualization of single amino acid variant deleteriousness in human proteins. *Nucleic Acids Research*, *45*(W1), W201–W206. https://doi.org/10.1093/nar/gkx390

Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., & Kircher, M. (2019). CADD: Predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research*, *47*(D1), D886–D894. https://doi.org/10.1093/nar/gky1016

Reva, B., Antipin, Y., & Sander, C. (2011). Predicting the functional impact of protein mutations: Application to cancer genomics. *Nucleic Acids Research*, *39*(17), e118. https://doi.org/10.1093/nar/gkr407

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W. W., Hegde, M., Lyon, E., Spector, E., Voelkerding, K., Rehm, H. L., & ACMG Laboratory Quality Assurance, C. (2015). 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. *Genetics in Medicine*, *17*(5), 405–424. https://doi.org/10.1038/gim.2015.30

Rogers, M. F., Shihab, H. A., Mort, M., Cooper, D. N., Gaunt, T. R., & Campbell, C. (2018). FATHMM-XF: Accurate prediction of pathogenic point mutations via extended features. *Bioinformatics*, *34*(3), 511–513. https://doi.org/10.1093/bioinformatics/btx536

Samocha, K. E., Kosmicki, J. A., Karczewski, K. J., O'Donnell-Luria, A. H., Pierce-Hoffman, E., MacArthur, D. G., Neale, B. M., & Daly, M. J. (2017). Regional missense constraint improves variant deleteriousness prediction. *bioRxiv*

Sasidharan Nair, P., & Vihinen, M. (2013). VariBench: A benchmark database for variations. *Human Mutation*, *34*(1), 42–49. https://doi.org/10.1002/humu.22204

Schwarz, J. M., Cooper, D. N., Schuelke, M., & Seelow, D. (2014). MutationTaster2: Mutation prediction for the deep-sequencing age. *Nature Methods*, *11*, 11–12. https://doi.org/10.1038/nmeth.2890

Shihab, H. A., Gough, J., Cooper, D. N., Stenson, P. D., Barker, G. L., Edwards, K. J., Day, I. N., & Gaunt, T. R. (2013). Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Human Mutation*, *34*(1), 57–65. https://doi.org/10.1002/humu.22225

Shihab, H. A., Rogers, M. F., Gough, J., Mort, M., Cooper, D. N., Day, I. N., Gaunt, T. R., & Campbell, C. (2015). An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics*, *31*(10), 1536–1543. https://doi.org/10.1093/bioinformatics/btv009

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., Weinstock, G. M., Wilson, R. K., Gibbs, R. A., Kent, W. J., Miller, W., & Haussler, D. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, *15*(8), 1034–1050. https://doi.org/10.1101/gr.3715005

Siepel, A., Pollard, K., & Haussler, D. (2006). New Methods for Detecting Lineage-Specific Selection In *Research in Computational Molecular Biology* (3909, pp. 190–205). Springer.

Sim, N. L., Kumar, P., Hu, J., Henikoff, S., Schneider, G., & Ng, P. C. (2012). SIFT web server: Predicting effects of amino acid substitutions on proteins. *Nucleic Acids Research*, *40*(Web Server issue), W452–W457. https://doi.org/10.1093/nar/gks539

Strande, N. T., Brnich, S. E., Roman, T. S., & Berg, J. S. (2018). Navigating the nuances of clinical sequence variant interpretation in Mendelian disease. *Genetics in Medicine*, *20*(9), 918–926. https://doi.org/10.1038/s41436-018-0100-y

Sundaram, L., Gao, H., Padigepati, S. R., McRae, J. F., Li, Y., Kosmicki, J. A., Fritzilas, N., Hakenberg, J., Dutta, A., Shon, J., Xu, J., Batzoglou, S., Li, X., & Farh, K. K. (2018). Predicting the clinical impact of human mutation with deep neural networks. *Nature Genetics*, *50*(8), 1161–1170. https://doi.org/10.1038/s41588-018-0167-z

Vaser, R., Adusumalli, S., Leng, S. N., Sikic, M., & Ng, P. C. (2016). SIFT missense predictions for genomes. *Nature Protocols*, *11*(1), 1–9. https://doi.org/10.1038/nprot.2015.123

Yang, H., Robinson, P. N., & Wang, K. (2015). Phenolyzer: Phenotype-based prioritization of candidate genes for human diseases. *Nature Methods*, *12*(9), 841–843. https://doi.org/10.1038/nmeth.3484

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Anderson, D., Lassmann, T. (2022). An expanded phenotype centric benchmark of variant prioritisation tools. *Human Mutation*, 43, 539–546. https://doi.org/10.1002/humu.24362