Supplementary Material for *Calibration of additional computational tools expands ClinGen recommendation options for variant classification with PP3/BP4 criteria*

Timothy Bergquist, Sarah L. Stenton, Emily A.W. Nadeau, Alicia B. Byrne, Marc S. Greenblatt, Steven M. Harrison, Sean V. Tavtigian, Anne O'Donnell-Luria, Leslie G. Biesecker, Predrag Radivojac, Steven E. Brenner, Vikas Pejaver, ClinGen Sequence Variant Interpretation Working Group

SUPPLEMENTARY MATERIALS AND METHODS

Terminology and definitions

As in our previous work,[1] we use the terms likelihood ratio and posterior probability interchangeably because the likelihood ratio is the ratio of the posterior odds to the prior odds, and since the prior probability of pathogenicity is fixed for this study, there is a one-to-one mapping between likelihood ratios and posterior probabilities.

Mapping computational tool outputs to calibration data sets

For AlphaMissense, we used three complementary mapping approaches. First, we linked precomputed scores to our data sets using chromosomal coordinates and Ensembl transcript identifiers as the key.[2] Second, to ensure that the correct isoform was being considered, we undertook the mapping based on the Ensembl transcript identifier and amino acid substitution. Third, we undertook an additional mapping based on UniProt protein identifiers, using the corresponding mapping file provided by AlphaMissense.[3] For ESM1b, we mapped precomputed scores to our data sets using the provided UniProt identifiers (with and without isoform-specificity) and amino acid substitutions. For variants that still remained unmapped, we used dbNSFP v4.4a[4] to reannotate our variant list with the most up-to-date UniProt annotations, which were in turn used to map precomputed scores to our data sets. For EVE, we first mapped variants using UniProt or Ensembl transcript identifiers and amino acid substitution. We further matched all remaining unmapped variants to the UniProt gene name and amino acid substitution. For VARITY, we first mapped precomputed scores to variants in our data sets using UniProt protein identifiers, without consideration of the specific isoform. We then mapped the remaining variants strictly using chromosomal coordinates.

Despite our best efforts, the gap in time between data set construction and tool publication meant that there inexorably would be irreconcilable differences among gene, protein and/or variant identifiers in our data sets compared to the files with precomputed scores for some tools. We expect this to be a major issue only if the differences in missing data were non-

random, which was not the case for all methods but one (average proportion of missing-at-random scores < 10%). In the case of EVE,[5] predictions were available only for a subset of genes in our calibration set, specifically leaving about half of the benign/likely benign variants in our data set unscored, potentially introducing sampling bias. For this reason, we could not reliably calibrate EVE in this study.

REFERENCES

1. Pejaver V, Byrne AB, Feng BJ, et al. Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for PP3/BP4 criteria. *Am J Hum Genet*. 2022;109(12):2163-2177. doi:10.1016/j.ajhg.2022.10.013

2. Martin FJ, Amode MR, Aneja A, et al. Ensembl 2023. *Nucleic Acids Res*. 2023;51(D1):D933-D941. doi:10.1093/nar/gkac958

3. The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res*. 2023;51(D1):D523-D531. doi:10.1093/nar/gkac1052

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

5. Frazer J, Notin P, Dias M, et al. Disease variant prediction with deep generative models of evolutionary data. *Nature*. 2021;599(7883):91-95. doi:10.1038/s41586-021-04043-8