Pervasive ancestry bias in variant effect predictors

Ankit K. Pathak[1], Nikita Bora[2], Mihaly Badonyi[1], Benjamin J. Livesey[1], Joanne Ngeow[2,3,4]* and Joseph A. Marsh[1]*

[1]*MRC Human Genetics Unit, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, UK*

[2]*Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore*

[3]*Oncology Academic Clinical Program, Duke-NUS Medical School, Singapore*

[4]*Cancer Genetics Service, Division of Medical Oncology, National Cancer Centre Singapore, Singapore*

**Correspondence to: joanne.ngeow@ntu.edu.sg; joseph.marsh@ed.ac.uk*

Abstract

Computational variant effect predictors (VEPs) are playing increasingly important roles in the interpretation of human genetic variants. We observe striking differences in the ways that many VEPs score missense variants from European compared to non-European populations. We advocate for the adoption of population-free VEPs, i.e. those not trained on human population or clinical variants, to improve health equity and enhance the accuracy of genetic diagnoses across diverse populations.

Main

The rapid advance of population sequencing technologies has led to the creation of an extensive catalogue of human genomic variants, significantly enhancing our capacity for diagnosis, treatment, and prevention of genetic diseases. Historically, genomic studies have predominantly concentrated on European populations, introducing biases in our understanding of genetic variation and its impacts[1]. Although recent shifts towards including more diverse cohorts are beginning to mitigate these biases, many sequence resources have still remained predominantly European. Increasing genomic inclusivity is crucial not only for promoting health equity[2] but also for enriching our understanding of human biology[3]. By examining a broader array of genetic diversity, we gain deeper insights into the complex mechanisms underlying genetic diseases across different populations.

Variant effect predictors (VEPs) are computational tools designed to predict the phenotypic impacts of variants, often in terms of potential pathogenicity[4]. Many such tools have been developed over the years, and they have seen continual gains in their ability to identify damaging missense variants[5]. While, currently, they cannot routinely be used as anything more than supporting evidence in making genetic diagnoses[6], they are widely used in clinical variant prioritisation.

Most currently used VEPs are trained using supervised learning methods on datasets of known human variants, including those known to be pathogenic and benign, or those observed in the human population. We refer to these as *population-based* VEPs and, while they are widely used, they are associated with concerns regarding data circularity, with methods being strongly biased by what they are trained on[5,7]. Given the historical focus of human sequencing and interpretation on individuals of European heritage, this could lead to ancestry bias in these VEPs. Even with the increase in non-European sequencing, European

variants may be more likely to have been investigated in more depth, and therefore more likely to be formally classified as pathogenic or benign. As VEPs become more commonly used in variant prioritisation and diagnosis, these biases increase the risk of misdiagnosis and might make clinicians altogether miss genetic disorders in non-European populations.

A second class of VEPs, the *population-free* methods should be free of ancestry bias, as they are not trained on clinical or human population variants. This group largely overlaps with unsupervised models, but excludes those that use human allele frequency as a feature, and can include supervised models trained on other data, *e.g.*, deep mutational scanning datasets. In recent years, several population-free VEPs have shown remarkable performance, often better than any population-based methods in terms of correspondence with functional assays and identification of pathogenic variants[5]. However, they are currently less widely used in clinical genome interpretation, likely in part due to not having been explicitly trained for predicting pathogenicity.

To investigate the issue of potential ancestry bias in VEPs, we compared how they score missense variants from different populations. We illustrate this for three population-free VEPs: EVE[8], CPT[9] and GEMME[10] (Fig. 1A). First, we compare the distributions of scores for variants observed in known disease genes from Europeans (EUR) to different populations, including Admixed American (AMR), East Asian (EAS) and South Asian (SAS) from gnomAD[11]; and Indigenous Mexican (IMX) and African (AMC) from Mexico City[12], filtering for allele frequency between 0.1-1% to ensure comparability between populations of different sizes (Figure S1). Our findings reveal largely similar distributions across these groups, with only minor variations. If we establish optimal pathogenicity thresholds for each VEP based on discrimination between known pathogenic and benign variants[13], then the relative fraction predicted pathogenic by each VEP is similar across groups; for example, CPT predicts 1.2 times more pathogenic variants in IMX than in EUR, and 0.9 times fewer in AMC.

In Fig. 1B, we present the same analysis on three population-based VEPs: ClinPred[14], BayesDel[15] and AlphaMissense[16]. Here, the differences in distributions are far larger. For example, ClinPred predicts 4.0 times more pathogenic variants in AMC than EUR, while BayesDel predicts 1.4 times more in IMX than EUR. AlphaMissense is worthy of further note, being largely an unsupervised model, but tuned to human allele frequencies, which has the potential to introduce bias in clinical variant assessment[17]. Overall, it shows much smaller differences than the other population-based VEPs, but still notably larger than the population-free models, suggesting that it is biased to some extent by the composition of the variants on which it is tuned.

In Fig. 2, we scale up this analysis to consider 15 population-free and 37 population-based VEPs across more ancestry groups from gnomAD[11], including African/African American (AFR) and Middle Eastern (MID). Here, the heatmap represents the log of the relative enrichment in predicted pathogenic variants compared to EUR, calculated the same way as in Figure 1. VEPs are sorted according to the absolute log of the relative enrichment, averaged across all populations, so that methods that show larger differences in non-EUR *vs* EUR across populations, regardless of direction of difference, will have larger values. Strikingly, the 24 methods showing the biggest differences between European and non-European populations are all population-based VEPs, strongly suggesting that many of these models are biased based on training populations.

One possible argument is that the population-based VEPs are simply better at identifying genuine differences between groups. Indeed, if tested purely for their ability to discriminate

between known pathogenic and benign variants (Figure S2), several of the top methods are the same that show the largest differences between populations. However, this is almost certainly confounded by circularity issues, as these top methods will have been directly trained on many of the tested variants and genes[7]. Several of these methods were shown to have large discrepancies between their performance in clinical variant classification and against functional assays, highly suggestive of overfitting to the clinical and population variants[5]. Moreover, despite these circularity issues, some of population-free VEPs perform better at pathogenic *vs* benign discrimination than population-based VEPs that show much larger population differences.

It is also notable that certain population-based VEPs are classifying fewer variants as predicted pathogenic, using our optimal pathogenicity threshold approach. For example, ClinPred predicts 2.26% of EUR variants to be pathogenic, compared to 19.81% for CPT. This is an inevitable result of their population-based design as they will be highly unlikely to score population variants present in their training data as pathogenic. Importantly, however, even if we set the threshold so that each model predicts the same proportion of EUR variants to be pathogenic, we see the same trend for population-based *vs* population-free VEPs (Figure S3).

Interestingly, the population-level bias in VEPs can go in both directions: either predicting variants from different non-European populations to be relatively more or less pathogenic. For example, several population-based VEPs find AFR variants to be relatively deficient in predicted pathogenicity compared to population-free methods, but IMX variants to be relatively enriched. There are multiple factors likely contributing to this. On one hand, European population variants are more likely to have been seen by the VEPs and therefore scored as benign. However, pathogenic variants in European populations are also more likely to have been classified as pathogenic[18]. In addition, although we have attempted to control for allele frequency at the population level to improve comparability between different ancestry groups, we do see that certain populations have more common variants than others (Figure S1). As AFR has the most common variants, VEPs that directly incorporate information about allele frequency may be more likely to score them as benign. Importantly, although there are clearly interesting differences between populations in terms of variant composition, we wish to emphasise that this is not the focus of our study.

While a few of the population-based VEPs appear to show very severe levels of bias in this study, many others show smaller differences between populations that nevertheless are larger than observed for the population-free methods. These biases are unlikely to strongly influence analyses involving rare variants, *e.g.*, severe disease caused by *de novo* mutations. However, as population-level analyses of variants become more common, it is important to consider the impacts that training on different populations can potentially have. While one potential strategy is to train models on more representative populations, the population variants will always lead to some form of bias. Therefore we suggest that the safest strategy is to use population-free VEPs to ameliorate these issues. It is also important to note that not all variant effect prediction strategies that incorporate population data are necessarily susceptible to this bias. For example, popEVE uses patterns of population variation to calibrate variant effect scores from gene to gene, but knowledge of specific population variants is not used in scoring[19]; it is consistent with the population-free VEPs in terms of its differences between populations. We also note the great potential of

high-throughput experimental approaches involving multiplexed assays of variant effect (MAVEs) for scoring population variants in an unbiased manner[20], although the number of genes for which these measurements are available is still very limited.

Methods

Population datasets

This study utilised two population sequencing datasets to gather genomic variants in various ancestries — gnomAD v4.1[11] and Mexico City Prospective Study (MCPS)[12]. Only single nucleotide variants with a "PASS" quality filter were extracted from the variant calling files provided by the two projects. Ensembl VEP 110[21] was then used to map the variants to UniProt reference sequences of releases 2023_03 (MCPS) and 2024_02 (gnomAD v4.1).

In order to build the European cohort, variants present in the non-Finnish European (NFE) population in gnomAD and the European (EUR) population in MCPS were purposed. For the non-European cohorts, variants from the following populations were used — African/African American (AFR), Admixed American (AMR), East Asian (EAS), Middle Eastern (MID), and South Asian (SAS) in gnomAD; African (AMC) and Indigenous Mexican (IMX) in MCPS. We excluded the Finnish and Ashkenazi Jewish populations in gnomAD from the analyses, given their genetic proximity to Europeans.

To draw a fair comparison, it was important to use similar sets of variants with respect to their penetrance in the populations to assess the predicted enrichment of pathogenic variants. Thus, for both European and non-European cohorts, only variants with an allele frequency between 0.1% and 1% in that population were considered. This meant that the number of variants used for the comparisons and their distribution with respect to allele frequencies were similar (Figure S1). Further, we only considered variants in genes with reported "pathogenic" or "likely pathogenic" missense variants in ClinVar, thus limiting our analyses to only known disease genes.

Optimal pathogenicity thresholds

The *scikit-learn* python package's *metrics.roc_curve* function was used to calculate the receiver operating characteristic (ROC) area under the curve (AUC) scores for each VEP using 47,336 "pathogenic" or "likely pathogenic" and 46,907 "benign" or "likely benign" labelled variants in ClinVar[22] (downloaded on 16th November 2023). The pathogenic variants were labelled as true positives and benign variants as true negatives. For VEPs with an inverted scale (lower scores being more pathogenic), the labels were inverted too. The threshold at which the difference between the True Positive Rate (TPR) and the False Positive Rate (FPR) was maximum was chosen as the optimal pathogenicity threshold for each VEP.

Log odds ratios

The determined optimal pathogenicity thresholds were used to evaluate each VEP's predicted number of pathogenic and benign variants for each population. Thereafter, to assess VEP-wise relative enrichment of predicted pathogenic variants in each non-European population compared to the European, the log odds ratio was calculated using *SciPy* python package's *stats.fisher_exact* function. This was equivalent to:

$$Log\ Odds\ Ratio\ =\ ln\left(\left(\frac{path_{non-EUR}}{ben_{non-EUR}}\right)/\left(\frac{path_{EUR}}{ben_{EUR}}\right)\right)$$

where,

$path_{non-EUR}$ = Number of predicted pathogenic variants in a non-European population

$ben_{non-EUR}$ = Number of predicted benign variants in a non-European population

$path_{EUR}$ = Number of predicted pathogenic variants in the European population

ben_{EUR} = Number of predicted benign variants in the European population

Variant effect predictors

A total of 73 VEPs were initially considered for this analysis. The VEP scores were collected using the in-house pipeline previously described[5]. To filter for VEPs with reasonably good performance, only VEPs with a ROC AUC score of ≥0.8 in distinguishing between ClinVar (likely) pathogenic and (likely) benign were chosen. In addition, VEPs with a coverage of less than 50% variants in any population were also filtered out, leaving 52 VEPs for the comparisons.

Acknowledgements

This project was supported by funding to JAM from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 101001169), a Lister Institute Research Prize Fellowship and by the Medical Research Council (MRC) Human Genetics Unit core grant (MC_UU_00035/9). JN is supported by an NMRC Clinician Scientist Award.

Data availability

Complete datasets associated with the analyses in this study are available at https://osf.io/sez98.

References

1. Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. *Nature* **538**, 161–164 (2016).
2. Madden, E. B. *et al.* Advancing genomics to improve health equity. *Nat Genet* 1–6 (2024) doi:10.1038/s41588-024-01711-z.
3. Sirugo, G., Williams, S. M. & Tishkoff, S. A. The Missing Diversity in Human Genetic Studies. *Cell* **177**, 26–31 (2019).
4. Livesey, B. J. *et al.* Guidelines for releasing a variant effect predictor. Preprint at https://doi.org/10.48550/arXiv.2404.10807 (2024).
5. Livesey, B. J. & Marsh, J. A. Variant effect predictor correlation with functional assays is reflective of clinical classification performance. Preprint at https://doi.org/10.1101/2024.05.12.593741 (2024).
6. Richards, S. *et al.* Standards and Guidelines for the Interpretation of Sequence Variants: A Joint Consensus Recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* **17**, 405–424 (2015).

7. Grimm, D. G. *et al.* The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. *Hum. Mutat.* **36**, 513–523 (2015).
8. Frazer, J. *et al.* Disease variant prediction with deep generative models of evolutionary data. *Nature* **599**, 91–95 (2021).
9. Jagota, M. *et al.* Cross-protein transfer learning substantially improves disease variant prediction. *Genome Biology* **24**, 182 (2023).
10. Laine, E., Karami, Y. & Carbone, A. GEMME: A Simple and Fast Global Epistatic Model Predicting Mutational Effects. *Mol Biol Evol* **36**, 2604–2619 (2019).
11. Chen, S. *et al.* A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* **625**, 92–100 (2024).
12. Ziyatdinov, A. *et al.* Genotyping, sequencing and analysis of 140,000 adults from Mexico City. *Nature* **622**, 784–793 (2023).
13. Greiner, M., Pfeiffer, D. & Smith, R. D. Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. *Preventive Veterinary Medicine* **45**, 23–41 (2000).
14. Alirezaie, N., Kernohan, K. D., Hartley, T., Majewski, J. & Hocking, T. D. ClinPred: Prediction Tool to Identify Disease-Relevant Nonsynonymous Single-Nucleotide Variants. *The American Journal of Human Genetics* **103**, 474–483 (2018).
15. Feng, B.-J. PERCH: A Unified Framework for Disease Gene Prioritization. *Human Mutation* **38**, 243–251 (2017).
16. Cheng, J. *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023).
17. Marsh, J. A. & Teichmann, S. A. Predicting pathogenic protein variants. *Science* **381**, 1284–1285 (2023).
18. Venner, E. *et al.* The frequency of pathogenic variation in the All of Us cohort reveals ancestry-driven disparities. *Commun Biol* **7**, 1–11 (2024).
19. Orenbuch, R. *et al.* Deep generative modeling of the human proteome reveals over a hundred novel genes involved in rare genetic disorders. 2023.11.27.23299062 Preprint at https://doi.org/10.1101/2023.11.27.23299062 (2023).
20. Dawood, M. *et al.* Defining and Reducing Variant Classification Disparities. 2024.04.11.24305690 Preprint at https://doi.org/10.1101/2024.04.11.24305690 (2024).
21. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biology* **17**, 122 (2016).
22. Landrum, M. J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42**, D980-5 (2014).

Figures

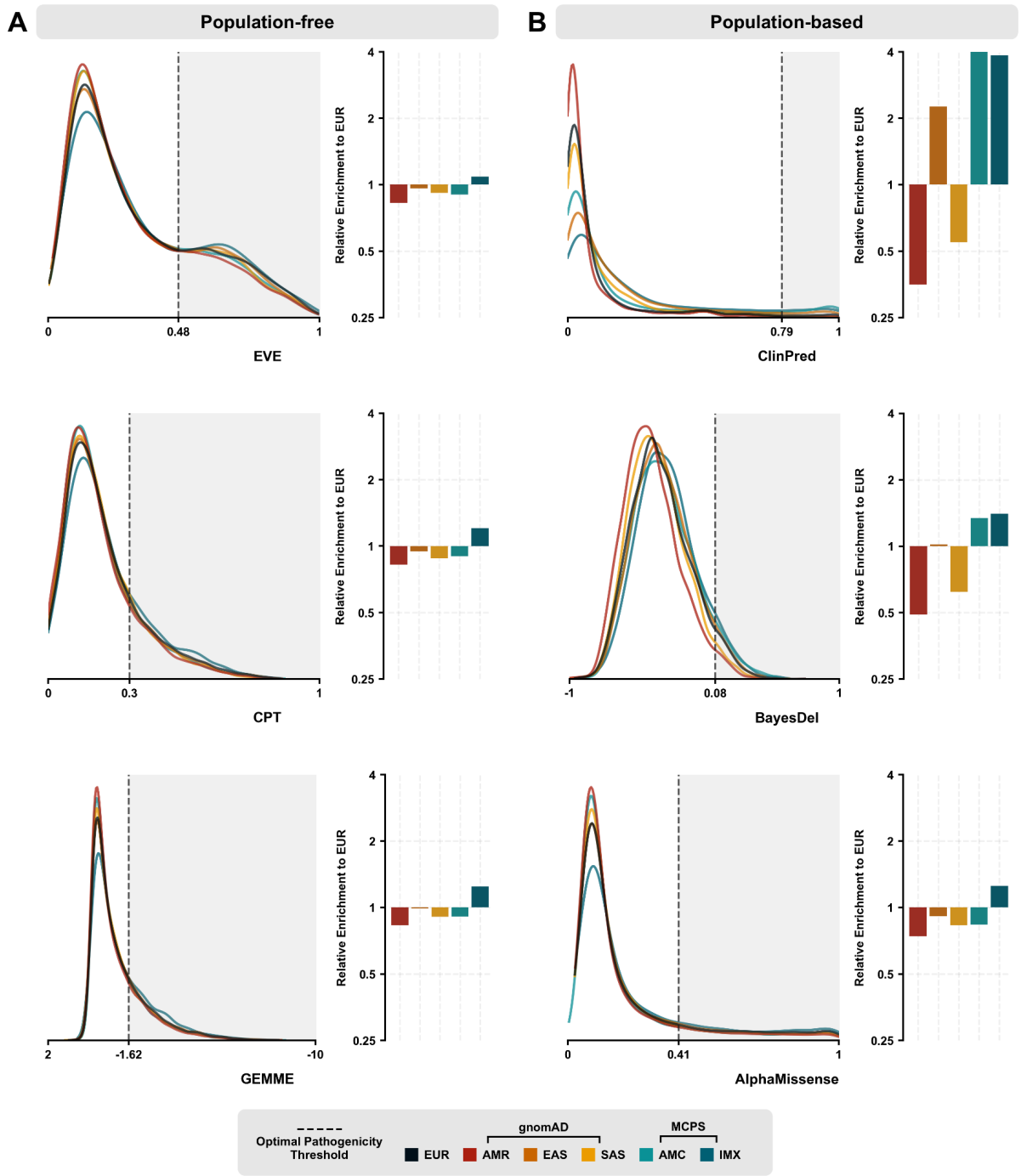

Figure 1. Distributions of scores and predicted pathogenicity for select VEPs across different populations. (A) Representative population-free VEPs. (B) Representative population-based VEPs. The left panels show the distribution of VEP scores across the European and five non-European populations. For each VEP, the optimal pathogenicity threshold is indicated, designating variants to the right of that point (shaded grey) as predicted to be pathogenic. The right panels illustrate the relative enrichment in predicted pathogenic variants in each population in comparison to the European.

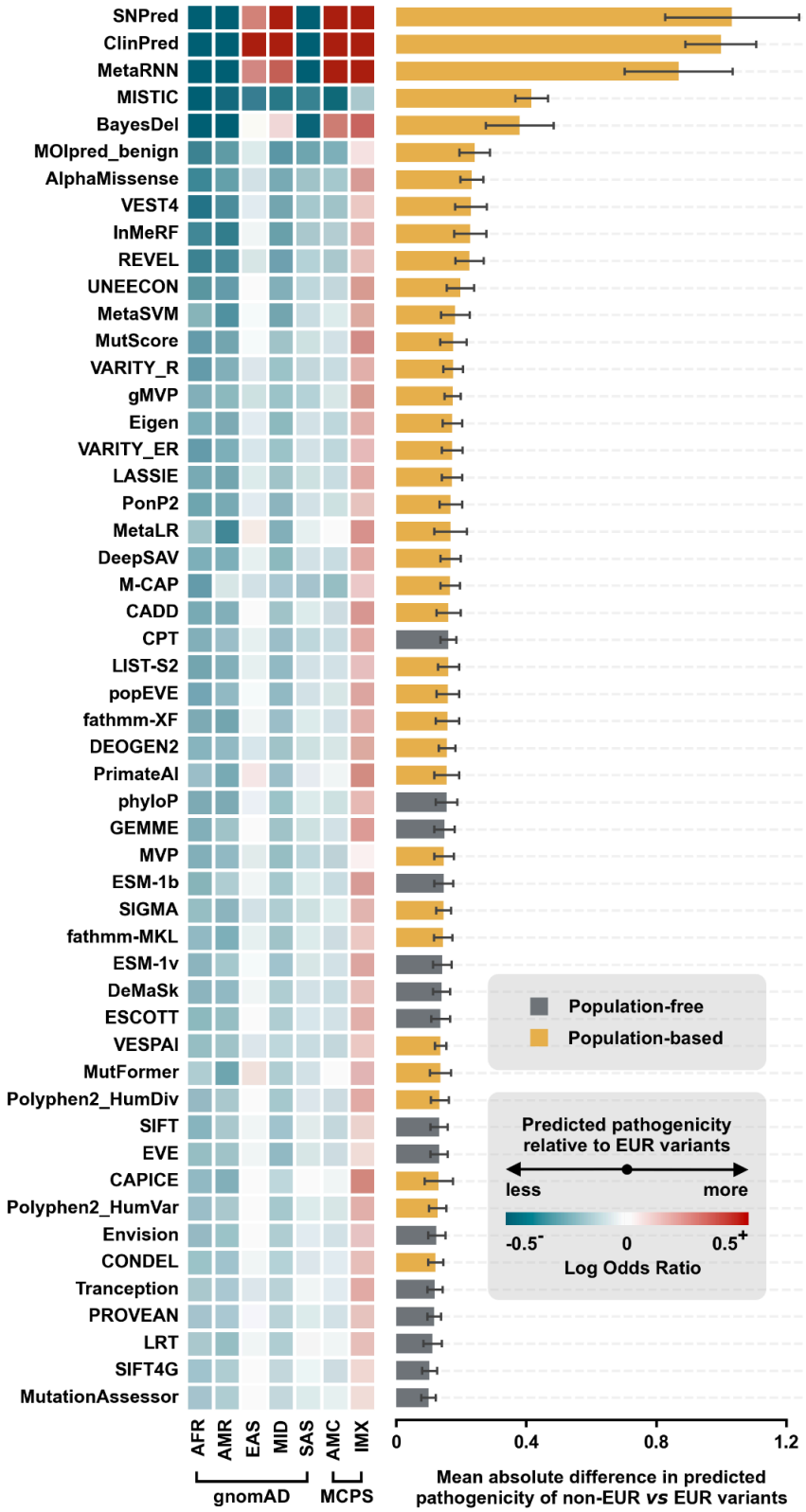

Figure 2. Relative enrichment of predicted pathogenic variants by all VEPs across all non-European populations compared to the European. The heatmap showcases the log odds ratio of predicted pathogenic variants in each non-EUR population *vs* EUR, calculated the same as the right panel of Figure 1. The bar plot indicates the mean of the absolute log odds ratio across all non-EUR populations, while error bars represent standard error.

Supplementary Figures

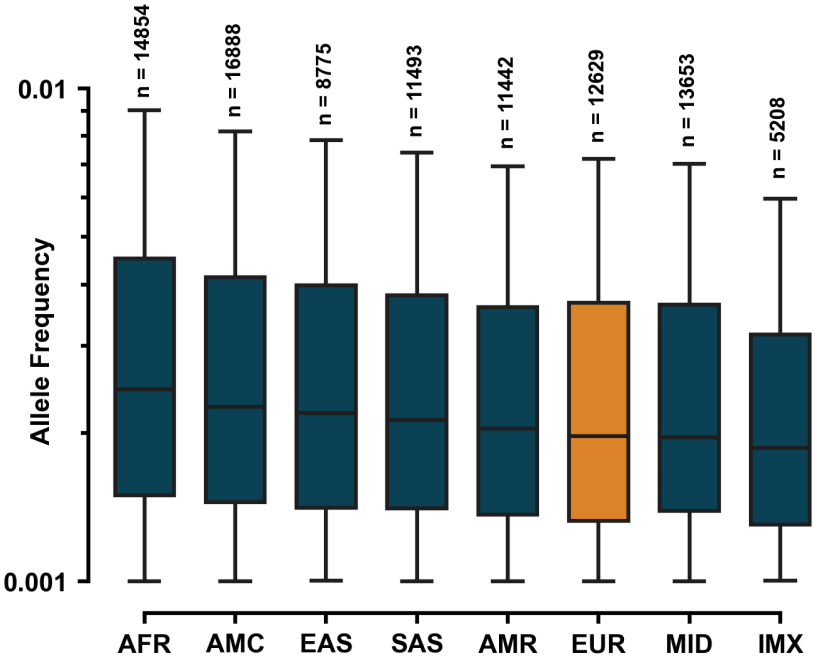

Figure S1. Distribution of allele frequencies of variants across each population. The box plot illustrates the distribution of allele frequencies for missense variants utilised in this study across each population, organised by descending median allele frequency. The number of variants considered in each population is also indicated. To ensure comparability across populations, only variants with allele frequencies ranging from 0.1% to 1% were considered. This approach prevents populations sequenced more extensively from skewing the data with an excess of rare variants, which would confound analyses of population differences.

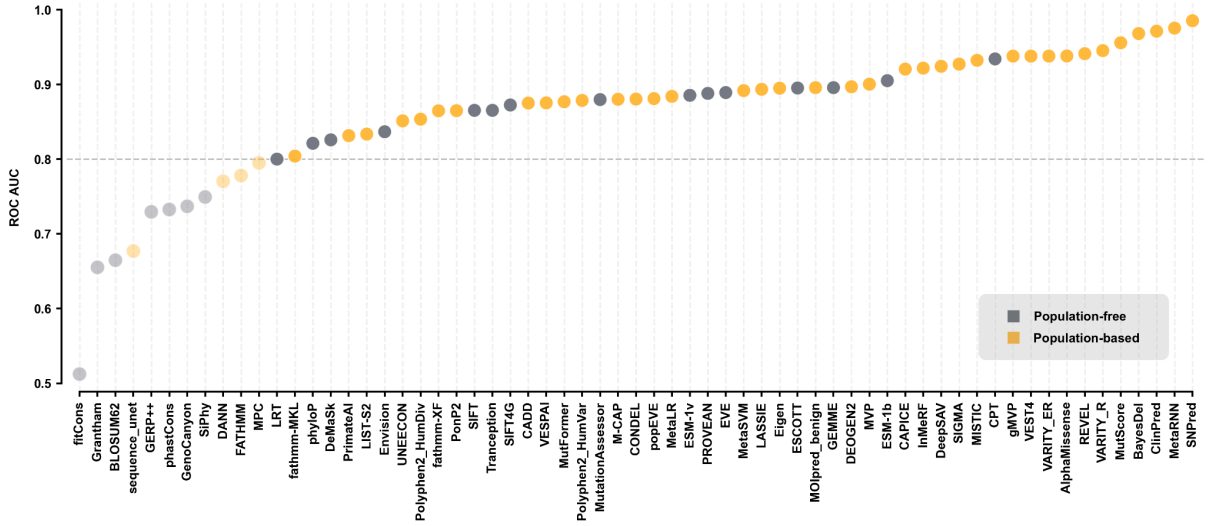

Figure S2. Performance of all VEPs in distinguishing between known pathogenic and benign variants. For each VEP, the receiver operating characteristic (ROC) area under the curve (AUC) was calculated for the discrimination between missense variants classified as (likely) pathogenic or (likely) benign in ClinVar. Only VEPs with a ROC AUC ≥0.8 were considered for the analyses.

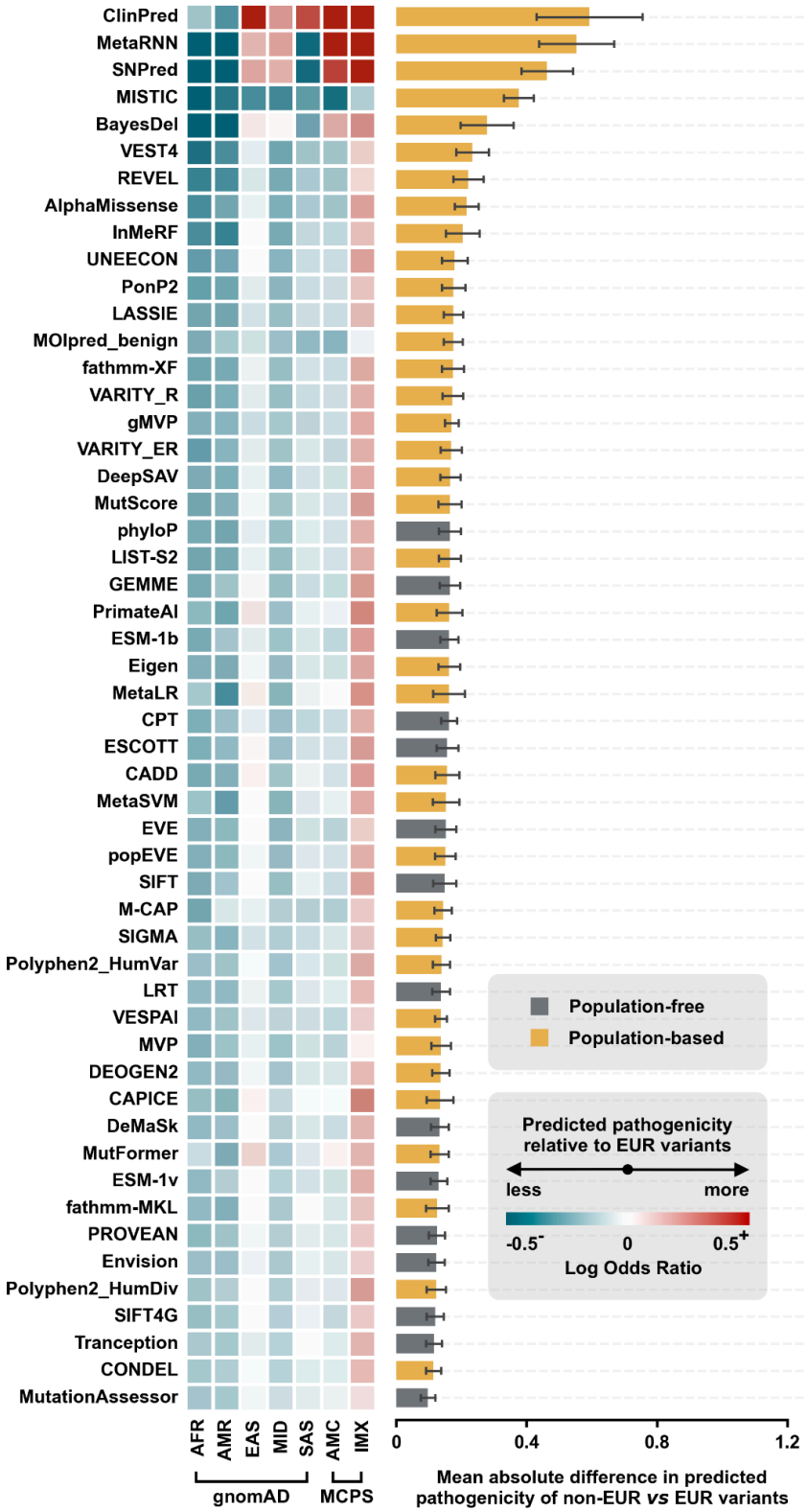

Figure S3. Relative enrichment of predicted pathogenic variants when defining thresholds so that 20% of EUR variants are predicted pathogenic. A custom pathogenicity threshold is defined for each VEP, specifically adjusted to classify 20% of variants in EUR as pathogenic. This controls for certain population-based VEPs which tend to predict far fewer variants to be pathogenic. Following this adjustment, the remaining analysis follows the same methodology as outlined for Figure 2.