nature biotechnology

Brief Communication

https://doi.org/10.1038/s41587-024-02511-w

A DNA language model based on multispecies alignment predicts the effects of genome-wide variants

Received: 31 October 2023

Accepted: 20 November 2024

Published online: 02 January 2025

Check for updates

Gonzalo Benegas[1,2], Carlos Albors[2,5], Alan J. Aw[3,5], Chengzhong Ye[3,5] & Yun S. Song[2,3,4] ✉

Protein language models have demonstrated remarkable performance in predicting the effects of missense variants but DNA language models have not yet shown a competitive edge for complex genomes such as that of humans. This limitation is particularly evident when dealing with the vast complexity of noncoding regions that comprise approximately 98% of the human genome. To tackle this challenge, we introduce GPN-MSA (genomic pretrained network with multiple-sequence alignment), a framework that leverages whole-genome alignments across multiple species while taking only a few hours to train. Across several benchmarks on clinical databases (ClinVar, COSMIC and OMIM), experimental functional assays (deep mutational scanning and DepMap) and population genomic data (gnomAD), our model for the human genome achieves outstanding performance on deleteriousness prediction for both coding and noncoding variants. We provide precomputed scores for all ~9 billion possible single-nucleotide variants in the human genome. We anticipate that our advances in genome-wide variant effect prediction will enable more accurate rare disease diagnosis and improve rare variant burden testing.

With the rising trend of whole-genome sequencing, there is a pressing need to understand the effects of genome-wide variants, which would lay the foundation for precision medicine[1]. In particular, predicting variant deleteriousness is key to rare disease diagnosis[2] and rare variant burden tests[3]. Indeed, a recent review highlighted the analysis of functional rare variants as the biggest contribution of human genetics to drug discovery[4].

Language models are gaining traction as predictors of deleteriousness, with their ability to learn from massive sequence databases and score variants in an unsupervised manner. Given the success of accurately scoring missense variants with protein language models[5–7], it is natural to consider scoring genome-wide variants with DNA language models. For this task, we recently developed the genomic pretrained network (GPN), a model based on a convolutional neural network trained on unaligned genomes, and showed that it achieves excellent variant effect prediction (VEP) results in the compact genome of *Arabidopsis thaliana*[8]. However, the human genome, which harbors a similar number of genes but interspersed over nearly 23-fold larger regions and containing many more repetitive elements, most of which may not be functional, is substantially harder to model. In fact, previous attempts at unsupervised VEP with human DNA language models (for example, Nucleotide Transformer[9]) showed inferior performance compared to simpler conservation scores. Increasing the scale of the model, data and computer improves performance but it can still be poor, even for a model trained for 28 days using 128 top-of-the-line graphics processing units (GPUs)[9].

To address the above challenge, we here introduce GPN-MSA (GPN with multiple-sequence alignment), a DNA language model that

[1]Graduate Group in Computational Biology, University of California, Berkeley, CA, US. [2]Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA, US. [3]Department of Statistics, University of California, Berkeley, CA, US. [4]Center for Computational Biology, University of California, Berkeley, CA, US. [5]These authors contributed equally: Carlos Albors, Alan J. Aw, Chengzhong Ye. ✉e-mail: yss@berkeley.edu

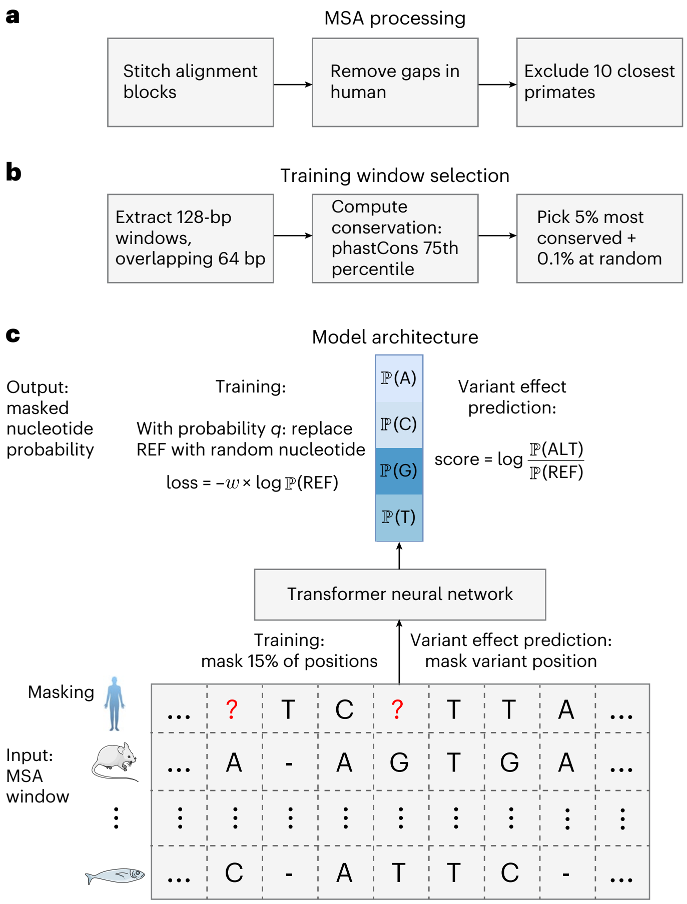

a
MSA processing
Stitch alignment blocks
Remove gaps in human
Exclude 10 closest primates
b
Training window selection
Extract 128-bp windows, overlapping 64 bp
Compute conservation: phastCons 75th percentile
Pick 5% most conserved + 0.1% at random
c
Model architecture
Output: masked nucleotide probability
Training:
With probability q: replace REF with random nucleotide
loss = −w × log ℙ(REF)
ℙ(A)
ℙ(C)
ℙ(G)
ℙ(T)
Variant effect prediction:
score = log ℙ(ALT)/ℙ(REF)
Transformer neural network
Training: mask 15% of positions
Variant effect prediction: mask variant position
Masking
Input: MSA window
… ? T C ? T T A …
… A - A G T G A …
⋮ ⋮ ⋮ ⋮ ⋮ ⋮ ⋮ ⋮ ⋮
… C - A T T C - …

Fig. 1 | Overview of GPN-MSA. a, MSA processing. Starting with a multiple-alignment format file, alignment blocks are stitched together following the order in the human reference. Columns with gaps in the human reference are discarded, followed by the removal of the ten primate species closest to human (chimp to squirrel monkey). **b**, Training window selection. For each 128-bp window along the genome, conservation is computed as the 75th percentile of phastCons. The top 5% conserved windows are chosen alongside a random 0.1% from the remaining windows. **c**, Model architecture. The input is a 128-bp MSA window where certain positions in the human reference are masked and the goal is to predict the nucleotides at the masked positions given the context across both columns (positions) and rows (species) of the MSA. During training, 15% of the positions are masked. During VEP, only the variant position is masked. The sequence of MSA columns is processed through a Transformer neural network, resulting in a high-dimensional contextual embedding of each position. Then, a final layer outputs four nucleotide probabilities at each masked position. The model is trained with a weighted cross-entropy loss, designed to downweight repetitive elements and upweight conserved elements (Methods). As data augmentation in nonconserved regions, before computing the loss, the reference is sometimes replaced by a random nucleotide (Methods). The GPN-MSA VEP score is defined as the log-likelihood ratio between the alternate (ALT) and reference (REF) allele. Mouse and fish icons are from Servier (https://smart.servier.com/).

is designed for genome-wide VEP and is based on the biologically motivated integration of MSA across diverse species using the flexible Transformer architecture[10]. We apply this modeling framework to humans using an MSA of diverse vertebrate genomes[11] and show that it outperforms not only recent DNA language models such as Nucleotide Transformer[9] and HyenaDNA[12] but also current widely used models such as CADD[13], phyloP[14,15], ESM-1b[6,16]), Enformer[17] and SpliceAI[18]. Our model took only 3.5 h to train on four NVIDIA A100 GPUs, which is a considerable reduction in the required computing resources compared to the aforementioned Nucleotide Transformer[9]. We anticipate that this massive reduction in computational footprint will enable the efficient exploration of new ideas to train improved DNA language models for genome-wide VEP.

GPN-MSA was trained on a whole-genome MSA of 100 vertebrate species (Supplementary Fig. 1), after processing (Fig. 1a) and filtering (Fig. 1b). It is an extension of GPN[8] to learn nucleotide probability distributions conditioned not only on surrounding sequence contexts but also on aligned sequences from related species that provide important information about evolutionary constraints and adaptation (Fig. 1c and Methods). It draws inspiration from the MSA Transformer[19], a protein language model trained on MSAs of diverse protein families; it was originally designed for structure prediction but was later shown to achieve excellent missense VEP performance[5]. In addition to the fact that our model operates on whole-genome DNA alignments, which comprise small, fragmented synteny blocks with highly variable levels of conservation and, hence, are considerably more complex than protein alignments, there are essential differences in the architecture and training procedure of GPN-MSA from the MSA Transformer (Methods).

By using the MSA as auxiliary information, GPN-MSA can accurately predict nucleotides from their context, especially in functional regions (Supplementary Table 1). At sites where the reference allele differs from the inferred ancestral allele[20], GPN-MSA usually favors the ancestral allele (Supplementary Table 2). However, predicting the human reference is just a pretext task. What we really care about is the likelihood assigned to human genetic variants that have not been seen during training. Conservation statistics computed on an MSA column, from simple frequencies to more complex phylogeny-aware *P* values[14], are intuitive and powerful measures of deleteriousness. GPN-MSA is designed to process conservation information across multiple MSA columns, as has been exploited by earlier models such as phastCons[21] based on a hidden Markov model. To illustrate GPN-MSA's power beyond single-column statistics and its ability to leverage genomic context, we note that, even at perfectly conserved positions, GPN-MSA assigns more deleterious scores to loss-of-function (for example, stop gain or loss and splice donor or acceptor variant) and missense variants compared to synonymous variants (Fig. 2a). Furthermore, variants with extreme GPN-MSA log-likelihood ratios tend to have lower minor allele frequencies (MAFs) than variants with extreme log-likelihood ratios based on MSA column frequencies, suggesting that GPN-MSA is a better estimator of deleteriousness (Fig. 2b).

We demonstrate the capability of GPN-MSA to improve the unsupervised deleteriousness prediction on several human variant datasets (Methods). We emphasize that only the reference genome is used to train GPN-MSA and that no human variant dataset is used in training. Nevertheless, GPN-MSA can still capture several functional attributes of variants, such as epigenetic marks and the impact of natural selection (Supplementary Fig. 2).

For evaluation, we first consider the classification of ClinVar[22] pathogenic versus common missense variants in gnomAD[23]. GPN-MSA substantially outperforms other human DNA language models such as Nucleotide Transformer[9], with the largest number of parameters (2.5 billion), as well as HyenaDNA[12], with the largest context size of 1 Mb (Fig. 2c and Extended Data Fig. 1a). We also find that GPN-MSA achieves improved performance compared to genome-wide predictors CADD[24] and phyloP[14,15], as well as the missense-specific ESM-1b[6,16]. These results are based on using common variants as controls instead of ClinVar benign-labeled variants, as recommended by the developers of CADD to reduce ascertainment bias[13]. When using benign-labeled variants in ClinVar as controls, the area under the receiving operating characteristic curve (AUROC) for every method is reduced and GPN-MSA performs marginally behind CADD and ESM-1b (Fig. 2d); regardless of the control set, the three methods perform very similarly on ClinVar missense variants.

Next, we consider the classification of somatic missense variants frequently observed across cancer tumors (COSMIC[25]) versus

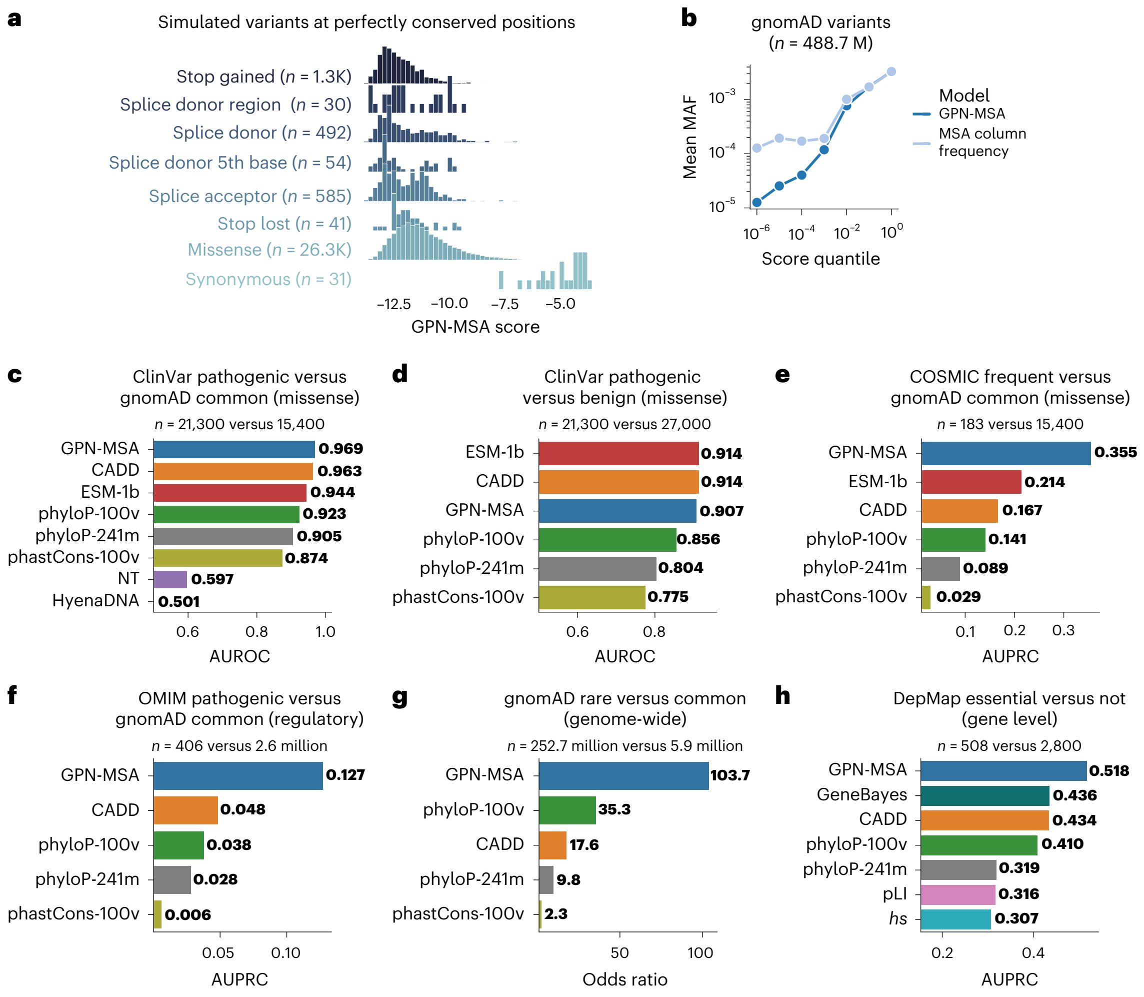


Fig. 2 | VEP results. a, Variant-type-specific distribution of GPN-MSA scores at positions in held-out chromosome 22 where the corresponding MSA columns have perfect conservation (that is, no variation) in the 89 nonhuman species seen by the model. **b**, Mean MAF for different score quantile bins ($[0, 10^{-6})$, $(10^{-6}, 10^{-5}]$, …, $(10^{-1}, 1]$) in the full set of gnomAD biallelic sites. The MSA column frequency score is the log-likelihood ratio based on the empirical column frequencies, with a pseudocount of 1. To break ties, we added a very small random number to each score (the pattern across random seeds was stable). **c**, Classification of ClinVar pathogenic versus gnomAD common missense variants. Exact sample size, $n = 21{,}273$ versus 15,402. NT, Nucleotide Transformer. **d**, Classification of ClinVar pathogenic versus ClinVar benign missense variants. Exact sample size, $n = 21{,}275$ versus 26,993. **e**, Classification of COSMIC frequent (frequency > 0.1%) versus gnomAD common missense variants. Exact sample size, $n = 183$ versus 15,399. **f**, Classification of OMIM pathogenic versus gnomAD common regulatory variants. We matched OMIM promoter variants with gnomAD upstream-of-gene variants, enhancer with intergenic and 'all' with the union of the matches of the specific categories. Exact sample size, $n = 406$ versus 2,573,918. **g**, Enrichment of rare (singletons) versus common (MAF > 5%) gnomAD variants in the tail of deleterious scores (the threshold was chosen such that each score made 30 false discoveries). Odds ratios and P values were computed using a one-sided Fisher's exact test. All shown odds ratios have a P value < 0.05. Exact sample size, $n = 252{,}706{,}195$ versus 5,894,721. **h**, Classification of DepMap essential versus nonessential genes using VEPs and selection constraint metrics (Methods). A gene is defined to be essential if >1,000 cell lines in DepMap assays depend on it, whereas it is defined to be nonessential if no cell line depends on it. Exact sample size, $n = 508$ versus 2,815.

gnomAD common missense variants. Because of the extreme class imbalance in this case, we focus on the precision and recall metrics. GPN-MSA achieves the highest performance, with substantial margins of improvement over other models (Fig. 2e and Extended Data Fig. 1b).

We also evaluate deep mutational scanning (DMS) experimental data[26] for 31 human proteins (Supplementary Table 3). GPN-MSA and CADD perform comparably on classifying variants labeled according to protein-specific binarization (Methods), with the former slightly outperforming the latter on the area under the precision–recall curve (AUPRC) metric (Supplementary Fig. 3). They both compare favorably with phyloP and phastCons. However, the protein language model ESM-1b achieves the best overall performance on this task; it likely benefits from modeling long-range interactions within each protein and training on diverse proteins across a much larger evolutionary timescale. Another challenge for genome-wide variant effect predictors on this task is that they expect additional context, such as introns, which is typically lacking in DMS experimental assays. Nevertheless, GPN-MSA performs better than ESM-1b on some proteins (for example, TAR DNA-binding protein (TADBP), for which the GPN-MSA AUROC is 0.83 while the ESM-1b AUROC is 0.75). An intriguing avenue for future research would be to explore the conditions under which one model outperforms another and integrate the strengths of both DNA and protein language models. As another note of caution, none of the models performs exceedingly well relative to ClinVar results.

Moving on to regulatory variants, we evaluate the classification of a curated set of variants implicated in Mendelian disorders (OMIM[27]) versus gnomAD common variants. We again consider precision and recall because of the extreme class imbalance and find that GPN-MSA achieves the best performance overall, as well as in each variant category (Fig. 2f and Extended Data Figs. 1c and 2). Nucleotide Transformer exhibits poor performance compared to other models (Extended Data Fig. 2). For several variant categories, CADD's precision increases from near zero as recall increases, which indicates that a substantial fraction of its top discoveries are actually false (Extended Data Fig. 1c). One example of a deleterious variant that was assigned an extreme score by GPN-MSA is `rs606231231`, lying in the well-known ZRS enhancer that controls the expression of *SHH* at the long range of 1 Mb (Supplementary Fig. 4). This variant is associated with polydactyly[28] and has been experimentally verified to alter gene expression in mouse limb[29]. Another example is `rs1367115848`, disrupting hepatocyte nuclear factor 4 binding at the *F7* promoter and causing severe factor VII deficiency[30] (Supplementary Fig. 5).

Following this, we further evaluate the enrichment of rare versus common gnomAD variants in the tail of the distribution of deleteriousness scores. Deleterious variants should be under purifying selection and, hence, their frequencies in populations should tend to be lower. Therefore, if a variant effect predictor is accurate, we expect rare variants to be enriched compared to common variants for extreme deleteriousness scores. GPN-MSA achieves the highest enrichment overall (Fig. 2g), as well as within most variant categories, with different margins (Extended Data Fig. 3 and Supplementary Fig. 6). In the case of intronic variants, it also outperforms SpliceAI[18], a state-of-the-art splicing predictor. There is one category where GPN-MSA performs behind CADD: splice-region variants (here, we group variants immediately close to the exon borders, such as splice donors and acceptors). To be clear, GPN-MSA generally assigns extreme scores to these variants (Extended Data Fig. 4a); the challenge is understanding which variants are not deleterious and more likely to be common in the population. CADD features useful for this task could potentially be integrated into GPN-MSA to improve performance. We note that the overall genome-wide performance in Fig. 2g is not merely an averaging of the performances in the different categories; it also involves scoring variants relative to each other across these categories. On a separate enrichment analysis of low-frequency versus common gnomAD variants in nonexonic regions, GPN-MSA achieves a substantially improved performance over Enformer[17] (Extended Data Fig. 5 and Supplementary Fig. 7). Despite its improved ability to distinguish rare or low-frequency versus common variants in gnomAD, how well GPN-MSA would perform in distinguishing deleterious rare variants from benign rare variants genome-wide is an important open question, the answer to which will require more comprehensive labeled data.

GPN-MSA also outperforms other methods when subsetting to putatively conserved, neutral or accelerated positions in the genome (Extended Data Fig. 6). While we observe that SpliceAI and Enformer, which are functional genomics models, perform worse than the simpler phyloP in deleteriousness prediction, we note that this is an application that they were not designed for. It is also worth noting that, although phyloP-241m (fit to the 241-way Zoonomia mammalian alignment) was recently proposed as a deleteriousness predictor[15], the older vertebrate phyloP-100v actually achieves better results in many of our benchmarks. On missense variants, GPN-MSA generally compares favorably with PrimateAI-3D[31] despite the latter being trained with allele frequency information, which gives it an advantage in our benchmarks (Supplementary Fig. 8).

Additionally, we evaluate our model on the classification of essential genes using the DepMap cancer dependency data[32]. This dataset contains gene essentiality measurements based on genome-scale CRISPR knockout screens on over 1,000 cancer cell lines. Essential genes are supposed to be more intolerant to deleterious mutations and, consequently, their variants tend to have overall higher impacts. We summarize VEPs into gene-level essentiality scores (Methods) and benchmarked them against several gene-level selection constraint metrics on how well each method classifies essential versus nonessential genes identified by DepMap assays. GPN-MSA outperforms other genome-wide variant effect predictors, as well as selection constraint metrics (pLI[33], heterozygous selection[34] and GeneBayes[35]) specifically designed for gene essentiality prediction (Fig. 2h).

To understand the importance of different components of our model, we perform an ablation study and assess the impact on VEP performance. Here, we summarize the main takeaways (Extended Data Table 1). Firstly, the inclusion of the MSA is critical for GPN-MSA's high performance. Secondly, the simple MSA column frequency (likewise for phyloP) does a good job in the ClinVar benchmark but this is not the case for the other benchmarks. ClinVar might be a relatively easy task because variants with pathogenic labels often lie at the extreme tail of deleteriousness, supported by multiple lines of evidence, and perhaps also because conservation scores have traditionally been used as part of labeling guidelines[36]. Thirdly, in our case, training on conserved regions is better than the usual approach of training on the whole genome. Lastly, with an MSA of diverse vertebrate genomes, we see diminishing returns of increased window size for deleteriousness prediction. More details are provided in the Methods.

Given the positive evaluation of GPN-MSA, we computed scores for all ~9 billion possible single-nucleotide variants in the human genome, which we make available along with a few recommendations. Examining the complete gnomAD variant set, there seems to be a near-linear relationship between the GPN-MSA score bin and the logarithm of the average minor allele frequency within that specific bin (Extended Data Fig. 7). We believe that the deleteriousness of GPN-MSA scores should be interpreted as a continuum; if a hard threshold is helpful, we recommend a cutoff around −7 on the basis of the distribution of scores in different datasets (Extended Data Fig. 8). Incidentally, the bimodality of score distribution for frequent variants in COSMIC suggests that many of them could be passenger mutations (Extended Data Fig. 8b). Additionally, we recommend using scores only to compare variants relative to each other and warn against interpreting them as calibrated fitness estimates (Methods). Predictions can be visualized as sequence logos[37] in the UCSC Genome Browser[38,39] (example in Extended Data Fig. 9).

To recapitulate, our main contributions are threefold. First, we propose the first DNA language model operating directly on a whole-genome alignment. Second, we demonstrate outstanding performance in humans on a number of clinically relevant VEP datasets. While there already exist many effective missense variant effect predictors, we anticipate that our ideas and readily available GPN-MSA scores will be particularly helpful to interpret variation in noncoding regions. Lastly, the general approach we developed for humans is computationally efficient, which would enable future research in the field.

In the rapidly advancing landscape of DNA language modeling, scaling up model and context sizes has been the primary avenue of exploration[9,12,40]. In contrast, our present work focuses on the explicit modeling of related sequences (known as retrieval augmentation in natural language processing[41]). This has led to a highly computationally efficient model and state-of-the-art VEP performance for genome-wide variants. It remains to be explored how useful GPN-MSA's learned representations would be for downstream applications (for example, for genome annotation or gene expression prediction). Expanding the context length, possibly through leveraging recent technical developments[12], might be beneficial for such tasks.

Our modeling approach also differs from earlier works, in that we train mostly on conserved regions of the genome to increase the proportion of functional as opposed to neutral sites. This strategy may potentially miss some functional sites, including those in primate or human-specific regions, fast-evolving regions or in hard-to-align regions. We also recognize the challenge in balancing data quality (that is, sequence information content) with data quantity during training, given that larger models have a higher risk of overfitting when trained on smaller amounts of data. We believe that these are critical considerations for researchers seeking to adopt our present strategy.

The masked language modeling objective can be too easy if sequences very similar to the human genome are included in the MSA, resulting in the learned probability distribution being not very useful for VEP. This observation led us to exclude most primate genomes during training. To tackle this limitation, we are actively exploring alternative training objectives that are aware of phylogenetic relationships. We are also exploring how best to integrate population genetic variation information instead of relying on a single reference genome.

In our view, one of the most promising applications of GPN-MSA is effective genome-wide rare variant burden testing, which has been mostly restricted to coding regions[42]. We envision that several other statistical genetics tasks can be empowered by GPN-MSA, such as functionally informed fine-mapping[43], polygenic risk scores[44] and related variant randomness tests[45] and variant prioritization in integrative analyses.

Sequence models (such as phyloP and GPN-MSA) might achieve better deleteriousness prediction results but are still less interpretable than functional genomics models such as SpliceAI and Enformer. While both functional genomics models and DNA language models have much room for independent improvement, it is likely that jointly modeling DNA sequence and functional genomics may have the biggest impact.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-024-02511-w.

References

1. Goldfeder, R. L., Wall, D. P., Khoury, M. J., Ioannidis, J. P. & Ashley, E. A. Human genome sequencing at the population scale: a primer on high-throughput DNA sequencing and analysis. *Am. J. Epidemiol.* **186**, 1000–1009 (2017).
2. Marwaha, S., Knowles, J. W. & Ashley, E. A. A guide for the diagnosis of rare and undiagnosed disease: beyond the exome. *Genome Med.* **14**, 23 (2022).
3. Lee, S., Abecasis, G. R., Boehnke, M. & Lin, X. Rare-variant association analysis: study designs and statistical tests. *Am. J. Hum. Genet.* **95**, 5–23 (2014).
4. Trajanoska, K. et al. From target discovery to clinical drug development with human genetics. *Nature* **620**, 737–745 (2023).
5. Meier, J. et al. Language models enable zero-shot prediction of the effects of mutations on protein function. In *Proc. Advances in Neural Information Processing Systems 34* (eds Ranzato, M. et al.) 29287–29303 (Curran Associates, Inc., 2021).
6. Brandes, N., Goldman, G., Wang, C. H., Ye, C. J. & Ntranos, V. Genome-wide prediction of disease variant effects with a deep protein language model. *Nat. Genet.* **55**, 1512–1522 (2023).
7. Jagota, M. et al. Cross-protein transfer learning substantially improves disease variant prediction. *Genome Biol.* **24**, 182 (2023).
8. Benegas, G., Batra, S. S. & Song, Y. S. DNA language models are powerful predictors of genome-wide variant effects. *Proc. Natl Acad. Sci. USA* **120**, e2311219120 (2023).
9. Dalla-Torre, H. et al. Nucleotide Transformer: building and evaluating robust foundation models for human genomics. *Nat. Methods* https://doi.org/10.1038/s41592-024-02523-z (2024).
10. Vaswani, A. et al. Attention is all you need. In *Proc. Advances in Neural Information Processing Systems 30* (eds Guyon, S. et al.) 6000–6010 (Curran Associates, Inc., 2017).
11. Armstrong, J., Fiddes, I. T., Diekhans, M. & Paten, B. Whole-genome alignment and comparative annotation. *Annu. Rev. Anim. Biosci.* **7**, 41–64 (2019).
12. Nguyen, E. et al. HyenaDNA: long-range genomic sequence modeling at single nucleotide resolution. In *Proc. 37th International Conference on Neural Information Processing Systems* (eds Oh, A. et al.) 43177–43201 (Curran Associates, Inc., 2023).
13. Rentzsch, P., Schubach, M., Shendure, J. & Kircher, M. CADD-Splice-improving genome-wide variant effect prediction using deep learning-derived splice scores. *Genome Med.* **13**, 31 (2021).
14. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010).
15. Sullivan, P. F. et al. Leveraging base-pair mammalian constraint to understand genetic variation and human disease. *Science* **380**, eabn2937 (2023).
16. Rives, A. et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl Acad. Sci. USA* **118**, e2016239118 (2021).
17. Avsec, Ž. et al. Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* **18**, 1196–1203 (2021).
18. Jaganathan, K. et al. Predicting splicing from primary sequence with deep learning. *Cell* **176**, 535–548.e24 (2019).
19. Rao, R. M. et al. MSA Transformer. In *Proceedings of the 38th International Conference on Machine Learning* (eds Meila, M. & Zhang, T.) (PMLR, 2021).
20. Paten, B. et al. Genome-wide nucleotide-level mammalian ancestor reconstruction. *Genome Res.* **18**, 1829–1843 (2008).
21. Siepel, A. et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
22. Landrum, M. J. et al. ClinVar: improvements to accessing data. *Nucleic Acids Res.* **48**, D835–D844 (2020).
23. Chen, S. et al. A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* **625**, 92–100 (2024).
24. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–D894 (2019).
25. Tate, J. G. et al. COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Res.* **47**, D941–D947 (2019).
26. Notin, P. et al. ProteinGym: large-scale benchmarks for protein fitness prediction and design. In *Proceedings of the Advances in Neural Information Processing Systems 37* (eds Oh, A. et al.) (NeurIPS, 2023).
27. Smedley, D. et al. A whole-genome analysis framework for effective identification of pathogenic regulatory variants in mendelian disease. *Am. J. Hum. Genet.* **99**, 595–606 (2016).
28. Albuisson, J. et al. Identification of two novel mutations in Shh long-range regulator associated with familial pre-axial polydactyly. *Clin. Genet.* **79**, 371–377 (2011).
29. Kvon, E. Z. et al. Comprehensive in vivo interrogation reveals phenotypic impact of human enhancer variants. *Cell* **180**, 1262–1271.e15 (2020).
30. Arbini, A. A., Pollak, E. S., Bayleran, J. K., High, K. A. & Bauer, K. A. Severe factor VII deficiency due to a mutation disrupting a hepatocyte nuclear factor 4 binding site in the factor VII promoter. *Blood* **89**, 176–182 (1997).
31. Gao, H. et al. The landscape of tolerated genetic variation in humans and primates. *Science* **380**, eabn8153 (2023).
32. The Dependency Map Consortium. DepMap 23Q4 public. *figshare* https://doi.org/10.25452/figshare.plus.24667905.v2 (2023).
33. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
34. Agarwal, I., Fuller, Z. L., Myers, S. R. & Przeworski, M. Relating pathogenic loss-of-function mutations in humans to their evolutionary fitness costs. *eLife* **12**, e83172 (2023).
35. Zeng, T., Spence, J. P., Mostafavi, H. & Pritchard, J. K.Bayesian estimation of gene constraint from an evolutionary model with gene features. *Nat. Genet.* **56**, 1632–1643 (2024).

36. 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).
37. Schneider, T. D. & Stephens, R. M. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* **18**, 6097–6100 (1990).
38. Kent, W. J. et al. The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
39. Nair, S. et al. The dynseq browser track shows context-specific features at nucleotide resolution. *Nat. Genet.* **54**, 1581–1583 (2022).
40. Fishman, V. et al. GENA-LM: a family of open-source foundational models for long DNA sequences. Preprint at *bioRxiv* https://doi.org/10.1101/2023.06.12.544594 (2023).
41. Borgeaud, S. et al. Improving language models by retrieving from trillions of tokens. In *Proc. 39th International Conference on Machine Learning* (eds Chaudhuri, K. et al.) 2206–2240 (PMLR, 2022).
42. Weiner, D. J. et al. Polygenic architecture of rare coding variation across 394,783 exomes. *Nature* **614**, 492–499 (2023).
43. Weissbrod, O. et al. Functionally informed fine-mapping and polygenic localization of complex trait heritability. *Nat. Genet.* **52**, 1355–1363 (2020).
44. Márquez-Luna, C. et al. Incorporating functional priors improves polygenic prediction accuracy in UK Biobank and 23andMe data sets. *Nat. Commun.* **12**, 6052 (2021).
45. Aw, A. J., McRae, J., Rahmani, E. & Song, Y. S. Highly parameterized polygenic scores tend to overfit to population stratification via random effects. Preprint at *bioRxiv* https://doi.org/10.1101/2024.01.27.577589 (2024).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

MSA processing

The `multiz`[46] whole-genome alignment of 100 vertebrates was downloaded from https://hgdownload.soe.ucsc.edu/goldenPath/hg38/multiz100way/maf/. Contiguous alignment blocks were stitched together using the `multiz` utility `maf2fasta` and any columns with gaps in human were removed. The ten primate species closest to human were removed. We also downloaded associated conservation scores phastCons[21] (https://hgdownload.soe.ucsc.edu/goldenPath/hg38/phastCons100way) and phyloP[14] (https://hgdownload.soe.ucsc.edu/goldenPath/hg38/phyloP100way).

Training region selection

Instead of training on the whole genome, we focused on the most conserved genomic windows, aiming to emphasize functionally important regions such as exons, promoters and enhancers. The conservation of a genomic window was defined as the 75th percentile of phastCons scores in the window. We then chose a cutoff; in our current experiments, we included the top 5% most conserved windows. We also included 0.1% of the remaining windows of the genome to ensure that there was no extreme distribution shift when performing VEP in nonconserved regions. The reverse complement of each selected window was added as data augmentation. Chromosome 21 was held out for validation (early stopping) and chromosome 22 was held out for evaluating language modeling performance metrics such as perplexity. We used all chromosomes for evaluating the separate task of VEP.

Model architecture

We adopted the general approach of masked language modeling[47]. As a general caveat, in this work, we did not systematically tune hyperparameters; thus, they are likely far from optimal. The input was a 128-bp MSA window where certain positions in the human reference were masked and the goal was to predict the nucleotides at the masked positions, given its context across both columns (positions) and rows (species) of the MSA. During training, 15% of the positions were masked. During VEP, only the variant position was masked. The one-hot encodings of nucleotides from different species at each position were first concatenated. Then, the sequence of MSA columns was processed through a Transformer neural network (RoFormer[48]) resulting in a high-dimensional contextual embedding of each position. Then, the final layer output four nucleotide probabilities at each masked position. The model was trained on the reference sequence with a weighted cross-entropy loss.

We downweighted repeats and upweighted conserved elements such that incorrect predictions in neutral regions were penalized less. We introduced a smoothed version of phastCons, phastCons_M, as the max of phastCons over a window of 7 nt. The goal was to give importance not only to conserved regions but also to regions immediately next to them. The loss weight w is defined as follows:

$$w \propto (0.1 \times \mathbb{1}\{\text{repeat}\} + \mathbb{1}\{\neg\text{repeat}\}) \times \max(\text{phyloP}, 1) \times (\text{phastCons}_M + 0.1),$$

which includes tenfold downweighting on repetitive elements[8] plus upweighting based on both phyloP and phastCons_M.

As data augmentation in nonconserved regions, before computing the loss, the reference was replaced by a random nucleotide with a certain probability q:

$$q = 0.5 \times \mathbb{1}\{\text{phastCons}_M < 0.1\}$$

As a result, the probability of keeping the reference nucleotide is $0.5 + \frac{0.5}{4} = 0.625$ and the probability of replacing it with each of the alternative nucleotides is $\frac{0.5}{4} = 0.125$. The intention is to guide the model to assign more neutral scores in nonconserved regions.

Our code is based on the Hugging Face Transformers library[49]. All models were trained with default hyperparameters (https://huggingface.co/docs/transformers/model_doc/roformer#transformers.RoFormerConfig) (for example, 12 layers with 12 attention heads each) except for those listed in Supplementary Table 4. The total number of parameters was approximately 86 million. We performed early stopping on the basis of validation loss. We trained the model for a max of 30,000 steps (~14 epochs) in approximately 3.5 h using four NVIDIA A100 GPUs.

The GPN-MSA VEP score is defined as the log-likelihood ratio between the alternate and reference allele. In our experiments, we averaged the predictions from the positive and negative strands. With our four NVIDIA A100 GPUs, we managed to score approximately 25 million variants per hour.

Differences between GPN-MSA and MSA Transformer

While MSA Transformer takes as input an arbitrary set of aligned sequences, GPN-MSA was trained on sequences from a fixed set of species. This allows simpler modeling of the MSA as a sequence of fixed-size alignment columns, reducing computation and memory requirements. VEP masking only the target sequence (in our case, human) is identical[5]. Because VEP was our main goal, during training, we also only masked positions from the target sequence. MSA Transformer, however, proposes masking MSA entries at random during training on the basis of results from structure prediction, their intended application.

Language modeling metrics

The median perplexity and accuracy were computed for the test chromosome (chr 22). Perplexity is defined as the exponentiation of the cross-entropy loss, which is equivalent to 1 over the probability given to the correct nucleotide. The ancestral alleles[20] were downloaded from https://ftp.ensembl.org/pub/release-109/fasta/ancestral_alleles/. The accuracy was computed for test chromosome positions where the ancestral and reference alleles differ.

Ablation study

We performed an ablation study to understand the impact of each of our design choices on VEP when modified independently (Extended Data Table 1). For each setting, three replicate models with different seeds were trained, where applicable. Because we fixed the remaining hyperparameters, results should be interpreted as differences given a similar training procedure and compute budget. There are many crucial hyperparameters worth investigating further. These include the size of the model, the number of iterations and the learning rate schedule, all of which would most certainly affect the performance of each ablated model.

While varying window size has been mainly motivated by the question of how much context can be leveraged by the model to make a prediction, it also affects two other components of the training workflow. Firstly, it affects which positions of the genome are used for training, given that we first split the genome into windows of a given size and filter them according to the 75th phastCons percentile; this statistic would be biased toward certain regions over others according to the window size. Secondly, it influences the number of data points (windows) or the total number of tokens (base pairs) used for training, which can affect optimization and generalization. With these considerations in mind, to study how much the model uses context, we decided to reduce the window size at VEP time rather than retrain from scratch. For the increased window size, we did retrain the model from scratch; however, for better performance, we encourage future studies to also vary other hyperparameters, particularly increasing the percentage of the genome used for training, to reduce the chance of overfitting.

Details on the ablations are as follows:

- w/o MSA: the model is only trained on the human sequence, without access to other species.

- MSA frequency: variants are scored using the log-likelihood ratio of observed frequencies in the MSA column, with a pseudocount of 1.
- Combined phyloP and phastCons: the same combination used to upweight the loss.
- Train on 50–100% most conserved: expand the training region from the smaller 5% most conserved to a larger set with less overall conservation.
- Include closest primates: do not filter out from the MSA the ten primates closest to human.
- 51 mammals: subset the vertebrate MSA to only mammals (51, besides human).
- 51 vertebrates: subset the vertebrate MSA from 89 to 51 vertebrates (besides human). The subset is made at random except for the closest species (bushbaby), which is deterministically chosen.
- Do not upweight conserved: do not upweight the loss function on conserved elements.
- Do not replace nonconserved: do not replace the reference in nonconserved positions with random nucleotides when computing the loss function.
- Increased window size (256): the whole training procedure (including window selection) is repeated with window size 256.
- Reduced window size (4–64): the default model is shown a smaller window at VEP time.

We now describe the results. While all metrics can distinguish large differences in performance, the ClinVar and gnomAD metrics are based on larger sample sizes and, therefore, should be more adequate for investigating more subtle differences. Modeling the single human sequence instead of the MSA had by far the biggest impact. We note that alignment-based methods, including simple ones such as MSA frequency, usually achieved a high performance in metrics for missense variants (ClinVar and COSMIC) but not for genome-wide variants (OMIM and gnomAD). Including primate species close to human or training on less conserved regions also had a large impact on performance. Of relatively minor impact were subsetting to 51 mammals or 51 vertebrates, removing the upweighting of conserved elements or removing the data augmentation procedure of replacing nucleotides in nonconserved positions. Increasing the window size provided at VEP from 4 to 128 showed diminishing returns. A model trained with the doubled window size of 256 actually showed worse gnomAD mean performance across random seeds but a comparable max performance. This instability across runs could potentially be reduced by increasing the percentage of the genome used for training.

VEP glossary

Variant consequences were obtained by running Ensemble VEP Release 109 (ref. 50) with arguments `–most_severe –distance 1000`.

For in silico mutagenesis analysis, we analyzed the scores for a random subset of 10 million simulated variants in test chromosome 22. Our data augmentation procedure of replacing the reference with another nucleotide in nonconserved regions caused an artificial shift in the distribution of scores (Extended Data Fig. 4). The peak observed in the distribution of scores (Extended Data Fig. 4a) roughly coincided with $\log \frac{0.125}{0.625} \approx -1.6$, related to the probability that the alternate and reference alleles were exchanged as data augmentation in nonconserved regions. More importantly, regardless of the previous point, we considered likelihood ratios to be systematically lower than the ratios expected based on differences in fitness. To illustrate this point, we note that the distribution of GPN-MSA scores for synonymous variants was centered around −3, meaning that the alternate allele was usually 20 times less likely than the reference allele, while we would expect them to be equally likely (corresponding to a score of 0, as observed in GPN[8]). The current training procedure still favors assigning a higher probability to the most frequent allele in the MSA even when the fitness is uniform, an issue that may be mitigated by explicitly modeling phylogenetic relationships. Nevertheless, it is clear from the benchmark results that the scores, used in a relative fashion, are very powerful at discriminating deleterious from neutral variants.

We also analyzed the scores at all positions in chromosome 22 where the aligned species (regardless of human, which is masked) had the same nucleotide.

We summarize datasets and their provenance, metrics used to evaluate each dataset and technical details in constructing VEP scores below.

VEP data sources:

- ClinVar[22]: downloaded release `20230730`.
- COSMIC[25]: downloaded `Cosmic_MutantCensus_v98_GRCh38.tsv.gz` and computed frequency as the proportion of samples containing the mutation, restricting to whole-genome or whole-exome samples.
- OMIM[27]: downloaded a set of curated pathogenic regulatory variants.
- gnomAD[23]: downloaded version `3.1.2` and filtered to variants with allele number of at least 25,000, besides the official quality-control flags. We defined common variants as those with MAF > 5% and rare variants as singletons.
- DMS[26]: downloaded the human protein data in ProteinGym version 0.1 (Supplementary Table 3). For genomic variant effect predictors, we took the most deleterious score among all the single-nucleotide variations causing the amino acid substitution reported in ProteinGym.
- DepMap[32]: downloaded the gene dependency data `CRISPRGeneDependency.csv` in the Public 23Q4 release. The data were further summarized into per-gene essentiality by counting the number of dependent cell lines. A cell line is defined as dependent on the gene if the probability of dependency is greater than 0.5.

Main VEP metrics:

- ClinVar: AUROC for classification of ClinVar 'pathogenic' versus gnomAD common missense variants.
- COSMIC: AUPRC for classifying COSMIC frequent (frequency > 0.1%) versus gnomAD common missense variants.
- OMIM: AUPRC for classification of OMIM pathogenic versus gnomAD common regulatory variants. We matched OMIM promoter variants with gnomAD upstream-of-gene variants, enhancer with intergenic and 'all' with the union of the matches of the specific categories.
- gnomAD: enrichment of rare versus common gnomAD variants in the tail of deleterious scores (the threshold was defined by allowing a small number of false discoveries, such as 30). We grouped certain variant consequences with small samples sizes as follows: 'splice-region' groups, Ensembl categories `splice_donor`, `splice_acceptor`, `splice_donor_5th_base`, `splice_donor_region`, `splice_region` and `splice_polypyrimidine_tract`; 'start-or-stop' groups, Ensembl categories `start_lost` and `stop_gained` or `stop_lost`.
- DepMap: AUPRC for classifying essential genes (number of dependent cell lines > 1,000) versus nonessential genes (number of dependent cell lines = 0). To summarize VEP scores into gene-level essentiality scores, we used the following procedure. For each gene, we considered annotations from the canonical Ensembl transcript in the gnomAD version 2 constraint table[33] and divided all variants into two groups: (1) exonic and splice donor or acceptor variants and (2) remaining variants. VEP scores were generated for all variants in each group. We then summarized the variant-level scores in

each group into a gene-level score by calculating the most deleterious kth quantile. We considered a range of k from 0.5 to 10^{-5}. We then selected the best-performing quantile in each group on the basis of essentiality prediction AUPRC for genes on chromosomes 17–22 (Supplementary Table 5). The final gene-level essentiality score was obtained by summing the selected best quantiles of the two groups after rank normalization. The presented benchmark results in Fig. 2h are based on genes on held-out chromosomes 1–16. This analysis included genes from the gnomAD version 2 constraint table where all benchmarked methods had available predictions at all positions. We further excluded genes whose genomic regions overlapped with other genes.

Additional VEP metrics:

- gnomAD (Enformer set): enrichment of low-frequency (0.5% < AF < 5%) versus common (MAF > 5%) gnomAD nonexonic variants. We used low-frequency instead of rare because of the lack of Enformer precomputed scores.
- ClinVar pathogenic versus benign: AUROC for classification of ClinVar ‘pathogenic’ versus ClinVar ‘benign’ missense variants.
- DMS: AUROC and AUPRC for binary classification where labels were defined by the protein-specific binarization cutoff provided in ProteinGym, except for TADBP. No models performed well on predicting the relative toxicity measurement for TADBP variants[51] but they worked much better on predicting the absolute value of relative toxicity. We used a cutoff of 0.125 on the absolute value of relative toxicity to binarize the DMS data for TADBP.
- gnomAD (ablation set): enrichment of rare versus common gnomAD variants in the tail of deleterious scores. Because of the large size of the full data, we subsetted the rare variants to match the number of common variants per consequence.

VEP scores:

- GPN-MSA: log-likelihood ratio between alternate and reference allele. Predictions from both strands were averaged.
- CADD: raw scores (version 1.6), negated (lower means more deleterious).
- phyloP-100v: computed on 100-vertebrate alignment, negated (lower means more deleterious).
- phastCons-100v: computed on 100-vertebrate alignment, negated (lower means more deleterious).
- phyloP-241m: computed on 241-mammal alignment, negated (lower means more deleterious).
- Nucleotide Transformer: the most powerful version (`2.5b-multi-species`), with 2.5 billion parameters and trained on 850 species. The central 6-mer was masked and the score was computed as the log-likelihood ratio between the alternate and reference 6-mer. Predictions from both strands were averaged. Given the high computational requirements, we only scored variants for ClinVar and a subset of OMIM variants.
- HyenaDNA: the most powerful version (`large-1m-seqlen-hf`), with 54.6 million parameters and a context length of 1 Mb. The log-likelihood ratio between the alternate and reference sequence was computed. Predictions from both strands were averaged. Given the very high computational requirements, we only scored variants for ClinVar.
- ESM-1b: precomputed log-likelihood ratios between alternate and reference alleles were obtained in protein coordinates[6]. For variants affecting multiple isoforms, the minimum (most deleterious) score was considered.
- SpliceAI: precomputed scores recommended for variant effect prediction (`spliceai_scores.masked.snv.hg38.vcf.gz`) were downloaded from https://basespace.illumina.com/s/otSPW8hnhaZR. The authors do not recommend any specific way of computing a single deleteriousness score. We scored variants using minus the maximum absolute delta in splice acceptor or donor probability in any gene.
- Enformer: precomputed scores for variants with MAF > 0.5% in any 1000 Genomes population[52] were downloaded from https://console.cloud.google.com/storage/browser/dm-enformer/variant-scores. The authors do not recommend any specific way of computing a single deleteriousness score. We scored variants using minus the norm of the 5,313 delta features (single-nucleotide polymorphism activity difference). We found that the L^1 norm works better than the L^2 or L^∞ norm.
- pLI: precomputed scores for the genes were downloaded from gnomAD version 2 and version 4 releases[23,33] under the `Constraint` sections.
- Heterozygous selection coefficients[34]: precomputed $\log_{10}$ maximum a posteriori estimates for the genes were obtained from the supplementary data in the original publication.
- GeneBayes[35]: posterior mean estimates of S_{het} for the genes were downloaded from https://github.com/tkzeng/GeneBayes.
- PrimateAI-3D: scores were obtained through email upon signing an academic license agreement with Illumina.

GPN-MSA captures variant functional impact

A variant’s impact on loss of fitness is mediated by genetic and functional pathways. To investigate whether GPN-MSA captures any functional impact of a variant, we performed functional enrichment analysis on the same random subset of 10 million simulated variants in chromosome 22 (see Methods, ‘VEP glossary’). We used 18 functional annotations obtained from the FAVOR database[53] (accessed through Harvard Dataverse on April 10, 2023), which measure both the impact of natural selection and gene-regulatory activity of a variant (Supplementary Table 6). For clarity, we collected the computational details of the functional annotations, as summarized below.

- B statistic[54], nucleotide diversity[55] and recombination rate[55] are mathematical quantities derived from evolutionary models and are computed directly on the genomic position of the variant. They provide population and genetic interpretation of the impact of natural selection on the variant.
- Epigenetic tracks, RNA sequencing (RNA-seq), DNAse sequencing (DNase-seq), percentage G+C and percentage CpG were all computed on genomic positions to be included as training features in CADD[24]. Specifically, ENCODE track features are not gene specific but are distributed as ‘bigWig’ value tracks along genomic coordinates. Values for each cell type for which a track is available are summarized to create a new genome coordinate-based track, which is subsequently assigned to the variant on the basis of its genomic position. Whenever a variant is not annotatable for a track (for example, RNA-seq level for a nonexonic variant), an `NA` value is assigned.

We note that, across all functional annotations, at least 69% of all variants (n = 6,854,981) were annotated. The average completeness rate across all functional annotations was 85%. This ensured that sample sizes were sufficiently large for statistical analyses to be well powered.

To investigate whether extreme values of GPN-MSA were associated with functional impact, we ran Mann–Whitney tests between the lowest (most deleterious) 1% GPN-MSA scoring (‘target’) variants and the remaining (‘background’) variants, across all 18 annotations. We found significant enrichment (P < 0.05 after controlling for family-wise error rate) of nine histone mark levels, as well as RNA-seq and DNase-seq levels in each dataset, and significant depletion of nucleotide diversity, recombination rate and B statistic (Supplementary Fig. 2). Significant depletions were observed of H3K9me3 and H3K27me3, two recognized gene repressors. These results suggest that extremely negative GPN-MSA scores could potentially prioritize variants with an impact on gene expression and regulation.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The whole-genome alignment was downloaded from https://hgdownload.soe.ucsc.edu/goldenPath/hg38/multiz100way/maf. ClinVar variants (release 20230730) were downloaded from https://ftp.ncbi.nlm.nih.gov/pub/clinvar/. COSMIC variants (v98) were downloaded from https://cancer.sanger.ac.uk/cosmic/download/cosmic. OMIM variants were obtained from the supplementary material in Smedley et al.[27]. gnomAD variants (version 3.1.2) were downloaded from https://gnomad.broadinstitute.org/data. ProteinGym variants (v0.1) were downloaded from https://proteingym.org/download. DepMap gene dependency data (Public 23Q4 release) were downloaded from https://depmap.org/portal/data_page. FAVOR annotations were downloaded from https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/1VGTJI.

Code availability

Code to reproduce all results is available from GitHub (https://github.com/songlab-cal/gpn)[56]. The pretrained model, training dataset, benchmark datasets, precomputed scores for all 9 billion possible single-nucleotide variants in the human genome and gene-level essentiality scores are available at https://huggingface.co/collections/songlab/gpn-msa-65319280c93c85e11c803887. Sequence logos derived from GPN-MSA's predictions can be visualized at https://genome.ucsc.edu/s/gbenegas/gpn-msa-sapiens.

References

46. Blanchette, M. et al. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* **14**, 708–715 (2004).
47. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (eds Burstein, J. et al.) 4171–4186 (Association for Computational Linguistics, 2018).
48. Su, J. et al. RoFormer: enhanced transformer with rotary position embedding. *Neurocomputing* **568**, 127063 (2024).
49. Wolf, T. et al. Transformers: state-of-the-art natural language processing. In *Proc. 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (eds Liu, Q. & Schlangen, D.) 38–45 (Association for Computational Linguistics, 2020).
50. McLaren, W. et al. The Ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).
51. Bolognesi, B. et al. The mutational landscape of a prion-like domain. *Nat. Commun.* **10**, 4162 (2019).
52. Consortium, G. P. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
53. Zhou, H. et al. FAVOR: functional annotation of variants online resource and annotator for variation across the human genome. *Nucleic Acids Res.* **51**, D1300–D1311 (2023).
54. McVicker, G., Gordon, D., Davis, C. & Green, P. Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genet.* **5**, e1000471 (2009).
55. Gazal, S. et al. Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. *Nat. Genet.* **49**, 1421–1427 (2017).
56. Benegas, G., Albors, C., Aw, A. J., Ye, C. & Song, Y. S. GPN repository. *GitHub* https://github.com/songlab-cal/gpn (2024).

Acknowledgements

We thank M. Kircher for helpful correspondence regarding CADD. This research is supported in part by the National Institutes of Health (grant R35-GM134922) and a grant from the Koret–UC Berkeley–Tel Aviv University Initiative in Computational Biology and Bioinformatics.

Author contributions

G.B. and Y.S.S. conceptualized and designed the study. G.B. developed and implemented the method. G.B. tested the method and analyzed data, with contributions from C.A., A.J.A. and C.Y. All authors contributed to designing benchmarks, interpreting the results and writing the paper. Y.S.S. supervised the project.

Competing interests

The authors declare no competing interests.

Additional information

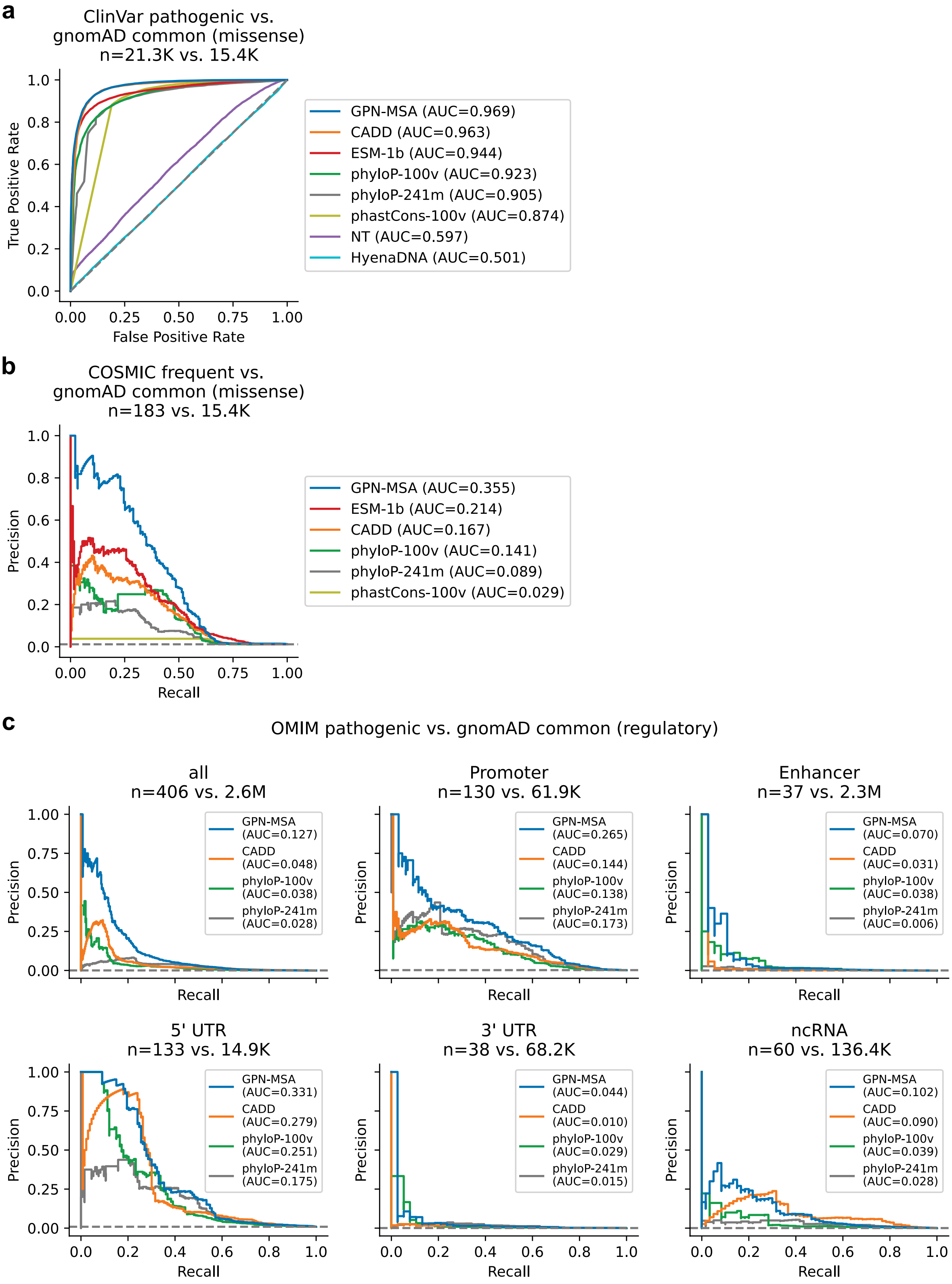
Extended data is available for this paper at https://doi.org/10.1038/s41587-024-02511-w.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-024-02511-w.

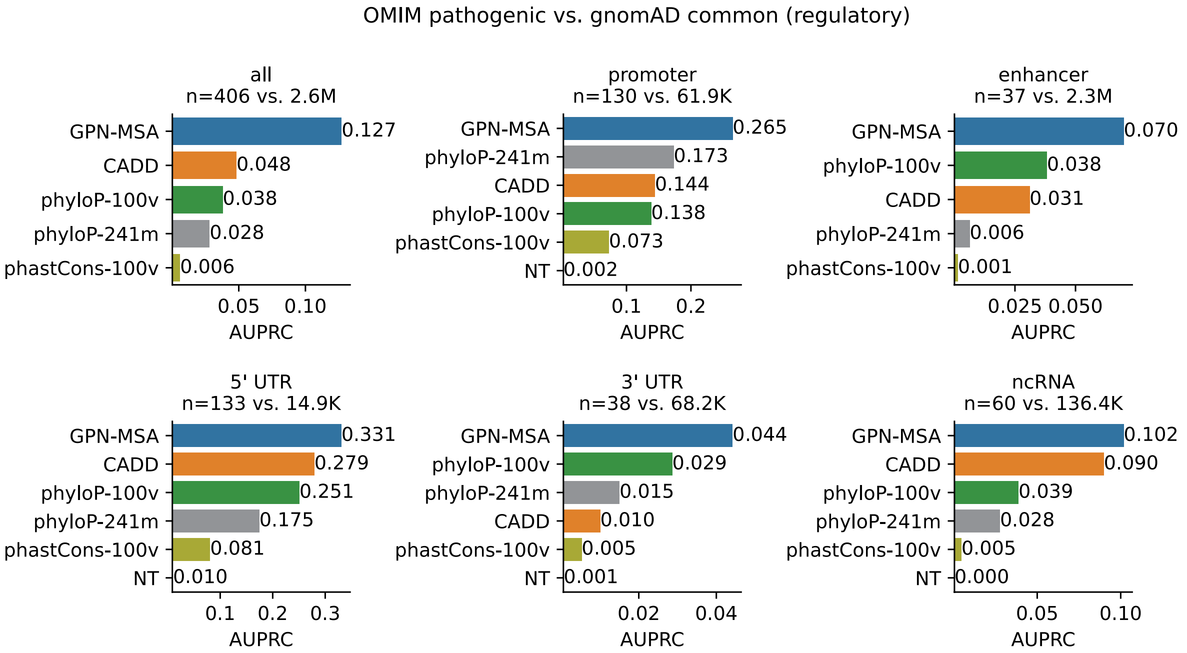
Correspondence and requests for materials should be addressed to Yun S. Song.

Peer review information *Nature Biotechnology* thanks Konrad Karczewski and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

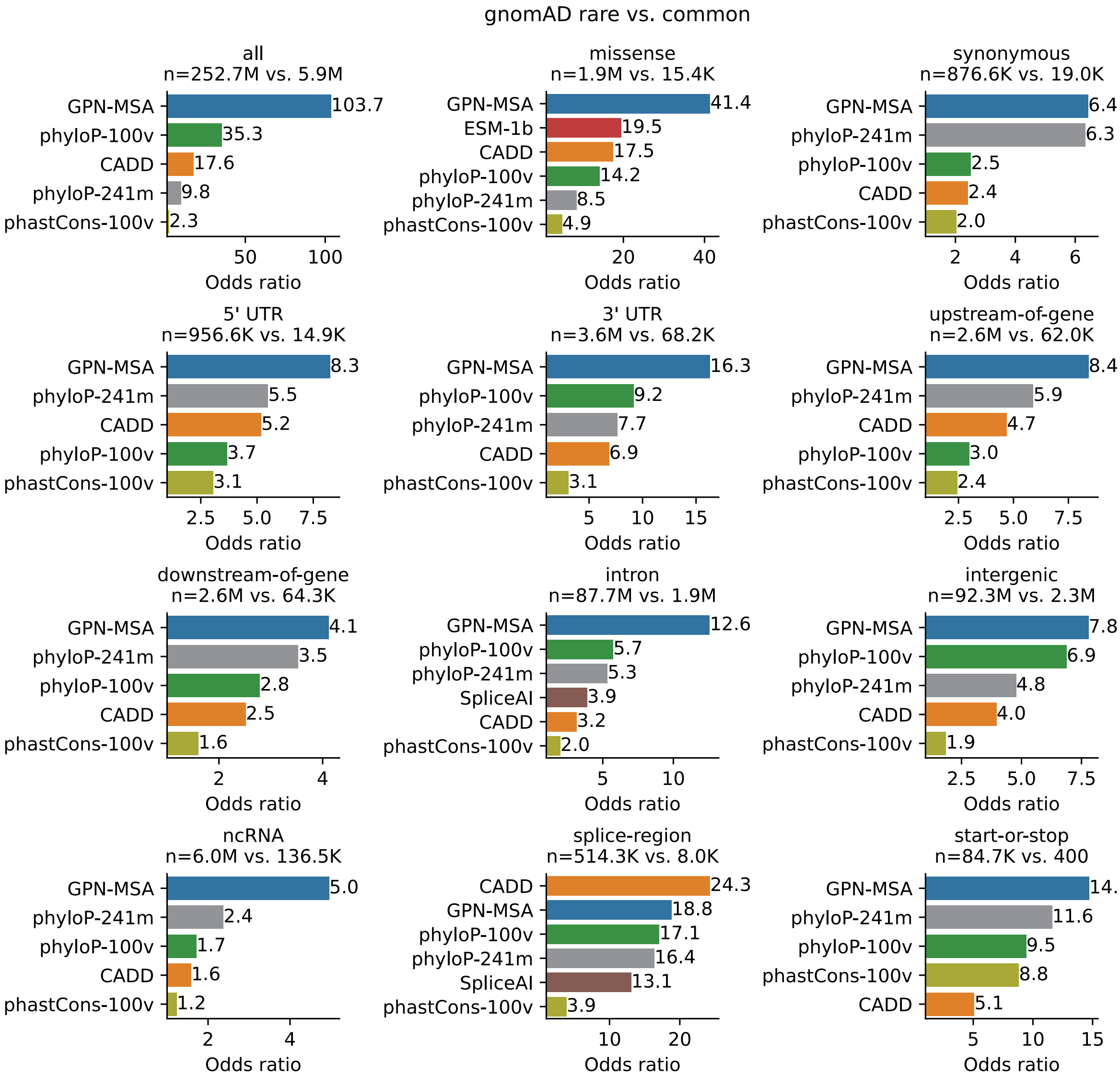
Reprints and permissions information is available at www.nature.com/reprints.



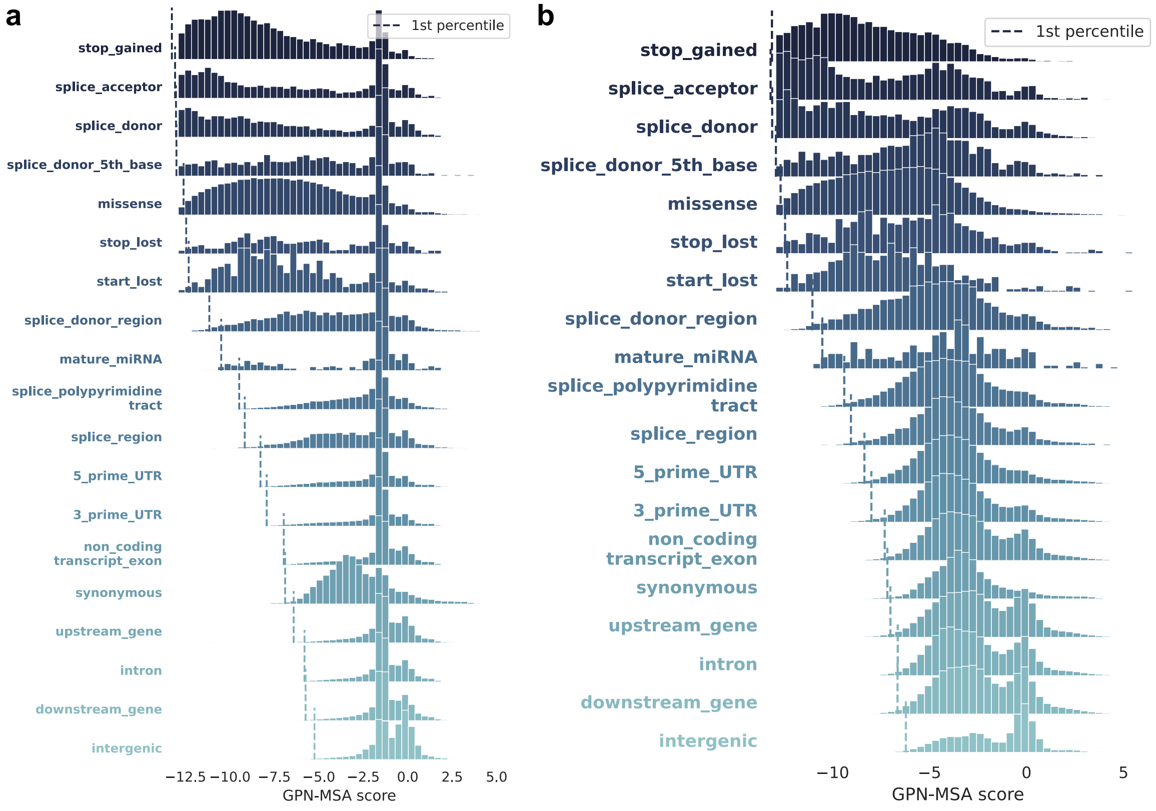
Extended Data Fig. 1 | Receiver Operating Characteristic and Precision-Recall curves for variant effect prediction. (**a**) Same setting as Fig. 2c. (**b**) Same setting as Fig. 2e. (**c**) Same setting as Fig. 2f.



Extended Data Fig. 2 | OMIM performance for specific non-coding variant categories. Area under the precision-recall curve (AUPRC) is considered because of the extreme class imbalance.

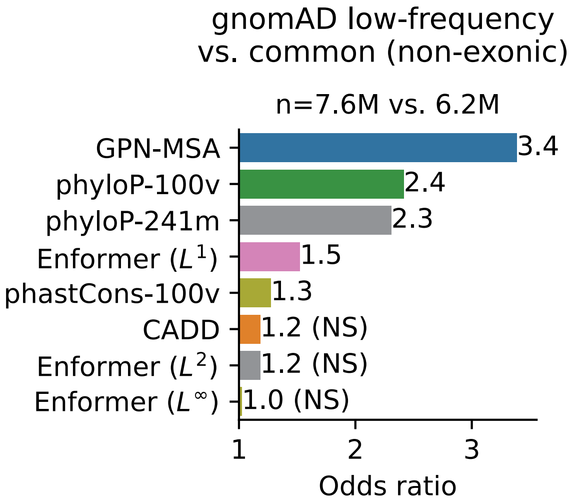


Extended Data Fig. 3 | gnomAD rare vs. common odds ratios for different variant categories. Enrichment of rare (singletons) vs. common (MAF > 5%) gnomAD variants in the tail of deleterious scores (the threshold was chosen such that each score makes 30 false discoveries). Odds ratios and p-values were computed using one-sided Fisher's exact test. All shown odds ratios have p-value < 0.05.

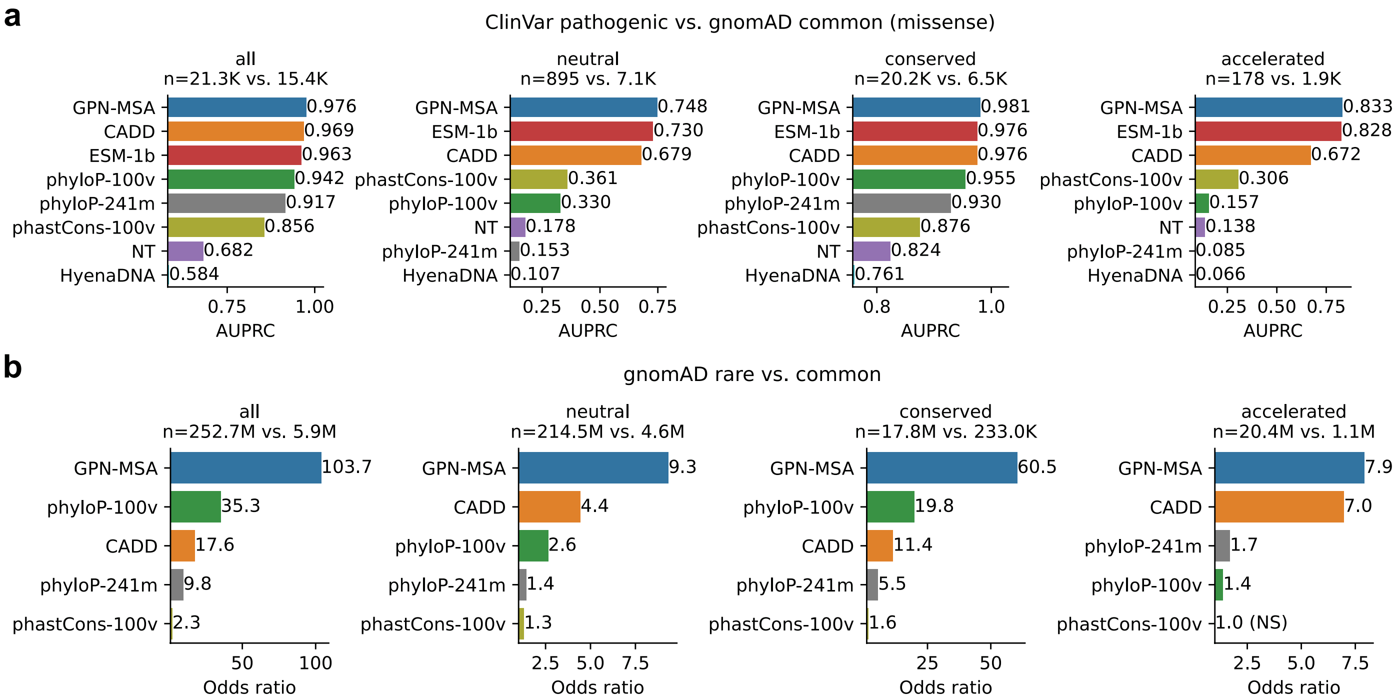


Extended Data Fig. 4 | In silico mutagenesis. Distribution of GPN-MSA scores of a random subset of 10M SNPs in held-out chromosome 22, across categories, sorted by first percentile (dashed vertical lines). (**a**) Default model. (**b**) Without random replacement with another nucleotide at non-conserved positions.

https://doi.org/10.1038/s41587-024-02511-w

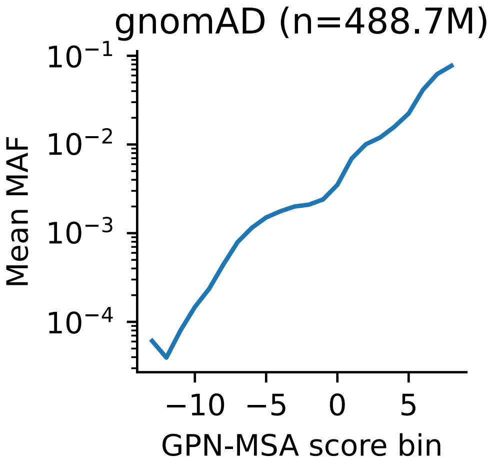


Extended Data Fig. 5 | Comparison with Enformer. Enrichment of low-frequency (0.5% < AF < 5%) vs. common (MAF > 5%) gnomAD non-exonic variants in the tail of deleterious scores (the threshold was chosen such that each score makes 30 false discoveries). Odds ratios and p-values were computed using one-sided Fisher's exact test. All shown odds ratios have p-value < 0.05, unless noted otherwise.

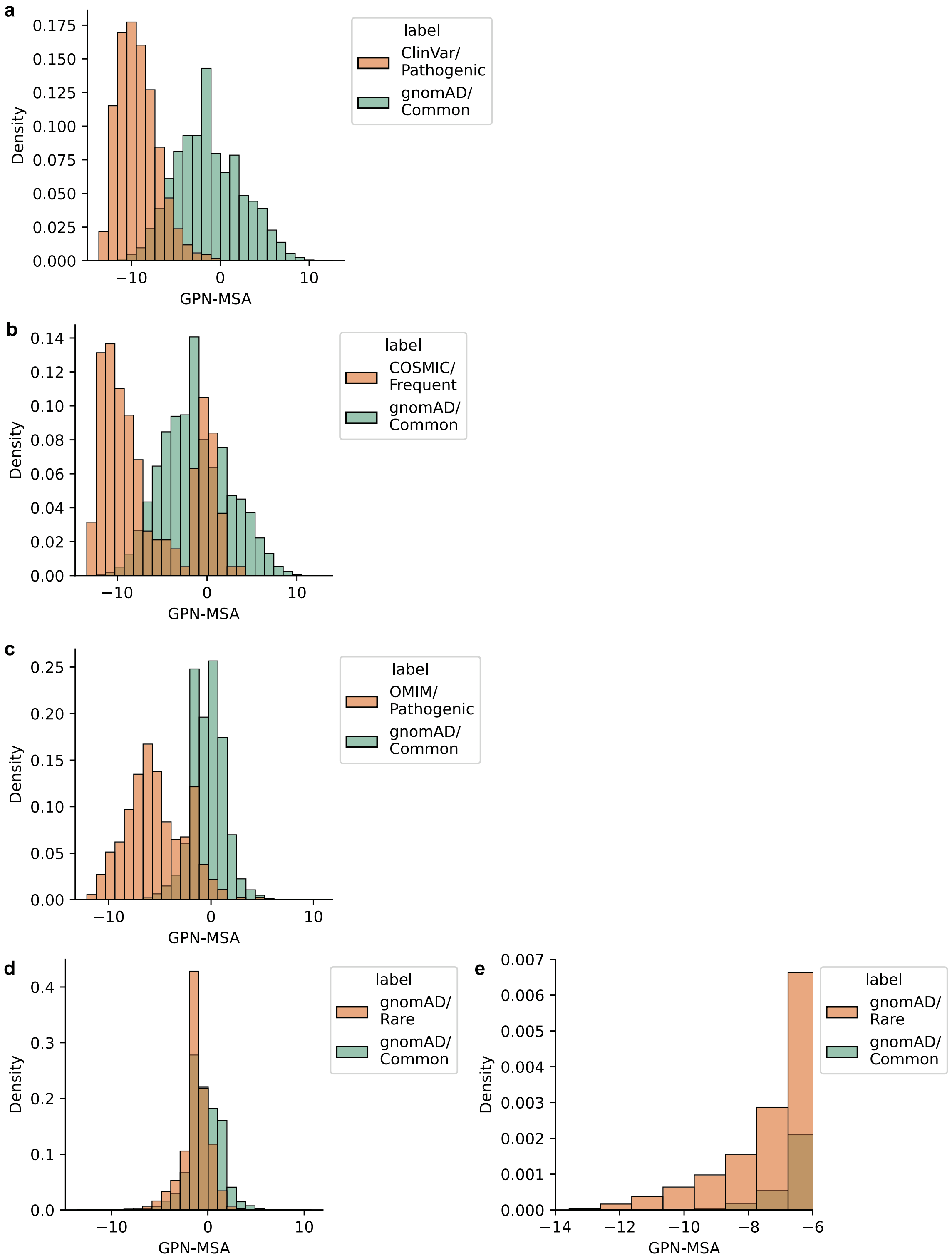


Extended Data Fig. 6 | Performance stratified by putative evolutionary class. Conservation and acceleration were defined using phyloP-241m cutoffs at $p < 0.05$. phyloP[14] is a statistical phylogenetic test. (**a**) Same setting as Fig. 2c, using AUPRC instead of AUROC since the class balance varies substantially. (**b**) Same setting as Fig. 2g. Enrichment of rare (singletons) vs. common (MAF > 5%) gnomAD variants in the tail of deleterious scores (the threshold was chosen such that each score makes 30 false discoveries). Odds ratios and p-values were computed using one-sided Fisher's exact test. All shown odds ratios have p-value < 0.05.

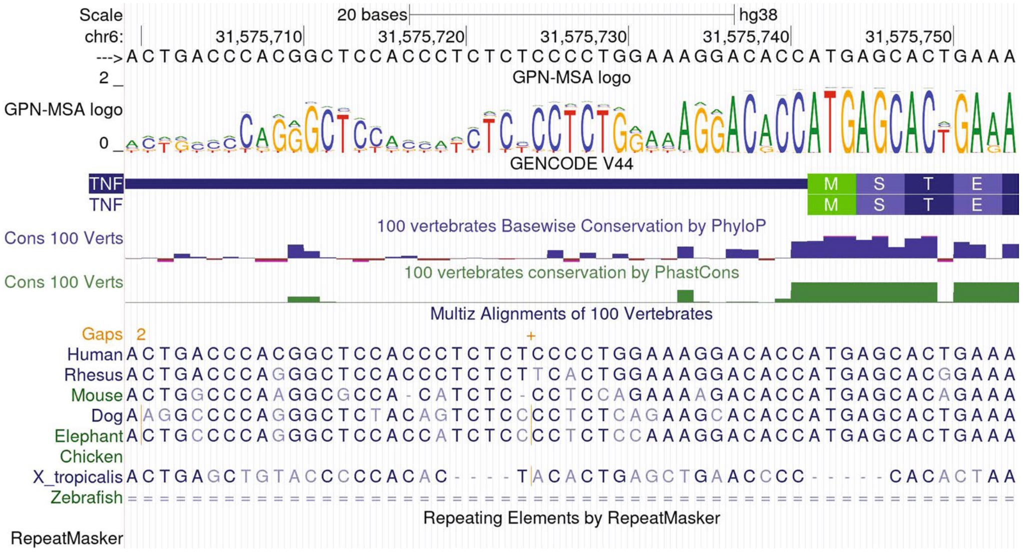
https://doi.org/10.1038/s41587-024-02511-w



Extended Data Fig. 7 | Mean minor allele frequency (MAF) vs. GPN-MSA scores in the full set of gnomAD bi-allelic sites. GPN-MSA score bins are [−13.5, −12.5), [−12.5, −11.5), …, [8.5, 9.5).



Extended Data Fig. 8 | Histogram of GPN-MSA scores. a) Scores for Fig. 2c. (**b**) Scores for Fig. 2e. (**c**) Scores for Fig. 2f. (**d**) Scores for Fig. 2g. (**e**) A zoomed-in version of (**d**) highlighting the left tail.


Scale
20 bases
hg38
chr6:
31,575,710
31,575,720
31,575,730
31,575,740
31,575,750
---> ACTGACCCACGGCTCCACCCTCTCTCCCCTGGAAAGGACACCATGAGCACTGAAA
2 _
GPN-MSA logo
0 _
GENCODE V44
TNF
TNF
M S T E
M S T E
Cons 100 Verts
100 vertebrates Basewise Conservation by PhyloP
Cons 100 Verts
100 vertebrates conservation by PhastCons
Multiz Alignments of 100 Vertebrates
Gaps 2
Repeating Elements by RepeatMasker
RepeatMasker

```
Human       ACTGACCCACGGCTCCACCCTCTCTCCCCTGGAAAGGACACCATGAGCACTGAAA
Rhesus      ACTGACCCAGGGCTCCACCCTCTCTTCACTGGAAAGGACACCATGAGCACGGAAA
Mouse       ACTGGCCCAAGGCGCCA-CATCTC-CCTCCAGAAAAGACACCATGAGCACAGAAA
Dog         AAGGCCCCAGGGCTCTACAGTCTCCCCTCTCAGAAGCACACCATGAGCACTGAAA
Elephant    ACTGCCCCAGGGCTCCACCATCTCCCCTCTCCAAAGGACACCATGAGCACTGAAA
Chicken
X_tropicalis ACTGAGCTGTACCCCACAC----TACACTGAGCTGAACCCC-----CACACTAA
Zebrafish   =======================================================
```

Extended Data Fig. 9 | GPN-MSA logo track on the UCSC Genome Browser. Shown region: chr6:31,575,700-31,575,754.

Extended Data Table 1 | Ablation study

	ClinVar		COSMIC		OMIM		gnomAD	
	mean	max	mean	max	mean	max	mean	max
Default	0.969	0.969	0.348	0.355	0.124	0.133	41.377	43.118
w/o MSA	0.584	0.585	0.012	0.012	0.000	0.000	2.773	3.093
MSA frequency (no neural net)	0.950	0.950	0.270	0.270	0.031	0.031	15.122	15.122
Combined phyloP and phastCons	0.928	0.928	0.141	0.141	0.039	0.039	13.919	13.919
Train on 50% most conserved	0.963	0.963	0.241	0.257	0.125	0.139	32.116	35.286
Train on 100% of genome	0.960	0.961	0.210	0.224	0.120	0.122	27.964	30.717
Include closest primates	0.954	0.956	0.214	0.219	0.139	0.156	24.674	26.721
51 mammals	0.964	0.965	0.323	0.325	0.073	0.076	41.381	43.408
51 vertebrates	0.967	0.967	0.342	0.347	0.101	0.108	41.624	46.792
Don't upweight conserved	0.967	0.967	0.303	0.309	0.133	0.152	37.094	39.991
Don't replace non-conserved	0.966	0.967	0.333	0.340	0.111	0.112	41.853	45.213
Window size = 256	0.969	0.969	0.366	0.376	0.119	0.128	37.908	42.957
Window size = 64	0.968	0.968	0.340	0.348	0.111	0.119	37.478	39.522
Window size = 32	0.967	0.967	0.323	0.328	0.108	0.121	36.664	38.866
Window size = 16	0.964	0.965	0.246	0.250	0.095	0.106	29.203	31.623
Window size = 8	0.961	0.961	0.188	0.209	0.080	0.086	21.159	21.943
Window size = 4	0.942	0.943	0.131	0.155	0.052	0.053	13.457	14.789

Performance of three random seeds of each independent ablation on four variant effect prediction metrics. ClinVar (AUROC): same setting as Fig. 2c. COSMIC (AUPRC): same setting as Fig. 2e. OMIM (AUPRC): same setting as Fig. 2f. gnomAD (odds ratio): same approach as Fig. 2g, including all variant types, but subsetting the rare variants to match the number of common variants. Ablations are detailed in Methods.

nature portfolio

Corresponding author(s): Yun S. Song

Last updated by author(s): Oct 24, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	□	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	□	A description of all covariates tested
□	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	□	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	□	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No data collection in this study.
Data analysis	All code is publicly available at https://github.com/songlab-cal/gpn. Ensemble VEP Release 109 was used. Huggingface Transformers version 4.29.2 was used.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The whole-genome alignment was downloaded from https://hgdownload.soe.ucsc.edu/goldenPath/hg38/multiz100way/maf. ClinVar variants (release 20230730) were downloaded from https://ftp.ncbi.nlm.nih.gov/pub/clinvar/. COSMIC variants (v98) were downloaded from https:

//cancer.sanger.ac.uk/cosmic/download/cosmic. OMIM variants were obtained from the supplementary material in Smedley et al 2016 [27]. gnomAD variants (version 3.1.2) were downloaded
from https://gnomad.broadinstitute.org/data. ProteinGym variants (v0.1) were downloaded
from https://proteingym.org/download. DepMap gene dependency data (Public 23Q4 release)
was downloaded from https://depmap.org/portal/data_page. FAVOR annotations were downloaded from https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/
1VGTJI.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We did not collect any data, therefore sample size calculation is not relevant. "n" in the manuscript refers to the number of variants available in published datasets.
Data exclusions	No data were excluded from the analyses.
Replication	No experimental findings were disclosed, hence no replication was performed.
Randomization	Randomization was not relevant for this study as it involved a reanalysis of published datasets.
Blinding	Not applicable. No new data was collected.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging