GPN-MSA: an alignment-based DNA language model for genome-wide variant effect prediction

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

[1]Graduate Group in Computational Biology, University of California, Berkeley
[2]Computer Science Division, University of California, Berkeley
[3]Department of Statistics, University of California, Berkeley
[4]Center for Computational Biology, University of California, Berkeley

Abstract

Whereas protein language models have demonstrated remarkable efficacy in predicting the effects of missense variants, DNA counterparts have not yet achieved a similar competitive edge for genome-wide variant effect predictions, especially in complex genomes such as that of humans. To address this challenge, we here introduce GPN-MSA, a novel framework for DNA language models that leverages whole-genome sequence alignments across multiple species and takes only a few hours to train. Across several benchmarks on clinical databases (ClinVar, COSMIC, and OMIM) and population genomic data (gnomAD), our model for the human genome achieves outstanding performance on deleteriousness prediction for both coding and non-coding variants.

[*] These authors contributed equally to this work.

[†] To whom correspondence should be addressed: yss@berkeley.edu

Introduction

With the rising trend in 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 highlights analysis of functional rare variants as the biggest contribution of human genetics to drug discovery [4].

Language models are gaining traction as deleteriousness predictors, 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 Pre-trained Network (GPN), a model based on a convolutional neural network trained on unaligned genomes, and showed that it achieves excellent variant effect prediction results in the compact genome of *Arabidopsis thaliana* [8]. The human genome – which harbors a similar number of genes but interspersed over nearly 23 times larger regions and contains much more repetitive elements, most of which may not be functional – is substantially harder to model, however. In fact, previous attempts at unsupervised variant effect prediction with human DNA language models (e.g., Nucleotide Transformer [9]) have shown inferior performance compared to simpler conservation scores. Increasing the scale of the model, data, and compute improves performance, but it can still be poor, even for a model trained for 28 days using 128 top-line graphics processing units (GPUs) [9].

To address the above challenge, we here introduce GPN-MSA, a novel DNA language model which is designed for genome-wide variant effect prediction and is based on the biologically-motivated integration of a multiple-sequence alignment (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 previous DNA language models but also current widely-used models such as CADD [12], phyloP [13], ESM-1b [6, 14], Enformer [15], and SpliceAI [16]. Our model takes only 4.75 hours to train on 4 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 variant effect prediction.

Results

GPN-MSA is trained on a whole-genome MSA of 100 vertebrate species (Figure 1a, full tree in Supplementary Figure S1), after suitable processing (Figure 1b) and filtering (Figure 1c). 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 (Figure 1d, Methods). It draws heavy inspiration from the MSA Transformer [17], 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 variant effect prediction performance [5]. Besides 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 considerably more complex than protein alignments – there are also essential differences in the architecture and training process of GPN-MSA from the MSA Transformer (Methods).

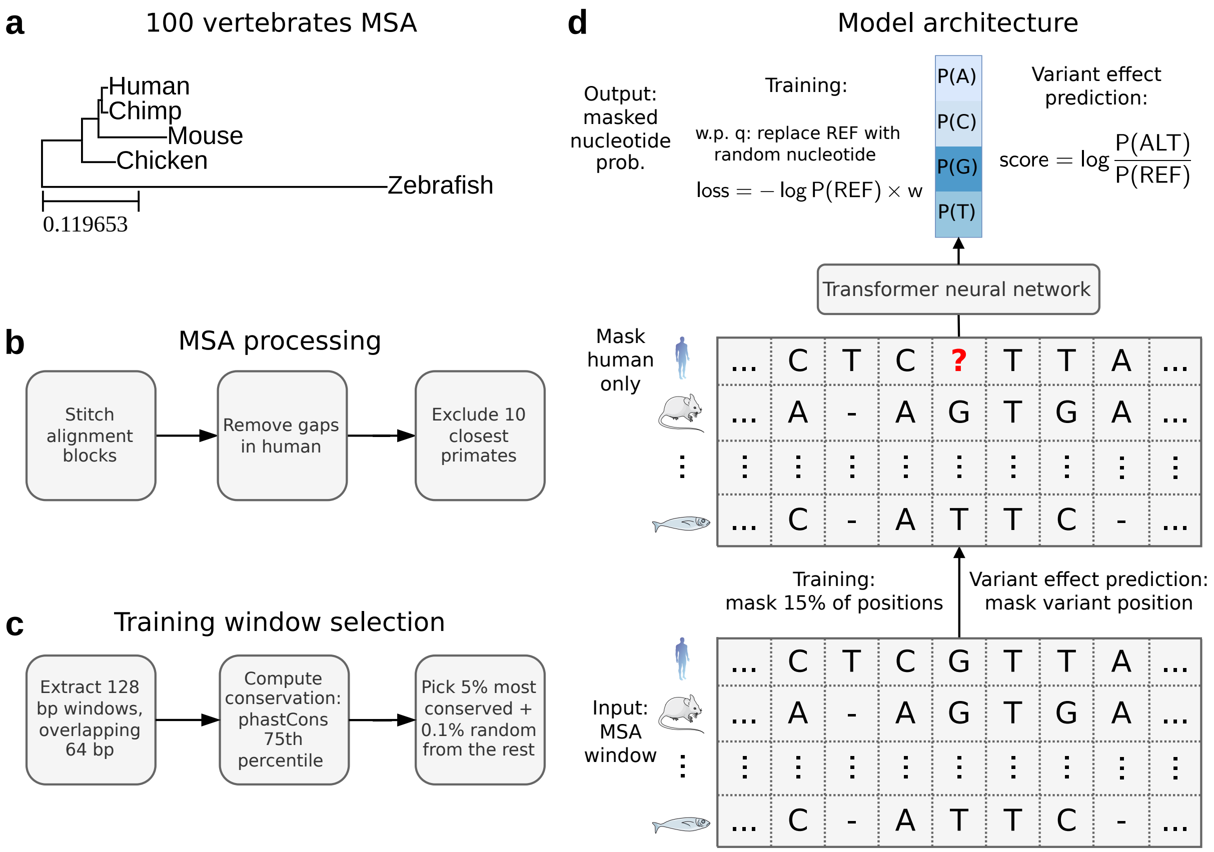
a
100 vertebrates MSA
Human
Chimp
Mouse
Chicken
Zebrafish
0.119653
b
MSA processing
Stitch alignment blocks
Remove gaps in human
Exclude 10 closest primates
c
Training window selection
Extract 128 bp windows, overlapping 64 bp
Compute conservation: phastCons 75th percentile
Pick 5% most conserved + 0.1% random from the rest
d
Model architecture
Output: masked nucleotide prob.
Training:
w.p. q: replace REF with random nucleotide
loss = − log P(REF) × w
P(A)
P(C)
P(G)
P(T)
Variant effect prediction:
score = log P(ALT)/P(REF)
Transformer neural network
Mask human only
Training: mask 15% of positions
Variant effect prediction: mask variant position
Input: MSA window
... C T C ? T T A ...
... A - A G T G A ...
... C - A T T C - ...
... C T C G T T A ...

Figure 1: **Overview of GPN-MSA. (a)** Subsampled phylogenetic tree of 100 vertebrate species constituting the whole-genome MSA (full tree in Supplementary Figure S1). **(b)** 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 10 primate species closest to human (Chimp to squirrel monkey). **(c)** 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. **(d)** Model architecture. The input is a 128-bp MSA window where certain positions in the human reference have been 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 variant effect prediction, 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 up-weight conserved elements (Methods). As data augmentation in non-conserved regions, prior to computing the loss, the reference is sometimes replaced by a random nucleotide (Methods). The GPN-MSA variant effect prediction score is defined as the log-likelihood ratio between the alternate and reference allele. REF: reference allele. ALT: alternate allele.

We demonstrate the capability of GPN-MSA to improve 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 utilized in training. Nevertheless, GPN-MSA can still capture several functional attributes of variants, such as epigenetic marks and the impact of natural selection (Supplementary Figure S2, Supplementary Figure S3).

For evaluation, we first consider the classification of ClinVar [18] pathogenic vs. common missense variants in gnomAD [19]. We use common variants as control instead of ClinVar benign-labeled variants, as recommended by the developers of CADD to reduce ascertainment bias [12]. We find that GPN-MSA achieves the best performance compared to genome-wide predictors CADD [20], phyloP [13], the Nucleotide Transformer (NT) [9], and the missense-specific ESM-1b [6, 14] (Figure 2a, Supplementary Figure S4a).

Next, we consider the classification of somatic missense variants frequently observed across cancer tumors (COSMIC, the Catalogue of Somatic Mutations in Cancer [21]) vs. gnomAD common missense variants. Because of the extreme class imbalance in this case, we focus on the precision and recall metrics. GPN-MSA again achieves the highest performance, with substantial margins of improvement over other models (Figure 2b, Supplementary Figure S4b).

Moving on to regulatory variants, we evaluate on the classification of a curated set of variants implicated in Mendelian disorders (OMIM, Online Mendelian Inheritance in Man [22]) vs. 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 (Figure 2c, Supplementary Figure S4c). 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 (Supplementary Figure S4c).

Lastly, we evaluate on the enrichment of rare vs. common gnomAD variants in the tail of deleteriousness scores. Deleterious mutations should be under purifying selection and hence their frequencies tend to be low in the population. 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 overall enrichment, as well as in each variant category, with different margins (Figure 2d). In the case of intron variants, it also outperforms SpliceAI [16], a state-of-the-art splicing predictor. We note that the overall performance 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 vs. common gnomAD variants in gene flanking and intergenic regions, GPN-MSA achieves a substantially improved performance over Enformer [15] (Figure 2e). 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 they were not designed for. It is also worth noting that although phyloP trained on 241 mammals (Zoonomia) was recently proposed as a deleteriousness predictor [23], the older vertebrate phyloP actually achieves better results in most of our benchmarks (Supplementary Figure S5).

To understand the importance of different components of our model, we perform an ablation study and assess the impact on variant effect prediction performance (Supplementary Figure S6). We find that the inclusion of the MSA is most critical and that different ways of prioritizing conserved regions can have a significant impact on the results.

GPN-MSA's predictions for every position of chromosome 6 can be visualized as sequence logos [24] in the UCSC Genome Browser [25, 26] (example in Supplementary Figure S7); we plan to release predictions for all ∼9 billion possible single nucleotide variants in the human genome using the final, revised model upon publication. We also provide a Jupyter notebook detailing how to run predictions on a given VCF file using our trained model.

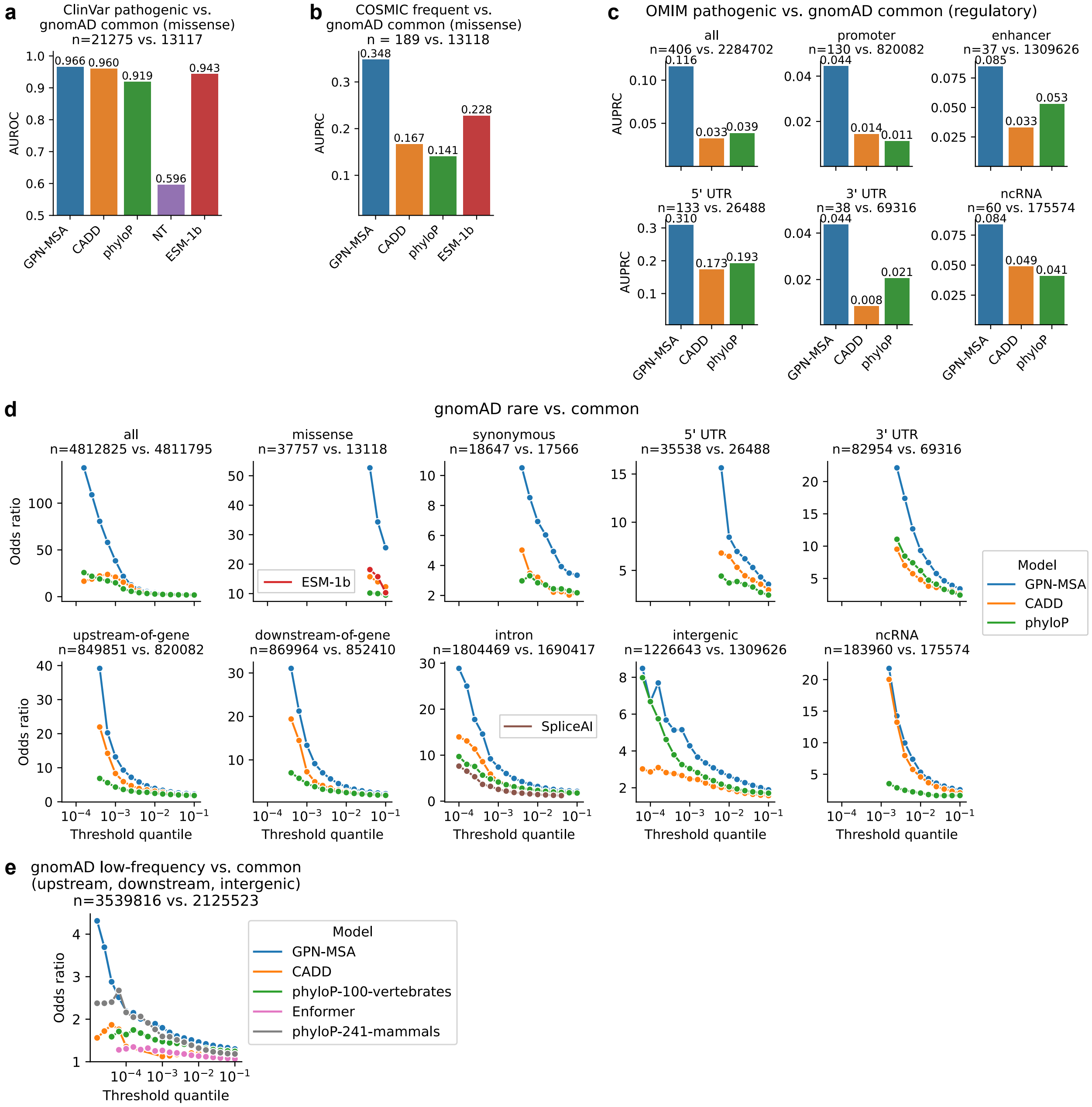

Figure 2: **Comparison of variant effect prediction results.** **(a)** Classification of ClinVar pathogenic vs. gnomAD common missense variants. NT: Nucleotide Transformer (version `2.5b-multi-species`). **(b)** Classification of COSMIC frequent (frequency $> 0.1\%$) vs. gnomAD common missense variants. **(c)** Classification of OMIM pathogenic vs. 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, after removing any overlap with missense variants. **(d)** Enrichment of rare (singletons) vs. common (MAF $> 5\%$) gnomAD variants in the tail of deleterious scores (defined using different threshold quantiles, e.g. the 10% most extreme scores are considered deleterious, or the 1% most extreme). In categories other than "all" or "missense", we removed any overlap with missense variants. Odds ratios and p-values were computed using Fisher's exact test. All shown odds ratios have p-value < 0.05. The minimum threshold was chosen such that no score has less than 10 counts in the contingency table. **(e)** Comparison with Enformer. Enrichment of low-frequency ($0.5\% <$ AF $< 5\%$) vs. common (MAF $> 5\%$) gnomAD flanking and intergenic variants in the tail of deleterious scores. We removed any overlap with

Figure 2 *(continued)*: missense variants. Enformer scores were calculated as L^2 norm of delta predictions. We used the same odds ratio plotting considerations as in (d) AUROC: area under the receiving operating characteristic curve. AUPRC: area under the precision-recall curve. MAF: minor allele frequency. AF: allele frequency. "phyloP" refers to the statistic computed on the 100 vertebrates alignment.

Discussion

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 state-of-the-art performance in humans on a number of clinically-relevant variant effect prediction datasets. Lastly, the general approach we have 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 avenues of exploration [9, 27, 28]. In contrast, in our work we focus on the explicit modeling of related sequences (known as retrieval augmentation in natural language processing [29]). This has led to a highly computationally efficient model and state-of-the-art variant effect prediction performance for both coding and non-coding variants. It remains to be explored how useful GPN-MSA's learned representations would be for downstream applications, e.g., for genome annotation or gene expression prediction. Expanding the context length, possibly through leveraging recent technical developments [27], might be beneficial for such tasks.

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 variant effect prediction. This observation has led us to exclude most primate genomes during training. To tackle this limitation, we are actively exploring alternative training objectives which 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 [30]. We envision that several other statistical genetics tasks can be empowered by GPN-MSA, such as functionally informed fine-mapping [31] and polygenic risk scores [32].

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.

Methods

MSA Processing

The `multiz` [33] 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 10 primate species closest to human were removed. We also downloaded associated conservation scores phastCons [34] `https://hgdownload.soe.ucsc.edu/goldenPath/hg38/phastCons100way` and phyloP [13] `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 there is no extreme distribution shift when performing variant effect prediction in non-conserved 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 possible testing (not actually used in this study).

Model Architecture

We adopt the general approach of masked language modeling [35]. As a general caveat, in this work we did not systematically tune hyperparameters, so they are likely far from optimal. The input is a 128-bp MSA window where certain positions in the human reference have been masked, and the goal is 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 are masked. During variant effect prediction, only the variant position is masked. The 1-hot encodings of nucleotides from different species at each position are first concatenated. Then, the sequence of MSA columns is processed through a Transformer neural network (RoFormer [36]) 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 on the reference sequence with a weighted cross-entropy loss.

Our considerations for the loss weight were the following: downweighting repeats and upweighting conserved elements (so wrong predictions in neutral regions are penalized less). We introduce a smoothed version of phastCons, phastCons_M, as the max of phastCons over a window of 7 nucleotides. The goal was to not only give importance to conserved regions, but 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 10-fold downweighting on repetitive elements [8] plus upweighting based on both phyloP and phastCons_M.

As data augmentation in non-conserved regions, prior to computing the loss, the reference is replaced by a random nucleotide with a certain probability q:

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

The intention is to guide the model to assign more neutral scores in non-conserved regions.

Our code is based on the Hugging Face Transformers library [37]. All models were trained with default hyperparameters [1] (e.g. 12 layers with 12 attention heads each) except for the ones listed in Supplementary Table S1. The total number of parameters is approximately 86 million. We performed early stopping based on validation loss. We manage to train the model in approximately 4.75 hours using 4 NVIDIA A100 GPUs.

The GPN-MSA variant effect prediction score is defined as the log-likelihood ratio between the alternate and reference allele. In our experiments, we average the predictions from the positive and negative strand. With our 4 NVIDIA A100 GPUs, we manage to score approximately 5 million variants per hour.

[1] https://huggingface.co/docs/transformers/model_doc/roformer#transformers.RoFormerConfig

Differences between GPN-MSA and MSA Transformer

While the MSA Transformer takes as input an arbitrary set of aligned sequences, GPN-MSA is 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. Variant effect prediction, masking only the target sequence (in our case, human), is identical [5]. Since variant effect prediction is our main goal, during training we also only mask positions from the target sequence. The MSA Transformer, however, proposes masking MSA entries at random during training, based on results from structure prediction, their intended application.

Ablation Study

We performed an ablation study to understand the impact of each of our design choices on variant effect prediction when modified independently (Supplementary Figure S6). For each setting, three replicate models with different seeds were trained, where applicable. Since we hold the rest of the hyperparameters fixed, results should be interpreted as differences given a similar training procedure and compute budget.

- 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.
- Train on 50% 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 10 primates closest to human.
- Don't upweight conserved: do not upweight the loss function on conserved elements.
- Don't replace non-conserved: do not replace the reference in non-conserved positions with random nucleotides when computing the loss function.

Modeling the single human sequence instead of the MSA has by far the biggest impact. Using the column-wide MSA frequencies as predictor also shows a large decrease in performance. Including primate species close to human, or training on less conserved regions, have a moderate impact on performance. Finally, of relatively minor impact are removing the upweighting of conserved elements or removing the data augmentation procedure of replacing nucleotides in non-conserved positions.

Variant Effect Prediction (VEP) glossary

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

VEP data sources:

- ClinVar [18]: downloaded release `20230730`.
- COSMIC [21]: 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 [22]: downloaded a set of curated pathogenic regulatory variants.
- gnomAD [19]: downloaded version `3.1.2` and filtered to autosomal variants with allele number of at least $2 \times 70\,000$, besides the official quality-control flags. In each autosomal chromosome, selected all common variants (minor allele frequency $> 5\%$) as well as an equally-sized subset of rare variants (singletons).

VEP metrics:

- ClinVar: area under the receiving operating characteristic curve (AUROC) for classification of ClinVar "Pathogenic" vs. gnomAD common missense variants.
- COSMIC: area under the precision-recall curve (AUPRC) for classifying COSMIC frequent (frequency $> 0.1\%$) vs. gnomAD common missense variants.
- OMIM: AUPRC for classification of OMIM pathogenic vs. 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, after removing any overlap with missense variants.
- gnomAD: enrichment of rare vs. common gnomAD variants in the tail of deleterious scores (defined using different threshold quantiles, e.g. the 10% most extreme scores are considered deleterious, or the 1% most extreme). In categories other than "all" or "missense", we removed any overlap with missense variants.

VEP scores:

- GPN-MSA: log-likelihood ratio between alternate and reference allele. Predictions from both strands were averaged.
- CADD: raw scores, negated so lower means more deleterious.
- phyloP: computed on 100 vertebrate alignment, negated so lower means more deleterious.
- Nucleotide Transformer (NT): the center 6-mer was masked and the score was computed as the log-likelihood ratio between alternate and reference 6-mer. Predictions from both strands were averaged. Given the high computational requirements, we only scored variants for the ClinVar metric. The performance of the four different models can be seen in Supplementary Figure S8.
- 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 minor allele frequency (MAF) greater than 0.5% in any 1000 Genomes population [38] were downloaded from `https://console.cloud.google.com/storage/browser/dm-enformer/variant-scores`. These were intersected with upstream-of-gene, downstream-of-gene and intergenic variants with gnomAD MAF greater than 0.5%. 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 (SNP Activity Difference or SAD). We found that the L^1 and L^2 norms seem to perform similarly, better than the L^∞ norm (Supplementary Figure S9).

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 separately on four datasets curated across four public variant interpretation databases, ClinVar, COSMIC, OMIM and gnomAD. We used 18 functional annotations obtained from the FAVOR database [39] (accessed via Harvard Dataverse on April 10, 2023), which measure both impact of natural selection and gene regulatory activity of a variant (see Supplementary Table S2). For clarity, we collect computational details of the functional annotations and summarize them below.

- B Statistic [40], nucleotide diversity [41] and recombination rate [41] are mathematical quantities derived from evolutionary models, and are computed directly on the genomic position of the variant. They provide population-genetic interpretation of the impact of natural selection on the variant.

- Epigenetic tracks, RNA-seq, DNAse-seq, percent GC and percent CpG were all computed on genomic positions, to be included as training features in CADD [20]. 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 based on its genomic position. Whenever a variant is not annotatable for a track (e.g., RNA-seq level for a non-exonic variant), an `NA` value is assigned.

We found evidence of GPN-MSA capturing gene regulatory activity and impact of natural selection. Across all four datasets, significant negative correlations were observed between GPN-MSA and 8 histone mark levels (not including H3K9me3 and H3K27me3, which are recognized gene repressors; see Supplementary Figure S2). Additionally, GPN-MSA was positively correlated with nucleotide diversity and B statistic — for the both of which a smaller value indicates stronger impact of natural selection. In general, the strongest correlations of any annotation were observed in the dataset consisting of ClinVar pathogenic and gnomAD common missense variants.

Next, 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 within each dataset, across all 18 annotations. Sample sizes were reasonably large between the target and background samples: the minimum sample size of any target set was 124. We found significant enrichment ($p < 0.05$ after controlling for FWER) of H4K20me1, a transcription activation mark, and RNA-seq levels in each dataset, and significant depletion of nucleotide diversity (Supplementary Figure S3). Interestingly, for H3K27me3, generally recognized as a gene repressor, all but the COSMIC pathogenic and gnomAD common missense dataset reported enrichment in the target variants. These results suggest that extremely negative GPN-MSA scores could potentially prioritize variants with impact on gene expression and regulation.

Code Availability

Code to reproduce all results is available at https://github.com/songlab-cal/gpn.

Data Availability

The processed whole-genome MSA is available at https://huggingface.co/datasets/songlab/multiz100way. The specific genomic windows used for training are available at https://huggingface.co/datasets/songlab/gpn-msa-sapiens-dataset. The variants used for benchmarking (including predictions) are available at https://huggingface.co/datasets/songlab/human_variants. Predictions for all 9 billion possible single nucleotide variants in the human genome will be provided with the final, revised model upon publication. Predictions with the draft model can be performed with the Jupyter notebook at https://github.com/songlab-cal/gpn/blob/main/examples/msa/vep.ipynb. Sequence logo derived from GPN-MSA's predictions (currently only chromosome 6) can be visualized at https://genome.ucsc.edu/s/gbenegas/gpn-msa-sapiens.

Model Availability

The pretrained model is available at https://huggingface.co/songlab/gpn-msa-sapiens.

Acknowledgements

We thank Dr. Martin Kircher for helpful correspondence regarding CADD. This research is supported in part by an NIH grant R35-GM134922 and a grant from the Koret-UC Berkeley-Tel Aviv University Initiative in Computational Biology and Bioinformatics.

References

[1] Rachel L Goldfeder, Dennis P Wall, Muin J Khoury, John PA Ioannidis, and Euan A Ashley. Human genome sequencing at the population scale: a primer on high-throughput DNA sequencing and analysis. *American Journal of Epidemiology*, 186(8):1000–1009, 2017.

[2] Shruti Marwaha, Joshua W Knowles, and Euan A Ashley. A guide for the diagnosis of rare and undiagnosed disease: beyond the exome. *Genome Medicine*, 14(1):1–22, 2022.

[3] Seunggeun Lee, Gonçalo R Abecasis, Michael Boehnke, and Xihong Lin. Rare-variant association analysis: study designs and statistical tests. *The American Journal of Human Genetics*, 95(1):5–23, 2014.

[4] Katerina Trajanoska, Claude Bhérer, Daniel Taliun, Sirui Zhou, J. Brent Richards, and Vincent Mooser. From target discovery to clinical drug development with human genetics. *Nature*, 620(7975):737–745, Aug 2023.

[5] Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alex Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in Neural Information Processing Systems*, 34, 2021.

[6] Nadav Brandes, Grant Goldman, Charlotte H. Wang, Chun Jimmie Ye, and Vasilis Ntranos. Genome-wide prediction of disease variant effects with a deep protein language model. *Nature Genetics*, Aug 2023.

[7] Milind Jagota, Chengzhong Ye, Carlos Albors, Ruchir Rastogi, Antoine Koehl, Nilah Ioannidis, and Yun S. Song. Cross-protein transfer learning substantially improves disease variant prediction. *Genome Biology*, 24(1):1–19, 2023.

[8] Gonzalo Benegas, Sanjit Singh Batra, and Yun S. Song. DNA language models are powerful predictors of genome-wide variant effects. *Proceedings of the National Academy of Sciences, in press*, 2023.

[9] Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza Revilla, Nicolas Lopez Carranza, Adam Henryk Grywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Hassan Sirelkhatim, Guillaume Richard, et al. The Nucleotide Transformer: Building and Evaluating Robust Foundation Models for Human Genomics. *bioRxiv*, pages 2023–01, 2023.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. *Advances in Neural Information Processing Systems*, 30, 2017.

[11] Joel Armstrong, Ian T Fiddes, Mark Diekhans, and Benedict Paten. Whole-genome alignment and comparative annotation. *Annual Review of Animal Biosciences*, 7:41–64, 2019.

[12] Philipp Rentzsch, Max Schubach, Jay Shendure, and Martin Kircher. CADD-Splice—improving genome-wide variant effect prediction using deep learning-derived splice scores. *Genome Medicine*, 13(1):1–12, 2021.

[13] Katherine S Pollard, Melissa J Hubisz, Kate R Rosenbloom, and Adam Siepel. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research*, 20(1):110–121, 2010.

[14] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.

[15] Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R Ledsam, Agnieszka Grabska-Barwinska, Kyle R Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods*, 18(10):1196–1203, 2021.

[16] Kishore Jaganathan, Sofia Kyriazopoulou Panagiotopoulou, Jeremy F McRae, Siavash Fazel Darbandi, David Knowles, Yang I Li, Jack A Kosmicki, Juan Arbelaez, Wenwu Cui, Grace B Schwartz, et al. Predicting splicing from primary sequence with deep learning. *Cell*, 176(3):535–548, 2019.

[17] Roshan M Rao, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. MSA Transformer. In *International Conference on Machine Learning*, pages 8844–8856. PMLR, 2021.

[18] Melissa J Landrum, Shanmuga Chitipiralla, Garth R Brown, Chao Chen, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Wonhee Jang, Kuljeet Kaur, Chunlei Liu, et al. ClinVar: improvements to accessing data. *Nucleic Acids Research*, 48(D1):D835–D844, 2020.

[19] Siwei Chen, Laurent C Francioli, Julia K Goodrich, Ryan L Collins, Masahiro Kanai, Qingbo Wang, Jessica Alföldi, Nicholas A Watts, Christopher Vittal, Laura D Gauthier, et al. A genome-wide mutational constraint map quantified from variation in 76,156 human genomes. *bioRxiv*, pages 2022–03, 2022.

[20] Philipp Rentzsch, Daniela Witten, Gregory M Cooper, Jay Shendure, and Martin Kircher. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research*, 47(D1):D886–D894, 2019.

[21] John G Tate, Sally Bamford, Harry C Jubb, Zbyslaw Sondka, David M Beare, Nidhi Bindal, Harry Boutselakis, Charlotte G Cole, Celestino Creatore, Elisabeth Dawson, et al. COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Research*, 47(D1):D941–D947, 2019.

[22] Damian Smedley, Max Schubach, Julius OB Jacobsen, Sebastian Köhler, Tomasz Zemojtel, Malte Spielmann, Marten Jäger, Harry Hochheiser, Nicole L Washington, Julie A McMurry, et al. A whole-genome analysis framework for effective identification of pathogenic regulatory variants in mendelian disease. *The American Journal of Human Genetics*, 99(3):595–606, 2016.

[23] Patrick F Sullivan, Jennifer RS Meadows, Steven Gazal, BaDoi N Phan, Xue Li, Diane P Genereux, Michael X Dong, Matteo Bianchi, Gregory Andrews, Sharadha Sakthikumar, et al. Leveraging base-pair mammalian constraint to understand genetic variation and human disease. *Science*, 380(6643):eabn2937, 2023.

[24] Thomas D Schneider and R Michael Stephens. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research*, 18(20):6097–6100, 1990.

[25] W James Kent, Charles W Sugnet, Terrence S Furey, Krishna M Roskin, Tom H Pringle, Alan M Zahler, and David Haussler. The human genome browser at UCSC. *Genome Research*, 12(6):996–1006, 2002.

[26] Surag Nair, Arjun Barrett, Daofeng Li, Brian J Raney, Brian T Lee, Peter Kerpedjiev, Vivekanandan Ramalingam, Anusri Pampari, Fritz Lekschas, Ting Wang, et al. The dynseq browser track shows context-specific features at nucleotide resolution. *Nature Genetics*, 54(11):1581–1583, 2022.

[27] Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Callum Birch-Sykes, Michael Wornow, Aman Patel, Clayton Rabideau, Stefano Massaroli, Yoshua Bengio, et al. HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution. *arXiv preprint arXiv:2306.15794*, 2023.

[28] Veniamin Fishman, Yuri Kuratov, Maxim Petrov, Aleksei Shmelev, Denis Shepelin, Nikolay Chekanov, Olga Kardymon, and Mikhail Burtsev. GENA-LM: A Family of Open-Source Foundational Models for Long DNA Sequences. *bioRxiv*, pages 2023–06, 2023.

[29] S Borgeaud, A Mensch, J Hoffmann, T Cai, E Rutherford, K Millican, G Driessche, JB Lespiau, B Damoc, A Clark, et al. Improving language models by retrieving from trillions of tokens. *arXiv preprint arXiv:2112.04426*, 2021.

[30] Daniel J Weiner, Ajay Nadig, Karthik A Jagadeesh, Kushal K Dey, Benjamin M Neale, Elise B Robinson, Konrad J Karczewski, and Luke J O'Connor. Polygenic architecture of rare coding variation across 394,783 exomes. *Nature*, 614(7948):492–499, 2023.

[31] Omer Weissbrod, Farhad Hormozdiari, Christian Benner, Ran Cui, Jacob Ulirsch, Steven Gazal, Armin P Schoech, Bryce Van De Geijn, Yakir Reshef, Carla Márquez-Luna, et al. Functionally informed fine-mapping and polygenic localization of complex trait heritability. *Nature Genetics*, 52(12):1355–1363, 2020.

[32] Carla Márquez-Luna, Steven Gazal, Po-Ru Loh, Samuel S Kim, Nicholas Furlotte, Adam Auton, and Alkes L Price. Incorporating functional priors improves polygenic prediction accuracy in UK Biobank and 23andMe data sets. *Nature Communications*, 12(1):6052, 2021.

[33] Mathieu Blanchette, W James Kent, Cathy Riemer, Laura Elnitski, Arian FA Smit, Krishna M Roskin, Robert Baertsch, Kate Rosenbloom, Hiram Clawson, Eric D Green, et al. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Research*, 14(4):708–715, 2004.

[34] Adam Siepel, Gill Bejerano, Jakob S Pedersen, Angie S Hinrichs, Minmei Hou, Kate Rosenbloom, Hiram Clawson, John Spieth, LaDeana W Hillier, Stephen Richards, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, 15(8):1034–1050, 2005.

[35] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[36] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021.

[37] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv preprint arXiv:1910.03771*, 2019.

[38] 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68, 2015.

[39] Hufeng Zhou, Theodore Arapoglou, Xihao Li, Zilin Li, Xiuwen Zheng, Jill Moore, Abhijith Asok, Sushant Kumar, Elizabeth E Blue, Steven Buyske, et al. FAVOR: functional annotation of variants online resource and annotator for variation across the human genome. *Nucleic Acids Research*, 51(D1):D1300–D1311, 2023.

[40] Graham McVicker, David Gordon, Colleen Davis, and Phil Green. Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genetics*, 5(5):e1000471, 2009.

[41] Steven Gazal, Hilary K Finucane, Nicholas A Furlotte, Po-Ru Loh, Pier Francesco Palamara, Xuanyao Liu, Armin Schoech, Brendan Bulik-Sullivan, Benjamin M Neale, Alexander Gusev, et al. Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. *Nature Genetics*, 49(10):1421–1427, 2017.

Supplementary Tables

Table S1: Training hyperparameters

Weight decay	0.01
Max steps	30 K
Batch size	2048
Learning rate	10^{-4}
Learning rate schedule	Cosine
Learning rate warmup steps	1 K

Table S2: Functional annotations considered in our analysis of functional enrichment. In particular, annotations relying on predictive models are not considered.

Functional Annotation	Category	Definition / Interpretation
Recombination Rate	Local Nucleotide Diversity	How likely a region tends to undergo recombination
Nuclear Diversity	Local Nucleotide Diversity	How likely the region diversifies
B Statistic	Local Nucleotide Diversity	Population-genetic quantity. Lower value means greater impact of selection on removing diversity
Percent GC	Epigenetics	Percent GC in $+/-$ 75bp window
Percent CpG	Epigenetics	Percent CpG in $+/-$ 75bp window
RNA-seq	Epigenetics	Max level over 10 cell lines (ENCODE)
DNase-seq	Epigenetics	Max level over 12 cell lines (ENCODE)
H3K4me1	Epigenetics	Max level over 13 cell lines (ENCODE)
H3K4me2	Epigenetics	Max level over 14 cell lines (ENCODE)
H3K4me3	Epigenetics	Max level over 14 cell lines (ENCODE)
H3K9ac	Epigenetics	Max level over 13 cell lines (ENCODE)
H3K9me3	Epigenetics	Max level over 14 cell lines (ENCODE)
H3K27ac	Epigenetics	Max level over 14 cell lines (ENCODE)
H3K27me3	Epigenetics	Max level over 14 cell lines (ENCODE)
H3K36me3	Epigenetics	Max level over 10 cell lines (ENCODE)
H3K79me2	Epigenetics	Max level over 13 cell lines (ENCODE)
H4K20me1	Epigenetics	Max level over 11 cell lines (ENCODE)
H2AFZ	Epigenetics	Max level over 13 cell lines (ENCODE)

Supplementary Figures

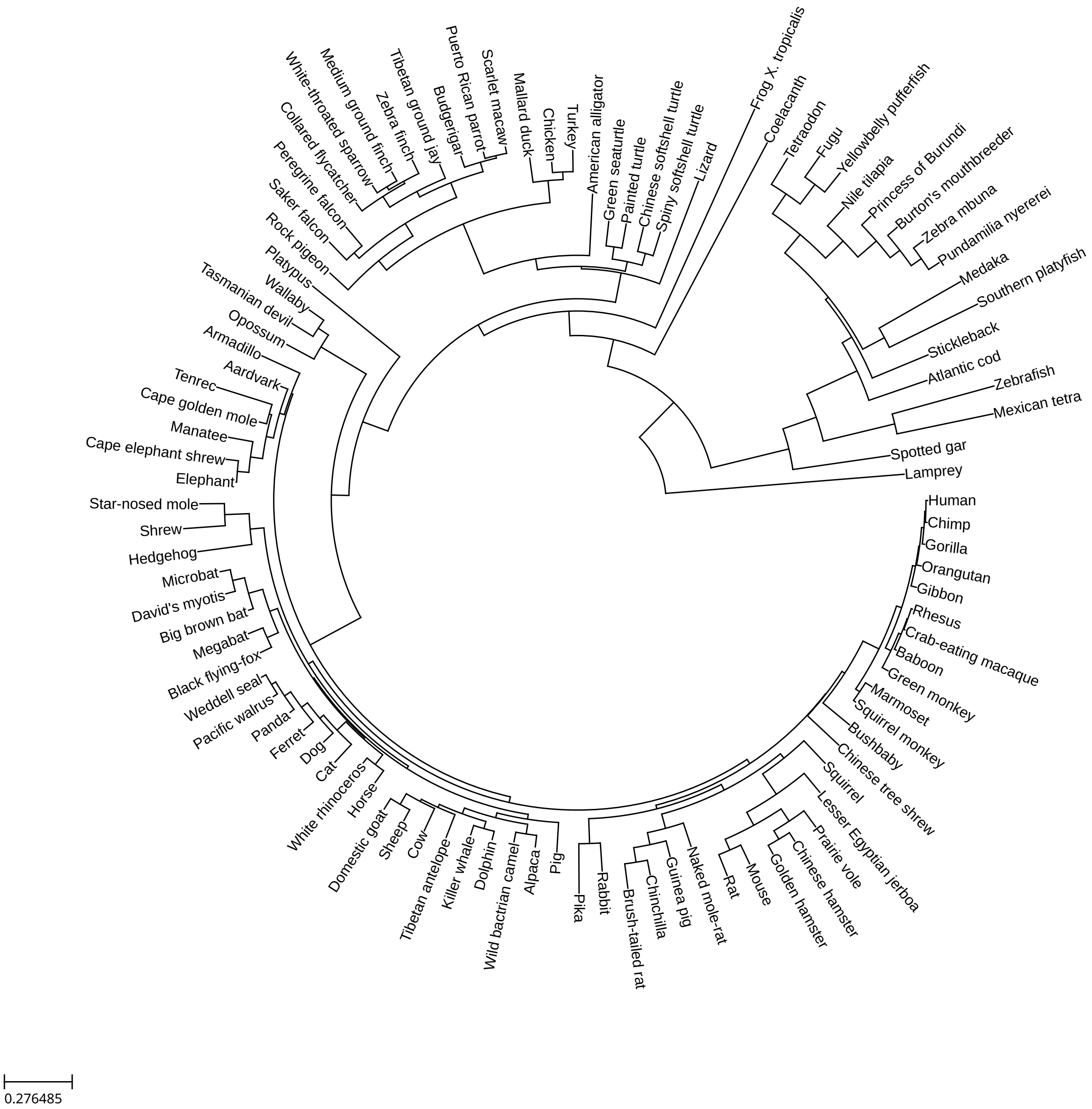

Figure S1: **Phylogenetic tree of 100 vertebrates.**

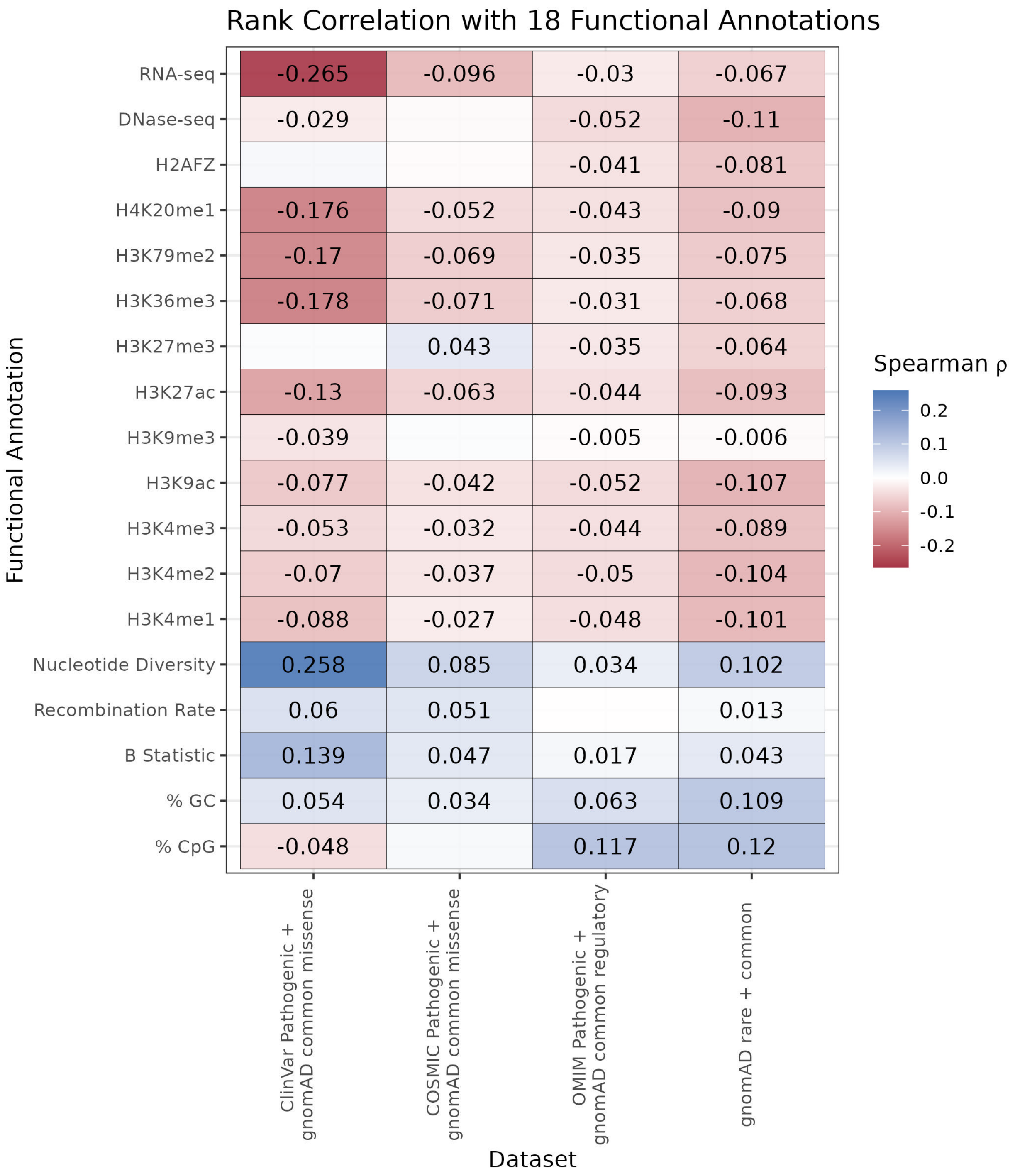

Figure S2: **Functional impact of GPN-MSA.** Rank correlation between GPN-MSA and 18 assay-based or biologically interpretable functional annotations, across four datasets. Only significant (with FWER controlled at 0.05) correlations are shown. For tracks like RNA-seq, which require the variant to be exonic, variants without the annotation are not included in correlation computations.

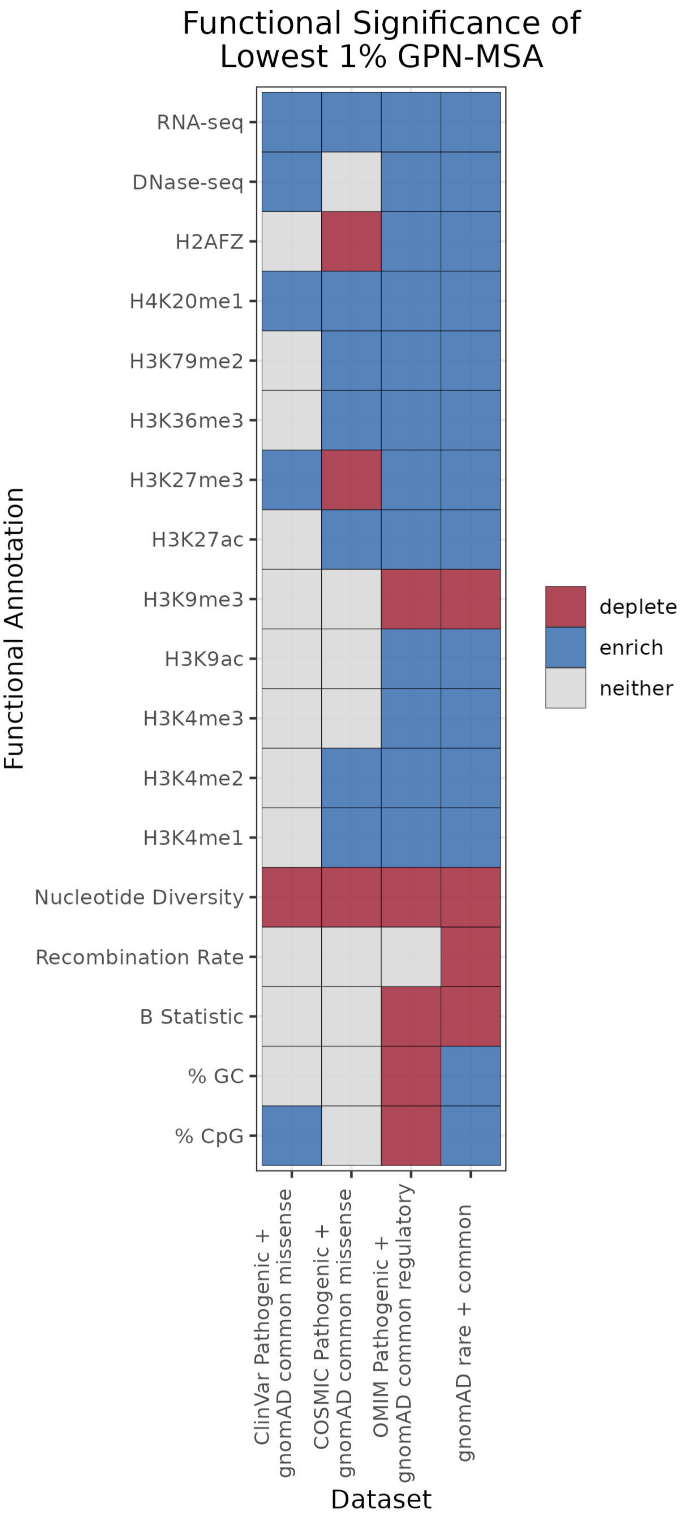

Figure S3: **Functional enrichment and depletion of deleterious tail.** Significant (Wilcoxon-Mann-Whitney test with FWER controlled at 0.05) enrichments and depletions of functional annotations between the deleterious GPN-MSA tail set of variants and background variants, across four datasets. For tracks like RNA-seq, which require the variant to be exonic, variants without the annotation are not included in the two-sample test.

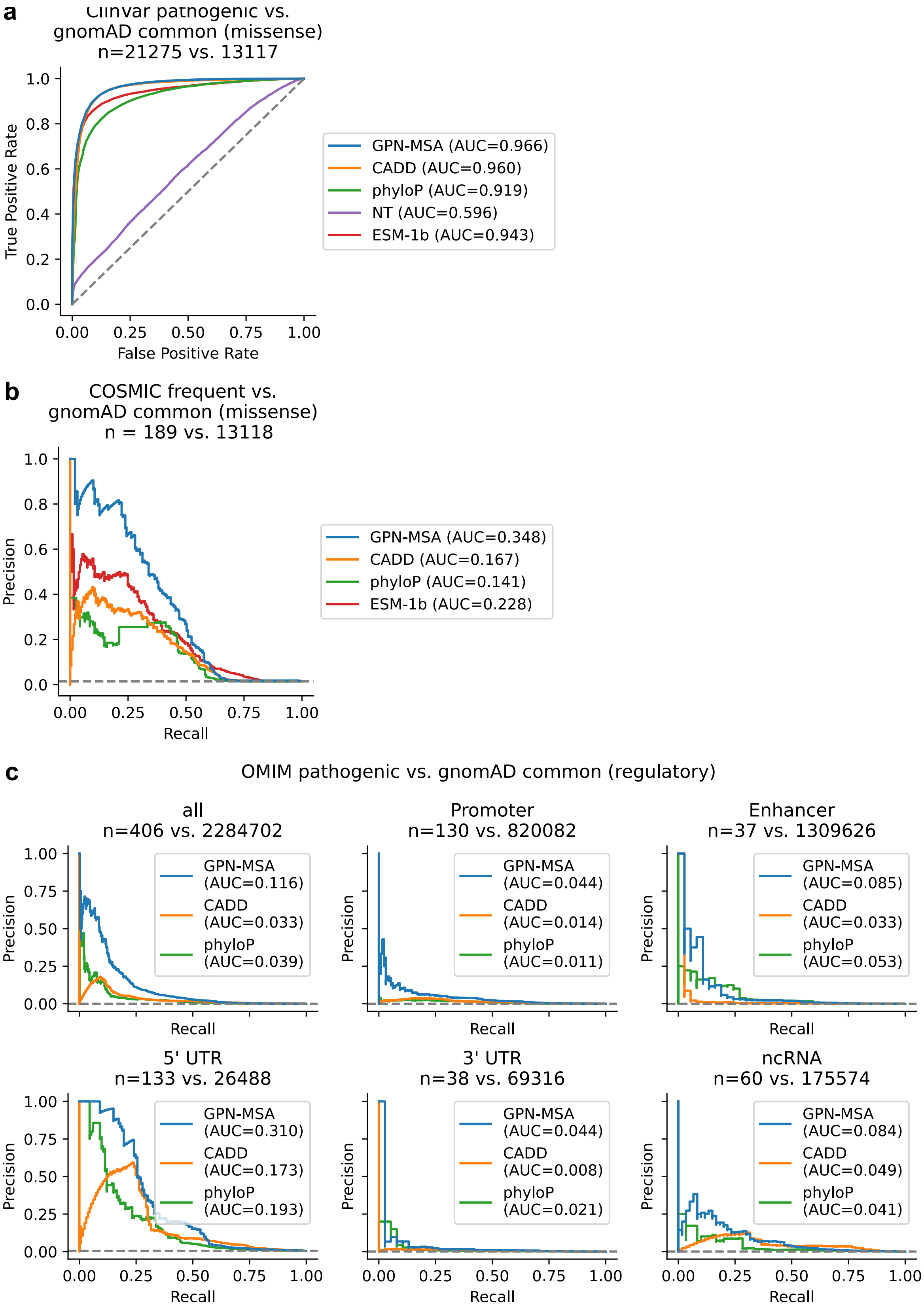

Figure S4: **Receiver Operating Characteristic and Precision-Recall curves for variant effect prediction.** **(a)** Same setting as Figure 2a. **(b)** Same setting as Figure 2b. **(c)** Same setting as Figure 2c.

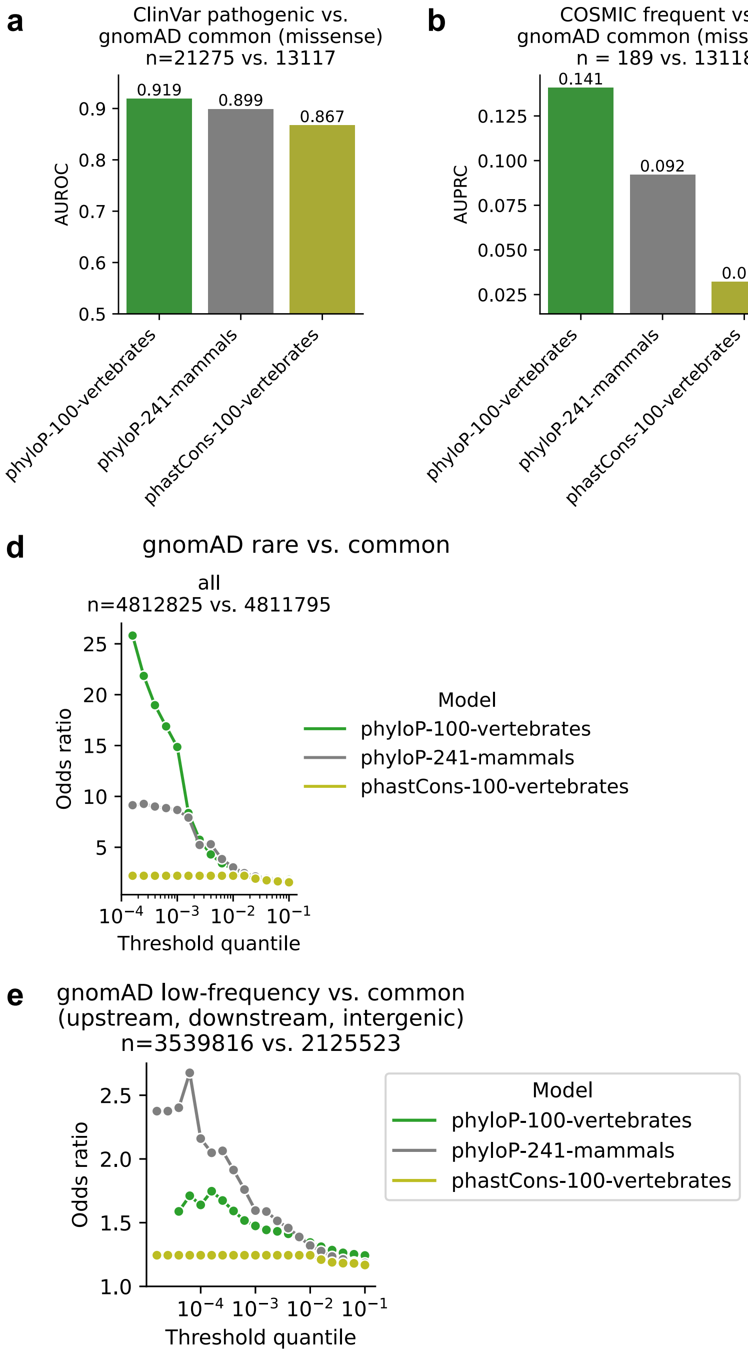

Figure S5: **Variant effect prediction with conservation scores.** **(a)** Same setting as Figure 2a. **(b)** Same setting as Figure 2b. **(c)** Same setting as Figure 2c. **(d)** Same setting as Figure 2d. **(e)** Same setting as Figure 2e.

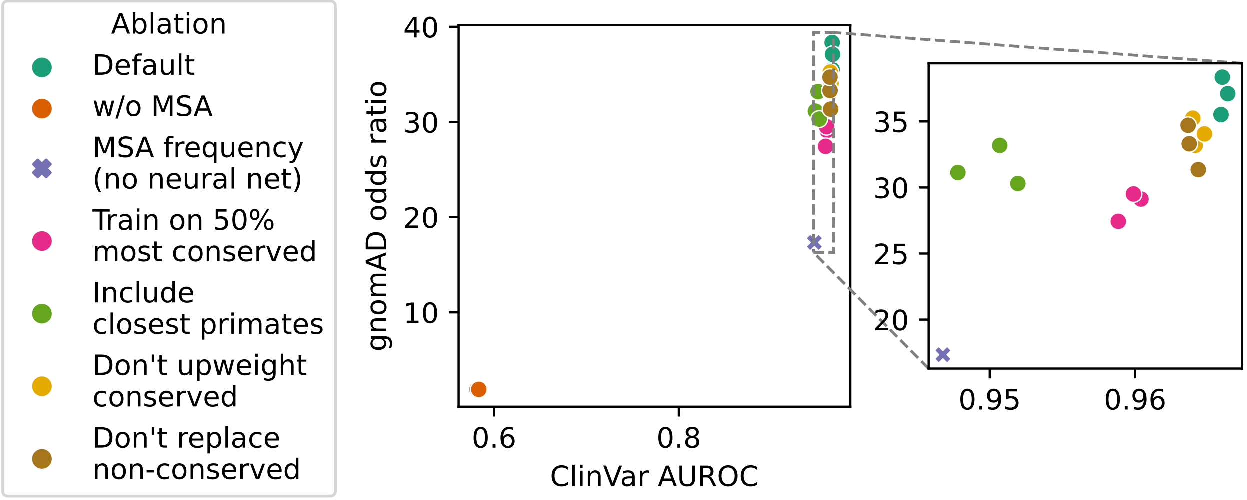

Figure S6: **Ablation study.** Performance of three random seeds of each independent ablation on two variant effect prediction metrics. ClinVar AUROC: same setting as Figure 2a. gnomAD odds ratio: same setting as Figure 2d (threshold quantile = 10^{-3}). Ablations include: training solely on the human sequence (w/o MSA), scoring variants based on MSA column frequencies (MSA frequency), expanding training to include 50% most conserved regions, including nearest primates in MSA, not upweighting conserved elements, and not replacing non-conserved positions when calculating loss. Further details in Methods.

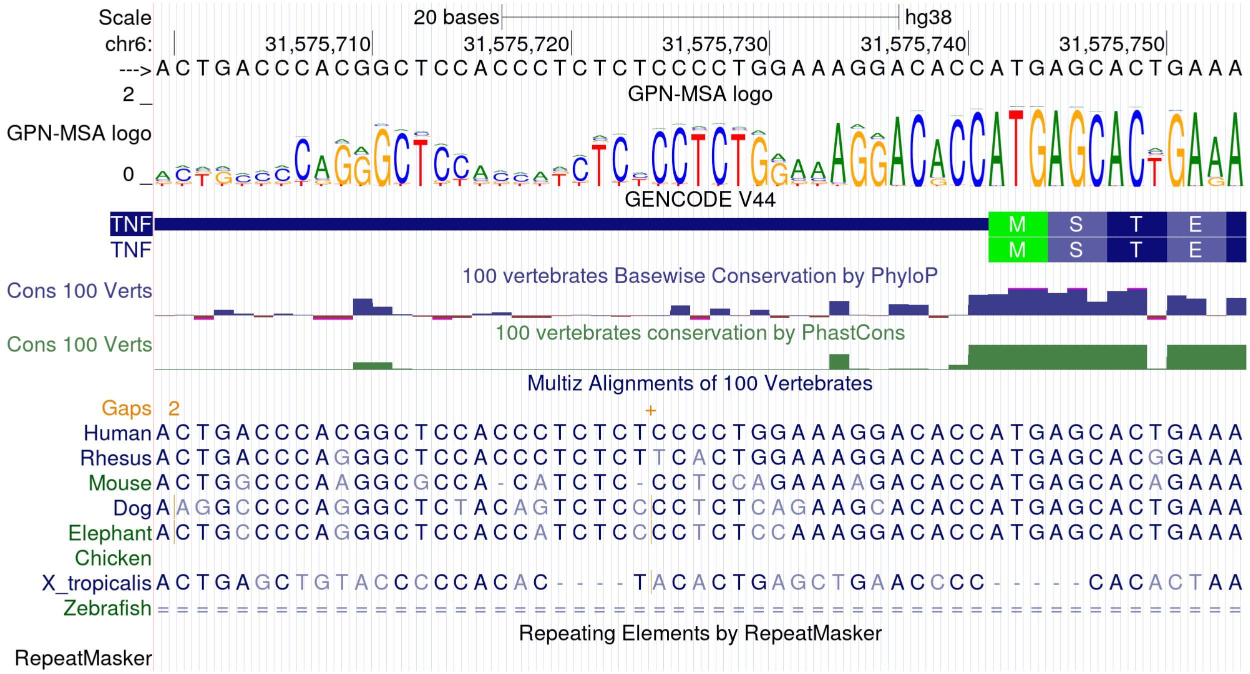
Scale
20 bases
hg38
chr6:
31,575,710
31,575,720
31,575,730
31,575,740
31,575,750
---> ACTGACCCACGGCTCCACCCTCTCTCCCCTGGAAAGGACACCATGAGCACTGAAA
2
GPN-MSA logo
0
GPN-MSA logo
GENCODE V44
TNF
TNF
M S T E
M S T E
Cons 100 Verts
100 vertebrates Basewise Conservation by PhyloP
Cons 100 Verts
100 vertebrates conservation by PhastCons
Multiz Alignments of 100 Vertebrates
Gaps 2
Human ACTGACCCACGGCTCCACCCTCTCTCCCCTGGAAAGGACACCATGAGCACTGAAA
Rhesus ACTGACCCAGGGCTCCACCCTCTCTTCACTGGAAAGGACACCATGAGCACGGAAA
Mouse ACTGGCCCAAGGCGCCA-CATCTC-CCTCCAGAAAAGACACCATGAGCACAGAAA
Dog AAGGCCCCAGGGCTCTACAGTCTCCCCTCTCAGAAGCACACCATGAGCACTGAAA
Elephant ACTGCCCCAGGGCTCCACCATCTCCCCTCTCCAAAGGACACCATGAGCACTGAAA
Chicken
X_tropicalis ACTGAGCTGTACCCCCACAC----TACACTGAGCTGAACCCC-----CACACTAA
Zebrafish ===
Repeating Elements by RepeatMasker
RepeatMasker

Figure S7: **GPN-MSA logo track on the UCSC Genome Browser.** Shown region: `chr6:31,575,700-31,575,754`.

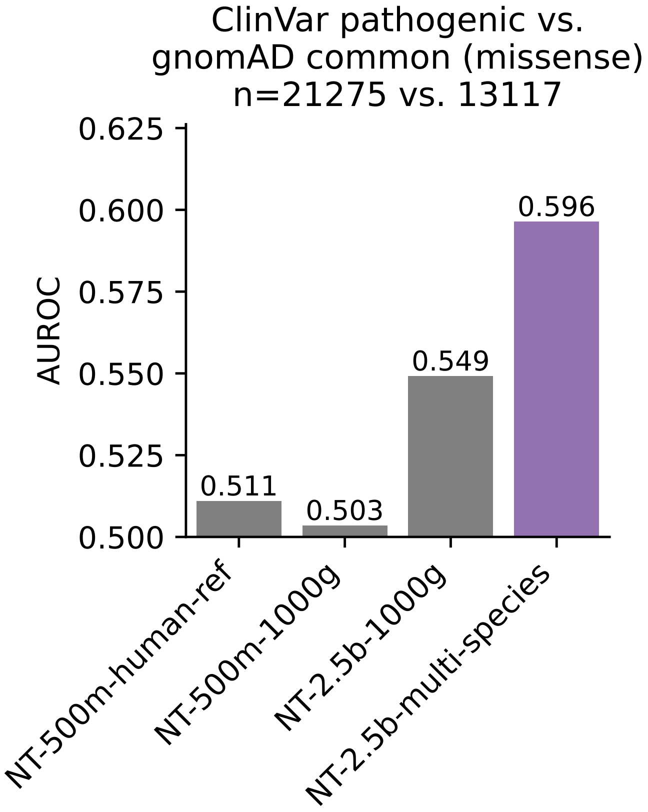

Figure S8: **Variant effect prediction with Nucleotide Transformer models.** Same setting as Figure 2a.

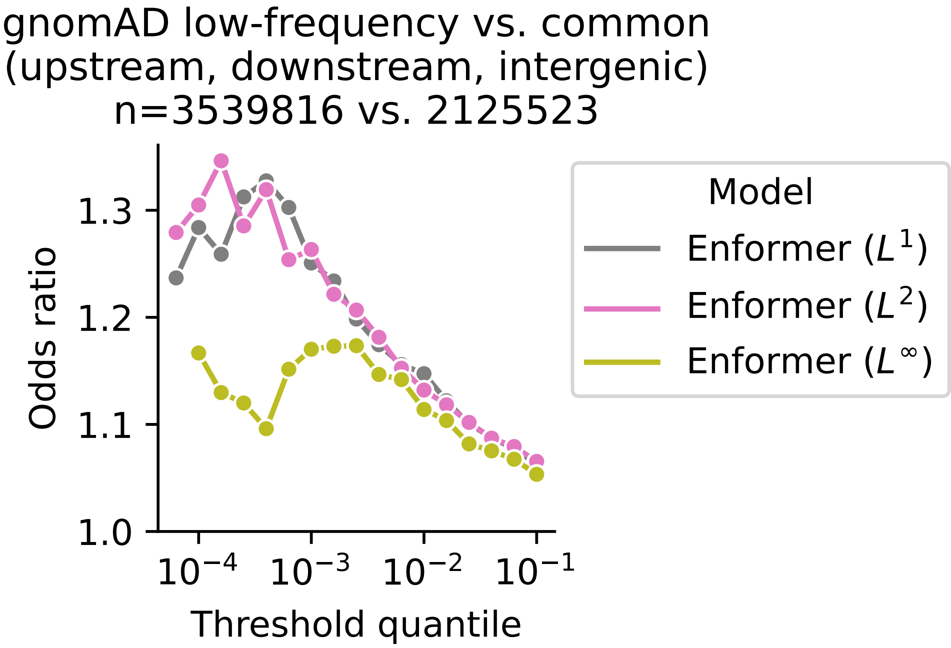

Figure S9: **Variant effect prediction with different norms of Enformer delta predictions.** Same setting as Figure 2e.