The Nucleotide Transformer: Building and Evaluating Robust Foundation Models for Human Genomics

Hugo Dalla-Torre[1], Liam Gonzalez[1], Javier Mendoza Revilla[1], Nicolas Lopez Carranza[1], Adam Henryk Grzywaczewski[2], Francesco Oteri[1], Christian Dallago[2 3], Evan Trop[1], Hassan Sirelkhatim[2], Guillaume Richard[1], Marcin Skwark[1], Karim Beguir[1], Marie Lopez*† [1], Thomas Pierrot*† [1]

[1]**InstaDeep** [2]**Nvidia** [3]**TUM**

Abstract

Closing the gap between measurable genetic information and observable traits is a longstanding challenge in genomics. Yet, the prediction of molecular phenotypes from DNA sequences alone remains limited and inaccurate, often driven by the scarcity of annotated data and the inability to transfer learnings between prediction tasks. Here, we present an extensive study of foundation models pre-trained on DNA sequences, named the Nucleotide Transformer, integrating information from 3,202 diverse human genomes, as well as 850 genomes from a wide range of species, including model and non-model organisms. These transformer models yield transferable, context-specific representations of nucleotide sequences, which allow for accurate molecular phenotype prediction even in low-data settings. We show that the representations alone match or outperform specialized methods on 11 of 18 prediction tasks, and up to 15 after fine-tuning. Despite no supervision, the transformer models learnt to focus attention on key genomic elements, including those that regulate gene expression, such as enhancers. Lastly, we demonstrate that utilizing model representations alone can improve the prioritization of functional genetic variants. The training and application of foundational models in genomics explored in this study provide a widely applicable stepping stone to bridge the gap of accurate molecular phenotype prediction from DNA sequence alone.

Introduction

Foundation models in artificial intelligence (AI) refer to large models incorporating millions of parameters and trained on vast amounts of data, which can be adapted for an array of predictive purposes. These models have deeply transformed the AI field with notable examples in natural language including the so-called language models (LMs) BERT [1] and GPT-3 [2]. LMs have gained considerable popularity over the past years, owing to their capacity to be trained on unlabeled data to create general-purpose representations that can solve downstream tasks. One way they achieve a general understanding of language is by solving billions of cloze tests in which, given a sentence with some blanked-out words, they are rewarded by suggesting the correct word to fill the gap. This approach is referred to as masked language modelling [1]. Early examples of foundation models leveraging this objective in biology were trained on protein sequences by tasking LMs to uncover masked amino acids in large protein sequence datasets [3, 4, 5]. Trained protein LMs applied to downstream tasks, in what is called transfer learning, showed an aptitude to compete and even surpass previous methods for protein structure [3, 4] and function predictions [6, 7], even in data scarce regiments [8].

Beyond protein sequences, the dependency patterns embedded in DNA sequences are fundamental to the understanding of genomic processes, from the characterization of regulatory regions to the impact of individual variants in their haplotypic context. Along this line of work, specialized deep learning (DL) models were trained to identify meaningful patterns of DNA, for example, to predict gene expression from DNA sequences alone [9, 10, 11, 12, 13], with recent work combining convolutional neural

*Equal Supervision
†Corresponding authors: t.pierrot@instadeep.com & m.lopez@instadeep.com

networks (CNN) with transformer architectures enabling the encoding of regulatory elements as far as 100 kilobases (kb) upstream [14]. Modern genomics research has led to a tremendous increase in the volume of sequencing data available, from repeat measurements of single organisms (e.g., to study population variations [15]), to assays measuring multiplexed samples (e.g., to measure the biodiversity of earth's and human gut microbiomes [16, 17]). This deluge of data poses both an opportunity and a challenge; on the one hand, abundant genomic data able to uncover the intricate patterns of natural variability across species and populations is vastly available; on the other hand, powerful deep learning methods, that can operate at large scale, are required to perform accurate signal extraction from this large amount of unlabelled genomic data. Large foundation models trained on sequences of nucleotides appear to be the natural choice to seize the opportunity in genomics research, with attempts to explore different angles recently proposed [18, 19, 19, 20].

With the Nucleotide Transformer we crucially add to recent attempts to encode genomics through foundation models. We built four distinct foundation language models of different sizes, ranging from 500M up to 2.5B parameters, and pre-trained them on three different datasets encompassing the Human reference genome, a collection of 3,200 diverse human genomes, and 850 genomes from several species (Table 6). Once trained, we leveraged representations, or embeddings, of each of these models to further train downstream models on 18 genomics tasks. We then explored the models' attention maps, perplexities and performed data dimensionality reduction on embeddings to decipher the sequence features learned during pre-training. Finally, we probed the embeddings' ability to model the impact of functionally important genetic variants in humans. As a result, we demonstrate the benefits of building and pre-training foundational language models in genomics, across different genomic datasets and parameter sizes, with the ability to surpass specialized solutions.

Results

The Nucleotide Transformer models outperformed or matched 15 of 18 genomic baselines after fine-tuning

We developed a collection of transformer DNA language models, dubbed Nucleotide Transformer, able to learn general nucleotide sequence representations from unlabeled genomic data. After pre-training, model representations were used to solve domain-specific tasks with simple models based on a small number of parameters, such as logistic regression, across 18 different genomic prediction tasks (Fig. 1a) and for which baseline performance metrics were available [21, 22, 23, 24, 25] (Fig. 1 b). Assembling a set of standardized downstream tasks is paramount to evaluating and comparing language models after the self-supervised training step. Motivated by the lack of such benchmark in genomics, we first compiled a collection of 18 datasets, based on five peer-reviewed research studies, into a common format to facilitate experimentation and increase reproducibility.

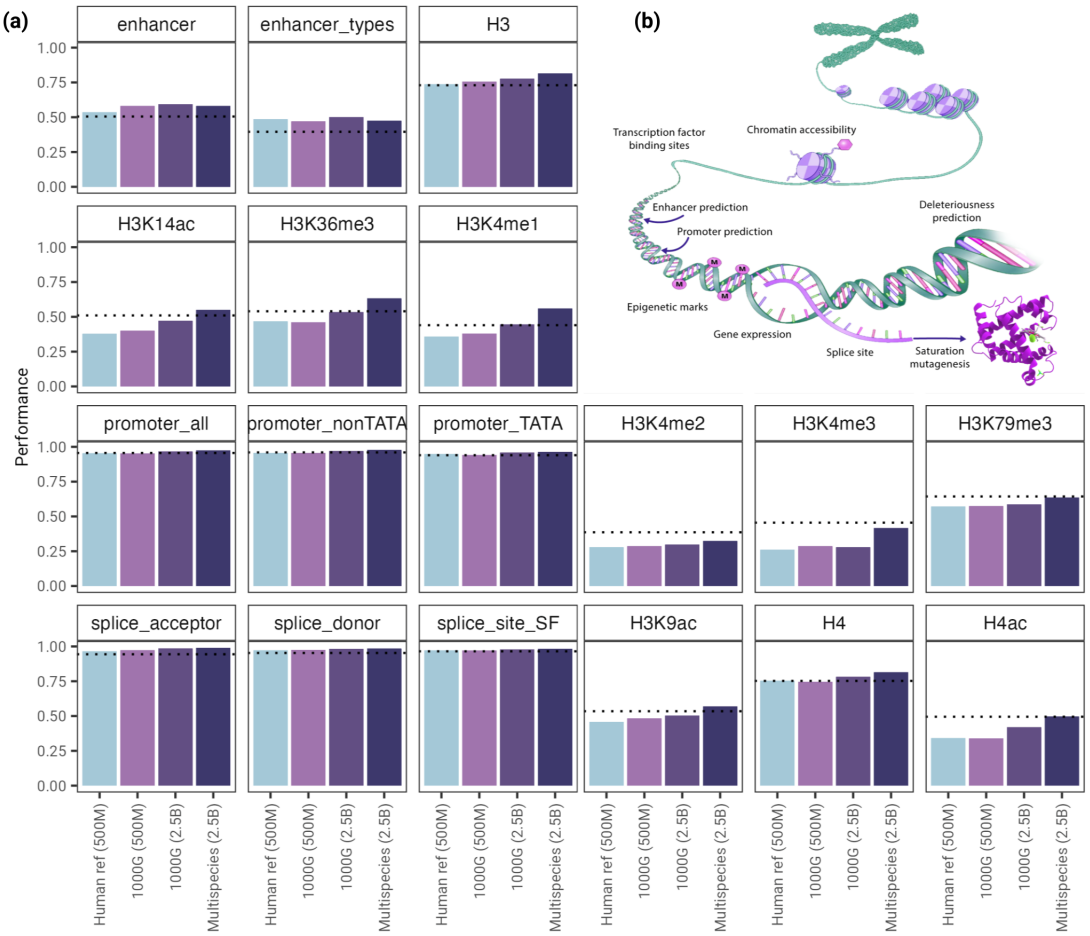

Figure 1: **2.5B Multispecies Nucleotide Transformer model beats 15 out of 18 downstream tasks using fine-tuning. a)** The downstream tasks evaluated in this study visually contextualized to the genome. **b)** Performance results on test sets across all downstream tasks for each Nucleotide Transformer model (from smallest 500M human reference genome, to largest 2.5B multi-species). Results refer to fine-tuning (probing results in Suppl. Fig. 2). To unify different metric choices reported in baselines, the term *performance* stands in for MCC, F1-score, or accuracy in order of preference and availability (details in Table 5).

The Nucleotide Transformer models were evaluated after self-supervised training through two different techniques: probing and fine-tuning. Probing refers to the use of learned LM embeddings of DNA sequences as features to simpler models for predicting genomic tasks. Specifically, we probed ten arbitrarily chosen layers of the LMs using either a logistic regression model or a small multi-layer perceptron (MLP) model composed of up to two hidden layers (Fig. 2). In agreement with recent work [26], we observe that the best performance for a given task is both model and layer dependent (Table 9). Moreover, we confirm that the best model performance is never obtained using embeddings from the final layer, as usually done in earlier work [5]. For example, in the H3K4me1 histone occupancy classification task, we observe a relative difference between the highest and lowest performing layer as high as 38%, indicating that the representation quality varies significantly across the layers of the transformer model (Suppl. Fig. 1).

Using probing, our models outperformed or matched 11 out of 18 baselines (Table 4). This number increases to 15 out of 18 after fine-tuning (Fig. 1). Even though fine-tuning was not extensively studied in previous work [5], possibly due to its high compute requirements, we overcame this limitation by relying on a recent parameter efficient fine-tuning technique [27] requiring only 0.1% of the total model parameters to be tuned. This technique allowed for faster fine-tuning on a single GPU, in addition to reducing the storage needs by 1000-fold over all fine-tuning parameters, and increasing performance. Coincidentally, while using simple downstream models on embeddings might appear faster and less compute-intensive, in practice we observed that rigorous probing was more compute-intensive compared to fine-tuning, as the choice of the layer, downstream model, and hyperparameters strongly impacted the performance. Moreover, a smaller variance in performance was observed when fine-tuning.

As an example, when classifying weak and strong enhancers using sequences alone, the combination of the best LM and downstream model reached performance (metrics in Table 5) of 0.424 before, and 0.501 after fine-tuning, thus outperforming the previous state-of-the-art baseline of 0.395 using LSTM-CNN [22] (Table 4). On promoter detection, we report performance of 0.966 and 0.974 before and after fine-tuning respectively, compared to 0.956 for the baseline [24]. For splice site detection, performance increased from 0.762 to 0.983 when fine-tuning compared to the baseline performance of 0.965 achieved by SpliceFinder [23] (Table 4). This task is also the most affected by the downstream model type, with performance increasing from 0.590 to 0.762 when considering a logistic regression as opposed to an MLP.

Parameter scaling and increasing data diversity improve performance

Motivated by trends in natural language processing (NLP), where increasing the volume and diversity of the datasets, as well as the size of the transformer model yields better results [28], we trained large models from 500M up to 2.5B parameters on datasets of increasing size. We first constructed a pre-training dataset with sequences from the Human reference genome to establish our baseline LM (see Methods). We prepared additional training sets with sequences from 3,200 genetically diverse human genomes [15], spanning 27 populations across the world (Table 2), which increased the data size but also included 98% redundancy amongst samples. Finally, we developed a custom multispecies dataset composed of reference sequences of 850 species (Fig. 2a; Table 6; see Methods). We observed that scaling the model size and adding diversity to the dataset sequences increases the average performances of the downstream tasks (Fig. 2c).

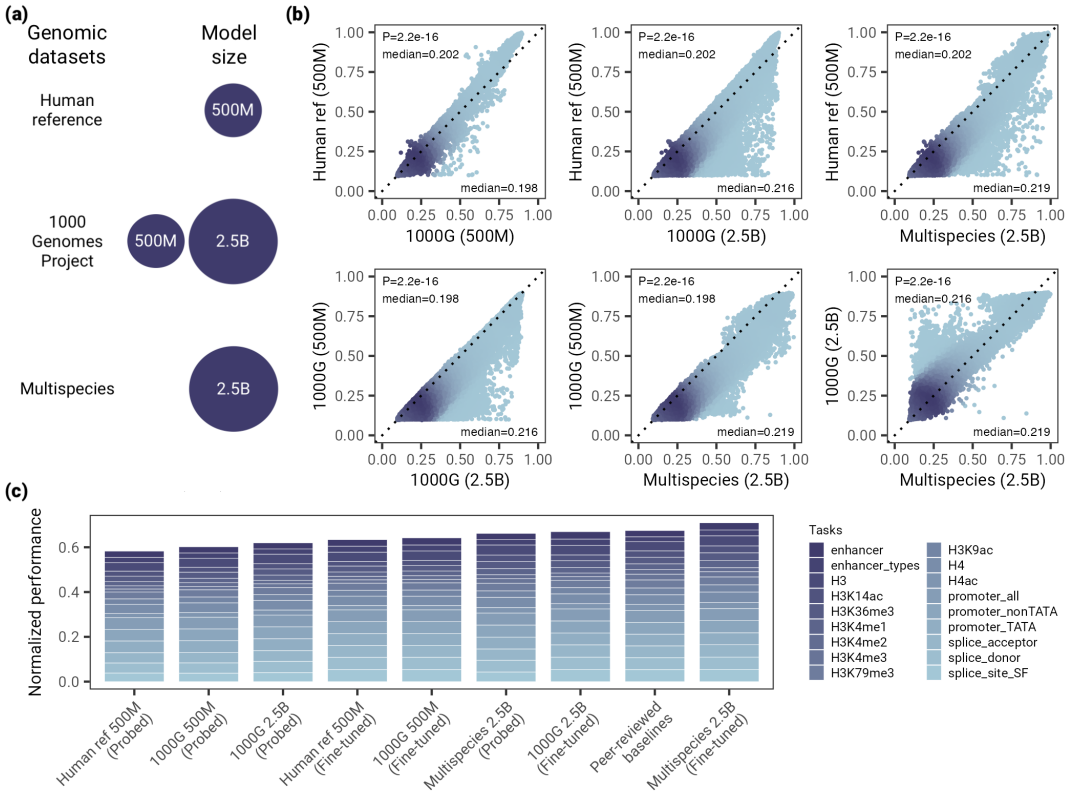

Figure 2: **Increasing model size and data diversity improves downstream performance and increases token reconstruction accuracy. a)** Visual representation of models parameter sizes across datasets (larger circles indicate larger models). **b)** Pairwise comparison of token reconstruction accuracy across models. P-values refer to a two-sided Wilcoxon signed rank test between models. Median values for the two models compared are shown. **c)** Normalized sum of performance across 18 downstream tasks for all Nucleotide Transformer models after probing and fine-tuning. Normalized sum is obtained by summing performances over tasks for each model and dividing by the number of tasks.

As a benchmark, the 500M parameter model trained on the Human reference genome achieved a compound performance (in absolute best values) of 0.634 across tasks, whereas a model with the same number of parameters trained on the 1000G dataset achieved a performance of 0.641. When increasing the size of the model from 500M to 2.5B parameters on the 1000G dataset, compound performance increased to 0.669. While adding intra-species diversity slightly helped the overall performance on the designated tasks, adding inter-species diversity at a fixed model scale yielded better results. Indeed, the 2.5B model on the multispecies dataset yields a compound performance of 0.709. Only this combination of largest model size (2.5B), inter-species diversity (multispecies dataset) and fine-tuning resulted in the compound performance of 0.709 surpassing baselines at 0.674 (Fig. 2c).

To further investigate the benefits of increasing the number of parameters of the models and genomic diversity of the datasets, we assessed the ability of the models to reconstruct masked nucleotides. Specifically, we partitioned the Human reference genome into non-overlapping 6kb sequences, tokenized the sequence into 6-mers, randomly masked a number of tokens, and estimated the proportion of tokens correctly reconstructed (Fig. 2b). We observed that the reconstruction accuracy of the 500M Human reference model showed a higher median accuracy than the 500M 1000G model (median=0.202 versus 0.198, P<2.2e-16, two-sided Wilcoxon rank sum test); however, it was lower than the accuracy obtained by the 2.5B 1000G model (median=0.216) and the 2.5B Multispecies model (median=0.219), which illustrates the impact of jointly increasing the model size and diversity of the dataset. Interestingly, while the 2.5B Multispecies model showed the best reconstruction accuracy overall, a number of sequences showed much higher reconstruction accuracy for the 2.5B 1000G model, which likely points to particular characteristics of human sequences that were better learned by this model.

The Nucleotide Transformer models learnt to detect known genomic elements and human genetic variation

To gain insights into the sequence elements that the nucleotide transformer is utilizing when making predictions, we explored different aspects of the transformer language model architecture. First, we evaluated the ability of the models to reconstruct tokens containing natural variation present in human populations based on their haplotypic background (Fig. 3a). We measure the effectiveness of sequence reconstruction by recording the model perplexity scores over single nucleotide polymorphisms (SNPs) occurring at 1%, 5%, 10% and 50% frequencies. To evaluate the impact of genomic background and genetic structure on the mutation reconstruction, we considered an independent dataset of genetically diverse human genomes, originating from 7 different meta-populations [29] (see Methods). SNP tokens are centered in sample-specific 6kb sequences. The perplexity is computed at the token position after masking it. A lower perplexity value is indicative of more accurate sequence reconstruction. Across all populations and SNPs frequencies, we consistently observe that the 2.5B parameter models, with median perplexity ranging from 48.4±4.1 (2SD) to 95.4±7.2, outperform their 500M counterparts, for which the perplexity values fluctuate between 69.1±5.2 and 118.2±8.2. These results are in line with the observed reconstruction accuracies over human genome sequences, and confirm the impact of model size on performance.

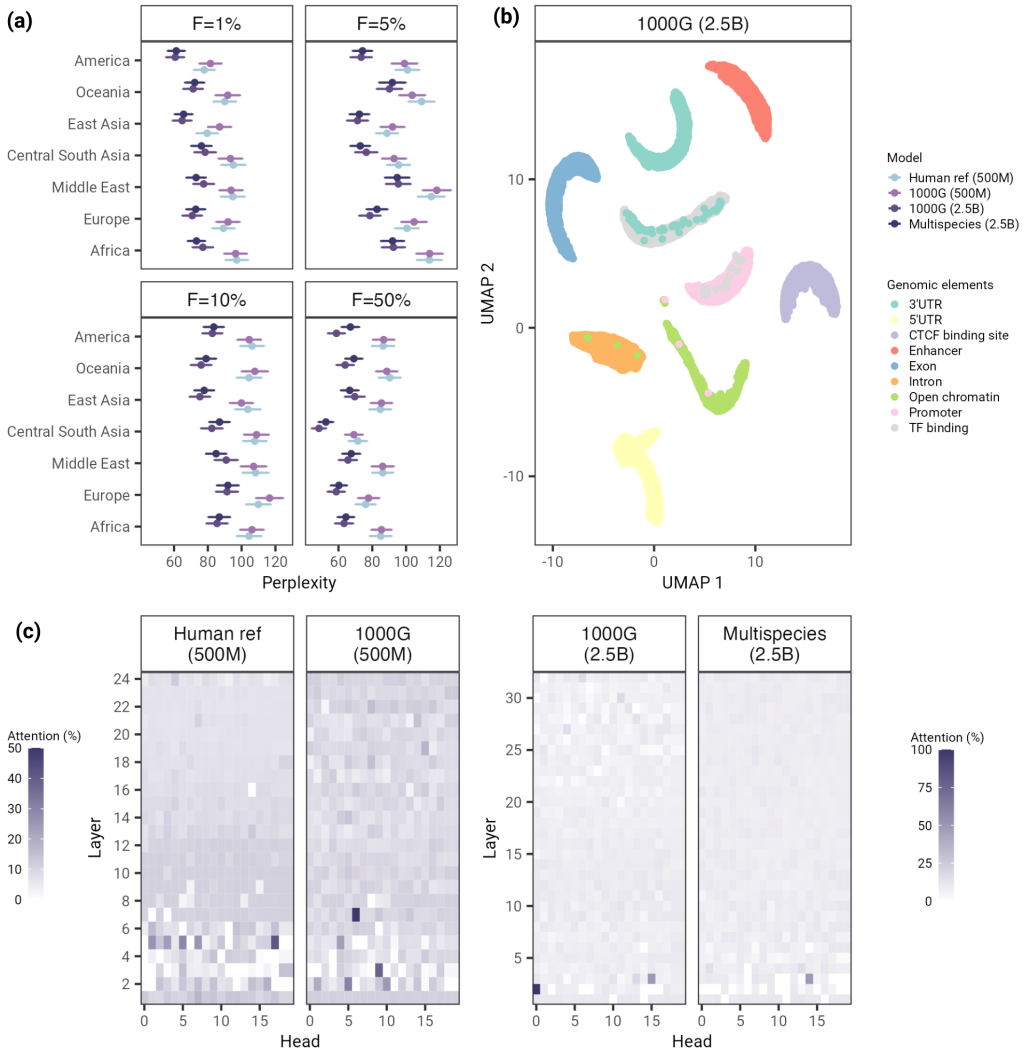

Figure 3: **The Nucleotide Transformer models acquired knowledge about genomic elements and human genetic variation. a)** Reconstruction perplexity of the Nucleotide Transformer models on SNPs at 1%, 5%, 10% and 50% frequency across worldwide human populations. **b)** U-MAP projections of embeddings of nine regulatory elements from layer four of the 2.5B model trained on the 1000G dataset. **c)** Attention percentages per head and layer for all Nucleotide Transformer models computed on enhancers.

Next, we sought to determine whether the model embeddings encode representations of the different genomic sequence features. We considered a total of 90,000 sequences annotated by Ensembl [30] as: "5' UTR", "3' UTR", "exon", "intron", "enhancer", "promoter", "CTCF binding site", "open chromatin", and "transcription factor binding sites". For each sequence category, we examined 10,000 sequences of length 6kb, where the element is randomly positioned in the sequence. We computed embeddings at three arbitrary layers for all Nucleotide Transformer models and averaged them only over the genomic element indexes. The resulting data was projected through Uniform Manifold Approximation and Projection (UMAP) to lower dimensions. We report the visualization of the embeddings obtained at one layer over two UMAP components of the 2.5B parameters model trained over the 1000G dataset (Fig. 3b). We observe that the "enhancers", "5' UTR regions", "CTCF binding sites", and "exons"sequence clusters are perfectly separated from the other genomic features. Interestingly, sequences corresponding to "open chromatin"and "introns"features show limited overlap, while "promoters", "TF binding sites"and "3' UTR regions"exhibit substantial co-clustering, probably owing to the versatility of sequence function in regulatory portions of the genome. We observed similar trends for all models, but not in all layers (Suppl. Fig. 13 and Suppl. Fig. 14), confirming the layers' specialization, as well as the models' ability to capture genomic patterns during self-supervised training.

Finally, to understand whether the model encodes for genomic features, we computed the attention percentages for all heads and layers of the Nucleotide Transformer models over the same nine genomic features, following the methodology introduced in previous work [31]. An attention head is thought to recognize some specific elements if its attention percentage is significantly greater than the naturally occurring frequency of the element in the pre-training dataset (Suppl. Fig. 12). We performed two proportions z-test to determine whether any of these elements were detected due to chance across models and heads, and corrected for multiple testing using Bonferroni. Of 480 heads for the 500M models, with a significance level of 0.05, we observed that 27 and 24 heads specialized in the detection of enhancers for the 500M Human reference and 1000G models. In the case of the larger 2.5B models with 640 heads, 28 and 26 detected enhancers for either the 1000G and multispecies datasets. Interestingly, maximum attention percentages over enhancers were higher for the larger models, with values reaching up to almost 100% attention for the 2.5B 1000G model (Fig. 3c). Similar results were observed for all nine genomic sequence annotations (Suppl. Fig. 3 -11), with a maximum number of 117 out of 640 heads detecting introns for the 2.5B Multispecies model. Altogether, these results demonstrate that DNA sequences are adequately recapitulated by language models, both at the scale of large genomic regions and individual variants in diverse populations. Remarkably, custom representations can be obtained for any region of the genome with mutations which could provide substantial improvement in assessing the pathogenicity of mutations in their haplotypic context.

The Nucleotide Transformer embeddings predict the impact of mutations

Finally, we assessed the ability of our transformer models to differentiate between different types of genetic variants, and to prioritise those of functional importance. In order to do so, we tested divergence between alternative allele and reference embeddings for mutations of interest using four metrics (see Methods; dubbed "divergence score").

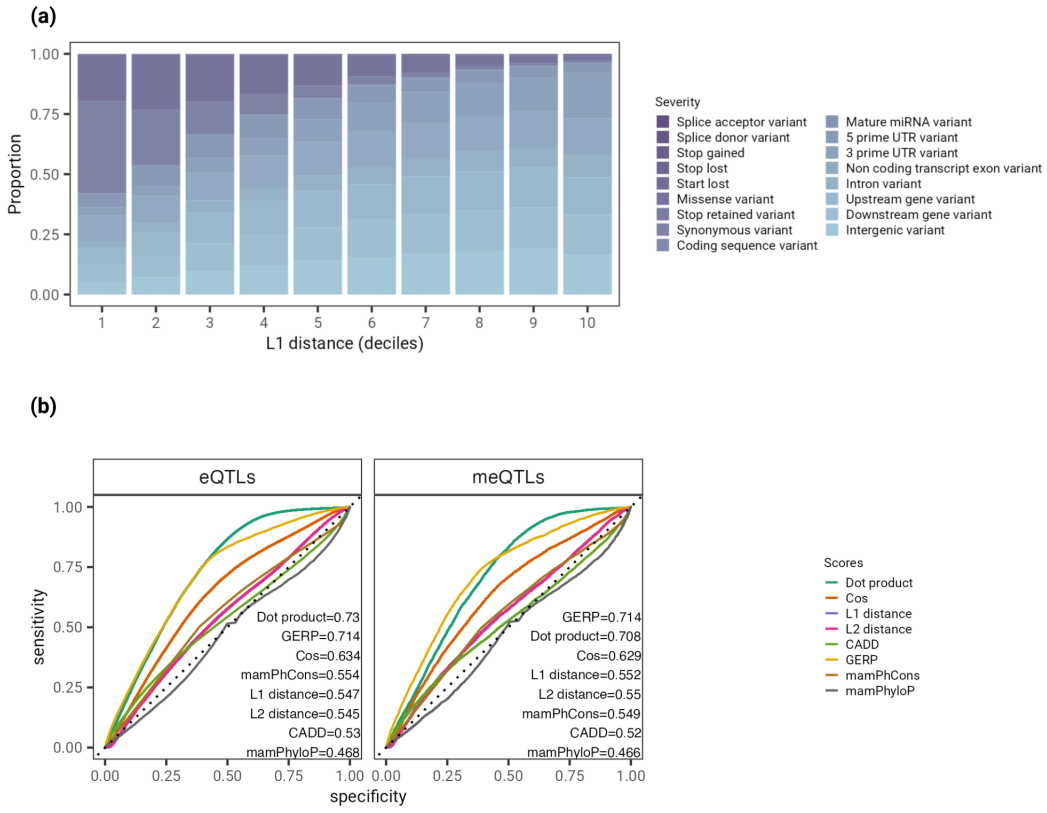

Figure 4: **Prioritizing functional genetic variants based solely on DNA sequence information. a)** Proportion of variant consequence terms across deciles based on the L1 distance metric for the Multispecies (2.5B) model. The consequence terms are shown in order of severity (more severe to less severe) as estimated by Ensembl. **b)** Comparison of different distance metrics for the Multispecies (2.5B) model to other methods for prioritizing functional variants based on GRASP eQTLs and meQTLs, against 1000 Genomes Project common SNPs. Model performance is measured with the area under the receiver operating characteristic curve (AUC).

We first considered all segregating 1000 Genomes Project SNPs from chromosome 22 and investigated the distribution of these embedding-based divergence scores across 17 variant categories (e.g. stop gained, missense, intergenic). For the four models, the scores among the different categories were significantly different (P-value = 2.22e−16, K-S rank sum test). The Multispecies (2.5B) model using the L1 (Manhattan) distance showed the strongest ability to differentiate between categories (Fig. 4a). Notably, scores computed from this model at coding variants, i.e. those with a potential functional impact on protein function, showed significantly lower scores compared to intergenic variants (median of 675 and 622 for missense and synonymous variants, respectively, versus median of 780 for intergenic variants, P-value = 2.22e−16, two-sided Wilcoxon rank sum test). This illustrates how embedding-based distances alone encodes information about variant functional categories. Nonetheless, we also found that other types of variants, including those with high potential disruption on protein, such as stop gain mutations, showed similar scores compared to intergenic variants (P-value >0.05, two-sided Wilcoxon rank sum test), which suggests that this score is not necessarily associated with mutation severity.

Lastly, we investigated the ability of the divergence scores to prioritize functional variants, and compared these against scores that measure levels of genomic conservation, as well as a score that leverages both conservation and genomic functional features (including genic effects, regulatory element annotations, among others). Specifically, we assessed the ability of the divergence scores to classify genetic variants exerting regulatory effects on gene expression (i.e., expression quantitative trait loci [eQTLs]) and genetic variants associated with DNA methylation variation (i.e., methylation quantitative trait loci [meQTLs]) (Fig. 4b). As the scores computed from the Multispecies (2.5B) model showed the best ability to differentiate variant categories, we relied on divergence scores only from this model. Notably, the performance of divergence scores surpassed that of previous scores in prioritizing eQTLs (dot score product, AUC=0.73) and achieved an AUC close to the best-performing score in prioritizing meQTLs (dot score product, AUC=0.708 versus GERP score, AUC=0.714). Overall, these results illustrate how divergence scores, extracted solely from embeddings of DNA sequences, can help reveal and contribute to understanding the potential biological consequences of variants associated with a disease or phenotype.

Discussion

Evolution informs representations: models trained on genomes from different species top charts. We explored the impact of different datasets to pre-train equally sized transformer models. Both intra- (i.e. when training on multiple genomes of a single species) and inter-species (i.e., on genomes across different species) variability play an important factor driving accuracy across tasks (Fig. 2). Notably, the models trained on genomes coming from different species perform well on categorical human genomics downstream tasks, as well as on human variant prediction, even when compared to models trained exclusively on the human genome (Fig. 2). This could indicate that the genome LMs capture a signal of evolution so fundamental across species that it better generalises to shared functions.

Model scale drives performance. The Nucleotide Transformer models trained ranged from 500 million up to 2.5 billion parameters, which is five times larger than DNABert [18] and ten times larger than the Enformer [14]. As has been the case in NLP [28], results in the genomic space confirmed that increasing model size yields better performance. Training the largest parameter model required a total of 128 GPUs across 16 compute nodes for 28 days. Significant investments were made to engineer efficient training routines that took full advantage of the infrastructure, highlighting the need for both specialized infrastructure and dedicated software solutions. Once trained however, these models can be used for inference at a relatively low cost, benefiting the research community as a whole.

Diverse datasets, large models and fine-tuning: best results require it all. Previous work in biological LM development evaluated downstream performance exclusively probing the last transformer layer [5], most likely due to its perceived ease of use, relative good performance and low computational complexity. As in this study we sought to push the limits of downstream accuracy, we performed computationally expensive, rigorous probing of different transformer layers, downstream models and