Pre-training Genomic Language Model with Variants for Better Modeling Functional Genomics

Tianyu Liu[1,2], Xiangyu Zhang[2], Rex Ying[3], Hongyu Zhao[1,2*]

[1]Interdepartmental Program in Computational Biology & Bioinformatics, Yale University, New Haven, 06511, CT, USA.
[2]Department of Biostatistics, Yale University, New Haven, 06511, CT, USA.
[3]Department of Computer Science, Yale University, New Haven, 06511, CT, USA.

*Corresponding author(s). E-mail(s): hongyu.zhao@yale.edu;
Contributing authors: tianyu.liu@yale.edu; xiangyu.zhang@yale.edu; rex.ying@yale.edu;

Abstract

Sequence-to-function models can predict gene expression from sequence data and be used to link genetic information with transcriptomics data to understand regulatory processes and their effects on complex phenotypes. The genomic language models are pre-trained with large-scale DNA sequences and can generate robust representations of these sequences by learning the genomic context. However, few studies can estimate the predictability of gene expression levels and bridge these two classes of models together to explore individualized gene expression prediction. In this manuscript, we propose UKBioBERT as a DNA language model pre-trained with genetic variants from UK BioBank. We demonstrate that UKBioBERT generates informative embeddings capable of identifying gene functions, and improving gene expression prediction in cell lines, thereby enhancing our understanding of gene expression predictability. Building upon these embeddings, we combine UKBioBERT with state-of-the-art sequence-to-function architectures, Enformer and Borzoi, to create UKBioFormer and UKBioZoi. These models exhibit better performance in predicting highly predictable gene expression levels and can be generalized across different cohorts. Furthermore, UKBioFormer effectively captures the relationship between genetic variants and expression variations, enabling in-silico mutation analyses. Collectively, our findings underscore the value of integrating genomic language models and sequence-to-function approaches for advancing functional genomics research.

Keywords: DNA Sequence Model, Large Language Model, Sequence-to-Function Models, Functional Genomics, Gene Expression

1 Introduction

Deciphering DNA as a language is a challenging but essential task to understand the complexity in genetics and genomics [1–8]. Previous research has demonstrated that the functional elements in genomes, such as cis-regulatory elements (CREs) and trans-regulatory elements, may be affected by many factors including distance to targeted genes as well as the context of their nearby genomics regions [9, 10]. These properties suggest the existence of polysemy and distant semantic relationships in DNA sequence, which are also found in natural language [11]. Therefore, modeling DNA as a language of life can also be based on the advanced techniques in Natural Language Processing (NLP) [12], and many foundation models based on the architecture of Transformer [13] have been proposed to model DNA sequences, and led to better understanding of functional genomics in different species [14] or cell lines [15].

There are two major types of genomic language models (gLMs), either learning the context of DNA sequences by masked language modeling (MLM) or learning the causal relationship in nucleotide sequences by causal language modeling (CLM). The representative base model for the first class is the bidirectional transformer (BERT) [16]. For example, DNABERT [11], DNABERT-2 [17], and DNABERT-S [14] utilize BERT as the base model and pre-train it with DNA sequences from the reference genome [18]. Furthermore, GENA-LM [19] pre-trains the base BERT model with both reference genome and single nucleotide polymorphisms (SNPs) from the 1000 Genomes Project [20], and provides models with different scales for users. Training the model based on BERT can enhance our understanding of the context information of one DNA sequence and thus it will generate better presentations for applications, in principle. The representative base model for the second class is a generative pre-trained transformer (GPT) [21]. As an example, DNAGPT [22] utilizes GPT-based models for pre-training and it can generate new DNA sequences. HyenaDNA [2] and Evo [23] are based on the Hyena convolution network [24], but their pre-training strategies are CLM, so they are also focusing on generating new DNA sequences or other large molecules (e.g., RNA or amino acid sequences).

Despite the exciting progress made by gLMs, there are concerns about applying these foundation models in genomic research. First, CLM-based DNA language models suffer from hallucination when generating new DNA sequences [25], and are not good at modeling the genomic information with the consideration of context, as shown in the recent benchmarking analysis of DNA sequence embeddings from various models [26]. Second, although MLM-based DNA language models are better candidates for learning the context information as well as encoding sequence information into the embedding space, the reliability of generated embeddings is affected by multiple factors, including the quality of pre-training datasets, the choice of pre-training task, the evaluation metrics, and the context window size, can all affect the model performance [26, 27].

Moreover, gLMs such as GPN [28] and GENA-LM consider including SNPs [29] to enhance model performance based on the principle of data augmentation, but there is no demonstration of the necessity of using variants versus using reference genome with the random augmentation approach [30]. Furthermore, the downstream applications considered by these methods are still based on reference genome sequences, and thus the pre-training data have a weak link with real application scenarios. These settings may limit the contribution of introducing human variants in model pre-training, although it is well-known that many genetic variants play an important role in genetic regulation, disease onset, and progression [31, 32].

To better model DNA as a language as well as address the limitations of the existing methods, we propose UKBioBERT, a foundation model pre-trained with both reference genome and variant information from UK BioBank [33]. We gather variants from approximately 300,000 UK Biobank participants with European ancestry and select the best approach to leveraging the advantages of pre-trained weights from other gLMs to formalize our models efficiently. We propose a new metric to evaluate the quality of gLMs by paying attention to the ability of a model to distinguish sequences with different functions. Furthermore, we investigate the contribution of genetic variants by considering downstream applications including cell-type-specific gene expression prediction and individualized gene expression prediction, which are more related to sequence-to-function analysis. The representations of UKBioBERT improved over baseline models, and explain the difficulties of expression prediction across individuals and cohorts. We also demonstrate that gene expression prediction at the individual level can be further improved by fine-tuning Enformer [30] with individual genome information and embeddings from UKBioBERT, denoted as UKBioFormer. Finally, we analyze the interpretability of the embeddings from UKBioBERT and identify both common and rare variant effects validated by expression quantitative trait loci (eQTL) information based on UKBioFormer. Overall, UKBioBERT and UKBioFormer offer useful tools for researchers to explore the power of variant modeling with sequences from different population-level datasets.

2 Results

Method overview. The development of UKBioBERT is inspired by the training process of GPN [28] and Gena-LM [19], as well as data augmentation from deep learning [34]. We pre-trained UKBioBERT based on the checkpoint from DNABERT2 [17] and genomic information with variants from UK BioBank with a self-supervised learning framework. Recent benchmarking analyses demonstrated that DNABERT2 performed well for several gene-centric tasks, such as gene finding, variant effect prediction, and gene expression prediction [35–37]. These observations suggest the usefulness of modeling genomic language with the awareness of frequency and context differences contributed by variants. The embeddings from UKBioBERT can help state-of-the-art (SOTA) method EPInformer [15] for better-predicting gene expression captured by different techniques at the cell-type level. These contributions are summarized in Figure 1 (a). Moreover, the embeddings generated from individual genomic information can also explain the predictability of gene expression levels from a novel angle.

Furthermore, we selected the optimal models from a collection of UKBioBERT and combined the embeddings with Enformer to fine-tune the fused model UKBioFormer for gene expression prediction and eQTL identification at the individual level. These contributions are summarized in Figure 1 (b). Details of our selected tools are further explained in the Methods section.

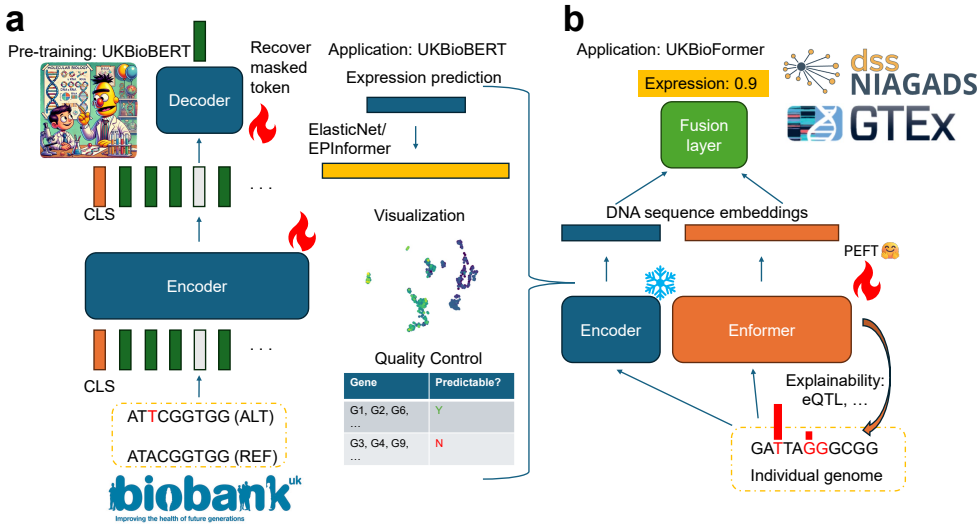

Fig. 1 The landscape of UKBioBERT and UKBioFormer. The fire logo represents the adjustment of model parameters, and the ice logo represents the freeze of model parameters. The logos of UK BioBank, GTEx, and ROSMAP come from the official websites. (a) The pre-training stage and related downstream applications of UKBioBERT. We utilize masked language modeling to pre-train UKBioBERT. The cartoon logo is generated by GPT-4o [38]. (b) The fine-tuning stage and related downstream applications of UKBioFormer. We fuse Enformer and UKBioBERT to enhance gene expression prediction.

Pre-trained UKBioBERT learns better gene representations. Previous genomic language models evaluated their pre-training stage based on the validation loss and further investigated the expressive power of their pre-trained model based on downstream applications of the DNA sequence embeddings, such as enhancer identification and species classification [17, 35]. However, these tasks did not consider an important problem in the pre-training stage of foundation models, known as model collapse or context collapse. Model collapse is defined as that the pre-trained model cannot distinguish between data that differ in input features. Furthermore, this problem is also reflected in the embedding space of low-performance models, that is, the embeddings are similar for DNA sequences containing different contexts. Since the identity of a given DNA sequence can be varied, and we mainly consider gene expression levels, we focus on the function of genes as the main labels. The function of genes is known to be highly associated with expression levels, and thus gene function is a suitable label to describe the level of model collapse. Specifically, we compared

gLMs by using different models to embed DNA sequences with different functional labels and tested the clustering performance based on these labels. To make a fair comparison, we subset the DNA sequences of human genes from a reference genome and pair the sequences with the corresponding functional labels from [39], as inputs and observed labels. We considered Normalized Mutual Information (NMI), Adjusted Rand Index (ARI), and Average Silhouette Width (ASW) as evaluation metrics [40, 41] and averaged them as a score to compare different gLMs, with a higher average score representing a better model. We selected gLMs [2, 11, 17, 19, 30, 42–50] from review and benchmarking papers detailed in the Methods section with UKBioBERT based on both scores and model scales. We also compared the embeddings generated from UKBioBERT with those from pre-trained DNABERT2 as the base model, those with random initialization, those with contrastive learning pre-training [51, 52], and those with LD score regression pre-training, shown in Extended Data Figure 1 (a). According to this figure, choosing a base model trained with a reference genome is helpful in learning a good gene representation, and MLM is best for pre-training. One possible reason is that the masked modeling approach is closer to the biological interpretation of variants, rather than statistical-based or computational-based methods such as LD score modeling and contrastive learning. We follow the default pooling approaches from these methods to generate the embeddings, while the embeddings from UKBioBERT benchmarked in this section are generated based on the embeddings of the CLS token. Details of the comparison process are explained in the Methods section.

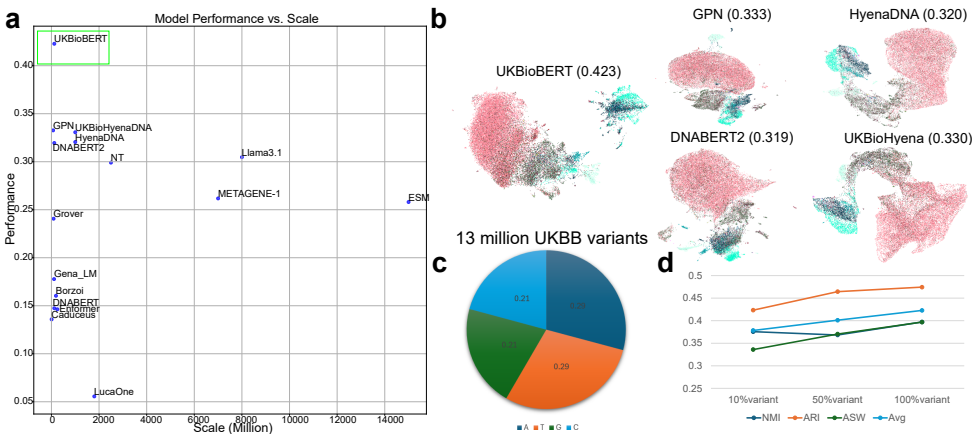

Fig. 2 Main results of model pre-training. (a) The scatter plot for the relationship between clustering performances and model scales. We highlight the best model UKBioBERT. (b) The UMAP visualization of UKBioBERT and other baselines with top performances. The average clustering scores are annotated with UMAP panels. (c) The statistics of the human variant corpus used for pre-training. (d) The relationship between model performance and training steps. The gene functional annotation list can be found in Extended Data Figure 1 (b).

Our results are shown in Figures 2 (a) and (b). According to Figure 2 (a), UKBioBERT has the highest average score in the representation of the genes with

their corresponding functions, without including the functional labels in the model pre-training stage. Furthermore, UKBioBERT has a relatively feasible parameter size compared with billion-level models, which improved its efficiency. UKBioBERT can also encode DNA sequences with arbitrary lengths by automatically dealing with the length of the input sequence. Moreover, we found that fine-tuning LLMs with DNA sequences and taking the embeddings do not help the model learn the functional information of genes, demonstrated by the low score of embeddings from Llama 3.1 7B [49, 50]. Models pre-trained with gene expression information did not perform well in this case, and one explanation could be that their attention mechanism assigned higher weights for the transcription starting site (TSS) rather than other regions, while the other explanation is that the random data augmentation used in the pre-training process is not as good as variant-based data augmentation. To visualize the differences of gene embeddings, we further show the sequence embeddings from UKBioBERT and other top baselines using uniform manifold approximation and projection (UMAP) [53] in Figure 2 (b). According to this figure, the panel corresponding to UKBioBERT clearly identifies the protein-encoding genes (with the pink labels) and other genes with different functions with specific clusters. The achievement of UKBioBERT can be attributed to the richness of the training dataset, whose statistics are shown in Figure 2 (c) and we utilized over 13 million variants for the continued pre-training stage. We also presented the Avg score vs. training step curve in Extended Data Figure 2 (a), which shows that the optimal checkpoint is obtained at the early stage of pre-training, and the reduction of clustering score can be treated as a signal of overfiting. Finally, we explored the data scaling laws in the pre-training stage to investigate if including more variants can improve the quality of gene representations generated by UKBioBERT, shown in Figure 2 (d). According to this figure, incorporating more human variants can improve the identification of gene functions and thus the quality of representations will improve as data scale increases based on UK BioBank data. However, we also intend to maintain a balance between the pre-training quality and the ability to predict gene expression. Therefore, we further evaluated the performances of using gene embeddings to predict individual gene expression levels as well as the quality of the embedding based on DNABERT2 pre-trained with gnomAD [54] (a dataset contains 0.7 billion human variants), and the results are summarized in Extended Data Figure 1 (a) and Extended Data Figure 2 (b). Based on these figures, our current setting has already optimized the representations of genes for both the gene function separation task and the expression prediction task. Therefore, we demonstrated the contribution of incorporating human variants into the model pre-training stage within a specific range of data contexts. The UMAPs for all baselines as well as the meaning of annotations are summarized in Extended Data Figures 1 (b) and (c). Therefore, the embeddings from UKBioBERT have the best performance in uncovering the gene functions directly from the sequence information, and thus can be treated as good representations with the preservation of genomic context.

Embeddings from UKBioBERT can enhance model performance for gene expression prediction at the cell-type level. Transformer-based methods have been widely shown to effectively predict gene expression levels from sequences.

Furthermore, recent research argued that incorporating and modeling external information, such as the interaction between promoter and potential enhancers, epigenetic signals, as well as chromatin contact data, can further improve the model's performance in predicting gene expression measured by different techniques, for example, CAGE-seq and RNA-seq [15]. For example, the authors of EPInformer demonstrated the advantages of EPInformer versus other transformer-based genomic deep-learning models in predicting gene expression levels by training genes on different chromosomes. Therefore, we extended this concept to a broadened domain and demonstrated that we could incorporate more informative signals to enhance EPInformer, including the DNA sequence embeddings from UKBioBERT as well as the gene description embeddings from scELMo [55]. The concept of DNA sequence embedding was discussed in our previous section, where our sequence embeddings allowed us to capture the functional information in the corresponding genomic region under the awareness of the whole DNA sequence library. Moreover, we could also utilize Large Language Models to generate and encode the descriptions of gene functions, and such text-based gene embeddings can also introduce the numerical representation of gene functions to help predict the gene expression profiles. Therefore, we improved EPInformer by modeling the default genomic information, the sequence embeddings from promoter and enhancers, as well as text-based gene embeddings, to formulate a new model and train it based on the same datasets. Due to data access limitations, we focused only on human cell lines with available promoter and enhancer information and validated our model with the corresponding CAGE-seq and RNA-seq results. We considered K562 and GM12878 cell lines, where we can retrieve enough information to implement the default mode of EPInformer. We also considered modeling silencers under the same framework, but there was no improvement after including the silencer information [56], shown in Extended Data Figure 3. Our hyper-parameter setting and data splits (12-fold cross-validation) follow the same settings from EPInformer, and thus our results can directly reveal the performance difference.

In Figure 3 (a), we visualize the comparisons among different modes of EPInformer, including the mode without pre-trained convolution network for identifying the enhancers (A), the default mode (B), EPInformer with DNA sequence embeddings generated by HyenaDNA [2] (C), EPInformer with DNA sequence embeddings from UKBioBERT (D), EPInformer with text-based embeddings from scELMo (E), EPInformer with embeddings from UKBioBERT and embeddings from scELMo (F). According to this figure (each dot represents the Pearson Correlation Coefficient (PCC) between predicted and observed values evaluated), the joint mode F outperformed the default mode of EPInformer, and also improved the method compared with other ablation versions. For the testing results based on CAGE-seq from GM12878, our method had an average PCC over 0.9, significantly outperforming other methods. Furthermore, our method also had a lower variance in 3 out of 4 datasets compared with the raw mode, which suggests that the joint mode has better stability. To further investigate the properties of model outputs, we visualize the predicted and observed gene expression levels from these 12 folds in Figure 3 (b), as well as their distributions. These figures are annotated with a set of overall PCCs, which are all higher than 0.8. Moreover, the predicted gene expression levels show a smoother distribution,

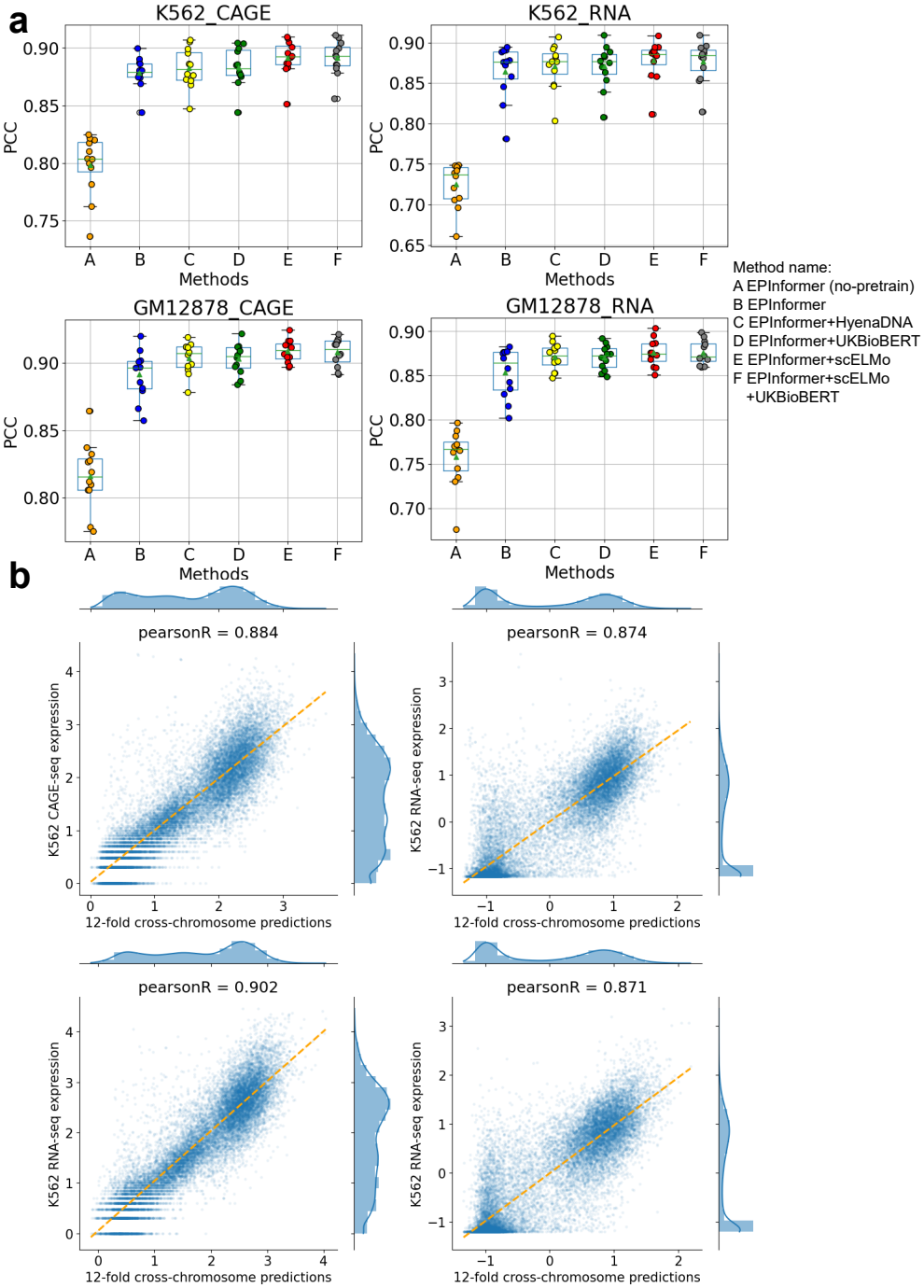

Fig. 3 The results of optimizing gene expression prediction in the K562 and GM12878 cell lines with UKBioBERT. (a) Comparisons of different methods for gene expression prediction across techniques and cell lines. The metric used in this section is PCC. (b) The scatter plot and distribution plot of the best performer (combination E) for the relationship between observed expression and predicted expression. The joint PCC is annotated and the yellow line is fit with two sets of expression levels.

especially in the prediction of RNA-seq data, which also suggests that the RNA-seq data might contain noise which could be reduced by our method.

Embeddings from UKBioBERT can recover the predictability of gene expression levels across different conditions. Since UKBioBERT is pre-trained with variants from large-scale population-level data, we also assumed that the DNA sequence embeddings from UKBioBERT can help us predict the gene expression profiles at the individual level, and further help explain the predictability of specific genes. Here we considered baselines including the traditional method ElasticNet [57], the pre-trained model with zero-shot learning mode Enformer, and fine-tuned Enformer based on personalized data, known as Personalized Enformer (Performer) [7]. To predict gene expression levels, we used the embeddings from UKBioBERT as inputs and ElasticNet as the predictor for evaluation, and chose 41 representative genes for analysis. These genes were either considered in previous research [7] or have gene expression level heritability estimated based on Linear Mixed Models (LMMs) [58]. In this study, we collected 670 samples from GTEx data with paired RNA-seq and WGS data, and performed 5-fold cross-validation to report the testing performances measured by three metrics: Pearson Correlation Coefficient (PCC), R^2 score (R2), and mean squared error (MSE).

The PCC between the predicted and observed expression levels is shown in Figure 4 (a). According to this figure, ElasticNet, UKBioBERT, and Performer all outperformed Enformer, suggesting that the zero-shot learning mode of Enformer cannot directly predict individualized gene expression levels well. Furthermore, these three methods showed similar results, supported by the strong correlation of different scores (PCC=0.988, p-value=2.19e-34 between UKBioBERT and ElasticNet, and PCC=0.991, p-value=1.92e-36 between UKBioBERT and Performer). We found that this is not only reflected in the results of PCC for the two groups of genes but also shown by studies on the R2 score and MSE, shown in Extended Data Figures 4 (a)-(d). On average, ElasticNet had slightly better performance, similar to the conclusion from a previous study [7]. However, we also found that no model can perfectly predict the expression of all genes and there is a significant difference in the prediction performance across different genes. We first tried to use heritability to explain this observation, but Figure 4 (b) showed that the heritability of the selected gene group did not have an apparent correlation with prediction performance. We further computed the correlation between metrics and heritability based on both PCC and Spearman Correlation Coefficient (SCC), but both results did not show significant correlation (p-value=0.19 for PCC, p-value=0.17 for SCC), shown in the left two panels of Figure 4 (c). Therefore, gene heritability does not explain the prediction differences across genes. Interestingly, by visualizing the relationship between the gene embeddings from UKBioBERT and expression levels, we found that the individuals with similar expression patterns are more likely to be clustered in one group, especially for genes with good prediction performance, shown in Extended Data Figure 6. Therefore, the clustering performance of individuals based on expression levels may be a good explanation for gene predictability. We calculated the clustering metrics based on NMI for all 41 genes and visualized the strong relationship between prediction performances and clustering metrics in the middle panels of Figure 4 (c), which supports our inference.

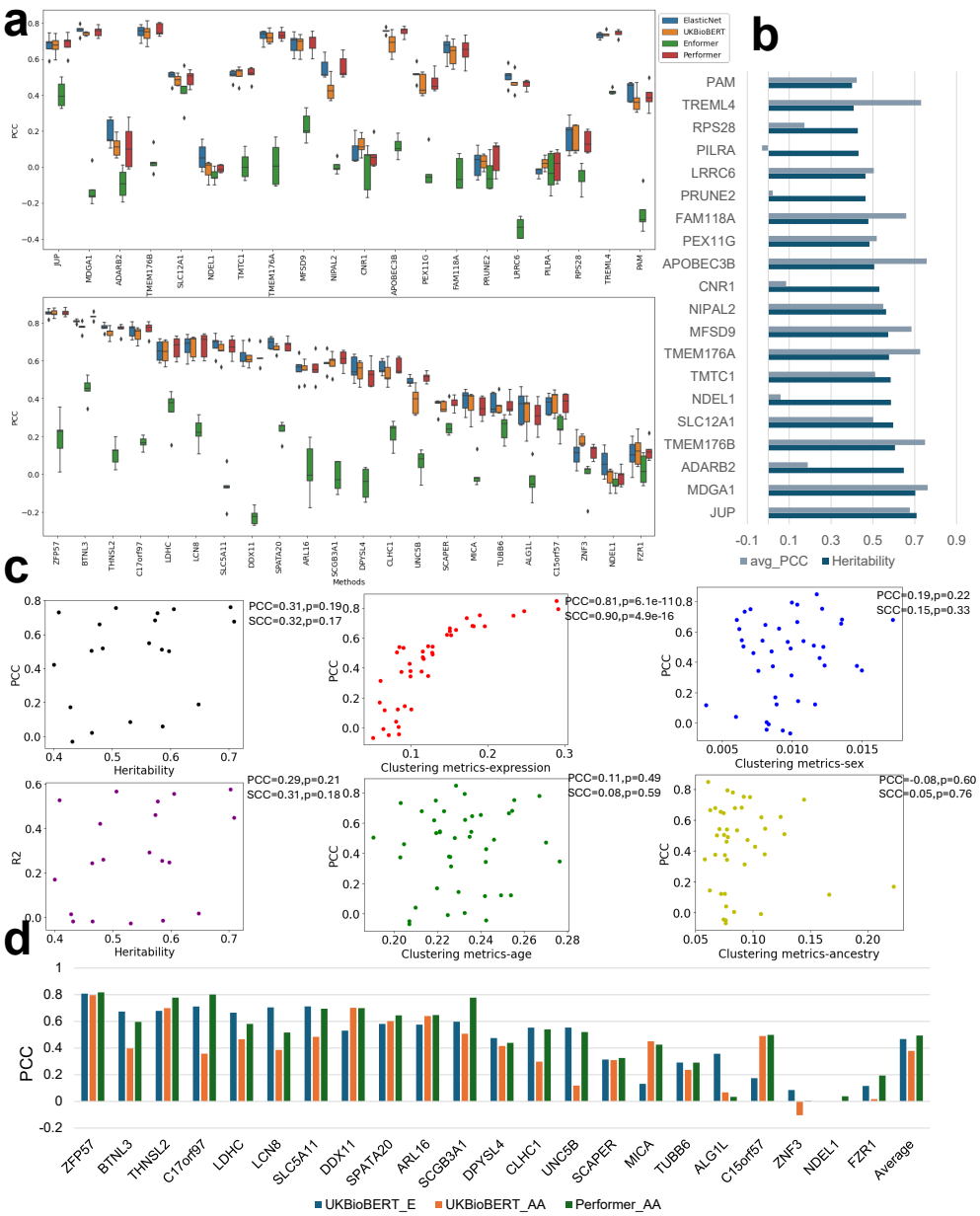

Fig. 4 Performance comparisons of gene expression prediction among UKBioBERT and selected baselines. (a) Comparison of PCC across two sets of genes for the selected methods. (b) Relationship between averaged PCC across five folds and heritability for selected genes. (c) Panel list for the relationship between scores of metrics and metadata information of individuals in GTEx. The metrics include PCC and R2, while the metadata includes heritability, cluster-based expression levels, age, sex, and ancestry. (d) Comparison of PCC for models tested with individuals from different ancestries across selected genes.

To investigate biological reasons for this strong relationship, we also analyzed the correlation between PCCs and other genome-level factors, such as gene location, number of unique DNA sequences, number of enhancers/promoters, and gene function score (GIFts) from GeneCard [59, 60], as discussed in Appendix A. Based on our analysis, the GIFts scores have a significant negative correlation with prediction performance (Extended Data Figure 11 (e)). It is also difficult to predict expression levels for genes with more functions because multiple gene functions cannot be fully explained only by sequence information. Therefore, before starting prediction, we recommend visualizing such relationships for the training datasets in low-dimensional space such as UMAP, computing the clustering scores, and searching the GIFts score to determine whether the specific gene may be predicted from sequences or not.

The gene embeddings from UKBioBERT can also be used to analyze the relationship between prediction performance and other metadata or environmental factors, such as age, gender, and ancestry. We also visualize such relationships between the PCC and these three factors in the middle and right panels of Figure 4 (c), showing that the accuracy of gene expression prediction at the population level is not strongly correlated with the above factors. The prediction performance is only related to how easily the corresponding gene sequences can be distinguished, and it is biased to attribute the expression of a specific gene to factors such as ancestry.

Moreover, we tested different pooling methods to generate the gene embeddings based on UKBioBERT for gene expression prediction, including CLS token representation (CLS), mean pooling representation (Mean), and max pooling representation (Max). We performed tests based on the top five genes ranked by PCCs and our results implied that the Mean pooling and Max pooling are comparable, and they both surpassed the CLS pooling approach, shown in Extended Data Figure 7 (a). This observation is consistent with previous analysis of using LLM embeddings for regression problems in medical applications [61]. We also explored possible data-efficient learning designs for predicting gene expression levels with the gene ZFP57, which has the best prediction performance, and model UKBioBERT. We subsampled the training datasets with different proportions of individuals and tested the prediction performances, shown in Extended Data Figure 7 (b). According to this figure, using all the samples had the best performance (highest average PCC). Furthermore, Extended Data Figure 7 (c) demonstrates that using the full genome information from both two parents can generate better prediction performance than using partial information. Therefore, our current implementation is optimal at the data level and we suggest including as many samples as possible for better prediction.

Fine-tuned UKBioFormer predicts more accurate gene expression levels across individuals compared with baselines. The previous section mainly focused on gene expression predictability by the representations from UKBioBERT, which did not cover two major challenges of personalized gene expression prediction: cross-cohort expression prediction and group-level gene expression prediction. The cross-cohort expression prediction focuses on predicting gene expression for unseen cohort-level data, such as sequence data from individuals in a new study or of different ancestry [57]. The group-level gene expression focuses on extending the training design of prediction at the single-gene level to multiple genes which belong to the same pathway or

share the same enhancers, starting from genes pre-trained with Enformer. We considered these two gene expression prediction problems (cross-cohort and group-level) in this section. In addition to the four baseline models presented previously, we considered two new models specifically designed for cross-individual prediction, denoted as UKBioFormer and UKBioZoi. UKBioFormer integrates Enformer and UKBioBERT for more accurate predictions with a parameter-efficient-fine-tuning (PEFT) design, and UKBioZoi follows the same idea but the base model is Borzoi. We used the same metrics to evaluate model performances.

To perform a systematical evaluation, we utilized the same two sets of genes and visualized the prediction metrics in Figures 5 (a), (b), Extended Data Figures 4 (a)-(d), and Extended Data Figures 5 (a) and (b). According to these figures, UKBioFormer had better performance than Performer and UKBioZoi on average, and its performance is consistent with our previous analysis based on gene predictability. If we selected genes with good predictability (PCC larger than 0.6), UKBioFormer had better performances than Performer in 63.3% genes. Furthermore, UKBioFormer had more genes with improvement over Performer, shown in Extended Data Figure 5 with higher x-axis value. Furthermore, UKBioFormer had a more efficient design as its running time and GPU memory consumption are lower than Performer and ElasticNet, shown in Extended Data Figures 8 (a)-(c). Although UKBioZoi had a slightly worse performance, it has shorter training time and less memory cost. Therefore, we recommend using UKBioFormer first for personalized gene expression prediction, and if the computation resource becomes a limitation, using UKBioZoi as an alternative plan.

We also considered expression predictions across cohort-level data, for example, predicting the gene expression levels for African American (AA) individuals based on a model trained with European (E) individuals. According to Figure 4 (d) and Extended Data Figure 9, compared to training on Europeans and testing on Europeans, the prediction performance of UKBioBERT on the AA population was overall low, with prediction correlation negative for some genes. This may be due to the limitation of the embedding-based model. However, if we utilized Performer or UKBioFormer to make the same predictions, there was an obvious improvement in the prediction results based on the AA population, without modifying the training datasets. This result can be explained by the fact that sequence-based models can better capture variants that are correlated between groups while having a larger receptive field and therefore more generalization ability. Our analysis for the Religious Orders Study and rush Memory and Aging Project (ROSMAP) [62] dataset can be found in Appendix B.

It is well known that genes do not act alone, with genes of similar or complementary functions to regulate biological processes [63]. This is partly captured by gene-gene interaction [64]. Therefore, we hypothesize that training genes with similar functions together might boost each gene's prediction performance. Furthermore, it has been demonstrated that joint training of different genes might improve prediction of some regulatory factors [7], and thus we explored if we can train UKBioFormer as a group. Therefore, we considered three different gene groups to examine this assumption, including the enhancer-shared group, the pathway-shared group, and the Enformer-pretrained group. The enhancer-shared group contains ZFP57 and other genes sharing the same enhancers in the blood tissue (n=8). The pathway-shared group contains

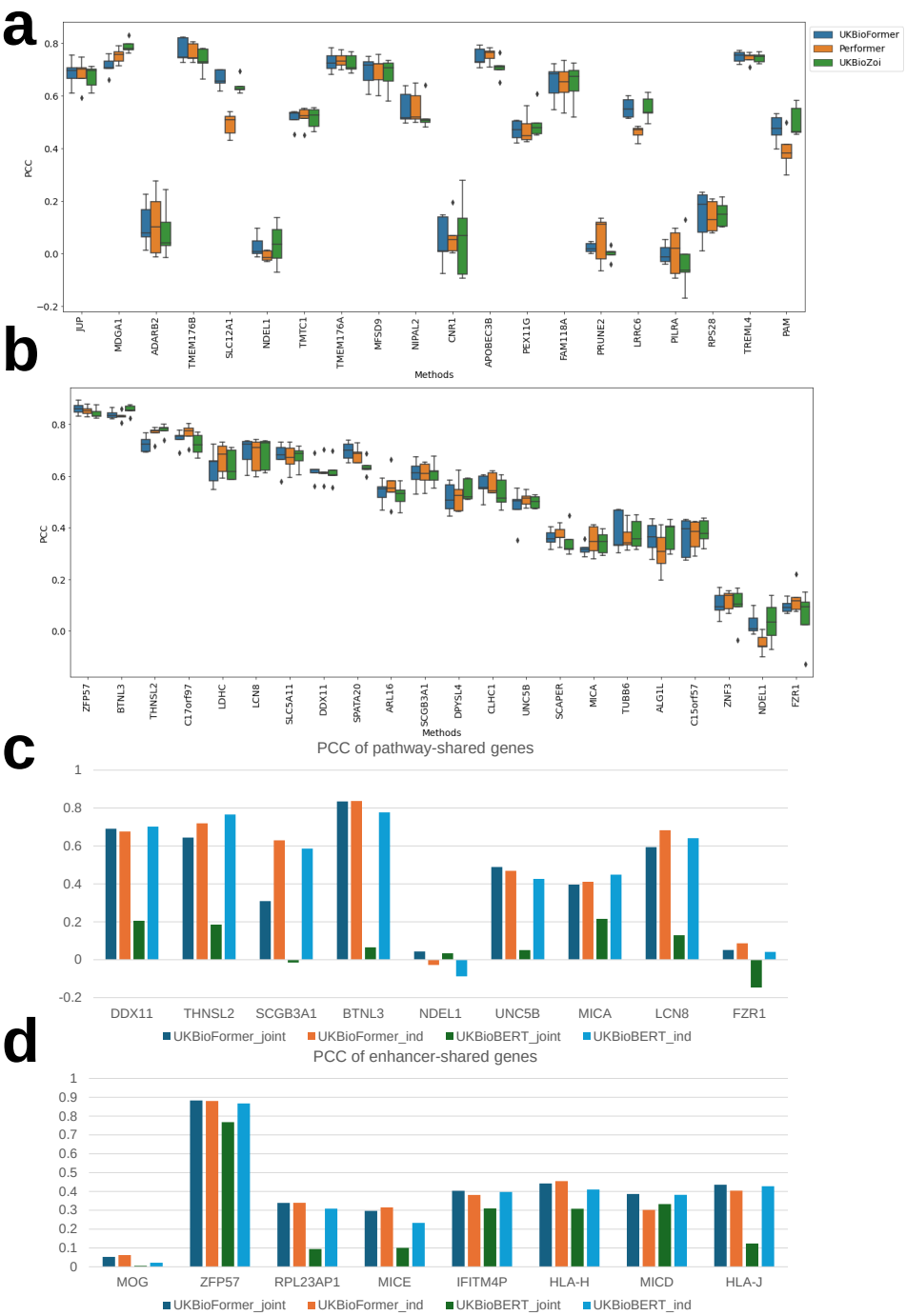

Fig. 5 Performance comparisons among UKBioFormer and selected baselines. (a) Comparison of PCC across genes from group 1 for selected methods. (b) The comparison of PCC across genes from group 2 for selected methods. (c) Comparison between joint training (UKBioBERT_joint, UKBioFormer_joint) and independent training (UKBioBERT_ind, UKBioFormer_ind) for pathway-shared genes measured by PCC. (d) Comparison between joint training (UKBioBERT_joint, UKBioFormer_joint) and independent training (UKBioBERT_ind, UKBioFormer_ind) for enhancer-shared genes measured by PCC.

genes in the same Gene Ontology (GO) pathway [65, 66] physiological response to stimulus (GO:0050896) (n=9). The last group contains genes used in the pre-training stage of Enformer (n=300). According to Figures 5 (c) and (d), the group-level training might not always benefit the expression prediction for genes in either enhancer-shared group or pathway-shared group, by comparing the group-level prediction performance with separate prediction performance using the same testing dataset. If we only use UKBioBERT, using a group of genes at the same time for training can even reduce the model's performance. Moreover, according to Extended Data Figures 10 (a)-(c), the prediction performance based on the Enformer-pretrained group is also not promising, as many genes have a PCC score lower than 0.5. Therefore, the training model in the group-level genes for expression prediction is still challenging and could be affected by many limitations, for example, the low quality of training data as well as the under-fiting of UKBioFormer. However, limited by the current data access and computation resources, these challenges are difficult to overcome. The results from Performer also suggest that the single-gene-level model can explain more variability than the group-gene-level model. Therefore, we believe that the current optimal plan is still training single-gene-level models for genes with good predictability.

Regarding the ablation studies and hyper-parameter tuning of UKBioFormer, we summarized the results based on the expression profiles of ZFP57 in Extended Data Figures 12 (a)-(d). In Extended Data Figure 12 (a), we demonstrated that the default setting referred from [7] to set epochs=100 is optimal, as increasing the epochs does not improve the model performance. We set the batch size to the maximum allowable value of the allocated GPU (8 for A40 and 12 for H100 or A100). Reducing the epoch is also not required and recommended as we have a validation dataset to select the model with the best weights. Extended Data Figure 12 (b) shows that using a lower learning rate can help improve the prediction performances. Extended Data Figure 12 (c) shows our results for the PEFT design. According to this figure, reducing the number of transformer layers can improve model performance, while other PEFT settings such as low-rank adaption (LoRA) [67] may reduce model performance. Furthermore, running PEFT design by reducing the number of transformer layers also outperformed the raw mode of Performer, and thus the optimal setting of PEFT design is to use the one-layer transformer architecture during the fine-tuning process. In Extended Data Figure 12 (d), we compared the default setting of the optimizer (Adam [68]) and loss function (mean squared error loss) with other ablation settings, by alternating optimizers (AdamW [69] or MUON [70]) or loss functions (pcc_mse, huber_loss [71], or group_loss [7]), we did not see significant improvements, and thus the default setting of UKBioFormer is optimal. Using gradient accumulation (GA) to increase the batch size also does not improve model performance, shown in Extended Data Figure 12 (E). UKBioFormer also converged faster than Performer, shown in Extended Data Figure 12 (f). We finally compared UKBioFormer and UKBioZoi with implementation of Borzoi based on Flash Attention [72, 73], known as UKBioFlashZoi (the base model is FlashZoi [74]), and the results are shown in Extended Data Figure 13. According to this figure, UKBioFormer had the best performance, while UKBioZoi performed slightly better than UKBioFlashZoi.

UKBioFormer can identify the direction of eQTL effects with the help of neural network explainability. Linking the variant with gene expression is a very challenging but important task, which can help us understand gene regulation. Previous research for eQTL inference requires access to matched genotype and expression data, which is expensive and time-consuming. Therefore, one of the major contributions introduced by the sequence-to-function model is to infer eQTL without having to measure hundreds to thousands of individual gene expression profiles, claimed by Enformer. However, the analysis of Performer shows that only using pre-trained Enformer cannot fully explain the effects of most eQTLs, and thus the fine-tuning process of the sequence-to-function model for better inferring eQTL is required and encouraged. Moreover, the fine-tuned model might also capture eQTL better than ElasticNet, especially for rare variants. Therefore, we hypothesize that UKBioFormer should also have the ability to identify eQTL with the correct direction and magnitude for the predictable genes. Furthermore, with the explainability tools designed for the sequence-to-function model, we can also compute gradients and perform In Silico Mutagenesis (ISM) to analyze the effects of different variants at the same locus, allowing us to uncover the role of such locus in regulating gene expression levels. The approaches we used to interpret UKBioFormer are summarized in Figure 6 (a).

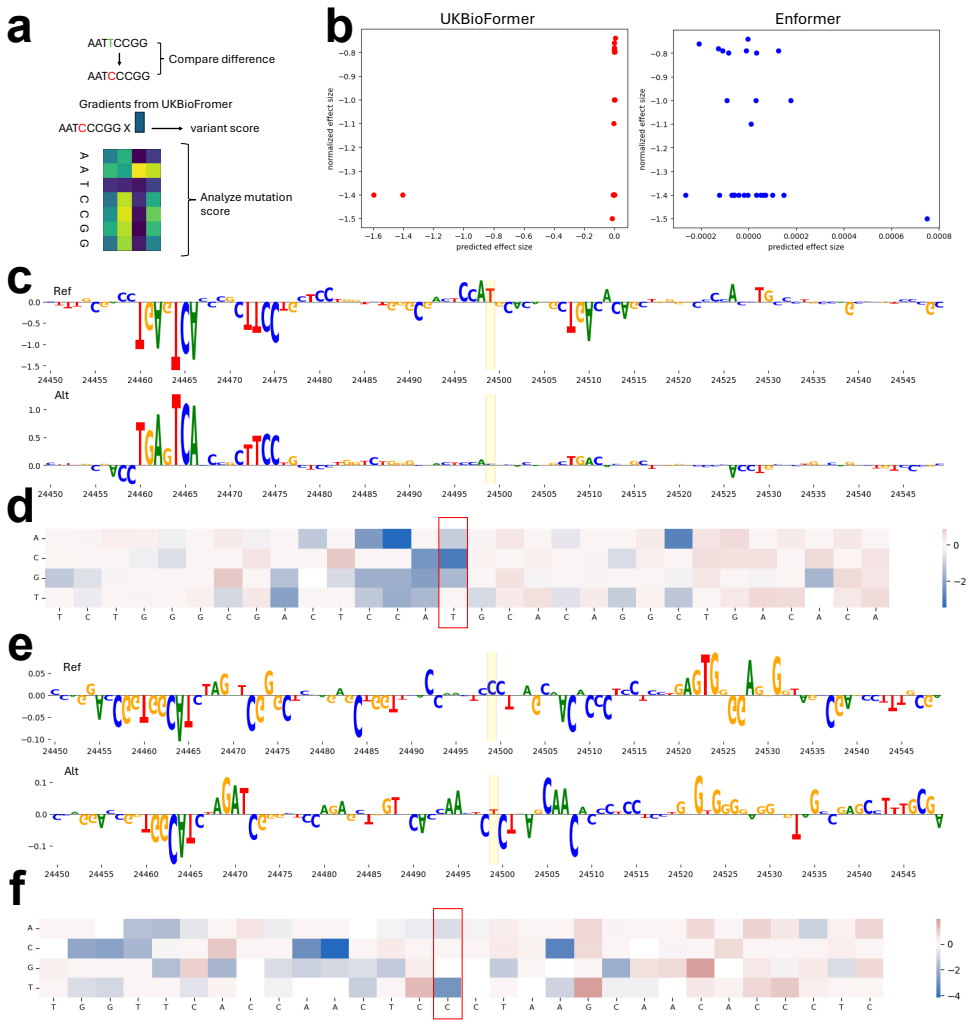

Fig. 6 Understanding variant effects based on fine-tuned UKBioFormer. (a) The schematic diagram of the analysis of variance effects. We considered three approaches to analyzing variant effects, including effect size prediction, integrated gradients computation, and In Silico Mutagenesis (ISM). (b) The relationship between predicted effect size and normalized effect size based on the results from UKBioFormer and Enformer. (c) The visualization of base importance based on the reference sequence and the alternative sequence of eQTL with code rs9910080. (d) The results of ISM operation at the region near the location of rs9910080. (e) The visualization of base importance based on the reference sequence and the alternative sequence of eQTL with code rs9903086. (f) The results of ISM operation at the region near the location of rs9903086.

To identify eQTLs based on UKBioFormer, we focused on the gene JUP as UKBioFormer has good predictions based on the testing datasets of JUP, and thus its expression levels are predictable. We further utilized the same training dataset from GTEx and selected the best model with the explainability mode based on the same

validation dataset from blood. The estimated eQTL from the same dataset can be downloaded from the GTEx website, with information on reference alleles, alternative alleles, and normalized effect size. Firstly, we utilized the ISM method to predict the effect size of the top 30 eQTLs among 120 eQTLs with UKBioFormer ranked by p-value. Notably, 71% of predicted eQTLs have the correct signs, which is higher than the probability of random guess (50%) as well as the proportion estimated from Enformer (53%) or Performer (68%). Furthermore, we found that the prediction results of UKBioBERT for eQTL analysis also recovered the relationship between effect size and variance. We visualize the relationship between effect size and standard error (se) reported by GTEx database in Extended Data Figure 14, which shows a positive correlation between absolute effect size and se (PCC=0.49, p-value=1.72e-23). Such relationship was also reflected in the Extended Data Figure 9 from [75]. We observed that eQTL with high se computed by UKBioFormer are more likely to have correct directions (79% predicted eQTLs have correct directions for the high se group, and 64% predicted eQTLs have correct directions for the low se group), which matched the observed condition. We also plotted the relationship between predicted and observed effect size (normalized effect size) in Figure 6 (b) based on UKBioFormer (left panel) and Enformer (right panel), and UKBioFormer presented better alignment with observed effect size. We could also identify two representative eQTLs with perfect matching scores, including variant rs9910080 and variant rs9903086, which are also highlighted in the same figure. These two variants overlap with enhancer GH17J041791 of gene JUP, and thus biologically they have the potential to affect gene expression levels. Second, we obtained the contribution of different base pairs to the prediction by running the gradient attribution method for a specific sequence region (inputs $\times$ gradients), the regions close to the TSS, and containing eQTLs of JUP. According to Figure 6 (c), we observed that the change of base pair for eQTL will cause an obvious gradient difference, and changing the selected base from a reference allele (T) to an alternative allele (C) results in a decrease in predicted gene expression, which matches the direction of this eQTL. We further considered running ISM to estimate the change of different alleles and their effects on gene expression levels, and Figure 6 (d) presents that the alternative allele also has the lowest ISM score. Such conclusions can also be derived from our analysis of the other variant in Figures 6 (e) and (f). The results of base importance and ISM scores for selected variants but computed based on Enformer and Performer are shown in Extended Data Figures 15 (a)-(d), which implied worse performance than UKBioFormer as the estimated variant effect is wrong in both magnitude and direction. Therefore, by using UKBioFormer to predict gene expression and capture the motifs associated with this expression phenotype, we can explore the effects of different mutations on gene expression and thus understand genetic relationships at a deeper level.

3 Discussion

Sequence-to-function models have become important for understanding how genetic information can be executed or interpreted in various biological processes. This can help us further investigate the role of variants in expression regulation as well as gene

heritability and predictability. However, few sequence-to-functions models focus on understanding individual genome information, and these models lack the design to capture the contributions of variants far from TSS as well as genome context. Furthermore, these models are large and inefficient, which creates challenges when deploying them in single-GPU platforms. Furthermore, incorporating genomic language models for interpreting variant effects at the individual level is also lacking in studies, even if this could be a promising direction as an enhancement of sequence-to-function models.

Therefore, to overcome these challenges, we have proposed two sets of models to better understand functional genomics with the help of cohort-level genomic data. We leverage UK BioBank database and extract the large-scale variants to train a gLM based on DNABERT2, known as UKBioBERT. We demonstrated that UKBioBERT can generate a better representation for genes with functional annotations by benchmarking the clustering performances with other gLMs. Such evaluation of function preservation in the latent space can also be used as a new metric for model pre-training. We also found that the embeddings from UKBioBERT are more informative to visualize individual information, enhance gene expression prediction in different cell lines, and provide insights to gene predictability, which is not precisely explained by heritability. Therefore, variant information offers informative data augmentation for gLM training.

To link UKBioBERT with sequence-to-function models, we fused the embeddings from UKBioBERT with models including Enformer or Borzoi and proposed two new sequence-to-function models, known as UKBioFormer and UKBioZoi, to predict gene expression for different individuals more efficiently. We demonstrated that these two models had good performance in the genes with high predictability, and their results also had high correlations with ElasticNet and outperformed Performer. Based on our results, we recommend using UKBioZoi when computation resource is a challenge, while UKBioFormer is expected to achieve a better performance. Moreover, we showed that using UKBioFormer can also make cross-cohort expression prediction, for example, training a model from individuals with European ancestry and making prediction in African-American individuals, which demonstrates that UKBioFormer is less affected by the biases brought by different variants across cohorts. We also discussed in detail the design options for UKBioFormer, including how to choose the loss functions, optimizers, hyper-parameters, and plans of PEFT. Therefore, work may lead to future development to combine these two types of models to make further improvements or more interesting applications.

Finally, the explainability of UKBioFormer can also explain the role of variants in regulating gene expression levels with good predictability. We demonstrated that UKBioFormer has the ability to predict the correct direction of over 70% eQTLs, and can also match the observed effect size well for certain important eQTLs. Furthermore, we can also utilize UKBioFormer to test the effect of base changes on the role of surrounding bases in gene expression regulation and select target variants to test their effects experimentally. Therefore, we believe that UKBioFormer will open more possibilities to explore how variants affect gene expressions.

For future work, we plan to tackle one challenge discussed in this study, namely, the fog of multi-gene training. We found that training the model with multiple genes

cannot always improve expression prediction for all the selected genes in different scenarios. This could be a problem due to the drawbacks in model design, the limitation of computation resources, or the noise in data. As a next step, we plan to focus on cell-type information across individuals and incorporate more modalities as training datasets to see if we can overcome this challenge in the cell-type-specific context. Moreover, it is also promising to link the discoveries of UKBioBERT and UKBioFormer with therapeutic research in diseased samples, which might inspire novel treatment plans for diseases associated with genetics.

4 Methods

Problem statement. Here we intend to first have a model $\mathcal{M}$, which can model the DNA sequence dataset $\mathcal{D}$ by learning a good representation of arbitrary sequences. This approach can be implemented by the self-supervised approach (e.g., mask language modeling strategy [16]) or supervised approach (e.g., predicting Linkage Disequilibrium (LD) score [33]) at the pre-training stage. Furthermore, we also intend to have the other model $\mathcal{P}$, which is built based on the pre-trained model $\mathcal{M}$ and should predict the gene expression levels based on the corresponding sequence information of the given gene in the population level. To achieve this goal, we can train the model $\mathcal{P}$ by minimizing the distance between the observed gene expression levels y and predicted gene expression levels $\hat{y}$.

Implementation of UKBioBERT. Here we discuss the model architecture and pre-training strategies for UKBioBERT. We utilize DNABERT2 [17] (which is a Bidirectional Encoder Representation from Transformers [16] model pre-trained with reference genome), and continue pre-training our method based on the sequences with variants. We create a function to modify the reference genome with matched variant information, as shown in Algorithm 1. The variants come from UK BioBank [33] database's EUR population. Our pre-training strategy is masked language modeling, which is determined after comparing it with the contrastive-learning-based pre-training approach and predicting the LD score approach. We split the training, validation, and testing dataset for UKBioBERT to select the optimal checkpoint, and then test the clustering score. The loss function for model training is to classify the mask token by minimizing the cross-entropy loss.

We follow the default setting of DNABERT2 for hyper-parameter selection. UKBioBERT supports sequences with different lengths, as long as the user has paired GPU capacity. Based on our experiments, an H100 GPU can support one 100K DNA sequence as input. We also pre-train HyenaDNA [2] with a causal language modeling approach based on sequence information from UK BioBank data as a baseline, which is denoted as UKBioHyenaDNA.

Applications of UKBioBERT. We extract the embeddings of DNA sequences from UKBioBERT for downstream applications.

1. Visualization. We utilize the DNA sequence embeddings corresponding to gene functional labels or individual labels for visualizing the DNA sequence in the lower dimension, which shows reasonable clustering performances.

Algorithm 1 Sequence editing functions.

F1: Replacing. Input: Raw DNA sequence R, alternative variant v, and variant index i.
Output: Modified DNA sequence R_r.
Compute: $R[i] = \text{v}$, $R_r = R$
Return R_r

F2: Inserting. Input: Raw DNA sequence R, alternative variant v, and variant index i.
Output: Modified DNA sequence R_s.
Compute: $R_1 = R[: i], R_2 = R[i :]$.
Compute: $R_2 = Append(v, R2)$.
Compute: $R_s = R_1 || R_2$.
Return R_s

F3: Deleting. Input: Raw DNA sequence R, deleted base(s) v, and deleted index i.
Output: Modified DNA sequence R_d.
Compute: $R_1 = R[: i], R_2 = R[i + len(v) :]$.
Compute: $R_d = R_1 || R_2$.
Return R_d

2. Cell-state-specific expression prediction. UKBioBERT can generate sequence embeddings to help EPInformer [15] in predicting gene expression levels for known cell lines, including K562 and GM12878 [76]. These cell lines contain accessible information on enhancers, chromatin interactions, etc. We encode one promoter sequence and 60 enhancer sequences with length 2000 (the default setting of EPInformer) and merge the embeddings with the default input of EPInformer to provide extra information for model training. Our embeddings can significantly improve model performance, compared with the default mode.

3. Individual gene expression prediction. Leveraging gene embeddings from UKBioBERT, we can also use the encoded gene sequence from Whole Genome Sequencing data [77] of individuals to predict the corresponding expression levels in the selected sample for the corresponding gene. Our regressor is ElasticNet [78] and we perform cross-validation to select the best hyper-parameter setting and make predictions for five different testing sets.

4. Individual stratification. The gene embeddings from UKBioBERT can further help us understand the differences across individuals and how they affect the prediction performances. We consider different labels, including expression levels, ancestries, and genders to explore the possibility of using UKBioBERT to distinguish them. Our results show that UKBioBERT can link the predictability of gene expression levels with the embedding quality, which further helps us to determine which set of genes can be successfully predicted.

Finally, we intend to highlight that UKBioBERT is powerful in generating sequence embeddings, and thus its application is not limited to expression-related analysis.

Implementation of UKBioFormer. We build a new transformer-based model, named as UKBioFormer, based on Enformer [30] and UKBioBERT. The improvement of UKBioFormer is discussed in two sections, the design contribution as well as the engineering contribution. UKBioZoi follows a similar idea but its base model is Borzoi, so we do not plan to discuss it in detail.

Regarding the design contribution, previous research demonstrated that the attention mechanism applies a higher weight to the transcription starting site (TSS) and to regions in close proximity to the TSS, which might not model correctly for the contribution of far-away genetic information or variants [7, 79]. Therefore, incorporating the gene embeddings from UKBioBERT can further help the model to capture the nucleotide information in marginal regions, as we encode DNA sequences by averaging the embeddings of different tokens and our pre-training strategy does not include the TSS information to reduce inductive bias. In summary, we believe that UKBioFormer can make a good balance between learning the local and global genetic information for expression prediction.

Regarding the engineering contribution, we consider pre-training, parameter efficient fine-tuning (PEFT) [80], and feature fusion to enhance the efficiency and capacity of UKBioBERT. First, we do not train UKBioFormer from scratch but utilize the model weights from Enformer as initial parameters. According to [7], this approach might at least help in predicting the expression levels from genes pre-trained based on the CAGE-seq of Enformer, and thus starting from a good initial weight can also help the model perform better transfer learning. Second, we consider applying the PEFT approach for fine-tuning UKBioFormer with lower resource requirements as well as possibly better model performances. Using PEFT can also support the model with longer sequences as inputs, compared with the 49k length used by Performer for fine-tuning, our UKBioFormer can accept over 100k length sequences. In addition, we have tried the low-rank adaption (LoRA) [67] approach and pruning approach to optimize the parameter usage of UKBioFormer, and found that pruning the transformer layers can improve the training efficiency as well as performances better than LoRA in our Results section. Compared with UKBioBERT, UKBioFormer also has a faster convergence speed, as discussed in the Results section. Third, we demonstrated that Fusion is an effective approach to combine the gene embeddings from UKBioBERT for our learning target. Here we finally extract the outputs from the regression head of gene embeddings from pruned Enformer (which is training) and weighted average the outputs from such gene embeddings and the outputs from the regression head based on the gene embeddings from UKBioBERT (which is frozen). The weight is learnable and always in the range of (0,1). This approach can not only guarantee that our models will capture the local and global genetic information but also will not affect the default abilities of interpreting UKBioFormer.

The loss function of the fine-tuning method is mean squared error loss, which is determined after comparing different kinds of loss functions. Here we evaluated loss functions including mean squared error (MSE) loss, the joint loss between MSE and PCC (pcc_mse), huber loss (huber_loss) [71], and group loss (group_loss) [7]. To construct pcc_mse, we train the model by maximizing the correlation between the observed and predicted gene expression levels in the training dataset while minimizing

the MSE between them. To construct huber_loss, we minimize the balanced MSE loss between the observed and predicted gene expression levels. To construct group loss, we not only minimize the MSE between the observed and predicted gene expression, but minimize the MSE between the difference of individuals in two groups to let the model also learn the variability existing in individuals.

Most of our hyper-parameters follow the settings from [7]. We have a smaller batch size because we do not have the computation resources used in this manuscript (8*H100). However, our model's performance for cross-individual prediction is comparable with their setting. Our method is implemented based on gReLU [81].

Applications of UKBioFormer. Here we discuss the application scenarios of UKBioFormer, which cover the important tasks in understanding functional genomics, ranging from expression prediction to expression quantitative trait locus (eQTL) identification.

1. Cross-individual gene expression prediction. By fine-tuning UKBioFormer with training individuals and validation individuals, we can obtain an optimized model for specific genes' expression prediction. We can further test our model's performance based on the DNA sequences from the testing individuals. We demonstrated that our methods can outperform baseless ones for predicting the gene expressions with good predictability.

2. Reducing biases in expression prediction. We also demonstrate that our methods can improve prediction across individuals with different confounder variables, such as ancestry. We train our model with individuals from European (E) ancestry and test the model with individuals from African American (AA) ancestry. It shows that our fine-tuning approach achieves better performance for certain genes, compared with selected baselines. We also notice that for this application, using all of the transformer layers provided a more stable and better prediction on average.

3. Testing the synergetic effect of the training model at a group level. We also explore the potential contributions brought by training genes with similar features in a group and investigate if this approach can enhance the prediction performances for all selected genes or not. Here we consider three different groups, including enhancer-shared group, pathway-shared group, and Enformer-pretrained group. The first group contains genes sharing the same enhancer in the blood tissue, the second group contains genes in the same Gene Ontology (GO) pathway, and the last group contains genes pre-trained in the Enformer model, provided by [7, 30]. We compare the joint-training results with individual-training results to analyze the prediction results across different genes, and make conclusions based on our prediction results accordingly.

4. Interpreting mutation effects. We can also utilize tools to interpret the contributions of mutation to the change of gene expression prediction. Here we collect the reference sequences and alternative sequences and utilize In Silico Mutagenesis (ISM) [30, 81] to compute the weights for different sequences, By comparing the vector of weeds for these two cases we can understand whether the given variant or mutation can enhance or reduce gene expression levels.

5. Identifying eQTL. The traditional approach utilizes a statistical framework to estimate the contribution of a single variant for affecting gene expression prediction,

known as eQTL inference. Different eQTLs might have different directions and magnitudes for the targeted gene. We demonstrated that our methods can help predict the directions of eQTLs for genes with good predictability, and improve the prediction of magnitudes for certain eQTLs with strong effect size (which is defined as a large absolute value). However, there are still various challenges for further identifying reliable eQTLs.

Metrics. For the model pre-training stage, we evaluate the quality of embeddings based on the preservation of the gene function context. The idea is to run a clustering algorithm based on gene embeddings from different foundation models and compare the cluster labels with annotated gene functional labels. We access the optimal resolution of the Leiden clustering algorithm for each method's generation to perform a fair comparison. The metrics we considered included Normalized Mutual Information, Adjusted Rand Index, Average Silhouette Width, and their average score Avg. These metrics are computed based on scikit-learn [40] and scIB [41]. Details are discussed below.

- Normalized Mutual Information (NMI): NMI is a score to evaluate the performance of biological information conservation. We compute this score based on the mutual information between the optimal Leiden clusters and the known cell-type labels and then take the normalization.
- Adjusted Rand Index (ARI): ARI is a score to evaluate the performance of biological information conservation. ARI is used to evaluate the agreement between optimal Leiden clusters and gene labels.
- Average Silhouette Width (ASW): We have gene type ASW for this metric. For one cell, ASW calculates the ratio between the inner cluster distance and the intra cluster distance for this gene. Therefore, higher ASW_{cell} means better biological information conservation. To make them consistent, we take the normalization, that is:
$$ASW = \frac{ASW^{raw} + 1}{2}.$$

All of the selected metrics in this section range from 0 to 1, and thus we can also average them to evaluate the effect jointly, which can be represented as Avg $\in (0, 1)$.

For the cell-type-specific gene expression task, we utilize the Pearson Correlation Coefficients (PCCs) between the measured gene expression levels and observed gene expression levels as the metric, which follows the default choice of EPInformer.

For the individual gene expression task, we utilize PCCs, R^2 score, and mean square error (MSE) as metrics for evaluation. Higher PCCs and R^2 scores represent a better model, while higher MSE represents a worse model. All the metrics are implemented based on scikit-learn or Scipy [82].

For the eQTL discovery task, we take the observed eQTL information from GTEx website [77], including the variant information and normalized effect size (NES). We compare the signs and absolute values of NES and predicted effect size to perform the comparison.

Baselines. Here we discuss the baselines used in this study for different tasks.

For the model pre-training stage, we consider various foundation models, which are trained based on different strategies as well as base models. The first group is multi-task biological sequence foundation models, including DNABERT [11], DNABERT2 [17], GPN [28], HyenaDNA [2], nucleotide transformer (NT) [42], Grover [43], Gena_LM [19], Caduceus [44], ESM [45], METAGENE-1 [46], and LucaOne [47]. Most of the models are pre-trained with DNA sequences, while LucaOne takes multi-omic-sequence information in the pre-training stage and ESM utilizes the gene's corresponding protein sequence as inputs. If they have multiple versions, we utilize the model with the largest parameter size. The embeddings from ESM are extracted based on the file provided by KGWAS [83]. These models generate sequence embeddings or novel DNA sequences, which can be used for further applications. The second group is an expression-based large-scale sequence model, including Enformer [30] and Borzoi [48]. These models are trained to predict gene expression levels under different technologies (CAGE-seq or RNA-seq), and the learned embeddings will encode the functional information of the selected sequences. The embeddings and outputs of these models are often used for tasks related to gene expression prediction. The last group is the general foundation model, for example, Llama 3.1 [49, 50] was pre-trained with multi-language data, which is a powerful open-source tool for tasks related to natural language processing. Since Llama 3.1 does not provide direct access for its generated embeddings, we pre-train Llama 3.1 with DNA sequence information from UK BioBank and use AngIE [84] to generate the embeddings from Llama 3.1.

For the cell-type-specific gene expression prediction, we consider EPInformer [15] and its variants as baselines. EPInformer is a transformer-based predictor leveraging the information of promoter sequences, enhancer sequences, as well as epigenomic information for enhancing model performance in gene expression, enhancer identification, and variant effect estimation based on data from cell lines. In addition to our final design, we compared EPInformer which incorporates different features, including the combination of EPInformer and HyenaDNA or scELMo [55].

For individual gene expression prediction, we include ElasticNet and UKBioBERT as baselines. ElasticNet is widely used in predicting gene expression at the individual levels based on SNPs. We also utilize UKBioBERT to generate sequence embeddings and these embeddings can be treated as inputs of ElasticNet to make predictions. We use the cross-validation setting of ElasticNet and 2000 maximal epochs for model training [7], and we only consider single-gene model for both ElasticNet and SuSiE [85] in our experiment. We also consider the variants of Borzoi as UKBioZoi and include them in our benchmarking analysis.

Details of the parameter setting of baseline models are summarized in our codes.

Datasets. The dataset used in pre-training can be downloaded from the UK BioBank website. The gene expression and sequence information of cell lines K562 and GM12878 can be downloaded from the links provided by EPInformer. The dataset used in individual gene expression prediction can be downloaded from GTEx and ROSMAP websites. Access to datasets from UK BioBank, GTEx, and ROSMAP requires extra requests. We summarized the dataset information in Supplementary File 1.

Code availability and reproductivity. For pre-training data pre-processing, we reply on Yale High-performance Computing Center (YCRC) and utilize one CPU with

up to 800 GB. For model pre-training, we utilize one NVIDIA A100 (H100) GPU with up to 100 GB RAM. For model fine-tuning, we utilize one NVIDIA A100 (H100/A40) GPU with up to 70 GB RAM.

The codes of this project can be found in https://github.com/HelloWorldLTY/UKBioLM. The license is MIT license.

5 Ethics and Inclusion.

Although UKBioBERT and UKBioFormer are not biased to gender, races, and other factors, the users are solely responsible for the content they generate with models in UKBioBERT and UKBioFormer, and there are no mechanisms in place for addressing harmful, unfaithful, biased, and toxic content disclosure. Any modifications of the models should be released under different version numbers to keep track of the original models related to this manuscript.

The target of current UKBioBERT and UKBioFormer only serves for academic research. The users cannot use it for other purposes. Finally, we are not responsible for any effects of the use of the model.

6 Acknowledgments.

We thank Shiron Drusinsky and Sean Whalen for the suggestions of data preprocessing and the reproduction of Performer and ElasticNet, and Dr. Zihan Dong and Leqi Xu for the suggestions of using UK BioBank data.

7 Author contributions.

T.L. proposed the study. T.L. and X.Z. ran all the experiments. R.Y. and H.Z. provided the computation resources. T.L., H.Z., and R.Y. wrote the manuscript. H.Z. supervised this project.

8 Competing interests.

The authors declare no competing interests.

References

[1] Yang, Z.: A space-time process model for the evolution of dna sequences. Genetics **139**(2), 993–1005 (1995)

[2] Nguyen, E., Poli, M., Faizi, M., Thomas, A., Wornow, M., Birch-Sykes, C., Massaroli, S., Patel, A., Rabideau, C., Bengio, Y., et al.: Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. Advances in neural information processing systems **36** (2024)

[3] Reinagel, A., Bray Speth, E.: Beyond the central dogma: model-based learning of how genes determine phenotypes. CBE—Life Sciences Education **15**(1), 4 (2016)

[4] Zhou, J., Troyanskaya, O.G.: Predicting effects of noncoding variants with deep learning–based sequence model. Nature methods **12**(10), 931–934 (2015)

[5] Zhou, J., Theesfeld, C.L., Yao, K., Chen, K.M., Wong, A.K., Troyanskaya, O.G.: Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. Nature genetics **50**(8), 1171–1179 (2018)

[6] Ramprasad, P., Pai, N., Pan, W.: Enhancing personalized gene expression prediction from dna sequences using genomic foundation models. Human Genetics and Genomics Advances (2024)

[7] Drusinsky, S., Whalen, S., Pollard, K.S.: Deep-learning prediction of gene expression from personal genomes. bioRxiv, 2024–07 (2024)

[8] Zhao, H., Liu, T., Huang, T., Lin, Y., Ying, R.: Unicorn: Towards universal cellular expression prediction with an explainable multi-task learning framework (2025)

[9] Ji, Y., Mishra, R.K., Davuluri, R.V.: In silico analysis of alternative splicing on drug-target gene interactions. Scientific reports **10**(1), 134 (2020)

[10] Kim, S., Wysocka, J.: Deciphering the multi-scale, quantitative cis-regulatory code. Molecular cell **83**(3), 373–392 (2023)

[11] Ji, Y., Zhou, Z., Liu, H., Davuluri, R.V.: Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. Bioinformatics **37**(15), 2112–2120 (2021)

[12] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., *et al.*: Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38–45 (2020)

[13] Vaswani, A.: Attention is all you need. Advances in Neural Information Processing Systems (2017)

[14] Zhou, Z., Wu, W., Ho, H., Wang, J., Shi, L., Davuluri, R.V., Wang, Z., Liu, H.: Dnabert-s: Learning species-aware dna embedding with genome foundation models. ArXiv (2024)

[15] Lin, J., Luo, R., Pinello, L.: Epinformer: a scalable deep learning framework for gene expression prediction by integrating promoter-enhancer sequences with multimodal epigenomic data. bioRxiv, 2024–08 (2024)

[16] Kenton, J.D.M.-W.C., Toutanova, L.K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of naacL-HLT, vol. 1,

p. 2 (2019). Minneapolis, Minnesota

[17] Zhou, Z., Ji, Y., Li, W., Dutta, P., Davuluri, R.V., Liu, H.: Dnabert-2: Efficient foundation model and benchmark for multi-species genomes. In: The Twelfth International Conference on Learning Representations

[18] Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., *et al.*: The sequence of the human genome. science **291**(5507), 1304–1351 (2001)

[19] Fishman, V., Kuratov, Y., Shmelev, A., Petrov, M., Penzar, D., Shepelin, D., Chekanov, N., Kardymon, O., Burtsev, M.: Gena-lm: a family of open-source foundational dna language models for long sequences. bioRxiv, 2023–06 (2023)

[20] Siva, N.: 1000 genomes project. Nature biotechnology **26**(3), 256–257 (2008)

[21] Radford, A.: Improving language understanding by generative pre-training (2018)

[22] Zhang, D., Zhang, W., He, B., Zhang, J., Qin, C., Yao, J.: Dnagpt: a generalized pretrained tool for multiple dna sequence analysis tasks. bioRxiv, 2023–07 (2023)

[23] Nguyen, E., Poli, M., Durrant, M.G., Kang, B., Katrekar, D., Li, D.B., Bartie, L.J., Thomas, A.W., King, S.H., Brixi, G., *et al.*: Sequence modeling and design from molecular to genome scale with evo. Science **386**(6723), 9336 (2024)

[24] Poli, M., Massaroli, S., Nguyen, E., Fu, D.Y., Dao, T., Baccus, S., Bengio, Y., Ermon, S., Ré, C.: Hyena hierarchy: Towards larger convolutional language models. In: International Conference on Machine Learning, pp. 28043–28078 (2023). PMLR

[25] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y.J., Madotto, A., Fung, P.: Survey of hallucination in natural language generation. ACM Computing Surveys **55**(12), 1–38 (2023)

[26] Marin, F.I., Teufel, F., Horlacher, M., Madsen, D., Pultz, D., Winther, O., Boomsma, W.: Bend: Benchmarking dna language models on biologically meaningful tasks. In: The Twelfth International Conference on Learning Representations (2023)

[27] Benegas, G., Ye, C., Albors, C., Li, J.C., Song, Y.S.: Genomic language models: opportunities and challenges. Trends in Genetics (2025)

[28] Benegas, G., Batra, S.S., Song, Y.S.: Dna language models are powerful predictors of genome-wide variant effects. Proceedings of the National Academy of Sciences **120**(44), 2311219120 (2023)

[29] Schork, N.J., Fallin, D., Lanchbury, J.S.: Single nucleotide polymorphisms and the future of genetic epidemiology. Clinical genetics **58**(4), 250–264 (2000)

[30] Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J.R., Grabska-Barwinska, A., Taylor, K.R., Assael, Y., Jumper, J., Kohli, P., Kelley, D.R.: Effective gene expression prediction from sequence by integrating long-range interactions. Nature methods **18**(10), 1196–1203 (2021)

[31] Rehm, H.L., Berg, J.S., Brooks, L.D., Bustamante, C.D., Evans, J.P., Landrum, M.J., Ledbetter, D.H., Maglott, D.R., Martin, C.L., Nussbaum, R.L., *et al.*: Clingen—the clinical genome resource. New England Journal of Medicine **372**(23), 2235–2242 (2015)

[32] Li, Z., Subasri, V., Stan, G.-B., Zhao, Y., WANG, B.: GV-rep: A large-scale dataset for genetic variant representation learning. In: The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2024). https://openreview.net/forum?id=CW9SJyhpVt

[33] Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., *et al.*: Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS medicine **12**(3), 1001779 (2015)

[34] Li, B., Hou, Y., Che, W.: Data augmentation approaches in natural language processing: A survey. Ai Open **3**, 71–90 (2022)

[35] Marin, F.I., Teufel, F., Horlacher, M., Madsen, D., Pultz, D., Winther, O., Boomsma, W.: BEND: Benchmarking DNA language models on biologically meaningful tasks. In: The Twelfth International Conference on Learning Representations (2024). https://openreview.net/forum?id=uKB4cFNQFg

[36] Tang, Z., Somia, N., Yu, Y., Koo, P.K.: Evaluating the representational power of pre-trained dna language models for regulatory genomics. bioRxiv (2024)

[37] Feng, H., Wu, L., Zhao, B., Huff, C., Zhang, J., Wu, J., Lin, L., Wei, P., Wu, C.: Benchmarking dna foundation models for genomic sequence classification. bioRxiv (2024)

[38] Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)

[39] Theodoris, C.V., Xiao, L., Chopra, A., Chaffin, M.D., Al Sayed, Z.R., Hill, M.C., Mantineo, H., Brydon, E.M., Zeng, Z., Liu, X.S., *et al.*: Transfer learning enables predictions in network biology. Nature **618**(7965), 616–624 (2023)

[40] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., *et al.*: Scikit-learn: Machine learning in python. the Journal of machine Learning research **12**, 2825–2830 (2011)

[41] Luecken, M.D., Büttner, M., Chaichoompu, K., Danese, A., Interlandi, M., Müller, M.F., Strobl, D.C., Zappia, L., Dugas, M., Colomé-Tatché, M., *et al.*: Benchmarking atlas-level data integration in single-cell genomics. Nature methods **19**(1), 41–50 (2022)

[42] Dalla-Torre, H., Gonzalez, L., Mendoza-Revilla, J., Lopez Carranza, N., Grzywaczewski, A.H., Oteri, F., Dallago, C., Trop, E., Almeida, B.P., Sirelkhatim, H., et al.: Nucleotide transformer: building and evaluating robust foundation models for human genomics. Nature Methods, 1–11 (2024)

[43] Sanabria, M., Hirsch, J., Joubert, P.M., Poetsch, A.R.: Dna language model grover learns sequence context in the human genome. Nature Machine Intelligence **6**(8), 911–923 (2024)

[44] Schiff, Y., Kao, C.H., Gokaslan, A., Dao, T., Gu, A., Kuleshov, V.: Caduceus: Bi-directional equivariant long-range dna sequence modeling. In: Forty-first International Conference on Machine Learning

[45] Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C.L., Ma, J., *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), 2016239118 (2021)

[46] Liu, O., Jaghouar, S., Hagemann, J., Wang, S., Wiemels, J., Kaufman, J., Neiswanger, W.: Metagene-1: Metagenomic foundation model for pandemic monitoring. arXiv preprint arXiv:2501.02045 (2025)

[47] He, Y., Fang, P., Shan, Y., Pan, Y., Wei, Y., Chen, Y., Chen, Y., Liu, Y., Zeng, Z., Zhou, Z., et al.: Lucaone: Generalized biological foundation model with unified nucleic acid and protein language. bioRxiv, 2024–05 (2024)

[48] Linder, J., Srivastava, D., Yuan, H., Agarwal, V., Kelley, D.R.: Predicting rna-seq coverage from dna sequence as a unifying model of gene regulation. Nature Genetics, 1–13 (2025)

[49] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)

[50] Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)

[51] Kim, T., Yoo, K.M., Lee, S.-g.: Self-guided contrastive learning for BERT sentence representations. In: Zong, C., Xia, F., Li, W., Navigli, R. (eds.) Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing

(Volume 1: Long Papers), pp. 2528–2540. Association for Computational Linguistics, Online (2021). https://doi.org/10.18653/v1/2021.acl-long.197 . https://aclanthology.org/2021.acl-long.197/

[52] Fang, H., Wang, S., Zhou, M., Ding, J., Xie, P.: Cert: Contrastive self-supervised learning for language understanding. arXiv preprint arXiv:2005.12766 (2020)

[53] McInnes, L., Healy, J., Melville, J.: Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426 (2018)

[54] Chen, S., Francioli, L.C., Goodrich, J.K., Collins, R.L., Kanai, M., Wang, Q., Alföldi, J., Watts, N.A., Vittal, C., Gauthier, L.D., *et al.*: A genomic mutational constraint map using variation in 76,156 human genomes. Nature **625**(7993), 92–100 (2024)

[55] Liu, T., Chen, T., Zheng, W., Luo, X., Zhao, H.: scelmo: Embeddings from language models are good learners for single-cell data analysis. bioRxiv, 2023–12 (2023)

[56] Zeng, W., Chen, S., Cui, X., Chen, X., Gao, Z., Jiang, R.: Silencerdb: a comprehensive database of silencers. Nucleic acids research **49**(D1), 221–228 (2021)

[57] Gamazon, E.R., Wheeler, H.E., Shah, K.P., Mozaffari, S.V., Aquino-Michaels, K., Carroll, R.J., Eyler, A.E., Denny, J.C., Consortium, G., Nicolae, D.L., *et al.*: A gene-based association method for mapping traits using reference transcriptome data. Nature genetics **47**(9), 1091–1098 (2015)

[58] Wheeler, H.E., Shah, K.P., Brenner, J., Garcia, T., Aquino-Michaels, K., Consortium, G., Cox, N.J., Nicolae, D.L., Im, H.K.: Survey of the heritability and sparse architecture of gene expression traits across human tissues. PLoS genetics **12**(11), 1006423 (2016)

[59] Safran, M., Rosen, N., Twik, M., BarShir, R., Stein, T.I., Dahary, D., Fishilevich, S., Lancet, D.: The genecards suite. Practical guide to life science databases, 27–56 (2021)

[60] Harel, A., Inger, A., Stelzer, G., Strichman-Almashanu, L., Dalah, I., Safran, M., Lancet, D.: Gifts: annotation landscape analysis with genecards. BMC bioinformatics **10**, 1–11 (2009)

[61] Gao, Y., Myers, S., Chen, S., Dligach, D., Miller, T., Bitterman, D., Churpek, M., Afshar, M.: When raw data prevails: Are large language model embeddings effective in numerical data representation for medical machine learning applications? In: Findings of the Association for Computational Linguistics: EMNLP 2024, pp. 5414–5428 (2024)

[62] Bennett, D.A., Buchman, A.S., Boyle, P.A., Barnes, L.L., Wilson, R.S., Schneider, J.A.: Religious orders study and rush memory and aging project. Journal of Alzheimer's disease **64**(s1), 161–189 (2018)

[63] Davidson, E., Levin, M.: Gene regulatory networks. Proceedings of the National Academy of Sciences **102**(14), 4935–4935 (2005)

[64] Cordell, H.J.: Detecting gene–gene interactions that underlie human diseases. Nature Reviews Genetics **10**(6), 392–404 (2009)

[65] Aleksander, S.A., Balhoff, J., Carbon, S., Cherry, J.M., Drabkin, H.J., Ebert, D., Feuermann, M., Gaudet, P., Harris, N.L., *et al.*: The gene ontology knowledgebase in 2023. Genetics **224**(1), 031 (2023)

[66] Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., *et al.*: Gene ontology: tool for the unification of biology. Nature genetics **25**(1), 25–29 (2000)

[67] Hu, E.J., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. In: International Conference on Learning Representations

[68] Kingma, D.P.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

[69] Loshchilov, I.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)

[70] Jordan, K., Jin, Y., Boza, V., Jiacheng, Y., Cecista, F., Newhouse, L., Bernstein, J.: Muon: An optimizer for hidden layers in neural networks (2024). https://kellerjordan.github.io/posts/muon/

[71] Huber, P.J.: Robust estimation of a location parameter. In: Breakthroughs in Statistics: Methodology and Distribution, pp. 492–518. Springer, ??? (1992)

[72] Dao, T., Fu, D.Y., Ermon, S., Rudra, A., Ré, C.: FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In: Advances in Neural Information Processing Systems (NeurIPS) (2022)

[73] Dao, T.: FlashAttention-2: Faster attention with better parallelism and work partitioning. In: International Conference on Learning Representations (ICLR) (2024)

[74] Hingerl, J.C., Karollus, A., Gagneur, J.: Flashzoi: An enhanced borzoi model for accelerated genomic analysis. bioRxiv, 2024–12 (2024)

[75] Aguet François 1 Brown Andrew A. 2 3 4 Castel Stephane E. 5 6 Davis Joe R. 7 8 He Yuan 9 Jo Brian 10 Mohammadi Pejman 5 6 Park YoSon 11 Parsana Princy 12

Segrè Ayellet V. 1 Strober Benjamin J. 9 Zappala Zachary 7 8, G.C.L., Addington Anjene 15 Guan Ping 16 Koester Susan 15 Little A. Roger 17 Lockhart Nicole C. 18 Moore Helen M. 16 Rao Abhi 16 Struewing Jeffery P. 19 Volpi Simona 19, N., 16, P.S.L..B.M.E..B.P.A., 137, N.C.F.N.C.R., *et al.*: Genetic effects on gene expression across human tissues. Nature **550**(7675), 204–213 (2017)

[76] Souza, N.: The encode project. Nature methods **9**(11), 1046–1046 (2012)

[77] Consortium, G.: The gtex consortium atlas of genetic regulatory effects across human tissues. Science **369**(6509), 1318–1330 (2020)

[78] Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society Series B: Statistical Methodology **67**(2), 301–320 (2005)

[79] Sasse, A., Ng, B., Spiro, A.E., Tasaki, S., Bennett, D.A., Gaiteri, C., De Jager, P.L., Chikina, M., Mostafavi, S.: Benchmarking of deep neural networks for predicting personal gene expression from dna sequence highlights shortcomings. Nature Genetics **55**(12), 2060–2064 (2023)

[80] Han, Z., Gao, C., Liu, J., Zhang, J., Zhang, S.Q.: Parameter-efficient fine-tuning for large models: A comprehensive survey. arXiv preprint arXiv:2403.14608 (2024)

[81] Lal, A., Gunsalus, L., Nair, S., Biancalani, T., Eraslan, G.: grelu: A comprehensive framework for dna sequence modeling and design. bioRxiv, 2024–09 (2024)

[82] Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., *et al.*: Scipy 1.0: fundamental algorithms for scientific computing in python. Nature methods **17**(3), 261–272 (2020)

[83] Huang, K., Zeng, T., Koc, S., Pettet, A., Zhou, J., Jain, M., Sun, D., Ruiz, C., Ren, H., Howe, L.J., et al.: Small-cohort gwas discovery with ai over massive functional genomics knowledge graph. medRxiv, 2024–12 (2024)

[84] Li, X., Li, J.: AoE: Angle-optimized embeddings for semantic textual similarity. In: Ku, L.-W., Martins, A., Srikumar, V. (eds.) Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1825–1839. Association for Computational Linguistics, Bangkok, Thailand (2024). https://doi.org/10.18653/v1/2024.acl-long.101 . https://aclanthology.org/2024.acl-long.101

[85] Wang, G., Sarkar, A., Carbonetto, P., Stephens, M.: A simple new approach to variable selection in regression, with application to genetic fine mapping. Journal of the Royal Statistical Society Series B: Statistical Methodology **82**(5), 1273–1300 (2020)