Gao *et al. Genome Biology* (2024) 25:310
https://doi.org/10.1186/s13059-024-03449-7

Genome Biology

METHOD

Open Access

EpiGePT: a pretrained transformer-based language model for context-specific human epigenomics

Zijing Gao[1†], Qiao Liu[2*†], Wanwen Zeng[2], Rui Jiang[1*] and Wing Hung Wong[2,3*]

†Zijing Gao and Qiao Liu contributed equally to this work.

*Correspondence: liuqiao@stanford.edu; ruijiang@tsinghua.edu.cn; whwong@stanford.edu

[1] Ministry of Education Key Laboratory of Bioinformatics, Bioinformatics Division at the Beijing National Research Center for Information Science and Technology, Center for Synthetic and Systems Biology, Department of Automation, Tsinghua University, Beijing 100084, China
[2] Department of Statistics, Stanford University, CA, Stanford 94305, USA
[3] Department of Biomedical Data Science, Bio-X Program, Center for Personal Dynamic Regulomes, Stanford University, Stanford, CA 94305, USA

Abstract

The inherent similarities between natural language and biological sequences have inspired the use of large language models in genomics, but current models struggle to incorporate chromatin interactions or predict in unseen cellular contexts. To address this, we propose EpiGePT, a transformer-based model designed for predicting context-specific human epigenomic signals. By incorporating transcription factor activities and 3D genome interactions, EpiGePT outperforms existing methods in epigenomic signal prediction tasks, especially in cell-type-specific long-range interaction predictions and genetic variant impacts, advancing our understanding of gene regulation. A free online prediction service is available at http://health.tsinghua.edu.cn/epigept.

Keywords: Epigenomics, Gene Regulation, 3D genome, Transformer, Language model

Background

A fundamental but largely unresolved problem in genomics is to decode the information residing in the non-coding part of the human genome. It remains incompletely understood how regulatory elements govern gene expression in different contexts [1], and how noncoding variants may disrupt the underlying regulatory syntax of DNA [2]. Fortunately, recent advances in epigenome sequencing [3, 4] have resulted in the accumulation of data useful for the study of these questions, including chromatin accessibility, DNA methylation, histone modifications, and 3D chromatin interaction. Thus, there is great interest in performing systematic analysis of these data to enhance our ability to interpret the non-coding part of the genome [5–11].

The inherent similarities between natural language and biological sequences have also stimulated interest in developing large language models (LLM) for the interpretation of genome sequences. As is well known, the development of LLM has been the main driving force behind many recent breakthroughs in artificial intelligence such as ChatGPT, leading to numerous applications in bioinformatics [12, 13]. The architecture of

the LLM is a multilayer transformer network, and the model is trained on a very large corpus of natural language data. Such pre-trained models can be readily tailored or adapted to various downstream tasks. Considering DNA sequences as the texts in the genomic language, similar transformer-based approaches have been used to model DNA sequences [14, 15]. For example, the Enformer model [16] takes the DNA sequence of a large genomic region as input and predict thousands of epigenomic features across cellular contexts covered by the training data. Although already useful in many applications, such models relying on only DNA sequences as input are not capable of predicting the function of sequences in new cellular contexts. Furthermore, despite the importance of 3D chromatin contacts in gene regulation, 3D interaction data have not been included in the training of current genomic LLMs. Therefore, there is an urgent need to further develop the core technologies of genomic LLMs to overcome these limitations.

In this paper, we present EpiGePT, a transformer-based model for human epigenomics prediction with the following new capabilities. First, the inability to make predictions in novel contexts has greatly limited the applicability of current methods, EpiGePT removes this limitation by making both the input and output context-dependent, where the context is represented by a TF-profile vector specifying the expression of key transcription factors (TFs) in that context. This choice is motivated by the fact that reference gene expression data are available for many cellular contexts that are important in development and diseases, but for which few epigenomic features have been measured. We note that the reference TF expression profile has been used to represent cellular context in earlier works on accessibility prediction [17, 18], but this idea has not been explored for the development of genomic LLMs. Second, a new learning algorithm is developed to enable the inclusion of 3D chromatin contact data in the training data. In this way, EpiGePT can predict 3D genome features such as enhancer-promoter interactions that are known to be important for gene regulation but are not modeled in current genomic LLMs. Besides, by using a masked training strategy, EpiGePT can be trained on a diverse set of contexts even if different sets of epigenomics signals are available in different contexts.

There is a profound difference in training strategy between EpiGePT and current genomic LLMs. Each input genomic region provides an example for training in current LLMs such as the Enformer. In contrast, each combination of input region and cellular context provides an example for training in EpiGePT, thus providing a much larger number of examples available for model training. As for training data sets, since most cellular contexts that have epigenomic data will also have expression data, we can use most available epigenomic data, such as those used by the Enformer, to train our model. In a series of experiments, we illustrate that our model is superior to existing methods in epigenomic signal prediction, long-range chromatin interaction prediction, and the variant effect prediction.

Results

Overview of EpiGePT

EpiGePT is a genomic language model for cross-cell-type prediction of chromatin states by multi-task learning based on genome-wide pre-training on epigenomic data (Fig. 1 and Additional file 1: Fig. S1). The model is composed of four modules, including a

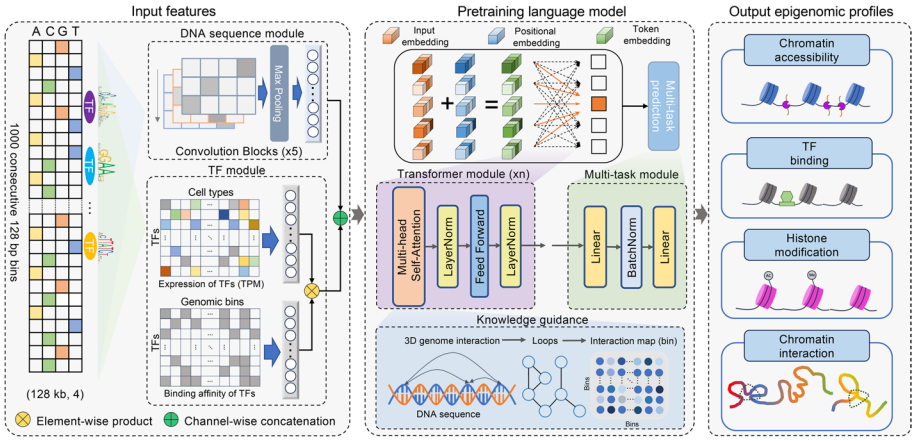

Fig. 1 Overview of EpiGePT model for multiple epigenomic signals prediction. The EpiGePT model consists of four modules, namely the Sequence module, the TF module, the Transformer module, and the Multi-task prediction module. The sequence module comprises multiple layers of convolution applied to the one-hot encoded DNA sequence input. The input sequence length consists of 1000 genomic bins of 128 bp for the prediction of multiple signals and 50 bins of 200 bp for the prediction of DNase signal alone. The TF module encompasses the binding status and expression of 711 transcription factors. The Transformer module consists of a series of consecutive transformer encoders, while the multi-task module is composed of a fully connected layer. Additionally, the EpiGePT framework integrates an optional knowledge guidance module that enhances the interpretability of the model by incorporating three-dimensional chromatin interaction data into the attention layer, thus improving its understanding of regulatory mechanisms

sequence module, a TF module, a transformer module, and a prediction module. The sequence module is responsible for processing the long DNA sequence of interest (e.g., 128 kb) by employing a series of convolutional and pooling blocks (e.g., 5) to extract a comprehensive set of sequence features. The TF module is specifically designed to represent a cellular context by a TF-profile vector, which specifies the state of a few hundred TFs in that context. The features computed by the sequence and TF modules are then fed as input tokens to the transformer module, where each token corresponds to a genomic bin (e.g., a 128-bp window) in the original DNA sequence. The transformer module leverages self-attention mechanisms to learn the relationships among the input bins, enabling the model to make predictions of multiple chromatin states given the context information from the TF module. Importantly, by including a novel loss term that involves the self-attention weights, EpiGePT is capable of learning from data on context-specific chromatin interactions. Since 3D interaction is known to be a key mechanism in gene regulation, the ability to learn from interaction data is an attractive feature of our approach. Finally, the fourth module in EpiGePT is a predictive module which predicts epigenomic signals and chromatin interactions based on the output of the transformer module.

Genome-wide prediction of epigenomic signals

To assess the performance on predicting epigenomic signals, we first compared EpiGePT to task-specific models that are specifically designed for predicting a single epigenomic signal. Taking the chromatin accessibility for instance, the performance of EpiGePT was compared against existing task-specific models such as BIRD [19], ChromDragoNN [17], and DeepCAGE [18]. The widely available public DNase-seq [20] data across 129

cellular contexts on 1,175,374 genomic regions were collected and preprocessed from ENCODE project [21] (see "Methods"). Performance is evaluated in three prediction settings: (i) "cross-region" setting where the predictive model is tested on new genomic regions not seen in training, (ii) "cross-cell type" setting where the model is tested on new cell types, and (iii) "cross-both" setting where testing is done on new regions in new cell types (Additional file 1: Fig. S2, Text S1). In each setting, we employed three evaluation metrics, namely Pearson correlation coefficient (PCC), Spearman correlation coefficient (SCC), and prediction square error (PSE), to assess the similarity between the predicted and true values of the DNase signals (see "Methods"). The results, presented in Fig. 2a and Additional file 1: Fig. S3-S4, showed that EpiGePT consistently outperformed

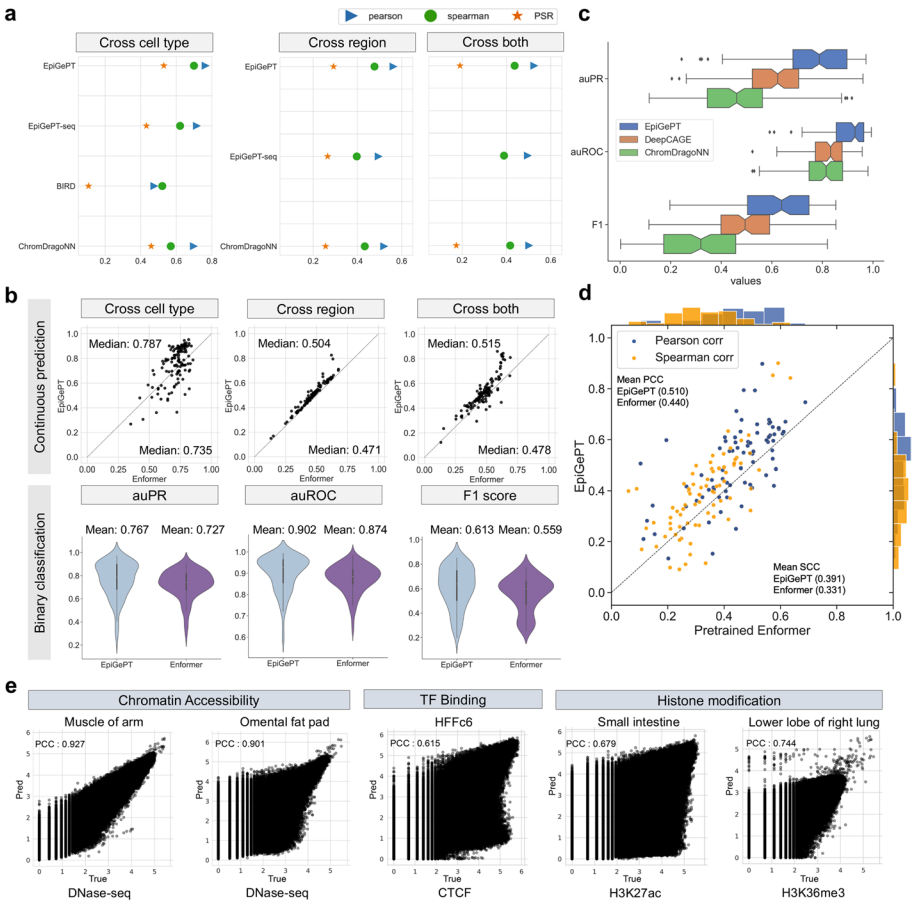

Fig. 2 Performance of EpiGePT and baseline methods on the benchmark experiment. **a** EpiGePT and baseline methods were compared in terms of their regression performance for DNase signal regression across cell types, genomic regions, and combined cell type and genomic regions. **b** Comparison of EpiGePT and Enformer performance. Each point in the scatter plot represents the performance of Enformer on the data of a specific cell type (*x*-axis) compared to the performance of EpiGePT (*y*-axis). The top three graphs represent the prediction of continuous DNase signals (PCC), while the bottom three graphs represent the binary classification of chromatin accessibility regions. **c** EpiGePT and baseline methods' performance on binary prediction of DNase-seq signals. **d** EpiGePT demonstrates more excellent performance in predicting diverse epigenetic signals across various cell types, compared with the pre-trained Enformer on 78 genomic tracks across 19 unseen cell types. The orange points represent SCC, and the blue points represent PCC. **e** EpiGePT cross-cell-type predictions compared to experimental signals visualized for a representative example. The predictions specific to DNase are based on the hg19 reference genome, while predictions for multiple epigenomic profiles are conducted using the hg38 reference genome

baseline methods including BIRD [19], and ChromDragoNN [17] by a relatively large margin under the above settings. For example, EpiGePT achieved a cross-cell type prediction PCC of 0.787, demonstrated a 6.9% higher performance than the best baseline method, ChromDragoNN. In addition, we also evaluated the prediction of binary chromatin accessibility status, i.e., predicting whether a peak exists within the corresponding genomic bin (>50% overlap). For binary prediction, EpiGePT again achieved a superior performance with an average auPRC (area under the precision-recall curve) of 0.767 compared to 0.623 of DeepCAGE [18] and 0.476 of ChromDragoNN [17] (Fig. 2c). Finally, we compared EpiGePT and ChromDragoNN [17] in the binary classification of functional regions versus nonfunctional regions, using the functional chromatin status derived from ChromHMM [22] annotations as ground truth (Additional file 1: Text S2). EpiGePT achieved an average 8.1% higher auROC (area under the receiver operating characteristic curve) than ChromDragoNN [17], and an average 2.3% higher macro-auROC than ChromDragoNN [17] (p-value < 0.001 under one-sided Wilcoxon signed rank test) in a finer-grained classification for different types of regulatory elements (Additional file 1: Fig. S5a-b). We also compare the consistency and discrepancy of different methods by treating the predicted label of one method as ground truth and evaluate on other methods (Additional file 1: Text S3, Fig. S6). These results demonstrate that EpiGePT provides better predictions than task-specific models.

Next, we compared EpiGePT with a state-of-the-art genomic LLM, Enformer [16], in two different ways. First, we trained an Enformer model from scratch with only the aforementioned DNase-seq data (Additional file 1: Text S4, Fig. S7). EpiGePT demonstrates a 3.3 to 5.2% higher performance than Enformer in terms of the median PCC under the three prediction settings (Fig. 2b). Additionally, EpiGePT also outperforms a retrained Enformer model from scratch with eight different epigenomic signals, achieving a 12.3% improvement in the average PCC across all genomic tracks (Additional file 1: Fig. S8a-b). Second, we compared EpiGePT directly to the pretrained Enformer model provided by the original paper. To do this, we collected eight different epigenomic signals from 104 different cellular contexts (see "Methods"). We first left out 13 of these contexts where HiChIP data are also available for downstream chromatin interactions validation. Then, EpiGePT model was trained across 72 training cellular contexts (without using HiChIP-based chromatin contacts data in the training) and subsequently compared against pre-trained Enformer on the remaining 19 test cellular contexts, on 15,870 training genomic regions with 128 kbp length. Since most of the cellular contexts have missing epigenomic signals, we designed a masked training strategy to handle this issue (see "Methods"). Under the test cellular contexts, EpiGePT exhibited superior performance with higher PCC than Enformer in 60 out of 78 matched epigenomic signals across 19 test cellular contexts by achieving an average PCC of 0.510, compared to 0.440 of Enformer (Fig. 2d and Additional file 1: Fig. S8c-d). For DNase-seq specifically, the average PCC of EpiGePT reached 0.710 and the average SCC reached 0.664 across 7 cell types, compared to the average PCC of 0.455 and the average SCC of 0.488 of Enformer. In the above comparison, we are in fact comparing out-sample prediction by EpiGePT to in-sample prediction by Enformer. The favorable results achieved by EpiGePT in this experimental setting suggest that our model enables prediction in novel contexts without sacrificing performance. To illustrate the prediction performance further, several

tracks of predicted chromatin states and the corresponding ground truth chromatin states are displayed in Fig. 2e.

We further explored EpiGePT's applicability to other species using zero-shot learning and fine-tuning strategies. Given the high conservation of TFs between mouse and human [23], we adapted EpiGePT to mouse data, identifying 688 overlapping TFs and setting the expression of non-overlapping TFs to zero. Using chromatin accessibility data from three organ-level mouse datasets (brain, lung, kidney) as examples, we employed OpenAnnotate [24, 25] to annotate 247,750 peaks for brain data, 243,900 peaks for lung data, and 180,400 peaks for kidney data, respectively. First, in the zero-shot experiments, EpiGePT demonstrated PCC ranging from 0.330 to 0.422 (Additional file 1: Fig. S9). For example, the brain dataset achieved a PCC of 0.422, and a SCC of 0.379 on the test chromosomes (chromosomes 18, 19, X, and Y). Following fine-tuning, the model's predictive accuracy improved significantly, even with limited data, with PCC increases of 18.1, 18.4, and 19.5% observed in brain, kidney, and lung datasets, respectively (Additional file 1: Fig. S10a, S10c). We further explored the EpiGePT's cross-cell type prediction capabilities on mouse data. The model fine-tuned on lung data achieved a PCC of 0.490 on brain data, and the kidney-fine-tuned model reached a PCC of 0.448, both outperforming the zero-shot predictions (Additional file 1: Fig. S10b). Additionally, we demonstrated the importance of TF module in a fine-tuning setting on mouse data (Additional file 1: Fig. S11). These comprehensive experiments highlight the pretrained EpiGePT model's substantial potential for cross-species applications.

EpiGePT enables the prediction of chromatin interactions

We examined the capacity of EpiGePT for predicting long-range chromatin interactions, which is important for understanding chromatin architecture and relations between regulatory elements and target genes. We employed several experimental settings to examine the ability of EpiGePT in capturing long-range chromatin interactions. In setting (A), we directly utilized the self-attention weights extracted from the pretrained EpiGePT model (without including HiChIP data in the training) to predict enhancer-promoter (E-P) interactions and silencer-promoter (S-P) interactions. In setting (B), we integrated HiChIP-derived 3D chromatin contacts into the training of the model and then use the model to predict E-P interactions in novel contexts not seen in the training. In setting (C), we designed a pretrain-finetune strategy for EpiGePT model to predict E-P interactions. The results under each setting are discussed below.

Setting (A): prediction by EpiGePT not trained with 3D contact data. In this setting, we use the cell-type specific self-attention scores to predict chromatin interactions, including E-P and S-P interactions (see "Methods"). Two sets of interactions containing 664 and 5091 candidate element-gene interactions, obtained by CRISPRi [26] experiments on K562 cell line, were collected and further filtered and divided into positive and negative samples, for use as ground truths to evaluate E-P prediction performance. In the Gasperini et al. [27] dataset, EpiGePT consistently outperformed Enformer by achieving the highest auPRC in most cases (Fig. 3a). For instance, EpiGePT achieved auPRC of 0.647 to 0.887 for identifying enhancer-gene transcription start site (TSS) pairs in different distance groups (Fig. 3a and Additional file 1: Fig. S12a-b). In the Fulco et al. [28] dataset, EpiGePT also outperformed other competing methods. For example,

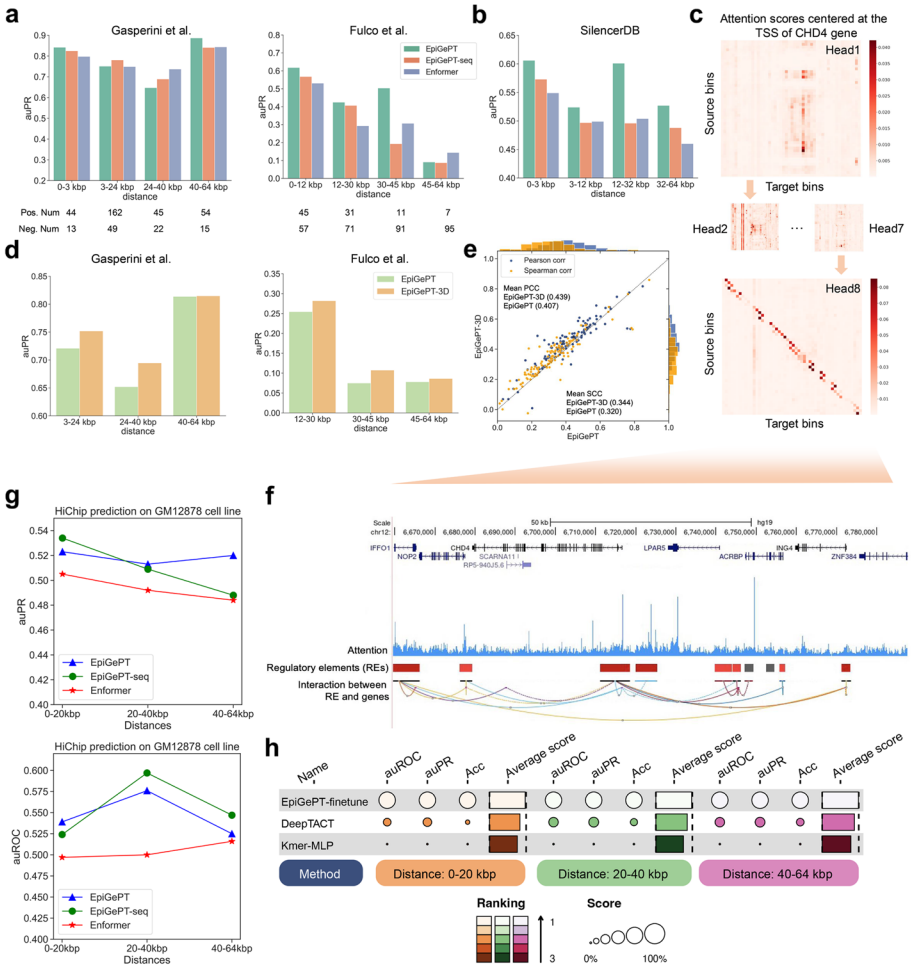

Fig. 3 Application of self-attention mechanism in EpiGePT for long-range chromatin interaction identification. **a** The performance (auPRC) of attention score of EpiGePT in distinguishing enhancer-gene pairs at different distance ranges on two different datasets. **b** The performance (auPRC) of attention score of EpiGePT in distinguishing silencer-gene pairs at different distance ranges based on the data from SilencerDB [29]. **c** Heatmap of the self-attention matrix of each attention head centered at the TSS of the *CHD4* gene, the (*i*, *j*) element in the matrix denotes the average attention score between the *i*th genomic bin and the *j*th genomic bin across all layers. **d** The performance (auPR) of self-attention scores of EpiGePT and EpiGePT-3D in identifying enhancer-promoter interactions across different distance ranges on the K562 cell type. **e** The predictive performance (blue points denote PCCs and orange points denote SCCs) of EpiGePT with knowledge guidance across 19 cell types and 15,870 long sequences (128 kbp). **f** Attention scores centered at the TSS of the *CHD4* gene, and putative enhancer regions in its vicinity. **g** The performance (auROC and auPR) of attention score of EpiGePT in distinguishing HiChIP loops of H3K27ac at different distance ranges on GM12878 cell line. **h** The performance table (auROC and auPRC) of the fine-tuned EpiGePT model and baseline methods (DeepTACT and Kmer-MLP) in distinguishing enhancer-gene pairs at various distance ranges (0–20 kbp, 20–40 kbp, and 40–64 kbp) on K562 cell line under a fivefold cross validation setting. The size of the circles represents the rank of different methods, with larger circles indicating higher performance and a better rank. A score of 0% represents the lowest-performing method, while 100% represents the highest-performing method. The sizes of the remaining circles are determined based on their percentage scores. The color also reflects the rank, with lighter colors indicating better performance. The average score bar reflects the mean of the three metrics

EpiGePT achieves an auPRC of 0.504, compared to 0.307 of Enformer in the 30–45 kbp group (Fig. 3a). Next, to assess performance on S-P interactions, we downloaded putative silencers from the SilencerDB [29] and used the TSS of annotated nearest gene as the potential target. We selected the same number of negative pairs randomly while

conserving the distance distribution. The results show that EpiGePT achieved a better performance in distinguishing positive S-P pairs from negative pairs than Enformer. For instance, EpiGePT achieves an auROC of 0.575 in long-range S-P interactions (32–64 kbp) compared to 0.483 of Enformer (Fig. 3b, Additional file 1: Fig. S12c). Finally, to assess performance in predicting chromatin interactions, we collected HiChIP [30] loops on K562 and GM12878 cell lines from the HiChIPdb [31]. EpiGePT achieves a superior performance by discerning HiChIP loops from randomly selected loops with the same distance distribution. For instance, EpiGePT achieves an auPRC of 0.520 for long range loops (40–64 kbp) prediction in GM12878 cell line, surpassing that of Enformer (0.484) by a large margin (Fig. 3g, Additional file 1: Fig. S12d). These results clearly demonstrated the utility of EpiGePT attention scores in capturing functional chromatin interactions.

To better understand the self-attention mechanism of EpiGePT, we showed the attention weights (averaged across heads) for the bin containing the TSS of the gene *CHD4*. The attention weights were computed based on the pretrained EpiGePT model with K562 cell line as the context of interest. We also display chromatin interactions detected under K562 as well as regulatory element annotations from the GeneHancer [32]. It is seen that both the interaction data and the regulatory element annotations are consistent with the attention weights learned by EpiGePT (Fig. 3c and Fig. 3f).

Setting (B): Prediction by EpiGePT-3D, which include Hi-C data in its training. The above results suggest that in a good transformer-based genomic language model, the attention weight given by one bin to another bin (within the input region) should be consistent with the strength of 3D interaction between them. Thus, when experimental data on 3D interaction are available, we can leverage this data to improve the learning of the parameters of our genomic language model, by penalizing parameter values that resulted in poor correlation between the attention weights and the interaction data (see "Methods"). To obtain such training data, we collected 4,107,687 H3K27ac-based HiChIP loops across 13 cell lines or tissues from HiChIPdb [31], which denote potential E-P interactions. Setting aside loops from K562 and GM12878 cell lines as test data, other HiChIP loops are incorporated into the training. The resulting model is denoted as EpiGePT-3D. We found that adding 3D interaction data in the training can lead to a noticeable improvement for cross-cell-type prediction (3.3% higher in PCC) (Fig. 3e). Moreover, EpiGePT-3D demonstrated improved predictive performance on E-P interactions identified by HiChIP loops in new cellular contexts (Fig. 3d, Additional file 1: Fig. S13). For instance, the auPRC increased from 0.652 to 0.695 for Gasperini et al.'s dataset, which is on a context not covered by the Hi-C data in the training, in 24–40 kbp group when incorporating 3D genome data.

Setting (C): Prediction by fine-tuning pretrained EpiGePT. Fine-tuning is an strategy that transfers the knowledge of a pretrained model to new tasks, which is particularly prevalent in language models such as GPT [33] and BERT [34]. Here, we explore the performance of fine-tuning given a pretrained EpiGePT model on downstream tasks, such as predicting 3D genome interaction. Specifically, we fixed the weights of the pretrained EpiGePT model and trained an additional finetune network utilizing the last hidden states of the Transformer module for predicting E-P interactions. We compared EpiGePT with fine-tuning strategy (EpiGePT-finetune) to two baselines, DeepTACT

[35] and a k-mer frequency-based method[29] with HiChIP H3K27ac loops from K562 and GM12878 cell lines (see "Methods"). The results illustrate that EpiGePT-finetune exhibited a superior classification performance across diverse distance ranges and positive–negative sample ratios compared to baselines (Additional file 2: Table S1-S4). For example, EpiGePT-finetune achieved an average auROC of 0.982, surpassing 0.886 of DeepTACT [35] and 0.694 of Kmer by a large margin in the GM12878 cell line within the 0–20 kbp distance range at a positive–negative sample ratio of 1:1 (Fig. 3h, Additional file 1: Fig. S14-S15). This significant improvement demonstrates the power of fine-tuning a base pretrained genomic language model on a downstream task with limited data. In addition, we utilized the hg38 genome and evaluated the performance of EpiGePT-3D, EpiGePT after fine-tuning, with baseline methods. The results indicate that EpiGePT-3D achieved a slight overall advantage in predicting chromatin interactions under the same training conditions (Additional file 1: Fig. S16-S17). Furthermore, EpiGePT-3D also slightly outperformed the pretrained Enformer (Additional file 1: Fig. S18) in predicting epigenomic tracks, further validating the improvements brought by the 3D genome data.

In summary, we designed three settings for chromatin interaction prediction and have discussed and summarized the details and differences of each setting in relation to various application scenarios (Additional file 1: Text S5, Additional file 2: Table S5). We also validated the improvements brought by incorporating 3D genome data, suggesting that EpiGePT-3D pretrained on larger datasets has significant potential as more comprehensive 3D genome data become available.

EpiGePT unveils the regulatory relationships between TFs and target genes

In this section, we further explored the TF module to see whether EpiGePT is able to learn the regulatory relationships between TFs and target genes (TGs). We defined gradient importance scores (GIS) based on the absolute gradient values of predicted epigenomic signals with respect to the expression of a TF in the input TF profile, to rank the TFs for their potential to regulate a given TG (see "Methods"). Particularly, we use the TF profile of embryonic stem cell (ESC) to specify the context in the EpiGePT model. We selected the important ESC regulator *POU5F1* as the target gene and calculated the GIS for identifying TF-TF interactions (see "Methods"). Multiple potential regulators for POU5F1 identified by EpiGePT in ESC context are consistent with literatures, such as *ESRRB-POU5F1* [36] (rank 2nd), and *ETV5-POU5F1* [37] (rank 5th). Next, we focus on *ESRRB* which plays essential role for balancing pluripotency of ESCs [38]. Treating ESRRB as the target gene, our GIS-based ranking identified several key TFs, such as *POU5F1* and *REST*, that have significantly higher ranks than other TFs (Fig. 4a, Additional file 1: Fig. S19a). By using ChIP-seq data of *POU5F1* for validation, we observed significantly higher GIS in bins overlapping with the ChIP-seq data (Additional file 1: Fig. S19b, p-value < 0.00018 under one-sided Mann–Whitney U test). Next, we visualized the TF ranks obtained from eight epigenomic profiles across 1000 bins surrounding the TSS of *ESRRB*. By averaging ranks across these signals and bins among all the 711 TFs, the important ESC regulator *POU5F1* ranks 3 out of 711 (Fig. 4b). We further collected the top 5% of TFs for each bin and conducted gene ontology (GO) enrichment analysis based on these TF coding genes. Interestingly, the GO terms enriched also included biological processes of embryonic cell differentiation and development.

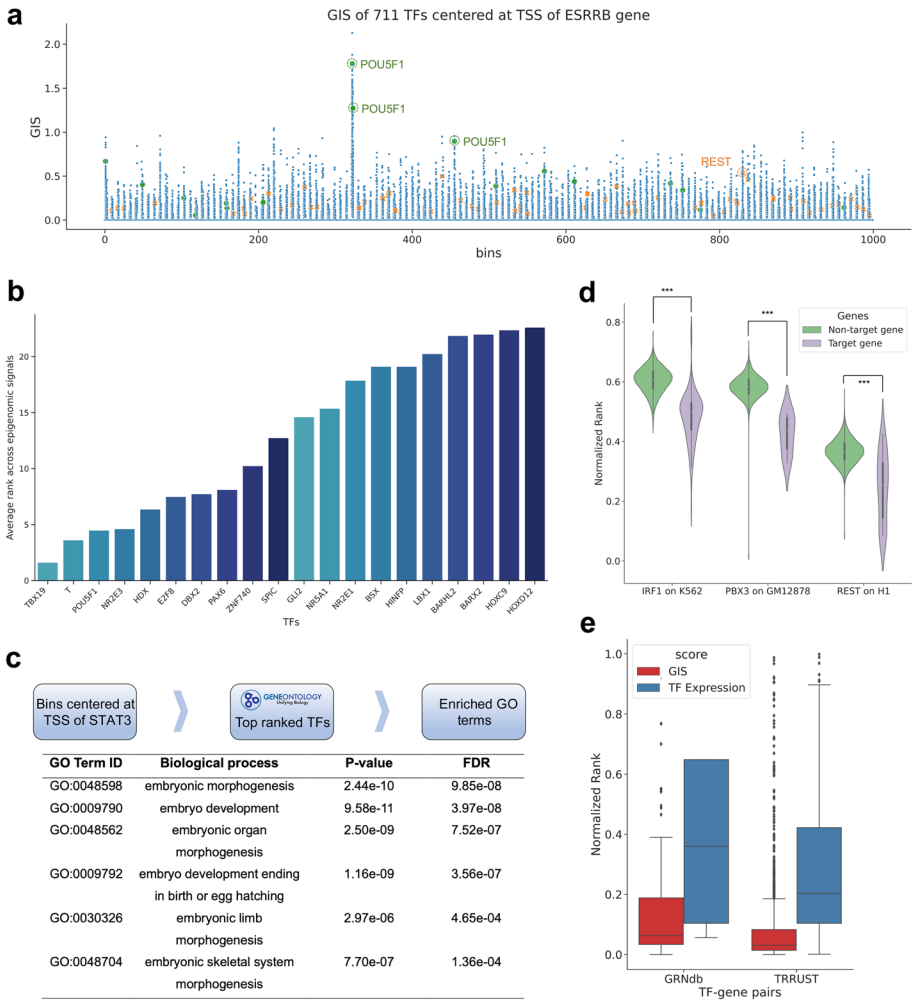

Fig. 4 Gradient importance scores (GIS) uncover regulatory transcription factors. **a** Genomic regions around TSS of the *ESRRB* gene and TF expression data on ESC were used in EpiGePT. The scatter plot represents the GIS scores of 711 TFs on each genomic bin. Each dot represents the GIS score of a core TF on a specific genomic bin. Two important ESC regulators *REST* and *POU1F5* are highlighted. **b** Bar plot of the top 5% ranked TFs, based on the average ranks from the GIS of eight epigenomic signals across bins (below). **c** Based on the top 5% ranked TFs in 128-kbp region centered at TSS of the *ESRRB* gene, gene ontology enrichment analysis revealed significant enrichment in biological processes related to embryonic development and cellular differentiation. **d** Based on TF ChIP-seq data, all 23,635 human genes were classified into target genes and non-target genes. The results revealed that TFs exhibited significantly higher ranks on potential target genes compared to non-target genes. **e** The distribution of the rank of TFs in the GIS and expression value among the 2705 TF-gene pairs from the TRRUST database and 1066 TF-gene pairs derived from genotype-tissue expression (GTEx) data of the liver sourced from the GRNdb database

However, using the top 5% of TFs with high expression in ESCs resulted in lower significance for biological processes associated with embryonic cell development (Fig. 4c and Additional file 1: Fig. S20), which again demonstrates the effectiveness of the GIS-based ranking. Furthermore, we use TF-TG relationships from either ChIP-seq data or external databases as ground truth to validate the TF-TG relationships inferred by EpiGePT. We defined potential TF-target gene pairs based on TF ChIP-seq data specific to certain cell types among all human genes (see "Methods"). The results demonstrated a significant higher rank of TF-target gene pairs, compared to TF-non-target gene pairs based

on the integrated GIS-based ranking (Fig. 4d, p-value < 0.001 under one-sided Mann–Whitney U test). Second, we collected TF-TG regulatory network data from two publicly available databases. We obtained a total of 1066 TF-TG pairs from the GRNdb [39] database based on liver-specific GTEx data, and 2705 TF-TG pairs from the TRRUST [40] database after filtering. Then we calculated the rank of each TF based on either integrated GIS or the TF expression value by using the liver expression as the TF reference profile. Interestingly, we found that the median ranking percentile of TFs from TRRUST was 3.1%, significantly higher than the percentile of 20.4% based on expression values (Fig. 4e, p-value < 1e − 5 under one-sided Wilcoxon signed rank test). With a similar result was obtained using another database GRNdb, where EpiGePT is seen to achieve a median ranking percentile of 6.3%, compared to 36.0% by gene expression value. For instance, *TMEM55B*, which plays a significant role in lysosome movement, and is regulated by sterol response element binding factor 2 (*SREBF2*) [41]. Consistently, GIS ranking identified *SREBF2* as the top-ranked TF associated with *TMEM55B*. Overall, the validation results from both ChIP-seq datasets and external databases support the effectiveness of GIS in identifying context-specific TF-TG relationships.

EpiGePT improves variant effect prediction

Context-specific prediction of the functional impact of genetic variants is important for genetic studies. To test the utility of EpiGePT in this task, we first collected an eQTLs dataset [42] that contains 20,913 causal and non-causal variant-gene pairs across 49 different tissues from the supplementary data of Wang et al. [42]. EpiGePT, EpiGePT-seq (i.e. EpiGePT without the TF module) and Enformer were then applied to estimate the context-specific log-ratio scores (LOS) between the alternative DNA sequence and the reference DNA sequence (see "Methods", Fig. 5a). Finally, a random forest classifier is trained based on these LOS's to distinguish causal variant-gene pairs from non-causal pairs. The experimental results show that better prediction performance can be achieved when the LOS is based on EpiGePT than when the LOS is based on Enformer. For example, in the lung tissue, EpiGePT achieved an auPRC of 0.922, compared to 0.873 of Enformer, for the classification of casual SNPs vs non-causal SNPs. To verify the effectiveness of TF module, we replace the TF reference profile of lung with a less relevant cell type, stomach, and the auPRC decreases from 0.922 to 0.892 (Fig. 5b). Similar results were seen for other tissue contexts—across 48 tissues, EpiGePT-seq achieved an average auPRC of 0.910, compared to 0.898 of Enformer (Additional file 1: Fig. S5c). The above experiments demonstrated the usefulness of EpiGePT in assessing variant effects.

To further evaluate the performance of EpiGePT in predicting disease-associated variants, we extracted 52,876 pathogenic SNPs from the ClinVar [43] database and 418,863 benign SNPs from the ClinVar database, also with 84,095 benign SNPs from the ExAC database [44] as positive and negative sets, respectively. We defined a 128-kbp region surrounding each pathogenic SNP as the risk region. We extracted all benign or likely benign SNPs that fall within the risk region as the positive samples. As the relevant tissue or cell type information is not available, we concatenated the LOS of the eight epigenomic signals and the self-attention scores, across multiple cellular contexts, and then evaluated whether the constructed features are beneficial in distinguishing pathogenic SNPs from benign ones in a classification analysis. To achieve this, we augmented the

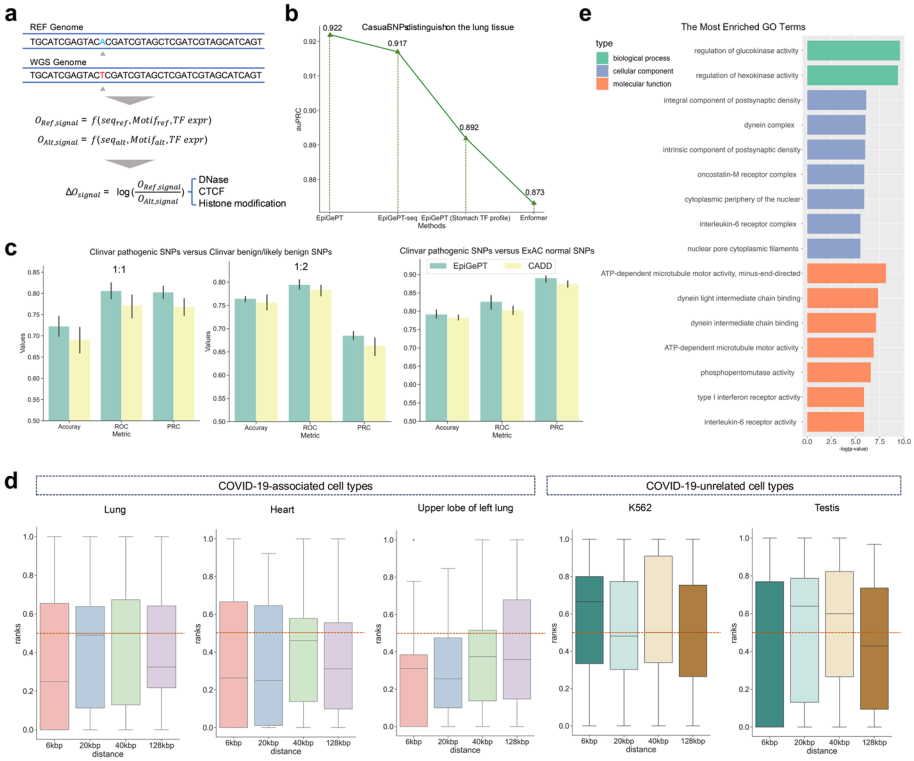

Fig. 5 Variant effect prediction of EpiGePT. **a** The LOS for each epigenomic signal is calculated by the log change fold of the predicted epigenomic signal for reference genome and WGS genome. **b** The performance of EpiGePT and Enformer in discriminating causal SNPs on the Lung tissue. **c** The three subplots from left to right respectively depict the classification results for disease-related SNPs and benign SNPs down-sampled sourced from the ClinVar database, with balanced positive and negative samples (1:1 and 1:2 ratio), as well as normal SNPs sourced from the ExAC database with a MLP classifier. **d** The ranked position of COVID-19-related GWAS data among surrounding benign SNPs based on their LOS, as determined using different tissue or cell-type expression data. The results were stratified based on the distance range of the risk region. The resulting mean and median ranks were both below 0.5. **e** Enrichment result (biological process, cellular component, and molecular function) of the nearest genes of the COVID-19 associated SNPs with the max LOS

popular CADD-derived features (CADD [45] scores) by concatenating them to the EpiGePT-derived features discussed in the above, to obtain a comprehensive feature vector (see "Methods"). Subsequently, we compared the performance of the multi-layer perceptron (MLP) classifier based on the comprehensive feature vector to that based on CADD-derived features alone. The results demonstrated that incorporating EpiGePT-derived features significantly enhance the performance in predicting pathogenic SNPs. Specifically, when the positive-to-negative sample ratio was set to be 1:1, the average auROC increased from 0.772 to 0.806, and the average accuracy increased from 0.690 to 0.723 (Fig. 5c). This observation indicates that features extracted by EpiGePT provide a valuable complement to CADD scores, enabling a more comprehensive interpretation of disease-associated variants.

EpiGePT prioritizes potential SNPs associated with comorbidities of COVID-19

We investigated whether using EpiGePT to predict variant effects could help in the discovery of key SNPs related to COVID-19. COVID-19 is an infectious disease caused

by the SARS-CoV-2 virus, which emerged in late 2019 and quickly spread around the world, causing a global pandemic [46]. In order to validate the ability of EpiGePT in identifying key SNPs, we collected GWAS data from a COVID-19 genetic study [47], including 9484 variants derived from 4933 patients with confirmed severe respiratory symptoms and 1,398,672 control individuals without COVID-19 symptoms. To validate the ability of the model to identify COVID-19-associated SNPs, we firstly defined a risk region around the selected COVID-19-associated SNPs and computed the rank of the variant score of pathogenic SNPs within the surrounding benign SNPs from the ClinVar database. Note that the expected percentile rank for random guessing (uniform distribution) is 0.5 (see "Methods"). Previous studies [48, 49] suggested that COVID-19 infection could potentially impair the function of the heart or the lungs, leading to congestive heart failure or decreased lung function. Interestingly, we found that the average rank of COVID-19-associated SNPs was 0.250 when lung expression data was employed for the TF reference profile and a 6-kbp risk region was examined (Fig. 5d, p-value < 0.05 under one-sided Binomial exact test). However, when we employed the expression data from less relevant contexts, such as K562 cells or Testis cells, the median rank is close to random guessing (i.e. 0.5), indicating its ineffectiveness in discerning SNPs pertinent to COVID-19. These results suggest that EpiGePT is able to prioritize the COVID-19-associated SNPs thus shedding lights on finding the potential disease-associated variants and the relevant tissue contexts.

Next, we examine whether the genes close to max-LOS SNPs are likely be associated with biological processes and functions relevant to COVID-19, when compared with genes close to low scores SNPs or not closed to associated SNPs. Since the genetic pathology of COVID-19 is not yet clear and the earliest lesion is in the lungs, we ranked all 9484 possible SNPs using lung expression data as the TF reference profile. We then identified the SNPs with the highest ranks and performed GO enrichment analysis on nearest genes of the top-30 scored SNPs (Fig. 5e). The enrichment results revealed potential biological processes that are relevant to COVID-19, such as the regulation of glucokinase activity which is associated with the homeostasis of human blood glucose [50]. Notably, diabetes mellitus, a condition closely associated with hyperglycemia, is a typical comorbidity of COVID-19 [51]. However, GO enrichment analysis based on the nearest genes of the lowest-scored 30 SNPs resulted in enrichment outcomes that were less relevant to COVID-19 or its complications (Additional file 1: Fig. S21). Among the potential genes around the top-10 scored SNPs, we identified that the *TBC1D4* gene, which regulates glucose homeostasis, is potentially associated with COVID-19 comorbidities. Our findings are consistent with previous research by Pellegrina et al. [52] and highlights the potential of our EpiGePT approach in discovering new genetic markers that may be implicated in the pathogenesis of COVID-19. Overall, our EpiGePT model provides new perspectives for understanding how the genetic variants could contribute to the COVID-19 susceptibility and severity.

Model ablation analysis

To verify the roles of the main modules in the model, we first conducted ablation experiments on the model architecture (Additional file 1: Fig. S22). For TF module ablation, the results compared to EpiGePT without TF module (EpiGePT-seq) and the inclusion

of the TF module led to improvement in cross-cell-type prediction of DNase signals, with a median PCC of 0.787 of EpiGePT, compared to 0.74 for EpiGePT-seq. We additionally examined the impact of the TF module by employing three methods, namely replacing TF scores with zero, replacing TF scores with random noise, and removing motif binding scores. The results again confirmed the positive impact of the TF module. For sequence module ablation, we trained a TF-only model without the sequence module. The results indicated that removing the sequence module resulted in an average decrease of 0.084 in the PCCs of the epigenomic signals on a cell-type-wise basis (Additional file 1: Fig. S22a). For multi-task module ablation, we trained eight separate predictive models for each of the eight epigenomic signals. In the case of the H3K4me1 signal prediction, the performance of the single-task prediction model exhibited an average PCC decrease from 0.408 to 0.329 compared to the multi-task prediction model. Similarly, the overall prediction performance for the eight signals declined by 0.074 (Additional file 1: Fig. S22b). This decrease may be attributed to the intricate nature of gene regulation that multiple epigenomic signals can synergize with each other, allowing their joint modeling to gain deeper biological insights.

Second, we conducted three ablation experiments on the TF module to address scenarios where TF profiles are sparse or missing. We evaluated the performance of pretrained EpiGePT with a different number of randomly missing TFs by imputing missing TF expression values with zero, the mean or the median of reference TF profiles. Results consistently showed that a larger number of missing TF profiles leads to a more significant decrease in the model performance, highlighting the importance of the TF module (Additional file 1: Fig. S23). To handle the scenarios where the TF expression data are not available, we also simulated scenarios using similar cellular contexts and sequencing data from different samples of the same cell type (Additional file 1: Fig. S24-S25). Moreover, we constructed a comprehensive TF reference dataset and provided guidance on how to construct the TF reference profile for specific cell types based on the large amount of existing sequencing data [53, 54] (Additional file 1: Text S6, Fig. S26-28).

Third, we conducted ablation experiments on the number of training cellular contexts to assess the impact of cell type quantity on performance improvement in predicting epigenomic signals using EpiGePT. The performance of EpiGePT continues to increase as the more training contexts are incorporated (Additional file 1: Fig. S29-S30). Even with a single training context, the incorporation of the TF module yields a 1.2% average improvement in PCC compared to the best sequence-based method. These extensive experiments emphasized the central role of TF module to account for the context-specific gene regulation and leading to a better prediction performance even with a limited number of training cellular contexts.

Online prediction tool for EpiGePT

In order to facilitate the utilization of EpiGePT for the prediction of multiple chromatin states of any cellular context and genomic regions, we have developed a user-friendly web server, named EpiGePT-online (http://health.tsinghua.edu.cn/epigept) (Additional file 1: Text S7). The web server was developed using PHP, JavaScript and HTML, which provides an interactive web interface for efficiently online prediction of epigenomic profiles (Fig. 6). The web server includes a built-in kernel that

encompasses the framework for data preprocessing, TF motif binding scores calculating, and prediction of epigenomic signals for both hg19 and hg38 human reference genome. Users can obtain the predicted signals for multiple genomic regions by submitting a region file and a TF expression file in Numpy or CSV formats, or predicted signals for a specific region by submitting a TF expression file (Additional file 1: Fig. S31). We provided TF expression profile across more than 100 cellular contexts from ENCODE on the download page. Users can download the results in csv format for further applications such as genetics analysis. Furthermore, we provide a case application of the EpiGePT-online to enable users to quickly learn how to use our website (Additional file 1: Text S8). We anticipate that this web server will assist researchers in deepening their understanding of gene regulatory mechanisms.

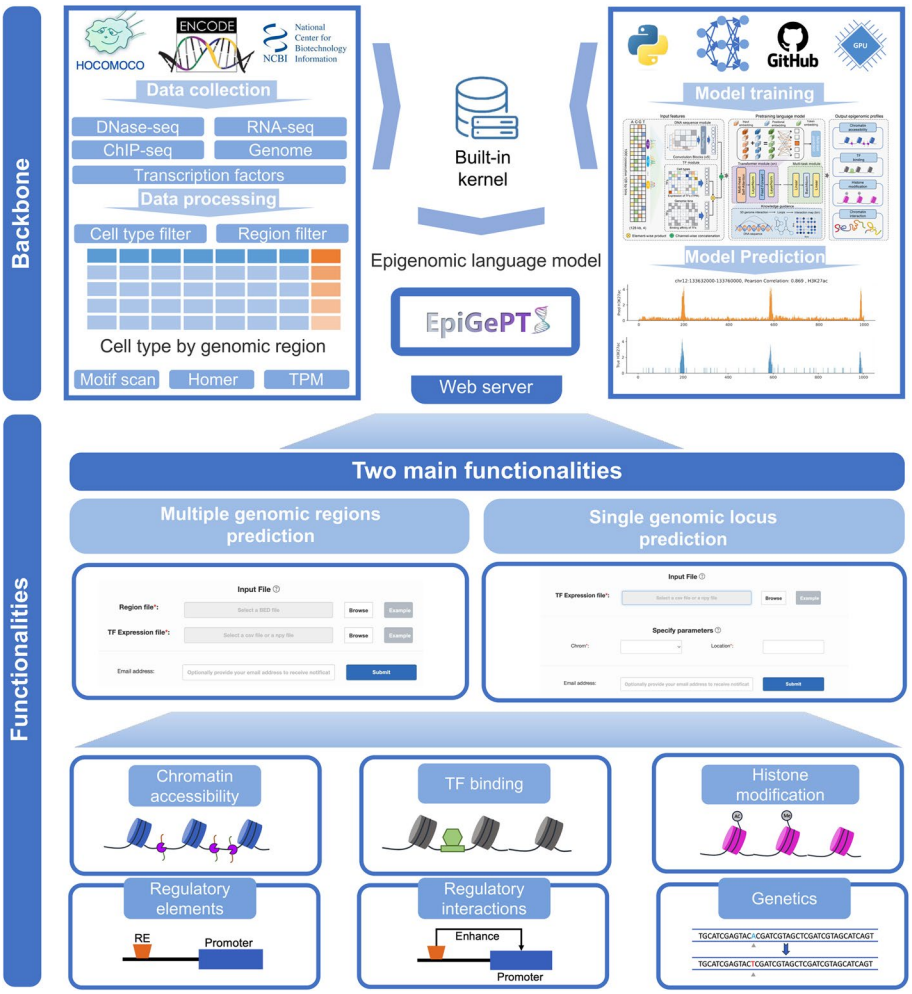

Fig. 6 Overview of the online prediction web server of EpiGePT. We collected eight types of epigenetic genome modification signals and corresponding expression data of transcription factors in different cell types or tissues from the ENCODE project. Based on these data, we trained the EpiGePT model and deployed it as a built-in kernel on an Apache server. Users without much coding experience can also access the web server in two ways to obtain the eight types of epigenetic genome modification signals for specified cell types and genomic regions without programming or installation

Discussion

There exist several extensions and refinements that can be applied to further improve the EpiGePT model. Firstly, the incorporation of chromatin regulators (CRs) as trans-acting factors into the TF module could enhance the modeling of regulated transcription processes, thereby increasing the accuracy of the predictions. Second, the integration of DNA methylation information [55, 56] while modeling DNA sequences allows for a more comprehensive and accurate decoding of the epigenomic language, providing a more comprehensive model of gene regulation states compared to the analysis solely based on DNA sequences. Third, advancements in sequencing technologies have led to the generation of extensive multi-omics data, spanning from molecules to tissues [57]. Integrating single-cell level data into EpiGePT as pseudo-bulk data can further expand the training context, facilitating a comprehensive understanding of regulatory heterogeneity in a finer resolution. Fourth, the 3D genome data in a limited number of contexts demonstrate improvement in cross-cell-type prediction power. Model performance may be further improved by imputing the missing 3D genome data through 3D genome data generative models [58, 59]. Fifth, despite the promising performance of EpiGePT in human epigenomics, there are limitations when applying it across species. We discussed the potential and effectiveness of transferring EpiGePT to other type of species such as plants [60] and fungi [61, 62] (Additional file 1: Text S9, Fig. S32).

Conclusions

In this paper, we introduced a pretrained transformer-based language model for epigenomics. Compared with the existing task-specific models and sequence-based language model, EpiGePT has the added capability to make predictions on novel contexts. Furthermore, EpiGePT is able to incorporate a new type of data (3D genome interaction data) during model training, which enables the identifying functional regulatory interactions such as enhancer-promoter interactions. EpiGePT demonstrates state-of-art performance in diverse experimental settings compared to existing methods. Based on the predicted epigenomic features and 3D interactions from EpiGePT, we performed two investigations on how information is encoded in the human genome sequence: First, we identify the interactions of cis-regulatory elements and their target genes with the help of self-attention mechanism in EpiGePT. Through direct utilization of self-attention scores, model fine-tuning, and leveraging 3D genome interactions, we validated the capacity of EpiGePT to capture regulatory interactions. Second, to assist the identification and interpretation of human disease-associated SNPs, we estimate the effect of a variant on the epigenomic features around the variant, based on the LOS computed by the outputs of EpiGePT. Such variant effect prediction by EpiGePT establishes a foundation for understanding the underlying relationship between genetic variations and disease mechanisms.

Based on EpiGePT, users are able to predict multiple chromatin profiles in different cell lines or tissues, which could provide a foundation for biological discovery, decoding transcriptional regulation mechanisms and investigating disease mechanisms. We anticipate EpiGePT will furnish researchers with valuable insights into understanding regulatory mechanisms.

Methods

Data processing

Chromatin accessibility data and expression data

We used three different datasets in the experiments. For chromatin accessible data, we downloaded DNase bam files across 129 human biosamples from ENCODE [21] project (Additional file 2: Table S6). We divided the human hg19 genome into 200-bp non-overlapping bins, and we assigned the label for each bin in each cell type. For the regression design, we pooled the bam files of multiple replicates for a cell type, and obtain the raw read count n_{lk} for bin l in cell type k. We normalized the raw read count in order to eliminate the effect of sequencing depths, in the form of $\widetilde{n}_{lk} = Nn_{lk}/N_k$, where N_k denotes the total number of pooled reads for cell type k and $N = \min_k N_k$ denotes the minimal number of pooled reads across all cell types. The normalized read counts are further log transformed with pseudo count 1, which represent the continuous level of chromatin accessibility. For binary classification design, we assigned a binary label y_{lk} to 1 if the number of raw read counts of the bin l in the cell type k greater than 30, which represent the bin is an accessible region in this cell type, resulting in the identification of regions as accessible in 13% on average and 8% at median in the screened genomic regions across 129 cell types. The proportion of open regions varies among different cell types, and the average openness level mentioned above is generally consistent with that maintained in ChromDragoNN [17].

RNA-seq data of the 711 human TFs were downloaded and extracted from the ENCODE project (Additional file 2: Table S7- S8). We perform log transformation with pseudo count 1 and quantile normalization based on TPM values. The normalized TPM values were averaged across replicates and mean expression profile after normalization of each cell type was finally used to calculated of the transcription feature.

Multiple chromatin signals data

For the human reference genome hg19 (GRCh37), DNase-seq, RNA-seq and ChIP-seq data were also downloaded from ENCODE project (Additional file 2: Table S9-S11). We applied the same process to these data as above, and finally we obtained the 8 epigenomic signals of 13,300,000 bins of 128 bp in 28 cell types. The continuous level of chromatin signals we extracted were "DNase", "CTCF", "H3K27ac", "H3K4me3", "H3K36me3", "H3K27me3", "H3K9me3," and "H3K4me1," which includes crucial epigenetic modifications and markers for gene regulation and transcription.

For the collected data of human reference genome hg38 (GRCh38), we adopted a data collection strategy that includes missing data. Specifically, within a particular tissue or cell type, we ensured the presence of at least one ChIP-seq signal. Then, epigenomic profiles of 8 signals for 15,870,000 bins of 128 bp across 104 cell types were obtained (Additional file 2: Table S12-S14).

Model architecture

Sequence module

As shown in Fig. 1 and Additional file 1: Fig. S1a, the sequence module receives a one-hot matrix (A = [0,0,0,1], C = [0,1,0,0], G = [0,0,1,0], T = [0,0,0,1]) of size (128,000,4) as input, representing a sequence of 128 kilobase pairs (kbps) and contains five 1-dimentional

convolutional blocks to extract DNA sequence features. Each block includes a convolutional layer and a maxpooling layer (Additional file 1: Fig. S1b). The first convolutional layer considers the input channels as 4 and performs convolution along the sequence direction. The input sequence features are one-hot embeddings of size $L \times 4$, where L denotes the length of the input long-range DNA sequence. After 5 maxpooling layers, the output size of sequence feature is $L/N \times C$, where C denotes the hyper-parameter for sequence embedding and N denotes the length of locus to predict. We set C to 256 in the pre-training stage of chromatin accessibility prediction experiments. Rectified linear units (ReLU) are used after each convolution operation for keeping positive activations and setting negative activation values to zeros. By reducing the input length by 128 times through pooling operations, this module effectively compresses the input information while retaining essential features. Sequence features were then concatenated with TF expression features, and we finally obtained a vector of size $L/N \times (C + n_{TF})$, where n_{TF} denotes the dimension of the TFs features after padding. In our model, after adding padding to the 711 TFs, the n_{TF} is set to 712. Therefore, the input token number for the transformer module is 1000, and each token embedding has a dimensionality of 968.

Transformer module

We utilize the transformer module to integrate information from both the sequence and TFs, enabling the capturing of long-range interactions between genomic bins. We applied N_t layers of Transformer encoder with n_h different attention heads to the token embedding sequence. The input word embedding (X) of the transformer encoder is a genomic bin sequence with dimensions ($Sequencelength, embeddingdim$). Specifically, this dimension is (1000, 968) in EpiGePT, indicating that input genomic bin sequence has a length of 1000, and each genomic bin has an embedded representation that combines the sequence information with cell-type-specific features with dimension of 968. For position embedding, we employed absolute position embedding to represent the positional information of the 1000 genomic bins in the input 128-kbp DNA sequence, with dimensions of (1000, 968). Each Transformer encoder includes a multi-head self-attention mechanism and a feed-forward neural network. For self-attention in each head, the calculation is based on the matrix operation.

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

For multi-head attention, Transformer encoder learns parameter matrices $W_i^Q \in \mathbb{R}^{d_{model} \times d_K}$, $W_i^K \in \mathbb{R}^{d_{model} \times d_K}$ and $W_i^V \in \mathbb{R}^{d_{model} \times d_V}$ for the i_{th} head and concatenate the multiple heads to do the projection, then learns parameter matrices $W^O \in \mathbb{R}^{n_h d_v \times d_{model}}$ to obtain the output of multi-head attention layer.

$$Q_i = XW_i^Q, K_i = XW_i^K, V_i = XW_i^V$$

$$A_i = softmax(\frac{Q_i {K_i}^T}{\sqrt{d_k}})$$

$$head_i = Attention(Q_i, K_i, V_i) = A_i V_i$$

$$MultiHead(Q, K, V) = Concat\left(head_1, \cdots, head_{n_h}\right) W^O$$

where d_{model} denotes the dimension of token embedding X, which is 968 in EpiGePT X denotes the embeddings from the sequence module for the first attention layer or the output of previous attention layer. n_h denotes the number of head in Transformer encoder, which is 8 in EpiGePT, and $d_K = d_V = d_{model}/n_h = 121$. The matrix A_i is called the self-attention matrix for head i. The outputs of n_h heads are then concatenated, and a mapping function represented by W^O is applied to obtain the output of the multi-head attention. After passing through an add & norm layer, the multi-head attention output is used as input to the feed-forward layer, where more comprehensive features of the input sequence are extracted. The above describes the computational workflow of a single Transformer encoder layer. We set N_t to 16 for the chromatin accessible prediction experiments, N_t to 12 for the chromatin state classification and multiple chromatin signals prediction experiments.

Prediction module

For regression model, the output layer uses a linear transformation and use mean square error (MSE) as the loss function. For classification model, the output layer uses a linear transformation combined with a sigmoid function, and use the cross-entropy loss for classification experiments.

TF module

For binding status, we scanned the input bins for potential binding sites for a set of 711 human TFs from HOCOMOCO database [63] with the tool Homer [64] (Additional file 2: Table S7-S8, S10-S12, S14-S16). We then selected the maximum score of reported binding status for each TF to obtain a vector of 711 dimensions as the motif feature for each DNA bin. For gene expression, we focused on log-transformed TPM values of the 711 TFs and obtained a vector of 711 dimensions after quantile normalization as the expression feature. With these data, we combined the two vectors of motif and expression features by taking the element-wise product, and we concatenated the result to the output of sequence module.

Model evaluation

To evaluate our model, we applied five-fold cross-validation in the different experiments on cell-type level. For chromatin accessible experiments, the 129 cell lines are partitioned into a training set and a testing set randomly.

Cell-type-wise metrics are defined to evaluate our method in different experiments, which were calculated with the data within a test cell type across all genomic locus. For binary classification design, we used cell-type-wise auPRC and auROC to evaluate our EpiGePT. Let $Y_{L\times K}$ and $\widehat{Y}_{L\times K}$ be the true and predicted matrix, where L denotes the number of locus and K denotes the number of test cell types. We calculated the auPRC and auROC for each $(y_{1i}, y_{2i}, \cdots, y_{Li})$ and $(\widehat{y}_{1i}, \widehat{y}_{2i}, \cdots, \widehat{y}_{Li})$. For multiple classification, we use macro average of the auPRC and auROC to evaluate

the classification performance, which compute the metric independently for each class and then take the average hence treating all classes equally. For regression design, we used three metrics for model evaluation, which are cell-type-wise PCC, SCC, and prediction squared error. Prediction square error (PSR) is calculated as $PSR = 1 - \sum_k \sum_l \left(y_{lk} - \widehat{y}_{lk}\right)^2 / \left(y_{lk} - \bar{y}_{*k}\right)^2$, where $\bar{y}_{*k} = \sum_l y_{lk}/L$ denotes the mean of the true level of the response in the cell type k.

To compare the performance of our method with other baseline methods, we conducted hypothesis testing on the metrics based on cell types. Since the metrics on a given cell type across different methods are paired data and the statistical distribution is unknown, we employed both Binomial and Wilcoxon tests, with the alternative hypothesis being that EpiGePT outperforms the other methods. If we reject the null hypothesis, it provides compelling evidence to support the claim that EpiGePT performs better than the other methods.

To evaluate the computational efficiency, we recorded the running time of a single epoch of EpiGePT and baseline methods (Additional file 1: Text S10). Compared to traditional CNN models such as DeepCAGE [18] and ChromDragoNN [17], as well as larger sequence models like Enformer, EpiGePT demonstrates a balance between high computational efficiency and performance.

Model training strategy

As our proposed model is designed for cross-cell-type prediction of epigenomic signals by multi-task learning, some of the target epigenomic signals are missing in the existing ENCODE database. For instance, there are 104 cellular contexts with both gene expression and at least one of the epigenomic data. However, this number will decrease from 104 to 28 if we consider eight epigenomic signals simultaneously. The proposed model takes each cellular context and genomic region pair as a training instance, which ensures the availability of a very large number of training instances. To utilize the data from the cellular contexts where some signals are not available (missing data), we will use a new training strategy to handle the missing data where the loss function is designed as

$$L = \frac{1}{J}\sum_{j=1}^{J} \frac{1}{|B_i|} \sum_{k=1}^{K} ||y_{i,j,k} - \widehat{y}_{i,j,k}||_2^2 \cdot I(k \in B_i)$$

where $y_{i,j,k}$ and $\widehat{y}_{i,j,k}$ denote the k^{th} true and predicted signal from the j^{th} genomic bin in the i^{th} context, and $I(\cdot)$ is an indicator function and B_i denotes the index set that contains all available signals in the i^{th} context. We update the parameters in the model through stochastic gradient descent based on minibatches. We utilized the Adam optimizer with a batch size of 10 and a learning rate set to 5×10^{-5}. This training strategy provide us with a significantly larger training sample size and allows us to utilize much more available data from the public databases, and we enable EpiGePT to learn broader patterns of epigenetic states across diverse cell types. The total parameter size of EpiGePT is 71.3M, making it feasible to train and make inference even on a single GPU. A comprehensive benchmark study of the computational resources used in training EpiGePT is provided in Additional file 1: Text S11, Additional file 2: Table S17-S19 and Additional file 1: Fig. S33.

Transferring EpiGePT to mouse data

The ENCODE project [21] includes a substantial amount of bulk-level epigenomic and gene expression experimental data for mouse. We took the DNase-seq and RNA-seq data from ENCODE project as illustrative examples. Specifically, we collected the bulk data from three organ-level datasets for zero-shot experiments and fine-tuning experiments: brain (ENCSR000COE), lung (ENCSR000CNM), and kidney (ENCSR000CNG). For DNase-seq data, we annotated the chromatin openness in genomic regions using OpenAnnotate [24, 25]. The foreground openness in these three organs was annotated on the regions based on the corresponding narrow peak files (ENCFF941CCB, ENCFF268DLZ, and ENCFF587XGH), from which genomic regions were selected for prediction. Then, each peak was extended to a fixed length centered on the original peak center. After processing, we obtained 247,750 peaks for brain data, 243,900 peaks for lung data, and 180,400 peaks for kidney data. For RNA-seq data, we used RNA-seq data from the same samples (ENCFF455OZW, ENCFF641SMI, and ENCFF876OPZ), with TPM values representing the expression levels of corresponding TFs. For TF binding status, due to the high conservation of TF binding between humans and mouse, we utilized the same homer tool [64] and motif database with human to scan for TF binding status on peaks. We identified 688 corresponding TFs in mouse RNA-seq data using the gene annotation file (vM25) from the GENCODE project (https://www.gencodegenes.org) from the 711 selected TFs of humans, while the expression values of the remaining TFs were set to zero. A more detailed description of the necessary conditions based and data preparation is provided in Additional file 1: Text S9.

Incorporation of 3D chromatin interaction data

With the emergence of methodologies like Hi-C and HiChIP for genome-wide chromatin interaction measurement, a substantial volume of 3D chromatin interaction data has been produced across various cellular contexts. Clearly, this data can provide highly valuable information for identifying functional elements in the genome and for understanding gene regulation, but this information has not been captured by current genomic LLMs such as the Enformer [16] or earlier CNN-based genomic models [5, 17, 18, 65].

We propose here to exploit the self-attention weights of the transformer model to design a learning strategy that would allow EpiGePT to capture interaction information from Hi-C or HiChIP data. Specifically, we propose to use the ground truth 3D genome interaction to guide the self-attention matrices in the transformer module during the training process. First, we obtained loop information at 5k resolution from the HiChIPdb database [31]. Given potential noise within HiChip data, we selectively filtered potential H3K27ac-based HiChip loops using a stringent q-value threshold of 0.001. This curation aimed to utilization of highly confident loops, safeguarding the model's ability to capture regulatory information without interference from noise. In this way, we acquired corresponding HiChip loop data for 13 out of 104 cell types. Next, we mapped these loops onto the genomic bins used for pre-training. Specifically, we employed the normalized count as a metric to gauge the likelihood score for each loop. During the mapping process, we aggregated all loops based on this score to each

specific genomic bin, and then we obtained the HiChIP interaction matrix H_i Based on the self-attention matrix $A^i_{p,q} \in R^{J \times J}$ and the HiChIP interaction matrix H_i from the i^{th} cell type/tissue where p, q are indexes for transformer layer and multi-heads, we apply a row-wise normalization to H_i (row sum to 1) to obtain $\widetilde{H}_i$ and average the self-attention matrices across the heads in the last transformer layers to obtain $\widetilde{A}^i$. Since elevated attention weights are expected between regions that interacts in 3D, we will compute a new loss term CSL, which is defined as cosine similarity loss between the rows of $\widetilde{H}_i$ and $\widetilde{A}^i$. Through the guidance of 3D genome interaction data, our approach can learn a more comprehensive model for gene regulation. For example, it will enable prediction of cell-type-specific enhancer-promoter interaction, which is a task beyond current models such as the Enformer. Note that the CSL term does not alter the architecture of the model. It simply put some soft constraints on the attention weights according to the experimental data on chromatin interactions, so that the optimized model will give predictions that are more consistent with the context-specific interaction data. For the results in Fig. 3, the weight α for 3d genome loss during training EpiGePT-3D was chosen as 2.

Fine-tuning for predicting E-P interaction

For the fine-tuning process, we kept the parameters of the pre-trained model fixed without making any updates. For the specific fine-tuning task of chromatin interaction prediction based on HiChIP data, the multi-task prediction module was replaced with a two-layer MLP network, containing 256 hidden nodes for each layer. During the training process, only the weights in the MLP network in the prediction module were updated. Notably, when utilizing HiChIP data at a resolution of 5 k, both the enhancer and promoter anchors spanned 5 kbp. Then we use a region extending 128 kbp from the center of the anchor of the neighboring gene, as input region for EpiGePT. Consequently, a 968-dimensional feature vector for each genomic bin was derived from the output of the last transformer encoder layer. These feature vectors from all bins within the two anchors were concatenated, resulting in a high-dimensional vector of size 76,472. To ensure the fairness of validating EpiGePT-finetune in capturing E-P interaction relationships, we fine-tuned the model separately on the HiChIP data of each cell line during the fine-tuning process. The test cell lines K562 and GM12878 were excluded from the pretrained EpiGePT training cell types.

Baseline methods

Four baselines were introduced for epigenetic signals prediction. BIRD [19] is a multiple linear regression model that only takes gene expression data as input and makes predictions on a fixed locus. ChromDragoNN [17] is a deep neural network that takes gene expression of 1630 TFs and DNA sequence as input. Specifically, ChromDragoNN [17] uses a ResNet [66] to extract sequence features and use linear transformation to combine the TF gene expression feature and sequence feature to make the final prediction. DeepCAGE [18] is a deep densely connected convolutional network for predicting chromatin accessibility. Enformer [16] is a deep neural network that integrates convolutional neural network and transformer, and only takes DNA sequence as input. Enformer takes DNA sequence of length 196 kbp as input to predict 5313 genomic tracks of human and

1643 tracks of mouse genome simultaneously. Enformer can only model and predict cell types in the training data and cannot be applied to new cell types. In order to ensure fairness in some of the benchmark experiment, we retrained the Enformer model with the same input and output data as EpiGePT with Pytorch-lightning and made modifications on the number of encoder layers when reproduce the Enformer model (Supplementary Text S4). Besides, comparison with the pretrained Enformer model was also provided in Fig. 2d where we strictly used the ENCODE experiment ID to obtain the matched experiments for comparison.

Two baseline methods were introduced for predicting HiChIP interaction. DeepTACT [35] is a deep learning method for predicting 3D chromatin contacts using both DNA sequence and chromatin accessibility. We adopted the structure of DeepTACT [35] and kept the anchor length at 5 k. The input to the model consists of two anchor sequences represented as one-hot matrices and the two openness scores of the two anchors on the corresponding cell type extracted from OpenAnnotate [24]. Regarding the Kmer features [67], K is chosen as 5 to extract sequence features. For each anchor, a vector of dimension $4^5 = 1024$ was obtained. Further training was performed using an MLP with a hidden layer dimension of 256.

Prediction of 3D genome interaction

We collected cis-regulatory elements-gene pairs in K562 cells from other studies and public database to demonstrate the interpretability of self-attention mechanisms in the EpiGePT. Enhancers and silencers are typical *cis*-regulatory elements known play important roles in transcriptional control during normal development and disease. For enhancers, we downloaded enhancer-gene pairs from two studies: Gasperini et al. [27] and Fulco et al. [28], both of which were tested using a CRISPRi [26] assay perturbation. Two datasets contain 664 and 5091 element-gene interactions. For silencers, we obtained and random sampled 831 validated silencers-gene pairs with distance within 64 kbp in K562 cells curated from high-throughput experiments from SilencerDB [29]. As there are no experimentally validated interaction relationships between these silencers and genes, we generated silencer-gene pairs by associating the nearest neighbor genes for classification purposes. Similarly, negative samples were generated by constructing DNase-seq, ATAC-seq, and nearest genes using the same approach. Ultimately, we obtained a dataset comprising 1662 silencer-gene pairs, encompassing both positive and negative instances.

To obtain scores for regulatory element-gene pairs, we first used the region extending 128 kbp from the TSS of the gene as input and extracted the token where the interacting regulatory elements reside, so that we could filter out regulatory element-gene pairs that were located further than 64 kbp apart. Subsequently, we stratified the remaining pairs based on their distance. Since the positive and negative sample ratios varied across datasets, we adopted different stratification strategies for different distance ranges (Fig. 3). Next, we averaged the attention matrices of the Transformer encoder across all layers and heads. The summed attention scores from other tokens to the key token containing the center of the regulatory element were used as the attention score of this element-gene pair. This score represents the attention value that the enhancer receives in the region around the TSS of the gene. We also calculated the attention score from the bin

containing the center of the regulatory element to the bin containing the TSS, which only slightly affects the experimental results of regulatory element prioritization.

We collected 5 k resolution data from the HiChIPdb (http://health.tsinghua.edu.cn/hichipdb) database, specifically from K562 and GM12878 cell lines. We filtered the data to include only loops where at least one anchor falls within a gene region. We stratified the loops based on distance into three categories: 0–20 kbp, 20–40 kbp, and 40–64 kbp. For each distance category, we selected 2000 positive pairs with most significant q-value. To ensure consistency in the distance distribution, we selected negative pairs by fixing a gene and choosing anchors at equidistant locations in the opposite direction. These are then used as test data to evaluate the prediction methods.

Gradient importance scores

EpiGePT possesses the capability to assign priority rankings to transcription factors by utilizing GIS, taking into account specific cell types and chromatin regions. The GIS were employed to identify potential functional relationships between specific TFs and target genes. Specifically, for a given TF-target gene pair, the TSS of genes were used as central loci, and the regions spanning 128 kbp upstream and downstream of the TSS were selected as input. Next, we selected bins with motif binding scores indicating potential binding for the given TF. For these selected bins, we calculated the GIS for the predictions of eight epigenomic signals, for each of 711 core TFs.

$$GIS_{ijk} = \frac{1}{|\zeta|}\sum_{l\in\zeta}\left|\frac{\partial\widehat{y}_{ljk}}{\partial tf_{ij}}\right|$$

where i denotes the ith TF in the set of core TFs, j denotes the jth cell type, k denotes the kth predicted epigenomic signal, and ζ denotes the set of genomic bins that have binding for the given TF. In the calculation of the gradient, $\widehat{y}_{ljk}$ denotes the predicted value of the kth epigenomic signal by the model using the expression in the jth cell type at the lth bin. On the other hand, tf_{ij} denotes the product of the expression of ith TF in the jth cell type and the corresponding TF binding score.

If we consider the GIS for the prediction of all 8 epigenomic signals simultaneously, we can prioritize the TFs by calculating their ranks based on each signal separately. Then, we can calculate an integrated gradient importance score (IGIS) for each TF by averaging the ranks from all 8 signals.

$$IGIS_{ij} = \frac{1}{8}\sum_{k} rank(GIS_{ijk})$$

Both the GIS and the IGIS are capable of capturing the significance of a transcription factor (TF) in regulating a specific gene within the context of a specific cell type. Consequently, these scores hold potential value in the discovery of TFs that play crucial roles in the regulation of specific genes, thereby contributing to our understanding of essential regulatory mechanisms.

In the context of validating TF-TG pairs in the GRNdb and TRRUST databases, we opted to utilize liver expression data as a representative example due to the unavailability of cell type information for TRRUST. Furthermore, in this experimental setup, the tf_{ij}

denotes the expression of *i*th TF in the *j*th cell type and ζ denotes the set of genomic bins that have binding for the TF of the given TF-target gene pair.

Potential TF-target gene pairs from ChIP-seq data

In this study, we utilized three distinct cell types to conduct a comprehensive screening of TF-target gene pairs and non-target gene pairs across the human genome. Initially, we obtained the narrow peak files (ENCFF388AJH, ENCFF717IXP, and ENCFF885KLR) from ChIP-seq experiments across three cell types from the ENCODE project. Subsequently, we examined the number of peaks within a 128-kbp region both upstream and downstream of the TSS for each gene. Different thresholds were applied to the ChIP-seq data of various TFs. Genes lacking any peaks within the defined region were classified as non-target genes, while genes surpassing the threshold in terms of peak counts were designated as target genes. Specifically, for the aforementioned three cell types, threshold values of 10, 15, and 6 were respectively employed. Finally, the IGIS approach was employed to determine the corresponding ranks of TFs in the TF-target gene pairs.

Pathogenic SNP prioritization

We collected single-nucleotide polymorphisms (SNPs) data from the ClinVar and ExAC databases, which include both potentially pathogenic and benign SNPs. To evaluate the ability of EpiGePT to predict variant effects, we computed the log-ratio scores (LOS) for multiple chromatin signals using EpiGePT on these SNPs. Subsequently, we utilized these scores to distinguish between pathogenic and benign SNPs. The LOS for each chromatin signal was defined by computing a forward pass through the model using the reference and alternative alleles.

$$\Delta O_{signal} = log(\frac{output(I_{alt})}{output(I_{ref})})$$

where I_{ref} denotes the input DNA sequence based on the reference genome, and I_{alt} denotes the input DNA sequence containing variants. Each chromatin epigenomic profile in each cell line or tissue predicted by EpiGePT can be used to compute a specific variant score. We did not take the absolute value in this calculation, so the resulting LOS indicates the direction of change in the model output after the appearance of the variant. In addition to the predicted chromatin signals output by the eight models, attention score changes based on self-attention are also noteworthy. We computed the log-ratio scores for attention by summing the attention scores of the 10 bins upstream and downstream of the locus of the SNP, to evaluate the effect of the variant.

$$\Delta O_{attention} = \sum_{i=-5}^{5} |log\left(\frac{attn(bin_i)_{I(alt)}}{attn(bin_i)_{I(ref)}}\right)|$$

where i represents the index of the neighboring bins relative to the locus of the SNP. To avoid the variant effects of different bins from cancelling each other out during the summation process, we computed the absolute value of the change in attention scores for each bin and then summed the scores of the 10 adjacent bins centered at the SNP position. For the classification of pathogenic SNPs, we calculated these nine LOS for

attention separately for each of the 28 tissues or cell lines in training data. As a result, we obtained a feature vector of 252 dimensions for each SNP. Then a classifier with 252 features computed by EpiGePT and 52 annotations from CADD score as inputs are used to predict pathogenic SNPs against benign or likely benign SNPs. Here, we employed MLP as classifier to validate the effectiveness of the features we obtained. A five-fold cross-validation experiment is employed for validation, and we utilize two different positive-to-negative sample ratios, namely 1:1 and 1:2. For each sample ratio, we randomly sample 32,000 positive samples. The effectiveness of the variant score in identifying pathogenic SNPs is evaluated using the area under the auROC and the auPRC. Additionally, we also utilized the logistic regression (LR) as the classifier, consistent with the LR classifier used in CADD, and found a similar improvement when predicting pathogenic SNPs.

COVID-19-associated SNPS prioritization

We applied the same method to calculate the LOS of the 8 epigenomic signals for the COVID-19 GWAS data. The absolute values of the scores were summed as the overall score for each SNP. Then, we use the absolute sum as the effect score of the SNP and prioritize the COVID-19-associated SNPs based on this score. For each significant SNP associated with COVID-19 severity obtained from the GWAS data, we selected normal SNPs within a 128-kb region around the SNP as background to calculate the rank of the LOS for the COVID-19 associated SNP in this region. Furthermore, we calculated the LOS for all 9484 COVID-19 associated SNPs and ranked them accordingly. The top 10 SNPs with the highest LOS were selected, which are considered to have potential genetic associations with COVID-19 severity and complications.

GTEx classification

We collected eQTL data from the supplementary materials of Wang et al. [42]. In their study, the authors identified causal eQTLs through statistical fine-mapping, using a posterior inclusion probability (PIP) threshold of >0.9 for putative causal variants based on expression modifier score (EMS), and a PIP threshold of <0.9 for putative non-causal variants. To validate the ability of EpiGePT to distinguish potential causal variants, we perform a classification task on these variants. For each variation, 128-kbp sequence regions near it were selected as the input of the model, and a score of variation was given by EpiGePT model. For each variant under each tissue, we can obtain an 8-dimensional vector of genomic features including DNase, CTCF, and other ChIP-seq signals. Based on the LOS, separate random forest classifiers consisting of 10 decision trees are trained for each tissue in order to distinguish between causal and non-causal variants. The models are evaluated using fivefold validation on each tissue, with area under the auPRC and auROC as metrics for assessing their ability to distinguish between causal and non-causal variants.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-024-03449-7.

Additional file 1: Text S1: Data splitting strategy for model training. Text S2: Data processing for ChromHMM annotation data. Text S3: Details of the comparison matrix in functional chromatin status prediction experiments. Text S4:

Implementation of Enformer model and Enformer+. Text S5: Description and comparison of EpiGePT with different settings. Text S6: Guidance on how to use EpiGePT in two scenarios without tissue-specific TFs data. Text S7: System design and implementation of the web server. Text S8: Case application of the EpiGePT-online. Text S9: Conditions and data preparation for retraining the EpiGePT model on other species. Text S10: Running time of the EpiGePT and baseline methods. Text S11: Computational resources used to train the model. Fig S1: Model architecture of EpiGePT for multiple epigenomic signals prediction. Fig. S2: Three data partitioning strategies for model training and testing. Fig. S3: EpiGePT's performance in predicting DNase-seq and other epigenetic signals. Fig. S4: Comparison of cross-cell-type chromatin accessibility (DNase-seq profiles) prediction performance between EpiGePT and baseline methods. Fig. S5: Performance of EpiGePT and baseline methods on chromatin states classification and causal variants classification. Fig. S6: The average predictive performance of the remaining methods when each model serves as the ground truth. Fig. S7: The performance comparison of retrained Enformer with different numbers of encoder layers for predicting chromatin accessibility. Fig. S8: Performance of EpiGePT in cross-cell-type prediction. Fig. S9: Zero-shot learning of EpiGePT on chromatin accessibility (DNase-seq) prediction on mouse data. Fig. S10: Model fine-tuning EpiGePT on chromatin accessibility (DNase-seq) prediction on mouse data. Fig. S11: The cross-chromosome predictive performance of pretrained EpiGePT and EpiGePT-seq after fine-tuning on mouse data. Fig. S12: The performance (auROC) of attention score of EpiGePT in distinguishing regulatory element-gene pairs at different distance ranges. Fig. S13: Incorporating 3D genomic information from HiChip data enhances the predictive performance of EpiGePT on E-P regulatory interaction on K562 cell line. Fig. S14: The fine-tuning performance of the EpiGePT model on predicting potential enhancer-promoter regulatory networks. Fig. S15: The ROC and PR curves of the EpiGePT model on predicting potential enhancer-promoter regulatory networks. Fig. S16: Comparison of the performance between fine-tuned EpiGePT-3D, EpiGePT and baseline methods in predicting HiChIP loops (ROC curves). Fig. S17: Comparison of the performance between fine-tuned EpiGePT-3D, EpiGePT and baseline methods in predicting HiChIP loops (PR curves). Fig. S18: The performance comparison of original Enformer with EpiGePT and EpiGePT-3D on 78 genomic tracks. Fig. S19: The GIS of ChIP-seq overlapped bins versus non-overlapped bins of POU5F1 centered at the TSS of ESRRB. Fig. S20: Gene ontology enrichment analysis based on the top 5% TFs with high expression in ESCs. Fig. S21: Enrichment result (Cellular component and Molecular function) of the nearest genes of the COVID-19 associated SNPs with the low LOS. Fig. S22: Ablation analysis of the EpiGePT model. Fig. S23: The simulation evaluated the changes in PCC for cross-cell type predictions under different numbers of missing TF profiles with three replacement strategies. Fig. S24: In three sets of similar cell types, the changes in predictive performance of eight types of epigenomic signals were simulated under different replacement strategies (similar, mean, median) when TF profiles were missing. Fig. S25: In the scenario of two TF profiles from the same cell type but different samples, the changes in DNase-seq predictive performance are simulated when missing TF profiles are replaced with the TF profile from a different sample. Fig. S26: The pie charts illustrate the distribution of RNA-seq data from the ENCODE project. Fig. S27: The pie charts illustrate the distribution of human and mouse single-cell RNA-seq data across various tissues from the CELLxGene database. Fig. S28: The pie charts illustrate the distribution of human and mouse single-cell RNA-seq data across various cell types from the CELLxGene database. Fig. S29: The Pearson correlation coefficient (PCC) for predicting chromatin accessibility (DNase-seq) under different numbers of training cell lines/tissues. Fig. S30: The Spearman correlation coefficient (SCC) for predicting chromatin accessibility (DNase-seq) under different numbers of training cell lines/tissues. Fig. S31: Case application of the EpiGePT-online. Fig. S32: The overlap between human and plant gene annotations from the Ensembl Plant project in terms of the 711 selected TFs. Fig. S33: The computational resource usage, including peak GPU memory and time for training an epoch on an NVIDIA GeForce RTX 4090.

Additional file 2: Table S1: Comparison of the performance between fine-tuned EpiGePT and baseline methods in predicting HiChIP loops on the K562 cell line (positive-to-negative sample ratio of 1:1). Table S2: Comparison of the performance between fine-tuned EpiGePT and baseline methods in predicting HiChIP loops on the GM12878 cell line (positive-to-negative sample ratio of 1:1). Table S3: Comparison of the performance between fine-tuned EpiGePT and baseline methods in predicting HiChIP loops on the K562 cell line (positive-to-negative sample ratio of 1:2). Table S4: Comparison of the performance between fine-tuned EpiGePT and baseline methods in predicting HiChIP loops on the GM12878 cell line (positive-to-negative sample ratio of 1:2). Table S5: Description and comparison of EpiGePT with different settings. Table S6: The information of DNase-seq bam file across 129 biosamples from the ENCODE project. Table S7: The information of RNA-seq tab-separated values (tsv) file across 129 biosamples from the ENCODE project. Table S8: The preprocessed expression data of 711 human transcription factors from the ENCODE project across 129 biosamples. Table S9: The information of DNase-seq, CTCF and other six Histone markers bam file across 28 cell lines or tissues from the ENCODE project (hg19). Table S10: The information of RNA-seq tab-separated values (tsv) file across 28 cell lines or tissues from the ENCODE project (hg19). Table S11: The preprocessed expression data of 711 human transcription factors from the ENCODE project across 28 cell lines or tissues (hg19). Table S12: The information of RNA-seq tab-separated values (tsv) file across 104 cell lines or tissues from the ENCODE project (hg38). Table S13: The information of DNase-seq, CTCF and other six Histone markers bam file across 104 cell lines or tissues from the ENCODE project (hg38). Table S14: The preprocessed expression data of 711 human transcription factors from the ENCODE project across 104 cell lines or tissues (hg38). Table S15: The order and names of epigenomes of the expression matrices across 56 epigenomes from the ROADMAP project. Table S16: The preprocessed expression data of 642 human transcription factors across 56 epigenomes from the ROADMAP project. Table S17: The computational resource usage, including peak GPU memory and peak RAM consumption (NVIDIA GeForce RTX 3090). Table S18: The computational resource usage, including peak GPU memory and peak RAM consumption (NVIDIA GeForce RTX 4090). Table S19: The computational resource usage, including peak GPU memory and peak RAM consumption (NVIDIA RTX A6000).

Additional file 3: Review history.

Acknowledgements
Not applicable.

Review history
The review history is available as Additional file 3.

Peer review information
The Editorial Board Member Nicolae Radu Zabet and Wenjing She were the primary editors of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions
Q.L., R.J., and W.H.W. conceived and supervised the project. Z.G. and Q.L. designed and implemented the method and conducted experiments. Z.G. and Q.L. and W.Z. performed data analysis. Z.G. developed and built the webserver. Z.G., Q.L., W.Z., R.J., and W.H.W. drafted the manuscript. All authors read and approved the final manuscript.

Funding
Q.L. was partially supported by the National Human Genome Research Institute (NHGRI) of the National Institutes of Health (NIH) under Award Number K99HG013661. Q.L., W.Z., and W.H.W. were partially supported by NIH grants R01 HG010359, P50 HG007735, and NSF DMS 1952386. Z.G. and R.J. were supported by the National Key Research and Development Program of China [2021YFF1200902; 2023YFF1204802], the Beijing Natural Science Foundation [L242026], and the National Natural Science Foundation of China [62273194].

Data availability
We provide the detailed information of multiple chromatin signals of whole genome, motif binding status, and expression data of 711 TFs in the corresponding cell lines/tissues, which are used in EpiGePT. Specifically, information on the processed data related to DNase-seq predictions is provided in Additional file 2: Tables S6–S8. For predictions based on multiple epigenomic profiles using the hg19 reference genome, the processed data information is provided in Additional file 2: Tables S9–S11. For predictions using the hg38 reference genome, the processed experimental data information is provided in Additional file 2: Tables S12–S14. The high-throughput validated silencers of K562 cell line are download from SilencerDB (http://health.tsinghua.edu.cn/silencerdb) database [29]. The HiChIP data for training EpiGePT-3D and E-P interaction prediction are downloaded from HiChIPdb (http://health.tsinghua.edu.cn/hichipdb) database (FitHiChIP 5 k loops) [31]. The processed DNase-seq peak and ATAC-seq peak data are obtained from the SilencerDB (http://health.tsinghua.edu.cn/silencerdb/analysis.php). Enhancer-gene pairs of CRISPRi [28] experiments are obtained from the supplementary information of Gasperini et al. [27] and Fulco et al. [28]. The regulatory network data for transcription factors and target genes were obtained from the TRRUST [40] database (https://www.grnpedia.org/trrust/) and the GRNdb [39] database (http://www.grndb.com). The annotated chromatin states for whole genome are downloaded from the ROADMAP epigenomics project (https://egg2.wustl.edu/roadmap/web_portal/chr_state_learning.html) [68]. The RNA-seq read counts matrix for protein coding genes (Additional file 2: Table S15-S16) used for the prediction of the chromatin 15-states annotated by ChromHMM are downloaded from the ROADMAP [68] project (https://egg2.wustl.edu/roadmap/data/byDataType/rna/expression/57epigenomes.N.pc.gz). The eQTL data are collected from the supplementary data 3 of Wang et al. [42]. The GWAS data of COVID-19 are download from the Release 4 of COVID-19 Host Genetics Initiative [47] (https://www.covid19hg.org/). In the section of regulatory relationships between TFs and target genes, the DNase-seq data for mouse are obtained from the ENCODE project [21] (ENCSR000COE, ENCSR000CNM, ENCSR000CNG), the narrow peak files are obtained from the ENCODE project [21] (ENCFF941CCB, ENCFF268DLZ, ENCFF587XGH), and the RNA-seq data also came from the ENCODE project [21] (ENCFF455OZW, ENCFF-641SMI, ENCFF876OPZ). In the ablation experiments addressing missing TFs, the expression data from different samples were sourced from the ENCODE project [21] (ENCSR029FTY, ENCSR410DUZ, ENCSR899SWV, ENCSR436ZKE). We also provide the processed data for model training and pretrained models at the download page of EpiGePT-online (http://health.tsinghua.edu.cn/epigept/download.php). The code of EpiGePT is freely available at GitHub (https://github.com/ZjGaothu/EpiGePT) [69] and Zenodo [70] under the MIT license.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors have declare that they have no competing interests.

Received: 27 March 2024 Accepted: 28 November 2024
Published online: 18 December 2024

References

1. Preissl S, Gaulton KJ, Ren B. Characterizing cis-regulatory elements using single-cell epigenomics. Nat Rev Genet. 2023;24:21–43.
2. O'Malley RC, Huang S-sC, Song L, Lewsey MG, Bartlett A, Nery JR, Galli M, Gallavotti A, Ecker JR. Cistrome and epicistrome features shape the regulatory DNA landscape. Cell. 2016;165:1280–92.

3. Vandereyken K, Sifrim A, Thienpont B, Voet T. Methods and applications for single-cell and spatial multi-omics. Nat Rev Genet. 2023;24:1–22.
4. Wang KC, Chang HY. Epigenomics: technologies and applications. Circ Res. 2018;122:1191–9.
5. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning–based sequence model. Nat Methods. 2015;12:931–4.
6. Xu C, Liu Q, Zhou J, Xie M, Feng J, Jiang T. Quantifying functional impact of non-coding variants with multi-task Bayesian neural network. Bioinformatics. 2020;36:1397–404.
7. Liu Q, Gan M, Jiang R. A sequence-based method to predict the impact of regulatory variants using random forest. BMC Syst Biol. 2017;11:7.
8. Chen KM, Wong AK, Troyanskaya OG, Zhou J. A sequence-based global map of regulatory activity for deciphering human genetics. Nat Genet. 2022;54:940–9.
9. Sahu B, Hartonen T, Pihlajamaa P, Wei B, Dave K, Zhu F, Kaasinen E, Lidschreiber K, Lidschreiber M, Daub CO, et al. Sequence determinants of human gene regulatory elements. Nat Genet. 2022;54:283–94.
10. Liu Q, Xia F, Yin Q, Jiang R. Chromatin accessibility prediction via a hybrid deep convolutional neural network. Bioinformatics. 2018;34:732–8.
11. Song S, Cui H, Chen S, Liu Q, Jiang R. EpiFIT: functional interpretation of transcription factors based on combination of sequence and epigenetic information. Quant Biol. 2019;7:233–43.
12. Wang J, Cheng Z, Yao Q, Liu L, Xu D, Hu G. Bioinformatics and biomedical informatics with ChatGPT: year one review. Quant Biol. 2024;12:345–59.
13. Zhang S, Fan R, Liu Y, Chen S, Liu Q, Zeng W. Applications of transformer-based language models in bioinformatics: a survey. Bioinform Adv. 2023;3:vbad001.
14. Ji Y, Zhou Z, Liu H, Davuluri RV. DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. Bioinformatics. 2021;37:2112–20.
15. Zhou Z, Ji Y, Li W, Dutta P, Davuluri R, Liu H. Dnabert-2: efficient foundation model and benchmark for multi-species genome. arXiv preprint arXiv:2306.15006. 2023;2023–06. https://arxiv.org/abs/2306.15006.
16. Avsec Ž, Agarwal V, Visentin D, Ledsam JR, Grabska-Barwinska A, Taylor KR, Assael Y, Jumper J, Kohli P, Kelley DR. Effective gene expression prediction from sequence by integrating long-range interactions. Nat Methods. 2021;18:1196–203.
17. Nair S, Kim DS, Perricone J, Kundaje A. Integrating regulatory DNA sequence and gene expression to predict genome-wide chromatin accessibility across cellular contexts. Bioinformatics. 2019;35:i108–16.
18. Liu Q, Hua K, Zhang X, Wong WH, Jiang R. DeepCAGE: incorporating transcription factors in genome-wide prediction of chromatin accessibility. Genomics Proteomics Bioinformatics. 2022;20:496–507.
19. Zhou W, Sherwood B, Ji Z, Xue Y, Du F, Bai J, Ying M, Ji H. Genome-wide prediction of DNase I hypersensitivity using gene expression. Nat Commun. 2017;8:1–17.
20. Song L, Crawford GE. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. Cold Spring Harbor Protocols. 2010;2010:pdb.prot5384.
21. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57.
22. Ernst J, Kellis M. Chromatin-state discovery and genome annotation with ChromHMM. Nat Protoc. 2017;12:2478–92.
23. Breschi A, Gingeras TR, Guigó R. Comparative transcriptomics in human and mouse. Nat Rev Genet. 2017;18:425–40.
24. Chen S, Liu Q, Cui X, Feng Z, Li C, Wang X, Zhang X, Wang Y, Jiang R. OpenAnnotate: a web server to annotate the chromatin accessibility of genomic regions. Nucleic Acids Res. 2021;49:W483–90.
25. Gao Z, Jiang R, Chen S. OpenAnnotateApi: Python and R packages to efficiently annotate and analyze chromatin accessibility of genomic regions.Bioinformatics. Advances. 2024;4:vbae055.
26. Larson MH, Gilbert LA, Wang X, Lim WA, Weissman JS, Qi LS. CRISPR interference (CRISPRi) for sequence-specific control of gene expression. Nat Protoc. 2013;8:2180–96.
27. Gasperini M, Hill AJ, McFaline-Figueroa JL, Martin B, Kim S, Zhang MD, Jackson D, Leith A, Schreiber J, Noble WS. A genome-wide framework for mapping gene regulation via cellular genetic screens. Cell. 2019;176(377–390): e319.
28. Fulco CP, Nasser J, Jones TR, Munson G, Bergman DT, Subramanian V, Grossman SR, Anyoha R, Doughty BR, Patwardhan TA. Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. Nat Genet. 2019;51:1664–9.
29. Zeng W, Chen S, Cui X, Chen X, Gao Z, Jiang R. SilencerDB: a comprehensive database of silencers. Nucleic Acids Res. 2021;49:D221–8.
30. Mumbach MR, Rubin AJ, Flynn RA, Dai C, Khavari PA, Greenleaf WJ, Chang HY. HiChIP: efficient and sensitive analysis of protein-directed genome architecture. Nat Methods. 2016;13:919–22.
31. Zeng W, Liu Q, Yin Q, Jiang R, Wong WH. HiChIPdb: a comprehensive database of HiChIP regulatory interactions. Nucleic Acids Res. 2023;51:D159–66.
32. Fishilevich S, Nudel R, Rappaport N, Hadar R, Plaschkes I, Iny Stein T, Rosen N, Kohn A, Twik M, Safran M. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. Database. 2017;2017:bax028.
33. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. 2018.
34. Devlin J, Chang MW, Lee K, Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. 2018;2018–10. https://arxiv.org/abs/1810.04805.
35. Li W, Wong WH, Jiang R. DeepTACT: predicting 3D chromatin contacts via bootstrapping deep learning. Nucleic Acids Res. 2019;47:e60–e60.
36. van den Berg DL, Snoek T, Mullin NP, Yates A, Bezstarosti K, Demmers J, Chambers I, Poot RA. An Oct4-centered protein interaction network in embryonic stem cells. Cell Stem Cell. 2010;6:369–81.
37. Zhang J, Cao H, Xie J, Fan C, Xie Y, He X, Liao M, Zhang S, Wang H. The oncogene Etv5 promotes MET in somatic reprogramming and orchestrates epiblast/primitive endoderm specification during mESCs differentiation. Cell Death Dis. 2018;9:224.
38. Levy SH, Cohen SF, Arnon L, Lahav S, Awawdy M, Alajem A, Bavli D, Sun X, Buganim Y, Ram O. Esrrb is a cell-cycle-dependent associated factor balancing pluripotency and XEN differentiation. Stem Cell Reports. 2022;17:1334–50.
39. Fang L, Li Y, Ma L, Xu Q, Tan F, Chen G. GRNdb: decoding the gene regulatory networks in diverse human and mouse conditions. Nucleic Acids Res. 2021;49:D97–103.

40. Han H, Cho J-W, Lee S, Yun A, Kim H, Bae D, Yang S, Kim CY, Lee M, Kim E. TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. Nucleic Acids Res. 2018;46:D380–6.
41. Willett R, Martina JA, Zewe JP, Wills R, Hammond GR, Puertollano R. TFEB regulates lysosomal positioning by modulating TMEM55B expression and JIP4 recruitment to lysosomes. Nat Commun. 2017;8:1580.
42. Wang QS, Kelley DR, Ulirsch J, Kanai M, Sadhuka S, Cui R, Albors C, Cheng N, Okada Y. Leveraging supervised learning for functionally informed fine-mapping of cis-eQTLs identifies an additional 20,913 putative causal eQTLs. Nat Commun. 2021;12:3394.
43. Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Hoover J. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. 2016;44:D862–8.
44. Karczewski KJ, Weisburd B, Thomas B, Solomonson M, Ruderfer DM, Kavanagh D, Hamamsy T, Lek M, Samocha KE, Cummings BB. The ExAC browser: displaying reference data information from over 60 000 exomes. Nucleic Acids Res. 2017;45:D840–5.
45. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 2019;47:D886–94.
46. Li J, Lai S, Gao GF, Shi W. The emergence, genomic diversity and global spread of SARS-CoV-2. Nature. 2021;600:408–18.
47. org C-HGlab. The COVID-19 host genetics initiative, a global initiative to elucidate the role of host genetic factors in susceptibility and severity of the SARS-CoV-2 virus pandemic. Eur J Hum Genet. 2020;28:715–8.
48. Wang W, Wang C-Y, Wang S-I, Wei JC-C. Long-term cardiovascular outcomes in COVID-19 survivors among non-vaccinated population: a retrospective cohort study from the TriNetX US collaborative networks. EClinicalMedicine. 2022;53:101619.
49. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, Zhang L, Fan G, Xu J, Gu X. Clinical features of patients infected with 2019 novel coronavirus in Wuhan China. The lancet. 2020;395:497–506.
50. Agius L. Targeting hepatic glucokinase in type 2 diabetes: weighing the benefits and risks. Diabetes. 2009;58:18–20.
51. Singh AK, Gupta R, Ghosh A, Misra A. Diabetes in COVID-19: prevalence, pathophysiology, prognosis and practical considerations. Diabetes Metab Syndr. 2020;14:303–10.
52. Pellegrina D, Bahcheli AT, Krassowski M, Reimand J. Human phospho-signaling networks of SARS-CoV-2 infection are rewired by population genetic variants. Mol Syst Biol. 2022;18: e10823.
53. d Galbraith E, sc Merleau N, mcdonald Smith B. The human cell count and size distribution. Proc Natl Acad Sci U S A. 2023;120:e2303077120.
54. CZI Cell Science Program, Abdulla S, Aevermann B, Assis P, Badajoz S, Bell SM, Bezzi E, Cakir B, Chaffer J, Chambers S. CZ CELLxGENE discover: a single-cell data platform for scalable exploration, analysis and modeling of aggregated data. Nucleic Acids Res. 2024. https://doi.org/10.1093/nar/gkae1142.
55. Loyfer N, Magenheim J, Peretz A, Cann G, Bredno J, Klochendler A, Fox-Fisher I, Shabi-Porat S, Hecht M, Pelet T. A DNA methylation atlas of normal human cell types. Nature. 2023;613:355–64.
56. Li S, Zeng W, Ni X, Liu Q, Li W, Stackpole ML, Zhou Y, Gower A, Krysan K, Ahuja P, et al. Comprehensive tissue deconvolution of cell-free DNA by deep learning for disease diagnosis and monitoring. Proc Natl Acad Sci U S A. 2023;120: e2305236120.
57. Gao Z, Chen X, Li Z, Cui X, Jiang Q, Li K, Chen S, Jiang R. scEpiTools: a database to comprehensively interrogate analytic tools for single-cell epigenomic data. J Genet Gen. 2024;51:462–5.
58. Liu Q, Lv H, Jiang R. hicGAN infers super resolution Hi-C data with generative adversarial networks. Bioinformatics. 2019;35:i99–107.
59. Liu Q, Zeng W, Zhang W, Wang S, Chen H, Jiang R, Zhou M, Zhang S. Deep generative modeling and clustering of single cell Hi-C data. Brief Bioinform. 2023;24:bbac494.
60. Yates AD, Allen J, Amode RM, Azov AG, Barba M, Becerra A, Bhai J, Campbell LI, Carbajo Martinez M, Chakiachvili M. Ensembl genomes 2022: an expanding genome resource for non-vertebrates. Nucleic Acids Res. 2022;50:D996–1003.
61. Basenko EY, Pulman JA, Shanmugasundram A, Harb OS, Crouch K, Starns D, Warrenfeltz S, Aurrecoechea C, Stoeckert CJ Jr, Kissinger JC. FungiDB: an integrated bioinformatic resource for fungi and oomycetes. Journal of Fungi. 2018;4: 39.
62. Alvarez-Jarreta J, Amos B, Aurrecoechea C, Bah S, Barba M, Barreto A, Basenko EY, Belnap R, Blevins A, Böhme U. VEuPathDB: the eukaryotic pathogen, vector and host bioinformatics resource center in 2023. Nucleic Acids Res. 2024;52:D808–16.
63. Kulakovskiy IV, Vorontsov IE, Yevshin IS, Sharipov RN, Fedorova AD, Rumynskiy EI, Medvedeva YA, Magana-Mora A, Bajic VB, Papatsenko DA. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. Nucleic Acids Res. 2018;46:D252–9.
64. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell. 2010;38:576–89.
65. Kelley DR, Reshef YA, Bileschi M, Belanger D, McLean CY, Snoek J. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. Genome Res. 2018;28:739–50.
66. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). 2016. p. 770–8. https://doi.org/10.1186/s13059-022-02799-4.
67. Chor B, Horn D, Goldman N, Levy Y, Massingham T. Genomic DNA k-mer spectra: models and modalities. Genome Biol. 2009;10:1–10.
68. Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, Kellis M, Marra MA, Beaudet AL, Ecker JR. The NIH roadmap epigenomics mapping consortium. Nat Biotechnol. 2010;28:1045–8.
69. Gao Z, Liu Q, Zeng W, Jiang R, Wong WH. EpiGePT: a pretrained transformer-based language model for context-specific human epigenomics. GitHub. 2024. https://github.com/ZjGaothu/EpiGePT.
70. Gao Z, Liu Q, Zeng W, Jiang R, Wong WH. EpiGePT: a pretrained transformer-based language model for context-specific human epigenomics. Zenodo. 2024. https://doi.org/10.5281/zenodo.14201753.

Publisher's Note