Published online 10 May 2022

Nucleic Acids Research, 2022, Vol. 50, No. 14 ***e81***
https://doi.org/10.1093/nar/gkac326

Integrating convolution and self-attention improves language model of human genome for interpreting non-coding regions at base-resolution

Meng Yang[1,2,*,†], Lichao Huang[1,†], Haiping Huang[1], Hui Tang[1], Nan Zhang[1], Huanming Yang[6,7], Jihong Wu[3,4,5,*] and Feng Mu[1,*]

[1]MGI, BGI-Shenzhen, Shenzhen 518083, China, [2]Department of Biology, University of Copenhagen, Copenhagen DK-2200, Denmark, [3]Department of Ophthalmology, Eye & ENT Hospital, Shanghai Medical College, Fudan University, Shanghai, China, [4]Shanghai Key Laboratory of Visual Impairment and Restoration, Science and Technology Commission of Shanghai Municipality, Shanghai, China, [5]Key Laboratory of Myopia (Fudan University), Chinese Academy of Medical Sciences, National Health Commission, Shanghai, China, [6]BGI-Shenzhen, Shenzhen 518083, China and [7]Guangdong Provincial Academician Workstation of BGI Synthetic Genomics, BGI-Shenzhen, Shenzhen, 518120, China

Received September 22, 2021; Revised February 22, 2022; Editorial Decision April 19, 2022; Accepted May 09, 2022

ABSTRACT

Interpretation of non-coding genome remains an unsolved challenge in human genetics due to impracticality of exhaustively annotating biochemically active elements in all conditions. Deep learning based computational approaches emerge recently to help interpret non-coding regions. Here, we present LOGO (Language of Genome), a self-attention based contextualized pre-trained language model containing only two self-attention layers with 1 million parameters as a substantially light architecture that applies self-supervision techniques to learn bidirectional representations of the unlabelled human reference genome. LOGO is then fine-tuned for sequence labelling task, and further extended to variant prioritization task via a special input encoding scheme of alternative alleles followed by adding a convolutional module. Experiments show that LOGO achieves 15% absolute improvement for promoter identification and up to 4.5% absolute improvement for enhancer-promoter interaction prediction. LOGO exhibits state-of-the-art multi-task predictive power on thousands of chromatin features with only 3% parameterization benchmarking against the fully supervised model, DeepSEA and 1% parameterization against a recent BERT-based DNA language model. For allelic-effect prediction, locality introduced by one dimensional convolution shows improved sensitivity and specificity for prioritizing non-coding variants associated with human diseases. In addition, we apply LOGO to interpret type 2 diabetes (T2D) GWAS signals and infer underlying regulatory mechanisms. We make a conceptual analogy between natural language and human genome and demonstrate LOGO is an accurate, fast, scalable, and robust framework to interpret non-coding regions for global sequence labeling as well as for variant prioritization at base-resolution.

INTRODUCTION

In 2003, the Human Genome Project (HGP) successfully digitalized the ‘book of life’. It is convinced that biological structure and function are intrinsically encoded in the primary genome sequence. The non-coding regions, accounting for over 98% of the whole genome, implement significant yet largely unknown regulatory functions. Recent large consortia projects, including the ENCyclopedia of DNA Elements (ENCODE) (1,2), Roadmap Epigenomics (3), and the Genomics of Gene Regulation (GGR), have produced large amount of experimental mapping readouts to help annotate non-coding genome in specific tissues or cell-lines. On the other hand, Genome-wide association studies (GWAS) have discovered that the vast majority (>90%) of associated genome loci for complex disease and traits fall in non-coding regions (4). Hence, it is of exceptional utility to explore these datasets and derive novel hypothe-

[*] To whom correspondence should be addressed. Tel: +86 0755 36352505; Email: yangmeng1@genomics.cn
Correspondence may also be addressed to Jihong Wu. Tel: +86 021 64377134; Email: jihongwu@fdu.edu.cn
Correspondence may also be addressed to Feng Mu. Tel: +86 0755 36352505; Email: mufeng@mgi-tech.com
[†] The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

sis to interpret non-coding regions. Unlike the protein coding region where there is a clear genetic code, incorporating broader sequence context is critical to understand functional effects of regulatory variants, which requires more powerful and semantic-rich representational model to capture higher-order complexity in the region. Deep learning has transformed ranges of tasks in computer vision and natural language processing (NLP). In bioinformatics field, deep learning based computational methods have also been proposed in various applications, such as predicting molecular phenotypes based on raw DNA sequence as input and achieving better performance than traditional machine learning approaches, as referred to an excellent review paper (5). One classical model is DeepSEA (6), pioneering to apply deep convolutional neural network (CNN) architecture to extract features of genome sequence given 1000-bp context and train on chromatin profiles in a supervised multitask learning manner. DeepSEA can able to predict the binary presence or absence of 919 chromatin marks.

The inherent sequential nature of the genome is analogous to documents composed of words, characters and phrases. The exciting advance in the NLP field has shed light on using similar strategy to extract general and transferable information from biological sequence. Neural network model was introduced into NLP since 2013. Word2vec (7) was proposed to learn distributional vector embeddings of each word to capture their similarities given the sentence context. Word2vec uses multilayer perceptron (MLP) (8) to predict neighboring words given center word (called skip-gram) or predict center word given neighboring words (called 'Continuous Bag of Words', CBOW). The learned word vectors can then be directly queried for downstream text classification tasks. Word2vec essentially relies on modelling co-occurrence probabilities without considering word position information and static embeddings cannot handle words with multiple meanings, so-called polysemous words. Traditional CNN-based feature extractors rely on local parameter sharing and the pooling operation may lead to loss of global information. Recurrent Neural Network (RNN) (9,10) is an alternative architecture to process sequential data. RNN can capture position dependency information via passing the memory state from previous elements. RNN's fundamental constraint of sequential operation leads to difficulty of parallelization and faces the risk of vanishing gradient when processing longer sequence. In 2017, Transformer (11) has emerged as a powerful architecture that relies completely on attention mechanism to draw global dependencies in Seq2Seq modelling task. In the encoder part, self-attention mechanism relates different positions across a single sequence to compute a contextualized representation with better parallelization. Transformer can tackle long-range dependency without position bias, outperforming CNNs or RNNs in many global sequence classification tasks. On the other hand, CNN is better at capturing locality.

Since 2018, a new wave of pre-trained language models using self-supervision techniques have emerged as a core trend in NLP, including RNN-based ELMo (12), ULMFiT (13), Transformer-based OpenAI-GPT (14) and Google-BERT (15). Instead of conventional left-to-right unidirectional modeling, BERT, which stands for Bidirectional Encoder Representations from Transformers, leveraged a multilayer bidirectional Transformer architecture to pre-train on large unlabeled corpora by jointly incorporating both left and right contexts. The pre-trained model learns contextualized token embeddings through two proxy training objectives: MLM (Masked Language Model), predicting randomly masked tokens and NSP (Next Sentence Prediction), predicting whether two sentences follow each other. The pre-trained BERT can then be easily fine-tuned to various downstream NLP tasks and obtain new state-of-the-art results competing with human performance. Thereafter, a series of pre-trained models spring up to further improve performance, such as XLNet (16), UniLM (17), MASS (18), MT-DNN (19), XLM (20), ALBERT (21), RoBERTa (22) and ELECTRA (23). A comprehensive review can be found in an integrative reference (24). One representative model, ALBERT, a lite version of BERT, establishes better results with significantly reduced model size through factorized embeddings and cross-layer parameter sharing techniques. Unlike models trained on general domain corpora, SciBERT (25) and BioBERT (26) are proposed based on BERT backbone and trained on a large amount of multidisciplinary scientific literature and biomedical text corpus, respectively. The domain-specific BioBERT achieves dramatic improvement in biomedical text-mining tasks, such as name entity recognition, relation extraction and question answering. BioBERT is comprised of 12 layers with hidden size of 768, 12 heads and 12 attention heads in each layer. Recently, Transformer was reported by Facebook to learn protein structure and function (27). DNABERT (28) is recently proposed to learn the human genome and is composed of 12 Transformer layers with 768 hidden units and 12 attention heads in each layer, which is configured as a heavy version of BERT. DNABERT is fine-tuned on several functional sequence recognition tasks and splicing site identification. However, features learned by Transformer are too general and not specific or sensitive to a single or few changes in the sequence, unable to satisfy the needs of interpreting human genome at base-resolution. Recently, Facebook and Google both propose introducing the concept of convolution into Transformer architecture to bring soft inductive bias with better locality, namely ConViT (29) and CoAtNets (30).

Motivated by these observations, in this study we develop LOGO, a pre-trained language model with much lighter architecture than the pioneering DNABERT, which is composed of only two layers with 256 hidden units and 8 heads (embedding size is set to 128), to learn contextualized representations of reference genome hg19. The main intuition to choose these hyperparameters is to keep LOGO as lightweight as possible to save GPU memories without compromising performance. We implement ablation studies covering both pre-training and downstream fine-tuning tasks and demonstrate how to determine the optimal choice of hyperparameters. Details can be found in Supplementary Tables S2–S4. LOGO with 3-mer tokenization contains around 1 million parameters while DNABERT contains 100 million parameters. DNABERT leverages span masking strategy and consumes 25 days on 8 RTX2080TI GPUs to implement pre-training. LOGO has much faster pre-training speed, both due to lighter ALBERT-like structure as well as a random masking strategy following dividing the tokenized sequence into k-mers k-stride groups. LOGO accepts fixed-length DNA sequence input (1000 and 2000 bp) while DNABERT uses variable length input ranging from 5 bp to 500 bp. LOGO shows substantially more effective parameter efficiency than DNABERT. To demonstrate the versatility of LOGO, we implement fine-tuning for multiple downstream tasks and obtain excellent performance from aspects of accuracy, speed, scalability, and robustness. Sequence-level classification tasks include promoter prediction, promoter-enhancer interaction prediction and chromatin features prediction. Another key innovation of LOGO is that we introduce a novel encoding scheme for alternative alleles and leverage a hybrid architecture by mixing convolution and self-attention to alleviate the locality-insensitivity issue of Transformer and facilitate functional prioritization of noncoding variants. DNABERT only reports high-attention variants extracted from Transformer encoder while LOGO leverages convolution operation and forces the model to see the nucleotide change with allelic-effects. We also propose a framework to embed prior knowledge into LOGO and explore better performance over original settings. LOGO provides a unified framework not only for sequence labelling or motif identification task, but also for SNP or indel prioritization to interpret non-coding regions at base-resolution.

MATERIALS AND METHODS

Architecture of the pre-training model

The pre-training model leverages the encoder part of Transformer architecture and learns representations of input sequences via multi-head self-attention mechanism. We follow the BERT notation conventions and denote the vocabulary embedding size as E, the number of encoder layers as L, and the hidden size as H. Each training instance is started with a 100-bp bin as described above and extended forward along the reference genome until reaching 1000-bp length. The model processes the genome into sequential segments of k-mer tokens, and each token is labelled by a unique vocabulary ID as input. The size of token embedding has length $E = 128$. Token embeddings are summed with position embeddings and fed into Transformer encoder network. We leverage the ALBERT strategy to untie the input token embedding size E from the hidden layer size H in the Transformer encoder. The encoder is composed of a stack of $L = 2$ identical layers. Each layer has two sub-layers. The first is a multi-head self-attention layer, and the second is a position-wise fully connected feed-forward layer. A residual connection is added to each sub-layer, followed by layer normalization, leading to output of each sub-layer being LayerNorm (x + Sublayer(x)), where Sublayer(x) is the function implemented by the sub-layer itself. Each hidden sub-layer produces vector outputs with dimension of H = 256 = dmodel. An attention function can be described as mapping a query and a set of key-value pairs to an output, where the queries (Q), keys (K), values (V), and output are all vectors. The output is computed as a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key. For each self-attention layer, the input consists of queries and keys of dimension d_k, and values of dimension d_v. The dot products of the query with all keys, divided by $\sqrt{d_k}$, are fed into a $softmax$ layer to obtain the weights on the values. The attention functions on a set of queries are computed simultaneously, packed together into a matrix Q. The keys and values are also packed together into matrices K and matrices V. The matrix of outputs is computed as:

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

To increase the capacity of the model, the input of each hidden layer is processed by multiple attention heads, which means on each projected version of queries, keys and values, the attention function is performed A(number of heads) times in parallel. The outputs of each head are concatenated and projected, resulting in the final values, as depicted:

$$\begin{aligned} Multi\,Head(Q, K, V) &= Concat(head_1, \cdots, head_h)W^O, \\ &\times where\ \ head_i \\ &= Attention(QW_i^Q, KW_i^K, VW_i^V) \end{aligned}$$

$$W_i^Q \in \mathbb{R}^{d_k \times d_{model}},\ \ W_i^K \in \mathbb{R}^{d_k \times d_{model}},\ \ W_i^V \in \mathbb{R}^{d_v \times d_{model}},\ \ W^O \in \mathbb{R}^{hd_v \times d_{model}}$$

We employ $A = 8$ heads. For each of these, we use $d_k = d_v = \frac{d_{model}}{A} = 32$. Due to the reduced dimension of each head, the total computational cost is close to that of single-head attention with full dimensionality. Following each attention layer, the fully connected feed-forward network is applied to each position separately and identically, which consists of two linear transformations with a $ReLU$ activation in between.

$$FFN\ (x) = \max(0,\ xW_1 + b_1)\,W_2 + b_2$$

After a forward pass through $L = 2$ layers, a final classification layer is used to project the hidden state (d_{model}) to output classes of dimension equal to k-mer vocabulary size.

Motivated by ALBERT architecture, we use a factorization of token embedding parameters. By using this decomposition, the embedding parameters are reduced from $O(V \times H)$ to $O(V \times E + E \times H)$. We also enforce sharing all parameters across two layers motivated by improved parameter efficiency of ALBERT. In original BERT model, for a given token, its input representation is constructed by summing the corresponding token, segment, and position embeddings. Position embeddings are used to capture relative positions of each input token within a sequence. Since we discard NSP task, we do not use segment embeddings in pre-training stage. One contribution of this work is that we demonstrate a method to incorporate prior knowledge into the language model. Knowledge layer is introduced and encoded in one-hot format. For example, if we have M knowledge items to label the input sequence, then a M-dimension one-hot knowledge vector is introduced and concatenated with input sequence vectors. For example, if a sequence is labelled by an annotation knowledge, all k-mers spanning from annotation start position to end position will be recorded as '1' for this type of knowledge, and k-mers of other positions will be recorded as '0'. Knowledge embeddings are learned by the model and the dimensions are set as the same as token embeddings size. In this study, knowledge embeddings are only used in the fine-tuning stage in promoter prediction task.

Pre-training

We define similar self-supervised loss for Masked Language Model [MLM] pre-training task as in BERT and discard Next Sentence Prediction (NSP) task. In the pre-training stage, to balance the computation burden and representation utility, we generate and evaluate four types of *k*-mer ($k = 3, 4, 5, 6$) with 1-stride settings to tokenize the genome (Data generation and tokenization in Supplementary Text S1). For each *k*-mer setting of 1000-bp sequence, *k* sets of input will be all used as training instances. For each set, we use similar masking strategy as in BERT. The masked token will be represented as [MASK]. We randomly masked 15% of *k*-mer tokens for prediction, 80% of which are replaced with [MASK], 10% are replaced by a random token from the vocabulary and another 10% remain unchanged. The original token at masked position will be predicted with cross-entropy loss. The pre-training loss is the sum of the mean masked LM likelihood. We follow the BERT notation conventions and denote the vocabulary embedding size as E, the number of encoder layers as L, and the hidden size as H. In LOGO model, each *k*-mer of input sequence will be represented as 128-dimension vocabulary token embedding vectors. The embedding size of hidden layers is set to be larger than input token embedding size as in ALBERT, since hidden layers are meant to learn context-dependent representations. All embeddings and model weights are expected to be learned by the model from MLM task. We use four Nvidia Tesla V100 SXM3 32G GPU to train the model. Because the number of training records exceeds 180 million (3-mer: 60 million×3, 4-mer: 60 million×4, 5-mer: 60 million×5, 6-mer: 60 million×6), to speed up training, we convert all data to Tensorflow tfrecord and adopt Tensorflow's '*Multi Worker Mirrored Strategy*' strategy to support multi-machine and multi-GPU training. Parallel training technique is used on four GPUs to support large batch size. Hyperparameters are summarized as below: layers (L) = 2; token embedding size (E) = 128; hidden dimension size (H) = 256; attention heads (A) = 8; batch size (BSZ) = 512 for each GPU, 512×4 = 2048 for 4 GPUs; steps-per-epoch = 4000; maximum epochs = 100; sequence length = 333, 250, 200, 166 tokens for 3-mer, 4-mer, 5-mer, 6-mer setting, respectively to encode 1000-bp input sequence. We use an Adam optimizer with learning rate = 0.00001. Other hyperparameters are set as default in ALBERT.

Analysis of T2D-related GWAS variants

We download all T2D-associated SNPs from GWAS Catalog (2020–05-14 version) and obtain corresponding LD SNPs from LDlink, resulting in 156 175 SNPs (*P*-value ranging from 9×10^{-6} to 6×10^{-447}). To make fair comparison with DeepSEA, we use the same approach to compute chromatin effects of variants. For each SNP, we extract the 1000-bp or 2000-bp context sequence centered on that variant based on hg19 reference genome (SNPs locates at the 500th position). A pair of 1000-bp sequences centered on both reference allele and alternative allele at the variant position are used to calculate the probabilities for each chromatin feature. Absolute differences between probability values and relative log fold changes of odds are calculated following DeepSEA pipelines. Both forward and complementary sequences are computed and averaged to obtain the predicted chromatin effects. The magnitude of the predicted chromatin effect on a chromatin feature for an SNP is computed as the product of the absolute difference between probability values and the relative log fold change of odds. We use the same protocol as in DeepSEA to obtain negative non-functional SNPs which contains 1 000 000 negative SNPs randomly chosen from 1000 Genomes Project. We calculate chromatin effects for these negative SNPs to generate the empirical background distribution and use the same E-value definition to evaluate significance of variant effects. For each chromatin feature, E-value is computed as the proportion of negative SNPs with higher predicted chromatin effect magnitude on the same chromatin feature. We use fine-tuned LOGO-919, LOGO-2002 (Pretrained LOGO with 2000-bp context fine-tuned on 2002 chromatin features from ExPecto, $n = 690$ TF, 334 DHSs and 978 HM, respectively) and LOGO-3357 (Pretrained LOGO with 2000-bp context fine-tuned on 3357 chromatin features after integrating EpiMap features with ExPecto features, $n = 826$ TF, 668 DHSs and 1863 HM, respectively) to calculate E-values for 919, 2002 and 3357 chromatin features, respectively (data details in Supplementary Text S4 and Supplementary Text S5). A variant is considered as putative functional significant if at least one chromatin feature's E-value is equal or less than a certain threshold, i.e. 1×10^{-5}, which might be used to infer underlying regulator disruption mechanism. Profile of Thurner islet chromatin states is downloaded from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5828664/bin/elife-31977-fig3-data5.zip. Profile of Varshney islet chromatin states is downloaded from https://theparkerlab.med.umich.edu/data/papers/doi/10.1073/pnas.1621192114/ after consultation with Dr. Narisu from Francis Collins Lab via email.

Model architecture for LOGO-E2E

To fine-tune on variant prioritization task in an end-to-end manner, we modify LOGO architecture to accommodate signed allelic information. We stack three layers to encode input sequence. The first layer is called 'Ref layer'. We tokenize each 1000-bp context sequence extracted from hg19 reference genome using 6-mer-1-stride and feed it into 'Ref layer' via concatenating six sets of 6-mer [Ref] tokens in an interlaced manner. This novel operation is introduced to encode input sequence at base-resolution without compromising representation capacity of *k*-mer strategy. The second layer is called 'Alt' layer, we use this layer to encode allelic information at certain position. Only changed position compared to 'Ref layer' will have input value with corresponding 6-mer [Alt] token, other positions corresponding to the context sequence are set to [zero]. In this way, we explicitly encode the alternative allele to enforce the model to see directional alteration. The third layer is called ‘Type’ layer to encode variant type. In this paper, we do not evaluate SNVs and indels simultaneously, so we set [Type] token at corresponding position equal to 1 and other positions are set to [zero]. Each variant with surrounding context of certain length will be encoded as a matrix input containing ‘Ref’, ‘Alt’ and ‘Type’ information, which is formatted as a ‘npz’ compressed file. One-dimension convolutional layer is added after token embeddings and then fed into Transformer architecture. Three kernels with different sizes (2,3,5) are introduced to capture multi-scale features. Through experiments, it is found that this method reduces the weight updating frequency and makes the fluctuation range more stable during the fine-tuning process of the model. The final hidden state corresponding to learned [CLS] token embeddings is followed by a classification layer with sigmoid output of deleteriousness effect of target variant. Binary cross entropy loss is used to calculate loss function. We download LOGO-919 model weights and perform LOGO-E2E fine-tuning on HGMD training sets using 1 Nvidia TITAN Xp Pascal GPU. We use batch size of 64 and $L = 2$, $E = 128$, $H = 256$, $A = 8$. We use Adam optimizer with initial learning rate of 0.00001, and other parameters are set as default, and use early stopping strategy and stop training when validation loss no longer decreases for three consecutive epochs.

Model architecture for LOGO-C2P

For LOGO-C2P, one-dimension convolutional layer is added after token embeddings and then fed into Transformer architecture. Three kernels with different sizes (2,3,5) are introduced to capture multi-scale features. We follow similar pipelines as in DeepSEA for functional SNP prioritization architecture and firstly use previously trained LOGO-919/LOGO-2002 to generate chromatin effect features for both reference and alternative alleles. We then conduct the same absolute difference and relative log fold change transformation as DeepSEA and feed these features into boosted model to train the classifier at the second stage. We assess different model performances of whether or not preserving four base-level evolutionary feature used by DeepSEA, including PhastCons scores for primates (excluding humans), PhyloP scores for primates (excluding humans), and GERP++ neutral evolution and rejected substitution scores. We use well-trained LOGO-919, LOGO-2002 and DeepSEA to generate chromatin features for each target variant, convert these features into DMatrix format, and train a regularized logistic regression model, using the XGBoost v0.9 implementation (https://github.com/tqchen/xgboost). It is worth mentioning that we discard z-score transformation as used in DeepSEA classifier. We argue that the tree-based approach does not require normalization as stated by original XGBoost author. The model is trained with L1 regularization parameter (alpha) = 20 and L2 regularization parameter (lambda) = 2000 for iterations = 1000. Other hyperparameters are set as: Step-size shrinkage parameter(eta):0.1booster:‘gbtree’, objective:‘binary:logistic’, loss:‘error’. We set early stopping rules when validation error no longer decreases for 200 epochs and preserve the best model weight. 1 Nvidia TITAN Xp Pascal GPU is used.

Benchmarking of classifier performance on HGMD, ClinVar and GWAS variants

For HGMD regulatory variants, the performance of each model is estimated by 10-fold cross-validation. For filtered 3498 regulatory variants, we construct negative controls from 1000 Genomes Project SNPs using two schemes: random sampling (unrestricted, 3690 negatives), and matched to positive ones within 1 kb (restricted, 3034 negatives). We fine-tune LOGO-E2E and LOGO-C2P on the HGMD dataset. We also retrain DeepSEA based classifier on the HGMD dataset with or without incorporating four evolutionary features. CADD-precomputed scores are downloaded from http://krishna.gs.washington.edu/download/CADD/v1.3/whole_genome_SNVs.tsv.gz, FunSeq2 precomputed scores are downloaded from http://org.gersteinlab.funseq.s3-website-us-east-1.amazonaws.com/funseq2.1.2/hg19_NCscore_funseq216.tsv.bgz, GERP precomputed scores are downloaded from http://mendel.stanford.edu/SidowLab/downloads/gerp/hg19.GERP_scores.tar.gz, LINSIGHT precomputed scores are downloaded from http://genome-mirror.cshl.edu/, CDTS metrics are downloaded from http://www.hli-opendata.com/noncoding. DeepSEA functional scores are computed locally based on 919 chromatin effect predictions and four evolutionary information–derived scores. DeepSEA functional significance score for a variant is defined as the product of the geometric mean E-value for predicted chromatin effects and the geometric mean *E*-value for evolutionary conservation features. We also assess DeepSEA functional score without 4 evolutionary features. The direction of different scores for all metrics is modified to ensure lower rank represents higher probability of pathogenicity. For held-out ClinVar test set, negative controls are subsampled by bootstrapping 10 times. For held-out GWAS variants, positive samples are subsampled by bootstrapping 10 times. To compare all methods, we compute false-positive versus true-positive rates for the complete range of score thresholds. Area under the receiver operating characteristic (AUROC) is used for benchmarking. Data processing can be found in Supplementary Text S6.

Benchmarking of classifier performance on CADD indels

We download the dataset from CADD Developmental release: v1.4, resulting in 3 675 207 indels. This dataset is less biased and contains much larger examples than manually curated ClinVar or HGMD. CADD is partially trained on this dataset, containing 1 837 708 *proxy-neutral* variants and 1 837 499 simulated *de novo proxy-deleterious* variants. The former sets emerge since the last human-ape ancestor and are fixed in human populations, which are considered neutral (or, at most, weakly deleterious). The latter are considered free of selective pressure including both neutral and deleterious indels. We fine-tune LOGO-E2E on these CADD Indels and use the expert-curated dataset as held-out test set. We download known indels from NCBI/NIH

ClinVar database (2020-10-03 release), which leads to 5556 pathogenic (defined as pathogenic and likely pathogenic in ClinVar) and 313 benign indels (defined as benign and likely benign in ClinVar), respectively. Due to class imbalance, we subsample positive indels five times to construct balanced test sets and benchmark against CADD, LINSIGHT and DeepSEA. Area under the receiver operating characteristic (AUROC) is used to benchmark different methods.

RESULTS

LOGO learns contextualized representation of k-mers of human reference genome and achieves state-of-the-art performance in promoter prediction task

The backbone of LOGO processing flow is motivated by recently emerged Transformer-based bidirectional encoder model (15,21) (Figure 1A). We conduct pre-training on human reference genome hg19 comprised of totally 3 billion base pairs. We segment both forward and complementary chain of whole genome sequence into 100-bp bins and get 60 million segments. For each bin, we extend forward to 1000-bp along the genome to create training instances, which are analogous to input sentences in the field of natural language.

Conventional one-hot encoded representation for each nucleotide has limited vocabulary size of five characters (i.e. A, G, C, T and Unknown/Undetected), which is considered as a semantically poor representation. k-mer encodes sequence into a certain length of successive nucleotides. For instance, a trinucleotide is a k-mer for which $k = 3$. To increase the information content, we tokenize each sequence in the way of *k*-mer representation. The intuition is that each nucleotide is not independent such as codon rules in coding region and regulatory motifs in non-coding region. Recent phrase-level or entity-level masking strategies used by the NLP community proved to be better. BERT or ALBERT generally allows maximum sequence length of 512 tokens, thus k-mer setting can dramatically reduce the number of tokens required to incorporate a 1000-bp context. Token vocabulary size equals to 5^k when using k-mer strategy. 7-mer or longer settings result in unbearable computation burden and memory overflow due to explosive vocabulary size. To balance the computation consumption and representation capacity, we evaluate four types of *k*-mer ($k = 3$, 4, 5, 6) to tokenize the genome. For any given sequence, different sets of k-mer representations can be generated when choosing different split positions. To avoid biased segmentation and further augment training data, we slide 1-bp (*k*-mer-1-stride) for each 1000-bp sequence to generate multiple sets ($n = k$) of *k*-mer tokens as input training instances.

Before being fed into the Transformer network, each token representation is created by summing its corresponding token and position embeddings. Token embeddings are learned through projecting the k-mer vectors into a distributional embedding space. To utilize the order of each sequence, we inject absolute position information of each input sequence and make the model learn position embeddings of the same size as token embeddings. During pre-training stage, we only adopt 'masked language model'(MLM) task to train a bidirectional representation of the human genome. 15% of tokens are randomly masked in each input sequence and the pre-training objective is to predict the masked token by a Softmax layer over the vocabulary. We choose 15% masking ratio according to the empirical choice as in the original BERT model. However, we apply masks at random positions across the genome instead of at fixed positions, which is expected to inject sampling diversities into the model. We denote the *k*-mer tokens embedding size as E, the number of encoder layers as L, and the hidden layer embedding size as H, the number of self-attention heads as A. Visualization of model architecture can be seen in Figure 1B. Hyperparameters are set as follows, $E = 128$; $L = 2$; $H = 256$ and $A = 8$. We pre-train LOGO with *k*-mer tokenization ($k = 3$, 4, 5, 6) on hg19 for a maximum of 50 epochs on four Nvidia Tesla V100 32G GPU. Model hyperparameters are determined by choosing a model size as minimal as possible without compromising the performance. Detailed hyperparameters and pre-training configuration can be seen in Supplementary Table S1 and corresponding ablation studies on how to choose these hyperparameters can be found in Supplementary Tables S2–S4.

Bigger k leads to larger vocabulary size, therefore requiring increased model parameters, more memory usage and longer convergence time. We assess the pre-training performance based on the accuracy (ACC) of masked tokens prediction. 3-mer tokenization achieves the highest MLM accuracy with a minimum training time spent per epoch (Figure 2A, B). For all k-mer settings, LOGO can achieve inflection point of pre-training accuracy after five epochs, though already surpasses 0.875 when training less than one epoch, revealing recurring sequence patterns of human genome is effectively learned. One epoch training time for 3-mer tokenization setting is around 11.4 h, and we reach accuracy plateau (ACC = 0.893) after 15 epochs. One epoch training time for 6-mer tokenization setting is around 70.8 h, and we reach accuracy plateau (ACC = 0.887) after 25 epochs. However, accuracy at the pre-training stage is not directly correlated with utility of specific fine-tuning tasks. We assess all four *k*-mer settings for different downstream tasks and only report the best one. Other details of the pre-training assessment can be found in Supplementary Table S5.

We first evaluate the utility of pre-trained LOGO on human promoter prediction tasks via fine-tuning. Data processing details can be found in Supplementary Text S2. Computational identification of promoters is analogous to sequence labeling or sentence classification task in NLP. Umarov *et al.* have developed CNN-based deep learning models, DeeReCT-PromID (31), to predict human RNA pol II promoters, outperforming other previous prediction tools. For benchmark purposes, we generate datasets in the same way with DeeReCT-PromID and define a positive promoter region from −200 bp to +400 bp window around all potential Transcription Starting Site (TSS) from EPDnew Database (32). Promoters with TATA-box (TATA+) and without TATA-box (TATA-) are assessed separately and afterwards jointly (Both), which leads to 2067, 14 388 and 16 455 positive sequences, respectively. Negative ones are constructed by randomly sampling outside the promoter region without containing a known TSS. Leveraging previously pre-trained LOGO model weights as initialization, we simply plug in these 600-bp sequence inputs, and to-

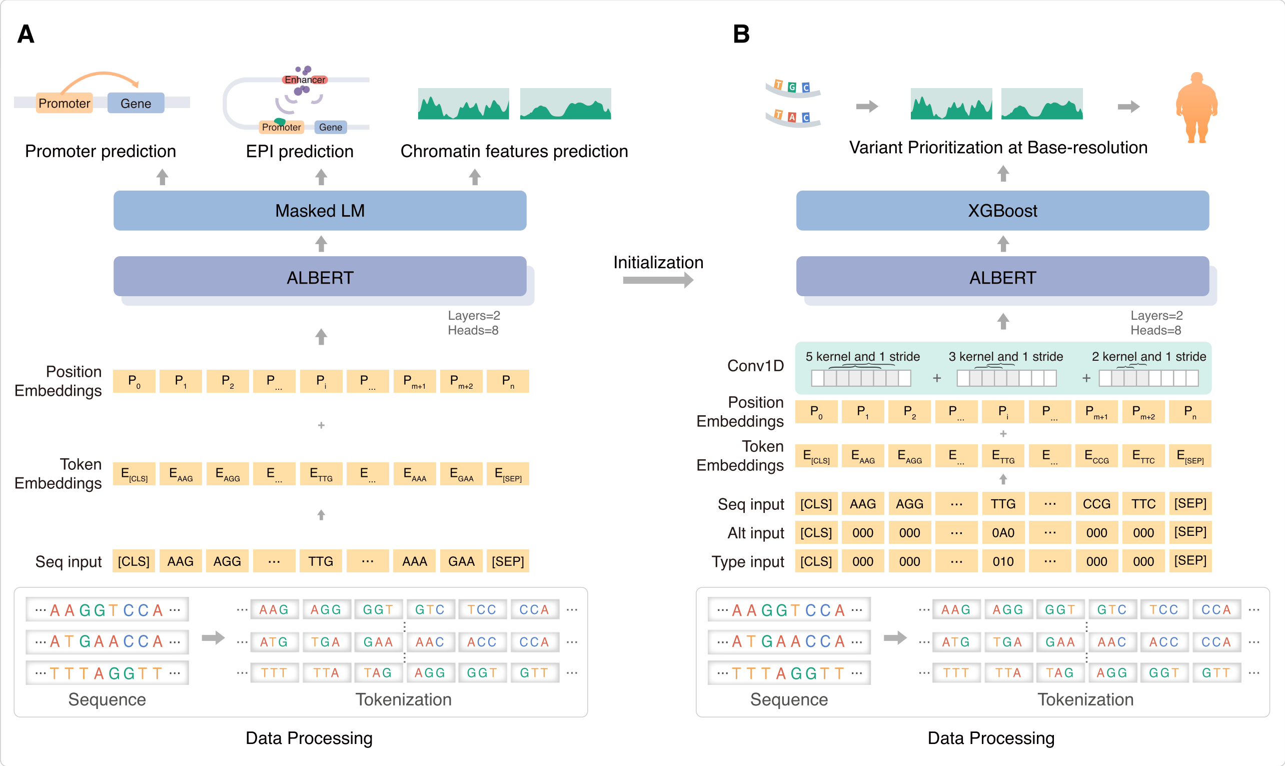
A
Promoter
Gene
Enhancer
Promoter
Gene
Promoter prediction
EPI prediction
Chromatin features prediction
Masked LM
ALBERT
Layers=2
Heads=8
Position Embeddings
P_0 P_1 P_2 $P_{...}$ P_i $P_{...}$ P_{m+1} P_{m+2} P_n
+
Token Embeddings
$E_{[CLS]}$ E_{AAG} E_{AGG} $E_{...}$ E_{TTG} $E_{...}$ E_{AAA} E_{GAA} $E_{[SEP]}$
Seq input
[CLS] AAG AGG ··· TTG ··· AAA GAA [SEP]
···AAGGTCCA···
···ATGAACCA···
···TTTAGGTT···
Sequence
··· AAG AGG GGT GTC TCC CCA ···
··· ATG TGA GAA AAC ACC CCA ···
··· TTT TTA TAG AGG GGT GTT ···
Tokenization
Data Processing
Initialization
B
T G C
T A C
Variant Prioritization at Base-resolution
XGBoost
ALBERT
Layers=2
Heads=8
Conv1D
5 kernel and 1 stride
3 kernel and 1 stride
2 kernel and 1 stride
+
+
Position Embeddings
P_0 P_1 P_2 $P_{...}$ P_i $P_{...}$ P_{m+1} P_{m+2} P_n
+
Token Embeddings
$E_{[CLS]}$ E_{AAG} E_{AGG} $E_{...}$ E_{TTG} $E_{...}$ E_{CCG} E_{TTC} $E_{[SEP]}$
Seq input
[CLS] AAG AGG ··· TTG ··· CCG TTC [SEP]
Alt input
[CLS] 000 000 ··· 0A0 ··· 000 000 [SEP]
Type input
[CLS] 000 000 ··· 010 ··· 000 000 [SEP]
···AAGGTCCA···
···ATGAACCA···
···TTTAGGTT···
Sequence
··· AAG AGG GGT GTC TCC CCA ···
··· ATG TGA GAA AAC ACC CCA ···
··· TTT TTA TAG AGG GGT GTT ···
Tokenization
Data Processing

Figure 1. Overview of LOGO. LOGO is firstly pre-trained on human reference genome hg19 and then fine-tuned on several downstream tasks. (**A**) LOGO uses self-attention based Transformer architecture, a light version language model (ALBERT) with only 2 self-attention layers with 256 hidden unites and 8 heads. The input genome sequence is segmented by sequential k-mer tokens (k = 3, 4, 5, 6). Token embeddings under 1000-bp or 2000-bp context are learned via masked language model (LM) task. The pre-training objective is to predict the randomly masked tokens by a Softmax layer over the vocabulary. Token embeddings and position embeddings are summed up and fed into ALBERT network. For sequence labelling tasks at global sequence level, including promoter identification, enhancer-promoter interaction (EPI) and chromatin features prediction, [CLS] token is used as global features extracted by LOGO, standing for aggregated representations of each input sequence for sequence classification task. [SEP] token stands for the end of each input sequence (Methods). (**B**) For variant prioritization task, LOGO utilizes a multi-stream scheme to encode reference allele, alternative allele, and corresponding altered position as input. A convolutional layer is added to introduce locality to capture allelic-effect at base-resolution. Pre-trained LOGO weights are used as model initializations for variant prioritization task (Materials and Methods).

kenize them via different k-mer-1-stride settings (k = 3, 4, 5, 6) and feed them into LOGO. When conducting sequence classification tasks, model input starts with a token [CLS] as in BERT and ALBERT. We use the final hidden vector of the [CLS] token as the aggregated representation for classification tasks and fine-tune model parameters in an end-to-end manner, which only introduces a few extra parameters in the final classification layer with sigmoid outputs. We use batch size of 128, set early-stop rules and fine-tune the model at most 20 epochs. The average training time per epoch is only around 45 s, which demonstrates excellent efficiency of 'pre-training and fine-tuning' paradigm. The best hyper-parameters are chosen based on the validation sets (Materials and Methods). We evaluate different k-mer settings of LOGO on promoter prediction tasks. LOGO significantly outperforms DeeReCT-PromID in all-settings as evaluated by Precision, Recall and F1-score metrics, as shown in Figure 2D and Supplementary Table S6. LOGO with 5-mer setting (LOGO-5-mer) achieves 15.0% point absolute improvement of mean F1-score than CNN-based DeeReCT-PromID (10-fold cross-validation) in case 'Both'. LOGO has demonstrated its powerful representation utility, which suggests bidirectional attention-based architecture confers an advantage to capture complex semantics of promoter structure over CNN-based model. We also show that pre-training generally benefits downstream promoter identification and chromatin feature identification compared with end-to-end supervised learning with random initializations (Supplementary Tables S7, S8).

We further explore a framework to integrate prior knowledge into LOGO on promoter prediction task. GenBank (33) contains rich functional annotations of the human genome sequence, including CDS, exon, gene, promoter, enhancer, silencer, pseudogene, insulator, conserved region, protein binding site, DNAse I hypersensitive site, nucleotide cleavage site and so on. These annotations can be regarded as prior knowledge of sequence inputs. We download 11 annotations terms from GenBank, i.e. 'CDS', 'exon', 'enhancer', 'insulator', 'conserved_region', 'protein_binding_site', 'pseudogene', 'DNAseI_hypersensitive_site', 'nucleotide_cleavage_site', 'silencer' and 'gene'. Annotations of 'promoter' are abandoned to avoid direct label leakage. We generate annotation labels for

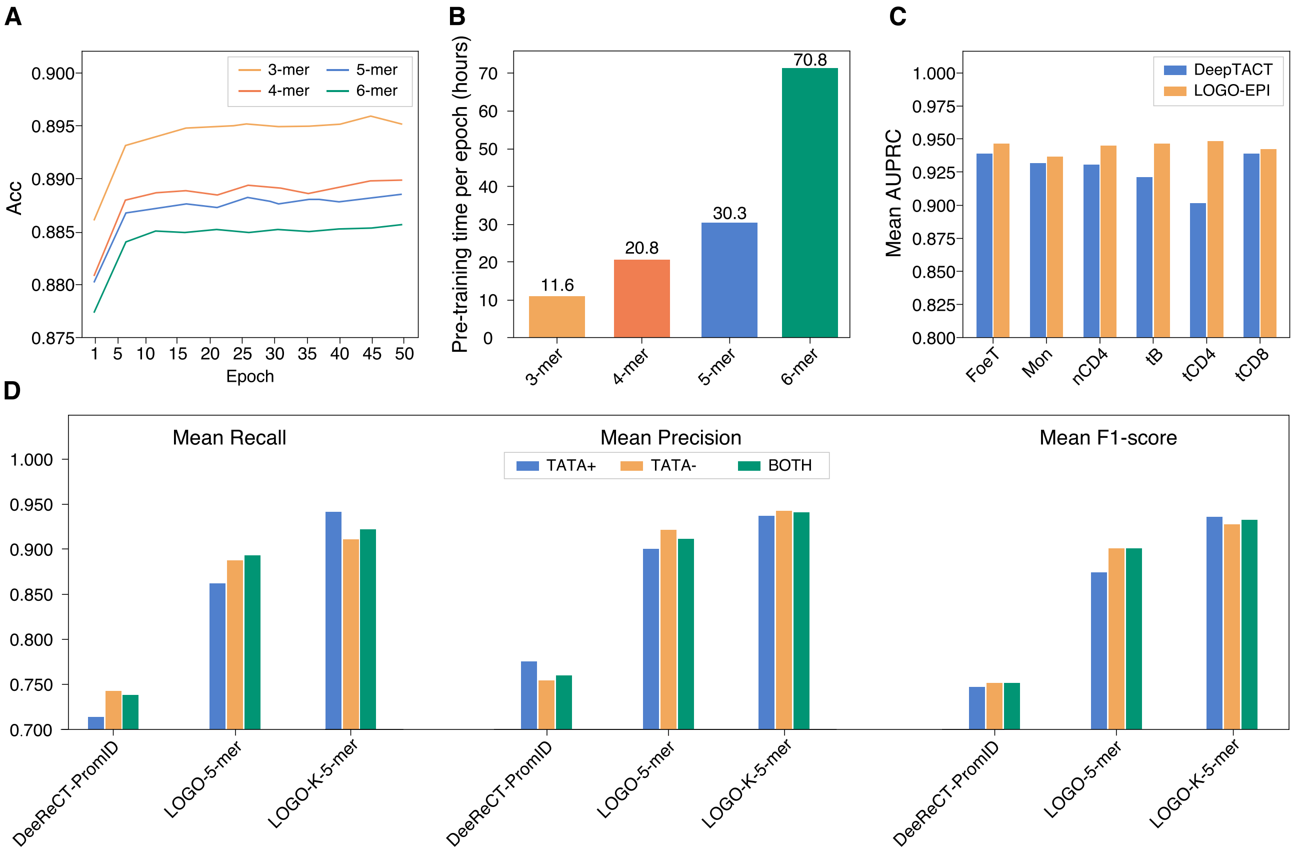

Figure 2. LOGO learns contextualized representation of k-mers of the human reference genome and achieves state-of-the-art performance for promoter prediction and enhancer-promoter interaction prediction. (**A**) Pre-training accuracy (ACC) plateaus after five epochs for all k-mer settings. 3-mer tokenization achieves the highest ACC. (**B**) LOGO pre-training time of one epoch for all *k*-mer settings is plotted. (All settings are trained on four Nvidia Tesla V100 32G GPU). Larger *k* leads to longer training time due to larger vocabulary size. (**C**) Pre-trained LOGO is fine-tuned on enhancer-promoter interaction prediction task (LOGO-EPI) and evaluated against DeepTACT on promoter capture Hi-C (PCHi-C) datasets in six different cell types, including fetal thymus (FoeT, $n = 6676$), monocytes (Mon, $n = 8062$), naïve CD4+ T cell (nCD4, $n = 8712$), total B cells (tB, $n = 9036$), total CD4+ T cell (tCD4, $n = 8282$) and total CD8+ T cell (tCD8, $n = 8140$). Mean area under precision-recall curve (AUPRC) are evaluated using 10-fold cross-validation. (**D**) Pre-trained LOGO using 5-mer tokenization (LOGO-5-mer) is fine-tuned on promoter prediction task and evaluated against DeeReCT-PromID on promoter sequences from EPDnew Database, including ones with TATA-box (TATA+, $n = 2067$), without TATA-box (TATA–, $n = 14\,388$) and jointly (both, $n = 16\,455$). Knowledge embedded LOGO (LOGO-K-5-mer) further boost performance. 11 annotations terms from GenBank, i.e. 'CDS', 'exon', 'enhancer', 'insulator', 'conserved_region', 'protein_binding_site', 'pseudogene', 'DNAseI_hypersensitive_site', 'nucleotide_cleavage_site', 'silencer' and 'gene' are introduced in one-hot encoded format as knowledge input. Metrics of mean Recall, mean Precision and mean F1-score are evaluated using 10-fold cross-validation.

each input sequence in a start-to-end spanning mode based on the hg19 coordinate. We propose a knowledge-enabled LOGO by adding input layers of one-hot encoded annotations and concatenating them with *k*-mer inputs (Supplementary Figure S1). Knowledge embeddings, position embeddings and token embeddings are summed up and then fed into Transformer network for fine-tuning tasks in an end-to-end manner (Figure 1B). Knowledge embedded LOGO with 5-mer setting (LOGO-K-5-mer) achieves F1-score of 0.933, yielding extra 3.2% absolute improvement than LOGO-5-mer (Figure 2D). We demonstrate the configurability and utility of knowledge-embedded framework for genome sequence labelling. We caution that this attempt is preliminary and might introduce position bias or indirect label leakage into the model. We envision that rationally incorporating experimentally validated human knowledge can assist in developing better sequence representation model for scientific discovery.

LOGO can be used to predict regulatory interactions between enhancer-promoter sequence pairs

Predicting 3D chromatin contacts between promoters and enhancers is critical to understand transcriptional regulation in specific cell-lines or tissues. Computational approach is needed to improve the resolution of Hi-C data and detect genome-wide physical interactions at corresponding regulatory elements. This task is analogous to general inter-sentence modelling in NLP, such as sentence pairs in paraphrasing, hypothesis-premise pairs in entailment, and question-passage pairs in the question-answering task. We draw lessons from the NLP field and consider 3D chromatin contacts prediction as a sequence pairing problem.

Li *et al*. proposed DeepTACT (34), a CNN and RNN mixed deep learning model with one attention layer to predict enhancer-promoter interactions (EPI). DeepTACT leverage both raw sequence and chromatin accessibility information, but we only benchmark LOGO against Deep-

TACT version without accessibility information input due to unavailability of processed chromatin features. Data for promoter-enhancer interaction and fine-tuning can be found in Supplementary Text S3. We retrain DeepTACT and fine-tune LOGO (LOGO-EPI) on the same bootstrapped dataset provided by the author of DeepTACT, which contains three parts: 2000-bp window enhancer sequences (35), 1000-bp window promoter sequences (36) and paired enhancer–promoter interaction (EPI) labels from promoter capture Hi-C (PCHi-C) experiments in six different cell types (37), i.e. fetal thymus (FoeT), monocytes (Mon), naïve CD4+ T cell (nCD4), total B cells (tB), total CD4+ T cell (tCD4) and total CD8+ T cell (tCD8). Details can be found in Supplementary Table S9. Similar data augmentation technique is applied to generate larger positive training examples. The performance of each model is evaluated by 10-fold cross-validation. LOGO-EPI uses 6-mer setting to tokenize input sequences. We add one 1D convolution operation for each input promoter or enhancer sequence before being fed into the Transformer network. The underlying intuition is to avoid large fluctuations of token embeddings during the fine-tuning stage and ensure certain disparities among tokens (Supplementary Figure S2). The learned representations of paired promoter and enhancer sequences are concatenated and fed into the final binary classification layer (model architecture seen in Figure S1). LOGO-EPI achieves 0.23–4.47% absolute improvement than DeepTACT on AUPRC for six cell lines (Figure 2C). LOGO-EPI outperforms DeepTACT most significantly for tCD4 and LOGO-EPI yields more consistent performance while DeepTACT fluctuates across different cell lines (AUPRC details can be found in Supplementary Table S10).

LOGO achieves superior performance on chromatin features prediction with significantly reduced model size and much less training time than previous models

Next, we move on to compare LOGO against CNN-based DeepSEA (6) to predict chromatin features from DNA sequences. Unlike fully supervised training manner as DeepSEA, we fine-tune the pre-trained LOGO with chromatin features prediction task and demonstrate higher accuracy with significantly improved scalability. To make a proper comparison as well as demonstrate model scalability, we use three sets of chromatin profiles with some overlaps; the first one is the same as the original DeepSEA paper with 919 chromatin features, the second one is 2002 chromatin features expanded by DeepSEA developer group reported in ExPecto (38), and we construct the third one of 3357 chromatin features by integrating ExPecto's 690 transcriptional factors (TF) binding features with recently released 2850 EpiMap (39) (for epigenome integration across multiple annotation projects) features after deduplication. Data details can be found in Supplementary Text S4.

In the first task, we use the same training, validation, and test sets as in DeepSEA. LOGO-919 obtains 0.70%, 0.70% and 0.80% absolute improvement of median AUROC than downloaded DeepSEA for predicting 690 TF binding, 125 DNase hypersensitive sites (DHSs) and 104 histone modification marks (HM), respectively (Figure 3A). The maximum increase is for transcription repressor ZNF274 binding in HepG2 cell line (AUROC = 0.703 by LOGO-919 versus AUROC = 0.582 by DeepSEA). LOGO's model architecture and training strategy confer huge advantage on computation efficiency and memory consumption over traditional deep CNN-based architecture completely trained in a multitask supervised manner. LOGO has a much smaller parameter size compared to DeepSEA. LOGO-919 contains around 1.52 million parameters, which is 34x fewer than DeepSEA's 52.8 million parameters (Figure 3C). LOGO-919 obtains better performance than downloaded DeepSEA after 33 hours of pre-training and fine-tuning on 4 Nvidia Tesla V100 GPU (around 110 h on 1 Nvidia TITAN Xp Pascal GPU). We also retrain DeepSEA from scratch on 1 Nvidia Titan Pascal GPU and stop after 1 month (720 hours) and reproduce slightly poorer performance than the downloaded version, which indicates LOGO-919 takes at least 6× shorter training time than DeepSEA. (Figure 3D). The improvement in parameter efficiency is the most critical advantage of LOGO framework, which gives LOGO superior advantage to extend to ever-growing chromatin maps. Interestingly, more complex models sometimes lead to inferior performance for chromatin features prediction. We train another 8-layer LOGO model with around 20M parameters, the same order of magnitude with DeepSEA. However, the results are even worse, which again justifies our choice of model hyperparameters (details can be seen in Supplementary Table S2, Table S3, Table S4, Table S15). The learned semantic-rich representation for k-mer tokens in a self-supervised manner alleviates the excessive needs of cumbersome model fitting for different supervised tasks from scratch. To demonstrate this concept, we conduct the second experiment using 2002 chromatin features, including 690 TF binding, 334 DHSs and 978 HM features as reported in ExPecto model (Dataset details of the number of chromatin features can be found in Supplementary Table S16). ExPecto also used CNN-based architecture and extend DeepSEA by doubling the number of convolution layers to increase model depth to satisfy doubled learning objectives, ending up with around 150 million parameters, nearly 3-fold of DeepSEA. We retrain the chromatin marks prediction part of ExPecto and stop after 1000 h. Compared with DeepSEA, the number of learning tasks for Expecto is doubled, while model parameters tripled, which shows severe lack of scalability. For fair comparison against ExPecto, we incorporate a 2000-bp context window while remaining other settings unchanged and fine-tune LOGO-2002 within 66 hours. We choose the time upper bound according to the intuition of doubled training time (66 h versus 33 h) for doubled tasks (2002 features versus 919 features) (Figure 3D). The model size only marginally increases (1.87 million parameters) due to the longer input context and additional parameters of the final classification layer (Figure 3C). The model backbone remains unchanged, and the results show LOGO-2002 can achieve comparable median AUROC with retrained ExPecto on held-out chromosome within 66 hours. The median AUROC for TF, DHSs and HM is 0.954, 0.913, 0.883 respectively (Figure 3B). We demonstrate that LOGO can scale easily via pre-training and fine-tuning paradigm with benefits of computational speed and reduced parameteri-

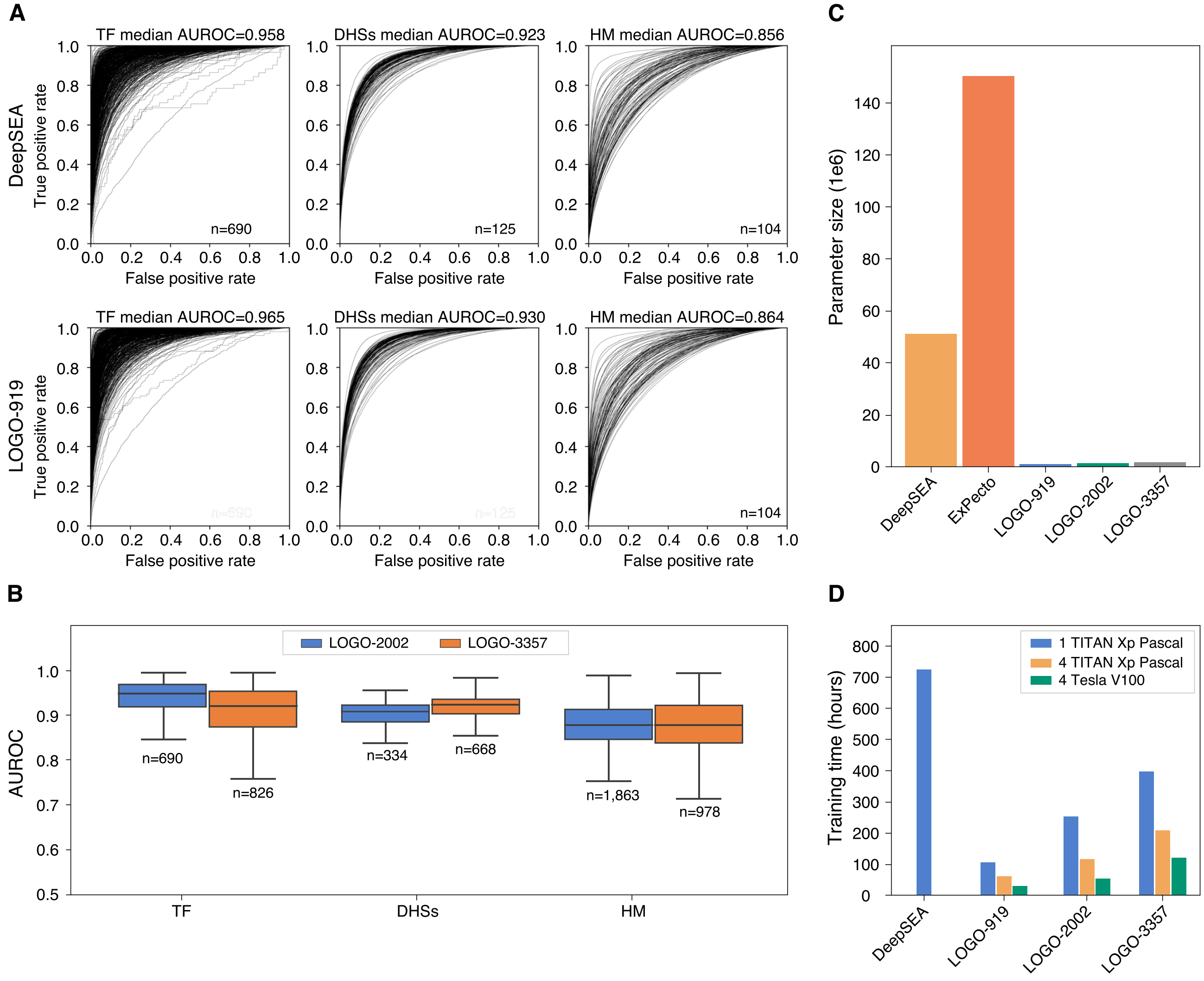

Figure 3. LOGO fine-tuned on chromatin profiles outperforms DeepSEA with significantly reduced parameter size and consumes much less training time. (**A**) Receiver operating characteristic (ROC) curve is plotted to compare predictive power between DeepSEA (top) and LOGO-919 (down) for 690 Transcription factor binding (TF), 125 DNase hypersensitive sites (DHSs) and 104 histone modification marks (HM) on held-out chromosome using 1,000-bp context window. Metrics of median area under receiver operating curve (AUROC) for all TF, DHSs and HM are displayed above each curve. (**B**) Boxplot shows AUROC for three types of features predicted by LOGO-2002 and LOGO-3357 (Methods). Box plots show median, upper, and lower quartiles, and highest and lowest values excluding outliers. (**C**) Plot shows parameter size of DeepSEA, ExPecto, LOGO-919, LOGO-2002 and LOGO-3357. (**D**) Plot shows comparison of training time among DeepSEA, LOGO-919, LOGO-2002 and LOGO-3357. Training time for DeepSEA is recorded as duration of reproducing DeepSEA from scratch using 1 TITAN Xp Pascal GPU, with slightly lower model performance than downloaded version. Training time of LOGO-919, LOGO-2002 and LOGO-3357 include both pre-training (about 2 epochs) and fine-tuning. Three sets of GPU configurations used to fine-tune LOGO are indicated by different colors.

zation. It is noted that DNABERT contains >100 million parameters while LOGO only contains 1 million parameter, which domonstrates LOGO's superior efficiency of parameter sharing among attention layers. In addition, LOGO predicts chromatin features in a jointly multi-task manner while DNABERT only supports TF-binding site prediction one TF by one TF, which is considered as a much simpler task and cannot effectively transfer the knowledge among different chromatin annotations. In light of LOGO's superior performance for sequence feature identification over DeepSEA, we further check whether LOGO can recapitulate those 4 representative variants reported by the original DeepSEA paper, i.e. chr1:109817590 G > T; chr16:209709 T > C; chr10. 23508363 A > G and chr16:52599188 C > T. We collect 1000 bp of DNA sequences centered around each variant and implement 'in silico' saturated mutagenesis by LOGO to scan all potential single-nucleotide substitutions and evaluate these mutation effects for binding events. LOGO is able to identify canonical motifs such as TTGCTCAA for CEBPB (HepG2), TGATAA for GATA1 (K562), GTAAATA for FOXA1 (HepG2) and GTACATA for FOXA2 (HepG2). Detailed results can be found in Supplementary Figure S10.

Incorporating more comprehensive chromatin profiles and using task-specific features have both been reported useful for functional analysis of noncoding variants (40,41). Abundant experimental mappings of human epigenomes are continuously accumulating chromatin profiles for more cell types and tissues. Further expanded chromatin features require larger CNN-based models with explosive parameters, while LOGO can be easily extended to more chromatin features with marginally increased parameters. LOGO demonstrates its powerful scalability and easy deployment, which is of critical importance to tackle even larger-scale functional maps. To further prove this concept, in the third experiment, we utilize the most comprehensive chromatin profiles from EpiMap and integrate them with all TF features from ExPecto. We construct datasets of a 2000-bp context window paired with a label vector for 3357 chromatin features using Selene (42). We fine-tune LOGO-3357 using the same model architecture as LOGO-2002, achieving median AUROC of 0.926, 0.928, 0.883 for 826 TF, 668 DHSs and 1863 HM features respectively (Figure 3B). The slightly lower performance than LOGO-919 and LOGO-2002 is mainly due to less stringent dataset construction and less precise position calibration for EpiMap related features. LOGO-3357 has nearly 2.22 million parameters, which is still significantly fewer than DeepSEA and ExPecto (Figure 3C), again demonstrating its scalability to triple chromatin features without the need of increasing model size substantially or extending disproportionate training time. All training details can be found in Supplementary Table S11.

LOGO can be used to predict functional effects of noncoding variants at base-resolution and provides mechanistic insights for investigating complex diseases

The associated loci identified by GWAS provide abundant information regarding the genetic basis of human complex diseases and traits. Nevertheless, owing to Linkage Disequilibrium (LD), it remains challenging to identify high-resolution causal variants in an interpretable manner (43). Variants from GWAS catalog predominantly consist of marginally associated variants that have not been fine-mapped. We attempt to extend LOGO to prioritize noncoding functional variants for complex diseases based on the predicted signals of the above three sets of chromatin features. We anticipate that, if a complex disease-related variant exerts its effect via disruption of TF binding motif or via alteration of DNA accessibility or histone modification, this SNP can be identified *de novo* from sequence by LOGO. We choose type-2 diabetes (T2D) as an example to test this hypothesis and construct evaluation datasets from the latest published literatures and GWAS resources. (Data Details see in Methods). We employ the same DeepSEA E-value metric to estimate the regulatory potential of a SNP by comparing the allele-specific probabilities per SNP to one million random SNPs from the 1000 Genome Project (Phase 3). In order to provide the community with a comprehensive catalog of LOGO annotated regulatory genome, we have implemented LOGO for all dbSNP reported noncoding SNPs and derive functional scores for each variant (Details can be found in Supplementary Text S8, Supplementary Table S17, Supplementary Figure S13), which can be accessed at https://github.com/melobio/LOGO.

First, we demonstrate LOGO can be used to prioritize putative causal regulatory variants from GWAS reported T2D-associated loci. We hypothesize that if LOGO is fine-tuned on more comprehensive chromatin profiles, it can identify more functional variants within LD blocks. We download all T2D-associated SNPs from GWAS Catalog (44) (2020-05-14 version, *P*-value ranging from 9×10^{-6} to 6×10^{-447}) and corresponding LD SNPs ($r^2 > 0.2$) from LDlink (45), resulting in 156,175 SNPs after deduplication. A variant is considered as functional significant if at least one chromatin feature's *E*-value is equal or less than 1×10^{-5} (6,46). LOGO-3357 can identify more functional SNPs ($n = 14\,764$) than LOGO-2002 ($n = 729$) and LOGO-919 ($n = 374$), details can be found in Supplementary Table S12. Within the 71 GWAS Catalog lead SNPs identified by LOGO-3357, 30 of them reach genome-wide significance (*P*-value $< 5\times10^{-8}$) in at least one GWAS. We divide all functional significant SNPs identified by LOGO-2002 and LOGO-3357 into two groups ($r^2 \geq 0.5$ and $r^2 < 0.5$), we compare the mean activated chromatin features (*E*-value $< 1\times10^{-5}$) of each group and discover that SNPs with higher LD activate more chromatin marks (*P*-value = 0.00093 for LOGO 3357, *P*-value = 0.02 for LOGO 2002, by Mann–Whitney *U*-test, details can be found in Supplementary Figure S3). Inspired by Basenji (47) and Enformer (48), we also implement saturation mutagenesis to interpret several T2D-related SNPs in a visually interpretable manner. We refer to a European T2D fine-mapping study conducted by Mahajan (49), who characterized 51 variants with posterior probability of association (PPA) >80% by incorporating islet-specific epigenome information. For example, PPA of rs963740 boosts from 50.3% to 87.9% by fGWAS (50) (Supplementary Figure S9), a statistical package for integrating regulatory annotations into GWAS. LOGO correctly predicted strong alteration of related chromatin feature caused by minor allele relative to the major allele, the most activated feature also indicates rs11257655 position overlapping with strong islets enhancer region and modulating the known motif of the transcription factor FoxA2. LOGO suggests that perturbed TF binding within islets as a potential etiological mechanism for T2D (Figure 4E). Another example variant consistent with Mahajan's finding (49) is SNP rs963740 (located in the DLEU1 locus), which can be found in Supplementary Figure S9.

We demonstrate that LOGO has the potential of fine-mapping causal variants within LD block in an explainable manner and the model fine-tuned on more chromatin features provides more functional attributions. We further conduct tissue-enrichment analysis for all putative functional variants identified by LOGO-2002. Hypergeometric test is used to evaluate whether activated chromatin features are enriched in certain categories. We find that these SNPs are functionally enriched in 18 categories out of total 27 with activation signals, including smooth muscle ($n = 51$), lymphoblastoid ($n = 45$), adipose ($n = 20$), muscle ($n = 63$), spleen ($n = 10$), and liver ($n = 48$), (–log(*P*-value) = 11.1, 5.3, 5.0, 3.2, 3.1 and 1.3 respectively), which is consistent with years of pathogenesis research that insulin mainly acts on liver, muscle and adipose as T2D-relevant tissues (Figure 4A, Supplementary Figure S8). Figure 4A has some abbreviations as follows: T2D, type 2 diabetes; EEM, extra-embryonic membranes; ESC, embryonic stem cell; ES-deriv, embryonic stem-derived; iPSC, induced pluripotent stem cell; HSC, hematopoietic stem cell. Recent integrative epigenomics study (39) (EpiMap) lever-

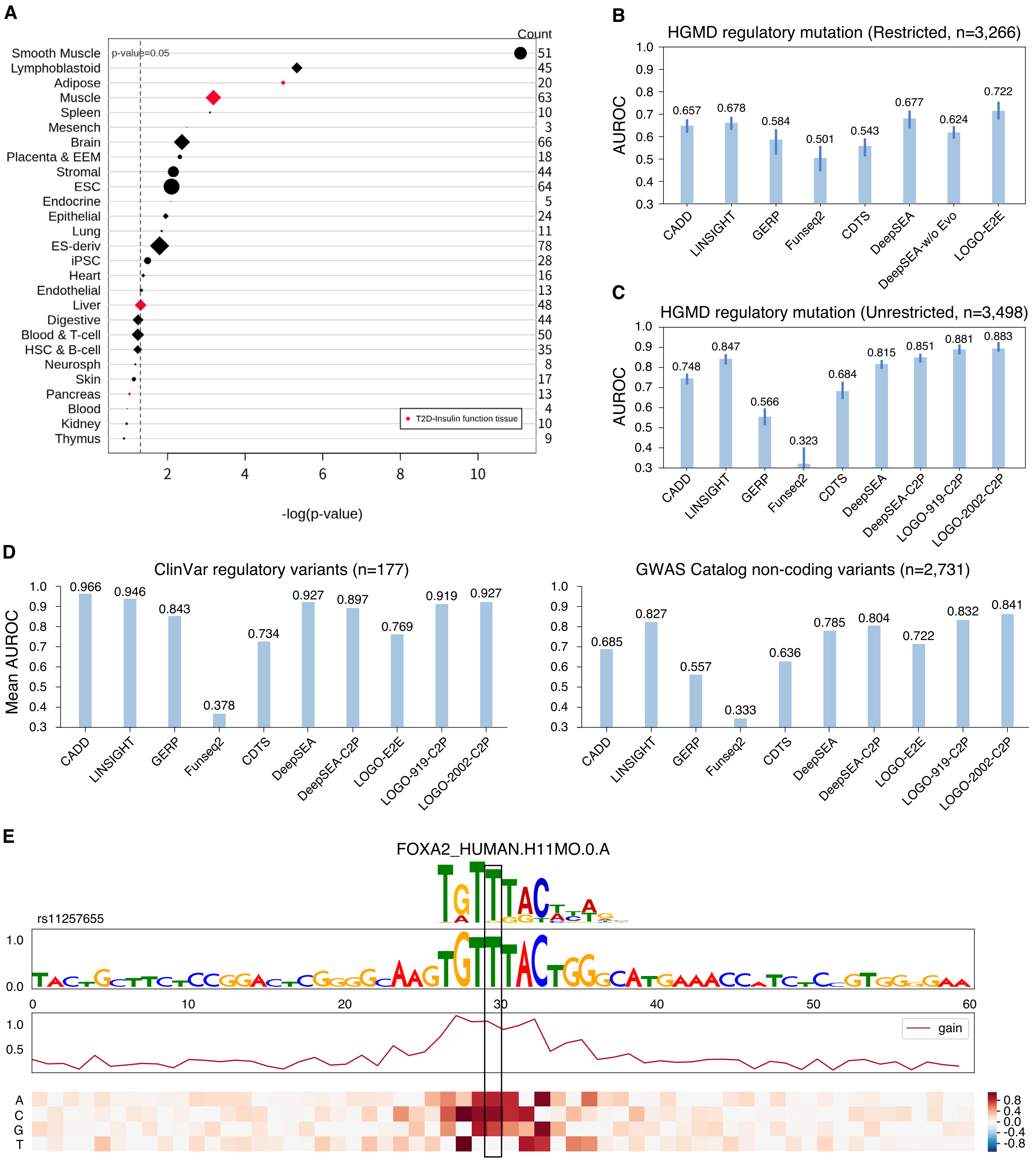

Figure 4. LOGO can be used to infer underlying regulatory mechanisms of T2D GWAS signals and prioritize functional variants both inherited diseases and complex traits or diseases. (**A**) Tissue enrichment results for significant T2D-related variants identified by LOGO-2002 located in promoter/enhancer state regions, sorted by -log(P-value) (Hypergeometric test). The size of circle/diamond represents the number of activated chromatin marks in corresponding tissue or cell types, also displayed on the right side of the plot. Red symbols indicate four well-known tissue types related to T2D. Tissue or cell types enriched by EpiMap for T2D GWAS signals are represented by diamond shape. (**B**) Prediction power of various models for prioritizing HGMD regulatory mutations ($n = 3266$) against negative controls ($n = 3266$) in restricted scenario that negative samples matched to positive ones within 1 kb. (**C**) Prediction power of various models for prioritizing HGMD regulatory variants ($n = 3498$) against negative controls ($n = 3690$) in unrestricted scenario of random sampling. Mean AUROC of 10-fold cross-validation for B and C are reported, Error bars represent standard deviations. (**D**) Comparison of model performance by metric of mean AUROC on the held-out test set from inherited disease domain (top: $n = 177$ stringent ClinVar regulatory variants, negative controls are bootstrapped 10 times) and complex trait or disease domain (bottom: $n = 2731$ genome-wide significant non-coding variants from GWAS Catalog, positive variants are bootstrapped 10 times). (**E**) LOGO prediction for rs11257655 captures its influence on CDC123-CAMK1D locus. rs11257655 is associated with T2D disease and overlapped with islets epigenome map. In silico mutagenesis of the region surrounding rs11257655 reveals an affected transcription factor motif.

ages enhancer sharing tree to investigate tissue enrichment of T2D-related SNPs and indicates that T2D is polyfactorial trait enriched in up to 18 tissue categories out of 33 tested. Sequence-based LOGO-2002 shows similar diversity of enrichment with significant tissue overlaps. Islets only constitute ∼1% of the pancreas and specific annotations of islet epigenome are absent in ENCODE and Roadmap Epigenomes Project (51). Thus, analyzing the pancreas organs alone fails to provide reliable information of islet epigenomes. Thurner (51) and Varshney (52) have specifically annotated promoter/enhancer state of islets, which are regarded as typical T2D relevant cell types. Specifically, 10 functional SNPs identified by LOGO-3357 are overlapping with islets-specific promoter/enhancer state with at least five activated chromatin features (E-value $< 1\times10^{-5}$) (Table 1). The disruption of regulatory function is consistent with a previous report that parts of T2D-related risk variants are considered to act through primary effects on beta-cell function. For instance, T2D-risk allele at rs9693089 (FAM167A locus) locates at the active enhancer state of islet sample identified by Varshney (52) and has been reported to be associated with very low-density lipoprotein (VLDL) synthesis by Kraja (53). The corresponding activated chromatin features include H3K4me3, H3K4me1, H3K27ac. Even though LOGO-3357 is not specifically trained on islet chromatin marks, this experiment demonstrates that deep learning based methods have the potential of providing extra informativeness using sequence alone as input (54).

Second, we demonstrate that LOGO can be a sequence-based tool to help interpret those GWAS signals with possible population bias or sample size limitations. The statistical power of GWAS relies heavily on sample size, allele frequency and effect size of candidate SNPs (55). GWAS with inadequate sample size can result in a multitude of nominally significant loci (P-value < 0.05). This problem is mainly mitigated by expanding the sample size or conducting meta-analysis across cohorts or ethnic groups. Sequence based LOGO model is expected not to be affected by allele frequency or population bias and can evaluate both common and rare variants *ab initio*. We illustrate this potential using the following examples. In study GCST005414 (56), rs340874 (PROX1 locus) reaches nominally significant (P-value $= 1\times10^{-7}$). However, this SNP achieves genome-wide significant in study GCST009379 (49)/GCST006867 (57) (P-value $= 2\times10^{-22}$ and 8×10^{-18}, respectively) with larger sample size. PROX1 has been reported to be associated with after-meal metabolism (58), non-esterified fatty acids, and glucose metabolism (59), which is also validated in both Japanese and Chinese populations (60,61) (MAF = 0.376). LOGO-3357 can directly identify rs340874 as a functional significant SNP (1 activated feature with E-value $\leq 1\times10^{-5}$, transcription factor POLR2A). PROX1 is reported to be a target gene of the POLR2A transcription factor from the ENCODE Transcription Factor Targets dataset. Rs896854 (TP53INP1 locus) is perceived to be associated with lipid levels of the Chinese population with nominal significant signal (P-value $= 2\times10^{-6}$) in the study GCST004894 (62) and genome-wide significant signal (P-value $= 1\times10^{-9}$) in the study GCST000712 (63). Rs516946 (ANK1 locus) is reported in several independent studies to be correlated with decrease of insulin level and dysfunction of pancreatic islet cells at a nearby site (64). Again, LOGO-3357 can identify both rs896854 (1 activated feature with E-value $\leq 1\times10^{-5}$, Blood & T-cell, H3K9me3) and rs516946 (1 activated feature with E-value $\leq 1\times10^{-5}$, Other, DNase-seq) to be functional. Furthermore, we evaluate another 43 regulatory variants with posterior probability of association (PPA) >80% in a recent fine-mapping study (49). 2 SNPs are identified as functional significant by LOGO-3357 (rs340874 at PROX1 locus and rs76549217 at ANKH locus). Another largest-scale T2D meta-analysis study accumulates 1.4 million samples and discovered 318 new loci (65), out of which LOGO-3357 can identify 14 SNPs to be functional. It is worth mentioning that all these 14 SNPs do not reach genome-wide significance in other populations except European ancestry. This result further indicates the unbiased predictive power of LOGO. 16 reported SNPs validated by LOGO-3357 with corresponding activated features are listed in Supplementary Table S14. We demonstrate that LOGO fine-tuned on chromatin features can help interpret GWAS non-coding SNPs and provide hints regarding underlying tissue-specific regulatory mechanism.

Introducing locality-sensitive encoding scheme and convolution facilitates prioritizing functional variants for both inherited diseases and complex traits or diseases

Next, we evaluate whether fine-tuning LOGO can be used to develop functional predictor of pathogenic regulatory single-nucleotide variants (SNV) or common GWAS phenotype-associated SNPs. We define two schemes of fine-tuning: end-to-end training on the binary label of deleteriousness (LOGO-E2E) and two-stage training of chromatin features prediction followed by variant prioritization (LOGO-C2P) (6,40). Perturbation of molecular phenotypes can serve as an indicator of potential deleteriousness inspired by DeepSEA. We compare LOGO against six common predictors, including evolution-based method (GERP) (66), sequence-based predictor based on chromatin effect signals with four evolutionary conservation features (DeepSEA) (6), functional genome-based method (Funseq2) (67), evolutionary method incorporating functional genome features (LINSIGHT) (68), machine learning based classifier (CADD) (69) and genome diversity metric (CDTS) (70). It is noted that DeepSEA and CADD can provide allele-specific evaluations, whereas others assign identical scores to all alternative variants. Our predictors, LOGO-E2E and LOGO-C2P, are designed to capture allelic effect.

For variants associated with inherited human diseases, we extract a dataset from Human Gene Mutation Database (version 2019–03) (71) to define positive examples of stringent regulatory mutations. We construct negative controls from 1000 Genomes Project (72) SNPs by stringent frequency and population control, resulting in 3498 pathogenic regulatory mutations and 3,690 negatives (total 7,188 variants, details can be found in Supplementary Figure S6). We use 10-fold cross-validation to make a robust comparison. For each fold, test variants are ensured to be scorable across methods. To increase stringency, we consider two schemes of negative sets selection: random sampling (unrestricted), and negative samples matched to

Table 1. Significant SNPs identified by LOGO-3357 overlapped with islet promoter/enhancer regions[a]

RS ID	Locus	Significant marks[b]	min E-value[c]	GWAS *P*-value[d]	GWAS odds[d]	1000G AF	Paper (PMID)
rs9693089 chr8:11298385 A-G	*FAM167A*	H3K4me3 H3K4me1 H3K27ac	0.000001	-	-	0.68111	23192668
rs4735337 chr8:95973465 T-C	*TP53INP1*	H3K4me3 H3K4me1 DNase-seq	0.000001	-	-	0.552516	25393876
rs1126899 chr7:130021488 G-C	*CPA1*	H3K4me1	0.000001	-	-	0.582268	-
rs11774700 chr8:118220270 T-C	*LOC105375716*	HNF4G	0.000001	-	-	0.271965	21188353
rs163800 chr20:57578508 T-C	*CTSZ*	H3K27ac	0.000001	-	-	0.000399361	29795304
rs4383259 chr19:53661337 A-G	*ZNF347*	H3K4me1 H3K27ac	0.000001	-	-	0.752196	24306210
rs3176447 chr1:51433687 T-A	*CDKN2C*	DNase-seq	0.000001	-	-	0.0740815	21145615
rs11671664 chr19:46172278 G-A	*GIPR*	H3K27me3	0.000001	3E-12	4.22[2.73–5.71]	0.155152	27480816
rs998451 chr2:135429288 G-A	*TMEM163*	CEBPB	0.000001	-	-	0.10643	24843659
rs1776897 chr6:34195011 G-T	-	EP300	0.000004	-	-	0.776757	27104953

[a] Islet promoter and enhancer regions are annotated by Thurner (51) and Varshney (52).
[b] Significant Marks means all chromatin marks with E-value $< 1 \times 10^{-5}$.
[c] Min E-value means the minimum E-value of corresponding chromatin mark activated by LOGO-3357.
[d] GWAS *P*-value and GWAS odds are only shown for lead SNP reported from GWAS Catalog.

positive ones within 1 kb (restricted, total 6532 variants). Dataset construction details are illustrated in Supplementary Text S6.

For LOGO-E2E, we use three layers to encode variant presence and allelic information at specific position, including the Ref layer, Alt layer and Variant Type layer. Ref layer is used to encode 1,000-bp context with 6-mer-1-stride setting. (Model architecture details can be found in Supplementary Figure S4) Alt layer is used to encode alternative allele at certain position to enforce the model to see directional alteration. Another Variant Type layer is set as default for SNV. By this means, we explicitly encode the alternative allele, ensure the ALT allele is always the effect allele. Each variant with surrounding context of certain length will be encoded as a matrix input containing ‘Ref’, ‘Alt’ and ‘Type’ information. 1-dimension convolutional layer is added before fed token embeddings into LOGO to learn the binary deleteriousness effect of the target variant. Fine-tuning LOGO in this way is expected to learn allelic pathogenicity. For LOGO-C2P, we follow similar pipelines in DeepSEA’s functional SNP prioritization part and firstly use previously trained LOGO-919/LOGO-2002 to generate chromatin effect features for both reference and alternative allele. We then conduct the same absolute difference and relative log fold change transformation as DeepSEA and feed these features into boosted logistic regression model to train the classifier at the second stage. It is worth mentioning that we discard the z-score transformation used in DeepSEA-C2P classifier. We also assess the difference between preserving or removing evolutionary conservation features. The original scores of LINSIGHT, CADD, FunSeq2, GERP, CDTS and DeepSEA functional significant score are used to obtain the binary classification result with full range of thresholds.

In the end-to-end setting, LOGO-E2E outperforms all other methods in restricted negative control scenario (AUROC = 0.722) (Figure 4B) and performs the second in the scenario of unrestricted negative control (AUROC = 0.823). It is consistent with previous finding that restricted scenario poses more difficulties for distinguishing functional sites from surroundings than separating functional regions from genome background. Nonetheless, LOGO-E2E leverages an Alt token layer to enforce the model to explicitly encode allele position and directional mutation event to be distinguished from nearby unchanged context, which equips the model with allelic specificity under 1,000-bp context. For the less challenging unrestricted task, LOGO-E2E performs slightly worse than LINSIGHT(AUROC = 0.847), one possible reason might be that LOGO has not been trained on population genomic data with conservation information to witness enough genome diversity from human and other related outgroup species. To overcome these shortcomings, LOGO-2002-C2P incorporates four evolutionary conservation features as in DeepSEA (PhastCons scores (73), PhyloP scores (74), and GERP++ neutral evolution (75) and rejected substitution scores (66)) and achieves the highest performance (AUROC = 0.883) (Figure 4C) in the scenario of the unrestricted negative control.

It is noted that all compared methods except DeepSEA-C2P are not specifically trained on the HGMD dataset. To avoid potential over-fitting controversy and assess the generalizability of LOGO-C2P, we extract from ClinVar database (76) to define an independent test set with 177 highly confident non-coding pathogenic SNVs (Data details in Methods and Supplementary Figure S5, all splicing variants removed). Analysis with LOGO-E2E, LOGO-919-C2P and LOGO-2002-C2P are compared with that of CADD, LINSIGHT, GERP, Funseq2, CDTS, DeepSEA and DeepSEA-C2P (Figure 4D). DeepSEA and DeepSEA-w/o Evo means DeepSEA derived functional significant score including and excluding four evolutionary features, respectively. DeepSEA-C2P represents a logistic regression model based on 919 chromatin marks and 4 evolutionary features. LOGO-E2E means LOGO fine-tuned on binary label of allelic deleteriousness in an end-to-end manner. LOGO-919-C2P and LOGO-2002-C2P stand for LOGO fine-tuned on prediction of 919/2002 chromatin features with four evolutionary features followed by allele-specific variant prioritization via logistic regression model. LOGO-2002-C2P ranks third (AUROC = 0.927) and significantly outperforms CDTS (AUROC = 0.734) but is slightly worse than LINSIGHT (AUROC = 0.946) and CADD(AUROC = 0.966). CDTS solely relies on 11,257 whole-genome sequences to obtain 7-mer constraint under 550-bp context of human species, whose lack of inter-species conservation leads to poorer performance to evaluate fitness consequence of inherited disease related variants. LOGO-E2E (AUROC = 0.769) is only trained on 3498 HGMD variants yet performs better than genome diversity based CDTS, which again proves end-to-end fine-tuning architecture captures some intrinsic features of non-coding genome by only using a few annotated examples. LOGO-C2P is only trained on HGMD dataset and proved to be well generalizable on the ClinVar dataset. LINSIGHT is trained on human polymorphism data from 54 unrelated individuals and three outgroup species divergence data from aligned primates genomes conditioned on 48 genomic features, revealing the utility of incorporating genome diversity information to interpret non-coding genome. CADD is trained with more than 60 genome annotations on a much larger dataset (n = 30 million) containing fixed or nearly fixed variants in human populations but is absent in human-ape ancestor as proxy-neutral variants and matched proxy-deleterious variants, which is essentially designed for binary classification of fitness consequence. The superior performance of LOGO-C2P, LINSIGHT and CADD shows that evolutionary information is likely to be powerful to identify regulatory pathogenic variants that tend to be under strong purifying selection.

We conduct another benchmark experiment to prioritize complex trait or disease-associated variants. GWAS variants are generally of weaker functional impact than HGMD mutations. We construct a positive test set by extracting all genome-wide significant variants (P-value $< 5\times10^{-8}$) replicated in at least two independent studies from GWAS Catalog followed by retaining SNPs overlapped with ENCODE candidate cis-Regulatory Elements (ccREs) (2) and fixation index (F_{ST}) lower than 0.01 to ensure little genetic differentiation (77,78). We ensure that all test variants have never been used in previous HGMD experiment, resulting in 2,731 positive GWAS SNPs and 704 negative controls. We bootstrap 10 times to obtain balanced held-out test set of 1408 variants (Data details in Supplementary Text S6 and Supplementary Figure S6). All predictors show reduced performance. Compared with more deleterious HGMD mutations under significant purifying selection, common GWAS-associated variants have smaller effect size with lower evolutionary conservation, thereby plausibly more difficult to predict. LOGO-C2P-2002 is the top performer (AUROC = 0.841) (Figure 4D) across all methods We show that LOGO-C2P-2002 has the advantages of considering both chromatin effects and evolutionary constraint at base-resolution. Though LOGO-C2P is solely fine-tuned on HGMD mutations, the result proves its domain transferability from inherited diseases to common phenotypes. The second-best predictor is LOGO-919-C2P (AUROC = 0.832), indicating the benefit of broad collection of chromatin features. LOGO-919-C2P outperforms DeepSEA-C2P (AUROC = 0.804), which again demonstrates the edge of attention-based Transformer over CNN-based architecture. For these two independent evaluations, LOGO-C2P performs relatively better than CADD and LINSIGHT in GWAS domain than ClinVar domain, which suggests that chromatin features are more informative for complex traits while evolutionary information is more important for inherited diseases. This is consistent with the hypothesis that highly deleterious mutations of genetic diseases are subject to stronger selection than complex disease loci (79). Recent EpiMap results also emphasize the central role of dense, rich, high-resolution epigenomic annotations to investigate regulatory circuitry of complex disease. LOGO-C2P exhibits its capability of integrating sequence context, regulatory annotation, and evolutionary constraint either explicitly or implicitly at different levels. It is noted that CDTS, which solely relies on human genetic diversity, shows poorer performance in both rare and common disease scenarios. We argue that the statistical test of 7-mer regional tolerance is not powerful enough to capture complex semantics underlying human genome sequence, even though more than 10,000 human genomes are incorporated (70). For the variant prioritization task, we also benchmark against Basenji and achieve better performance on a small dataset. Details can be found in Supplementary Figure S12.

In order to demonstrate the utility of incorporating 1D convolution, we conduct ablation experiments on GWAS Catalog SNPs benchmark dataset and obtain better result than convolution-excluded version (Details can be found in Supplementary Text S7, Supplementary Table S13 and Supplementary Figure S11). The convolution module adds three kinds of channel information, 2-mer, 3-mer and 5-mer, which can effectively provide more diverse contextual information and help the model clearly distinguish the variant change before and after the mutation. In addition, the convolution operation reduces the weight updating frequency and makes parameter updating more stable during the fine-tuning process, which is well suited for variant effects prediction task at base-resolution.

Furthermore, we explore LOGO performance of prioritizing pathogenicity of small insertion or deletion variants (Indels). We fine-tune LOGO in a similar way with LOGO-E2E using 3-mer tokenization with 1000-bp context (LOGO-E2E-Indel) on a much larger dataset from CADD Developmental release: v1.4 with 3,675,207 indels, including similar number of human-derived variants and simulations (69). We evaluate model performance of LOGO-E2E-Indel against LINSIGHT, CADD and DeepSEA-w/o Evo (excluding evolutionary features) on independent test set, consisting of 5869 non-coding Indels (<48 bp) from ClinVar recent release (clinvar_20201003), including 5556 positive samples (defined as pathogenic and likely pathogenic in ClinVar) and 313 negative samples (defined as benign and likely benign in ClinVar). LOGO-E2E-Indel achieves the best performance (AUROC = 0.743) across all compared methods (Supplementary Figure S7). These results indicate that LOGO-E2E can effectively utilize the learned semantic representations from pre-training and shows stronger generalizability for downstream classification tasks than CADD, which is trained in a fully supervised manner.

DISCUSSION

Genome sequence contains tremendous biological information regarding the species to which it belongs. Even though a multitude of high-throughput biochemical assays have been used to characterize the sequences, the complex nature of genome poses tremendous challenges to well interpret it. It is impractical to exhaustively perform functional annotations at every position in all conditions, and current assay design is believed to only cover the tip of the iceberg due to the limitations of existing hypothesis. A substantial gap remains between the outcomes of these experiments and a comprehensive understanding of the whole genome, especially those regulatory regions. New computational approaches are in pressing need to help interpret the underlying code. Motivated by recent huge progress in the field of NLP and CV, we propose a light language model called LOGO, utilizing ALBERT-version Transformer architecture for sequence labelling, and integrating convolution with a novel input encoding scheme for base-resolution interpretation.

Learning from raw reference genome successfully equips the model with strong adaptability across various downstream tasks by fine-tuning. No explicit annotation label is given during pre-training stage, and we have shown that the intrinsic bidirectional representations learned by the model can easily extend to sequence labelling tasks. In chromatin features prediction task, LOGO achieves higher accuracy than DeepSEA with significantly reduced parameters in much shorter computing time. Facing the needs of continuously growing number of functional annotations, we demonstrate that supervised multitask learning incurs problem of parameter explosion and tedious architecture tuning, while LOGO can efficiently extend to more abundant features with marginally increased parameters and trivial modification. Sequence-based chromatin effects prediction is informative to characterize GWAS SNPs via identifying certain disruption of regulatory function. These results offer a strong justification that developing pre-trained language model can enable accurate, fast, scalable, and robust genome modelling. The community can benefit from simply and economically fine-tuning the pre-trained LOGO for specific chromatin profiles of intertest with trivial effort. By initializing model with pre-trained weights, only one additional output layer needs to be modified instead of extensive architecture tuning. We also show that fine-tuning LOGO with an explicit ref/alt token encoding strategy and convolutional operation proves powerful to prioritize functional non-coding variants associated with human disease at base-resolution.

It is noted that LOGO is only trained on human reference genome hg19. We envision that introducing genome diversity in pre-training stage can further boost representation power. This can be done by feeding LOGO with all currently identified variants across human populations and from other related outgroup species, which is expected to automatically learn evolutionary conservation and context-dependent constraint across the genome. The learned representation will in turn facilitates variant function prediction and evolutionary landscape discovery. We make an analogy between biological sequence and human language that genome possesses diversified combinations of words or phrases without compromising intrinsic grammar constraints. Overall, LOGO offers a versatile strategy to represent both global and local patterns of the human genome and sheds light on unearthing more value of ever-growing WGS data in the boom of national genome project.

We hypothesize that there exist many dimensions not yet captured by LOGO. The intrinsic property of naïve self-attention based Transformer model leads to its incompetency to capture even longer-range context (80), which is essential for modeling distal regulatory dependency across human genome. Recently, Enformer (47) combines dilated convolutions and Transformers to model interactions up to 100 kb away and successfully links remote enhancers to target genes. In the future, we would like to explore whether designing advanced LOGO via incorporating hierarchical interactive mechanism (81) would solve the problem. For example, encoding a 1 Mb DNA sequence with 'sentence Transformer' and then feeding it to 'document Transformer'. LOGO can also be fine-tuned on tissue or cell-type specific expression profiles to investigate variant effects, potentially shedding light on eQTL fine-mapping and cis-regulatory evolution. Furthermore, there could be alternative ways to construct underlying vocabulary and define pre-training objectives with a further optimized masking strategy. We already show that injecting knowledge post hoc into the model can help boost performance. On the other hand, we anticipate that a large amount of existing somewhat noisy knowledgebase can be utilized to further boost the effectiveness of deep learning model. For example, sequence annotation databases (82,83) and biological networks (84) can be introduced systematically and structurally to guide self-supervised representation learning of genome sequence or inspire novel knowledge-guided masking strategy design. This will in turn help construct a better downstream prediction model in a more interpretable manner. In addition, LOGO can be reconfigured into a generative version, potentially be used to improve *in silico* mutagenesis efficiency and assess artificially designed sequences in the field of genome editing and synthetic biology. Integrating adversarial feedback loop of functional constraint into language model can potentially aid perturbation experiment and rational *de novo* design of new regulatory circuit (85,86).

DATA AVAILABILITY

All datasets in this study are derived from published resources and can be generated following protocols described in Methods. Demo data is available on Github at https://github.com/melobio/LOGO. LOGO functional scores for all noncoding SNPs can be found on figshare at https://doi.org/10.6084/m9.figshare.19149827.v2.

CODE AVAILABILITY

LOGO is written in Python. The source code is available on Github at https://github.com/melobio/LOGO.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Zhenzhong Lan from Westlake University for discussions regarding pre-trained language model. We thank Peng Cheng Laboratory, Shenzhen, Guangdong, China for providing GPU resources to train the model.
Author contributions: M.Y. conceived the problem and designed the study. H.M.Y. and F.M. supervised the work. M.Y., L.C.H., H.P.H., and H.T. performed deep learning experiments. L.C.H, H.T. and N.Z. performed bioinformatics analysis. J.H.W. provided detailed guidance on analysis of disease-related regulatory mutations. M.Y. wrote the manuscript, and other authors made modifications.

FUNDING

This research is supported by Guangdong Provincial Academician Workstation of BGI Synthetic Genomics [2017B090904014]; Jihong Wu is supported by Program of Shanghai Academic Research Leader [20XD1401100] and Program for Outstanding Medical Academic Leader [2019LJ01].
Conflict of interest statement. F.M. declares the following competing interests: stock holdings in MGI.

REFERENCES

1. ENCODE Project Consortium. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57.
2. Moore,J.E., Purcaro,M.J., Pratt,H.E., Epstein,C.B., Shoresh,N., Adrian,J., Hardison,R.C., Gingeras,T.R., Stamatoyannopoulos,J.A. and Weng,Z. (2020) Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature*, **583**, 699–710.
3. Kundaje,A., Meuleman,W., Ernst,J., Bilenky,M., Yen,A., Heravi-Moussavi,A., Stamatoyannopoulos,J.A., Wang,T. and Kellis,M. (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.
4. Visscher,P.M., Wray,N.R., Zhang,Q., Sklar,P., McCarthy,M.I., Brown,M.A. and Yang,J. (2017) 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet*, **101**, 5–22.
5. Eraslan,G., Avsec,Ž., Gagneur,J. and Theis,F.J. (2019) Deep learning: new computational modelling techniques for genomics. *Nat. Rev. Genet*, **20**, 389–403.
6. Zhou,J. and Troyanskaya,O.G. (2015) Predicting effects of noncoding variants with deep learning–based sequence model. *Nat. Methods*, **12**, 931–934.
7. Mikolov,T., Chen,K., Corrado,G. and Dean,J. (2013) Efficient estimation of word representations in vector space. arXiv doi: https://arxiv.org/abs/1301.3781, 07 September 2013, preprint: not peer reviewed.
8. Cybenko,G.(1989) Approximation by superpositions of a sigmoidal function. *Math. Control Signals Syst.*, **2**, 303–314.
9. Elman,J.L. (1990) Finding structure in time. *Cogn. Sci.*, **14**, 179–211.
10. Hochreiter,S. and Schmidhuber,J. (1997) Long short-term memory. *Neural Comput.*, **9**, 1735–1780.
11. Vaswani,A., Shazeer,N., Parmar,N., Uszkoreit,J., Jones,L., Gomez,A.N. and Polosukhin,I. (2017) Attention is all you need. arXiv doi: https://arxiv.org/abs/1706.03762, 06 December 2017, preprint: not peer reviewed.
12. Peters,M.E., Neumann,M., Iyyer,M., Gardner,M., Clark,C., Lee,K. and Zettlemoyer,L. (2018) Deep contextualized word representations. arXiv doi: https://arxiv.org/abs/1802.05365, 22 March 2018, preprint: not peer reviewed.
13. Howard,J. and Ruder,S. (2018) Universal language model fine-tuning for text classification. arXiv doi: https://arxiv.org/abs/1801.06146, 23 May 2018, preprint: not peer reviewed.
14. Radford,A., Narasimhan,K., Salimans,T. and Sutskever,I. (2018) Improving language understanding by generative pre-training. https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf.
15. Devlin,J., Chang,M.W., Lee,K. and Toutanova,K. (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv doi: https://arxiv.org/abs/1810.04805, 24 May 2019, preprint: not peer reviewed.
16. Yang,Z., Dai,Z., Yang,Y., Carbonell,J., Salakhutdinov,R. and Le,Q.V. (2019) Xlnet: generalized autoregressive pretraining for language understanding. arXiv doi: https://arxiv.org/abs/1906.08237, 02 January 2020, preprint: not peer reviewed.
17. Dong,L., Yang,N., Wang,W., Wei,F., Liu,X., Wang,Y. and Hon,H.W. (2019) Unified language model pre-training for natural language understanding and generation. arXiv doi: https://arxiv.org/abs/1905.03197, 15 October 2019, preprint: not peer reviewed.
18. Song,K., Tan,X., Qin,T., Lu,J. and Liu,T.Y. (2019) Mass: masked sequence to sequence pre-training for language generation. arXiv doi: https://arxiv.org/abs/1905.02450, 21 June 2019, preprint: not peer reviewed.
19. Liu,X., He,P., Chen,W. and Gao,J. (2019) Multi-task deep neural networks for natural language understanding. arXiv doi: https://arxiv.org/abs/1901.11504, 30 May 2019, preprint: not peer reviewed.
20. Lample,G. and Conneau,A. (2019) Cross-lingual language model pretraining. arXiv doi: https://arxiv.org/abs/1901.07291, 22 January 2019, preprint: not peer reviewed.
21. Lan,Z., Chen,M., Goodman,S., Gimpel,K., Sharma,P. and Soricut,R. (2019) Albert: a lite bert for self-supervised learning of language representations. arXiv doi: https://arxiv.org/abs/1909.11942, 09 February 2020, preprint: not peer reviewed.
22. Liu,Y., Ott,M., Goyal,N., Du,J., Joshi,M., Chen,D. and Stoyanov,V. (2019) Roberta: a robustly optimized bert pretraining approach. arXiv doi: https://arxiv.org/abs/1907.11692, 26 July 2019, preprint: not peer reviewed.
23. Clark,K., Luong,M.T., Le,Q.V. and Manning,C.D. (2020) Electra: Pre-training text encoders as discriminators rather than generators. arXiv doi: https://arxiv.org/abs/2003.10555, 23 March 2020, preprint: not peer reviewed.
24. Qiu,X., Sun,T., Xu,Y., Shao,Y., Dai,N. and Huang,X. (2020) Pre-trained models for natural language processing: a survey. *Sci. China: Technol. Sci.*, **63**, 1872–1897.
25. Beltagy,I., Lo,K. and Cohan,A. (2019) SciBERT: a pretrained language model for scientific text. arXiv doi: https://arxiv.org/abs/1903.10676, 10 September 2019, preprint: not peer reviewed.
26. Lee,J., Yoon,W., Kim,S., Kim,D., Kim,S., So,C.H. and Kang,J. (2020) BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, **36**, 1234–1240.
27. Rives,A., Goyal,S., Meier,J., Guo,D., Ott,M., Zitnick,C.L. and Fergus,R. (2021) Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl Acad. Sci. U.S.A.*, **118**, e2016239118.

28. Ji,Y., Zhou,Z., Liu,H. and Davuluri,R.V. (2021) DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. *Bioinformatics*, **37**, 2112–2120.
29. d'Ascoli,S., Touvron,H., Leavitt,M., Morcos,A., Biroli,G. and Sagun,L. (2021) Convit: improving vision transformers with soft convolutional inductive biases. *PMLR*, **139**, 2286–2296.
30. Dai,Z., Liu,H., Le,Q.V. and Tan,M. (2021) CoAtNet: marrying convolution and attention for all data sizes. *NeurIPS*, **34**, 3965–3977.
31. Umarov,R., Kuwahara,H., Li,Y., Gao,X. and Solovyev,V. (2019) Promoter analysis and prediction in the human genome using sequence-based deep learning models. *Bioinformatics*, **35**, 2730–2737.
32. Dreos,R., Ambrosini,G., Groux,R., Cavin Périer,R. and Bucher,P. (2017) The eukaryotic promoter database in its 30th year: focus on non-vertebrate organisms. *Nucleic Acids Res.*, **45**, D51–D55.
33. Benson,D.A., Cavanaugh,M., Clark,K., Karsch-Mizrachi,I., Ostell,J., Pruitt,K.D. and Sayers,E.W. (2018) GenBank. *Nucleic Acids Res.*, **46**, D41–D47.
34. Li,W., Wong,W.H. and Jiang,R. (2019) DeepTACT: predicting 3D chromatin contacts via bootstrapping deep learning. *Nucleic Acids Res.*, **47**, e60.
35. Noguchi,S., Arakawa,T., Fukuda,S., Furuno,M., Hasegawa,A., Hori,F. and Wolvetang,E. (2017) FANTOM5 CAGE profiles of human and mouse samples. *Scientific Data*, **4**, 170112.
36. Cunningham,F., Amode,M.R., Barrell,D., Beal,K., Billis,K., Brent,S. and Flicek,P. (2015) Ensembl 2015. *Nucleic Acids Res.*, **43**, D662–D669.
37. Javierre,B.M., Burren,O.S., Wilder,S.P., Kreuzhuber,R., Hill,S.M., Sewitz,S. and Fraser,P. (2016) Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. *Cell*, **167**, 1369–1384.
38. Zhou,J., Theesfeld,C.L., Yao,K., Chen,K.M., Wong,A.K. and Troyanskaya,O.G. (2018) Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat. Genet.*, **50**, 1171–1179.
39. Boix,C.A., James,B.T., Park,Y.P., Meuleman,W. and Kellis,M. (2021) Regulatory genomic circuitry of human disease loci by integrative epigenomics. *Nature*, **590**, 300–307.
40. Zhou,J., Park,C.Y., Theesfeld,C.L., Wong,A.K., Yuan,Y., Scheckel,C. and Troyanskaya,O.G. (2019) Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk. *Nat. Genet.*, **51**, 973–980.
41. Richter,F., Morton,S.U., Kim,S.W., Kitaygorodsky,A., Wasson,L.K., Chen,K.M. and Gelb,B.D. (2020) Genomic analyses implicate noncoding de novo variants in congenital heart disease. *Nat. Genet.*, **52**, 769–777.
42. Chen,K.M., Cofer,E.M., Zhou,J. and Troyanskaya,O.G. (2019) Selene: a pytorch-based deep learning library for sequence data. *Nat. Methods*, **16**, 315–318.
43. Schaid,D.J., Chen,W. and Larson,N.B. (2018) From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.*, **19**, 491–504.
44. Buniello,A., MacArthur,J.A.L., Cerezo,M., Harris,L.W., Hayhurst,J., Malangone,C. and Parkinson,H. (2019) The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.*, **47**, D1005–D1012.
45. Machiela,M.J. and Chanock,S.J. (2015) LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics*, **31**, 3555–3557.
46. Arloth,J., Eraslan,G., Andlauer,T.F., Martins,J., Iurato,S., Kühnel,B. and Mueller,N.S. (2020) DeepWAS: multivariate genotype-phenotype associations by directly integrating regulatory information using deep learning. *PLoS Comput. Biol.*, **16**, e1007616.
47. Kelley,D.R., Reshef,Y.A., Bileschi,M., Belanger,D., McLean,C.Y. and Snoek,J. (2018) Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Res.*, **28**, 739–750.
48. Avsec,Ž., Agarwal,V., Visentin,D., Ledsam,J.R., Grabska-Barwinska,A., Taylor,K.R., Assael,Y. and Jumper,J. (2021) Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods*, **18**, 1196–1203.
49. Mahajan,A., Taliun,D., Thurner,M., Robertson,N.R., Torres,J.M., Rayner,N.W. and McCarthy,M.I. (2018) Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat. Genet.*, **50**, 1505–1513.
50. Pickrell,J.K. (2014) Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am. J. Hum. Genet.*, **94**, 559–573.
51. Thurner,M., Van De Bunt,M., Torres,J.M., Mahajan,A., Nylander,V., Bennett,A.J. and McCarthy,M.I. (2018) Integration of human pancreatic islet genomic data refines regulatory mechanisms at type 2 diabetes susceptibility loci. *Elife*, **7**, e31977.
52. Varshney,A., Scott,L.J., Welch,R.P., Erdos,M.R., Chines,P.S., Narisu,N. and NISC Comparative Sequencing ProgramNISC Comparative Sequencing Program. (2017) Genetic regulatory signatures underlying islet gene expression and type 2 diabetes. *Proc. Natl. Acad. Sci. U.S.A.*,**114**, 2301–2306.
53. Kraja,A.T., Borecki,I.B., Tsai,M.Y., Ordovas,J.M., Hopkins,P.N., Lai,C.Q. and Arnett,D.K. (2013) Genetic analysis of 16 NMR-lipoprotein fractions in humans, the GOLDN study. *Lipids*, **48**, 155–165.
54. Dey,K.K., Van de Geijn,B., Kim,S.S., Hormozdiari,F., Kelley,D.R. and Price,A.L. (2020) Evaluating the informativeness of deep learning annotations for human complex diseases. *Nat. Commun.*, **11**, 4703.
55. Sham,P.C. and Purcell,S.M. (2014) Statistical power and significance testing in large-scale genetic studies. *Nat. Rev. Genet*, **15**, 335–346.
56. Bonàs-Guarch,S., Guindo-Martínez,M., Miguel-Escalada,I., Grarup,N., Sebastian,D., Rodriguez-Fos,E. and Torrents,D. (2018) Re-analysis of public genetic data reveals a rare X-chromosomal variant associated with type 2 diabetes. *Nat. Commun*, **9**, 321.
57. Xue,A., Wu,Y., Zhu,Z., Zhang,F., Kemper,K.E., Zheng,Z. and Yang,J. (2018) Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. *Nat. Commun.*, **9**, 2941.
58. Adamska-Patruno,E., Godzien,J., Ciborowski,M., Samczuk,P., Bauer,W., Siewko,K. and Kretowski,A. (2019) The type 2 diabetes susceptibility PROX1 gene variants are associated with postprandial plasma metabolites profile in non-diabetic men. *Nutrients*, **11**, 882.
59. Kretowski,A., Adamska,E., Maliszewska,K., Wawrusiewicz-Kurylonek,N., Citko,A., Goscik,J. and Gorska,M. (2015) The rs340874 PROX1 type 2 diabetes mellitus risk variant is associated with visceral fat accumulation and alterations in postprandial glucose and lipid metabolism. *Genes Nutr*, **10**, 4.
60. Fujita,H., Hara,K., Shojima,N., Horikoshi,M., Iwata,M., Hirota,Y. and Kadowaki,T. (2012) Variations with modest effects have an important role in the genetic background of type 2 diabetes and diabetes-related traits. *J. Hum. Genet.*, **57**, 776–779.
61. Hu,C., Zhang,R., Wang,C., Wang,J., Ma,X., Hou,X. and Jia,W. (2010) Variants from GIPR, TCF7L2, DGKB, MADD, CRY2, GLIS3, PROX1, SLC30A8 and IGF1 are associated with glucose metabolism in the Chinese. *PLoS One*, **5**, e15542.
62. Zhao,W., Rasheed,A., Tikkanen,E., Lee,J.J., Butterworth,A.S., Howson,J.M. and EPIC-Interact ConsortiumEPIC-Interact Consortium. (2017) Identification of new susceptibility loci for type 2 diabetes and shared etiological pathways with coronary heart disease. *Nat. Genet.*, **49**, 1450.
63. Voight,B.F., Scott,L.J., Steinthorsdottir,V., Morris,A.P., Dina,C., Welch,R.P. and Thorsteinsdottir,U. (2010) Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat. Genet.*, **42**, 579.
64. Harder,M.N., Ribel-Madsen,R., Justesen,J.M., Sparsø,T., Andersson,E.A., Grarup,N. and Pedersen,O. (2013) Type 2 diabetes risk alleles near BCAR1 and in ANK1 associate with decreased β-cell function whereas risk alleles near ANKRD55 and GRB14 associate with decreased insulin sensitivity in the danish inter99 cohort. *J. Clin. Endocrinol. Metab.*, **98**, E801–E806.
65. Vujkovic,M., Keaton,J.M., Lynch,J.A., Miller,D.R., Zhou,J., Tcheandjieu,C. and Saleheen,D. (2020) Discovery of 318 new risk loci for type 2 diabetes and related vascular outcomes among 1.4 million participants in a multi-ancestry meta-analysis. *Nat. Genet*, **52**, 680–691.
66. Davydov,E.V., Goode,D.L., Sirota,M., Cooper,G.M., Sidow,A. and Batzoglou,S. (2010) Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.*, **6**, e1001025.

67. Fu,Y., Liu,Z., Lou,S., Bedford,J., Mu,X.J., Yip,K.Y. and Gerstein,M. (2014) FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol.*, **15**, 480.
68. Huang,Y.F., Gulko,B. and Siepel,A. (2017) Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat. Genet.*, **49**, 618–624.
69. Rentzsch,P., Witten,D., Cooper,G.M., Shendure,J. and Kircher,M. (2019) CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.*, **47**, D886–D894.
70. Di Iulio,J., Bartha,I., Wong,E.H., Yu,H.C., Lavrenko,V., Yang,D. and Telenti,A. (2018) The human noncoding genome defined by genetic diversity. *Nat. Genet.*, **50**, 333–337.
71. Stenson,P.D., Ball,E.V., Mort,M., Phillips,A.D., Shiel,J.A., Thomas,N.S. and Cooper,D.N. (2003) Human gene mutation database (HGMD®): 2003 update. *Hum. Mutat.*, **21**, 577–581.
72. 1000 Genomes Project Consortium. (2015) A global reference for human genetic variation. *Nature*, **526**, 68.
73. Siepel,A., Bejerano,G., Pedersen,J.S., Hinrichs,A.S., Hou,M., Rosenbloom,K. and Haussler,D. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.
74. Pollard,K.S., Hubisz,M.J., Rosenbloom,K.R. and Siepel,A. (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.*, **20**, 110–121.
75. Cooper,G.M., Stone,E.A., Asimenos,G., Green,E.D., Batzoglou,S. and Sidow,A. (2005) Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.*, **15**, 901–913.
76. Landrum,M.J., Lee,J.M., Riley,G.R., Jang,W., Rubinstein,W.S., Church,D.M. and Maglott,D.R. (2014) ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.*, **42**, D980–D985.
77. Zhou,L. and Zhao,F. (2018) Prioritization and functional assessment of noncoding variants associated with complex diseases. *Genome Medicine*, **10**, 53.
78. Wright,S. (1965) The interpretation of population structure by F-statistics with special regard to systems of mating. *Evolution*, **19**, 395–420.
79. Quintana-Murci,L. (2016) Understanding rare and common diseases in the context of human evolution. *Genome Biol.*, **17**, 225.
80. Tay,Yi, Bahri,D., Metzler,D., Juan,Da-C, Zhao,Z. and Zheng,C. (2020) Synthesizer: rethinking self-attention for transformer models.*PMLR*, **139**, 10183–10192.
81. Wu,C., Wu,F., Qi,T. and Huang,Y. (2021) Hi-Transformer: hierarchical interactive transformer for efficient and effective long document modeling. *ACL*, **2**, 848–853.
82. Ward,L.D. and Kellis,M. (2016) HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. *Nucleic Acids Res.*, **44**, D877–D881.
83. Lonsdale,J., Thomas,J., Salvatore,M., Phillips,R., Lo,E., Shad,S. and Moore,H.F. (2013) The genotype-tissue expression (GTEx) project. *Nat. Genet.*, **45**, 580–585.
84. Greene,C.S., Krishnan,A., Wong,A.K., Ricciotti,E., Zelaya,R.A., Himmelstein,D.S. and Troyanskaya,O.G. (2015) Understanding multicellular function and disease with human tissue-specific networks. *Nat. Genet.*, **47**, 569–576.
85. Alley,E.C., Khimulya,G., Biswas,S., AlQuraishi,M. and Church,G.M. (2019) Unified rational protein engineering with sequence-based deep representation learning. *Nat. Methods*, **16**, 1315–1322.
86. Repecka,D., Jauniskis,V., Karpus,L., Rembeza,E., Rokaitis,I., Zrimec,J. and Zelezniak,A. (2021) Expanding functional protein sequence spaces using generative adversarial networks. *Nat. Mach. Intell.*, **3**, 324–333.