Nucleic Acids Research, 2025, **53**, gkae1310
https://doi.org/10.1093/nar/gkae1310
Advance access publication date: 15 January 2025
Methods

GENA-LM: a family of open-source foundational DNA language models for long sequences

Veniamin Fishman[1,2,*,†], Yuri Kuratov[1,3,†], Aleksei Shmelev[1,4], Maxim Petrov[1], Dmitry Penzar[1], Denis Shepelin[1], Nikolay Chekanov[1], Olga Kardymon[1,*] and Mikhail Burtsev[5,*]

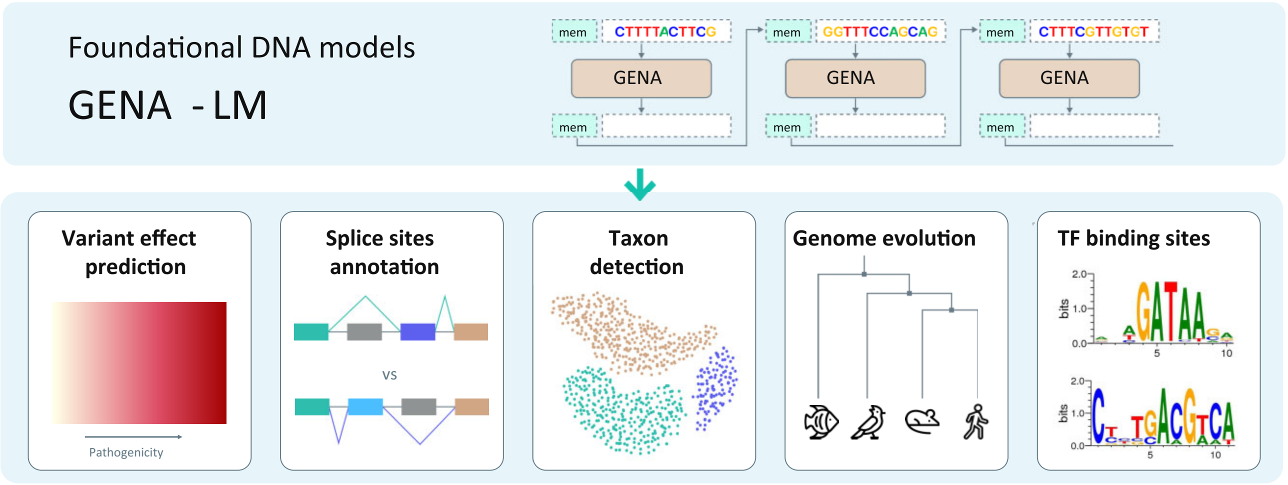
[1] AIRI, Presnenskaya embankment, 6 st22, Moscow, 123112, Russia
[2] Institute of Cytology and Genetics, Prospekt Akademika Lavrent'yeva, 10, Novosibirsk, 630090, Russia
[3] Moscow Institute of Physics and Technology, 9 Institutskiy per., Dolgoprudny, Moscow, 141701, Russia
[4] HSE University, International laboratory of statistical and computational genomics, Moscow, 109028, Russia
[5] London Institute for Mathematical Sciences Royal Institution, 21 Albemarle St, London W1S 4BS, UK

[*] To whom correspondence should be addressed. Email: mb@lims.ac.uk
Correspondence may also be addressed to Veniamin Fishman. Email: minja-f@ya.ru
Correspondence may also be addressed to Olga Kardymon. Email: kardymon@airi.net
[†] The first two authors should be regarded as Joint First Authors.

Abstract

Recent advancements in genomics, propelled by artificial intelligence, have unlocked unprecedented capabilities in interpreting genomic sequences, mitigating the need for exhaustive experimental analysis of complex, intertwined molecular processes inherent in DNA function. A significant challenge, however, resides in accurately decoding genomic sequences, which inherently involves comprehending rich contextual information dispersed across thousands of nucleotides. To address this need, we introduce GENA language model (GENA-LM), a suite of transformer-based foundational DNA language models capable of handling input lengths up to 36 000 base pairs. Notably, integrating the newly developed recurrent memory mechanism allows these models to process even larger DNA segments. We provide pre-trained versions of GENA-LM, including multispecies and taxon-specific models, demonstrating their capability for fine-tuning and addressing a spectrum of complex biological tasks with modest computational demands. While language models have already achieved significant breakthroughs in protein biology, GENA-LM showcases a similarly promising potential for reshaping the landscape of genomics and multi-omics data analysis. All models are publicly available on GitHub (https://github.com/AIRI-Institute/GENA_LM) and on HuggingFace (https://huggingface.co/AIRI-Institute). In addition, we provide a web service (https://dnalm.airi.net/) allowing user-friendly DNA annotation with GENA-LM models.

Graphical abstract



Received: September 16, 2024. **Editorial Decision:** December 8, 2024. **Accepted:** December 26, 2024

Introduction

The encoding of genetic information by DNA is a principal subject in biology, involving both straightforward and complex systems of translation and epigenetic coding, respectively. While the translation of messenger RNA to amino acid sequences employs a widely accepted genetic code, other forms of encoding, notably the epigenetic code, are more challenging (1). DNA sequences dictate functional genome elements, including promoters, enhancers and transcription factor (TF) binding sites, among others. However, the diversity and redundancy of their underlying motifs challenge their detection within vast eukaryotic genomes, complicating insights into non-coding genome evolution and interpretations of human genomic variants, given the yet-to-be-fully-unraveled complexity of the epigenetic code.

The advent of next-generation sequencing and additional high-throughput technologies has catalyzed the accumulation and public deposition of extensive databases, rich with functional genomic elements, enabling the broad application of computational methods to large-scale genomic data analysis (2). We, along with others (3), have successfully employed machine-learning methods, including ensemble learning (4) and convolutional neural networks (5,6), for this purpose. However, while potent, these approaches encounter constraints in identifying long-range dependencies within DNA sequences, a common phenomenon in human and other eukaryotic genomes (7). Recent strategies employing transformer neural network-based approaches seek to surmount these constraints (8), with cutting-edge transformer architectures showcasing the capability to infer-specific epigenetic properties and gene expression levels from DNA sequences with exceptional precision (8). However, training models tailored to specific tasks demands substantial computational resources, and their inference capabilities are inherently constrained by the targets represented in the training dataset.

Transfer learning, especially through pre-training, has been widely adopted in natural language processing (NLP) for its capacity to enhance computational efficiency and performance in scenarios with limited target data (9–14). Models pre-trained on substantial unlabeled datasets can be fine-tuned or utilized as feature extractors for new tasks, frequently outperforming models trained on task-specific datasets, particularly when those datasets are smaller. The application of this approach to bioinformatics is exemplified by the development of DNABERT (15), a BERT-like transformer neural network (14,16) pre-trained on the human genome to predict subsequences from context, and subsequently fine-tuned for downstream tasks such as promoter activity prediction and TF binding. While DNABERT signifies a promising advance, its applicability is hindered by an input size cap of 500 base pairs (bp), and recent DNABERT extension DNABERT-2 (17) also has an input length limit of ∼1–4 kb. This input length limitation restricts the ability of the models to capture the extended contexts vital for various genomic applications.

Enhancing input size for transformer models has recently been addressed through several developments, including sparse attention, effective attention and recurrence. Sparse attention techniques, which utilize either predefined or learned attention patterns like sliding window or block-diagonal, linearize the quadratic dependency of full attention on input length (18–22). Conversely, linear attention methods approximate full token-to-token interactions through softmax linearization (23,24). In the domain of recurrent models, inputs are segmented and sequentially processed, with intersegment information relayed through prior hidden states (25,26) or specialized memory (27–29). Notably, the recently introduced recurrent memory transformer (RMT) architecture facilitates information aggregation from both long (29) and extremely long input sequences (30), spanning thousands to millions of elements, respectively.

In this work, we introduce GENA language model (GENA-LM), a family of transformer-based foundational DNA models. After fine-tuning for predictive analysis of various functional genomic elements—including promoter activity, splicing, polyadenylation sites, enhancer annotations and chromatin profiles—GENA-LM models demonstrate state-of-the-art performance for a significant fraction of tasks, achieving top average performance relative to other models. Moreover, our augmentation of GENA-LM with the RMT enables tackling genomic tasks that require substantial input sequence lengths. We also explore new applications of GENA-LM, such as identifying DNA motifs essential for TF binding and assessing mutation effects in promoters and splice sites to aid in the prioritization of clinical variants. To broaden the model's utility beyond human genomes, we have developed and released species-specific models for yeast, Arabidopsis and Drosophila, as well as a multispecies model. To facilitate sequence annotation for the thousands of unannotated sequences now available, we have developed GENA-Web (https://dnalm.airi.net), a web service that generates various annotations based on DNA sequence input. We contribute to the research community by open-sourcing the GENA-LM family on GitHub (https://github.com/AIRI-Institute/GENA_LM) and providing pre-trained models (prefixed with `gena-lm-`) on HuggingFace (https://huggingface.co/AIRI-Institute).

Materials and methods

Datasets

Genomic datasets for language model pre-training

Dataset sources

Human T2T v2 genome assembly was downloaded from NCBI (acc. GCF_009914755.1). Genomic datasets used to train multispecies models were downloaded from ENSEMBL release 106 (https://ftp.ensembl.org/pub/release-106/). The list of species is provided in Supplementary Table S4. For the 1000-genome dataset, we used gnomAD v3.1.2 data. For taxon-specific models, we used the following resources:

(1) Arabidopsis model: Data were obtained from (31) and contain chromosome-level genomes of 32 *Arabidopsis thaliana* ecotypes.
(2) Yeasts model: Data were obtained from (32) and include telomere-to-telomere assemblies of 142 yeast strains.
(3) Drosophila model: Data were obtained from Progressive Cactus alignment of 298 drosophilid species generated by (33).

Genomic datasets preprocessing

To prepare genomic datasets for our training corpus, we processed each record in the genomic FASTA files. We excluded

Table 1. Parameters of downstream tasks datasets

Downstream task	Input length (bp)	Number of targets	Task
Promoters prediction (300)	300	2	Classification
Promoters prediction (2000)	2000	2	Classification
Promoters prediction (16 000)	16 000	2	Classification
Splice site prediction	15 000	3 per token / bp	Multiclass classification
Drosophila enhancers prediction	249	2	Regression
Chromatin profiling (1000)	1000	919	Multilabel classification
Chromatin profiling (8000)	8000	919	Multilabel classification
Polyadenylation sites prediction	443	1	Regression

contigs with the substring 'mitochondrion' in their identifiers and those shorter than 10 kb. From the remaining sequences, we divided them into 'sentences' spanning 500–1000 bp—the sentence length being randomized—and compiled 'documents' with 50–100 consecutive sentences. This approach follows the data processing in BigBird (21). Data augmentation incorporated reverse-complement sequences, and we applied a stochastic shift for some documents to include overlapping genomic sequences.

For the 1000-genome Single Nucleotide Polymorphism (SNP) augmentation, nucleotide substitution was executed, replacing reference alleles with alternative ones sourced from individual samples of 1000-genome cohort. In order to maintain the haplotype structure, each sample was processed individually. This meant that for every genomic region, multiple sequences were derived, each resulting from swapping reference alleles with sample-specific alternative variants from singe individual. We limited our focus to genomic regions where the proportion of positions with a noted variant for a given sample exceeded 0.01. No allele frequency filter was applied.

Train and test split

For our initial models, *bert-base* and *bigbird-base-sparse*, we hold out human chromosomes 22 (CP068256.2) and Y (CP086569.2) as the test datasets for the masked language modeling (MLM) task. In contrast, for subsequent models, identifiable by the 't2t' suffix in their names, we hold out human chromosomes 7 (CP068271.2) and 10 (CP068268.2) for testing. All remaining data were used for training.

Models focusing exclusively on human data were trained using the preprocessed Human T2T v2 genome assembly combined with its 1000-genome SNP augmentations, totaling $\approx 480 \times 10^9$ bp. On the other hand, multispecies models incorporated both the human-only and multispecies data, aggregating to roughly $\approx 1072 \times 10^9$ bp.

The data splitting strategy for downstream tasks was anchored to methodologies previously described in literature relevant for each particular downstream task. Comprehensive specifics for each task are provided in their respective dedicated sections.

Sequence tokenization

We employed Byte-Pair Encoding (BPE) tokenization (34) for our models, setting the dictionary size to 32 000 and initializing with a character-level vocabulary comprised of ['A', 'T', 'G', 'C', 'N']. Our study utilized two distinct tokenizers:

(1) The first tokenizer, trained exclusively on the human T2T v2 genome assembly, is denoted as 'T2T split v1' in Table 2.

(2) The second tokenizer, trained on a mixture of human-only and multispecies data sampled equally, is labeled 'T2T+1000G SNPs+Multispecies'.

Both tokenizers incorporate special tokens: CLS, SEP, PAD, UNK and MASK. Notably, the 'T2T+1000G SNPs+Multispecies' tokenizer integrates a preprocessing step to manage extensive gaps: sequences with over 10 consecutive 'N' characters are consolidated into a singular '–' token.

Downstream task datasets

A concise overview of the dataset parameters for downstream tasks is presented in Table 1. A comprehensive description follows.

Promoters prediction

For the task of predicting promoters, we sourced human sequences located upstream of TSS (transcriptional start sites) from the EPDnew database (https://epd.epfl.ch/EPDnew select.php). Sequences of lengths 300, 2000 and 16 000 bp were extracted, with each dataset being processed and assessed independently. For negative samples generation, we randomly selected genomic locations outside promoter sequences, ensuring that negative and positive samples do not overlap for maximum promoter length (16 kb). The entire dataset was segregated by sequence into training, validation and testing sets. The objective of this task is a binary classification: determining the presence or absence of a promoter within a given region.

Splice site prediction

To predict splice donor and acceptor sites, we replicated the dataset from (35), utilizing the original scripts provided by the authors. We adhered to the same training and testing splits as outlined in (35). In this dataset, a central 5000-bp target region is bracketed by 10 000 bp of context, with 5000 bp on each side. Splice site annotations within the target region are aligned to token positions. Tokens overlapping with either splice-donor or splice-acceptor sites are designated as positive samples for their respective splicing annotation class. Subsequently, both the target and its context were tokenized independently. If the combined length diverged from the model's input size, adjustments were made through either padding or truncation. In the event of truncation, sequences furthest from the target region's midpoint were first removed. We demarcated the context and target sequences using SEP tokens. Through this procedure, the target's size matched the model's input token count. However, the computational loss did not account for tokens representing either context or padding. This challenge is a multiclass, token-level classification task encompassing three categories: splice donor, splice acceptor and none.

Drosophila enhancers prediction

Candidate sequences, along with their associated housekeeping and tissue-specific activity in Drosophila cells, were sourced from the Stark Lab repository (https://data.starklab.org/almeida/DeepSTARR/Data/). These datasets are partitioned into training, validation and testing sets, consistent with those used for training the DeepSTARR model (36). The task at hand involves a two-class regression, wherein each 249-bp sequence is predicted to produce two continuous scores: one for housekeeping enhancer activity and another for developmental enhancer activity.

Chromatin profiling

We gained the DeepSEA dataset (37) from its original repository (http://deepsea.princeton.edu/media/co de/deepsea train bundle.v0.9.tar.gz). This dataset outlines the chromatin occupancy profiles of various genomic features, encompassing histone marks (HMs), TFs and DNAse I hypersensitivity (DHS) regions. The dataset comprises DNA sequences of 1000 bp, with a central 200-bp target region flanked by 400-bp contexts on either side. Each feature's occupancy is quantified over this 200-bp target. Additionally, we trialed an expanded context of 7800 bp (yielding a total input length of 8000 bp). To elongate the DNA context, we aligned the input DNA segments to the hg19 genome using *bwa fastmap*. Surrounding sequences at mapped sites were then harvested. Sequences that either failed remapping or aligned too proximate to a chromosome's terminus to permit extension were omitted, though these comprised <1% of the dataset. Our partitioning for training, validation and testing adhered to the divisions presented in the original DeepSEA dataset. The challenge is a multilabel classification, with class count reflecting the unique epigenetic profiles identified in DeepSEA (919 in total).

Polyadenylation sites prediction

For predicting polyadenylation sites, we employed the APARENT dataset (38) (available at https://github.com/johli/aparent). This dataset characterizes the frequency with which transcription machinery recognizes specific nucleotide sequences as polyadenylation signals. Utilizing the scripts published by the authors, we extracted the target values and delineated the training and testing datasets. Furthermore, we retrieved APARENT predictions (noted under the field *iso_pred*) to gauge the performance of the APARENT model. We tokenized the sequences from both upstream and downstream segments of the 5'-untranslated regions individually, and they were demarcated using a SEP token. This study focuses on regression analysis targeting 256-bp sequences.

Nucleotide Transformer dataset

The dataset and literature scores were obtained from the Nucleotide Transformer (NT version 2 scores) (39).

HyenaDNA species classification dataset

The dataset was reconstructed based on the description provided in (40). Genomes from five species (human, lemur, mouse, pig and hippo) were downloaded from NCBI (RefSeq assemblies GCF_000001405.40, GCF_020740605.2, GCF_000001635.27, GCF_000003025.6 and GCF_030028045.1, respectively). Four chromosomes (chromosomes 1, 3, 12 and 13) were used for models evaluation, other chromosomes were utilized during training. We sampled sequences from chromosomes randomly, using the uniform distribution. We used a five-way classification and reported top-1 accuracy. For each task length, we collected a total of 50 000 DNA subsequences from each species, ensuring a comprehensive dataset for our analysis.

Models architecture and training

DNA language models based on transformer architecture

We trained and expanded upon several transformer models, drawing inspiration from both BERT (14) and BigBird (21) architectures. These adapted models are consistently referred to as `GENA-LM` throughout this manuscript. Key distinctions between these architectures can be found in Table 2. A comprehensive breakdown of parameters and specific combinations for each model is available in Supplementary Table S5. Additionally, we enhanced BERT-based models with pre-layer normalization (41). In instances where the layer normalization is applied even to the final layer output, it is distinctly mentioned as *lastln* in the model names. For precise parameter details, refer to Supplementary Table S5.

All models were pre-trained using the MLM objective. During this process, the sequence was tokenized and flanked by the special tokens, CLS and SEP. In alignment with the BERT pre-training methodology, 15% of the tokens were randomly selected for prediction. Among these, 80% were replaced with MASK tokens, 10% were swapped with random tokens and the remaining 10% were retained unchanged. Training extended for 1–2 million steps, utilizing a batch size of 256 and operated on 8 or 16 NVIDIA A100 GPUs. We employed the FusedAdam implementation of the AdamW optimizer (42), made available through Nvidia Apex (https://github.com/NVIDIA/apex). The initial learning rate was set at 1×10^{-4}, including a warm-up phase. For most models, we adopted a linear learning rate decay, but in cases where pre-training diverged, we manually adjusted the learning rate.

GENA-LM fine-tuning

In our standard procedure, we tokenize input sequences and prepend and append them with the service tokens CLS and SEP, respectively. To ensure compatibility with the model's input requirements, sequences are either padded or truncated as needed. For datasets necessitating specialized tokenization, the specific preprocessing steps are detailed in the relevant dataset section.

Tokenized sequences were provided as inputs to downstream models. These models utilized one of the pre-trained GENA-LM architectures, augmented with a single fully connected output layer. The dimensions of this layer are denoted by `(hidden_size, target_size)`. Here, `hidden_size` refers to the hidden unit size specific to the GENA-LM model (refer to Supplementary Table S5), while `target_size` is specified in the description of each downstream task dataset discussed earlier. For single-label, multiclass classification tasks, we implemented a softmax activation function on the final layer, paired with cross-entropy loss. In contrast, multilabel, multiclass classification tasks employed a sigmoid activation function on the last layer, combined with a binary cross-entropy with logits loss. Regression tasks did not necessitate any activation function on the last layer and utilized mean squared error as the loss function. To address sequence

Table 2. Overview of the GENA-LM foundational DNA language models

Model	Architecture	Maximum seq len, tokens (≈ bp)	Tokenizer data	Training data
DNABERT	**BERT-12L**	**512 (512)**	**3,4,5,6-mer**	**GRCh38.p13**
GENA-LM models:				
bert-base	BERT-12L	512 (4500)	T2T split v1	T2T split v1
bert-base-t2t	BERT-12L	512 (4500)	T2T+1KG+M	T2T+1KG
bert-base-lastln-t2t	BERT-12L	512 (4500)	T2T+1KG+M	T2T+1KG
bert-base-t2t-multi	BERT-12L	512 (4500)	T2T+1KG+M	T2T+1KG+M
bert-base-t2t-yeast	BERT-12L	512 (4500)	T2T+1KG+M	Yeast
bert-base-t2t-fly	BERT-12L	512 (4500)	T2T+1KG+M	Drosophila
bert-base-t2t-athaliana	BERT-12L	512 (4500)	T2T+1KG+M	*A. thaliana*
bert-large-t2t	BERT-24L	512 (4500)	T2T+1KG+M	T2T+1KG
bigbird-base-sparse	BERT-12L, RoPE DS sparse attention	4096 (36 000)	T2T split v1	T2T split v1
bigbird-base-sparse-t2t	BERT-12L, RoPE DS sparse attention	4096 (36 000)	T2T+1KG+M	T2T+1KG
bigbird-base-t2t	BERT-12L HF sparse attention	4096 (36 000)	T2T+1KG+M	T2T+1KG

This table delineates the specifications of pre-trained GENA-LM models, highlighting variations in pre-training data, layer count, attention type and sequence length. Models archived on the HuggingFace model hub adhere to a consistent naming convention, prefixed by AIRI-Institute/gena-lm-. Models based on the BERT architecture utilize pre-layer normalization (41), with lastln indicating the application of layer normalization to the output of the terminal layer. 'T2T split v1' alludes to initial experiments using a non-augmented T2T human genome assembly split. The term '1KG' is shorthand for 1000G SNPs augmentations, while 'M' denotes the inclusion of multispecies data. The designations 'DS sparse' and 'HF sparse' are references to the DeepSpeed sparse attention and HuggingFace BigBird implementations, respectively. The abbreviation 'RoPE' signifies the adoption of rotary position embeddings (43) as an alternative to BERT's absolute positional embeddings. The models were structured with either 12 (denoted as BERT-12L) or 24 (denoted as BERT-24L) layers, comprising 110M and 336M parameters, respectively.

classification and regression tasks, we used the hidden state of the CLS token from the final layer. Meanwhile, for token-level classification tasks, such as splice site prediction, all hidden states from the ultimate layer were employed. Both the weights of the final fully connected layer and the parameters of the entire GENA-LM were fine-tuned during this process. Learning rate warm-up (44) was consistently applied across all tasks. The optimal number of training and warm-up steps was determined empirically for each individual task.

GENA-LM fine-tuning with recurrent memory

The recently introduced RMT presents a novel approach to extend the context length of pre-trained models (29). Unlike traditional transformers, which exhibit quadratic computational complexity in their attention layers, the RMT employs a recurrent mechanism to efficiently manage elongated sequences. This recurrent design ensures constant memory consumption and linear computational scaling with context length. To process input, the RMT divides the sequence into distinct segments, processing them in a sequential manner. Special memory tokens are integrated into the input of each segment. For a given segment, the outputs linked to its memory tokens are subsequently utilized as input vectors for memory tokens for the succeeding segment. By this method, a multilayer transformer, such as the pre-trained GENA-LM, functions as a single recurrent cell, addressing one segment at a time.

For both promoter and splice site prediction tasks, we segmented the input sequence into units, with each containing 512 tokens (∼4.5 kb). The initial 10 tokens of every sequence were allocated for memory tokens. Segments were processed in a sequential manner, where outputs from the memory tokens of one segment are used as the input memory tokens of the subsequent segment. During the training phase, gradients were allowed to propagate from the final segment to the initial one through these memory tokens. We did not impose any restrictions on the number of unrolls in backpropagation through time, allowing gradients to flow uninterrupted from the final to the initial segment. The initial states designated for memory tokens were randomly initialized, and further refined during the fine-tuning process. For the task of promoter prediction, we restricted loss computation to only the last segment. Conversely, for splice site prediction, the loss was determined for every individual segment. The training employed the AdamW optimizer and learning rates of {1e−04, 5e−05, 2e−05}. With a batch size set at 128, the training was terminated when there were no discernible improvements in validation scores. The results for the promoter prediction task are presented as averages over five folds. Meanwhile, the splice site prediction task results are averages across three runs, each employing a distinct random initialization. Training scripts are accessible within our provided codebase.

For the species classification task, we used *gena-lm-bert-base-t2t* model that has been augmented with RMT (eight segments) during the pre-training phase. The processes of fine-tuning were enhanced through the application of a curriculum learning strategy. This meant that our initial step included fine-tuning the model on DNA subsequences of 1000 bp in length (single segment). Following this initial phase, we proceeded to extend the fine-tuning process to handle longer DNA subsequences while using the model weights from the 1000-bp fine-tuned model as initial weights, increasing the challenge to a length of 32 kb (eight segments). Continuing with this progressive training methodology, we further advanced our model's capabilities by eventually fine-tuning it to efficiently process and analyze DNA subsequences extending up to 50 kb in length (12 segments). This gradual fine-tuning approach, in line with the principles of curriculum learning, facilitated the model in sequentially mastering tasks of escalating complexity, thereby enhancing its analytical precision and performance on genetic classification tasks.

Phylogenetic analysis using GENA-LMs without recurrent memory

For our phylogenetic analysis, we randomly sampled 500 subsequences from each genomic sequence, as detailed in Supplementary Table S3. To ensure representative sampling across entire genomes, the probability of selecting a sequence from a specific chromosome was proportionate to the chromosome's length. In instances not otherwise specified, we utilized the embedding of the CLS token from the final layer. Sequences shorter than 5 kb were processed using the *bert-large-t2t* model, whereas sequences exceeding this length were analyzed with the *big-bird-base-t2t* model to accommodate the extended context. For classifying species, we employed the HistGradientBoostingClassifier from the sklearn library, retaining its default parameters.

Classification of human promoter mutations

We curated records from the ClinVar database as of 1 July 2024, applying several filters to ensure data quality and relevance. Only records with >1 piece of evidence (i.e. filtering by the *single_submitter* field) and no conflicting interpretations of significance (as indicated in the *conflicting_interpretations* field) were included. We retained only variants with consequences classified as either benign or pathogenic. To focus on regulatory variants, we excluded variants overlapping exons, as defined by Gencode V45. Additionally, we filtered out variants located on sex chromosomes and mitochondrial variants. From the filtered dataset, we specifically selected those single-nucleotide substitutions that overlap with promoters within the 2 kb EPDnew promoter dataset described previously.

For each variant, we selected all overlapping promoters and computed the log odds value as follows: $OR = \log(\frac{p}{1-p})$, where p represents the promoter presence probability derived from the *gena-lm-bert-large-t2t* model, which was fine-tuned on a 2-kb length human promoter dataset. The OR was calculated for both the reference sequence and the sequence containing the mutation, and their absolute difference was utilized as the mutation score. If a single mutation overlapped several promoters, the highest score among them was used.

Cross-species epigenetic analysis

For the cross-species analysis of H3K27ac and CTCF binding sites, we collected Chromatin Immunoprecipitation followed by Sequencing (ChIP-seq) data from NCBI, with accession numbers listed in Supplementary Table S6, and uniformly processed them using the MACS3 software. For each dataset, we filtered out peak calls located on scaffolds shorter than 200 kb, and from the remaining data, we randomly selected 2000 binding sites as positive samples. All genomic regions located at least 8-kb away from any positive sample were designated as negative samples, and 2000 negative samples were randomly chosen for each dataset. We next computed the center of each sample and collected 2000 bp of the flanking sequences (±1000 bp) to provide contextual information.

Each genome was randomly split into five folds, ensuring that sequences from different folds did not overlap and that chromosomes were uniformly distributed across the folds. We fine-tuned the *gena-lm-bert-base-t2t* model using the human SRR10182244 dataset for H3K27ac and the human SRR26329064 dataset for CTCF. When assessing performance on human datasets, we utilized sequences exclusively from one fold (fold 1), which was held out during training. For non-human species, since their data were not included during the human model's fine-tuning, we performed evaluations on each of the five folds. Consequently, results for each human dataset evaluation are depicted by a single point in Figure 3A and B, while results for non-human dataset evaluations are represented by five points.

Cross-species promoter inference

We evaluated the *gena-lm-bert-large-t2t* model, which was fine-tuned on human promoter sequences, using data from seven species: macaque, mouse, rat, dog, zebrafish, chicken and *Caenorhabditis elegans*. For each species, we downloaded promoter sequences from EPDnew and prepared the data into five folds, following the same procedure used for the human dataset. For each species, we conducted 25 evaluation experiments by cross-applying the models trained on each of the five human dataset folds to each of the five species-specific dataset folds. For human data, we provide five evaluation results obtained on each of the human dataset folds.

Token attribution analysis

We employed the Integrated Gradients algorithm (45) to conduct token attribution analysis. For epigenetic data analysis, we utilized the *bigbird-base-sparse-t2t* model, which was fine-tuned on the standard DeepSEA dataset with sequences of 1000 bp. Despite the dataset comprising over 900 features, our analysis specifically targeted six key features: ATF1, CTCF, GATA2, H3K27me3, H3K9me3 and H3K4me1 ChIP-seq profiles from untreated K562 cells. For each genomic feature, we randomly chose 3000 nucleotide sequences that encompassed ChIP-seq peaks. Subsequent tokenization of these sequences adhered to the same methodology as that applied in the chromatin profile fine-tuning task. With default parameters set, token attribution values were derived. For motif analysis, we leveraged the XSTREME tool (46). Both FIMO and XSTREME assessments sourced motifs from the HOCOMOCO v11 database (47).

For token importance analysis concerning promoter mutations, we utilized the *gena-lm-bert-large-t2t* model fine-tuned on the 2-kb length promoter dataset as previously described. For each mutation, we identified all overlapping promoter regions and calculated token importance scores using Integrated Gradients. For each sample, the top-1 percentile of tokens were designated as 'highly important', while the remainder were classified as 'not important'. We then overlapped mutations with these tokens and categorized each mutation as either 'overlapping highly important token' or 'overlapping not important token'. If mutation overlapped tokens from both classes, the 'highly important token' class was assigned. Finally, we compared the distribution of pathogenic and benign mutations across the 'overlapping highly important token' and 'overlapping not important token' classes using a chi-squared test.

Code availability

The code to generate the findings of this manuscript is available in the 'supplementary code' section and on

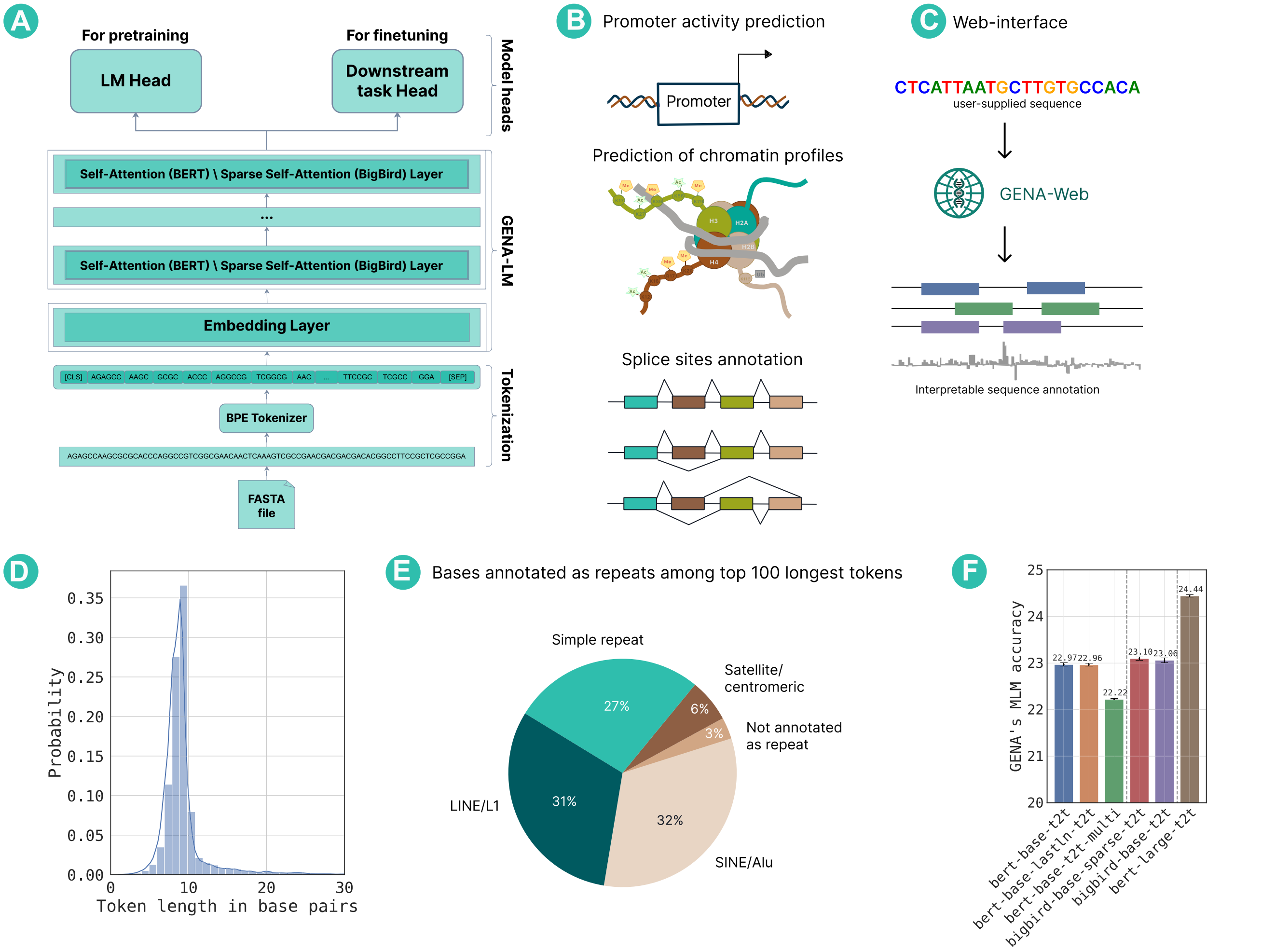


Figure 1. The GENA-LM family of foundational DNA language models. (**A**) The GENA-LM transformer-based architecture is pre-trained on DNA sequences using an MLM objective. GENA-LMs encompass a variety of models that differ in their pre-training data and architecture, as detailed in Table 2. All models adhere to the same workflow: DNA sequences are tokenized using a BPE algorithm before being processed through transformer layers, which generate representations of the input sequences that are suitable for downstream applications. Post pre-training, this foundational DNA model incorporates a downstream task-specific head, which utilizes DNA representations to address specific genomic tasks during the fine-tuning process. (**B**) GENA's evaluation tasks include predictions related to promoter and enhancer activities, splicing sites, chromatin profiles and polyadenylation site strength (not all shown). (**C**) Task-specific fine-tuned models can be queried via web service (https://dnalm.airi.net/). (**D**) Post-BPE tokenization, the median token length stands at nine bp, as reflected in the token length distribution. (**E**) Illustration of repetitive element representation for the 100 longest tokens. (**F**) GENA's model accuracies for pre-training on the MLM task demonstrate that models with a higher parameter count achieve superior performance.

our GitHub repository: https://github.com/AIRI-Institute/GENA_LM. Additionally, our trained models can be found on HuggingFace under the prefix 'gena-lm': https://huggingface.co/AIRI-Institute/.

Results

Family of pre-trained transformer-based GENA-LM models

In this study, we introduce a new universal transformer model tailored for nucleic acid sequences, which offers several improvements over existing models such as DNABERT (15) and BigBird (21) (as depicted in Figure 1A). To ensure its versatility across various applications, we pre-trained our model using multiple datasets and diverse input sequence lengths.

In the data preprocessing phase, we have extended established pipelines by integrating BPE for sequence tokenization (Figure 1A, bottom panel). The essence of BPE is that it constructs a sequence dictionary to pinpoint the most frequently occurring subsequences within the genome. This results in tokens of diverse lengths, ranging from a single base pair up to 64 bp. In our tests, the median token length was determined to be 9 bp (Figure 1D). Interestingly, our BPE vocabulary revealed tokens of significant biological relevance. For example, the longest tokens were often indicative of familiar repetitive elements, such as LINEs or simple repeats (Figure 1E). The tokenization approach we adopted is greedy, starting with the longest sequences in the dictionary and tokenizing them first. Employing non-overlapping tokens, as opposed to the overlapping k-mers used in earlier studies, allows for the analysis of more extended sequence fragments while maintaining the same model input size. To put it in perspective, 512 overlapping 6-mers represent 512 bp, but 512 non-overlapping BPE tokens can represent ∼4.5 kb. This is a crucial factor when dealing with expansive and intricate genomes like that of humans. Nevertheless, it is worth noting that the model's granularity is confined to the resolution of these individual tokens, which might pose constraints for certain applications.

Our second enhancement pertains to the diversification in the implementation of the attention mechanism. The foundational GENA models utilize a conventional attention mechanism, which empowers the model to discern relationships between every pair of tokens in the input sequences. Conversely, sparse GENA models incorporate a sparse attention mechanism. This approach extends the permissible length of the input sequence by constraining the overall number of connections. Nevertheless, it retains the capability to understand relationships between distant sequence elements. In the case of recurrent GENA models (see 'Handling even longer sequences with recurrent memory' section), the transformer is supplemented with memory capabilities. This modification facilitates the processing of even longer inputs by segmenting them.

Through the integration of BPE tokenization and the sparse attention mechanism, we are able to train models that can handle input sequences of ∼4.5 kb (512 tokens with full attention) and 36 kb (4096 tokens with sparse attention). The incorporation of recurrent memory further expands this capacity, allowing for the processing of input sequences spanning hundreds of thousands of base pairs.

For model training, we utilized the MLM task, a prevalent technique in NLP wherein the model predicts a masked token based on its surrounding sequence context. Unlike previous studies that used the hg38 genome assembly (15,21), we trained all our models using the more recent human T2T genome assembly, setting our experiment apart. To mitigate the risk of overfitting to the reference genome, we incorporated common variants from the 1000-genome project database into some of our models. Additionally, we enriched our training dataset with genomes from diverse species, encompassing standard model organisms such as mice, fruit flies, nematode worms and baker's yeast, as well as others covering the entire spectrum of eukaryotic taxa. For a detailed methodology, see 'Materials and methods' section.

Throughout the manuscript, we collectively refer to our suite of developed models as GENA-LMs. Each specific model is designated by its label as shown in Table 2. While each model has its unique merits and constraints, we wish to highlight the following:

(1) The *gena-lm-bert-base-t2t* model: This model emulates the BERT transformer architecture, serving as a benchmark for subsequent models.
(2) The *gena-lm-bert-base-t2t-yeast/fly/athaliana/multi* models: These models include multispecies or taxon-specific data during pre-training while using the same BERT architecture as a model described above.
(3) The *gena-lm-bert-large-t2t* model: With the most significant parameter count (336M) and an input capacity of 4.5 kb, it stands out in terms of complexity.
(4) The *gena-lm-bigbird-base-sparse-t2t* models: These models, although having fewer parameters than the *gena-lm-bert-large-t2t*, boast a more extended input sequence length of 36 kb.

Upon evaluating the performance of our models in the MLM task (Figure 1F), we observed that models with sparse attention slightly outperformed their full-attention counterparts limited to 512 tokens. This underscores the role of contextual information in the training regimen. Nonetheless, it is imperative to note that while achieving commendable scores in the MLM task is encouraging, it does not necessarily guarantee optimal translation of the learned DNA representations to downstream applications. Consequently, our study delves into the comprehensive assessment of GENA-LMs across a spectrum of biologically relevant tasks to explore their merits and constraints.

GENA-LM performance on different genomic tasks

To evaluate the foundational GENA-LM models, we selected a range of genomic challenges that have recently been addressed using artificial intelligence (Figure 1B). These challenges encompass (i) prediction of polyadenylation site strength; (ii) forecasting of chromatin profiles, which includes histone marks (HMs), DHS sites and TF binding sites, among others; (iii) identification of splicing sites. In addition, we employed a comprehensive set of 18 benchmarks recently developed by (39). The datasets for these tasks are derived from human genomic data. To explore the performance of models when applied to non-vertebrate species, we introduced challenges including (iv) determining the activity of housekeeping and developmental enhancers in a STARR-seq assay in Drosophila cells (36) and (v) estimation of DNA sequence promoter activity in humans, flies, yeasts and plants.

Prediction of chromatin profiles

The major feature of GENA-LMs is their ability to process long DNA sequences, ranging from 4 to 36 kb. Consequently, we benchmarked GENA-LMs on tasks where understanding long-range dependencies in DNA sequences is crucial for accurate prediction. In genomics, these long-range dependencies are particularly significant for various epigenetic features, which makes predicting a locus's epigenetic states based on its sequence a significant challenge. To assess the capabilities of the GENA-LM transformers in addressing this, we used DeepSEA dataset(37). This dataset encompasses over 900 cell-type-specific chromatin profiles, which are grouped into DHS sites, HMs and TF binding sites. In the foundational DeepSEA challenge, chromatin mark signals were predicted for each 200-bp genomic segment, informed by both its sequence and an additional 800-bp context derived from its flanking regions (±400 bp).

When deploying GENA-LMs for this challenge (see DeepSEA section of Table 3), we discovered that transformer models markedly surpassed the performance metrics previously achieved by the convolutional neural network, DeepSEA. Notably, for TF and DHS profiles, GENA-LMs delivered scores that eclipsed those reported for the BigBird architecture, even though BigBird utilized an expanded 8-kb context (leading GENA-LM average Receiver Operating Characteristic Area Under the Curve (ROC AUC) on a 1-kb context for TF: 96.81 ± 0.1 versus BigBird's 96.1; for DHS: 92.8 ± 0.03 versus BigBird's 92.3). Furthermore, the performance metrics for GENA-LMs were either on par with or exceeded those recently reported for the Nucleotide Transformer (39). They also proved superior to the outcomes of the DNABERT architecture when trained on 1-kb input lengths.

To ensure a more equitable comparison between the BigBird and GENA-LM architectures, we adapted the DeepSEA dataset to incorporate expanded context sequences. This adaptation allowed us to match the 8-kb input length characteristic of the BigBird architecture. Intriguingly, the augmented

Table 3. Comparative performance analysis of GENA-LM models across multiple genomic tasks and input sizes

	GENA-LM models						DNABERT models	
Sub-task\input size	**bert-base-t2t**	**bert-base-lastln-t2t**	**bert-base-multi-t2t**	**bert-large-t2t**	**bigbird-base-t2t**	**bigbird-base-sparse-t2t**	**DNABERT**	**DNABERT-2**
				DeepSEA chromatin profiling, ROC AUC				
DHS\1kb	92.15	89.13	92.87	90.73	92.04	92.70	86.03	
DHS\8kb				87.85	92.26	92.06		
HM\1kb	85.17	86.05	82.65	86.64	84.72	85.31	86.13	79.14
HM\8kb				88.18	89.71	89.69		
TF\1kb	95.69	95.98	92.83	96.54	94.96	96.81	96.37	88.53
TF\8kb				95.18	96.24	96.40		
				EPDnew promoter activity, *F*1				
promoter\0.3 kb	89.88	90.08	89.70	90.62	89.98	87.57	93.26	89.44
promoter\2 kb	93.41	93.62	93.11	94.16	93.42	93.85	94.28	92.98
promoter\16 kb					93.15	93.40		
				Splice site annotation, AP score				
splice site\15 kb	92.63	92.56	91.42	93.6	94.78	94.7		

This table encapsulates results for tasks including: DeepSEA (37) chromatin profile prediction (DHS, HMs and TF binding sites); promoter activity prediction based on EPDnew dataset; and splice sites annotation based on SpliceAI dataset. All values are average of at least three runs.

context had differential effects on the prediction of various epigenetic profiles. For HMs, a marked performance improvement was evident, with an AUC reaching 89.71 ± 0.08. This was notably superior to the shorter context's score of 86.64 ± 0.08, the original DeepSEA findings (85.6), and the BigBird's result (88.70). However, for TF and DHS predictions, the extension in input length yielded only marginal enhancements in performance.

We analyzed the AUC variations across individual HMs to pinpoint which epigenetic profiles were responsible for the observed performance enhancement. Remarkably, there was a distinct divergence between narrow and broad HMs. While the narrow marks demonstrated marginal AUC improvements, the broad marks exhibited a pronounced increase when the context length was extended (Supplementary Figure S1). These observations reinforce our prior observation (5) that broad HMs necessitate an expansive context for precise prediction. This highlights the importance of handling extended input lengths for such tasks.

The varying performance metrics of distinct GENA-LMs across diverse epigenetic profiles and context lengths underscore that no singular model universally excels across all challenges. For TFs, the *gena-lm-bigbird-base-sparse-t2t* stands out on 1-kb inputs, with performance diminishing marginally when the input size increases. In contrast, for DHS, the *gena-lm-bert-large-t2t* model, boasting the highest parameter count, emerges as the optimal choice. Surprisingly, extending the context for this model results in a notable performance dip. For HMs, the optimal approach hinges on processing extended contexts with the *gena-lm-bigbird-base-t2t* model. Collectively, GENA-LMs outstrip competing models like BigBird, DNABERT and Nucleotide Transformer, marking a new performance state of the art for this task.

Promoter activity prediction

Promoter activity is an essential characteristic of genomic sequences, allowing them to drive the expression of genes. Although the basal promoter is a relatively short genomic region of 300 bp, surrounding context can substantially modify promoter activity. We assessed whether our models can discriminate human promoter sequences using promoter instances from the EPD dataset and juxtaposed them against non-promoter control samples. We observed that when the input sequence length was extended from 300 bp to 2 kb, there was a significant improvement in performance, as shown in the EPDnew section of Table 3. With 300-bp sequences, the DNABERT architecture emerged superior, registering an *F*1 score of 93.26, compared with the top-performing GENA-LM's *F*1 score of 90.62. However, when evaluating 2-kb sequences, the GENA-LM performance matched DNABERT results, recording an *F*1 score of 94.28 ± 0.65 and 94.16 ± 0.19 for DNABERT and *gena-lm-bert-large-t2t*, respectively (no significant difference, Wilcoxon test *P*-value=0.0625). This result was markedly higher than the DNABERT-2 model's score, which, when fine-tuned for the same input sequence length, achieved an *F*1 score of 92.98 ± 0.25.

In assessing the performance of GENA-LMs for predicting promoter activity, we observed the following: (i) Models with a greater number of parameters outperformed those with fewer. For instance, the *gena-lm-bert-large-t2t* surpassed the *gena-lm-bert-base-t2t*. (ii) The ability to handle longer input sequences due to the sparse attention mechanism gave certain models an edge over traditional full-attention BERT models. As a result, the *gena-lm-bigbird-base-sparse* outperformed the *gena-lm-bert-base-t2t*. Interestingly, models with shorter inputs but more parameters, such as the *gena-lm-bert-large-t2t*, still had superior performance over the *gena-lm-bigbird-base-sparse*. (iii) Incorporating multispecies training by using genomic sequences beyond just human data during pre-training did not result in improved performance, as seen when comparing the *gena-lm-bert-base-t2t-multi* with the *gena-lm-bert-base-t2t*.

Splice site annotation

We further optimized GENA-LMs to predict splice-donor and splice-acceptor sites within the human genome (Splice section of Table 3). The task required analyzing large contexts: a 15-kb input comprised of a central 5-kb target flanked by 5-kb sequences on either end. Notably, the task-specific convolutional neural network, SpliceAI, marginally surpassed GENA-LMs, registering a mean PR (precision-recall) AUC of 0.960 compared with 0.947 ± 0.002 for GENA-LMs.

For this task, models designed for longer sequence inputs, such as *gena-lm-bigbird-base-t2t*, outperformed those tailored for shorter inputs, even if the latter were equipped with more parameters, as in *gena-lm-bert-large-t2t*. This aligns with our earlier findings, suggesting that extending contextual information could be more beneficial than merely increasing the number of parameters. Consistent with our promoter analysis, multispecies models, like *gena-lm-bert-base-t2t-multi* (mean PR AUC of 0.914), did not enhance performance compared with their single-species counterparts, such as *gena-lm-bert-base-t2t* (mean PR AUC of 0.926).

Benchmarking GENA-LMs on short sequence tasks

To compare GENA-LM with with several recently developed DNA language models, including Nucleotide Transformer (39), DNABERT (15), DNABERT-2 (17), HyenaDNA (40) and fine-tuned versions of Enformer (8), we adopted a recent series of 18 benchmarks (39), which include relatively short sequence inputs ranging from 300 to 600 bp. According to the results presented in the Table 4, *gena-lm-bert-large-t2t* outperformed all other models, achieving the highest average score and the second-best average ranked score. Notably, *gena-lm-bert-large-t2t* outperformed Nucletide Transformer in multiple tasks, despite having substantially less parameters (330M versus 2500M). Similarly, GENA-LMs demonstrated performance *on par* with these models in another series of benchmarks focused on prediction of the human polyadenylation sites and Drosophila enhancer activity, as detailed in Supplementary Note 1 (Supplementary Figures S9, S10 and S11, respectively).

Identifying functional genomic elements with GENA-LMs

Spotting motifs for TFs binding

Modern techniques for analyzing deep neural networks allow us to assess the contribution of each input element to a model's downstream task performance. Such analyses offer valuable insights into the underlying mechanisms of biological processes. Take the ChIP-seq technique, for instance, a prevalent method for chromatin profiling. Its resolution is ∼100–200 bp. However, recognition motifs for the majority of DNA-binding proteins are significantly shorter, typically between 4 and 10 bp. Consequently, deducing precise binding locations from ChIP-seq data is challenging and often necessitates supplementary experiments (48).

To ascertain if GENA can enhance the resolution of experimental ChIP-seq data, we employed token importance scoring (45) on the *bigbird-base-sparse-t2t* model, which was fine-tuned using the DeepSEA dataset. This methodology assigns a significance value to each token within the input, based on its relevance to the prediction outcome. Here, we concentrate on the binding of three TFs: ATF1, CTCF and GATA2 in human K562 cells. Each of these factors possesses well-established DNA recognition motifs (Figure 2A). This allows for a comparison between tokens important for GENA's

Table 4. GENA-LM (*gena-lm-bert-large-t2t*) has the highest average score over the wide range of genomic tasks compared with other foundational DNA models

	Model													
Task\dataset	GENA-bert-large-t2t	HyenaDNA-1KB	HyenaDNA-32KB	DNABERT-1	DNABERT-2	Enformer	NT-HumanRef (500M)	NT-1000G (500M)	NT-1000G (2.5B)	NT-Multispec (2.5B)	NT-Multispec-v2 (500M)	NT-Multispec-v2 (250M)	NT-Multispec-v2 (100M)	NT-Multispec-v2 (50M)
H3	0.79	0.78	0.74	0.76	0.79	0.72	0.72	0.74	0.76	0.79	0.78	0.79	0.79	0.79
H3K14ac	0.60	0.61	0.4	0.4	0.52	0.29	0.37	0.38	0.45	0.54	0.55	0.54	0.52	0.51
H3K36me3	0.61	0.61	0.48	0.47	0.59	0.34	0.45	0.47	0.53	0.62	0.63	0.62	0.59	0.58
H3K4me1	0.53	0.51	0.38	0.4	0.51	0.29	0.37	0.38	0.42	0.54	0.55	0.54	0.52	0.52
H3K4me2	0.46	0.46	0.28	0.28	0.34	0.21	0.26	0.26	0.28	0.32	0.32	0.32	0.3	0.3
H3K4me3	0.55	0.55	0.29	0.26	0.35	0.16	0.24	0.24	0.31	0.41	0.41	0.4	0.38	0.33
H3K79me3	0.67	0.67	0.57	0.58	0.61	0.5	0.56	0.56	0.57	0.62	0.63	0.62	0.61	0.6
H3K9ac	0.61	0.58	0.47	0.5	0.54	0.42	0.45	0.48	0.49	0.55	0.56	0.55	0.54	0.52
H4	0.78	0.76	0.76	0.79	0.8	0.73	0.75	0.76	0.79	0.81	0.8	0.8	0.79	0.8
H4ac	0.59	0.56	0.38	0.36	0.46	0.27	0.33	0.34	0.41	0.49	0.5	0.5	0.48	0.46
Enhancers	0.55	0.52	0.49	0.5	0.52	0.45	0.5	0.51	0.54	0.55	0.55	0.54	0.5	0.52
Enhancers (types)	0.45	0.39	0.36	0.37	0.42	0.31	0.43	0.4	0.44	0.45	0.42	0.45	0.41	0.4
Promoters	0.94	0.92	0.91	0.92	0.94	0.91	0.9	0.9	0.93	0.95	0.95	0.95	0.94	0.92
Promoters (TATA)	0.91	0.88	0.87	0.91	0.91	0.92	0.89	0.87	0.91	0.92	0.93	0.92	0.91	0.89
Promoters (non-TATA)	0.94	0.92	0.91	0.92	0.94	0.91	0.91	0.9	0.93	0.95	0.95	0.95	0.94	0.92
Splicing (both)	0.91	0.94	0.94	0.96	0.91	0.77	0.96	0.95	0.96	0.97	0.97	0.97	0.97	0.96
Splicing (acceptors)	0.92	0.92	0.92		0.95	0.83	0.93	0.93	0.96	0.97	0.96	0.96	0.96	0.95
Splicing (donors)	0.91	0.9	0.92		0.93	0.81	0.94	0.94	0.96	0.97	0.97	0.97	0.96	0.95
Average	0.707	0.693	0.615	0.586	0.668	0.547	0.609	0.612	0.647	0.690	0.691	0.688	0.673	0.662

Datasets and performance of alternative models for 18 tasks are from (39). All values are MCCs. Scores for all models except for the GENA-LM model were taken directly from the benchmark: https://huggingface.co/spaces/hlnicholls/nucleotide_transformer_benchmark.

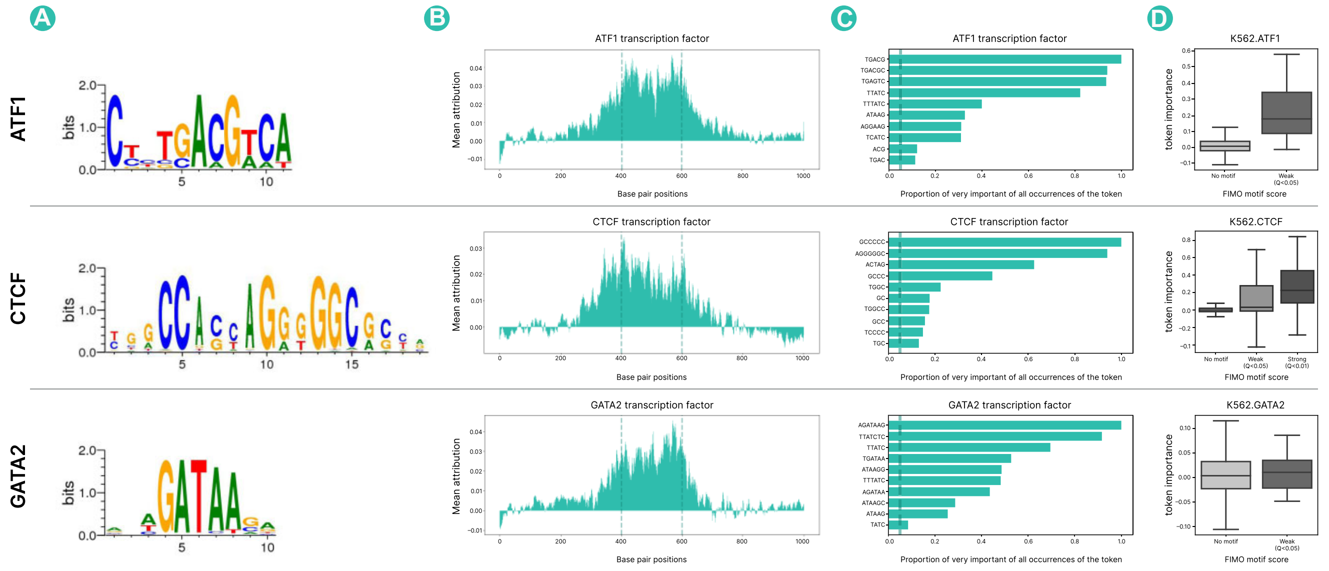


Figure 2. GENA-LM identifies DNA motifs essential for TF binding. In panels (A)–(D), each row pertains to a distinct factor, labeled to the left. (**A**) Logo representation of motifs for the three TFs considered in our analysis. (**B**) Profile of average token importance scores over the sequence length. Vertical dashed lines demarcate the 200-bp prediction region. (**C**) Bars represent the frequency of token occurrences in the 'highly important' category (tokens with scores in the top 5th percentile). The *X*-axis shows the proportion of these occurrences relative to all occurrences for that token. A vertical reference line marks the 0.05 fraction threshold; only tokens exceeding this fraction are displayed. (**D**) Boxplots detail the distribution of importance scores for tokens, categorized by different FIMO q-values. They display the median, interquartile range as well as the 5th and 95th percentiles.

predictions and the recognized sequence determinants associated with TF binding.

First, we examined the distribution of importance scores across the sequence length, as depicted in Figure 2B. It should be noted that during the fine-tuning process, the input consists of the DNA sequence from the 200-bp target region (where TF binding is anticipated) accompanied by an 800-bp contextual sequence. Our analysis, presented in Figure 2B, reveals a consistent pattern: the importance attributed to a token diminishes as its distance from the target region increases, a trend observed for all three TFs.

We subsequently sought to discern which sequences garnered high token importance scores. Tokens exceeding the 95th percentile of the importance score distribution were designated as 'highly important.' We then compiled tokens that consistently featured on this 'highly important' list. Upon visual examination (Figure 2C), we observed that these tokens frequently encompassed full or fragmented motifs of the target TFs. For instance, ATF1, which has a core motif of TGACG, prominently displayed a token matching this exact sequence among its highly important tokens. In the case of the GATA2 factor (core motif: GATAA), the token AGATAAG, incorporating the GATA2 motif, was most prevalent among the highly important tokens. As for CTCF, which boasts a motif more intricate and extended than its counterparts, the most recurrent highly important tokens primarily featured GC-rich subsequences of the motif.

To more comprehensively assess the congruence between known TF motifs and 'highly important' tokens, we employed the FIMO tool to annotate all DNA samples. FIMO is a bioinformatics software designed to identify specific motifs by leveraging the motif's position weight matrix (PWM). As depicted in Supplementary Figure S2, there is a discernible overlap between significant tokens and motifs detected by FIMO. Our statistical evaluation establishes a relationship between FIMO motif scores and token importance scores. Both robust motifs (FIMO q-value < 0.01) and more tenuous motifs (0.01 < FIMO q-value < 0.05) manifest markedly elevated token importance scores compared with tokens absent of any discerned motif (FIMO q-value > 0.01) (Figure 2D). It is worth noting, in the context of the GATA2 TF which is characterized by a shorter motif length, sequences with high FIMO q-values are absent. Nonetheless, we observed that the majority of the 'highly important' tokens encompass the core GATA2 motif, as delineated in Supplementary Figure S3.

While the results affirm that token importance mirrors the presence of established motifs for DNA-binding TFs, the congruence between FIMO-detected motifs and tokens with high scores is not absolute. This observation prompted us to delve into the nature of motifs encompassed by tokens vital for GENA model prediction, yet devoid of the target TF's annotated motif as per FIMO. Utilizing the *de novo* motif discovery tool XSTREME, we analyzed a subset of important tokens lacking a canonical motif (with FIMO target factor motif q-value > 0.05) and examined the enriched motifs therein. Intriguingly, for both CTCF and GATA2, XSTREME predominantly identified their respective motifs. In the case of important ATF1 tokens sans ATF1 motif, the secondary most abundant motif discerned belonged to the ATF family. This suggests that the rudimentary PWM statistics employed by FIMO might overlook biologically pertinent motif variants that diverge notably from the consensus represented by the PWM. Conversely, GENA-LM exhibits a promising potential in recognizing these variant motifs. When consolidated during XSTREME analysis, these diverse motif representations converge to echo the canonical motif's PWM. Moreover, our analysis revealed a significant presence of GATA2 motifs within tokens deemed essential for the ATF1 factor, hinting at a possible functional synergy between these TFs in K562 cells—a nuance discerned by GENA-LM. Given that ATF1 is an integral part of the AP-1 complex, our findings resonate with, and potentially elucidate, prior experimental data evidencing cooperation between GATA2 and the AP-1 complex (49).

Searching for DNA sequence determinants of hisone modifications

Having ascertained GENA's capacity at pinpointing DNA motifs crucial for TF binding, we turned our attention to discerning sequence determinants associated with HMs, which currently lack identifiable binding motifs. Our focus centered on H3K4me1, H3K9me3 and H3K27me3. These factors represent HMs with distinct and well-documented functional implications: H3K4me1 delineates active genomic regions, H3K9me3 signifies heterochromatin and H3K27me3 demarcates the suppressed 'facultative heterochromatin' including genes under developmental regulation, bound by polycomb group proteins. In our examination of these factors, we evaluated the distribution of importance scores relative to sequence length and accumulated tokens that frequently exhibited high importance values.

As depicted in Supplementary Figure S4, the distribution of token importance scores across sequence lengths for these epigenetic markers mirrors that of the previously discussed factors. Notably, specific tokens consistently emerged as highly significant in predicting these HMs, reminiscent of the motif-rich tokens previously identified as important for predicting associated TFs (see Supplementary Figure S5).

Prompted by our observations, we sought to determine if motifs enriched among the tokens with high importance scores were shared across these three factors. Extending the sequences of our selected highly important tokens by 4 bp, we then undertook a rigorous motif analysis using XSTREME. Our examination revealed several motifs of significance (see Supplementary Table S1), each aligning with known TFs. For the active promoter mark, H3K4me1, the significant tokens were found to encompass motifs corresponding to the GATA, JUN and FOSL TFs. These findings align with the documented roles of these factors in regulating transcription and influencing the oncogenic transformation of K562 cells (50,51). In the context of the H3K27me3 mark, which signifies facultative heterochromatin and designates functional elements repressed in specific cell lineages, our data from hematopoietic K562 cells indicated an enrichment of TF motifs not typically associated with blood cells. Examples include SNAI2, pivotal in epidermal cell differentiation, and ASCLI, a critical regulator of neurogenesis. This suggests that within this setting, GENA discerned genomic motifs that might be activated in alternative cell types but are designated for repression in the K562 lineage. Lastly, for the H3K9me3 mark indicative of constitutive heterochromatin, our analysis highlighted an enrichment of the ZNF274 TF motif. This is in agreement with its established role as a transcriptional repressor.

In summary, our findings indicate that beyond predicting the epigenetic profiles of a specific locus, GENA models can effectively identify the distinct subsequences that drive observed epigenetic signals. Such analysis holds potential to substantially augment the resolution of prevailing experimental approaches, like ChIP-seq, and to pinpoint TFs linked to particular HMs.

Evaluation of clinical relevance of mutations

The capability of GENA-LMs to accurately predict promoter activity inspired us to investigate whether model predictions could functionally characterize variants in human promoters. We compiled ClinVar mutations that overlap with promoter sequences and evaluated them using the odds ratio of promoter presence probability for wild-type versus mutated sequences. While not all pathogenic ClinVar variants overlapping promoters impact predicted promoter activity, a pathogenic-versus-benign classifier based on *gena-lm-bert-large-t2t* predictions achieved an AUC of 0.66 and an average precision (AP) of 0.59 (Supplementary Figure S6). Further, we applied the Integrated Gradients method (45) to identify input sequences tokens that are the most important for predicting promoter presence. Our analysis revealed that pathogenic mutations are enriched ∼2.5 times in the top percentile of the most important tokens (P-value $< 1e{-}15$, Supplementary Table S2). Comparable results were obtained with the *gena-lm-bigbird-base-sparse-t2t* model.

Similar to promoters, mutations at splice sites often have clinical implications in human diseases. To determine whether GENA-LMs can detect the impact of single-nucleotide perturbations at splice donor and acceptor sites, we conducted comprehensive *in silico* mutagenesis, evaluating the effects of every possible single-nucleotide substitution within a ±20-bp sequence surrounding splice sites. Despite the token-level resolution of the inputs, predictions from *gena-lm-bigbird-base-t2t* proved sensitive to single-nucleotide substitutions. We observed that the model distinctly identifies substitutions within canonical splice site motifs from other single-nucleotide variants (Supplementary Figure S7). While the latter rarely alter the model's prediction, mutations within canonical splice sites almost invariably abolish the predicted acceptor or donor site. Notably, the predicted effects of single-nucleotide substitutions align precisely with known splice site motifs: changes in the constant motif positions have more significant effects compared with the substitutions in the more variable regions of the motif. These findings highlight the potential of GENA-LMs for clinical interpretation of human genetic variants.

GENA-LMs beyond human species

Epigenetic annotation of non-human genomes using GENA-LMs

While the results discussed previously demonstrate the high efficacy of fine-tuned GENA-LM models in reconstructing and analyzing genomic features such as promoters or protein binding sites, task-specific training of these models necessitates the availability of experimentally measured data, which is often lacking for non-model species. We hypothesize that features relatively conserved across different species and cell types could be effectively predicted by a model that has been fine-tuned using human (or other reference) data. To explore this hypothesis, we collected experimentally measured profiles of insulatory protein CTCF binding sites, H3K27 acetylation (H3K27ac) HMs, and promoter activity across various animal species.

We initiated our study by evaluating a human-based promoter activity prediction model on data from seven species: macaque, mouse, rat, dog, zebrafish, chicken and *C. elegans*. For mammals (macaque, mouse, rat and dog), the model's performance closely mirrored that observed with human promoters, achieving an $F1$ score of ∼0.95, despite the absence of species-specific data during the model's fine-tuning (Figure 3A). The evaluation score decreased by 10 pt when applying the same model to phylogenetically more distant vertebrate species such as chicken and zebrafish, but it still achieved

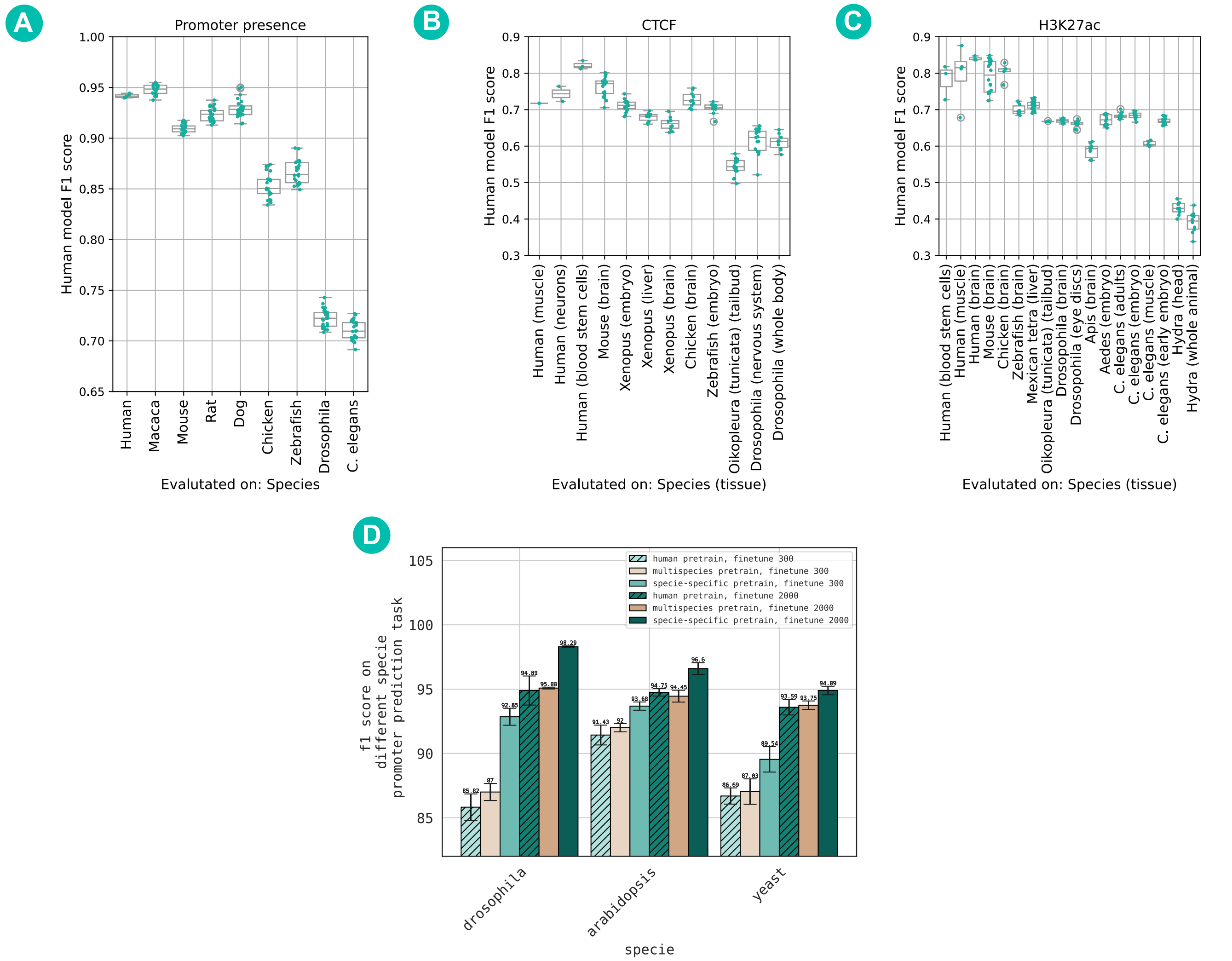


Figure 3. GENA-LMs demonstrate generalization across species. (A–C) GENA-LM fine-tuned on human promoters (**A**), CTCF (**B**) or H3K27 (**C**) binding sites evaluated on different species. (**D**) Effect of multispecies versus species-specific pertaining on promoter activity prediction.

a relatively high *F*1 score of around 0.85. However, for more distantly related species like the fruit fly or the flatworm *C. elegans*, the model's performance dropped significantly, resulting in an *F*1 score of ∼0.7. These results suggest that closely related mammals share similar promoter grammars, which can be captured by training on a human dataset, while more distant species exhibit gradual modifications in the DNA encoding of their promoters.

Then *gena-lm-base-t2t* was trained to differentiate between genomic sites with H3K27ac and CTCF binding sites and those without binding on human data, achieving *F*1 scores between 0.7 and 0.8 on held-out human sequences (Figure 3B and C). Applying this model to other species revealed that its performance correlates with the evolutionary distance between the target species and humans. For CTCF binding prediction, the model demonstrated similar performance in closely related species, such as mice, with a gradual decrease in other vertebrates and a more significant drop in invertebrate species such as Drosophila (Figure 3B and C).

While similar cross-species analyses for CTCF can be performed using a known motif, this is infeasible for H3K27ac due to the lack of any recognized motif. When evaluating the human H3K27ac model across species, we observed a substantial performance decline outside tetrapod species, and for evolutionarily distant species like Hydra *F*1 score drops to ∼0.4. The decline in model performance across species may serve as a measure of the interspecies differences in H3K27ac encoding. Thus, the use of GENA-LMs enables explorations into the evolution of sequence grammar. Furthermore, these results suggest that GENA-LMs can be used to infer epigenetic marks from genomic sequences in the absence of experimental data. Although the performance of such inferences is inferior to that of species-and cell type-specific models and may be limited to taxonomically close groups, it presents extensive opportunities for annotating available genomes, such as using the human model to annotate H3K27ac across several thousand available mammalian genomes.

Species-specific pre-training improves model quality

Based on the aforementioned experiments, it is clear that promoter grammar varies significantly across evolutionarily distant species such as humans, flies, yeasts and plants (52). As a result, learning species-specific information during pre-training could prove beneficial for accurate promoter presence prediction and other species-specific genomic tasks. However, pre-training species-specific GENA-LMs requires sufficient amount of data and computational resources. Thus, it is important to know whether transfer learning, when a model pre-trained on one data distribution is reused for tasks in similar domain, can substantially improve performance. Another viable strategy in this case involves generation of a universal foundational model trained on the mixture of species.

To study generalization capabilities of GENA-LM models, we utilized EPD promoter annotations for yeast, fly and Arabidopsis species to fine-tune universal, multispecies or species-specific models. We discovered that for all these species, *gena-lm-bert-base-lastn-t2t* pre-trained on human data can be effectively fine-tuned to provide reasonably accurate promoter classification (Figure 3D). When comparing datasets with varying promoter lengths (300 bp versus 2 kb), we observed a significant impact of contextual information on prediction accuracy, mirroring the dependency noted in human data. Compared with the model pre-trined on human data, multispecies pre-training proved to be benefitial: in five out of six experiments, we observed performance improvements ranging from 0.2 to 1.5 points when fine-tuning *gena-lm-bert-base-t2t-multi* model. It is important to note that the *gena-lm-bert-base-t2t-multi* training dataset includes dozens of genomes; hence, the amount of data from each species available during the pre-training phase was limited. Consequently, we pre-trained new taxon-specific models for yeast, flies and Arabidopsis. Fine-tuning these models to predict promoter activity in their respective taxa resulted in performance improvements of 2–7 points compared with the model pre-trained on human data (Figure 3D). Therefore, we conclude that taxon-specific models can significantly enhance DNA annotations by learning the unique DNA grammar of each species during pre-training. We have publicly released these taxon-specific models to facilitate further applications within the selected species.

Species classification based on embeddings from GENA-LMs

The universality of GENA-LMs in addressing various biological challenges suggests that the DNA embeddings of the pre-trained model encapsulate significant biological insights. Evolutionary distant species are known to exhibit divergence in their regulatory code and codon usage patterns. If GENA-LMs effectively capture these inherent biological characteristics during pre-training, one would expect that their embeddings could differentiate DNA sequences sourced from varying species without any additional fine-tuning. To evaluate this premise, we curated a set of 27 species spanning diverse taxonomic classifications, ranging from bacteria to humans (refer to Supplementary Table S3). These species also represent a broad spectrum of evolutionary divergence times, spanning from millions to over a billion years (as depicted in Figure 4A and B). We then examined the embeddings generated by inputting genomic DNA subsequences into pre-trained GENA-LMs. Our investigations encompassed a range of sequence lengths, beginning with the typical length of shotgun sequencing reads (100 bp) and culminating at 30 kb, a size consistent with reads from third-generation sequencing platforms.

Initially, we employed tSNE to project sequence embeddings derived from all genomes into a 2D space. This visualization reveals discernible clusters that mirror the phylogenetic relations between the species. Notably, distinct groupings emerged for bacteria, plants and yeasts, each isolated from the clusters representing animal genomes. Within the realm of animals, we could discriminate invertebrate species and various vertebrate classes. This evidence underscores that GENA-LM embeddings encapsulate nuances allowing for the differentiation of species based on their genomic sequences.

To deeply study these capabilities, we employed a Gradient Boosting algorithm for each of the 27 species pairs. Our aim was to achieve binary species classification leveraging the sequence embeddings.

The data in Figure 4C show the richness of information contained in GENA embeddings, enabling species differentiation based on their genomic DNA subsequences. The accuracy of classification is influenced by both the divergence time and the length of the input sequence, with the latter exerting a more pronounced effect. For species that are closely related (with divergence times ≤ 20 MYA), classification accuracy remains constrained ($F1$ score ≈ 0.7). However, for species diverging around 60–100 MYA—equivalent to the evolutionary separation among all mammalian species—employing the model that accepts longer sequence inputs boosts our classification capability, yielding an $F1$ score exceeding 0.8. Remarkably, for extensive divergence times ($\geq$200 MYA, reflecting the era of the last common ancestor of vertebrates), the classification's precision approaches perfection.

We next evaluated the classification efficacy of sequence embeddings derived from various layers and architectures of GENA-LMs. Across all models, embeddings sourced from the initial layers consistently delivered subpar performance. This performance incrementally improved, peaking around layers 9 to 12. Notably, for both *gena-lm-bert-large-t2t* (comprising 24 layers) and *gena-lm-bigbird-base-t2t* (with 12 layers), a minor performance dip was observed when utilizing embeddings from the final layers. This trend resonates with prior studies in NLP (53) and protein modeling (54). Such studies have posited that in transformer-based language models, the terminal layer embeddings encapsulate information tailored to the specific model training task. In contrast, embeddings from intermediary layers are more versatile, proving advantageous in knowledge transfer for tasks not explicitly addressed during the pre-raining phase. The depth of the layer is also indicative of the abstraction degree of the representations. While preliminary layers prioritize local level representations, the advanced layers capture intricate global features, such as binding sites and contact maps (54).

Collectively, our findings demonstrate that the sequence embeddings from pre-trained GENA-LMs encapsulate abundant biological insights, enabling the resolution of genomic challenges without the necessity for fine-tuning.

GENA-LM-based web service

Given the demonstrated potential of GENA-LMs in various genomic tasks, we aim to extend their accessibility through GENALM-Web, a web service designed for sequence annotation using DNA language models (Figure 1C). GENALM-Web incorporates several downstream tasks developed in this paper, such as promoter activity prediction, chromatin annotation, splice site inference and enhancer activity prediction for Drosophila sequences. Key features of the web service include the capability to handle exceptionally long inputs (up to 1 Mb), utilize extensive contextual information and conduct token importance analysis in real time. This last feature allows users to identify sequence regions responsible for specific features, even if these regions are located at distant genomic locations. The web service is accessible at

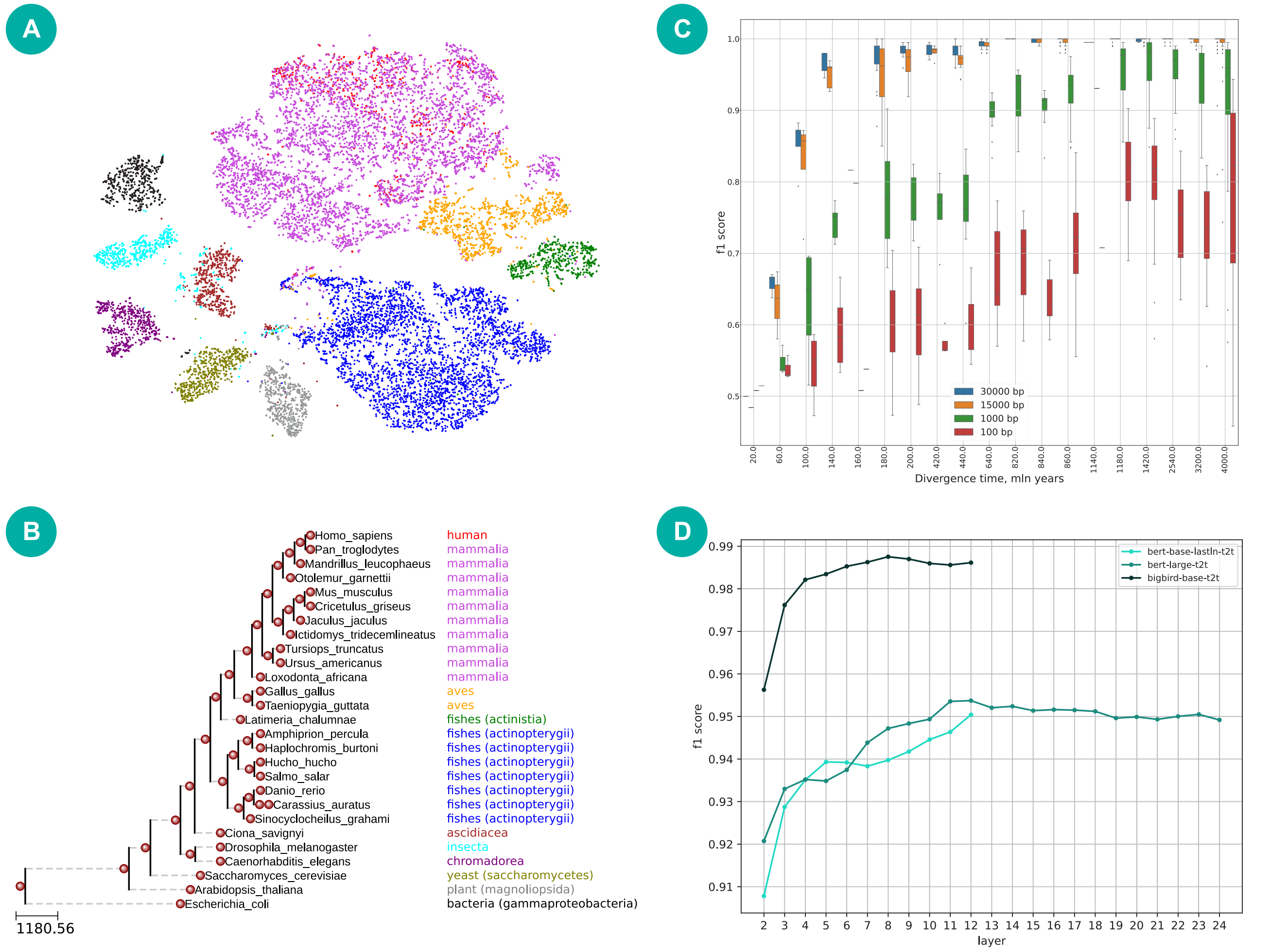


Figure 4. Sequence embeddings from pre-trained GENA-LMs facilitate species classification. t-Distributed Stochastic Neighbor Embedding (tSNE) projections (**A**) of sequences sampled from 27 species (**B**), representing a spectrum across the tree of life. (**C**) Classification performance for different sequence lengths plotted against divergence time. (**D**) Classification performance of embeddings taken from different layers of three models. Data are presented for sequence lengths of 5 kbp (for 'gena-lm-bert-base-lastln-t2t' and 'gena-lm-bert-large-t2t') and 30 kbp (for 'gena-lm-bigbird-base-t2t').

https://www.dnalm.airi.net, and documentation is available at service pages and in the supporting manuscript (55).

Handling even longer sequences with recurrent memory

While the integration of sparse attention techniques and BPE tokenization in GENA-LMs has substantially expanded the permissible DNA input length, the current limit (about 36 kb) may not sufficiently capture certain biological dependencies. Notably, the prediction of chromatin interactions (7), enhancer–promoter associations (4), gene expression (8) and other genomic phenomena necessitate the processing of contexts that extend beyond 30 kb. Additionally, our empirical analyses show improvements in promoter and splice site predictions as the context size expands from 512 to 4096 tokens (see 'GENA-LM performance on different genomic tasks' section). This indicates the potential benefits of further enhancing sequence length for these biological tasks.

To enhance the input capacity of GENA-LMs, we incorporated recurrent memory mechanisms. The RMT has been demonstrated as an efficient, plug-and-play method to handle extended input sequences using pre-trained transformer models (30). In this recurrent strategy, the input sequence is partitioned into segments which are processed one after the other (Figure 5A). Special memory tokens are introduced to each segment to pass information between consecutive segments, allowing them to use information from all previous segments. Thus, the entire pre-trained transformer effectively functions as a single recurrent unit.

RMT can be optionally incorporated during pre-training, enabling the model to learn the use of memory tokens at this stage. Alternatively, memory tokens can be introduced during the fine-tuning phase, using a model that was pre-trained without RMT. To assess these two training strategies and determine the optimal approach, we conducted pre-training experiments with *gena-lm-rmt-base-t2t* and *gena-lm-rmt-large-t2t* models. During pre-training, we evaluated how RMT augmentation influences MLM accuracy. As illustrated in Supplementary Figure S8, RMT augmentation enhances accuracy for the base-size model when provided with 8–10 segments of contextual information. However, the large-size model pre-trained without RMT augmentation achieves a substantially better score. Extending the large-size model's input with RMT does not improve the score (Supplementary Figure S8). These findings corroborate our previous

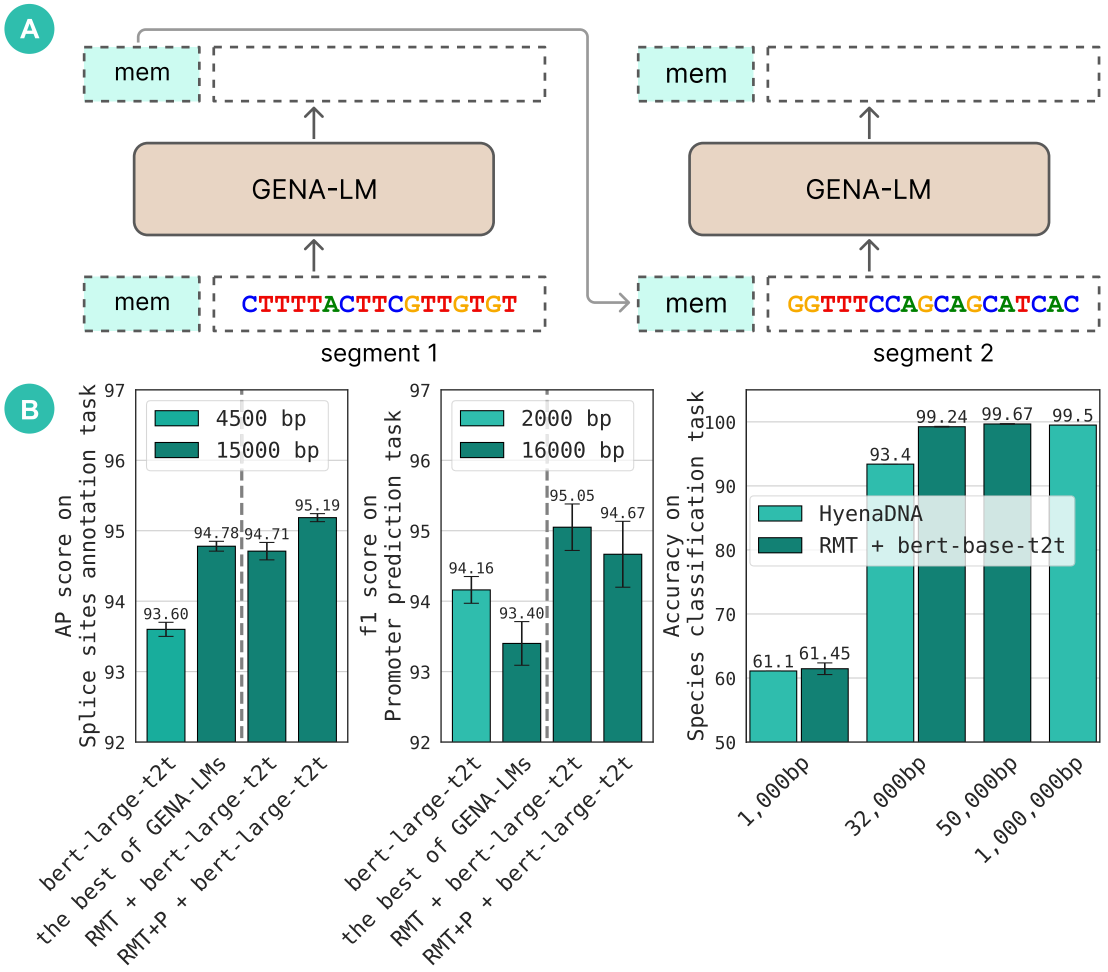


Figure 5. Leveraging recurrent memory to enhance the input capacity of GENA-LM models yields improved performance in downstream tasks. (**A**) The RMT architecture. A vocabulary of the model is augmented with a memory token denoted as 'mem' in the figure. Memory augmented model is fine-tuned to write relevant information in memory tokens and pass it to subsequent segments. (**B**) The augmentation of GENA-LM with RMT with 3× (left), 8× (center) and 50× (right) larger sequence lengths. Models with memory achieve superior results in splice site annotation and promoter prediction tasks when compared with all other GENA-LMs, including those utilizing sparse attention (Wilcoxon test *P*-value ≤0.043 in all comparisons). On the species classification task, RMT with GENA-LM outperforms the HyenaDNA model designed for long sequences. RMT+P refers to models that have not only been fine-tuned with RMT, but also pre-trained with it.

observation that *gena-lm-bert-large-t2t* outperforms *gena-lm-bigbird-base-sparse* in the MLM task Figure 1F, despite the latter's longer input length. Therefore, we propose that long-range dependencies beyond 4 kb are not necessary for accurate masked token prediction. However, these dependencies are important for downstream tasks.

For a comparative evaluation between RMT and other GENA models, we focused on tasks with inputs of moderate length (15–16 kb), which can be processed by sparse models. The *gena-lm-bert-large-t2t* model, when integrated with RMT, was fine-tuned on sequences of 16 kb for promoter prediction and 15 kb for splice site prediction. Inputs were divided into segments, with each segment comprising ∼512 tokens or about 4.5 kb. These segments also included memory tokens as part of the input, with 10 memory tokens used for each task. When contrasted with the original *gena-lm-bert-large-t2t* model, the sequence length processed by the *gena-lm-bert-large-t2t* + RMT increased substantially: from three to eight times (rising from 4.5 to 15 kb for the splice site prediction and from 2 to 16 kb for other tasks).

The expansion in input length significantly enhanced the performance of the *gena-lm-bert-large-t2t* model, as depicted in Figure 5B. Notably, models employing the RMT outperformed all other GENA-LMs, including those sparse variants of GENA-LM. While these sparse models can accommodate the input lengths featured in the aforementioned tasks, they have fewer parameters compared with *gena-lm-bert-large-t2t*. Thus, RMT allows combining models with the higher number of parameters and longer sequence inputs, achieving the best performance on the common biological tasks. Furthermore, RMT has no limit in a sequence length and could be used for even longer sequences. Sparse GENA-LMs, on the other hand, are limited to the lengths on which they were trained.

We next used splice sites and promoter activity prediction tasks to benchmark the effects of the RMT application during pre-training. This benchmark yields mixed results: for promoter activity prediction, limiting the RMT approach to the fine-tuning stage does not diminish performance compared with its use at both the pre-training and fine-tuning stages. However, for splice site annotation, pre-training with RMT proved beneficial, improving performance by ∼0.5 points.

Recently, the HyenaDNA team introduced a specific benchmark for DNA language models that process long input sequences (40). The authors demonstrated that HyenaDNA can classify five mammalian species (human, mouse, lemur, pig and hippo) based on their genomic sequences. Compared with the classification of a broader range of species using GENA-LM embeddings, as previously described, the species in this benchmark are phylogenetically closer, making classification more challenging. Original research (40) illustrated that classification accuracy is heavily dependent on the input DNA length, which increases gradually from 61.1 to 99.84 as sequence length scales from 1 kb to 1 Mb.

We compared HyenaDNA results with *gena-lm-bert-base-t2t* augmented with RMT. For sequence inputs of 1 kb, the classification accuracy of both models was relatively low (61.45 ± 0.91 for GENA and 61.1 for HyenaDNA). A significant enhancement in classification accuracy was observed when the sequence length was increased to 32 kb, with *rmt+gena-lm-bert-base-t2t* achieving 99.24 ± 0.06, thereby surpassing the performance of HyenaDNA (93.4), as shown in Figure 5 (right panel). Further extending the sequence length to 50 kb elevated the classification accuracy of *rmt+gena-lm-bert-base-t2t* to 99.67 ± 0.059, exceeding the accuracy HyenaDNA attained with 1000 kb sequences. This indicates that, within this experimental framework, RMT and the associated model architecture extract and leverage information from extended DNA sequences more efficiently than other technologies designed for processing long input sequences such as Hyena layers underlying HyenaDNA.

Discussion

Transformer architectures have garnered significant interest across diverse research domains, including genomics. They consistently achieve exemplary results in various biological tasks such as deciphering gene expression regulation in mammals (8) and *Escherichia coli* (56), predicting phenotypes from gene expression (57,58), deducing DNA methylation (59) and filling in missing genotypes (60), to name a few. Nevertheless, the challenge lies in training task-specific models for each distinct biological question. This process demands substantial time and resources. DNABERT, DNABERT-2 and similar foundational DNA models such as BigBird and Nucleotide Transformer provide a solution by offering a platform for refining universally applicable models without starting from scratch. The Nucleotide Transformer v2 (39) has incorporated rotary embeddings and gated linear units paired with swish activations, distinguishing it from its predecessor. This model has an input size of 12 kb. DNABERT v2 (17), while drawing upon the foundational DNABERT architecture, expands in terms of model parameters and employs BPE tokenization. However, its sequence input length remains below 4000 bp. Contrarily, HyenaDNA (40) introduces a novel architecture capable of handling vast DNA sequences, extending up to 1 million base pairs. Yet, benchmark results suggest an inverse relationship between HyenaDNA's performance and the input size used during its training, as noted by (39). Moreover, our benchmarking GENA-LMs augmented with RMT against HyenaDNA in species classification task indicate better performance of the former model.

A unique feature of HyenaDNA is its decoder-only configuration. Unlike the encoder-centric GENA-LMs, HyenaDNA does not generate sequence embeddings directly. Instead, it produces DNA sequences, making the derivation of class labels (for classification purposes) or quantitative targets (for regression) from its outputs a complex task. To predict specific DNA states with the HyenaDNA model, the authors utilized a DNA-alphabet encoding, obliging the model to understand this biologically unrelated nucleotide sequence interpretation.

We introduce GENA-LMs, a collection of open-source models boasting the most extensive input capacity among all accessible DNA transformers (for models starting with the `gena-lm-` prefix, visit https://huggingface.co/AIRI-Institute/) (10). The GENA-LM collection encompasses a spectrum of publicly accessible architectures, catering to researchers by offering tailored solutions for unique challenges. Moreover, GENA-LMs include several taxon-specific models that can improve performance in species-specific setups. Our rigorous benchmarking affirms that GENA-LMs not only surpass earlier pre-trained models but occasionally even rival the precision of task-specific convolutional neural networks.

In our comparison of various GENA-LMs, we investigated the influence of context length and the total number of model parameters on predictive accuracy. We found that the optimal balance between these two factors varies depending on the specific task. For instance, an extended context is vital for predicting promoter activity or deciphering widespread HM distributions, as previously indicated by (5). However, for certain tasks, a more concise context is adequate, making it more advantageous to augment the model's parameter count. The broad spectrum of GENA-LMs available offers researchers the flexibility to select a model best suited for their particular objective.

While GENA-LMs accommodate extensive input sizes, they occasionally fall short of the lengths required for peak accuracy in specific biological tasks. For example, research has shown that gene expression can be influenced by variants situated hundreds or even millions of base pairs distant from the promoter. This can be attributed to processes such as loop extrusion (61) and other 3D-genomic mechanisms (62). There are several strategies to address this constraint in GENA.

First, the RMT technique facilitates the processing of extensive sequence inputs using powerful models with a large number of parameters. Our benchmarks reveal that this approach delivers superior results for tasks where the biological signal spans a lengthy context. Notably, unlike transformer layers that exhibit a quadratic memory dependence on the number of tokens, the computational resources needed for RMT training and inference scale linearly with sequence length. RMT can be integrated with GENA-LMs not only during the fine-tuning phase of downstream tasks but also throughout the MLM pretraining stage. This could enhance learning operations on extended sequences, particularly for downstream task datasets that are of limited size. Furthermore, RMT models pre-trained on multiple segments can be utilized for a greater number of segments during inference (30). As such, RMT models are versatile enough to address a variety of downstream tasks, even for teams without access to cutting-edge computational infrastructure.

Second, the 3D proximity of chromatin can be determined using specialized models (7). This information can then be directly incorporated into transformer models, enabling them to capture long-range associations between functional genomic elements.

One limitation of GENA-LMs arises from the granularity imposed by the use of BPE tokenization, which confines predictions to specific tokens. To overcome this, exploring alternative DNA tokenization methods and developing low-level nucleotide embeddings could offer solutions for certain applications.

Beyond merely predicting specific biological signals, we demonstrate that GENA-LMs can also be harnessed to decipher and understand the functions of sequences underpinning these signals. An analysis of token importance revealed that GENA-LM accurately detected motifs corresponding to known TFs. Furthermore, it pinpointed TF binding sites crucial for specific HMs. There exists an array of factors that modify histones, termed histone 'writers', many of which are cell-type specific. Determining these factors and their corresponding genomic binding sites is a complex endeavor. In this context, we illustrate how GENA-LMs can aid in this task by discerning motifs deemed 'essential' for a specific HM within a particular cell type. However, it is pivotal to approach this method judiciously. The presence of enriched motifs may only indicate an association rather than a direct causal relationship. As an instance, while the enrichment of recognized activator factor motifs within H3K4me3-important tokens aligns with the understood biological roles of these factors, the presence of motifs specific to neural or dermal TFs within H3K27me3-important tokens in lymphoid K562 cells likely does not signify a direct causal role of these proteins in establishing the repressive H3K27me3 mark. We posit that these factors' targets were suppressed in blood lineage progenitors, implying that the enrichment of their motifs is a reflection of the developmental trajectory of these cells.

To sum up, our study provides compelling evidence that large language models trained on DNA sequences have the capability to generate useful biological insights. This not only presents an innovative method for solving an array of genomic challenges but also forges a pathway for a more nuanced understanding of genetic data. The transformative impact of language models has already been witnessed in protein biology, where they have brought about remarkable progress in predicting protein properties and engineering novel peptides with tailored functions (63–65). This is indicative of the potential these models hold, suggesting that their capabilities go beyond mere sequence analysis. With the exponential increase in multi-omics data—spanning genomics, transcriptomics, proteomics and metabolomics—it is imperative to have advanced analytical tools that can seamlessly integrate and interpret these vast and complex datasets. Language models, as demonstrated by our findings, appear poised to fill this role. As the nexus between computational techniques and biology strengthens, it is foreseeable that language models will be pivotal in ushering in a new era of DNA-based technologies.

Data availability

The code to generate the findings of this manuscript is available in the 'supplementary code' section, on our GitHub repository (https://github.com/AIRI-Institute/GENALM) and on Zenodo (https://doi.org/10.5281/zenodo.14394199). Additionally, our trained models can be found on HuggingFace under the prefix "gena-lm": https://huggingface.co/AIRI-Institute/.

Supplementary data

Supplementary Data are available at NAR Online.

Acknowledgements

Author contributions: M.B., O.K., D.S., V.F. and Y.K. conceived the study; Y.K. performed models pre-training and fine-tuning experiments, including experiments with RMT; V.F. prepared datasets, conducted bioinformatic analysis and provided biological interpretation of the results; A.S. performed models fine-tuning and benchmarking with help from M.P and Y.K.; M.P. and D.P. performed token importance analysis; D.P. analyzed ClinVar data and splice site mutations; N.C. contributed to dataset preparation; M.B. and O.K. supervised the study; all authors contributed to manuscript preparation.

Funding

Funding for open access charge: AIRI.

Conflict of interest statement

None declared.

References

1. Kim,S. and Wysocka,J. (2023) Deciphering the multi-scale, quantitative cis-regulatory code. *Mol. Cell*, **83**, 373–392.
2. Whalen,S., Schreiber,J., Noble,W.S. and Pollard,K.S. (2022) Navigating the pitfalls of applying machine learning in genomics. *Nat. Rev. Genet.*, **23**, 169–181.
3. Libbrecht,M.W. and Noble,W.S. (2015) Machine learning applications in genetics and genomics. *Nat. Rev. Genet.*, **16**, 321–332.
4. Belokopytova,P.S., Nuriddinov,M.A., Mozheiko,E.A., Fishman,D. and Fishman,V. (2020) Quantitative prediction of enhancer–promoter interactions. *Genome Res.*, **30**, 72–84.
5. Sindeeva,M., Chekanov,N., Avetisian,M., Shashkova,T.I., Baranov,N., Malkin,E., Lapin,A., Kardymon,O. and Fishman,V. (2023) Cell type-specific interpretation of noncoding variants using deep learning-based methods. *GigaScience*, **12**, giad015.
6. Penzar,D., Nogina,D., Meshcheryakov,G., Lando,A., Rafi,A.M., de Boer,C., Zinkevich,A. and Kulakovskiy,I.V. (2022) LegNet: resetting the bar in deep learning for accurate prediction of promoter activity and variant effects from massive parallel reporter assays. bioRxiv doi: https://doi.org/10.1101/2022.12.22.521582, 23 December 2022, preprint: not peer reviewed.
7. Belokopytova,P. and Fishman,V. (2021) Predicting genome architecture: challenges and solutions. *Front. Genet.*, **11**, 617202.
8. Avsec,Ž., Agarwal,V., Visentin,D., Ledsam,J.R., Grabska-Barwinska,A., Taylor,K.R., Assael,Y., Jumper,J., Kohli,P. and Kelley,D.R. (2021) Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods*, **18**, 1196–1203.
9. Pan,S.J. and Yang,Q. (2010) A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.*, **22**, 1345–1359.
10. Dai,A.M. and Le,Q.V. (2015) Semi-supervised sequence learning. In: Cortes, C., Lawrence,N., Lee,D., Sugiyama,M. and Garnett,R. (eds.) *Advances in Neural Information Processing Systems*. Vol. **28**, MIT press, USA, pp. 3079–3087.
11. Peters,M.E., Neumann,M., Iyyer,M., Gardner,M., Clark,C., Lee,K. and Zettlemoyer,L. (2018) Deep contextualized word representations. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long*

Papers). Association for Computational Linguistics, New Orleans, Louisiana, pp. 2227–2237.

12. Howard,J. and Ruder,S. (2018) Universal language model fine-tuning for text classification. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia, pp. 328–339.
13. Radford,A., Narasimhan,K., Salimans,T. and Sutskever,I. (2018). Improving language understanding with unsupervised learning. Technical report, OpenAI.
14. Devlin,J., Chang,M.-W., Lee,K. and Toutanova,K. (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, pp. 4171–4186.
15. 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.
16. Vaswani,A., Shazeer,N., Parmar,N., Uszkoreit,J., Jones,L., Gomez,A.N., Kaiser,Ł. and Polosukhin,I. (2017) Attention is all you need. In: *Advances in Neural Information Processing Systems*. Vol. **30**, MIT press, USA, pp. 5998–6008.
17. Zhou,Z., Ji,Y., Li,W., Dutta,P., Davuluri,R. and Liu,H. (2023) DNABERT-2: efficient foundation model and benchmark for multi-species genome. arXiv doi: https://arxiv.org/abs/2306.15006, 26 June 2023, preprint: not peer reviewed.
18. Guo,Q., Qiu,X., Liu,P., Shao,Y., Xue,X. and Zhang,Z. (2019) Star-transformer. arXiv doi: https://arxiv.org/abs/1902.09113, 25 February 2019, preprint: not peer reviewed.
19. Beltagy,I., Peters,M.E. and Cohan,A. (2020) Longformer: the long-document transformer. arXiv doi: https://arxiv.org/abs/2004.05150, 10 April 2020,preprint: not peer reviewed.
20. Ainslie,J., Ontanon,S., Alberti,C., Pham,P., Ravula,A. and Sanghai,S. (2020) ETC: encoding long and structured data in transformers. In: Webber,B., Cohn,T., He,Y. and Liu,Y. (eds.) *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, USA, pp. 268–284.
21. Zaheer,M., Guruganesh,G., Dubey,K.A., Ainslie,J., Alberti,C., Ontanon,S., Pham,P., Ravula,A., Wang,Q., Yang,L., *et al.* (2020) Big Bird: transformers for longer sequences. In: Larochelle, H., Ranzato,M., Hadsell,R., Balcan,M. and Lin,H. (eds.) *Advances in Neural Information Processing Systems*. Vol. **33**, Curran Associates Inc., USA, pp. 17283–17297.
22. Kitaev,N., Kaiser,L. and Levskaya,A. (2020) Reformer: the efficient transformer. In: *International Conference on Learning Representations*.
23. Choromanski,K.M., Likhosherstov,V., Dohan,D., Song,X., Gane,A., Sarlos,T., Hawkins,P., Davis,J.Q., Mohiuddin,A., Kaiser,L., *et al.* (2021) Rethinking attention with performers. In: *International Conference on Learning Representations*.
24. Katharopoulos,A., Vyas,A., Pappas,N. and Fleuret,F. (2020) Transformers are RNNs: fast autoregressive transformers with linear attention. In: Daumé,D. III and Aarti,S. (eds.) *Proceedings of the 37th International Conference on Machine Learning*, Vol. **119**. Proceedings of Machine Learning Research, pp. 5156–5165.
25. Dai,Z., Yang,Z., Yang,Y., Carbonell,J., Le,Q. and Salakhutdinov,R. (2019) Transformer-XL: attentive language models beyond a fixed-length context. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, pp. 2978–2988.
26. Rae,J.W., Potapenko,A., Jayakumar,S.M., Hillier,C. and Lillicrap,T.P. (2020) Compressive transformers for long-range sequence modelling. In: *International Conference on Learning Representations*.
27. Wu,Q., Lan,Z., Qian,K., Gu,J., Geramifard,A. and Yu,Z. (2022) Memformer: a memory-augmented transformer for sequence modeling. In: He,Y., Ji,H., Li,S., Liu,Y. and Chang,C.-H. (eds.) *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*. Association for Computational Linguistics, USA, pp. 308–318.
28. Hutchins,D., Schlag,I., Wu,Y., Dyer,E. and Neyshabur,B. (2022) Block-recurrent transformers. In: Oh, A.H., Agarwal,A., Belgrave,D. and Cho,K. (eds.) *Advances in Neural Information Processing Systems*, Vol. **35**, MIT press, USA.
29. Bulatov,A., Kuratov,Y. and Burtsev,M. (2022) Recurrent memory transformer. In: Koyejo,S., Mohamed,S., Agarwal,A., Belgrave,D., Cho,K. and Oh,A. (eds.) *Advances in Neural Information Processing Systems*, Vol. **35**, MIT press, USA, pp. 11079–11091.
30. Bulatov,A., Kuratov,Y., Kapushev,Y. and Burtsev,M. (2024) Beyond attention: breaking the limits of transformer context length with recurrent memory. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. **38**, AAAI Press, USA, pp. 17700–17708.
31. Kang,M., Wu,H., Liu,H., Liu,W., Zhu,M., Han,Y., Liu,W., Chen,C., Song,Y., Tan,L., *et al.* (2023) The pan-genome and local adaptation of *Arabidopsis thaliana*. *Nat. Commun.*, **14**, 6259.
32. O'Donnell,S., Yue,J.-X., Saada,O.A., Agier,N., Caradec,C., Cokelaer,T., De Chiara,M., Delmas,S., Dutreux,F., Fournier,T., *et al.* (2023) Telomere-to-telomere assemblies of 142 strains characterize the genome structural landscape in *Saccharomyces cerevisiae*. *Nat. Genet.*, **55**, 1390–1399.
33. Kim,B.Y., Gellert,H.R., Church,S.H., Suvorov,A., Anderson,S.S., Barmina,O., Beskid,S.G., Comeault,A.A., Crown,K.N., Diamond,S.E., *et al.* (2024) Single-fly assemblies fill major phylogenomic gaps across the Drosophilidae Tree of Life. *PLoS Biol.*, **22**, e3002697.
34. Sennrich,R., Haddow,B. and Birch,A. (2016) Neural machine translation of rare words with subword units. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, pp. 1715–1725.
35. Jaganathan,K., Kyriazopoulou Panagiotopoulou,S., McRae,J.F., Darbandi,S.F., Knowles,D., Li,Y.I., Kosmicki,J.A., Arbelaez,J., Cui,W., Schwartz,G.B., *et al.* (2019) Predicting splicing from primary sequence with deep learning. *Cell*, **176**, 535–548.
36. de Almeida,B.P., Reiter,F., Pagani,M. and Stark,A. (2022) DeepSTARR predicts enhancer activity from DNA sequence and enables the *de novo* design of synthetic enhancers. *Nat. Genet.*, **54**, 613–624.
37. Zhou,J. and Troyanskaya,O.G. (2015) Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods*, **12**, 931–934.
38. Bogard,N., Linder,J., Rosenberg,A.B. and Seelig,G. (2019) A deep neural network for predicting and engineering alternative polyadenylation. *Cell*, **178**, 91–106.
39. Dalla-Torre,H., Gonzalez,L., Revilla,J.M., Carranza,N.L., Grzywaczewski,A.H., Oteri,F., Dallago,C., Trop,E., Sirelkhatim,H., Richard,G., *et al.* (2024) The Nucleotide Transformer: building and evaluating robust foundation models for human genomics. *Nat. Methods*, https://doi.org/10.1038/s41592-024-02523-z.
40. Nguyen,E., Poli,M., Faizi,M., Thomas,A., Birch-Sykes,C., Wornow,M., Patel,A., Rabideau,C., Massaroli,S., Bengio,Y., *et al.* (2023) HyenaDNA: long-range genomic sequence modeling at single nucleotide resolution. arXiv doi: https://arxiv.org/abs/2306.15794, 27 June 2023, preprint: not peer reviewed.
41. Xiong,R., Yang,Y., He,D., Zheng,K., Zheng,S., Xing,C., Zhang,H., Lan,Y., Wang,L. and Liu,T. (2020) On layer normalization in the transformer architecture. In: Daumé,III and Aarti,S. (eds.) *Proceedings of the 37th International Conference on Machine Learning*, Vol. **119**, PMLR, USA, pp. 10524–10533.

42. Loshchilov,I. and Hutter,F. (2019) Decoupled weight decay regularization. In: *Proceedings of the 7th International Conference on Learning Representations*. ICLR, USA.
43. Su,J., Lu,Y., Pan,S., Wen,B. and Liu,Y. (2021) RoFormer: enhanced transformer with rotary position embedding. arXiv doi: https://arxiv.org/abs/2104.09864, 20 April 2021, preprint: not peer reviewed.
44. Goyal,P., Dollár,P., Girshick,R.B., Noordhuis,P., Wesolowski,L., Kyrola,A., Tulloch,A., Jia,Y. and He,K. (2017) Accurate, large minibatch SGD: training imagenet in 1 hour. arXiv doi: https://arxiv.org/abs/1706.02677, 08 June 2017, preprint: not peer reviewed.
45. Mukund,Sundararajan and Ankur Taly,Q.Y. (2017) Axiomatic attribution for deep networks. In: *Proceedings of the 34th International Conference on Machine Learning*. Vol. **70**, PMLR, USA, pp. 3319–3328.
46. Grant,C.E. and Bailey,T.L. (2021) XSTREME: comprehensive motif analysis of biological sequence datasets. bioRxiv doi: https://doi.org/10.1101/2021.09.02.458722, 03 September 2021, preprint: not peer reviewed.
47. Kulakovskiy,I.V., Vorontsov,I.E., Yevshin,I.S., Sharipov,R.N., Fedorova,A.D., Rumynskiy,E.I., Medvedeva,Y.A., Magana-Mora,A., Bajic,V.B., Papatsenko,D.A., *et al.* (2018) HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res.*, **46**, D252–D259.
48. Rhee,H.S. and Pugh,B.F. (2012) ChIP-exo: a method to identify genomic location of DNA-binding proteins at near single nucleotide accuracy. In: Ausubel, F.M. (ed.). *Current Protocols in Molecular Biology*. John Wiley and Sons, USA.
49. Kawana,M., Lee,M.E., Quertermous,E.E. and Quertermous,T. (1995) Cooperative interaction of GATA-2 and AP1 regulates transcription of the endothelin-1 gene. *Mol. Cell. Biol.*, **15**, 4225–4231.
50. Menendez-Gonzalez,J.B., Sinnadurai,S., Gibbs,A., Thomas,L.-A., Konstantinou,M., Garcia-Valverde,A., Boyer,M., Wang,Z., Boyd,A.S., Blair,A., *et al.* (2019) Inhibition of GATA2 restrains cell proliferation and enhances apoptosis and chemotherapy mediated apoptosis in human GATA2 overexpressing AML cells. *Sci. Rep.*, **9**, 12212.
51. Eferl,R. and Wagner,E.F. (2003) AP-1: a double-edged sword in tumorigenesis. *Nat. Rev. Cancer*, **3**, 859–868.
52. Camellato,B.R., Brosh,R., Ashe,H.J., Maurano,M.T. and Boeke,J.D. (2024) Synthetic reversed sequences reveal default genomic states. *Nature*, **628**, 373–380.
53. Rogers,A., Kovaleva,O. and Rumshisky,A. (2021) A primer in BERTology: what we know about how BERT works. *Trans. Assoc. Comput. Ling.*, **8**, 842–866.
54. Vig,J., Madani,A., Varshney,L.R., Xiong,C., Socher,R. and Rajani,N. (2021) {BERT}ology meets biology: interpreting attention in protein language models. In: *Proceedings of the International Conference on Learning Representations*.
55. Shmelev,A., Petrov,M., Penzar,D., Akhmetyanov,N., Tavritskiy,M., Mamontov,S., Kuratov,Y., Burtsev,M., Kardymon,O. and Fishman,V. (2024) GENA-Web - GENomic Annotations Web Inference using DNA language models. bioRxiv doi: https://doi.org/10.1101/2024.04.26.591391, 29 April 2024, preprint: not peer reviewed.
56. Clauwaert,J., Menschaert,G. and Waegeman,W. (2021) Explainability in transformer models for functional genomics. *Brief. Bioinform.*, **22**, bbab060.
57. Khan,A. and Lee,B. (2021) Gene transformer: transformers for the gene expression-based classification of lung cancer subtypes. arXiv doi: https://arxiv.org/abs/2108.11833, 26 August 2021, preprint: not peer reviewed.
58. Zhang,T.-H., Hasib,M.M., Chiu,Y.-C., Han,Z.-F., Jin,Y.-F., Flores,M., Chen,Y. and Huang,Y. (2022) Transformer for gene expression modeling (T-GEM): an interpretable deep learning model for gene expression-based phenotype predictions. *Cancers*, **14**, 4763.
59. Le,N. Q.K. and Ho,Q.-T. (2022) Deep transformers and convolutional neural network in identifying DNA N6-methyladenine sites in cross-species genomes. *Methods*, **204**, 199–206.
60. Mowlaei,M.E., Li,C., Chen,J., Jamialahmadi,B., Kumar,S., Rebbeck,T.R. and Shi,X. (2023) Split-transformer Impute (STI): genotype imputation using a transformer-based model. bioRxiv doi: https://doi.org/10.1101/2023.03.05.531190, 06 March 2023, preprint: not peer reviewed.
61. Kabirova,E., Nurislamov,A., Shadskiy,A., Smirnov,A., Popov,A., Salnikov,P., Battulin,N. and Fishman,V. (2023) Function and evolution of the loop extrusion machinery in animals. *Int. J. Mol. Sci.*, **24**, 5017.
62. Fishman,V.S., Salnikov,P.A. and Battulin,N.R. (2018) Interpreting chromosomal rearrangements in the context of 3-dimentional genome organization: a practical guide for medical genetics. *Biochemistry (Mosc.)*, **83**, 393–401.
63. Shashkova,T.I., Umerenkov,D., Salnikov,M., Strashnov,P.V., Konstantinova,A.V., Lebed,I., Shcherbinin,D.N., Asatryan,M.N., Kardymon,O.L. and Ivanisenko,N.V. (2022) SEMA: antigen B-cell conformational epitope prediction using deep transfer learning. *Front. Immunol.*, **13**, 960985.
64. Madani,A., Krause,B., Greene,E.R., Subramanian,S., Mohr,B.P., Holton,J.M., Olmos Jr,J.L., Xiong,C., Sun,Z.Z., Socher,R., *et al.* (2023) Large language models generate functional protein sequences across diverse families. *Nat. Biotechnol.*, **41**, 1099–1106.
65. Wang,Z., Combs,S.A., Brand,R., Calvo,M.R., Xu,P., Price,G., Golovach,N., Salawu,E.O., Wise,C.J., Ponnapalli,S.P., *et al.* (2022) LM-GVP: an extensible sequence and structure informed deep learning framework for protein property prediction. *Sci. Rep.*, **12**, 6832.

Received: September 16, 2024. **Editorial Decision:** December 8, 2024. **Accepted:** December 26, 2024