Advances in Computer, Signals and Systems (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/acss.2024.080709
ISSN 2371-8838 Vol. 8 Num. 7

# *Advances in foundation models for genomics: A detailed exploration of developments*

**Kaizhuang Jing[1,a], Shuangkai Han[1,b,*]**

*[1]School of Information Science and Technology, Yunnan Normal University, Kunming, China*
*[a]kzjing@foxmail.com, [b]han_skai@163.com*
*[*]Corresponding author*

***Keywords:*** Genomics, foundation models, large language models, DNA sequences, deep learning

***Abstract:*** Foundation models (FMs) are a class of deep learning models originating from natural language processing (NLP), trained on large-scale datasets through self-supervised techniques. After pre-training, these models can be fine-tuned with labeled data to accomplish a variety of downstream tasks. FMs have demonstrated outstanding performance across numerous NLP tasks and have been successfully applied in the fields of biology and medicine, exhibiting remarkable efficacy. However, despite the development of multiple FMs specifically tailored for genomics, referred to as genomic foundation models (GFMs), there remains a lack of systematic analysis of these models. This review provides an overview of the current applications and developments of GFMs, offering a comprehensive analysis of their strengths and weaknesses and categorizing their underlying principles. Given the inherent differences between DNA sequences and natural language, designing FMs suitable for genomics presents significant challenges. This paper aims to provide researchers with a detailed analytical report and valuable insights to guide the further development of high-quality GFMs.

## 1. Introduction

The Transformer model, introduced in 2017, catalyzed the creation of foundation models (FMs). FMs are large-scale pre-trained neural networks trained on vast amounts of data, which can be fine-tuned to perform a variety of downstream tasks. During pre-training, FMs acquire general knowledge or representations, which are then applied to downstream tasks through transfer learning during the fine-tuning step. The fine-tuned models synthesize knowledge representations from both tasks, thereby enhancing generalization performance on downstream tasks. The BERT model, the first FM, was designed based on the Transformer encoder architecture and served as the inspiration for many subsequent models, including RoBERTa, BART, and T5. These models have become prevalent in state-of-the-art (SOTA) natural language processing (NLP) model development and are collectively known as large language models. Although the BERT model was introduced in late 2018, these models gained broader recognition across various domains following the release of ChatGPT at the end of 2022. Additionally, they have been rapidly extended to different fields, such as text, images, videos, speech, tabular data, protein sequences, organic molecules, and reinforcement learning.

With the significant advancements of FMs in natural language understanding, there has been growing interest in applying FMs to understand and decode the genome. FMs, through numerical embeddings, can comprehend genomic sequences and be directly utilized for various genomic analysis tasks. These models can capture complex relationships and dependencies within DNA sequences, opening new avenues for understanding transcriptional regulation, non-coding genetic variants associated with human diseases and traits, and the functional effects of regulatory elements. Recent advancements in genomic language modeling have demonstrated its superiority in a range of downstream applications, including promoter prediction, DNA methylation prediction, chromatin state analysis, promoter-enhancer interaction prediction, TF-DNA binding prediction, and variant effect prediction, among others. These models provide researchers with powerful tools to understand the functional significance of different genomic elements and uncover critical biological processes and mechanisms.

The rapid development of numerous genomic foundation models (GFMs) has deepened our understanding of the genome while also contributing to an information deluge. Although comprehensive reviews of FMs within the broader healthcare domain are available, their coverage of genomics remains relatively superficial. Investigations specifically related to GFMs have identified only two pertinent studies thus far: the first[1] primarily reviewed literature up to 2023, while the second[2] focused on the applications and development processes. This paper aims to complement existing GFM reviews by providing a clearer and more concise understanding of GFM development, key development factors, challenges, and the latest trends in GFMs.

## 2. Development Review

Understanding the development history of GFMs aids in deeply comprehending the current state of the field and gaining insights into future directions. In Table 1, we have listed the existing GFMs and summarized their core designs. Next, we will briefly introduce these GFMs in chronological order according to their model architectures (including: Transformer, SSM, and other architectures).

### 2.1. Transformer-based architectures

GFMs based on the Transformer architecture are widely adopted by researchers. In these studies, to enhance predictive performance and generalizability, researchers have explored various approaches such as pure DNA input, multimodal input, and proxy input from natural language to construct GFMs. Although the Transformer architecture has been recognized for its powerful contextual capturing ability, the computational cost introduced by its quadratic complexity and the limitation of context window size are significant constraints for its application. To mitigate these limitations, researchers have attempted sparse attention mechanisms, lightweight architectures, and efficient tokenization methods. GFMs based on the Transformer architecture can also be further categorized according to the Transformer component used into encoder-based models and decoder-based models (including encoder-decoder models). The former excels at tasks requiring deep contextual understanding, while the latter is more adept at generative tasks.

#### 2.1.1. Encoder-based models

DNABERT is a BERT-based model designed to capture a global and transferable understanding of genomic DNA sequences. Through self-supervised learning and fine-tuning, the model demonstrated SOTA performance in tasks such as promoter prediction, splice site prediction, and transcription factor binding site prediction, and showed good cross-species generalization capabilities. DNABERT offers an interpretable method, allowing direct visualization of nucleotide-level importance and semantic relationships within the input sequences, providing new perspectives for studying genomic regulatory elements. However, DNABERT faces certain limitations in handling long sequences, particularly concerning the context window size, which affects the model's ability to capture long-range dependencies.

Table 1: Summary of GFMs covered in this review. The abbreviations used in this table are: BPE (Byte Pair Encoding), MLM (Masked Language Modeling), NSP (Next Sentence Prediction), CLM (Causal Language Modeling), CNN (Convolutional Neural Network), RNN (Recurrent Neural Network), SSM (State Space Model). The release dates refer to the earliest available time, including preprint versions.

| Model name | Tokenization method | Pre-training task | Model architecture | Release date |
|---|---|---|---|---|
| DNABERT | Overlapping k-mer | MLM | Transformer | 2020.09 |
| GeneBERT | Overlapping k-mer | MLM+NSP | Transformer | 2021.10 |
| GPN[9] | Nucleotide-level | MLM | CNN | 2022.08 |
| NT[3] | Non-overlapping k-mer | MLM | Transformer | 2023.01 |
| Species LM[4] | Overlapping k-mer | MLM | Transformer | 2023.01 |
| DNABERT-2[5] | BPE | MLM | Transformer | 2023.06 |
| HyenaDNA[7] | Nucleotide-level | CLM | SSM | 2023.06 |
| DNAGPT | Non-overlapping k-mer | CLM | Transformer | 2023.07 |
| GPN-MSA | Nucleotide-level | MLM | Transformer | 2023.10 |
| UTR-LM | Nucleotide-level | MLM | Transformer | 2023.10 |
| hgT5 | Unigram | T5 | Transformer | 2023.10 |
| MegaDNA | Nucleotide-level | CLM | Transformer | 2023.12 |
| Evo | Nucleotide-level | CLM | SSM+ Transformer | 2024.02 |
| Caduceus[8] | Nucleotide-level | MLM | SSM | 2024.03 |
| SpliceBERT[6] | Nucleotide-level | MLM | Transformer | 2024.04 |
| ChatNT | Overlapping k-mer | CLM | Transformer | 2024.05 |
| PlantCaduceus | Nucleotide-level | MLM | SSM | 2024.06 |
| CD-GPT | BPE | CLM | Transformer | 2024.06 |

GeneBERT is a multimodal model that combines BERT and Swin Transformer to enhance genomic data analysis. Its primary contribution lies in leveraging both one-dimensional gene sequences and two-dimensional interactions between transcription factors and gene regulatory regions. By employing three pre-training tasks—Masked Language Modeling (MLM), Next Sentence Prediction (NSP), and Sequence-Region Matching—the model enhances its robustness and generalization capabilities. Compared to previous models that performed well only in specific cell types, GeneBERT better captures gene expression regulatory mechanisms across different cell types, demonstrating superior performance in downstream tasks such as promoter prediction and disease risk assessment.

The NT[3] model shares a similar architecture with DNABERT but employs a larger model size (five times larger than DNABERT). It successfully generates transferable, context-specific nucleotide sequence representations using 3,202 human genomes and 850 genomes from various species. This approach achieved accurate predictions of molecular phenotypes in low-data environments and matched or exceeded the performance of specialized methods in 18 prediction tasks based solely on sequence representations. This study provides a significant foundation for accurately predicting molecular phenotypes from DNA sequences in genomics, showcasing the immense potential of FMs in bioinformatics. It also empirically demonstrated that increasing the model size can lead to better performance.

Species LM[4] is a species-aware DNA language model based on DNABERT, trained on genomic data from over 800 species spanning more than 500 million years of evolutionary history. The research indicates that this model can effectively distinguish between transcription factor and RNA-binding protein motifs and background non-coding sequences, demonstrating its powerful flexibility and ability to capture conserved regulatory elements, surpassing the limitations of traditional sequence alignment. This model not only reconstructs in vivo binding motif instances more accurately but also considers the evolution of motif sequences and their positional constraints. These findings suggest that species-aware training significantly improves the sequence representation capability for gene expression prediction and motif discovery.

DNABERT-2[5] is an extended version of DNABERT, replacing the traditional k-mer tokenization method with BPE as the new tokenization strategy, significantly enhancing computational efficiency and model performance. Additionally, the authors employed the ALiBi method, which removes strict input length limitations. They also recognized the lack of standardized benchmarks in the field of genomic understanding and subsequently created the Genomic Understanding Evaluation dataset, providing a comprehensive evaluation framework for future genomic models.

GPN-MSA is a model based on the Transformer architecture (RoFormer), aimed at improving the prediction accuracy of variant effects across the genome. Similar to GPN, this model leverages cross-species whole-genome sequence alignments to effectively predict both coding and non-coding variants in the human genome. Through evaluations on multiple clinical databases (ClinVar, COSMIC, OMIM), experimental functional assays (DMS, DepMap), and population genomic data (gnomAD), GPN-MSA demonstrated outstanding performance in predicting the pathogenicity of variants.

UTR-LM is a model based on the Transformer encoder, integrating sequence, secondary structure, and minimum free energy information through semi-supervised learning. The model successfully learned meaningful semantic representations related to the mRNA translation process. UTR-LM outperformed existing best baselines in tasks such as predicting mean ribosome loading, translation efficiency, mRNA expression level, and internal ribosome entry site, demonstrating its effectiveness in 5' UTR modeling. Furthermore, the study experimentally validated the model's predictive capabilities by designing 211 novel 5' UTR sequences with high translation efficiency.

SpliceBERT[6] is a BERT-based model specifically designed for analyzing primary RNA sequences from 72 vertebrate species. By pre-training on a diverse set of species sequences, the model effectively identifies evolutionarily conserved elements. This approach not only enhances the sequence modeling capability for RNA splicing but also exhibits outstanding performance in several downstream tasks, such as zero-shot prediction of variant impacts on splicing, human branch point prediction, and cross-species splice site prediction.

### 2.1.2. Decoder-based models

DNAGPT is a GPT-based model designed to address the challenges of extracting information from DNA sequences and to adapt to various tasks and data types. DNAGPT enhances the classical GPT model by training on over 200 billion mammalian base pairs, incorporating binary classification tasks (DNA sequence order) and numerical regression tasks (guanine-cytosine content prediction), along with a comprehensive token language. This allows DNAGPT to handle a diverse array of DNA analysis tasks while accommodating both sequence and numerical data processing. Through evaluations on tasks such as genomic signal and region recognition, mRNA abundance regression, and artificial genome generation, DNAGPT has shown superior performance compared to existing models tailored for specific downstream tasks, demonstrating the advantages of the newly designed model architecture in pre-training.

MegaDNA is a model based on the Multiscale Transformers architecture. The study demonstrated this model's foundational capabilities in predicting key genes, the effects of genetic variants, regulatory element activity, and the classification of unannotated sequences. Additionally, it can generate new sequences up to 96K base pairs long, which include functional regulatory elements and new proteins associated with bacterial viruses. This work not only advances the development of genomic language models but also opens new possibilities for synthetic biology applications. However, the paper also highlights limitations in the model's optimization at the gene and codon levels for efficient self-replication. Future research needs to delve into ethical, safety, and regulatory frameworks to ensure the responsible application of generative models in synthetic biology.

ChatNT is a multimodal conversational agent model designed to tackle complex biological tasks using NLP technology. By transforming genomic prediction tasks into a text-to-text format, the authors enabled users to interact with the model using English commands, thereby lowering the barrier to entry and enhancing accessibility. Additionally, ChatNT can handle DNA, RNA, and protein sequences simultaneously, demonstrating potential applications in transcriptomics and proteomics. This represents a significant step towards the development of a general-purpose biological AI system. However, the paper also notes that the current model's performance remains limited for specific tasks and requires fine-tuning for each task, which may hinder the model's generalization ability.

CD-GPT is a GPT-based model aimed at connecting different types of biological molecular sequences through the central dogma. The study's contribution lies in introducing a unified representation space and a shared multi-molecular vocabulary, effectively representing biological sequences and reducing their distance in the embedding space. Through extensive pre-training on comprehensive molecular-level data, CD-GPT exhibits outstanding performance in various predictive and generative tasks, including genome element detection, protein property prediction, and RNA-protein interaction identification. Moreover, the model can generate new protein sequences and perform reverse translation, showcasing its broad application potential in multi-omics analysis.

## 2.2. SSM-based architectures

GFMs based on the State Space Model (SSM) architecture offer an alternative to Transformer-based architectures. These models not only reduce the model size by several orders of magnitude while maintaining comparable performance, but they also outperform in certain tasks. Additionally, they exhibit lower time complexity and possess the capability to handle longer contexts, making them a promising research direction.

HyenaDNA[7] is a model based on SSM that enables long-range genomic sequence modeling at single-nucleotide resolution. Compared to traditional Transformer-based models, HyenaDNA can handle up to 1 million context lengths, which is 500 times the capacity of previous models, significantly enhancing the ability to model long-range interactions in DNA. Additionally, HyenaDNA leverages the advantages of implicit convolution, offering lower time complexity and faster training speeds. In long-range species classification tasks, HyenaDNA has demonstrated outstanding performance and achieved SOTA levels in multiple benchmark tests.

Caduceus[8] is a model based on SSM, designed to address key challenges in genomic sequence modeling, such as long-range token interactions, the influence of upstream and downstream regions, and DNA's reverse complementarity (RC). The authors developed the Caduceus model by extending the long-range Mamba blocks to support bidirectionality with BiMamba components and incorporating MambaDNA blocks that accommodate RC equivariance. This marks the first long-range DNA language model with RC equivariance and bidirectionality, significantly enhancing performance in downstream tasks. Notably, in long-range variant effect prediction tasks, Caduceus outperformed non-bidirectional or non-RC equivariant models by over tenfold.

PlantCaduceus is an extension of the Caduceus model, pre-trained on the genomes of 16 diverse angiosperm species. This pre-training allows PlantCaduceus to achieve efficient cross-species predictions with limited annotated data. The study shows that PlantCaduceus excels in transcription and translation modeling tasks on the maize genome, dating back approximately 160 million years, surpassing traditional supervised learning models. Additionally, the model can identify deleterious mutations across the entire genome without the need for multiple sequence alignments, and demonstrates significant enrichment in prioritizing rare allele mutations.

## 2.3. Other architectures

In addition to purely Transformer-based and SSM-based GFMs, researchers have also explored models based on convolutional neural networks (CNNs), recurrent neural networks (RNNs), and various hybrid architectures. These explorations provide important insights into the applicability and complementarity of different architectures.

GPN[9] is a CNN-based model that effectively predicts the effects of whole-genome variants through unsupervised pre-training on genomic DNA sequences. Researchers trained this model on unaligned reference genomes of *Arabidopsis thaliana* and related species, demonstrating its superiority in predicting functional impacts of variants, surpassing existing prediction tools based on conservation scores.

Evo is a hybrid model based on StripedHyena, combining Hyena and RoFormer. Evo was trained on 2.7 million prokaryotic and phage genomes, capable of zero-shot functional prediction across the three fundamental modes of the central dogma of molecular biology, performing comparably to leading domain-specific language models. Evo excels in multi-element generative tasks, capable of generating synthetic CRISPR-Cas molecular complexes and entire transposon systems. Leveraging information learned from whole genomes, Evo can predict gene essence at nucleotide resolution and generate coding-rich sequences up to 650 kb, several orders of magnitude longer than previous methods. Evo's advancements in multi-modal and multi-scale learning provide insights into better understanding and controlling biological complexity across multiple levels.

## 3. Key Development Factors

Based on existing GFM literature, the key factors in developing GFMs can be divided into three main components (see Figure 1): selection of pre-training datasets, tokenization methods, and model architectures and pre-training tasks.

## 3.1. Selection of pre-training datasets

The choice of pre-training datasets is crucial in the development of GFMs, as it determines the scope of general knowledge the model can acquire. The quality and quantity of these datasets directly impact the quality of the model's embedded representations. However, selecting appropriate training data requires deep domain-specific knowledge, especially in genomics, where there are no universally recognized, curated datasets akin to those in NLP (e.g., the Pile) or protein research (e.g., UniProt). In selecting datasets, considerations should include quality control, data duplication, quantity of data, and the selection of the data context window.

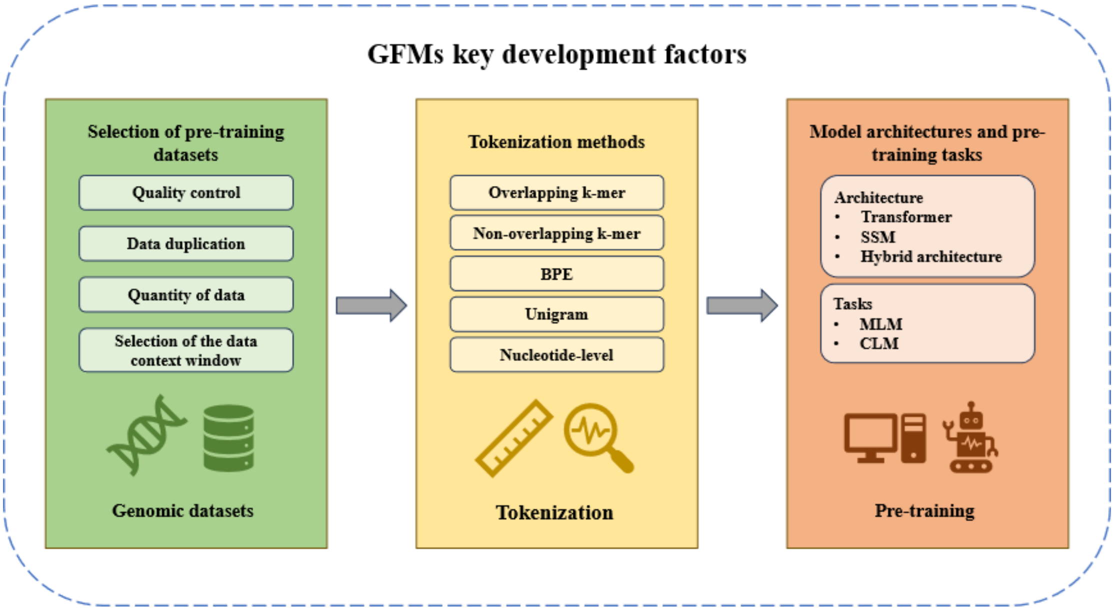

Figure 1: This figure illustrates the critical factors to be considered when developing GFMs. These factors directly impact the applicability and generalization capability of GFMs and determine their fine-tuning performance on specific downstream tasks (evaluation benchmarks). The key factors include dataset selection, tokenization methods, model architectures, and the design of pre-training tasks. The arrows in the figure indicate the dependencies and sequential relationships among these factors. By rationally designing and optimizing these elements, GFMs can better adapt to various genomics-related tasks, providing more accurate and reliable predictions.

Quality control: In protein research, quality control involves removing predicted pseudogenes or truncated proteins that no longer have a function. A recent study found that only 3.3% of bases in the human reference genome (the most commonly used GFM training dataset) are significantly constrained and likely functional. Typical genomic sequences used to train GFMs include a mix of functional and non-functional sites. A recommended solution is to apply base-pair level weighting to the training loss based on functional evidence.

Data duplication: In NLP and protein domains, filtering out duplicate sequences is standard practice, aiding in improving training efficiency and reducing overfitting. Given the high proportion of repeat sequences in eukaryotes, solutions include down-weighting or down-sampling strategies.

Quantity of data: To ensure sufficient data volume, one approach is to use sequence variants from the same species, although variations among individuals are typically limited. Another common method is to train across multiple species.

Selection of the data context window: Many interactions in the genome are limited to nearby locations (<6 kb), such as the motifs of transcription factor binding sites. However, long-range interactions also exist in the genome, such as those between exons of the same gene or between enhancers and promoters, which can span up to 1 Mb. Such extensive context lengths pose computational and statistical challenges that researchers are actively working to overcome. Additionally, regardless of the chosen context length, partitioning the genome into independent units remains challenging. For instance, an enhancer for one gene may be located within the intron of another gene.

## 3.2. Tokenization methods

The tokenization methods currently employed in GFMs include overlapping k-mer, non-overlapping k-mer, BPE, Unigram, and nucleotide-level.

Overlapping k-mer: This method effectively captures local sequence information but requires high computational and storage complexity due to the overlapping regions.

Non-overlapping k-mer: This approach reduces computational resource expenditure but may lead to the loss of some crucial local sequence information.

Byte Pair Encoding (BPE): BPE iteratively merges character pairs based on frequency distributions in the data, efficiently reducing the number of tokens and enhancing model training and inference efficiency. However, it may overlook certain semantic information.

Unigram: This method tokenizes based on the frequency distribution of individual nucleotides, preserving all sequence information but generating a large number of tokens, which leads to high input dimensions and increased computational complexity.

Nucleotide-level: This method decomposes the sequence into individual nucleotides, fully retaining the original information but resulting in extremely high input dimensions, thus incurring significant computational costs.

Each tokenization method has its own strengths and weaknesses, making it suitable for different genomic modeling and prediction tasks. Selecting an appropriate tokenization method requires balancing computational complexity, information retention, and model accuracy according to the specific needs of the task and the characteristics of the data. An optimal tokenization method can significantly enhance the performance and predictive capabilities of GFMs.

### 3.3. Model architectures and pre-training tasks

Currently, the architectures of GFMs primarily include Transformer-based architectures, State Space Model (SSM)-based architectures, and hybrid architectures. Transformer-based GFMs can be further divided into three categories: models based on Transformer encoders (similar to the BERT architecture), models based on Transformer decoders (similar to the GPT architecture), and full Transformer architecture models. SSM-based GFMs currently have two representative models: HyenaDNA, which is based on the Hyena hierarchy, and Caduceus, which is based on Mamba. The Transformer architecture leverages self-attention mechanisms but suffers from quadratic scaling issues. To address this, researchers have explored using SSMs as alternatives to Transformers, as SSMs can scale nearly linearly with sequence length, significantly improving the length and efficiency of sequence modeling. Additionally, there are hybrid architecture models that combine SSMs and Transformers, such as the EVO model. These hybrid architectures aim to integrate the advantages of different architectures to enhance overall performance.

The pre-training tasks for GFMs mainly include Masked Language Modeling (MLM) and Causal Language Modeling (CLM). The MLM task involves randomly masking parts of the input sequence and then training the model to predict the masked content based on the context. This task is primarily used to capture bidirectional dependencies within sequences and is commonly applied in Transformer encoder-based GFMs. The CLM task, also known as autoregressive language modeling, trains the model to generate or predict the next word or token in a sequence from left to right, emphasizing forward dependencies within the sequence. This task is often used in Transformer decoder-based and SSM-based GFMs. Both tasks require the model to predict data components given a context, thereby forcing the model to learn low-dimensional representations of the data. MLM generally excels in obtaining better representations and transfer learning capabilities, while CLM performs better in generative tasks.

## 4. Challenges and Future Directions

The primary challenges and future directions for GFMs encompass three main areas:

interpretability, pre-training task design.

## 4.1. Interpretability

To elucidate how these models generate predictions, particularly in genomics, a variety of methods have been developed. For instance, unsupervised clustering of the final layer embeddings in GPN[9] revealed distinct clusters for different genomic categories, such as coding sequences, introns, and untranslated regions. Similarly, SpliceBERT's[6] unsupervised clustering of embeddings for canonical splice sites and non-splice GT/AG sites demonstrated clear clusters corresponding to these two groups, indicating that the model captures key contextual patterns that determine genomic functional elements.

The attention mechanism in Transformer models aims to capture interaction patterns between input tokens. By interpreting attention weights or attention maps for a given input sequence, one can uncover the genomic features learned by the model. For example, in SpliceBERT[6], the attention weights between splice donors and acceptors were significantly higher than those between random site pairs, with even stronger interaction intensities in true donor-acceptor pairs.

Nucleotide reconstruction methods have also been employed to discover sequence motifs learned by the model. This approach has been utilized in GPN[9] to reveal significant patterns in reconstructed nucleotide distributions, particularly at functionally important sites like coding sequences and splice donor/acceptor sites. Moreover, tools like TF-MoDISco have been used to identify novel transcription factor binding sites, where the discovered sequence motifs matched those in known databases. Similarly, sequence motifs reconstructed by Species LM[4] matched binding sites of DNA and RNA-binding proteins in species not seen during training. These studies suggest that GFMs can not only capture functional genomic patterns but also reveal species-specific sequence motifs and regulatory code evolution.

It is important to note that while attention scores in Transformer models have been proposed as an interpretability method to address the black-box issue, several studies have shown that attention scores alone do not inherently possess interpretability. Other methods, such as attention flow and attention rollout, and layer-wise relevance propagation, have been successfully applied to interpret Transformer models, offering superior interpretability compared to mere attention scores. However, attention rollout methods cannot distinguish between positive and negative contributions, and attention flow methods are computationally complex. Layer-wise relevance propagation, by backpropagating the network's output prediction to the input layer, provides indications of feature importance and has demonstrated its superiority in Transformer models. Additionally, model-agnostic methods like LIME, SHAP, and weighted SHAP offer potential pathways for interpreting Transformer models.

## 4.2. Pre-training task design

The design of pre-training tasks for GFMs presents significant challenges. Ideally, pre-training allows deep learning models to capture universal data patterns; however, in the worst case, it may result in a waste of computational resources. Recent studies evaluating various GFMs on human genome prediction tasks have found that their performance generally does not surpass that of non-GFM baselines. These results are based on frozen embeddings, and a comprehensive evaluation involving full fine-tuning would provide more insights. Although GFMs are well-suited to demonstrate the value of transfer learning in less-studied organisms, delivering significant value in human genetics may require further innovation due to the availability of high-quality labeled data and well-designed models in this field. An important question is the extent to which the scaling hypothesis applies to GFMs—whether increasing the amount of unlabeled data and computational power will continuously enhance model performance.

## 5. Conclusion

GFMs have ushered in new possibilities for genomic tasks and the decoding of the DNA language. These models have emerged as powerful tools capable of extracting complex patterns, facilitating applications such as adaptive assessment, sequence design, and transfer learning. While the breakthroughs are indeed promising, it is imperative to maintain a balanced perspective, critically evaluating the applicability and limitations of these models, particularly concerning the rationality of pre-training designs. We must avoid overstating the novelty of GFMs or introducing unnecessary terminology.

## Acknowledgements

This work was supported by the Applied Basic Research Project in Yunnan Province (grant no. 202201AT070042), the project funding of the "Support Program of Xingdian Talents", the National Natural Science Foundation of China (grant no. 61862067, U1902201), Yunnan Provincial Science and Technology Department-Yunnan University Double First-Class Joint Fund Key Projects (grant no. 2019FY003027) and National Key R&D Program of China (grant no. 2022YFC2602500), Yunnan Normal University graduate research innovation fund project (grant no. YJSJJ23-B128).

## References

*[1] Consens M E, Dufault C, Wainberg M, et al. To transformers and beyond: large language models for the genome [J]. arXiv preprint arXiv:2311.07621, 2023.*
*[2] Benegas G, Ye C, Albors C, et al. Genomic Language Models: Opportunities and Challenges[J]. arXiv preprint arXiv:2407.11435, 2024.*
*[3] Dalla-Torre, H., Gonzalez, L., Mendoza Revilla, J., et al. The Nucleotide Transformer: Building and Evaluating Robust Foundation Models for Human Genomics. bioRxiv preprint, 2023.*
*[4] Karollus, A., Hingerl, J., Gankin, D., et al. Species-aware DNA language models capture regulatory elements and their evolu- tion. Genome Biology 25, 83, 2024.*
*[5] Zhou, Z., Ji, Y., Li, W., et al. DNABERT-2: Efficient foundation model and benchmark for multi-species genome. arXiv preprint arXiv:2306.15006, 2023.*
*[6] Chen, K., Zhou, Y., Ding, M., et al. Self-supervised learning on millions of primary RNA sequences from 72 vertebrates improves sequence-based RNA splicing prediction. Briefings in Bioinformatics 25, bbae163, 2024.*
*[7] Nguyen, E., Poli, M., Faizi, M., et al. HyenaDNA: Long- Range Genomic Sequence Modeling at Single Nucleotide Resolution. Advances in Neural Information Processing Systems vol. 36. Curran Associates, Inc. 43177–43201, 2023.*
*[8] Schiff, Y., Kao, C.-H., Gokaslan, A., et al. Caduceus: Bi-directional equivariant long-range DNA sequence modeling. arXiv preprint arXiv:2403.03234, 2024.*
*[9] Benegas, G., Batra, S. S., and Song, Y. S. DNA language models are powerful pre- dictors of genome-wide variant effects. Proceedings of the National Academy of Sciences 120, e2311219120, 2023.*