

Article

Benchmark of Genomic Language Models on Human and Rice Genomic Tasks

Xiaosheng Gao ^{1,2}, Shunyao Wu ^{1,*} and Weihua Pan ^{2,*}¹ College of Computer Science and Technology, Qingdao University, Qingdao 266071, China; 2018204621@qdu.edu.cn² State Key Laboratory of Genome and Multi-Omics Technologies, Shenzhen Branch, Guangdong Laboratory for Lingnan Modern Agriculture, Genome Analysis Laboratory of the Ministry of Agriculture and Rural Affairs, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen 518120, China

* Correspondence: wushunyao@qdu.edu.cn (S.W.); panweihua@caas.cn (W.P.)

Abstract

Genomic Language Models (GLMs), leveraging their vast parameter scales and the similarities between DNA sequences and natural languages, demonstrate immense potential in processing large-scale genomic data and elucidating gene regulation and evolutionary relationships. However, the cross-species generalization capability of large genomic models has not yet been systematically evaluated. This study addresses this critical gap by benchmarking five GLMs (DNABERT-2, GROVER, HyenaDNA, NT-V2, and AgroNT) and a CNN baseline model using human (*Homo sapiens*) and rice (*Oryza sativa*) genomes across four downstream tasks: promoter detection, transcription start site (TSS) scanning, species classification, and gene region identification, through both zero-shot testing and fine-tuning. During testing, factors such as hyperparameters, early stopping protocols, and computational resources were fixed to ensure fairness, enabling us to systematically evaluate their performance and cross-species generalization capabilities. The results were further analyzed from multiple mathematical and representational perspectives to provide a more rigorous and objective assessment of each model's performance. The results show that AgroNT consistently leads on rice tasks, while NT-V2 and DNABERT-2 achieved the best overall performance in fine-tuning and zero-shot experiments, respectively. Although their pretraining data did not include plants, they demonstrate excellent performance on rice-related tasks thanks to cross-species pretraining that enhances their generalization ability across human–rice domains. This benchmark study offers guidance on selecting appropriate genomic language models based on task characteristics and provides insights for future development in this field.

Keywords: genomic language models; benchmarking; genomics; cross-species generalization

Academic Editor: Leonel Pereira

Received: 22 December 2025

Revised: 2 February 2026

Accepted: 5 February 2026

Published: 10 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

The rapid advancement of high-throughput sequencing technologies has given rise to an immense volume of genomic data [1,2], creating an urgent need for efficient computational tools to further unravel its biological significance. Numerous studies have demonstrated that the similarities between DNA sequences and natural languages [3–5] endow genome language models (GLMs) with remarkable potential for processing large-scale genomic data and elucidating gene regulation and evolutionary relationships. In recent years, GLMs pretrained on single or multi-species genomic data have successfully captured contextual patterns within DNA sequences, achieving significant progress in

genomic tasks. For example, King et al. (2025) [6] fine-tuned the DNABERT-2 genomic language model and developed a colorectal cancer enhancer classification method using a large-scale dataset of 2.34 million sequences, which demonstrates the feasibility of identifying tumor-associated regulatory signals solely from DNA sequences. Li et al. (2025) [7] proposed HyenaCircle, which combines the HyenaDNA large model with Nanopore long-read sequencing data to predict eccDNA formation in the 1–5 kb range, achieving the first accurate prediction of ultra-long eccDNA in this interval. These cases illustrate the broad application prospects of GLMs, particularly in cancer regulation, mutation interpretation, and non-coding element analysis.

To compare the performance and representational capabilities of models, numerous benchmarking studies on GLMs have already emerged. Marin et al. (2024) [8] introduced the BEND benchmark framework, systematically evaluates six models (AWD-LSTM, Dilated ResNet, Nucleotide Transformer, DNABERT, GENA-LM, and HyenaDNA) across seven datasets covering biologically meaningful tasks, such as gene discovery, enhancer annotation, and chromatin accessibility prediction in biomedical contexts. Awasthi et al. (2025) [9] proposed an unsupervised framework based on numerical linear algebra metrics (RankMe, NESum, and StableRank) to assess the embedding quality of six models (NT, DNA-BERT2, HyenaDNA, Mistral DNA, GENA-LM, and GROVER), with downstream classification tasks validating the correlation between these unsupervised metrics and model performance. Feng et al. (2025) [10] compared five models (DNABERT-2, NT-V2, HyenaDNA, Caduceus, and GROVER) using zero-shot embeddings across multiple genomic tasks, revealing the impact of different model architectures, embedding strategies, and pre-training data on performance. These studies have advanced the understanding of GLMs' capabilities in genomic and biomedical tasks.

However, previous benchmarking studies have primarily compared models across their own independent task settings, and systematic cross-species comparisons under the same task setup remain largely unexplored. This gap is critical because precisely matching pre-training data to downstream tasks is regarded as the core strategy in GLM training. Moreover, genome language models pretrained on multi-species data hold promise for capturing shared genomic patterns across evolutionary distances, yet the specific benefits of such multi-species pre-training in downstream transfer performance still require rigorous evaluation.

To address this gap, this study evaluates the performance and cross-species generalization of GLMs by comparing the human (*Homo sapiens*) and rice (*Oryza sativa*) genomes. The human genome, one of the most comprehensively annotated, is approximately 3 Gb in size and features a complex regulatory network [11]. In contrast, rice, as one of the major staple crops, has a genome of about 400 Mb, characterized by high proportion of repetitive sequences [12]. The significant differences in genome size and regulatory mechanisms between these two species make them representative of typical differences between animal and plant genomes.

To evaluate the model performance from a more comprehensive perspective, this benchmark study evaluated the above tasks using two complementary paradigms: zero-shot testing and full-parameter fine-tuning. The primary objective of this benchmark study is to systematically evaluate the downstream task performance and cross-species generalization capabilities of five GLMs in the context of human and rice genomes, with a particular emphasis on investigating whether multi-species pre-training confers stronger cross-domain transferability than single-species pre-training. Five GLMs (DNABERT-2 [13,14], GROVER [15], HyenaDNA [16], NT-V2 [17], and AgroNT [18]) and a convolutional neural network (CNN) baseline were systematically evaluated across four downstream tasks: promoter detection, transcription start site (TSS) scanning, species classification, and gene region identification. Labels

for these tasks can be directly and reliably obtained from the DNA sequence itself and high-quality genome annotations; at the same time, these tasks span a gradient of scales from local motif recognition and gene structure parsing to cross-species discrimination, thereby comprehensively evaluating the model's ability to capture DNA sequence structure. Experiments utilized datasets sourced from Phytozome [19] and Ensembl [20], with fine-tuning performed on 8 NVIDIA RTX 4090 GPU and comprehensive evaluation conducted using multiple metrics to ensure fairness. To ensure the fairness of the benchmarking and to guarantee that the final task performance truly reflects the intrinsic quality of the models' pre-training rather than differences in downstream optimization, all experiments adopted the same fixed set of hyperparameters, including batch size, learning rate, early stopping criteria.

Experimental results demonstrate that AgroNT consistently achieves superior performance on rice tasks, underscoring the effectiveness of its pretraining on 48 plant species. DNABERT-2 and NT-V2 attained the highest average Matthews Correlation Coefficient (MCC) under zero-shot and fine-tuning conditions, respectively. This exceptional performance can be attributed to their multi-species pretraining and architectural improvements in positional encoding within Transformer models. Despite the absence of plant data in the pretraining corpora of NT-V2 and DNABERT-2, their multi-species pretraining enables the learning of universal sequence patterns across human and rice, facilitating adaptation to plant-specific tasks through fine-tuning and yielding performance approaching AgroNT's optimum across the four rice tasks. The poor performance of the non-pretrained CNN baseline highlights the critical importance of transfer learning. This study highlights the generalization capabilities of GLMs in cross-species tasks, providing insights for optimizing multi-species pretraining and task-specific fine-tuning strategies.

2. Methods

This study presents a systematic benchmarking of five genomic language models (DNABERT-2, GROVER, HyenaDNA, NT-V2, and AgroNT) alongside a convolutional neural network (CNN) baseline, focusing on cross-species generalization between human (*Homo sapiens*) and rice (*Oryza sativa*) genomes. All experiments were conducted under a standardized and fixed setting to ensure fairness, reproducibility, and objectivity in model comparison, with details provided in subsequent subsections. The overall workflow is illustrated in Figure 1.

Each model was evaluated on four representative downstream tasks that collectively assess different aspects of genomic understanding: Promoter detection, Transcription Start Site (TSS) Scanning, Species Classification and Gene Region Identification. Except for the species classification task which involves multiple species, all other experiments were conducted on rice (*Oryza sativa*) and human (*Homo sapiens*) genomes. All data were extracted from genome FASTA files and their corresponding GFF3 annotation files using industry-standard methods, with information from the annotation files serving as the ground truth for benchmark evaluation metrics. During model evaluation, we used scikit-learn [21] library functions to compute all metrics.

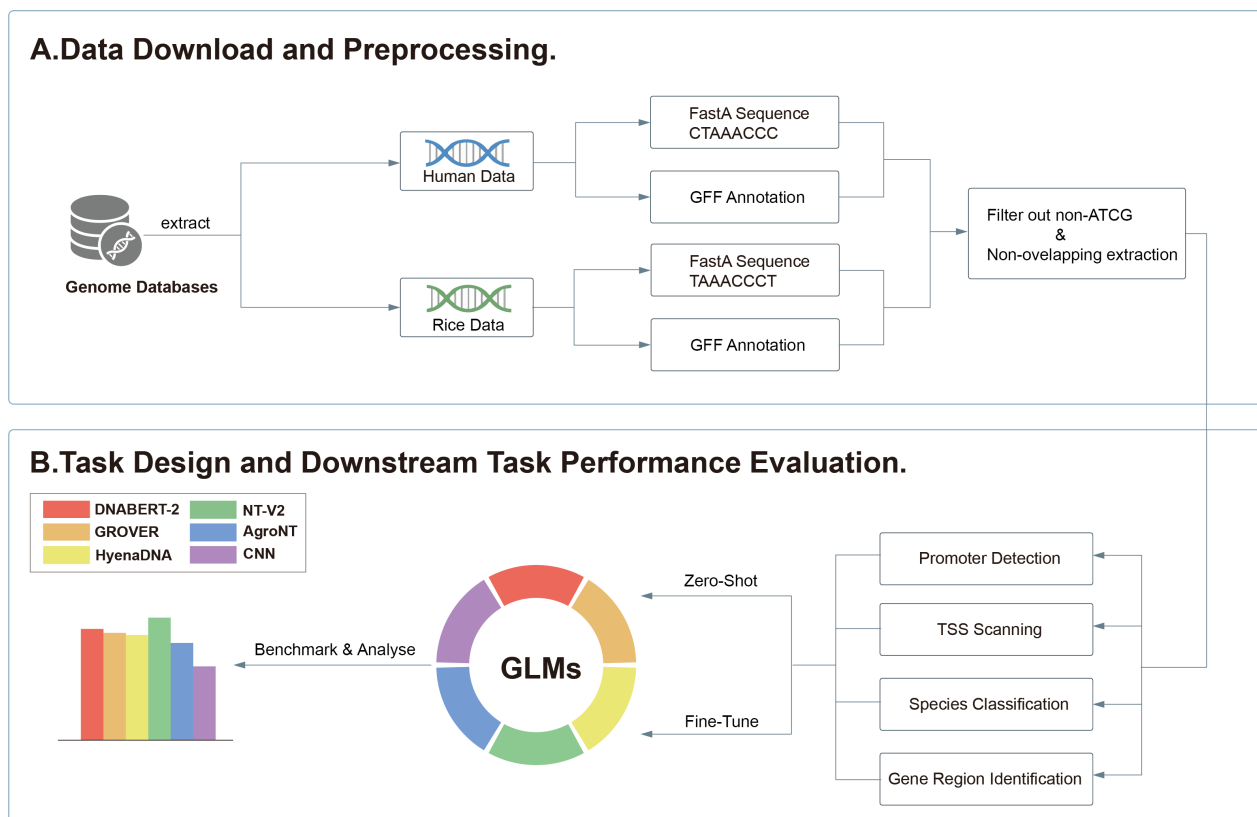


Figure 1. Overall Workflow Diagram. (A) Data Download and Preprocessing. Genomic sequences and annotation information for human and rice are extracted from genome databases. FASTA sequences are combined with GFF annotation files for processing. Non-ATCG characters are filtered out, and non-overlapping sequence extraction is performed. (B) Task Design and Downstream Task Performance Evaluation. Multiple DNA genomic language models (Genomic Language Models, GLMs) are compared, including DNABERT-2, GROVER, NT-V2, AgroNT, HyenaDNA, and a CNN baseline model. Performance is evaluated under two settings, zero-shot (Zero-Shot) and fine-tuning, across the following 4 tasks: Promoter Detection, Transcription Start Site (TSS) Scanning, Species Classification and Gene Region Identification. Bar charts show the comparative performance of each model on benchmark tasks, while a circular diagram indicates the GLMs included in the evaluation.

2.1. Models

This benchmark study evaluated five genomic language models (GLMs) and a convolutional neural network (CNN) baseline to assess their performance and efficiency in processing genomic data. Each model processes an input DNA sequence $S = (s_1, s_2, s_3, \dots, s_L)$, where $s_i \in \{A, T, C, G\}$, and L is the sequence length. Through a composition of tokenization $\rightarrow$ embedding $\rightarrow$ contextual encoding $\rightarrow$ classification head, the model maps S to a task-specific output y .

2.1.1. Tokenization Strategies

The above models employ three tokenization strategies to process the input S . Each strategy corresponds to a pre-trained vocabulary V , which converts the S into a discrete token sequence $t = (t_1, t_2, t_3, \dots, t_M)$, where $M \leq L$.

Byte Pair Encoding (BPE)

Byte-Pair Encoding (BPE) [22] constructs a vocabulary V_{BPE} by iteratively merging the most frequent adjacent symbol pairs in the training corpus. The initial symbol set is $\Sigma = \{A, T, C, G\}$. Formally, BPE learns a sequence of merge rules $R = \{(p_1, q_1), \dots\}$, where each pair (p_1, q_1) denotes merging tokens p_1 and q_1 into a new token; this merging

process continues until the vocabulary reaches a predefined size. During the inference stage, to generate t , starting from S and the initial symbol set Σ , merge operations are greedily applied in the order defined by R , until the predefined vocabulary size is reached, thereby yielding the final subword segmentation as Equation (1).

$$t \in \{t' \in \mathcal{V}_{BPE}^* \mid \text{concat}(t') = S\} \quad (1)$$

Equation (1) states that t is any valid subword segmentation from $\mathcal{V}_{BPE}$ whose concatenation reconstructs S .

BPE is adopted by models such as GROVER and DNABERT-2. It iteratively merges the most frequently co-occurring adjacent segments in the training corpus, often encoding common conserved sequence motifs into a single token or a very small number of tokens.

K-Mer Tokenization

The vocabulary for k-mer tokenization is the set of all possible k-mers, denoted as $V_{k-mer} = \{s_1 s_2 \dots s_k \mid s_i \in \{A, T, C, G\}\}$, with a vocabulary size of 4^k . During the inference stage, to generate the t , starting from S and V_{k-mer} , a fixed-length sliding window of size k is applied to segment S into overlapping or non-overlapping k-mers (in this benchmark study, all k-mer models use non-overlapping segmentation with $k = 6$). The resulting token sequence as shown in Equation (2).

$$t = (t_1, t_2, \dots, t_M), t_j = (S_{j \cdot k - k + 1}, S_{j \cdot k - k + 2}, \dots, S_{j \cdot k}), M = \left\lfloor \frac{L}{k} \right\rfloor \quad (2)$$

Equation (2) defines fixed-length, non-overlapping k-mer tokens. The j -th token contains bases from position $j \cdot k - k + 1$ to $j \cdot k$. If L is not divisible by k , the trailing fragment that is less than k in length is truncated.

K-mer tokenization is adopted by models such as NT-V2 and AgroNT. It achieves an effective balance between token compression ratio and sequence resolution, while preserving sufficient local resolution and offering high computational efficiency. This approach also avoids the additional overhead associated with BPE, including vocabulary imbalance and the learning of merge rules.

Single-Nucleotide Tokenization

Single-nucleotide tokenization treats each base as an independent token, with the vocabulary $V_{single} = \{A, T, C, G\}$, formally, given a sequence $S = (s_1, s_2, s_3, \dots, s_L)$. The resulting token sequence as shown in Equation (3).

$$t = (t_1, t_2, t_3, \dots, t_M), t_i = s_i, M = L \quad (3)$$

Equation (3) maps each nucleotide in S to a unique token.

Single-nucleotide tokenization is applied in models such as HyenaDNA. It preserves all local information and exhibits high sensitivity to single nucleotide polymorphisms.

The three tokenization strategies introduced above are selectively adopted by different GLMs during pretraining and inference. The following provides a detailed description of the models included in this benchmark study.

GROVER: GROVER employs BPE and Masked Language Modeling (MLM) for pre-training on the human genome, specifically designed to capture human-specific genomic patterns. It is a Transformer-based DNA language model that learns the probability distribution of bases in genomic sequences through an attention mechanism [23], capturing contextual patterns in the genome. This is mathematically expressed as Equation (4).

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (4)$$

Equation (4) calculates the multi-head self-attention, where Q , K , and V are the query, key, and value matrices, respectively, each of size $\mathbb{R}^{L \times d_k}$. These matrices are projected from the input embeddings, and d_k denotes the dimensionality of each attention head. $\sqrt{d_k}$ is used to normalize the dot products of the queries and keys, which helps stabilize the gradients during training.

DNABERT-2: Like GROVER, DNABERT-2 employs BPE and MLM for pretraining, but it uses a multi-species genomic dataset. It incorporates Attention with Linear Biases (ALiBi) [24] and Flash Attention [25] to optimize computational efficiency, enabling efficient processing of long genomic sequences. For S of length L , the attention score of the i -th query q_i and the key matrix K in Attention with Linear Biases (ALiBi) is calculated as Equation (5).

$$\text{Attention}_{\text{ALiBi}}(q_i, K, V) = \text{softmax}\left(\frac{q_i K^T}{\sqrt{d_k}} + m \cdot b_i\right) V, \quad (5)$$

$$b_i = [-(i-1), \dots, -1, 0, -1, \dots, -(L-1-i)]$$

Equation (5) modifies the standard attention by adding head-specific linear biases $m \cdot b_i$, where m is a head-specific constant from a geometric sequence, $q_i \in \mathbb{R}^{d_k}$ is the i -th query, and $K, V \in \mathbb{R}^{L \times d_k}$ are the key and value matrices, respectively. The slope m is learned for each attention head. The bias vector b_i imposes a distance-dependent penalty: past tokens are assigned negative values, the token itself is zero, and future tokens are positive. This forces the attention to decay linearly with distance, allowing extrapolation to longer sequences not seen during training without the need for positional encodings.

AgroNT: AgroNT is a Transformer-based DNA language model tailored for edible plant genome research. As a plant-specific version of the Nucleotide Transformer, it adopts a similar architecture, utilizing MLM for pretraining on genomic DNA from 48 plant species. DNA sequences are segmented into 6-mers to capture plant-specific genomic patterns.

NT-V2: NT-V2 is an improved version of the Nucleotide Transformer (NT), based on the Transformer architecture and pre-trained on multi-species genomes using MLM. NT-V2 employs 6-mer tokenization and combines Rotary Position Embeddings (RoPE) [26] and Gated Linear Units (GLU) with Swish activation to enhance computational efficiency. RoPE incorporates positional information into embeddings at each position i through rotation matrices, applied at every attention layer, enabling robust contextual modeling for long sequences (up to 2048 tokens, approximately 12 kb). This can be expressed as Equation (6).

$$\text{RoPE}(x_i, \theta) = x_i \cdot R_{\theta, m}, \quad R_{\theta, m} = \begin{bmatrix} \cos(m\theta) & -\sin(m\theta) \\ \sin(m\theta) & \cos(m\theta) \end{bmatrix} \quad (6)$$

Equation (6) encodes the absolute position i by applying a 2D rotation matrix $R_{\theta, m}$ to each pair of dimensions $(x_{i, 2m}, x_{i, 2m+1})$ of the embedding $x_i \in \mathbb{R}^d$, where θ is the base frequency and $m = 0, 1, \dots, d/2 - 1$ indexes the dimension pairs. The core advantage of this rotation scheme is that the relative distance $j - i$ between positions i and j naturally manifests as a fixed rotation angle in the attention inner product, independent of their absolute positions, thereby achieving true relative position encoding without explicit distance bias terms.

HyenaDNA: HyenaDNA utilizes the Hyena operator [27], designed specifically for modeling long genomic sequences. Unlike traditional self-attention mechanisms, it uses implicit convolution as Equation (7).

$$y_t = \sum_{s=1}^t h_{t-s} x_s y_t \quad (7)$$

Equation (7) calculates the output y_t at position t as the weighted sum of all previous inputs $x_s (s = 1, \dots, t)$ and a learnable filter h (with effective length up to t), efficiently

implemented via a structured state space model. This implicit convolution reduces the time complexity from $O(L^2)$ to $O(L \log L)$.

CNN: The Convolutional Neural Network (CNN) consists of three convolutional layers and serves as a non-pre-trained baseline to evaluate the advantages of pretraining. It uses single-nucleotide tokenization, requires low computational resources, and does not undergo pretraining. The CNN baseline architecture remains fixed across all tasks.

2.2. Dataset

Genomic FASTA files were downloaded from established genomic databases. Plant genomic data, including sequences and annotations for *Hordeum vulgare*, *Oryza sativa*, *Sorghum bicolor*, *Triticum aestivum*, and *Zea mays*, were sourced from Phytozome V12 unrestricted, a comprehensive plant comparative genomics resource. Human and mammalian data (*Homo sapiens*, *Bos taurus*, *Canis lupus familiaris*, *Mus musculus*, *Sus scrofa*) were obtained from Ensembl Release 114, ensuring high-quality annotations.

The preprocessing pipeline involved removing any information beyond the model's vocabulary (retaining only A, T, C, G nucleotides) to align with the tokenization requirements and prevent tokenization errors. For the gene region identification tasks, since sequences were extracted across the entire genome, data were sourced from Hardmasked FASTA files downloaded from the database to mitigate the impact of highly abundant repetitive sequences. To ensure robust evaluation, datasets for all tasks were carefully balanced. During training, the datasets were split into 80% training set, 10% test set, and 10% validation set; the same split was used for both zero-shot testing and fine-tuning experiments. Additionally, to prevent overlapping or adjacent sequences from appearing across different splits, a minimum distance of greater than 2000 bp between any two extracted sequences was enforced during random sampling.

2.3. Tasks

This study designs four downstream tasks to systematically evaluate model classification performance on genomic sequences. All tasks are constructed from high-quality annotated FASTA and GFF3 files.

2.3.1. Promoter Detection

This binary classification task aims to identify promoter sequences. Promoters are regulatory DNA regions that initiate transcription, being crucial for understanding gene regulatory networks. The task space is defined as $y = \{0, 1\}$, where 0 and 1 represent negative and positive samples, respectively. In this task, positive samples comprise sequences of length $L = 500$ bp immediately upstream of the transcription start site (TSS). This length is referenced from the promoter region definition used in DNABERT; negative samples are extracted from regions > 5000 bp upstream of any TSS to avoid overlap with potential regulatory elements, with $L = 500$ bp. Each category contains 20,000 sequences. This task evaluates models' ability to detect short sequence motifs with regulatory signals and measures their sensitivity in distinguishing functional from non-functional genomic regions.

2.3.2. Transcription Start Site (TSS) Scanning

This binary classification task aims to determine whether a sequence contains a transcription start site (TSS), which initiates transcription and defines gene boundaries. The task space is defined as $y = \{0, 1\}$, where 0 and 1 represent negative and positive samples, respectively. In this task, positive samples are sequences of length $L = 1024$ bp with a TSS randomly placed internally; negative samples are $L = 1024$ bp fragments from regions devoid of any TSS. The 1024 bp length strikes a balance between capturing local

sequence features and providing sufficient global contextual information for modeling complex transcriptional regulatory patterns. This task evaluates a model's ability to detect functional sites within longer sequences while measuring its capacity to capture positional dependencies and contextual signals inherent in genomic data.

2.3.3. Species Classification

This multi-class classification task aims to assign genomic sequences of length $L = 1024$ bp to their source species, supporting comparative genomics, phylogenetic inference, and species-specific adaptation studies. The task space is defined as $y = \{1, 2, 3, 4, 5\}$, where 1–5 represent five categories of plants and mammals, respectively. Plant data comprised five Poaceae species: barley (*Hordeum vulgare*), rice (*Oryza sativa*), wheat (*Triticum aestivum*), maize (*Zea mays*), and sorghum (*Sorghum bicolor*); mammalian data encompassed humans (*Homo sapiens*), cattle (*Bos taurus*), dogs (*Canis lupus familiaris*), mice (*Mus musculus*), and pigs (*Sus scrofa*). Each species contributed 20,000 randomly sampled sequences of length $L = 1024$ bp; this length captures local motifs and lineage-specific compositional biases. The initial design drew upon Mamba's [28] species classification task using humans and four great apes; however, due to extremely high sequence similarity, accuracy approached random guessing. A species ensemble exhibiting greater genomic divergence was ultimately selected to enhance task discriminability. This benchmark assesses models' ability to detect species-specific sequence patterns and generalize across diverse genomic architectures.

2.3.4. Gene Region Identification

This multi-class classification task aims to distinguish introns, exons, and intergenic regions within a single genome. The task space is defined as $y = \{1, 2, 3\}$, representing introns, exons, and intergenic regions respectively. This constitutes a fundamental step in genome annotation and gene structure analysis. Accurate classification of coding exons versus introns and intergenic non-coding DNA facilitates research into gene function and regulatory mechanisms. Sequence lengths L uniformly sampled from 512 to 2048 bp and the window is randomly positioned within the target region to capture mid-range compositional biases and splice-site proximity signals. Each category contains 20,000 sequences to ensure balanced representation. This task evaluates models' ability to distinguish genomic regions with differing functional roles and measures their capacity to model long-range dependencies and complex sequence contexts.

2.4. Zero-Shot Testing

To assess the intrinsic representational capabilities of the pre-trained genomic language models without any task-specific adaptation, we conducted zero-shot testing across all four downstream tasks. For each model, we first extracted sequence-level embeddings from the frozen model using mean pooling over the last hidden states, with attention mask-based weighting to exclude padding tokens. This mean-pooling strategy was selected due to its consistent superior performance in recent DNA foundation model benchmarks for sequence classification tasks [10]. To prevent label leakage, embeddings for the training set and test set were extracted in two separate passes.

The resulting embeddings were then used to train a simple random forest classifier on the training set embeddings and corresponding labels. Model performance was evaluated on the held-out test set embeddings using standard metrics. No parameters of the foundation models were updated during this process. The hyperparameters of the random forest classifier are detailed in Supplementary Table S17.

2.5. Fine-Tuning

Across all tasks and models, we adopted a globally fixed hyperparameter configuration to ensure fairness and objectivity in model comparison. Specifically, a uniform batch size of 16 and a learning rate of 2×10^{-5} were fixed for all models. A consistent early stopping strategy was applied: after each epoch, the F1-score on the validation set was computed, and training was stopped if the F1-score showed no improvement over three consecutive epochs. All experiments were conducted on 8 NVIDIA RTX 4090 GPUs from NVIDIA Corporation (Santa Clara, CA, USA) with identical training setups.

This fixed configuration prioritizes a fair and reproducible comparison of the models' pre-trained representation capabilities enhanced by fine-tuning, rather than introducing additional advantages through model-specific hyperparameter tuning or learning-rate sweeps.

3. Result

This study comprehensively evaluated six models—: DNABERT-2, HyenaDNA, GROVER, NT-V2, AgroNT, and a CNN baseline—across four downstream tasks: promoter detection, TSS scanning, species classification, and gene region identification.

Figure 2 shows the zero-shot performance on human (*Homo sapiens*) and rice (*Oryza sativa*) genomes, with average performance comparison between the two species. Figure 3 depicts the fine-tuned performance and Tables 1 and 2 reported the corresponding human–rice average comparison. Table 3 summarizes the average memory usage of each model across all tasks. Supplementary Tables S1–S16 present the performance metrics, including accuracy, F1-score, recall, precision, and MCC, for each model across these tasks. Supplementary Figures S1–S16 present the confusion matrices of each model on each task in the form of heatmaps.

Table 1. Specific Matthews Correlation Coefficient (MCC) values of each model for each downstream task (Zero-Shot Testing). Bold text highlights the model with the highest MCC performance per task.

	DNABERT-2	GROVER	HyenaDNA	NT-V2	AgroNT	CNN
Promoter Detection (Plant)	0.4483	0.4239	0.4246	0.4295	0.4868	0.5675
Promoter Detection (Human)	0.6692	0.7094	0.6232	0.6333	0.6274	0.6782
TSS Scanning (Plant)	0.5488	0.5147	0.3325	0.4957	0.5749	0.2788
TSS Scanning (Human)	0.8354	0.8064	0.7175	0.7837	0.7810	0.6937
Species Classification (Plant)	0.3573	0.3546	0.3074	0.3392	0.5074	0.1291
Species Classification (Human)	0.5266	0.3686	0.3291	0.4393	0.3108	0.1707
Gene Region Identification (Plant)	0.6980	0.6871	0.6649	0.6750	0.7171	0.6094
Gene Region Identification (Human)	0.6088	0.5961	0.4949	0.5438	0.5641	0.4584
AVG	0.5866	0.5576	0.4868	0.5424	0.5712	0.4482

Table 2. Specific Matthews Correlation Coefficient (MCC) values of each model for each downstream task (Fine-Tuning). Bold text highlights the model with the highest MCC performance per task.

	DNABERT-2	GROVER	HyenaDNA	NT-V2	AgroNT	CNN
Promoter Detection (Plant)	0.6381	0.6138	0.5659	0.6606	0.6936	0.5675
Promoter Detection (Human)	0.7408	0.7098	0.6766	0.7407	0.6907	0.6782
TSS Scanning (Plant)	0.6699	0.6197	0.3772	0.7007	0.6898	0.2788
TSS Scanning (Human)	0.8793	0.8496	0.7390	0.9065	0.8761	0.6937
Species Classification (Plant)	0.6417	0.6154	0.5446	0.6376	0.7463	0.1291
Species Classification (Human)	0.6749	0.6099	0.5694	0.6805	0.6159	0.1707

Table 2. *Cont.*

	DNABERT-2	GROVER	HyenaDNA	NT-V2	AgroNT	CNN
Gene Region Identification (Plant)	0.8318	0.8231	0.8090	0.8263	0.8685	0.6094
Gene Region Identification (Human)	0.6658	0.6376	0.6254	0.6945	0.6382	0.4584
AVG	0.7178	0.6849	0.6134	0.7309	0.7274	0.4482

Table 3. GPU memory consumption (in MiB) of each model on each task.

	DNABERT-2	GROVER	HyenaDNA	NT-V2	AgroNT
Promoter Detection (Plant)	4390	4550	3002	10,198	50,322
Promoter Detection (Human)	4102	4018	3002	10,198	50,322
TSS Scanning (Plant)	5796	7540	4886	13,844	76,662
TSS Scanning (Human)	5870	7418	4886	13,844	76,662
Species Classification (Plant)	5816	11,898	4886	13,844	76,662
Species Classification (Human)	5816	11,898	4886	13,844	76,662
Gene Region Identification (Plant)	8192	12,274	7978	23,850	91,678
Gene Region Identification (Human)	8454	12,274	7978	23,850	91,678
AVG	6055	8984	5188	15,434	73,831

In zero-shot testing, DNABERT-2 achieved the highest average MCC of 0.5866 across the eight subtasks, followed by AgroNT (0.5712) and GROVER (0.5576). NT-V2 (0.5424) and HyenaDNA (0.4868) performed relatively weaker. Among them, DNABERT-2 excelled particularly on human-related tasks, while AgroNT stood out on plant-related tasks, achieving the highest MCC in three subtasks (TSS scanning: 0.5749; species classification: 0.5074; gene region identification: 0.7171). GROVER attained the highest MCC (0.7094) on one human task (promoter prediction). Overall, zero-shot MCC values fluctuated between 0.3 and 0.8. Compared to random guessing, all models demonstrated positive discriminative ability, indicating that pre-trained embeddings are already able to capture some cross-species shared genomic patterns, although there remains a noticeable gap from the level required for direct practical application.

Fine-tuning significantly improved the performance of all models on most tasks. Compared with zero-shot results, the average MCC gain across models ranged from 0.13 to 0.19. Among them, NT-V2 exhibited the largest improvement and became the model with the highest average MCC after fine-tuning (0.7309), closely followed by AgroNT (0.7274) and DNABERT-2 (0.7178). AgroNT continued to show strong advantages on plant-related tasks, achieving the highest MCC in three subtasks (promoter prediction: 0.6936; species classification: 0.7463; gene region identification: 0.8685). It particularly excelled in grass family (Poaceae) species classification and gene region identification tasks, highlighting the capability of its pre-training on 48 plant genomes in capturing phylogenetic adaptation and understanding plant genomic architecture.

These results demonstrate that task-specific fine-tuning can fully unlock the potential of multi-species and species-specific pre-training. While pre-trained models already possess a certain degree of cross-species transferability in the zero-shot setting, task-specific fine-tuning remains a critical step to maximally release the power of multi-species and species-specialized pre-training.

Overall, performance disparities among models and between human and plant genomes on identical tasks stem from multiple factors, including pretraining data, model architecture and scale, and genomic structure.

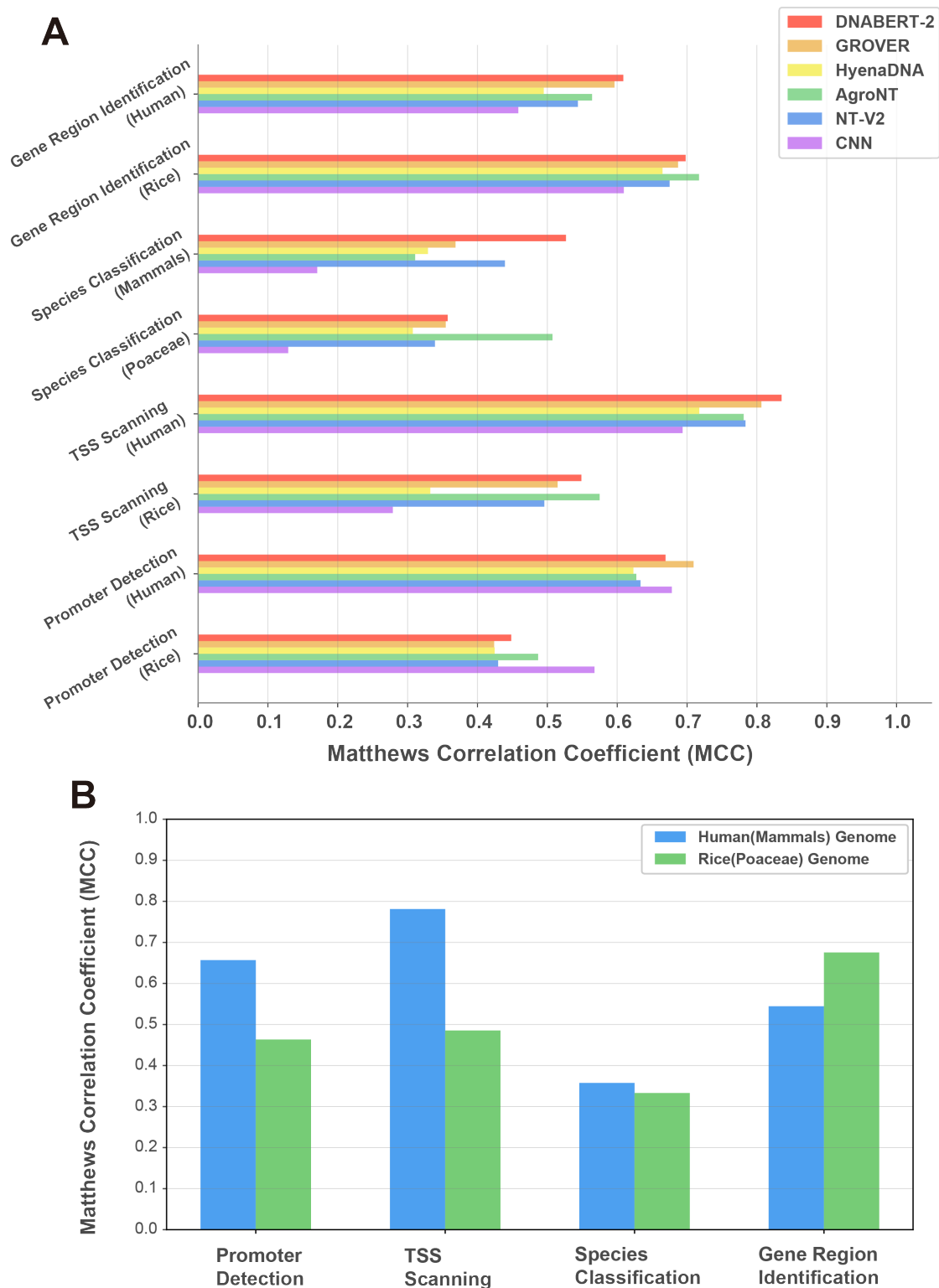


Figure 2. Performance of six models on downstream tasks for human (Mammals) and rice (Poaceae) genomes (Zero-Shot Testing). (A) Matthews Correlation Coefficient (MCC) of six models across four downstream tasks on human (*Homo sapiens*) and rice (*Oryza sativa*) genomes. species classification task includes additional mammalian and Poaceae genomes, hence group-level labels are used. Each subplot displays the MCC values for different models on specific tasks, reflecting their predictive performance on these tasks. (B) Comparison of average MCC across models for human and rice genomes, highlighting overall performance differences across the four downstream tasks.

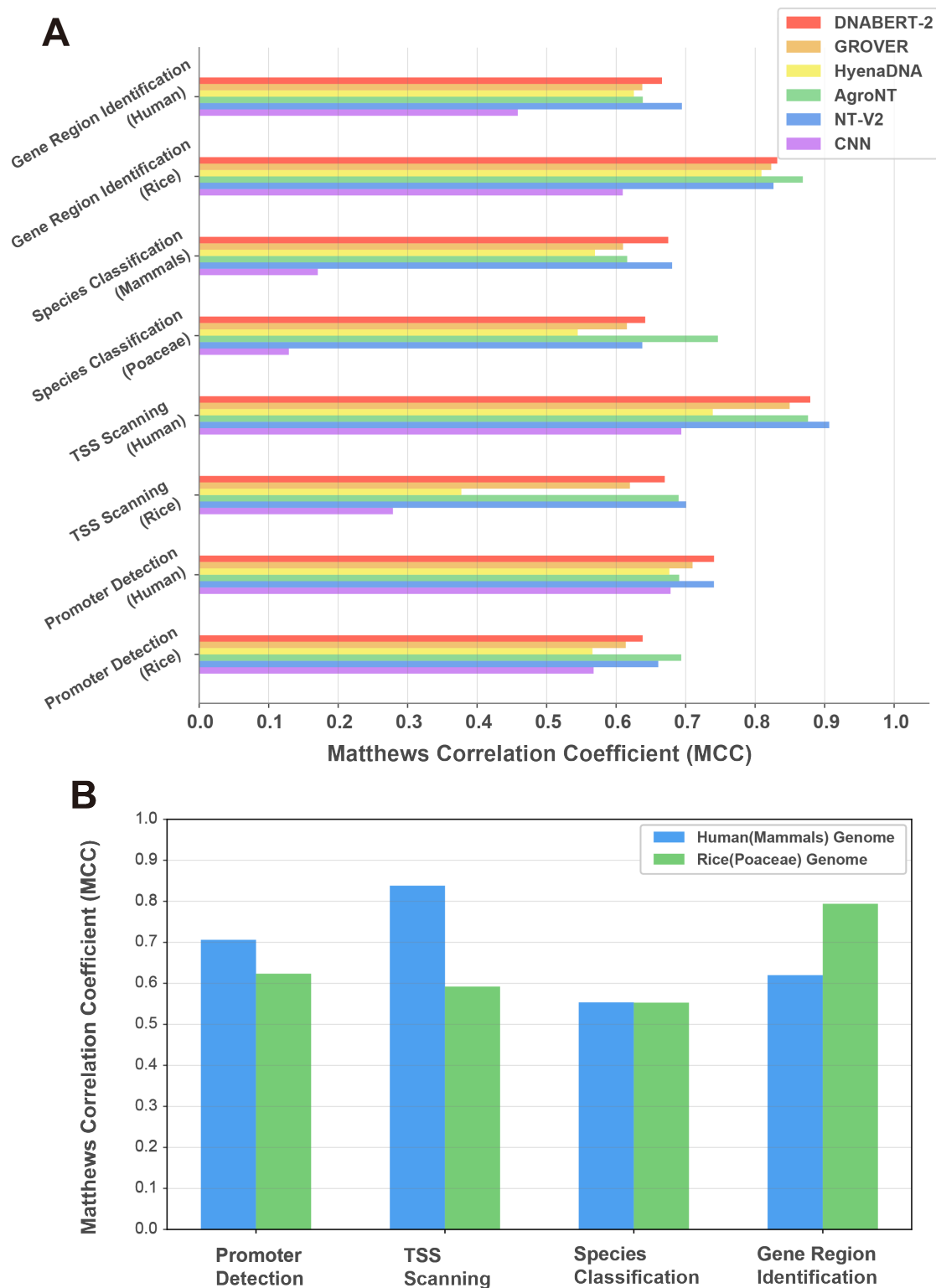


Figure 3. Performance of six models on downstream tasks for human (Mammals) and rice (Poaceae) genomes (Fine-Tuning). **(A)** Matthews Correlation Coefficient (MCC) of six models across four downstream tasks on human (*Homo sapiens*) and rice (*Oryza sativa*) genomes. Species classification task includes additional mammalian and Poaceae genomes, hence group-level labels are used. Each subplot displays the MCC values for different models on specific tasks, reflecting their predictive performance on these tasks. **(B)** Comparison of average MCC across models for human and rice genomes, highlighting overall performance differences across the four downstream tasks. Specific Matthews Correlation Coefficient (MCC) values of each model for each downstream task.

3.1. Performance Analysis

3.1.1. Pretraining Data and Generalization

The pre-training dataset determines the coverage of a model's initial embedding space, which in turn influences its generalization performance on the distribution of downstream tasks. In this benchmark, the involved models can be classified into three categories based on their pre-training datasets: GROVER and HyenaDNA, which were pre-trained on a single species; AgroNT, which was pre-trained on multiple species within a single domain (Edible Plants-related); NT-V2 and DNABERT-2, which were pre-trained across multiple domains and multiple species. Tables 4 and 5 respectively report the MCC ranking of each model on each task under zero-shot and fine-tuning settings.

Table 4. Zero-shot testing performance ranking of each model on each task.

	DNABERT-2	GROVER	HyenaDNA	NT-V2	AgroNT	CNN
Promoter Detection (Plant)	3	6	5	4	2	1
Promoter Detection (Human)	3	1	6	4	5	2
TSS Scanning (Plant)	2	3	5	4	1	6
TSS Scanning (Human)	1	2	5	3	4	6
Species Classification (Plant)	2	3	5	4	1	6
Species Classification (Human)	1	3	4	2	5	6
Gene Region Identification (Plant)	2	3	5	4	1	6
Gene Region Identification (Human)	1	2	5	4	3	6

Table 5. Fine-tuning performance ranking of each model on each task.

	DNABERT-2	GROVER	HyenaDNA	NT-V2	AgroNT	CNN
Promoter Detection (Plant)	3	4	6	2	1	5
Promoter Detection (Human)	1	3	6	2	4	5
TSS Scanning (Plant)	3	4	5	1	2	6
TSS Scanning (Human)	2	4	5	1	3	6
Species Classification (Plant)	2	4	5	3	1	6
Species Classification (Human)	2	4	5	1	3	6
Gene Region Identification (Plant)	2	4	5	3	1	6
Gene Region Identification (Human)	2	4	5	1	3	6

The results from the two groups of experimental tasks demonstrate that multi-domain and multi-species pre-trained models (DNABERT-2 and NT-V2) achieved the highest MCC in three tasks under both zero-shot and fine-tuning settings. Compared to single-species pre-trained models, they exhibited superior cross-species generalization after fine-tuning (DNABERT-2 and NT-V2 ranked in the top three in MCC across all tasks; although DNABERT-2's average MCC was slightly lower than that of AgroNT, its overall performance was more stable). Notably, despite the complete absence of plant genomes in the pre-training data of these two models, they still outperformed single-species human-pre-trained models on rice (*Oryza sativa*) related tasks. In particular, NT-V2 even surpassed AgroNT in the rice TSS scanning task. This superior performance likely stems from the diverse genomic corpora in their multi-species pre-training datasets, which collectively construct a broader and more representative embedding space encompassing sequence

evolutionary patterns from prokaryotes to eukaryotes and from lower to higher organisms, thereby significantly reducing the risk of overfitting to a single species.

To further visually illustrate the influence of pre-training data on the models' embedding spaces, this benchmark extracted embeddings from five models using 50,000 randomly sampled non-overlapping 1024 bp fragments each from the human and rice genomes. These embeddings were subsequently reduced to two-dimensional space via UMAP for visualization, as shown in Figure 4. The multi-species pre-trained models NT-V2 and DNABERT-2 were able to almost completely separate the two clusters of points, indicating that their embedding spaces successfully captured distinguishable deep cross-species patterns. AgroNT, as a single-domain multi-species pre-trained model, could also distinguish human from rice DNA reasonably well, but still exhibited substantial overlapping regions. In contrast, GROVER and HyenaDNA, which were pretrained solely on the human genome, showed highly overlapping point clouds for the two species, highlighting the cross-species generalization limitations inherent to single-species pre-training.

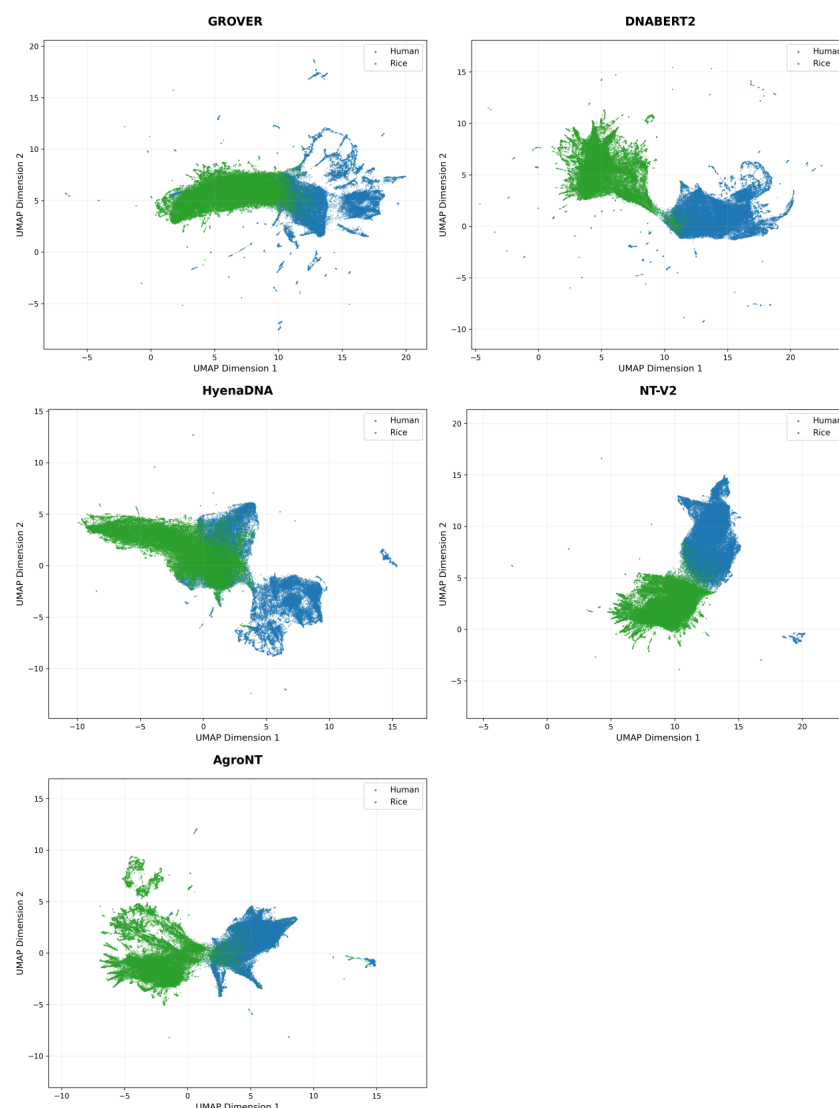


Figure 4. Comparison of Species Discriminability in the Embedding Spaces of 5 DNA Language Models. Embeddings were extracted from 50,000 randomly sampled non-overlapping 1024 bp DNA fragments each from the human and rice genomes, followed by UMAP dimensionality reduction to a two-dimensional projection. Blue points represent human DNA sequence fragments, and green points represent rice (*Oryza sativa*) DNA sequence fragments.

3.1.2. Model Architecture and Scale

Model architecture, parameter scale, and related attributes are also key factors influencing final performance. Table 6 presents the attributes of each model, including pretraining data, model architecture, tokenization strategies, etc.

Table 6. Overview of the genomic language models evaluated in this benchmark study. This table lists the tokenizer, parameter count, pretraining method, pretraining data, and model architecture for five models (DNABERT-2, GROVER, HyenaDNA, NT-V2, AgroNT). M, million.

	Tokenizer	Parameter	Pre-Train	Data	Architecture
DNABERT-2	BPE	120 M	MLM	Multi-Species	Transformer
GROVER	BPE	87 M	MLM	Human	Transformer
HyenaDNA	Single Nucleotide	1 M	NTP	Human	Hyena
NT-V2	6-mer	498 M	MLM	Multi-Species	Transformer
AgroNT	6-mer	985 M	MLM	48 Plants	Transformer

In this benchmark study, DNABERT-2, GROVER, NT-V2, and AgroNT are based on the Transformer architecture. This architecture leverages its core multi-head self-attention mechanism to effectively capture complex contextual dependencies throughout the entire sequence, which is particularly important for identifying dispersed regulatory elements. However, traditional positional encodings in Transformer models (such as fixed absolute positional encodings or sinusoidal/cosine encodings) have a fixed maximum position index during training. Once the input sequence length exceeds the pretraining length, frequency mismatch causes the attention distribution to become random, making stable length extrapolation difficult. Consequently, these models often struggle to reliably handle genomic data significantly longer than sequences seen during pretraining—a common scenario in genomic tasks.

To address this issue, DNABERT-2 and NT-V2 have introduced different optimizations to the standard Transformer architecture. NT-V2 incorporates Rotary Positional Embeddings (RoPE) and Gated Linear Units (GLU) with Swish activation. By applying frequency-based complex exponential rotation matrices to the query and key vectors, relative positional information is multiplicatively injected into the attention scores, making attention weights depend solely on the relative distance between positions and enabling efficient processing of sequences longer than those seen during pretraining. In contrast, DNABERT-2 adopts Attention with Linear Biases (ALiBi), which is mathematically equivalent to introducing a triangular bias that decays linearly with distance in the attention kernel. This approach achieves stable extrapolation to arbitrary sequence lengths without any learnable or fixed positional encodings and offers faster computation than RoPE. The excellent performance of these two models in this benchmark may be closely related to these architectural improvements.

A noteworthy phenomenon is the significant performance divergence between DNABERT-2 and NT-V2, both multi-species pretrained models. In zero-shot testing, DNABERT-2 achieved the highest average MCC, whereas NT-V2 showed only moderate to below-average performance in zero-shot settings but exhibited the largest gains after fine-tuning, surpassing all other models to become the top performer in average MCC.

This divergence may be attributed to differences in pretraining design. DNABERT-2 has a smaller parameter count and shorter pretraining sequence lengths, potentially focusing more on local motifs, but with relatively limited room for improvement during fine-tuning. NT-V2, with a larger parameter count and a context length four times that of DNABERT-2, learns richer and more distributed representations during pretraining. These representations are difficult to fully utilize through simple average pooling in zero-shot

testing and require task-specific fine-tuning to activate. Overall, DNABERT-2 excels in out-of-the-box usability, while NT-V2's design is more conducive to achieving greater improvements through fine-tuning, reflecting the trade-off between immediate generalization performance and adaptation potential in multi-species pretraining.

AgroNT has the largest parameter count in this benchmark (985 M) and consistently maintains advantages in plant tasks during both zero-shot testing and fine-tuning, with its human-task performance also showing certain improvements after fine-tuning.

Unlike the models above, HyenaDNA fundamentally modifies the attention mechanism structure by replacing explicit dot-product attention with the Hyena operator based on implicit long convolutions. It equivalently performs global convolution through parameterized implicit long filters (implemented via structured state-space recurrence), reducing the original $O(n^2)$ matrix computation complexity of attention to $O(n)$ or quasi-linear levels. This provides HyenaDNA with a significant computational advantage when processing ultra-long sequences (up to 1 M tokens), at which point training times for Transformer-based architectures typically become prohibitive. However, with only 1 M parameters, its representational capacity is severely limited, which likely explains its suboptimal performance on the relatively shorter sequence tasks examined in this study.

Finally, as a control, the non-pretrained CNN baseline model is trained from scratch using only task-specific data. Due to its limited receptive field, it struggles to capture the long-range dependencies commonly present in genomic sequences, resulting in significantly inferior performance compared to all pretrained models across all six tasks.

However, this baseline model still achieves relatively good performance on promoter detection tasks in both plants and humans. To investigate the reasons behind this phenomenon and to gain deeper insight into the biological signals captured by the CNN, we extracted the top 5 highest-activating 7-mer motifs from all filters in the first convolutional layer. The results are shown in Tables 7 and 8.

Table 7. Top 5 7-mer motifs with the highest activation strength from the first convolutional layer (across all filters) in human promoter data.

Rank	Motif	Score	Filter	Position
1	GCAGCTG	1.2274	63	110
2	GCAGCTG	1.2274	63	195
3	GCAGCGT	1.2274	63	392
4	GCAGCAG	1.2274	63	412
5	GCAGCCC	1.2274	63	210

Table 8. Top 5 7-mer motifs with the highest activation strength from the first convolutional layer (across all filters) in rice promoter data.

Rank	Motif	Score	Filter	Position
1	CTGTGGC	1.2317	31	434
2	CTGTTTT	1.2317	31	183
3	CTGTTTC	1.2317	31	98
4	CTGTCTT	1.2317	31	33
5	CTGTTTT	1.2317	31	51

In human promoters, the dominant motifs are GC-rich sequences, which are highly consistent with the structure of CpG islands and the widespread presence of Sp1 transcrip-

tion factor binding sites (GC-box) in TATA-less promoters [29]. In contrast, plant promoters are predominantly characterized by CTG-rich sequences or pyrimidine-rich motifs [30]. These motifs mainly originate from filter 31 and closely match the plant-specific Y-patch (pyrimidine patch) element. The Y-patch frequently substitutes for or complements the TATA box in plant core promoters. These findings indicate that the CNN successfully learns species-specific core promoter features.

The strong performance of CNNs on promoter detection tasks can be attributed to their three core mechanisms, local receptive fields, parameter sharing, and multi-scale feature extraction, which naturally align with the biological characteristic of DNA sequences: the presence of short yet biologically meaningful local patterns. This makes CNNs particularly effective at capturing common local regulatory elements in promoter regions (such as GC-box and Y-patch).

Additionally, model architecture and tokenization strategies may influence task performance through information compression and representational capabilities. For example, according to Dotan et al. (2024), subword units generated by the BPE algorithm are typically associated with biologically conserved sequence patterns, making BPE highly suitable for tasks requiring precise pattern recognition [26]. However, due to the task design in this benchmark, no obvious differences in performance attributable to different tokenization strategies were observed. The influence of tokenization strategies may primarily manifest in the final token length resulting from information compression. The vocabulary generated by the BPE algorithm achieves high compression ratios, effectively reducing computational complexity and storage requirements. k-mer tokenization divides sequences into fixed-length k-mers with computational complexity $O(n/k)$, making it highly suitable for large-scale data processing [27]. Single-nucleotide tokenization provides the finest granularity, producing the longest token sequences for a given input length, and is suitable for models like HyenaDNA with computational complexity lower than $O(n \log n)$. This approach effectively handles scenarios such as single-base mutations, enabling fine-grained annotation at the single-base level, something difficult to achieve with the other two tokenization methods.

3.1.3. Genomic Structure and Task Characteristics

In this benchmark study, zero-shot and fine-tuning tasks are designed with identical standards across human and plant genomes (including sequence length, sampling strategy, positive/negative sample definitions, and label space), thereby eliminating design bias. Consequently, observed performance differences primarily stem from the inherent characteristics of human and rice genomes as well as biological differences in their genomic structural distributions. Overall, plant genomes exhibit greater dynamism and variability in sequence space due to frequent polyploidy and rapid structural variation, whereas animal genomes tend toward conservation and structural stability [31–33]. Tables 9 and 10 lists the genome size, gene density, and average gene length for each genome used in the experiments.

In both zero-shot and fine-tuning settings, human tasks consistently achieved higher overall MCC in promoter identification and transcription start site (TSS) scanning. This advantage is likely attributable to more mature annotation resources and highly conserved regulatory motifs in the human genome, which result in lower variance among positive samples in the local pattern space. By contrast, rice tasks face greater challenges due to substantially higher genomic diversity. For example, rice has 42,189 genes with a gene density of 112.66 genes/Mb, whereas humans have only 21,570 genes and a gene density of just 6.96 genes/Mb. This implies that models must search for functional signals across a much broader sequence space.

Table 9. Genomic statistics of the species involved in this benchmark study. The table provides gene density, genome size, number of genes, exons, introns, and maximum/average gene, exon, and intron lengths for humans (*Homo sapiens*), pigs (*Sus scrofa*), mice (*Mus musculus*), dogs (*Canis lupus familiaris*), cattle (*Bos taurus*). Mb, megabase pairs.

	Humans	Pigs	Mice	Dogs	Cattle
Gene density (genes/Mb)	6.96	8.81	9.35	8.58	7.52
Genome size (Mb)	3099.75	2501.91	2728.22	2396.86	2770.69
Gene count	21,570	22,041	25,519	20,567	20,848
Exon count	2,165,096	567,727	1,298,921	539,879	680,837
Intron count	1,777,142	507,287	1,020,552	484,544	595,999
Max Gene length (bp)	2,473,539	2,361,238	2,960,899	2,057,682	2,323,167
Avg Gene length (bp)	64,434.27	48,581.21	42,447.59	52,315.71	54,534.4
Max Exon length (bp)	347,300	18,338	122,802	19,797	48,309
Avg Exon length (bp)	260.7	344.77	276.52	276.47	294.98
Max Intron length (bp)	1,240,120	1,111,590	2,908,816	1,089,409	1,284,691
Avg Intron length (bp)	8213.54	5335.66	6413.54	6142.65	6578.08
GC Content (%)	38.88	41.37	40.54	40.97	42.01

Table 10. Genomic statistics of the species involved in this benchmark study. The table provides gene density, genome size, number of genes, exons, introns, and maximum/average gene, exon, and intron lengths for rice (*Oryza sativa*), maize (*Zea mays*), sorghum (*Sorghum bicolor*), wheat (*Triticum aestivum*), barley (*Hordeum vulgare*). Mb, megabase pairs.

	Rice	Maize	Sorghum	Wheat	Barley
Gene density (genes/Mb)	112.66	30.7	48.66	156.66	8.22
Genome size (Mb)	374.47	2067.86	729.38	634.42	4833.79
Gene count	42,189	63,480	35,490	99,386	39,734
Exon count	259,392	356,887	205,840	1,727,928	1,715,777
Intron count	206,967	268,127	164,792	1,434,875	1,467,597
Max Gene length (bp)	57,094	191,581	93,619	35,335	820,598
Avg Gene length (bp)	2871.77	3001.87	3321.63	2551.95	6008.35
Max Exon length (bp)	15,363	117,494	9082	9275	23,453
Avg Exon length (bp)	317.78	348.72	338.35	266.18	278.87
Max Intron length (bp)	18,327	169,080	21,931	25,396	819,178
Avg Intron length (bp)	405.91	632.99	472.64	358.15	664.92
GC Content (%)	43.55	46.59	41.63	44.02	42.03

In species classification tasks, performance between the two groups was roughly comparable overall, with Mammals classification showing slightly higher average performance than Poaceae. However, this already reflects an adjustment in the mammalian selection from family-level to class-level taxonomy. This comparative result may arise from the relatively large evolutionary distances and more pronounced sequence pattern differences among the five plant species considered.

In the gene region identification task, rice tasks significantly outperformed human tasks under both zero-shot and fine-tuning conditions. This performance reversal stems from the greater discriminative power provided by the non-conserved nature of the rice genome. Rice genes are markedly shorter and more compact, with an average gene length of 2871 bp and an average intron length of 405 bp; in comparison, human genes average 64,434 bp in length with introns averaging 8213 bp. The high variation in intron length and sharper regional boundary signals in rice enable models to more accurately distinguish

exons, introns, and intergenic regions. Conversely, the human genome contains a large number of ultra-long introns, introducing substantial amounts of non-functional sequence that dilutes regional boundary signals, effectively equivalent to injecting high-entropy noise into the sequence, thereby significantly increasing modeling difficulty.

Additionally, in the gene region identification task, an examination of the confusion matrix heatmaps for gene region identification in the Supplementary Figures S7, S8, S15 and S16 reveals that, across nearly all experimental groups, the classification results for introns and intergenic regions are comparatively less distinct than those for exons, with the confusion rates between these two categories being significantly higher than the confusion rates between exons and other categories. The speculative explanation for this phenomenon may stem from the compositional similarity between introns and intergenic regions: both constitute non-coding regions, often enriched in repetitive sequences, characterized by nucleotide composition biases and low-complexity patterns, while lacking the codon usage preferences and conserved coding signals specific to exons. Consequently, models face greater difficulty in capturing discriminative features to separate these regions. In contrast, exons, due to their protein-coding function, typically exhibit more conserved motifs and periodic patterns (such as the triplet periodicity indicative of coding sequences), making them more readily identifiable by the model.

4. Discussion

This benchmark study systematically evaluated five genomic language models, namely DNABERT-2, GROVER, HyenaDNA, NT-V2 and AgroNT, together with a CNN baseline, across four downstream tasks on both human and rice genomes. These tasks collectively encompass sequence motif recognition, functional site detection, cross-species classification and gene structure parsing. The main findings are as follows. In zero-shot settings and fine-tuning experiments, DNABERT-2 and NT-V2 respectively achieved the highest average MCC, while AgroNT consistently showed superior performance on plant-related tasks. Compared with zero-shot results, fine-tuning produced substantial performance improvements across all models, with particularly notable gains observed for NT-V2. These outcomes highlight the strengths of broad multi-species pretraining, as seen in DNABERT-2 and NT-V2, and domain-specific plant pretraining, as seen in AgroNT, in achieving strong adaptation within specialized domains and effective cross-domain transfer.

Compared with recent comprehensive benchmark studies, the present work provides a complementary perspective. It places particular emphasis on fine-tuning performance and explicit cross-kingdom generalization between animals and plants. The evaluation is extended to plant genomes, and the results demonstrate that task-specific fine-tuning can significantly reduce domain gaps, especially for models that have been pretrained on multi-species or domain-specific data. At the same time, this study has several limitations. Although cross-species generalization was assessed, the core fine-tuning and downstream evaluation were primarily restricted to two representative species, human and rice. While the species classification task included additional species, broader phylogenetic coverage, such as more animal phyla, non-Poaceae plants or prokaryotes, is still needed to fully support claims of wide-ranging cross-kingdom transferability. In addition, the four downstream tasks mainly focused on sequence structure and classification. Future work could incorporate biologically more meaningful tasks to gain deeper insights into functional consequences. Furthermore, although the results clearly reveal model-specific advantages, more detailed biological interpretations would help provide a deeper mechanistic understanding.

Despite these limitations, the findings of this study offer practical guidance for future model development. The excellent performance of NT-V2 and AgroNT after fine-tuning

suggests that combining broad multi-species pretraining with domain-specific data (such as plant genomes) holds promise for obtaining more adaptive and robust representations. Meanwhile, smaller-scale models like DNABERT-2, although exhibiting relatively limited gains from fine-tuning, possess strong zero-shot usability that remains a valuable direction worthy of further exploration. If strict domain matching is pursued, training models on multiple genomes from the same domain rather than a single genome, may represent an effective strategy. Task-specific fine-tuning should be regarded as an essential step whenever conditions permit, given that in this benchmark study, the performance gap between zero-shot and fine-tuned settings remains substantial across nearly all models and tasks. Future developers may consider exploring hybrid pretraining strategies that balance breadth and depth, as well as architectural optimizations tailored to specific biological domains.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/app16041745/s1>. Supplementary Figure S1: Confusion matrix heatmap of Promoter Detection(Rice, Zero-Shot Testing). Supplementary Figure S2: Confusion matrix heatmap of Promoter Detection(Human, Zero-Shot Testing). Supplementary Figure S3: Confusion matrix heatmap of TSS Scanning(Rice, Zero-Shot Testing). Supplementary Figure S4: Confusion matrix heatmap of TSS Scanning(Human, Zero-Shot Testing). Supplementary Figure S5: Confusion matrix heatmap of Species Classification (Poaceae, Zero-Shot Testing). Supplementary Figure S6: Confusion matrix heatmap of Species Classification (Mammals, Zero-Shot Testing). Supplementary Figure S7: Confusion matrix heatmap of Gene Region Identification (Rice, Zero-Shot Testing). Supplementary Figure S8: Confusion matrix heatmap of Gene Region Identification (Human, Zero-Shot Testing). Supplementary Figure S9: Confusion matrix heatmap of Promoter Detection(Rice, Fine-Tuning). Supplementary Figure S10: Confusion matrix heatmap of Promoter Detection(Human, Fine-Tuning). Supplementary Figure S11: Confusion matrix heatmap of TSS Scanning(Rice, Fine-Tuning). Supplementary Figure S12: Confusion matrix heatmap of TSS Scanning(Human, Fine-Tuning). Supplementary Figure S13: Confusion matrix heatmap of Species Classification (Poaceae, Fine-Tuning). Supplementary Figure S14: Confusion matrix heatmap of Species Classification (Mammals, Fine-Tuning). Supplementary Figure S15: Confusion matrix heatmap of Gene Region Identification (Rice, Fine-Tuning). Supplementary Figure S16: Confusion matrix heatmap of Gene Region Identification (Human, Fine-Tuning). Supplementary Table S1: Benchmarking results of different models for Promoter Detection (Rice, Zero-shot Testing). Supplementary Table S2: Benchmarking results of different models for Promoter Detection (Human, Zero-shot Testing). Supplementary Table S3: Benchmarking results of different models for TSS Scanning (Rice, Zero-shot Testing). Supplementary Table S4: Benchmarking results of different models for TSS Scanning (Human, Zero-shot Testing). Supplementary Table S5: Benchmarking results of different models for Species Classification (Poaceae, Zero-shot Testing). Supplementary Table S6: Benchmarking results of different models for Species Classification (Mammals, Zero-shot Testing). Supplementary Table S7: Benchmarking results of different models for Gene Region Identification (Rice, Zero-shot Testing). Supplementary Table S8: Benchmarking results of different models for Gene Region Identification (Human, Zero-shot Testing). Supplementary Table S9: Benchmarking results of different models for Promoter Detection (Rice, Fine-Tuning). Supplementary Table S10: Benchmarking results of different models for Promoter Detection (Human, Fine-Tuning). Supplementary Table S11: Benchmarking results of different models for TSS Scanning (Rice, Fine-Tuning). Supplementary Table S12: Benchmarking results of different models for TSS Scanning (Human, Fine-Tuning). Supplementary Table S13: Benchmarking results of different models for Species Classification (Poaceae, Fine-Tuning). Supplementary Table S14: Benchmarking results of different models for Species Classification (Mammals, Fine-Tuning). Supplementary Table S15: Benchmarking results of different models for Gene Region Identification (Rice, Fine-Tuning). Supplementary Table S16: Benchmarking results of different models for Gene Region Identification (Human, Fine-Tuning). Supplementary Table S17: Random Forest classifier hyperparameters. The configuration was kept consistent across all experiments to ensure fair com-

parison and reproducibility. Supplementary Table S18: Confusion Matrix for Binary Classification. Rows represent the actual class, columns represent the predicted class, and entries (TP, FN, FP, TN) denote the number of samples predicted as positive and actually positive, predicted as negative but actually positive, predicted as positive but actually negative, and predicted as negative and actually negative, respectively. Supplementary Table S19: Confusion Matrix for Multi-Class Classification. Rows represent the actual class, columns represent the predicted class, and diagonal elements ($C_{ii,i} = A, B, \dots, N$) represent the number of correctly predicted samples for each class (true positives, TP). Off-diagonal elements ($C_{ij,i} \neq j$) represent the number of samples with actual class i predicted as class j , indicating misclassifications (false negatives, FN, for actual class i ; false positives, FP, for predicted class j).

Author Contributions: X.G. and W.P. conceived the study and designed the tasks. X.G. selected the models and performed the experiments. W.P. collected and processed the data. S.W. supervised the project and drafted the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key R&D Program of China (2025YFC3410300); the National Natural Science Foundation of China (Grant No. 32470678); the Agricultural Science and Technology Innovation Program (CAAS-ZDRW202503); the Youth Innovation Program of the Chinese Academy of Agricultural Sciences (Y2025QC36); the Innovation Program of Chinese Academy of Agricultural Sciences (CAAS-CSIAF-202301); Science and Technology Project of the Ministry of Agriculture and Rural Affairs, P.R.China.

Data Availability Statement: The code used for model fine-tuning, evaluation, and statistical analysis is publicly available at: <https://github.com/gxs1111/gxs1111-Rice-Human-GLM-Benchmark> (accessed on 2 February 2026). All processed benchmark datasets (promoter detection, TSS scanning, species classification, and gene region identification for both human and rice genomes) are publicly available on Hugging Face Datasets: <https://huggingface.co/datasets/gxs1220/rice-human-glm-benchmark-data> (accessed on 2 February 2026).

Acknowledgments: S.W. acknowledges the support of grant from Taishan Scholar Youth Expert program, and Youth Innovative Talents Program of Shandong Province of China. All authors acknowledge the Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences for providing computational resources.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Shendure, J.; Ji, H. Next-generation DNA sequencing. *Nat. Biotechnol.* **2008**, *26*, 1135–1145. [\[CrossRef\]](#)
- Reuter, J.A.; Spacek, D.V.; Snyder, M.P. High-throughput sequencing technologies. *Mol. Cell* **2015**, *58*, 586–597. [\[CrossRef\]](#)
- Brendel, V.; Busse, H.G. Genome structure described by formal languages. *Nucleic Acids Res.* **1984**, *12*, 2561–2568. [\[CrossRef\]](#)
- Head, T. Formal language theory and DNA: An analysis of the generative capacity of specific recombinant behaviors. *Bull. Math. Biol.* **1987**, *49*, 737–759. [\[CrossRef\]](#)
- Ji, S. The linguistics of DNA: Words, sentences, grammar, phonetics, and semantics. In *Molecular Strategies in Biological Evolution*; New York Academy of Sciences: New York, NY, USA, 1999; Volume 870.
- King, D.; Atlasi, Y.; Rafiee, G. DNABERT-2: Fine-Tuning a Genomic Language Model for Colorectal Gene Enhancer Classification. *arXiv* **2025**, arXiv:2509.25274. [\[CrossRef\]](#)
- Li, F.; Lu, W.; Bai, Y. HyenaCircle: A HyenaDNA-based pretrained large language model for long eccDNA prediction. *Front. Genet.* **2025**, *16*, 1641162. [\[CrossRef\]](#)
- Marin, F.I.; Teufel, F.; Horlacher, M.; Madsen, D.; Pultz, D.; Winther, O.; Boomsma, W. BEND: Benchmarking DNA Language Models on biologically meaningful tasks. In Proceedings of the International Conference on Learning Representations (ICLR), Vienna, Austria, 7–11 May 2024.
- Awasthi, R.; Mend Mend Arachchige, G.S.; Zhu, X. Unsupervised evaluation of pre-trained DNA language model embeddings. *BMC Genom.* **2025**, *26*, 710. [\[CrossRef\]](#) [\[PubMed\]](#)
- Feng, H.; Wu, L.; Zhao, B.; Huff, C.; Zhang, J.; Wu, J.; Lin, L.; Wei, P.; Wu, C. Benchmarking DNA foundation models for genomic and genetic tasks. *Nat. Commun.* **2025**, *16*, 10780. [\[CrossRef\]](#)

11. Consortium, I.H.G.S. Finishing the euchromatic sequence of the human genome. *Nature* **2004**, *431*, 931–945. [[CrossRef](#)] [[PubMed](#)]
12. International Rice Genome Sequencing Project; Sasaki, T. The map-based sequence of the rice genome. *Nature* **2005**, *436*, 793–800. [[CrossRef](#)] [[PubMed](#)]
13. Ji, Y.; Zhou, Z.; Liu, H.; Davuluri, R.V. DNABERT: Pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics* **2021**, *37*, 2112–2120. [[CrossRef](#)]
14. Zhou, Z.; Ji, Y.; Li, W.; Dutta, P.; Davuluri, R.; Liu, H. DNABERT-2: Efficient foundation model and benchmark for multi-species genome. *arXiv* **2023**, arXiv:2306.15006.
15. Sanabria, M.; Hirsch, J.; Joubert, P.M.; Poetsch, A.R. DNA language model GROVER learns sequence context in the human genome. *Nat. Mach. Intell.* **2024**, *6*, 911–923. [[CrossRef](#)]
16. Nguyen, E.; Poli, M.; Faizi, M.; Thomas, A.; Birch-Sykes, C.; Wornow, M.; Patel, A.; Rabideau, C.; Massaroli, S.; Bengio, Y.; et al. HyenaDNA: Long-range genomic sequence modeling at single nucleotide resolution. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2023.
17. Dalla-Torre, H.; Gonzalez, L.; Mendoza-Revilla, J.; Lopez Carranza, N.; Grzywaczewski, A.H.; Oteri, F.; Dallago, C.; Trop, E.; de Almeida, B.P.; Sirelkhatim, H.; et al. Nucleotide Transformer: Building and evaluating robust foundation models for human genomics. *Nat. Methods* **2025**, *22*, 287–297. [[CrossRef](#)] [[PubMed](#)]
18. Mendoza-Revilla, J.; Trop, E.; Gonzalez, L.; Roller, M.; Dalla-Torre, H.; de Almeida, B.P.; Richard, G.; Caton, J.; Lopez Carranza, N.; Skwark, M.; et al. A foundational large language model for edible plant genomes. *Commun. Biol.* **2024**, *7*, 835. [[CrossRef](#)]
19. Goodstein, D.M.; Shu, S.; Howson, R.; Neupane, R.; Hayes, R.D.; Fazo, J.; Mitros, T.; Dirks, W.; Hellsten, U.; Putnam, N.; et al. Phytozome: A comparative platform for green plant genomics. *Nucleic Acids Res.* **2012**, *40*, D1178–D1186. [[CrossRef](#)] [[PubMed](#)]
20. Dyer, S.C.; Austine-Orimoloye, O.; Azov, A.G.; Barba, M.; Barnes, I.; Barrera-Enriquez, V.P.; Becker, A.; Bennett, R.; Beracochea, M.; Berry, A.; et al. Ensembl 2025. *Nucleic Acids Res.* **2025**, *53*, D948–D957. [[CrossRef](#)]
21. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
22. Sennrich, R.; Haddow, B.; Birch, A. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Berlin, Germany, 2016.
23. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is All you Need. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2017.
24. Press, O.; Smith, N.A.; Lewis, M. Train Short, Test Long: Attention with Linear Biases Enables Input Length Extrapolation. *arXiv* **2021**, arXiv:2108.12409.
25. Dao, T.; Fu, D.; Ermon, S.; Rudra, A.; Ré, C. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2022.
26. Su, J.; Ahmed, M.; Lu, Y.; Pan, S.; Bo, W.; Liu, Y. RoFormer: Enhanced transformer with Rotary Position Embedding. *Neurocomputing* **2024**, *568*, 127063. [[CrossRef](#)]
27. Poli, M.; Massaroli, S.; Nguyen, E.; Fu, D.Y.; Dao, T.; Baccus, S.; Bengio, Y.; Ermon, S.; Ré, C. Hyena hierarchy: Towards larger convolutional language models. In *Proceedings of the 40th International Conference on Machine Learning*; PMLR: London, UK, 2023.
28. Gu, A.; Dao, T. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. *arXiv* **2024**, arXiv:2312.00752v2.
29. Deaton, A.M.; Bird, A. CpG islands and the regulation of transcription. *Genes Dev.* **2011**, *25*, 1010–1022. [[CrossRef](#)]
30. Brooks, E.G.; Elorriaga, E.; Liu, Y.; Duduit, J.R.; Yuan, G.; Tsai, C.-J.; Tuskan, G.A.; Ranney, T.G.; Yang, X.; Liu, W. Plant Promoters and Terminators for High-Precision Bioengineering. *BioDesign Res.* **2023**, *5*, 0013. [[CrossRef](#)]
31. Mukhopadhyay, P.; Ghosh, T.C. Relationship between gene compactness and base composition in rice and human genome. *J. Biomol. Struct. Dyn.* **2010**, *27*, 477–488. [[CrossRef](#)] [[PubMed](#)]
32. Reig-Valiente, J.L.; Marques, L.; Talon, M.; Domingo, C. Genome-wide association study of agronomic traits in rice cultivated in temperate regions. *BMC Genom.* **2018**, *19*, 706. [[CrossRef](#)]
33. Murat, F.; Van de Peer, Y.; Salse, J. Decoding plant and animal genome plasticity from differential paleo-evolutionary patterns and processes. *Genome Biol. Evol.* **2012**, *4*, 917–928. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.