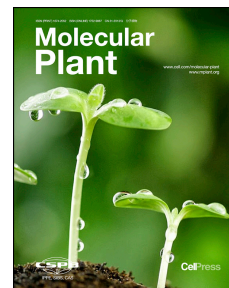


# Journal Pre-proof

PDLLMs: A group of tailored DNA large language models for analyzing plant genomes

Guanqing Liu, Long Chen, Yuechao Wu, Yangshuo Han, Yu Bao, Tao Zhang



PII: S1674-2052(24)00390-3

DOI: <https://doi.org/10.1016/j.molp.2024.12.006>

Reference: MOLP 1829

To appear in: *MOLECULAR PLANT*

Received Date: 6 July 2024

Revised Date: 21 October 2024

Accepted Date: 6 December 2024

Please cite this article as: Liu G., Chen L., Wu Y., Han Y., Bao Y., and Zhang T. (2025). PDLLMs: A group of tailored DNA large language models for analyzing plant genomes. Mol. Plant. doi: <https://doi.org/10.1016/j.molp.2024.12.006>.

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2024 The Author

**PDLLMs: A group of tailored DNA large language models for analyzing plant genomes**Guanqing Liu<sup>1,2</sup>, Long Chen<sup>1,2</sup>, Yuechao Wu<sup>1,2</sup>, Yangshuo Han<sup>1,3</sup>, Yu Bao<sup>1,2</sup>, Tao Zhang<sup>1,2,\*</sup>

<sup>1</sup>Jiangsu Key Laboratory of Crop Genomics and Molecular Breeding/Zhongshan Biological Breeding Laboratory/Key Laboratory of Plant Functional Genomics of the Ministry of Education, Agricultural College of Yangzhou University, Yangzhou 225009, China.

<sup>2</sup>Jiangsu Co-Innovation Center for Modern Production Technology of Grain Crops/Jiangsu Key Laboratory of Crop Genetics and Physiology, Yangzhou University, Yangzhou 225009, China.

<sup>3</sup>Chongqing Key Laboratory of Plant Resource Conservation and Germplasm Innovation, Integrative Science Center of Germplasm Creation in Western China (Chongqing) Science City, School of Life Sciences, Southwest University, Chongqing, 400715.

Correspondence: Tao Zhang (zhangtao@yzu.edu.cn)

Dear Editor,

Over the past two years, large language models (LLMs) have witnessed a notable achievement in the realms of artificial intelligence generated content (AIGC). These advanced AIGC systems are powered by state-of-the-art deep learning algorithms, which undergo training on vast amounts of data and possess the ability to comprehend extensive volumes of complex text. Recently, more and more LLMs have been developed and applied to a diverse range of tasks, such as interactive chat, auto-generate drawing, and even biological research (Kaddour et al., 2023).

General LLMs are initially designed for processing human languages, which utilize common characters as vocabulary, therefore they are not suitable for DNA sequence analysis. To realize learning and solving DNA analysis issues, some DNA LLMs have been developed by using k-mer of nucleotides as their vocabulary. These LLMs exhibited significant advancements in the applications for biology, among them, several representative LLMs have been developed, which are DNABERT/DNABERT-2 (Ji et al., 2021; Zhou et al., 2023) and Nucleotide Transformer (Dalla-Torre et al., 2024). However, these models are not designed for plant genomes. Recently, the InstaDeep institution developed a large language model named Agronomic Nucleotide Transformer (AgroNT) in editable plant genomes (Mendoza-Revilla et al., 2024). Although the model performs well in plant prediction tasks, the model size is too large to train on a personal computer. Taking into account the challenges posed by the large model size and limited plant genome optimization, we presented PDLLMs (Plant DNA Large Language Models), a group of DNA large language

models based on five foundational model architectures that are tailored for analyzing and processing plant genomes. These models utilize distinct architectural approaches to better accommodate the unique characteristics of plant genomic data. Besides, we constructed 10 plant datasets and fine-tuned 198 models for genomic prediction tasks.

### **Construction of foundation plant DNA language models**

The motivation behind this research is to develop a set of portable models that are able to study the genomics of crops, which are essential for food security and sustainable agriculture. To achieve this, we first collected genome sequences from 14 representative plant species (**Supplemental Table 1**), then we developed five types of plant-specific foundation DNA language models: Plant DNAMamba, Plant DNABERT, Plant DNAGPT, Plant DNAGemma, and Plant Nucleotide Transformer (NT) based on byte pair encoding (BPE), k-mer and single base tokenizers (**Figure 1A**). Notably, Plant DNAMamba is the newest addition, built upon a novel structured state space sequence model (SSM) called Mamba (Albert and Tri, 2023). The remaining four models leverage the popular transformer framework and draw inspiration from large language model like DNABERT and NT.

### **Fine-tune plant DNA language models for downstream genomic tasks**

After construction of these foundation models, we generated 10 plant datasets for different biological applications, including predictions on core promoter, sequence conservation, histone modification, open chromatin, lncRNAs and promoter strength in multiple plant species (**Figure 1B**). Foundation models were further fine-tuned on these datasets and 198 specific models were built for downstream genomic prediction tasks. Among these tasks, prediction of core promoter and sequence conservation are binary classification tasks with fixed sequence length, while classification of epigenetic tasks such as histone modification, open chromatin and lncRNAs are trained with variable sequence length. For open chromatin task, we predict for three classes including not, full and partial open chromatin regions (OCRs) based on the proportion of OCRs in a sequence (**Supplemental Figure 1**). Additionally, promoter strength predictions involve regression tasks. Detailed description of these tasks and corresponding datasets are available in **Supplemental Notes**.

### Statistics and comparison of foundation DNA language models

To evaluate the performance of our plant models, we compared these models across various downstream tasks using three different tokenizers (single base, 6-mer and BPE). In general, the plant DNAMamba model demonstrated overall superior performance among all the models considered, achieving the best performance in 5 out of 9 fine-tuning tasks, such as sequence conservation, H3K27ac, H3K4me3, H3K27me3, and open chromatin (**Figure 1E** and **Supplemental Figure 3**). Subsequently, we implemented several comparisons between our best models and other foundation models to assess their performances (**Figure 1C-1E**; **Supplemental Table 2**). First, we compared our models with AgroNT. The largest plant model AgroNT only obtained state-of-the-art predictions in promoter strength, however the model's parameters are ten times larger than our plant models, incapable of training on machines with low GPU memory sizes (**Figure 1C** and **1D**). In addition, our plant models achieved better performance than other non-plant LLMs in total nine fine-tuning tasks in most plant prediction tasks.

Based on the above results, it is evident that different tokenization methods can impact the performance of the same model architecture on various downstream tasks (**Supplemental Figure 3**). Considering DNAMamba as a highly promising foundation model in general purpose, we incorporated single base (one-hot like), 2 to 6-mers and BPE tokenization methods to comprehensively evaluate DNAMamba performance across various downstream tasks. The results showed that employing different tokenizer methods enhanced the performance of DNAMamba in downstream fine-tuning tasks (**Supplemental Figure 4**), which provide a reference for choosing different tokenizers across various genomic tasks.

We next explored the training and inference efficiency of each model. Except for the prediction performance, plant DNAMamba model has extra advantages in training and fine-tuning time (**Supplemental Figure 2**). When comparing the fine-tuning time of each model, Plant DNAMamba and DNABERT-2 models show the best efficiencies. Besides, all our plant models can be trained and inferred on consumer-grade GPUs (**Supplemental Figure 5**; **Supplemental Table 2**), making them accessible to researchers and scientists worldwide. Nevertheless, Plant DNAMamba cannot be trained on CPU and old GPU like the GTX1080Ti, which is released eight years ago. AgroNT failed to be trained on a consumer-grade GPUs without a parameter-efficient fine-tuning (PEFT)

method caused by out of memory error (**Supplemental Figure 2B**).

### Comparison between LLMs and other deep learning models

To further prove the predictive performance of our LLMs, we compared the best plant DNAMamba model with several non-LLM models in different downstream tasks. DNAMamba and the LSTM-based model PlncRNA-Hdeep (Meng et al., 2021) had comparable performance when predicting maize lncRNAs, however, PlncRNA-Hdeep performed worse in multiple species lncRNAs prediction (**Figure 1F**). In promoter strength prediction tasks, DNAMamba also outperformed the CNN models in 2 different systems (Jores et al., 2021) (**Figure 1G**). In addition, prediction results of chromatin accessibility in different tissues and species by DNAMamba were all better than the CNN-based model PlantDeepSEA (Zhao et al., 2021) (**Figure 1H**). These results showed that our plant DNA LLMs are versatile and robust, which defeat the specialized non-LLM models in various tasks using only one model.

### Application of plant foundation DNA language models

Large language models have versatile applications when applied to biology research, including protein structure prediction, identification of regulatory elements, gene expression prediction and others (Consens et al., 2023). Here, we demonstrated that our fine-tuned models are able to perform diverse prediction jobs. For example, fine-tuned models on histone modification enable the prediction of regions potentially enriched with histone modification. As is shown in **Supplemental Figure 6**, two regions enriched with H3K27ac at root and leaf are both predicted as H3K27ac enriched regions. Besides, for scientists who want to fine-tune or infer models locally, we provided both command-line scripts and a docker build for convenience (**Figure 1I**). Finally, we implemented an integrated web application specifically for predicting all the nine downstream tasks, a demo of online open chromatin prediction is displayed in the **Supplemental Figure 7**.

In conclusion, our research shows that foundation large language models can be effectively trained and inferred on CPUs and consumer-grade GPUs and exhibit remarkable performance in various plant genomic tasks. Besides, our constructed plant datasets will accelerate the development of other LLMs in the future. With the increase of computing power, complete sequencing of plant genomes

(Sun et al., 2022) and multi-modal data, it is expected that increasing model sizes, adding more plant genomes and integrating more omics data will further improve the abilities and applications of PDLLMs(Lam et al., 2024). We believe that our models will facilitate the discovery of important DNA elements and provide guidance for crop engineering, which opens up new possibilities for studying the genomics of crops and help in developing new crop varieties that are more resistant to diseases, pests, and environmental stress.

### **Data and code availability**

All the information of genome assemblies, gene annotations and omics data are detailed introduced in **Supplemental Table 1, 3** and **Supplemental Notes**. Codes for fine-tuning the models and inference are available at GitHub ([https://github.com/zhangtaolab/Plant\\_DNA\\_LLMs](https://github.com/zhangtaolab/Plant_DNA_LLMs)) and Gitee ([https://gitee.com/zhangtaolabyzu/Plant\\_DNA\\_LLMs](https://gitee.com/zhangtaolabyzu/Plant_DNA_LLMs)). The integrated web application is publicly accessible at <https://finetune.plantllm.org> and <https://bioinfor.yzu.edu.cn/llms/finetune>. We also put our web applications for online prediction each tasks, as well as all the trained plant DNA LLMs and plant genomics datasets on HuggingFace (<https://huggingface.co/zhangtaolab>) and ModelScope (<https://modelscope.cn/organization/zhangtaolab>), the lists of these models, datasets and web applications have been listed in the GitHub, Gitee and **Supplemental Table 4**. Hyperparameters used for pre-training and fine-tuning the models are listed in **Supplemental Table 5**.

### **Funding**

This work was financially supported by the National Natural Science Foundation of China (grant no. 32270585), the Key R&D Program of Jiangsu Province (Modern Agriculture) (BE2022335), the Project of Zhongshan Biological Breeding Laboratory (BM2022008-02), the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD) to T.Z.

### **Author contributions**

T.Z. conceived and designed the research. G.L., Y.W., Y.B. and L.C. collected the data for model training, T.Z. and G.L. preprocessed data and pretrained the models. G.L. performed model fine-tuning and comparison. L.C. and Y.W. helped test the models. G.L., L.C. and Y.H. developed the

web application. G.L. and T.Z. wrote and revised the manuscript. All authors read and approved the manuscript.

## Acknowledgments

No conflict of interest is declared.

## References

- Albert, G., and Tri, D.** (2023). Mamba: Linear-Time Sequence Modeling with Selective State Spaces. arXiv 10.48550/arXiv.2312.00752.
- Consens, M.E., Dufault, C., Wainberg, M., Forster, D., Karimzadeh, M., Goodarzi, H., Theis, F.J., Moses, A., and Wang, B.** (2023). To Transformers and Beyond: Large Language Models for the Genome. arXiv 10.48550/arXiv.2311.07621.
- 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.** (2024). Nucleotide Transformer: building and evaluating robust foundation models for human genomics. *Nat Methods* 10.1038/s41592-024-02523-z.
- Ji, Y., Zhou, Z., Liu, H., and Davuluri, R.V.** (2021). DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics* 37:2112-2120. 10.1093/bioinformatics/btab083.
- Jores, T., Tonnies, J., Wrightsman, T., Buckler, E.S., Cuperus, J.T., Fields, S., and Queitsch, C.** (2021). Synthetic promoter designs enabled by a comprehensive analysis of plant core promoters. *Nat Plants* 7:842-855. 10.1038/s41477-021-00932-y.
- Kaddour, J., Harris, J., Mozes, M., Bradley, H., Raileanu, R., and McHardy, R.** (2023). Challenges and Applications of Large Language Models. arXiv 10.48550/arXiv.2307.10169.
- Lam, H.Y.I., Ong, X.E., and Mutwil, M.** (2024). Large language models in plant biology. *Trends Plant Sci* 29:1145-1155. 10.1016/j.tplants.2024.04.013.
- Mendoza-Revilla, J., Trop, E., Gonzalez, L., Roller, M., Dalla-Torre, H., de Almeida, B.P., Richard, G., Caton, J., Carranza, N.L., Skwark, M., et al.** (2024). A foundational large language model for edible plant genomes. *Commun Biol* 7:835. 10.1038/s42003-024-06465-2.
- Meng, J., Kang, Q., Chang, Z., and Luan, Y.** (2021). PlncRNA-HDeep: plant long noncoding RNA prediction using hybrid deep learning based on two encoding styles. *BMC Bioinformatics* 22:242. 10.1186/s12859-020-03870-2.
- Sun, Y., Shang, L., Zhu, Q.H., Fan, L., and Guo, L.** (2022). Twenty years of plant genome sequencing: achievements and challenges. *Trends Plant Sci* 27:391-401. 10.1016/j.tplants.2021.10.006.
- Zhao, H., Tu, Z., Liu, Y.M., Zong, Z.X., Li, J.C., Liu, H., Xiong, F., Zhan, J.L., Hu, X.H., and Xie, W.B.** (2021). PlantDeepSEA, a deep learning-based web service to predict the regulatory effects of genomic variants in plants. *Nucleic Acids Research* 49:W523-W529. 10.1093/nar/gkab383.
- Zhou, Z., Ji, Y., Li, W., Dutta, P., Davuluri, R., and Liu, H.** (2023). DNABERT-2: Efficient Foundation Model and Benchmark For Multi-Species Genome. arXiv 10.48550/arXiv.2306.15006.

## Figure Legends

### Figure 1. Overview, statistics analysis and applications of plant DNA large language models.

(A) Pre-training of plant DNA large language models. The training process includes selection of representative plant reference genome and training data generation, plant specific tokenizer training, construction of five models based on different architectures. (B) Graphical representation of genetic and epigenetic features for downstream tasks that fine-tuned on all models. (C) Number of parameters and loading time of five trained models and three public models. (D) Memory used for loading models and maximum batch sizes that different models can be assigned (measured on Nvidia RTX4090 GPU). (E) Summary of best models' performances on different downstream tasks. Predictions of core promoter and sequence conservation are fixed-length binary classification tasks, all these tasks use F1 score as metric. Prediction of histone modifications and lncRNAs are variable-length binary classification tasks, all these tasks use F1 score as metric. Prediction of open chromatin is a variable-length multi-class classification task, which uses AUC score as metric. Predictions of promoter strength are regression tasks, which use R square as metric. Words under the scores represent the model's tokenizer and best scores are marked as red. (F) Performance of plant DNAMamba and PlncRNA-Hdeep in predicting lncRNAs. (G) Performance of plant DNAMamba and CNN models in predicting promoter strength in tobacco leaves and maize protoplasts. (H) Performance of plant DNAMamba and PlantDeepSEA in predicting chromatin accessibilities in different plant species. (I) Command-line tools, local build and online version of PDLLEs are provided for prediction.

### Supplemental Figure 1. Classification schematic of the open chromatin for prediction.

### Supplemental Figure 2. Pretraining and fine-tuning of the plant DNA LLMs.

(A) Total training days used for five plant LLMs. (B) Fine-tuning efficiency of all the selected DNA LLMs by three different GPUs. The efficiency is defined as how many input sequences are predicted by the model in one second. Note that Mamba model cannot be trained on the GTX1080Ti GPU, and AgroNT model failed to be fine-tuned on GTX1080Ti and RTX4070Ti without parameter-efficient strategy.

### Supplemental Figure 3. Comparison of plant DNA LLMs across different downstream tasks with 3 different tokenization methods.

Scores in the heatmap are exhibited as percentages. Binary classification tasks were measured with F1 score, multi-class classification task was measured with AUC score, regression tasks were measured with R square. For each tokenization methods, best model in each task was marked with white number. For each task, best model across all the tokenization methods was marked with blue box.

### Supplemental Figure 4. Performances of plant DNAMamba models with different tokenizers for predicting different downstream tasks.

Scores in the heatmap are exhibited as percentages. Binary classification tasks were measured with F1 score, multi-class classification task was measured with AUC score, regression tasks were measured with R square. For each tokenization methods, best model in each task was marked with

white number.

**Supplemental Figure 5. Inference time of all the models on different tasks using different devices.**

Eight models were inferred on nine downstream tasks. AMD Ryzen7 3700X CPU, Apple M2, Nvidia GTX1660 super, GTX1080Ti, RTX3080 and RTX4090 were tested for all the models. The efficiency is defined as how many input sequences are predicted by the model in one second. Each task was measured three times with different testing datasets, inference efficiencies are represented as mean  $\pm$  SD.

**Supplemental Figure 6. Illustration of the prediction task for H3K27ac histone modification.**

Two regions enriched with H3K27ac at different developmental stages are all detected as histone modification regions by plant DNA LLMs. Boxes with red dashed border represent the regions that are predicted by models.

**Supplemental Figure 7. A demo of the integrated web application of PDLLMs for open chromatin prediction.**

**Supplemental Table 1. Reference genomes and corresponding annotation information for models pre-training.**

**Supplemental Table 2. Statistics of all fine-tuned tasks and models.**

**Supplemental Table 3. Source data of fine-tuned tasks.**

**Supplemental Table 4. Download list of all pre-train models, fine-tune models and plant datasets and web applications for prediction.**

**Supplemental Table 5. Hyperparameters used for pre-training and fine-tuning the models.**

