植物学报 *Chinese Bulletin of Botany* 2026, **61** (4): 571–587, www.sciengine.com/CBB/home
URL: https://doi.org/10.3724/CBB-2026-0060

·特邀综述· 植物科学与人工智能融合研究专辑

植物基因组语言模型的研究进展与应用

张枭[1*], 李琼[1*], 王楠[2*]

[1]中国热带农业科学院热带作物品种资源研究所, 海口 571101; [2]中国农业科学院农业基因组研究所, 岭南现代农业科学与技术广东省实验室深圳分中心, 热带作物生物育种全国重点实验室, 农业农村部农业基因数据分析重点实验室, 深圳 518124

摘要 大语言模型(LLMs)的发展, 为基因组学研究提供了新路径。基因组语言模型(gLMs)的核心思想是将基因组数据视为具有内在语法的生物语言, 对其建模从而挖掘隐含的调控逻辑, 为解析遗传机制提供了新方法。该文围绕植物基因组语言模型的理论基础与研究现状展开综述, 介绍了基于跨物种预训练模型预测变异效应与功能元件、面向单细胞数据的细胞类型注释以及多模态数据融合3个重要研究方向, 讨论了植物基因组中高度重复序列和多倍体等特征对模型构建的影响, 阐述了部分已有建模策略与模型部署框架, 并展望了植物基因组语言模型的发展前景。

关键词 植物基因组学, 人工智能, 大语言模型, 多组学, 遗传变异

张枭, 李琼, 王楠 (2026). 植物基因组语言模型的研究进展与应用. 植物学报 **61**, 571–587.

大语言模型(large language models, LLMs)近年来不断发展, 以Transformer架构为核心的模型在自然语言理解和生成方面表现出色(Vaswani et al., 2017)。随着算力提升、训练数据规模扩大以及模型架构演进, 大语言模型已能够完成复杂推理和知识整合等任务(Wei et al., 2022)。植物基因组数据可被视为一种特殊的“生物语言”, 其中DNA序列和基因表达谱等均蕴含着复杂的遗传信息(Agathokleous et al., 2024)。将大语言模型引入植物基因组学研究, 有助于解码上述复杂信息, 加速优异种质资源的挖掘与利用, 为应对粮食安全以及气候变化等重大挑战提供新路径(Najafabadi and Jackson, 2026)。

大语言模型在植物基因组学研究中的应用可以从2个维度来理解。一个维度是对模式植物的纵向深度研究。经过长期研究, 以拟南芥(*Arabidopsis thaliana*)和水稻(*Oryza sativa*)等为代表的模式植物已在基因组学、表观组学和功能基因组学等方面积累了大量重要数据(Lan et al., 2021; Han et al., 2023; Wang et al., 2025)。通过大语言模型来整合这些数据, 有助于揭示围绕“中心法则”的调控机制(Barbadilla-Martinez et al., 2025)和理解植物生长发育规律(Xue et al., 2025)。另一个维度是横向比较研究, 即基于多物种序列的演化分析(Feng et al., 2020)。自然界中已开展了无数次的“演化实验”。目前, 数千个被测序的植物正是这一进程中的现存案例, 并为模型训练提供输入(Xie et al., 2024)。基于上述参考基因组构建跨物种的语言模型, 有助于理解调控序列的保守性以及功能元件的演化约束等规律(Benegas et al., 2025b)。

上述2个研究维度目前基于一套相似的模型构建框架, 大体包含4个部分。(1) 分词(亦称词元化): 将原始基因组序列拆解为词元(token), 每个词元会转化为高维向量, 称为词嵌入(embedding), 最终构成词嵌入矩阵, 作为模型训练的输入。(2) 架构设计: 根据序列长度和计算效率等需求, 选用单一或者混合模型架构, 确定模型理解序列的方式。(3) 自监督预训练: 基于选定的架构在无标注序列上学习基因组表征, 获得基础模型(foundation models)。(4) 下游任务迁移: 直接利用基础模型进行预测, 或者微调模型参数完成下游特定任务(图1)。

本文重点讨论跨物种基因组建模、单细胞组学解析及多模态数据融合等应用场景, 比较了不同模型的

收稿日期: 2026-03-16; 接受日期: 2026-06-01
基金项目: 国家自然科学基金(No.32302498)
* 通讯作者。E-mail: zhxiao7777@gmail.com; liqiong4416@126.com; wangnan05@caas.cn

© 植物学报 Chinese Bulletin of Botany

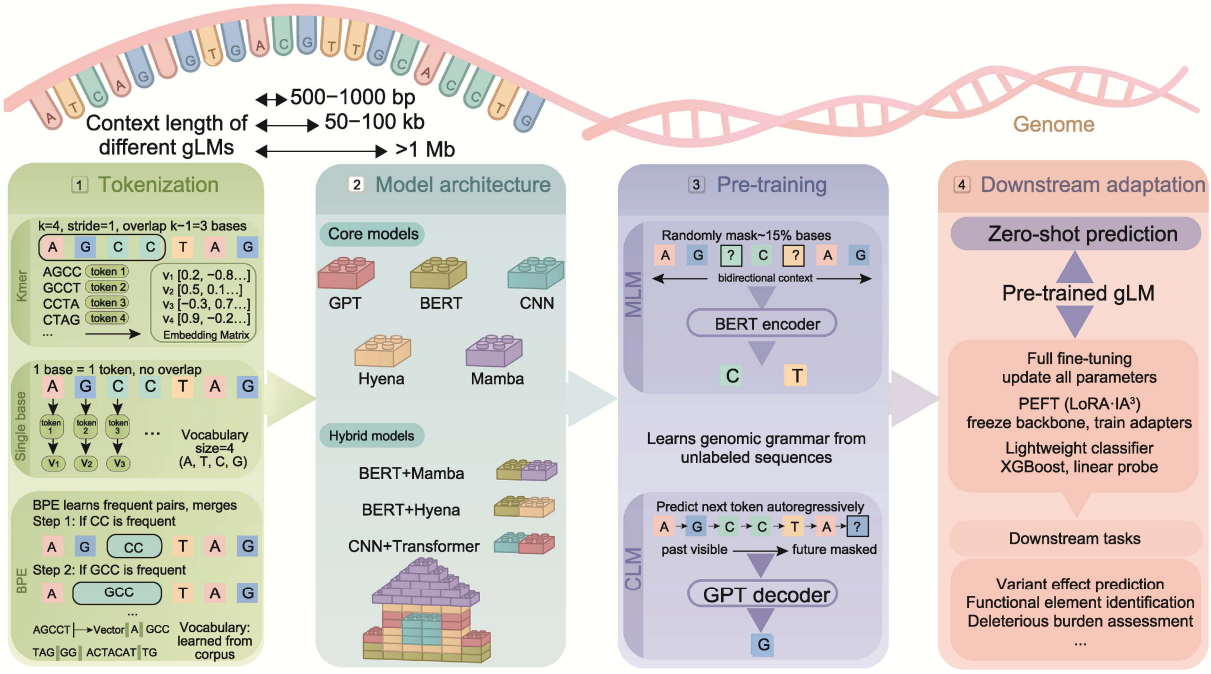

图1 基于DNA序列的植物基因组语言模型(gLMs)构建与应用流程

上下文长度，即模型单次处理的序列长度，在不同的模型中为数百碱基至1 Mb不等。(1) 分词(词元化): 原始DNA序列经分词切分为模型可处理的基本单元(词元)，模型每次处理的词元数量上限由上下文长度决定，主要包括3种分词策略——k-mer分词(滑动窗口，k=3，步长为1，相邻词元重叠k–1个碱基)、单碱基分词(每个碱基独立编码，词汇表大小为4)和BPE分词(基于语料库统计迭代合并高频碱基对，生成长度可变的词元)。各策略所得词元后续会映射为稠密的高维数值向量(词嵌入)，作为模型训练的输入表征。(2) 模型架构: 可选用编码器(BERT)、解码器(GPT)、卷积神经网络(CNN)或状态空间模型(Hyena和Mamba)作为基础构建模块，亦可采用混合架构(如BERT+Mamba、BERT+Hyena和CNN+Transformer)以兼顾局部特征提取与长程依赖建模。(3) 自监督预训练: 无须标注数据，通过两类任务使模型学习基因组序列的内在“语法”规律: 掩码语言建模(MLM，随机遮盖约15%的碱基，利用双向上下文预测被遮盖内容)和自回归因果语言建模(CLM，逐步预测下一个碱基，仅利用历史序列信息)，从而获得通用基因组表征。(4) 下游任务适配: 预训练模型可直接进行零样本预测，或通过全量微调、参数高效微调(PEFT，如LoRA和IA^3)及轻量级分类器(如XGBoost)等策略调整模型参数，以适配遗传变异效应预测、功能元件识别(启动子和增强子等)、有害负荷评估和剪接位点预测等具体下游任务。

Figure 1 Construction and application workflow of DNA sequence-based plant genomic language models (gLMs)

The context length, defined as the sequence length processed by the model in a single pass, ranges from several hundred bases to 1 Mb across different models. (1) Tokenization: The raw DNA sequence is segmented into basic units (tokens) that the model can process, with the maximum number of tokens per input determined by the context length. Three main tokenization strategies are used: k-mer tokenization (sliding window, k=3, stride=1, with k–1 overlapping bases between adjacent tokens), single-base tokenization (each base encoded independently, vocabulary size=4), and BPE tokenization (iteratively merging frequent base pairs based on corpus statistics to generate variable-length tokens). The resulting tokens are subsequently mapped to dense, high-dimensional numerical vectors (embeddings) as input representations for model training. (2) Model architecture: Core building blocks include encoders (BERT), decoders (GPT), convolutional neural networks (CNNs), and state space models (Hyena and Mamba). Hybrid architectures (e.g., BERT+Mamba, BERT+Hyena, CNN+Transformer) can also be adopted to balance local feature extraction with long-range dependency modeling. (3) Self-supervised pre-training: Without requiring labeled data, models learn the intrinsic "grammar" of genomic sequences through two training objectives: masked language modeling (MLM, randomly masking ~15% of bases and predicting them from bidirectional context) and causal language modeling (CLM, autoregressively predicting the next base using only past sequence information), yielding general-purpose genomic representations. (4) Downstream adaptation: Pre-trained models can be applied directly *via* zero-shot prediction, or fine-tuned using strategies such as full fine-tuning, parameter-efficient fine-tuning (PEFT, e.g., LoRA, IA^3), or lightweight classifiers (e.g., XGBoost) to adapt to specific downstream tasks, including variant effect prediction, functional element identification (promoters, enhancers, etc.), deleterious burden assessment, and splice site prediction.

性能和训练方式, 阐述了植物基因组高重复序列和多倍体等特有挑战, 最后展望了未来的发展方向。

1 模型的理论基础

1.1 基础模型基本概念

大语言模型通常指基于自监督预训练构建的基础模型, 具备理解与生成自然语言的能力。这类模型具有庞大的参数规模: 参数量可达十亿级别以上, 模型训练不依赖于大量标注数据, 而是从海量文本中学习泛化的语言表征。大语言模型通常遵循“尺度定律”(scaling laws), 即随着模型参数规模的扩大和训练数据量的增加, 其性能往往显著提升, 这种提升是非线性的, 达到某个临界点会出现涌现式能力跃迁(Kaplan et al., 2020; Wei et al., 2022)。然而, “大”模型这一概念并无严格的参数界限。当前普遍认为, 大语言模型的核心特征在于其可实现可扩展的知识编码与复杂推理能力, 而不仅仅取决于参数数量本身(Gou et al., 2021)。基于其强大的序列建模与知识表征能力, 大语言模型的应用正迅速扩展到自然语言处理以外的多个领域, 如基因组序列分析(Lam et al., 2024; Benegas et al., 2025b; Consens et al., 2025; Guo et al., 2025)。

自然语言处理(natural language processing, NLP)领域最具革命性的突破源于2017年提出的Transformer架构, 核心是其自注意力(self-attention)机制(Vaswani et al., 2017)。该架构在捕捉文本的长程依赖关系中展现出明显优势, 支持更广泛的序列上下文理解。这一进展也与GPU硬件算力的飞速发展密切相关; 注意力机制的计算是大规模矩阵运算, 与GPU的并行架构契合, 使得超大规模的模型训练变得可行。在架构层面, 本文不深入解释数学原理, 仅总结Transformer的核心机制: 引入位置编码(positional encoding)显式表示序列的顺序关系; 通过可学习的查询(query)、键(key)与值(value)矩阵实现自注意力机制; 通过多头注意力(Multi-Head Attention)机制并行捕捉多样化序列模式; 残差连接(residual connection)与层归一化(layer normalization)协同防止梯度消失(Vaswani et al., 2017) (表1)。在模型训练层面, 当前基于Transformer的预训练模型主要包括两类训练策略: 以BERT为代表的掩码语言建模(masked language modeling, MLM)方法, 通过随机遮盖输入中的单词, 使模型根据上下文进行预测训练, 因此在侧重于双向上下文理解的预测任务中具有优势(Devlin et al., 2019); 以GPT为代表的自回归因果语言建模(causal language modeling, CLM)方法, 从左到右逐个生成单词进行训练, 因此更适用于开放生成任务(Radford et al., 2018)。这两类模型普遍结合迁移学习(transfer learning)范式完成通用模型到专用模型的过渡, 利用少量标注数据对预训练模型进行微调, 即可适配特定的下游任务(Devlin et al., 2019) (图1)。

1.2 文本理解与基因组解读的异同

基因组语言模型的理论基础源于基因组序列与自然语言之间的结构相似性(Benegas et al., 2025b; Consens et al., 2025)。在自然语言中, 字母和汉字等最小语义单元按照特定的语法规则组合成更高层次的结构, 语法规则定义了词性和句式; 而在基因组序列中, 4种核苷酸(A、T、C、G)作为“字母”或“单词”, 依照遗传规则组成“句子”(基因), 最终形成“篇章”(基因组)。基因组的“语法”是一套复杂的调控规则, 还包括启动子和增强子等元件协同控制基因的表达。与自然语言处理的类比启发了基因组语言模型的构建, 自然语言文本经过分词(tokenization), 被切分成单词等基本语义单元后, 才能输入大模型进行预训练; 类似地, 基因组语言模型在训练时, 将DNA序列中的核苷酸碱基切分后作为建模的基本单元。在Transformer架构下, 两种模型均表现出捕捉上下文依赖的能力。这种上下文依赖性体现在: 自然语言中, 一个词的含义高度依赖其语境, 比如“小米”指作物还是公司, 取决于其在农业文献还是商业报道; 基因组中, 同一基因在不同细胞类型或环境下承载着截然不同的功能, 一个A-G突变(SNP)是否有功能也取决于所处的序列背景。将DNA视为一种“语言”, 并借鉴自然语言处理的方法进行研究, 有助于理解生命信息的编码规则。

尽管基因组序列与自然语言存在诸多相似, 基因组“语言”仍有其独特之处。例如, 它的“字母表”极为简洁, 仅由4种碱基构成, 却组合出远比人类语言复杂的调控规则和元件类型, 且这些“规则”历经数十亿年演化形成(Jumper et al., 2021)。基因的起止边界和调控元件的位置往往是隐含而模糊的, 未有像

表1 关键概念释义
Table 1 Definitions of key concepts

概念	释义
自监督训练 (self-supervised learning)	模型从数据中自动构造训练。例如，随机遮盖序列中部分碱基后，采用模型预测被遮盖内容，被遮盖的碱基本身即构成监督信号，无须人工标注。基于基因组序列进行植物基因组语言模型(gLMs)的预训练，不依赖稀缺且昂贵的湿实验标注数据
参数量与模型规模 (parameter and model scale)	参数是模型训练中学到的数值权重，决定如何将输入序列转化为预测结果。十亿参数即模型有 10^9 个可调节数值，相当于一个有十亿个变量的方程组
注意力机制 (attention mechanism)	Transformer 的核心计算单元。模型在处理序列中某个碱基时，会同时“回望”序列中所有其它位置，判断哪些位置与当前位置最相关，并据此分配不同的关注权重。具体实现上，通过查询(query)与键(key)矩阵计算序列各位置之间的相关性分数，再以此为权重对值(value)矩阵加权汇聚信息。多头注意力机制是指并行运行多组独立的注意力计算，每组自主学习序列中不同类型的依赖关系，各组结果最终拼接融合，输出统一的上下文感知表征
嵌入/向量表征 (embedding/vector representation)	将离散的输入转换为稠密高维数值向量的过程。在语言模型中，每个词元映射为高维空间的一组向量，语义相近的词向量也彼此靠近；在植物基因组模型中，功能相似的序列片段其向量亦相互靠近。嵌入是模型理解序列的第一步，也是下游任务计算(细胞类型注释和变异效应预测等)直接使用的特征表示
过拟合与泛化能力 (overfitting & generalization)	过拟合指模型在训练数据上表现极好，但遇到新数据时性能急剧下降，相当于“背答案”而非“理解规律”。泛化能力指将训练中学到的规律迁移至未见数据的能力，是衡量模型实用价值的核心指标。例如，用拟南芥(*Arabidopsis thaliana*)数据训练的模型能否准确预测玉米(*Zea mays*)变异效应。参数量过大且训练数据有限时过拟合风险上升
状态空间模型 (state space model, SSM)	Transformer之外的另一类序列建模架构。Transformer 的自注意力机制需两两比较序列所有位置对，计算量随序列长度增加呈平方增长。状态空间模型以线性递推或卷积替代注意力，将计算复杂度降至与序列长度呈线性关系，使超长基因组建模在算力上变得可行
零样本预测 (zero-shot prediction)	预训练模型在完全不使用目标任务标注数据的情况下，直接对新样本进行预测。模型仅凭在海量多物种基因组序列上学到的序列保守性与演化模式来推断变异的功能影响
迁移学习 (transfer learning)	指将已在某一领域大规模数据上训练好的模型迁移至新领域或新任务，使其适配目标任务的一类方法

自然语言中“空格”或“标点”那样明确的切分标志(Consens et al., 2025)。因此，不同模型采用不同的分词策略，有的以单碱基为最小建模单元，追求最小分辨率的序列表征，有的使用长度为k的重叠短序列(k-mer)或字节对编码(byte pair encoding, BPE)等基于统计的分词算法，将碱基组合作为最小建模单元(Ji et al., 2021; Zhou et al., 2024; Liu et al., 2025)。分词分辨率的不同直接影响模型对序列信息的感知和建模能力。

尽管基因组语言模型的预训练不依赖标注数据，但将预训练模型应用于具体下游任务时，常需要一定量的功能注释数据进行微调。与自然语言丰富的标注数据不同，带有准确功能注释的基因组数据极其稀缺，确定某个突变的功能或某个元件的角色，仍需依赖高成本的湿实验(如基因敲除)验证，这极大限制了高质量监督学习数据集的获取(Benegas et al., 2025b)。

1.3 植物基因组学与人工智能融合

近年来，植物基因组学与语言模型技术加速融合，这得益于人工智能领域模型架构的发展，也源于植物基因组学中海量、高质量数据的不断积累(Fan et al., 2025)。自2000年拟南芥基因组完成测序与组装以来，植物基因组学领域不断突破多项关键技术，包括高保真(HiFi)长读长测序、泛基因组构建、单细胞与空间转录组技术，以及端粒到端粒(T2T)的完整基因组组装(Jayakodi et al., 2025)。与此同时，人工智能领域也经历了从神经网络、生成对抗网络，到Transformer架构以及BERT、GPT和扩散模型等一系列突破性进展(图2) (Bengesi et al., 2024)。到2021年左右，两大领域开始深度结合，催生了DNABERT、Nucleotide Transformer和Evo等一批面向基因组学的语言模型(Ji et al., 2021; Nguyen et al., 2024; Dalla-Torre et al., 2025)。在植物科学中，AgroNT等模型基于48个作物基因组进行训练(Mendoza-Revilla et al., 2024);

近期提出的PDLLMs则整合了10个物种和9类下游任务，采用5种架构与3种分词策略，构建出192个DNA语言基础模型，在多样化下游任务上微调，展现出良好的泛化能力和应用价值(Liu et al., 2025) (图2)。

植物基因组学研究中，新方法通常从模式植物(如拟南芥和水稻)入手，验证其可行性，待方法成熟后再逐步推广至更多物种。从纵向深度来看，基于大语言模型整合基因组序列、转录组及表观遗传数据等多层次信息研究目前仍处于起步阶段。此类研究依赖于模式植物中已积累的高质量多组学数据(Benegas et al., 2025b)。从横向广度来看，面向跨物种比较的语言模型研究也已初步展开。当前植物领域已公开数千个物种的基因组组装序列，尤其是近年来基于HiFi测序技术完成的高质量参考基因组，为大规模预训练提供了重要的数据基础(Consens et al., 2025)。从陆地植物“生命之树”来看，虽然高质量参考基因组数量可观，但在系统覆盖与标准化程度上仍存在不足：截至2024年，约95%的属、70%的科与51%的目尚无参考基因组(Wang et al., 2026)。新近提出的植物星球计划(PLANeT)倡议在陆地植物演化的主要分支中系统生成高质量、标准化的参考基因组，结合人工智能解析跨谱系的保守调控原则与功能规律，服务于基础植物学研究(Wang et al., 2026)。

作物育种中的基因组预测(genomic prediction, GP)可显著提高育种效率，但其发展受限于传统分子标记的框架。该框架难以将基因组抽象为离散、位置

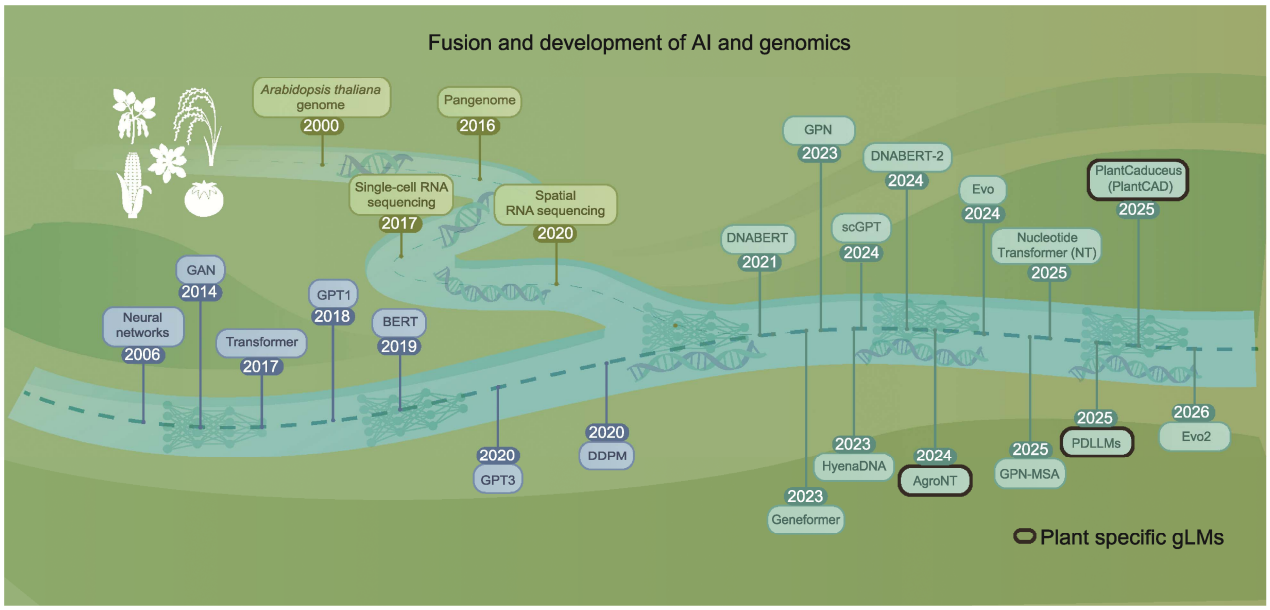

图2 人工智能(AI)与基因组学的技术融合与发展历程

基因组学技术自首次实现拟南芥全基因组组装(2000年)以来快速发展，相继涌现出泛基因组学、单细胞RNA测序和空间转录组等关键测序技术。与此同时，人工智能领域自2006年神经网络兴起后，陆续出现生成对抗网络(GAN)和Transformer架构等重要突破，并随之诞生了GPT、BERT系列模型以及去噪扩散模型(DDPM)等代表性算法。近5年来，两大技术领域深度融合，推动了一系列基因组学分析工具的创新，包括DNABERT和GPN等通用模型，以及近期面向植物研究的专用模型，如AgroNT、PlantCAD系列和PDLLMs。

Figure 2 Technological integration and evolution of artificial intelligence (AI) and genomics

Since the first draft genome assembly of *Arabidopsis thaliana* in 2000, genomics technologies have advanced rapidly, leading to the development of many important sequencing methodologies such as pangenomics, single-cell RNA sequencing and spatial transcriptomics. At the same time, the field of AI has witnessed major breakthroughs since the resurgence of neural networks in approximately 2006, including generative adversarial networks (GANs), the Transformer architecture and subsequent representative algorithms, such as the GPT and BERT models and denoising diffusion probabilistic models (DDPMs). Over the past five years, the deep integration of these two technological domains has led to the development of a range of genomics analysis tools, including general-purpose models such as DNABERT and GPN, as well as specialized models for plant research such as AgroNT, the PlantCAD series and PDLLMs.

无关的变异集合，也难以编码类似于结构变异和调控元件上下文等复杂信息。如果基于基因组语言模型框架，将复杂信息以嵌入向量(embeddings)的形式表示，便可作为基因组预测方法的输入，有望进一步提高预测准确度，进而提高育种效率(Alagarasan et al., 2026)。

受制于参考基因组之外的高质量多组学注释数据稀缺，植物领域有监督的序列到功能(sequence-to-function)预测模型目前仍处于起步阶段。相比之下，以AlphaGenome模型为代表的人类基因组研究取得显著成效，该模型通过在大规模实验功能注释数据上进行多任务联合训练，实现对转录水平、染色质可及性和剪接位点等多类功能特征的同步预测(Avsec et al., 2021, 2026)。

2 植物基因组语言模型的应用

植物基因组语言模型(Plant gLMs)已在多个关键研究场景中展现出应用潜力。本文将重点围绕以下方向展开讨论：跨物种演化与功能预测、单细胞测序的注释分析以及多模态基因组信息的融合。

2.1 跨物种基因组语言模型

在长期演化过程中，植物基因组形成了高度保守的调控“语法”，也积累了物种特异的序列特征。以这些多物种基因组数据为基础，构建预训练语言模型，具有重要应用价值。按照训练数据的物种范围，现有模型大致可分为3类(表2)。(1) 通用基因组语言模型，以人类或动物基因组为主要训练对象，部分加入植物基因组作为补充，如DNABERT系列(Ji et al., 2021; Zhou et al., 2024)、HyenaDNA (Nguyen et al., 2023)和Caduceus (Schiff et al., 2024); (2) 跨物种大规模训练模型，在数百至逾十万个物种基因组上联合预训练，如Nucleotide Transformer (Dalla-Torre et al., 2025)、Evo系列(Nguyen et al., 2024)和GPN-MSA (Benegas et al., 2025a); (3) 植物专用模型，仅使用植物基因组数

表2 跨物种模型比较
Table 2 Comparison of cross-species models

模型	架构类型	训练数据	参数量	上下文	主要任务	参考文献
DNABERT	BERT	人类基因组	~86 M	512 bp	启动子/剪接位点/转录因子(TF)结合位点识别、变异效应预测	Ji et al., 2021
HyenaDNA	Hyena	人类基因组	0.45 M–54.6 M	1 kb–1 Mb	长序列建模、染色质可及性预测	Nguyen et al., 2023
Caduceus	BiMamba	人类基因组	~7.7 M	~131 kb	长序列变异效应预测、调控元件识别	Schiff et al., 2024
DNABERT-2	BERT	多物种135个基因组	~117 M	~2.5 kb	多物种功能注释、变异效应预测	Zhou et al., 2024
Nucleotide Transformer (NT)	BERT	人类和850个物种	500 M/2.5 B	v1: ~6 kb v2: ~12 kb	变异效应预测、调控元件识别	Dalla-Torre et al., 2025
GPN-MSA	BERT+MSA	100余种脊椎动物全基因组比对	~10 M	128 bp	人类编码/非编码区有害变异预测	Benegas et al., 2025a
Evo	StripedHyena	原核生物和噬菌体	7 B	131 kb	基因组生成、跨域变异预测、CRISPR设计	Nguyen et al., 2024
Evo2	StripedHyena2	全生命域	40 B	1 Mb	大规模基因组生成与设计、致病变异识别	Brixi et al., 2026
GPN	CNN	十字花科8个物种	~10 M	512 bp	零样本变异效应预测、有害负荷评估	Benegas et al., 2023
AgroNT	BERT	48个植物物种	1 B	~6 kb	8类功能注释任务和变异优先级排序	Mendoza-Revilla et al., 2024
PlantCaduceus (PlantCAD)	BiMamba	16种被子植物	20 M–225 M	512 bp	翻译起始位点、转录终止位点、剪接位点识别、有害变异鉴定	Zhai et al., 2025b
PlantCAD2	Mamba2	65种被子植物	88 M/311 M/676 M	8192 bp	跨物种功能注释(零样本)和染色质可及性预测	Zhai et al., 2025a
PDLLMs	5种架构系统比较	10个植物物种	多规模	多尺度	9类分析任务(启动子强度、长链非编码RNA和组蛋白修饰等)	Liu et al., 2025

据训练, 如GPN (Benegas et al., 2023)、AgroNT (Mendoza-Revilla et al., 2024)以及PlantCaduceus系列(Zhai et al., 2025a, 2025b)。

基因组语言模型在架构选择与应用性能上存在差异。早期的DNABERT确立了基因组预训练的基础架构, 仅采用512 bp的上下文, 使其难以捕捉增强子等远程调控信号(Ji et al., 2021); Hyena和Mamba等状态空间模型的出现将序列上下文延伸至百万碱基级别, 对重复序列和远距离调控元件的建模有所改善(Nguyen et al., 2023; Gu and Dao, 2024); Nucleotide Transformer等通用模型以广泛的训练物种换取跨物种泛化能力(Dalla-Torre et al., 2025); AgroNT和PlantCaduceus等植物专用模型则通过聚焦植物基因组特征, 在变异效应预测等任务上表现出色(Mendoza-Revilla et al., 2024; Zhai et al., 2025b)。

在实际应用中, 研究者对于模型的选择需要在上下文长度、参数规模与训练数据范围三者之间进行权衡(Benegas et al., 2025b)。具体而言, Evo和PlantCAD2等长上下文模型在解析远程调控关系上更具优势, 但存在时间成本高以及计算开销大的缺点(Nguyen et al., 2024; Zhai et al., 2025a); GPN等轻量级模型推理速度快、部署成本低, 但对远程依赖信息的建模能力有限(Benegas et al., 2023)。此外, 不同模型所用评估数据集不同, 横向比较仍缺乏统一基础, 未来构建并使用一套植物标准化数据集也很重要(Liu et al., 2025)。

2.1.1 多物种语言模型的变异效应预测

全基因组关联研究(genome-wide association study, GWAS)极大地推动了对植物复杂性状(如耐旱性、抗病性和产量)遗传基础的解析(Benegas et al., 2025b)。然而, 基因组中广泛存在的连锁不平衡(linkage disequilibrium, LD)导致与非因果变异高度连锁的多个位点会同时与性状显示统计关联, 产生大量假阳性信号, 使得精准定位因果基因变得异常困难(Benegas et al., 2023)。传统的功能实验虽是鉴定基因功能的“金标准”, 但其成本高昂、通量低, 限制了大规模应用(Benegas et al., 2023)。基因组语言模型为解决这一难题提供了新思路, 模型在大量基因组序列(通常涵盖多物种以捕获演化信息)上进行预训练, 理解基因组的“演化语法”, 进而具备功能元件注释、序列演化约束评估以及遗传变异效应预测能力(Benegas et al., 2023)。预训练模型通过量化特定位点发生突变前后(即野生型与突变型序列)的模型输出差异, 评估该突变对功能的潜在影响, 以此筛选出可能性最大的因果变异(Mendoza-Revilla et al., 2024; Nguyen et al., 2024; Zhai et al., 2025b; Brixi et al., 2026)。以AgroNT模型为例, 该模型采用大规模模拟诱变(large-scale in silico mutagenesis)策略, 对高置信度增强子或高表达启动子序列中的每一个碱基位点, 依次替换为其它3种碱基, 并通过计算突变所引起的对数概率变化(log-fold change, LFC)来量化其对功能的影响(Mendoza-Revilla et al., 2024)。这种方法无须湿实验, 能够直接从序列中推断碱基级别的功能重要性, 为大规模、高通量地注释罕见变异与非编码变异的生物学效应提供了解决方案。

为应对连锁不平衡带来的挑战, GPN模型开发了一种整合LD信息的统计方法。该方法综合考虑目标位点所在连锁区块内所有变异的功能重要性, 以更准确地评估其因果可能性(Benegas et al., 2023)。具体而言, 首先计算目标变异位点周围一定范围内所有位点的连锁强度, 获取该区域内每个位点的功能重要性评分, 通过加权整合的方式计算目标位点的综合功能评分(Benegas et al., 2023)。此评分通过整合连锁区块和序列功能信息来缓解连锁不平衡对因果变异定位的干扰(图3A)。

2.1.2 识别有害负荷

识别基因组中的有害负荷(deleterious burden)是变异效应注释的关键, 对植物育种和濒危群体的保护具有重要意义。传统方法主要依赖多序列比对(MSA)和系统发育模型(如phyloP和phastCons), 通过估算演化速率来推断位点的保守性(Ramani et al., 2019)。然而, 功能保守但序列快速演化的调控元件难以进行可靠的序列比对, 其突变效应往往无法被准确评估(Benegas et al., 2025a)。跨物种基因组语言模型为此提供了新途径。与GWAS中估计变异效应类似, 此类模型通过从大量物种的基因组序列中学习规律, 对演化上高度保守的功能区域形成强烈的“序列期望”, 表现为对这些区域野生型核苷酸的高预测置信度(Schiff et al., 2024; Benegas et al., 2025a; Zhai et al., 2025b)。当此类关键位点发生突变时, 模型的预测概

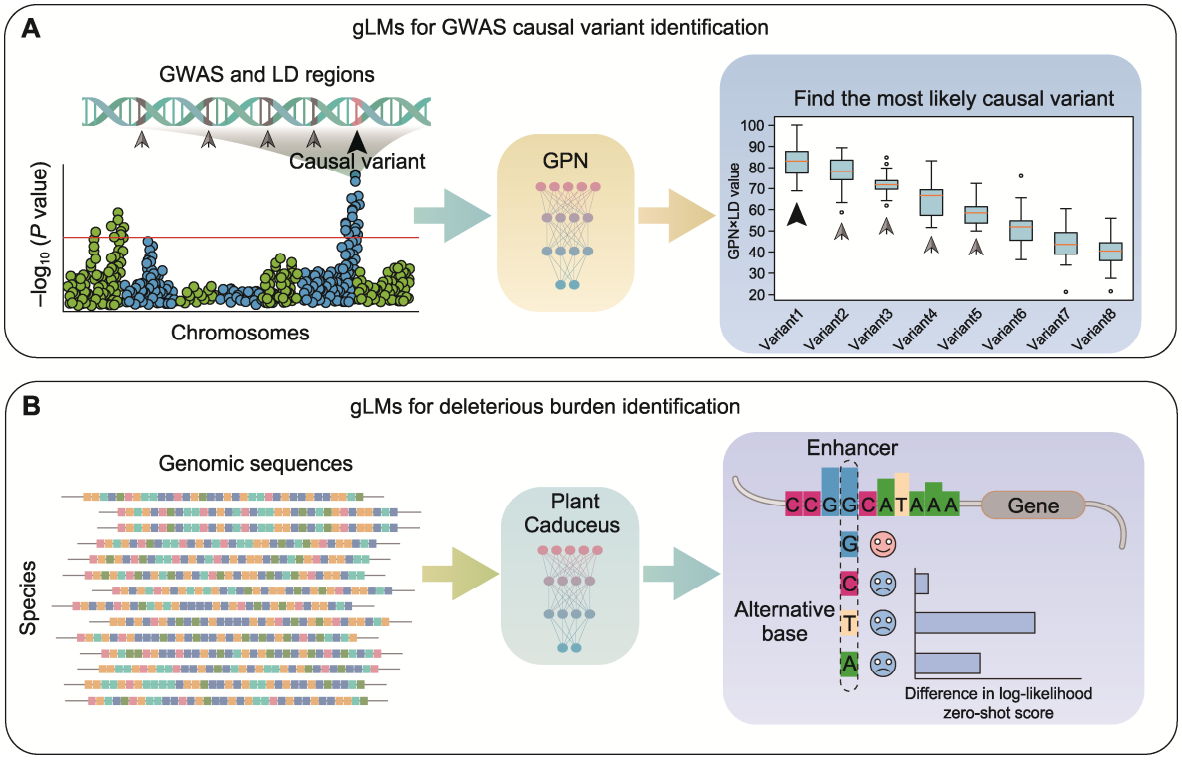

图3 基于基因组语言模型(gLMs)的变异效应预测

(A) 基于GPN模型的因果变异辨识(针对候选基因座中因连锁不平衡(LD)效应导致因果变异难以识别的问题，GPN模型通过结合其预测得分与LD信息，辅助筛选基因座中最有可能的因果变异，提升关键变异挖掘的准确性); **(B)** 基于PlantCAD模型的增强子与变异有害性预测(利用基因组序列构建PlantCAD模型，实现对增强子区域的识别与关键SNP定位。通过模型预测似然值的下降程度评估特定位点的进化约束，从而推断其变异影响(零样本评分))

Figure 3 Predicting variant effects based on genomic language models (gLMs)

(A) Causal variant identification based on the GPN model (causal variant identification within candidate loci is often confounded by linkage disequilibrium (LD), and the GPN model integrates prediction scores with LD information, to prioritize the most likely causal variants, thereby improving the accuracy of key variant identification); **(B)** Enhancer identification and deleterious variant prediction based on the PlantCAD model (this model is constructed from genomic sequences to identify enhancer regions and localize key SNPs. Evolutionary constraint at a given locus is assessed by the degree of decline in predicted log-likelihood, thereby inferring the effect (zero-shot score) of genetic variants)

率会出现扰动，其扰动幅度即预测似然性的下降程度与变异潜在的有害性呈强正相关(Schiff et al., 2024; Benegas et al., 2025a; Zhai et al., 2025b)。以PlantCaduceus模型为例，它能够基于多序列比对校准实现多物种全基因组范围内的潜在有害变异鉴定。该模型的预测结果得到了群体遗传学证据的支持，其所识别出的有害变异在自然群体中普遍表现出低等位基因频率的分布特征，符合负选择作用下的预期模式(Zhai et al., 2025b) (图3B)。

2.2 单细胞多组学数据解析

除基因组序列的建模外，gLMs的应用也延伸到单细胞转录组数据，此类模型被称为单细胞语言模型，核心是将基因表达水平抽象为一种“语言”，借此实现细胞类型识别和发育轨迹推断等多种分析任务。目前已发布了多个面向单细胞转录组数据的大规模预训练模型(表3)，如Geneformer、scBERT、CellPLM、scGPT、scFoundation和scExtract (Yang et al., 2022; Theodoris et al., 2023; Wen et al., 2024; Cui et al., 2024; Hao et al., 2024; Wu and Tang, 2025; Zhang et al., 2026)。在植物学领域，scPlantLLM已成功应用于拟南芥单细胞数据的聚类分析、细胞类型注释以及基因调控网络构建(Cao et al., 2025)。这些模型通常在海量单细胞数据上进行自监督预训练，并通

过下游微调适配细胞类型标注和功能预测等具体任务(Yang et al., 2022; Lan et al., 2024) (表3)。

在建模方式上，单细胞语言模型将每个基因ID视为一个词元，其表达量作为词元的数值属性，整个细胞的基因表达谱则构成描述其功能状态的“句子”，模型通过捕捉基因间的共表达模式以理解细胞背后的生物学特征(Lan et al., 2024)。由于细胞中基因的排列并不具有自然语言中的“语序”结构，模型通常依据基因表达量的大小对基因进行排序，并且以此为基础构建位置编码(positional encoding) (Yang et al., 2022; Li et al., 2024a)。预训练时，模型结合基因表达水平对细胞的高维表征进行动态调整(Yang et al., 2022; Cui et al., 2024)。在预训练任务的设计上，通常采用掩码语言建模策略，即随机遮盖15%的基因ID及其表达量，以训练模型恢复被遮盖信息的能力(Yang et al., 2022)。为降低连续表达值带来的计算复杂度，通常还引入分箱(binning)方法将其离散化处理(Yang et al., 2022; Hao et al., 2024)。通过训练，单细胞语言模型同时编码了基因的身份信息及其表达水平，支持细胞类型标注和功能预测等下游任务。

另一方面，以GenePT和scREADER为代表的模型通过融合文本知识来实现单细胞数据的分类任务，将基因的功能语义信息与表达水平信息进行整合(Li et al., 2024a; Chen and Zou, 2025)。该框架利用从大量植物学文献中提取的基因功能描述(如“参与光合作用”、“响应干旱胁迫”或“调控开花时间”)，将每个基因编码为包含丰富生物学语义的高维向量(Li et al., 2024a; Chen and Zou, 2025)。在这类模型中，细胞同时编码了基因身份、表达水平以及功能语义，在细胞类型标注任务中更具优势(图4)。

2.3 多模态基因组信息融合

植物基因组功能解析涉及DNA序列、转录组和表观遗传修饰等多层次信息，单一模态的建模往往难以全面捕捉基因组调控的复杂性。传统方法依赖多个相互独立的单模态模型分别处理各类预测任务，如转录因子结合位点识别、基因表达预测、RNA剪接分析和染色质可及性评估等，各模型之间缺乏信息交互，处于不同功能维度(Li et al., 2024b; Benegas et al., 2025b)。构建能够整合多层次基因组信息的多模态模型，是突破这一局限的关键方向。其建模的基本前提是同时具备基因组序列、转录组、组蛋白修饰和染色质可及性等多层次数据的系统性资源。目前具备这一条件的植

表3 单细胞模型的比较
Table 3 Comparison of single-cell models

模型	架构类型	训练数据	参数量	主要任务	参考文献
scBERT	BERT	人类scRNA-seq数据(150万细胞，无标注)	未报告	细胞类型注释、批次效应处理	Yang et al., 2022
Geneformer	BERT	人类单细胞转录组(3000万细胞)	v1: 10 M v2: 104 M/316 M	细胞类型注释、基因网络动力学预测、治疗靶点识别	Theodoris et al., 2023
CellPLM	Transformer	scRNA-seq+空间转录组数据	~80 M	细胞类型注释、去噪/补全、基因扰动预测	Wen et al., 2024
scGPT	GPT	人类单细胞转录组(3300万细胞)	~51 M	细胞类型注释、多批次整合、扰动响应预测和基因网络推断	Cui et al., 2024
scFoundation	非对称编码器-解码器	人类单细胞转录组(~5000万细胞)	~100 M	基因表达增强、药物响应预测、细胞类型注释、基因模块推断	Hao et al., 2024
scExtract	基于GPT-4	文献提取和scRNA-seq数据	基于GPT-4	全自动scRNA-seq数据注释与多数据集整合	Wu and Tang, 2025
GenePT	基于GPT-3.5	NCBI基因功能描述文本	基于GPT-3.5	细胞类型注释和基因属性预测(无须大规模scRNA-seq预训练)	Chen and Zou, 2025
scREADER	基于LLaMA-13B/GPT-3.5	基因功能描述，结合少量scRNA-seq	基于LLaMA-13B/GPT-3.5	跨物种细胞类型注释(利用基因功能语义跨物种迁移)	Li et al., 2024a
scPlantLLM	BERT	植物scRNA-seq数据(多物种，以拟南芥为主)	未报道	细胞聚类、细胞类型注释(零样本)、基因调控网络构建	Cao et al., 2025

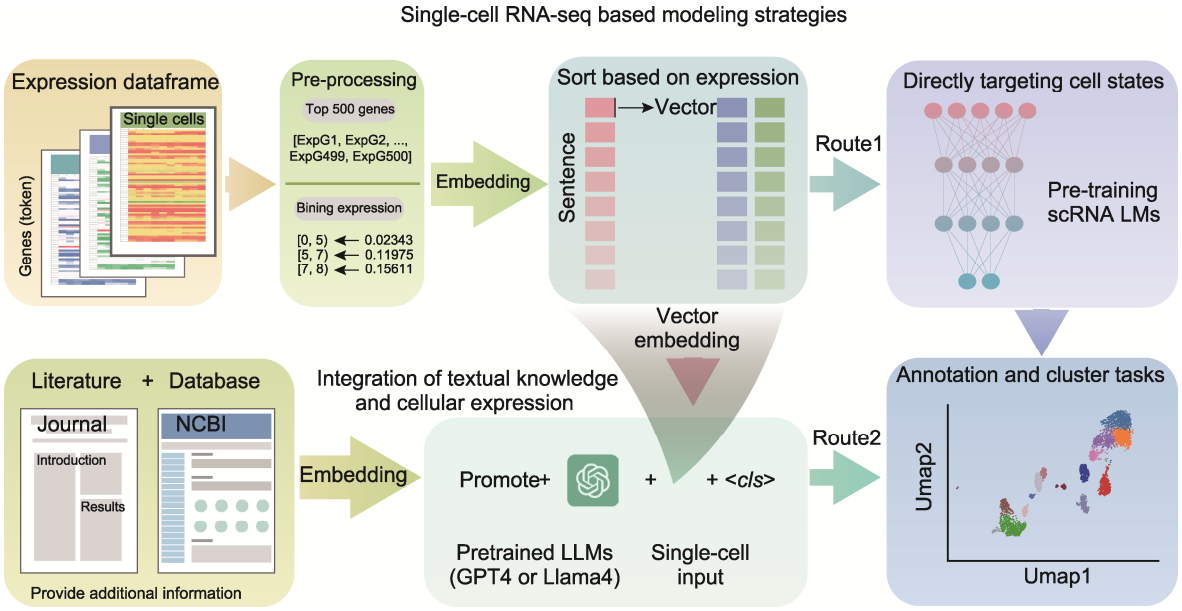

图4 单细胞RNA测序与语言模型融合的分析框架

图示单细胞转录组数据语言模型构建的两种策略。单细胞基因表达矩阵经预处理(如筛选高变异基因、表达量分箱离散化)后，通过排序嵌入或直接分箱转化为句子式向量。右侧为2条主要技术路径: Route 1 (直接靶向细胞状态)通过预训练单细胞语言模型，学习基因共表达模式，生成可用于细胞注释与聚类分析的嵌入表示; Route 2 (文本知识融合)则整合文献与数据库中的基因功能文本信息，结合预训练大语言模型(如GPT-4和Llama)，使模型在理解细胞的同时，能够参考基因背后的生物学先验知识。

Figure 4 An analytical framework that integrates single-cell RNA sequencing with language models

There are two strategies for constructing language models from single-cell transcriptomic data. The single-cell gene expression matrix is preprocessed (e.g., selection of highly variable genes, expression value binning and discretization) and then transformed into sentence-like vector representations *via* ordered embedding or direct binning. Two main technical routes are shown: Route 1 (direct modeling of cell states), a single-cell language model is pre-trained to learn gene co-expression patterns, generating embeddings suitable for cell type annotation and clustering analysis; Route 2 (integration of textual knowledge), textual information on gene functions from the literature and databases is incorporated, often using pre-trained large language models (e.g., GPT-4 or Llama), allowing models to draw on prior biological knowledge of genes when interpreting cell states.

物物种极为有限，仅拟南芥、水稻和玉米(*Zea mays*)等少数模式植物积累了相对完整的多组学数据(Lan et al., 2021; Han et al., 2023; Wang et al., 2025)；对绝大多数作物和野生植物而言，部分有基因组数据，但是表观遗传和转录调控等数据稀缺(Benegas et al., 2025b)。此外，不同组学数据的质量标准、实验平台与注释体系也存在差异。在此背景下，植物多模态数据融合的研究目前仍处于起步阶段。

基于自然语言训练的大语言模型进行文本处理融合，是一条值得探索的补充路径。植物学领域的长期研究积累了大量以文本形式存在的基因功能信息：基因功能描述记载了基因的分子功能与生物学过程，表型数据库记录了基因变异与性状的关联，实验文献则提供了不同条件下的基因表达与调控信息。将这些文本知识经大语言模型编码为统一语义表征，与基因表达数据联合注入模型推理，可在多组学数据缺失时提供跨模态的信息补充，如马铃薯(*Solanum tuberosum*)功能基因组AI知识库(Li et al., 2026)。未来，随着植物多组学数据和文本知识的积累，有望显著提升植物gLMs的多模态融合能力(Guo et al., 2025)。

3 植物基因组的特点和应用框架

3.1 重复序列

植物基因组通常有60%–80%的序列由重复元件构成。转座元件(transposable elements, TEs)是重复序列的主要组成部分(Kress et al., 2022)。这些元件在演化上高度动态，在不同物种乃至同一物种的不同品

种间都存在差异, 对基因组语言模型的训练构成了独特挑战(Benegas et al., 2023)。重复序列的序列特异性低, 生物学功能高度依赖其在基因组中的插入位置、调控环境以及与邻近基因的相互作用, 这种位置依赖性和上下文敏感性使得模型难以捕捉具有生物学意义的编码或调控信号(Karollus et al., 2024)。此外, 重复序列占比高, 基因组模型易偏向学习重复序列的特征, 从而降低对启动子和基因等关键功能元件的理解能力(Benegas et al., 2025b)。

为应对这一挑战, 目前已提出并应用了两种策略。其一是在构建训练数据集时采用屏蔽或降采样(down-sampling), 即识别并剔除重复序列区域, 或降低其被采样的概率(Zhai et al., 2025b)。其二是在模型训练阶段引入损失加权(down-weighting)机制, 调整重复序列区域的权重(Benegas et al., 2023, 2025a)。例如, 在PlantCaduceus和GPN等模型中, 被注释为重复区域的token权重被设置为0.1–0.2, 而非重复区域则保留权重1.0, 这使模型更专注于非重复区域的特征学习(Benegas et al., 2025a; Zhai et al., 2025b)。

长上下文建模技术为解析植物基因组中的重复序列提供了新思路。早期基于Transformer的模型受限于自注意力机制的计算复杂度, 输入长度通常被限制在数千碱基以内(Poli et al., 2023)。近年来, 状态空间模型(如Mamba)、Hyena架构(如HyenaDNA)与高效Transformer架构(如Longformer和ETC)已能够处理数万至数十万碱基的超长序列(Ainslie et al., 2020; Beltagy et al., 2020; Poli et al., 2023; Gu and Dao, 2024)。在植物基因组研究中, PlantCAD2模型采用Mamba2架构, 将上下文窗口扩展至8 kb。扩大的上下文窗口使得重复序列能够被置于更大的基因组语境中进行解析, 将分散的重复元件视为功能整体, 从而捕捉其与周边序列之间的位置依赖性调控关系(Zhai et al., 2025a)。

3.2 多倍体

多倍体是植物基因组的另一显著特征。许多现存作物为古多倍体或新形成的同源/异源多倍体。例如, 普通小麦(*Triticum aestivum*)为异源六倍体, 由A、B和D三套亚基因组融合而成。在演化过程中, 不同亚基因组之间常出现功能分化或亚基因组优势, 表现为表达水平、时序调控及环境应答等方面的差异(Levy and Feldman, 2022)。多倍体基因组语言模型的发展对作物育种具有重要实践意义。甘蔗(*Saccharum* spp.)、油菜(*Brassica napus*)和陆地棉(*Gossypium hirsutum*)等重要作物均为多倍体, 其表型优势与亚基因组之间的基因剂量效应和功能互补密切相关。面向多倍体的专用语言模型, 有望直接服务于基因组辅助育种, 为精准解析多倍体作物复杂性状遗传基础提供新的工具(Cheng et al., 2025)。

尽管一些基础模型使用了多倍体植物, 但仍然做了简化处理。例如, PlantCAD仅使用六倍体小麦的A亚基因组进行预训练, 规避了同源序列混淆问题, 但代价是丢失了亚基因组间的互作信息(Zhai et al., 2025b)。多倍体植物的建模可能受限于3个层面。第一, 同源亚基因组之间的序列高度相似, 导致模型难以区分来自不同亚基因组的序列, 注意力权重可能在同源区域间产生混淆。第二, 亚基因组优势是多倍体植物的重要特征, 但现有模型缺乏将亚基因组序列信息与表达调控进行关联。第三, 多倍体物种的全基因组高质量组装仍较困难, 能够利用的参考基因组相对稀缺, 限制了训练数据的规模。

GPN-MSA基于多序列比对构建(Benegas et al., 2025a), 未来也可以探索基于亚基因组比对的多序列输入方式, 将不同亚基因组的同源区段以比对格式输入模型, 使模型同时学习各亚基因组的特异性特征及其互作关系。在特征表示层面, 引入亚基因组标识符作为序列标签, 也可使模型明确区分序列来源。

3.3 科级基因组语言模型

跨物种基因组语言模型在科(family)水平上预训练有可能构成一个平衡的应用尺度。若所选物种过于单一或亲缘关系过近, 模型虽能捕捉该狭窄谱系内的细微特征, 却难以提炼出普适的生物学规律, 一旦迁移至远缘物种, 其性能往往急剧下降(Benegas et al., 2025b)。反之, 若物种选择过于广泛或亲缘关系过远, 模型虽被迫学习所有真核生物中最核心、最保守的“底层规律”, 却面临两大挑战: 不同类群间生物学机制差异巨大(如染色质结构、基因组织方式与调控逻辑), 模型难以从异构的数据中提取统一规律; 多样类群的数据可能相互干扰, 导致模型在具体类群上的预测精度下降(Karollus et al., 2024; Albors et al.,

2025)。因此，预训练物种的选择策略对模型构建至关重要。同一科的物种拥有较近的共同祖先，基因组在结构、基因内容及调控元件上具有同源性，使模型能够高效准确地学习(Albors et al., 2025)。同时，科内物种间存在足够的自然变异，迫使模型进行抽象与概括，从而获得较强的泛化能力(Zhai et al., 2025b)。在植物基因组语言模型中，可采用科内代表性物种的(亚)基因组进行建模，如禾本科或茄科。先前基于十字花科8个基因组构建的GPN模型，已初步验证科级预训练的可行性(Benegas et al., 2023)。科级模型有望成为未来构建层次化、模块化植物界模型生态系统的重要组成部分。

数据充分的前提下，在属内进行泛基因组级别建模也具有可行性，近期发布的OneGenomeRice (OGR)成为首个水稻专属的基础模型，通过对257份栽培稻和165份野生稻基因组进行训练，在长程调控建模和群体变异信号补充方面优于现有的跨物种模型(Qian et al., 2026)。跨物种基因组语言模型在不同尺度下的建模策略各有优势，选择时应取决于具体数据集以及生物学问题。

3.4 植物基础模型的部署策略

基因组语言模型的实际部署需综合考虑计算资源、任务特性与性能需求等多方面因素。从技术实现机制来看，主流部署方式主要包括预训练-微调(pre-training+fine-tuning)、检索增强生成(retrieval-augmented generation, RAG)以及模型蒸馏(model distillation)等策略(Hinton et al., 2015; Lewis et al., 2020; Gou et al., 2021)。

预训练-微调方法包括全量微调(full fine-tuning)与参数高效微调(parameter-efficient fine-tuning, PEFT)。全量微调更新模型全部参数，计算成本高且容易过拟合。PEFT则通过冻结主干参数、仅训练少量附加结构来提升效率(Houlsby et al., 2019)。例如，低秩适应(low-rank adaptation, LoRA)通过训练2个低秩矩阵模拟参数更新过程，成为当前的重要方法之一(Hu et al., 2022)。PlantCaduceus仅通过训练XGBoost分类器，在拟南芥数据微调后成功迁移应用到玉米基因组(Zhai et al., 2025b)。AgroNT在48个植物物种上进行预训练，其采用IA^3参数高效微调技术实现对不同任务的快速适应(Liu et al., 2022; Mendoza-Revilla et al., 2024)。

检索增强生成(RAG)通过在推理阶段引入外部知识库，增强模型在专业领域的表现能力(Lewis et al., 2020)。在生物信息学中，GenePT和scREADER等研究成功应用了RAG：模型从外部知识库中检索与目标基因相关的生物学文本描述(如GO注释、KEGG通路信息或表型数据)，将这些上下文与原始基因特征向量组合成增强提示，输入至生成模型进行后续预测(Li et al., 2024a; Chen and Zou, 2025)。类似的案例还有DeepPGDB和PlantGPT等(Li et al., 2025; Zhang et al., 2025)。

知识蒸馏是一种重要的模型压缩技术，旨在将大型模型的能力迁移至轻量化的模型。在蒸馏过程中，大型模型作为教师模型输出概率分布(“软标签”)充当监督信号，学生模型通过模仿该分布进行训练，以继承教师的推理能力(Gou et al., 2021)。

4 总结与展望

植物基因组语言模型为遗传和功能解析提供了新的方法，具有重要应用价值，但仍处于起步阶段。目前，功能标注数据的稀缺仍是制约其应用于植物领域的瓶颈。确定变异的表型效应或验证元件的调控功能，依赖成本高昂的湿实验，这使得可用于有监督微调的数据集难以获得，影响模型的应用。此外，当前模型存在可解释性问题，模型做出特定预测时，其依据的序列规律是否与已知生物学机制一致，目前仍缺乏系统验证。不同研究采用的评估数据集各异，缺乏标准化基准进行横向比较。在算力层面，尽管一些高效架构有望缓解长上下文预训练的硬件压力，但训练与部署成本仍是限制模型普及的因素。

展望未来，以下几个方向可能会有新的突破。科级系统发育建模方面，以禾本科和茄科等科为单位构建专用gLMs。引入亚基因组标识符、发展多亚基因组并行建模架构，实现多倍体整合建模。对多模态数据进行整合，将基因组序列、转录组和表观遗传修饰等多层次信息纳入模型训练框架。将文献知识编码为语义表征并进一步整合，使模型具备更全面的能力。

随着相关领域研究的不断深入，植物基因组语言模型有望从序列解析工具拓展至序列设计工具，可探索人工启动子和增强子等顺式元件的从头设计，服务合成生物学和精准育种。更长远来看，植物基因组语

言模型的应用将与现有的基因组选择和表型预测等育种环节深度结合, 为作物改良提供全新范式, 通过进一步整合环境和生态等数据, 有望参与气候适应性育种以及抗逆基因挖掘等更复杂的生物学任务。

作者贡献声明

张枭和王楠: 构思、绘图、撰写及修改论文; 李琼: 指导及审阅论文。

利益冲突: 所有作者均声明不存在利益冲突

Conflict of Interests: The authors declare that there is no conflict of interests

©The author(s) 2026. This is an open access article under the CC BY-NC-ND 4.0 license (https://creativecommons.org/licenses/by-nc-nd/4.0/)

参考文献

Agathokleous E, Rillig MC, Peñuelas J, Yu Z (2024). One hundred important questions facing plant science derived using a large language model. *Trends Plant Sci* **29**, 210–218.

Ainslie J, Ontañón S, Alberti C, Cvicek V, Fisher Z, Pham P, Ravula A, Sanghai S, Wang QF, Yang L (2020). ETC: encoding long and structured inputs in transformers. In: Proceedings of 2020 Conference on Empirical Methods in Natural Language Processing. Online: Association for Computational Linguistics. pp. 268–284.

Alagarasan G, Li HH, Xu BY, Crossa J, Varshney RK (2026). Genomic language model-based genomic prediction in plant breeding. *Trends Plant Sci* doi: 10.1016/j.tplants.2026.04.012.

Albors C, Li JC, Benegas G, Ye CZ, Song YS (2025). A phylogenetic approach to genomic language modeling. In: Proceedings of the 29th Annual International Conference on Research in Computational Molecular Biology. Seoul: Springer. pp. 99–117.

Avsec Ž, Agarwal V, Visentin D, Ledsam JR, Grabska-Barwinska A, Taylor KR, Assael Y, Jumper J, Kohli P, Kelley DR (2021). Effective gene expression prediction from sequence by integrating long-range interactions. *Nat Methods* **18**, 1196–1203.

Avsec Ž, Latysheva N, Cheng J, Novati G, Taylor KR, Ward T, Bycroft C, Nicolaisen L, Arvaniti E, Pan JS, Thomas R, Dutordoir V, Perino M, De S, Karollus A, Gayoso A, Sargeant T, Mottram A, Wong LH, Drotár P, Kosiorek A, Senior A, Tanburn R, Applebaum T, Basu S, Hassabis D, Kohli P (2026). Alphagenome: advancing regulatory variant effect prediction with a unified DNA sequence model. *Nature* **649**, 1206–1218.

Barbadilla-Martinez L, Klaassen N, van Steensel B, de Ridder J (2025). Predicting gene expression from DNA sequence using deep learning models. *Nat Rev Genet* **26**, 666–680.

Beltagy I, Peters ME, Cohan A (2020). Longformer: the long-document transformer. [2024-07-15]. https://doi.org/10.48550/arXiv.2004.05150.

Benegas G, Albors C, Aw AJ, Ye CZ, Song YS (2025a). A DNA language model based on multispecies alignment predicts the effects of genome-wide variants. *Nat Biotechnol* **43**, 1960–1965.

Benegas G, Batra SS, Song YS (2023). DNA language models are powerful predictors of genome-wide variant effects. *Proc Natl Acad Sci USA* **120**, e2311219120.

Benegas G, Ye CZ, Albors C, Li JC, Song YS (2025b). Genomic language models: opportunities and challenges. *Trends Genet* **41**, 286–302.

Bengesi S, El-Sayed H, Sarker K, Houkpati Y, Irungu J, Oladunni T (2024). Advancements in generative AI: a comprehensive review of GANs, GPT, autoencoders, diffusion model, and transformers. *IEEE Access* **12**, 69812–69837.

Brixi G, Durrant MG, Ku J, Naghipourfar M, Poli M, Sun G, Brockman G, Chang DL, Fanton A, Gonzalez GA, King SH, Li DB, Merchant AT, Nguyen E, Ricci-Tam C, Romero DW, Schmok JC, Taghibakhshi A, Vorontsov A, Yang B, Deng M, Gorton L, Nguyen N, Wang NK, Pearce MT, Simon E, Adams E, Amador ZJ, Ashley EA, Baccus SA, Dai HY, Dillmann S, Ermon S, Guo DL, Herschl MH, Ilango R, Janik K, Lu AX, Mehta R, Mofrad MRK, Ng MY, Pannu J, Ré C, St. John J, Sullivan J, Tey J, Viggiano B, Zhu K, Zynda G, Balsam D, Collison P, Costa AB, Hernandez-Boussard T, Ho E, Liu MY, McGrath T, Powell K, Pinglay S, Burke DP, Goodarzi H, Hsu PD, Hie BL (2026). Genome modelling and design across all domains of life with Evo 2. *Nature* **652**, 1349–1361.

Cao GS, Chao HY, Zheng WQ, Lan YM, Lu KY, Wang YY, Chen M, Zhang H, Chen DJ (2025). ScPlantLLM: a foundation model for exploring single-cell expression atlases in plants. *Genom Proteom Bioinf* **23**, qzaf024.

Chen YQ, Zou J (2025). Simple and effective embedding model for single-cell biology built from ChatGPT. *Nat Biomed Eng* **9**, 483–493.

Cheng L, Bao ZG, Kong QQ, Lassois L, Stein N, Huang SW, Zhou Q (2025). Genome analyses and breeding of polyploid crops. *Nat Plants* **11**, 1714–1728.

Consens ME, Dufault C, Wainberg M, Forster D, Karimzadeh M, Goodarzi H, Theis FJ, Moses A, Wang B (2025). Transformers and genome language models. *Nat Mach Intell* **7**, 346–362.

Cui HT, Wang C, Maan H, Pang K, Luo FN, Duan N, Wang B (2024). ScGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nat Methods* **21**, 1470–1480.

Dalla-Torre H, Gonzalez L, Mendoza-Revilla J, Lopez Carranza N, Grzywaczewski AH, Oteri F, Dallago C, Trop E, de Almeida BP, Sirelkhatim H, Richard G, Skwark MJ, Beguir K, Lopez M, Pierrot T (2025). Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nat Methods* **22**, 287–297.

Devlin J, Chang MW, Lee K, Toutanova K (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: Association for Computational Linguistics. pp. 4171–4186.

Fan WY, Guo ZW, Wang X, Zhang LK, Liu YH, Cai CC, Zhang K, Cheng F (2025). Deep learning applications advance plant genomics research. *Hortic Plant J* **11**, 1791–1806.

Feng SH, Stiller J, Deng Y, Armstrong J, Fang Q, Reeve AH, Xie D, Chen GJ, Guo CX, Faircloth BC, Petersen B, Wang ZJ, Zhou Q, Diekhans M, Chen WJ, Andreu-Sánchez S, Margaryan A, Howard JT, Parent C, Pacheco G, Sinding MHS, Puetz L, Cavill E, Ribeiro ÂM, Eckhart L, Fjeldså J, Hosner PA, Brumfield RT, Christidis L, Bertelsen MF, Sicheritz-Ponten T, Tietze DT, Robertson BC, Song G, Borgia G, Claramunt S, Lovette IJ, Cowen SJ, Njoroge P, Dumbacher JP, Ryder OA, Fuchs J, Bunce M, Burt DW, Cracraft J, Meng GL, Hackett SJ, Ryan PG, Jønsson KA, Jamieson IG, da Fonseca RR, Braun EL, Houde P, Mirarab S, Suh A, Hansson B, Ponnikas S, Sigeman H, Stervander M, Frandsen PB, van der Zwan H, van der Sluis R, Visser C, Balakrishnan CN, Clark AG, Fitzpatrick JW, Bowman R, Chen N, Cloutier A, Sackton TB, Edwards SV, Foote DJ, Shakya SB, Sheldon FH, Vignal A, Soares AER, Shapiro B, González-Solís J, Ferrer-Obiol J, Rozas J, Riutort M, Tigano A, Friesen V, Dalén L, Urrutia AO, Székely T, Liu Y, Campana MG, Corvelo A, Fleischer RC, Rutherford KM, Gemmell NJ, Dussex N, Mouritsen H, Thiele N, Delmore K, Liedvogel M, Franke A, Hoeppner MP, Krone O, Fudickar AM, Milá B, Ketterson ED, Fidler AE, Friis G, Parody-Merino ÁM, Battley PF, Cox MP, Lima NCB, Prosdocimi F, Lee Parchman T, Schlinger BA, Loiselle BA, Blake JG, Lim HC, Day LB, Fuxjager MJ, Baldwin MW, Braun MJ, Wirthlin M, Dikow RB, Ryder TB, Camenisch G, Keller LF, DaCosta JM, Hauber ME, Louder MIM, Witt CC, McGuire JA, Mudge J, Megna LC, Carling MD, Wang B, Taylor SA, Del-Rio G, Aleixo A, Vasconcelos ATR, Mello CV, Weir JT, Haussler D, Li QY, Yang HM, Wang J, Lei FM, Rahbek C, Gilbert MTP, Graves GR, Jarvis ED, Paten B, Zhang GJ (2020). Dense sampling of bird diversity increases power of comparative genomics. *Nature* **587**, 252–257.

Gou JP, Yu BS, Maybank SJ, Tao DC (2021). Knowledge distillation: a survey. *Int J Comput Vis* **129**, 1789–1819.

Gu A, Dao T (2024). Mamba: linear-time sequence modeling with selective state spaces. In: Proceedings of the 1st Conference on Language Modeling. Philadelphia: OpenReview.net. https://arxiv.org/abs/2312.00752v1.

Guo F, Guan RC, Li YH, Liu Q, Wang XW, Yang C, Wang JX (2025). Foundation models in bioinformatics. *Natl Sci Rev* **12**, nwaf028.

Han LQ, Zhong WS, Qian J, Jin ML, Tian P, Zhu WC, Zhang HW, Sun YH, Feng JW, Liu XG, Chen G, Farid B, Li RN, Xiong ZM, Tian ZH, Li J, Luo Z, Du DX, Chen SJ, Jin QX, Li JX, Li Z, Liang Y, Jin XM, Peng Y, Zheng C, Ye XN, Yin YJ, Chen H, Li WF, Chen LL, Li Q, Yan JB, Yang F, Li L (2023). A multi-omics integrative network map of maize. *Nat Genet* **55**, 144–153.

Hao MS, Gong J, Zeng X, Liu CM, Guo YC, Cheng XY, Wang TF, Ma JZ, Zhang XG, Song L (2024). Large-scale foundation model on single-cell transcriptomics. *Nat Methods* **21**, 1481–1491.

Hinton G, Vinyals O, Dean J (2015). Distilling the knowledge in a neural network. [2026-07-05]. https//doi.org/10.48550/arXiv.1503.02531.

Houlsby N, Giurgiu A, Jastrzebski S, Morrone B, De Laroussilhe Q, Gesmundo A, Attariyan M, Gelly S (2019). Parameter-efficient transfer learning for NLP. In: Proceedings of the 36th International Conference on Machine Learning. Long Beach: PMLR. pp. 2790–2799.

Hu EJ, Shen YL, Wallis P, Allen-Zhu Z, Li YZ, Wang SA, Wang L, Chen WZ (2022). LoRA: low-rank adaptation of

large language models. In: Proceedings of the 10th International Conference on Learning Representations. Online: OpenReview.net.

Jayakodi M, Shim H, Mascher M (2025). What are we learning from plant pangenomes? *Annu Rev Plant Biol* **76**, 663–686.

Ji YR, Zhou ZH, Liu H, Davuluri RV (2021). DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. *Bioinformatics* **37**, 2112–2120.

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D (2021). Highly accurate protein structure prediction with alphafold. *Nature* **596**, 583–589.

Kaplan J, McCandlish S, Henighan T, Brown TB, Chess B, Child R, Gray S, Radford A, Wu J, Amodei D (2020). Scaling laws for neural language models. [2026-07-15]. https://doi.org/10.48550/arXiv.2001.08361.

Karollus A, Hingerl J, Gankin D, Grosshauser M, Klemon K, Gagneur J (2024). Species-aware DNA language models capture regulatory elements and their evolution. *Genome Biol* **25**, 83.

Kress WJ, Soltis DE, Kersey PJ, Wegrzyn JL, Leebens-Mack JH, Gostel MR, Liu X, Soltis PS (2022). Green plant genomes: what we know in an era of rapidly expanding opportunities. *Proc Natl Acad Sci USA* **119**, e2115640118.

Lam HYI, Ong XE, Mutwil M (2024). Large language models in plant biology. *Trends Plant Sci* **29**, 1145–1155.

Lan W, He GH, Liu MY, Chen QF, Cao JY, Peng W (2024). Transformer-based single-cell language model: a survey. *Big Data Min Anal* **7**, 1169–1186.

Lan YH, Sun RK, Ouyang J, Ding WB, Kim MJ, Wu J, Li YH, Shi TL (2021). AtMAD: *Arabidopsis thaliana* multi-omics association database. *Nucleic Acids Res* **49**, D1445–D1451.

Levy AA, Feldman M (2022). Evolution and origin of bread wheat. *Plant Cell* **34**, 2549–2567.

Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Küttler H, Lewis M, Yih WT, Rocktäschel T, Riedel S, Kiela D (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc. pp. 9459–9474.

Li C, Long QQ, Zhou YC, Xiao M (2024a). SCREADER: prompting large language models to interpret scRNA-seq data. In: Proceedings of 2024 IEEE International Conference on Data Mining Workshops. Abu Dhabi: IEEE. pp. 665–672.

Li FP, Chen JX, Luo W, Liu JY, Chen GD, Shuai BY, Hou ZW, Gan ZP, Zhao HY, Zhan PL, Bi CW, Wang ZF, Hu HF, Wang SK (2025). DeepPGDB: a novel paradigm for AI-guided interactive plant genomic database. *Plant Commun* **6**, 101494.

Li JY, Jia YD, Li FT, Su XQ, Luo JL, Dong YR, Sun CY, Cui QH, Wang L, Li AX, Shang Y, Huang SW, Zhu YJ, Jia YX (2026). An AI-powered knowledge hub for potato functional genomics. *Plant Commun* **7**, 101730.

Li Q, Hu ZH, Wang YX, Li L, Fan YM, King I, Jia GJ, Wang S, Song L, Li Y (2024b). Progress and opportunities of foundation models in bioinformatics. *Brief Bioinform* **25**, bbae548.

Liu GQ, Chen L, Wu YC, Han YS, Bao Y, Zhang T (2025). PDLLMs: a group of tailored DNA large language models for analyzing plant genomes. *Mol Plant* **18**, 175–178.

Liu HK, Tam D, Muqeeth M, Mohta J, Huang TH, Bansal M, Raffel CA (2022). Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans: Curran Associates Inc. pp. 142.

Mendoza-Revilla J, Trop E, Gonzalez L, Roller M, Dalla-Torre H, de Almeida BP, Richard G, Caton J, Lopez Carranza N, Skwark M, Laterre A, Beguir K, Pierrot T, Lopez M (2024). A foundational large language model for edible plant genomes. *Commun Biol* **7**, 835.

Najafabadi MY, Jackson SA (2026). Hybrid AI in synthetic biology: next era in agriculture. *Trends Plant Sci* **31**, 555–565.

Nguyen E, Poli M, Durrant MG, Kang B, Katrekar D, Li DB, Bartie LJ, Thomas AW, King SH, Brixi G, Sullivan J, Ng MY, Lewis A, Lou A, Ermon S, Baccus SA, Hernandez-Boussard T, Ré C, Hsu PD, Hie BL (2024). Sequence modeling and design from molecular to genome scale with Evo. *Science* **386**, eado9336.

Nguyen E, Poli M, Faizi M, Thomas AW, Sykes CB, Wornow M, Patel A, Rabideau C, Massaroli S, Bengio

Y, Ermon S, Baccus SA, Ré C (2023). HyenaDNA: Long-range genomic sequence modeling at single nucleotide resolution. In: Proceedings of the 37th International Conference on Neural Information Processing System. New Orleans: Curran Associates Inc. pp. 1872.

Poli M, Massaroli S, Nguyen E, Fu DY, Dao T, Baccus S, Bengio Y, Ermon S, Ré C (2023). Hyena hierarchy: towards larger convolutional language models. In: Proceedings of the 40th International Conference on Machine Learning. Honolulu: PMLR. pp. 28043–28078.

Qian BL, Liang CW, Qin C, Liu C, Zhang CL, Xu CY, Li DX, Xue GR, He H, Zhang H, He HY, Chen DY, Xu JW, Zhang JY, Sun J, Shang LG, Jiang JL, Xia KK, Zhong LY, Chen LL, Fan LJ, Liu LQ, Qin MM, Li Q, Zhu SJ, Ma SW, Liu SP, Zhang SY, Fu S, Wei T, Xu XL, Jia XY, Xu X, Jing Y, Xu Y, Zhao YN, Xue YL, Guo YF, Xiao Z, Li ZR, Li ZF, Yue Z, Deng ZQ (2026). OneGenomeRice (OGR): a genomic foundation model for rice. bioRxiv https://doi.org/10.64898/2026.04.21.719822.

Radford A, Narasimhan K, Salimans T, Sutskever I (2018). Improving language understanding by generative pre-training. San Francisco: OpenAI. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

Ramani R, Krumholz K, Huang YF, Siepel A (2019). Phastweb: a web interface for evolutionary conservation scoring of multiple sequence alignments using phastCons and phyloP. *Bioinformatics* **35**, 2320–2322.

Schiff Y, Kao CH, Gokaslan A, Dao T, Gu A, Kuleshov V (2024). Caduceus: bi-directional equivariant long-range DNA sequence modeling. In: Proceedings of the 41st International Conference on Machine Learning. Vienna: PMLR. pp. 43632–43648.

Theodoris CV, Xiao L, Chopra A, Chaffin MD, Al Sayed ZR, Hill MC, Mantineo H, Brydon EM, Zeng ZX, Liu XS, Ellinor PT (2023). Transfer learning enables predictions in network biology. *Nature* **618**, 616–624.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017). Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc. pp. 6000–6010.

Wang L, Amar M, Arbizu CI, Bowman JL, Cannon CH, Cantrill DJ, Cao XF, Chen SL, Chen XY, Chen XM, Chong K, Deng XW, E WN, Guo HW, He C, Hollingsworth P, Hu Q, Kong HZ, Leitch IJ, Li DZ, Li JY, Liu HT, Lucas W, Mizrachi E, Rakotoarinivo M, Ren H, Rieseberg L, Ruan J, Scherson R, Schneider H, Soltis DE, Sun H, Silva-Filho MC, Smith S, Twyford A, Usadel B, Van de Peer Y, Van den Berg C, Varshney RK, Viccini LF, Wang QF, Wang XQ, Wang YD, Wendel J, Wing R, Wong GKS, Xing YW, Xu X, Yan NE, Yang HM, Zhang SZ, Zhu YX, Soltis PS, Weigel D, Huang SW (2026). PLANeT: understanding and leveraging the genome of land plants for a sustainable future. *Cell* **189**, 2519–2532.

Wang XY, Huang HW, Jiang SJ, Kang JM, Li DW, Wang KL, Xie S, Tong C, Liu CF, Hu GH, Li HQ, Li C, Yang LW, Ding YK, Li ST, Wang FM, Lohmann JU, Liang Z, Gu XF (2025). A single-cell multi-omics atlas of rice. *Nature* **644**, 722–730.

Wei J, Tay Y, Bommasani R, Raffel C, Zoph B, Borgeaud S, Yogatama D, Bosma M, Zhou D, Metzler D, Chi EH, Hashimoto T, Vinyals O, Liang P, Dean J, Fedus W (2022). Emergent abilities of large language models. [2026-07-15]. https://doi.org/10.48550/arXiv.2206.07682.

Wen HZ, Tang WZ, Dai XN, Ding JY, Jin W, Xie YY, Tang JL (2024). CellPLM: pre-training of cell language model beyond single cells. In: Proceedings of the 12th International Conference on Learning Representations. Vienna: OpenReview.net. https://doi.org/10.1101/2023.10.03.560734.

Wu YX, Tang FC (2025). scExtract: leveraging large language models for fully automated single-cell RNA-seq data annotation and prior-informed multi-dataset integration. *Genome Biol* **26**, 174.

Xie LJ, Gong XJ, Yang K, Huang YJ, Zhang SY, Shen LT, Sun YQ, Wu DY, Ye CY, Zhu QH, Fan LJ (2024). Technology-enabled great leap in deciphering plant genomes. *Nat Plants* **10**, 551–566.

Xue HC, Xu ZG, Liu YJ, Wang L, Ming X, Wu ZY, Lian H, Han YW, Xu J, Zhang ZD, Shao QL, Liu K, Wang FX, Wang AH, Zhao J, Zhang JF, Chen JH, Mao YF, Wang JW (2025). A unified cell atlas of vascular plants reveals cell-type foundational genes and accelerates gene discovery. *Cell* **188**, 6370–6390.

Yang F, Wang WC, Wang F, Fang Y, Tang DY, Huang JZ, Lu H, Yao JH (2022). ScBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nat Mach Intell* **4**, 852–866.

Zhai JJ, Gokaslan A, Hsu SK, Chen SP, Liu ZY, Marroquin E, Czech E, Cannon B, Berthel A, Romay MC, Pennell M, Kuleshov V, Buckler ES (2025a). PlantCAD2: a DNA foundation model for interpreting genomes across flowering plants. [2026-07-15]. bioRxiv https://doi.org/10.1101/2025.08.27.672609.

Zhai JJ, Gokaslan A, Schiff Y, Berthel A, Liu ZY, Lai W, Miller ZR, Scheben A, Stitzer MC, Romay MC, Buckler ES, Kuleshov V (2025b). Cross-species modeling of plant genomes at single-nucleotide resolution using a pretrained DNA language model. *Proc Natl Acad Sci USA* **122**, e2421738122.

Zhang RX, Wang Y, Yang WY, Wen J, Liu WZ, Zhi SP, Li GZ, Chai N, Huang JQ, Xie YY, Xie XR, Chen LT, Gu M, Liu YG, Zhu QL (2025). PlantGPT: an *Arabidopsis*-based intelligent agent that answers questions about plant functional genomics. *Adv Sci* **12**, e03926.

Zhang YJ, Venkatesh MS, Theodoris CV (2026). Discovery of candidate therapeutic targets with Geneformer. *Nat Protoc* doi: 10.1038/s41596-026-01364-8.

Zhou ZH, Ji YR, Li WJ, Dutta P, Davuluri RV, Liu H (2024). DNABERT-2: efficient foundation model and benchmark for multi-species genomes. [2026-07-15]. https://doi.org/10.48550/arXiv.2306.15006.

Plant Genomic Language Models: Advances and Applications

Xiao Zhang[1*], Qiong Li[1*], Nan Wang[2*]

[1]Tropical Crops Genetic Resources Institute, Chinese Academy of Tropical Agricultural Sciences, Haikou 571101, China; [2]Genome Analysis Laboratory of the Ministry of Agriculture and Rural Affairs, State Key Laboratory of Tropical Crop Breeding, Shenzhen Branch, Guangdong Laboratory of Lingnan Modern Agriculture, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen 518124, China

Abstract The development of large language models (LLMs) has opened new avenues for genomics research, leading to the emergence of genomic language models (gLMs). gLMs treat genomic data as a biological language with intrinsic grammar, learning the regulatory logic embedded within the genome and offering new approaches for the systematic dissection of genetic regulatory mechanisms. This review focuses on the theoretical foundations and current research landscape of plant genomic language models, and discusses major research directions: the application of cross-species pretrained models to variant effect prediction and functional element annotation, the development of models tailored to single-cell data, and strategies for integrating multimodal genomic information. We further discuss how plant-specific genomic features, particularly repetitive elements and polyploidy, affect model construction, and review modeling strategies and deployment frameworks. We conclude with a perspective on the future development of plant genomic language models.

Key words plant genomics, artificial intelligence, large language models, multi-omics, genetic variation

Zhang X, Li Q, Wang N (2026). Plant genomic language models: advances and applications. *Chin Bull Bot* **61**, 571–587.

* Authors for correspondence. E-mail: zhxiao7777@gmail.com; liqiong4416@126.com; wangnan05@caas.cn

(责任编辑: 朱亚娜)