



深度学习在人类基因组研究中的应用与挑战*

李郁葱^①, 屠凯岭², 谢丹^{①△}

1. 四川大学疾病分子网络前沿科学中心 组学技术与生物信息研究室, 四川大学生物治疗全国重点实验室, 四川大学华西医院(成都 610041); 2. 成都市第三人民医院(成都 610014)

【摘要】 近年来,随着高通量组学技术的发展,人类基因组数据迎来了爆炸式增长。通过海量数据挖掘人类基因组潜在作用已成为功能基因组研究的一项挑战。尽管传统统计方法在过去小规模数据分析中取得了成功,但其在分析效率、多维度数据整合方面存在明显局限,难以满足日益增长的数据分析需求。深度学习技术的引入为该领域研究提供了一个新的思路。本文系统综述了深度学习在人类基因组研究中的应用进展。研究表明,在标记数据完备的情况下,基于判别式深度学习的计算方法,如卷积神经网络(CNN)和长短期记忆神经网络(LSTM)架构在基因组变异发现任务中表现出较高的准确性和效率。而基于生成式深度学习的计算方法,如大语言深度学习模型架构(LLM)有效利用自监督预训练策略实现复杂基因组信息的整合,在功能性基因组序列注释中表现优秀。此外,本文还探讨了LLM深度学习模型在多组学数据整合及预测中的应用。未来,随着长片段、高维度数据的积累,深度学习技术有望实现更复杂混合基因组信息的整合,在人类基因组研究中发挥更加重要的作用。

【关键词】 深度学习 人类基因组 大语言模型 多组学数据整合 综述

Applications and Challenges of Deep Learning in Human Genome Research

Li Yucong^①, Tu Kailing², Xie Dan^{①△}. 1. Laboratory of Omics Technology and Bioinformatics, Frontiers Science Center for Disease-Related Molecular Network, State Key Laboratory of Biotherapy, West China Hospital, Sichuan University, Chengdu 610041, China; 2. The Third People's Hospital of Chengdu, Chengdu 610014, China

△ Corresponding author

This work was supported by the National Natural Science Foundation of China (No. 82173383, No. 82241238, No. 32300522) and the 1·3·5 Project for Discipline Excellence, West China Hospital, Sichuan University (No. ZYYC23024).

【Abstract】 In recent years, the advent of high-throughput omics technologies has fueled an explosive growth in human genomic data. Uncovering the latent functions within this vast data has become a significant challenge in functional genomics research. While traditional statistical methods have proved successful for analyzing smaller-scale datasets in the past, they exhibit clear limitations in analytical efficiency and integrating multi-dimensional data, struggling to meet the escalating demands of contemporary genomic analysis. The introduction of deep learning (DL) technologies offers a novel paradigm for this field. This review systematically examines the advances in applying deep learning to human genomics research. Studies demonstrate that when ample labeled data is available, discriminative DL computational methods—such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory networks (LSTMs)—achieve high accuracy and efficiency in genomic variant discovery tasks. Furthermore, generative DL methods, particularly Large Language Models (LLMs) leveraging self-supervised pre-training strategies, effectively integrate complex genomic information and exhibit superior performance in functional genomic sequence annotation and gene regulation studies. This review also explores the application of LLMs in multi-omics data integration and prediction. Looking ahead, the continued accumulation of long-read sequencing and high-dimensional data is expected to enable DL technologies to integrate increasingly complex and heterogeneous genomic information, playing an increasingly crucial role in human genomics research.

【Key words】 Deep learning Human genomics Large language models Multi-omics data integration Review

* 国家自然科学基金项目(No. 82173383, No. 82241238, No. 32300522)和四川大学华西医院学科卓越发展1·3·5工程项目(No. ZYYC23024)资助

△ 通信作者

出版日期: 2026-07-20

深入探究人类基因组对于解析复杂疾病至关重要, 主要涉及基因组变异发现和功能序列注释。在低数据量时代, 传统统计推断方法, 如GATK pipeline^[1]、Control free C^[2]和pindel^[3]等, 通过比对测序数据与参考基因组, 有效识别单核苷酸变异(SNVs)和小片段插入/缺失(indels), 为疾病相关基因研究提供支持。同时, 全基因组关联研究(GWAS)、表观全基因组关联研究(EWAS)和表达数量性状位点(eQTLs)等方法利用线性模型解析群体变异与人类表型之间的关联, 发现了大量与人类表型或下游基因表达起调控作用的位点, 揭示了基因组变异与复杂疾病之间的潜在联系^[4-5]。表观遗传学测序方法, 如HiC技术^[6]和Chip-seq^[7]技术, 揭示了DNA甲基化、组蛋白修饰等表观遗传学修饰在基因表达调控中的重要作用, 为理解基因组的动态调控机制提供了新的视角。这些技术的发展极大地推动了人类基因组研究的进展。

随着测序技术的发展和数据量的增加, 人类基因组研究进入新阶段。传统统计推断模型和机器学习方法在处理大规模、高维度数据时面临诸多挑战, 例如模型复杂度、计算资源需求以及对数据标注的依赖等。深度学习作为一种强大的数据分析工具, 在图像识别和自然语言处理等领域取得了突破性进展^[8]。在人类基因组研究中, 深度学习被广泛应用于基因组变异检测和功能序列注释。例如, DeepVariant^[9]和Clair^[10-12]系列工具通过卷积神经网络(CNN)和循环神经网络(RNN)架构, 更准确地识别基因组变异, DNABERT^[13]和Nucleotide Transformer^[14]利用大语言模型架构用于基因组序列编码和分析, 图神经网络(GNN)用于预测基因调控关系和细胞对扰动的响应^[15]。然而, 深度学习在人类基因组研究中也面临挑战, 如需要大量标注数据和模型的“黑箱”特性。如何提高模型的可解释性和泛化能力, 以及整合多组学数据以提升性能, 是当前研究的热点问题。

综上, 深度学习为人类基因组研究带来新机遇和挑战。本文系统介绍深度学习方法在基因组变异检测和功能序列注释中的应用, 分析其优势和局限性, 以期为生物信息学研究者提供参考, 推动人类基因组研究的进一步发展。

1 基于深度学习策略的基因组变异检测方法

1.1 深度学习在基因组变异检测中的应用

基因组变异, 包括引发遗传病的胚系变异和驱动肿瘤发生的体细胞变异^[10], 是理解人类疾病机理与群体多样性的基石。高通量测序是检测此类变异的核心技术,

其配套分析方法也经历了显著的演进。传统方法依赖序列比对特征的统计建模。例如, 利用测序读长与参考基因组局部序列差异信息针对点突变发现构建的GATK流程^[1], 基于局部测序深度变化趋势针对拷贝数目变异构建的control free-C算法^[2]以及围绕读长-基因组比对断点信息针对大片段结构变异(SVs)构建的无监督学习策略Sniffles^[16], cuteSV^[17]等。这些模型通常围绕二倍体基因组设计, 通过经典统计学模型提高变异检测的准确率。经典统计推断策略具有较好的模型可解释性, 但同时也存在复杂的参数估计过程以及模型泛化能力问题。随着研究的积累, 大量胚系变异的金标准数据集被不断发掘出来, 为优化这些统计模型提供了重要依据。

为了实现复杂变异的高精度识别并有效利用测序数据的完整信息, 人们开始利用大规模标注数据集建立深度学习模型。针对点突变检测的算法有基于卷积神经网络的DeepVariant^[9]以及基于长短期记忆网络(LSTM)架构的Clair系列算法^[10-12]。这些算法通过深度学习策略有效提取和整理序列比对的信息, 并通过大规模标注数据的有监督学习实现了点突变的检测。相比传统统计推断算法, 深度学习方法大幅提升了流程执行效率与自动化程度, 可处理复杂突变, 且不受CG含量、重复序列等局部序列特征影响; 同时依托高效的特征提取整合能力, 其能应用于长读长高错误率序列的突变检测, 泛化性能更优。与仅影响小范围基因组区间的点突变不同, SVs的影响范围广, 相关信号可跨越数千甚至数百万碱基。而传统基于局部比对的统计模型, 仅关注单个位点或小区的碱基支持、比对质量等信息, 难以整合分析这类长距离全局性信号模式, 因此SVs的检测, 需要综合考量读长比对的断点信息。针对此类问题, 叶凯团队提出将序列-参考基因组比对结果转换成图像的形式^[18], 通过CNN架构实现变异的检测, 为长读长技术解析复杂SV提供新范式。在泛基因组研究深入推进的背景下, 叶凯团队的深度学习方法Swave延续序列到图像的核心建模思路^[19], 引入“投影波”技术提炼点图信号以剔除重复序列噪声, 通过双向LSTM捕捉长程上下文信息, 为群体遗传多样性研究和致病性结构变异挖掘提供了更强大的技术支撑。

1.2 深度学习在基因组变异检测中的局限性

尽管深度学习方法在基因组变异检测中取得了显著进展, 但目前仍存在两个主要弊端。首先, 当前方法多依赖于序列比对特征的提取, 但一些特殊的序列如串联重复、节段性重复及低质量基因组拼接导致的比对错误尚未根本解决; 其次, 目前基于深度学习的突变检测方法多集中在针对胚系变异的检测中, 体细胞变异尤其是体细

胞结构变异检测滞后。上述问题的根源在于当前可用于解析特殊重复序列或大片段体细胞结构变异的长读长测序数据尚未形成广泛规模和可靠的标注数据集,诸如CNN和LSTM这类围绕分类任务设计的深度学习框架难以在标注数据缺失的情况下得到有效训练。深入解决此问题一方面需要依赖于后续研究在标注数据集中的进展,如优质基准细胞系或真实人类组织基因组的优质数据集,以及传统统计模型和实验验证证据所带来的结构变异精确标注^[20-21]。另一方面,低标注数据集依赖的深度学习方法,如自监督训练框架的开发是标注数据集缺乏背景下的解决方案之一。

2 基于深度学习策略的功能性基因组序列注释方法

2.1 DNA大模型训练方式

除了序列变异发现外,如何注释DNA序列的功能一直以来都是困扰遗传学研究的重要问题。传统遗传学研究主要通过线性回归模型(如GWAS, EWAS, eQTL等)挖掘变异与表型之间的关联性,以此来注释功能性基因组序列。过去已经通过这类研究手段发现了大量的功能性基因组位点。这种基于经典统计模型的方法擅长检测存在线性关系的调控元件,但受制于线性回归模型有限的输入信息,对于更加复杂非线性调控关系的检测能力较差。得益于注意力机制^[22]和自然语言处理模型^[23]的进展,通过大语言模型理解DNA的语义结构成为了基因组序列功能解析的一大研究方向。不同于传统监督学习框架,DNA大语言基础模型不再依赖标注数据集,而是直接通过自监督学习策略实现模型参数的优化。类似于经典自

然语言处理模型Transformer的训练过程,DNA大语言模型将参考基因组DNA视作语料库,以掩码或下一词元预测的自监督学习策略作为其预训练的主要手段实现模型参数优化,使得模型具备识别输入DNA序列特征的能力。而后针对特殊功能性基因组预测或DNA序列设计任务,通过小规模标签数据的微调快速实现。这样的训练设计使得这类模型可以从公共数据库中直接获取海量的训练数据。相较于传统方法,这些大语言模型的核心优势在于能够捕捉基因组中的长片段序列信息特征,避免了手动降维、特征筛选等可能造成信息丢失的步骤;同时,自监督学习策略避免了模型对标注数据集的依赖。

不同于传统自然语言处理模型直接使用语言中的词汇分词方法构建词向量的输入形式,基因组DNA需要进行针对性分词才能被进一步应用于后续模型。为了实现高效的序列编码,获得尽可能长的输入序列的特征识别能力,DNA大语言模型主要形成了3种主要的序列分词模式(图1)。其中以DNABERT^[13]为主要代表的k-mer分词,将长序列分成固定长度k的子串,但难以确定最佳k值,可能引入偏差。DNABERT-2^[24]使用字节对编码(BPE),迭代合并DNA数据中的邻近高频核苷酸片段,形成可变量度的单词。这种方法利用统计学策略有效降低了高频重现序列引入的信息冗余,但可能遗漏低频序列单元。除了这两种分词方法外,以HyenaDNA等^[25-27]为代表的DNA大语言模型直接使用单碱基分词策略,最大限度保留原始DNA序列信息,通过更改模型架构,融合注意力机制和卷积算子,可高效处理长序列直接使用单核苷酸编码。这种分词策略显著提高了输入序列的长度,但相对地,该种分词方法具有较大的计算复杂度。

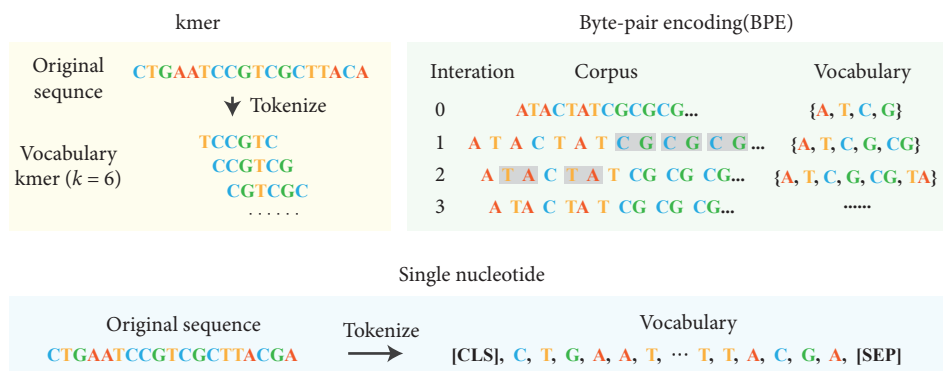


图 1 DNA大模型的主要分词方法

Fig 1 The main tokenization of the DNA large models

2.2 常用大模型介绍

在基因组研究常用的几个大模型架构中,Transformer^[22]采用编码器-解码器架构,应用中多简化为仅编码器

(BERT)^[23]或仅解码器(GPT)^[28]结构。编码器负责将基因组序列,如碱基序列、变异位点矩阵进行特征提取,转化为富含语义的高维特征向量,解码器则基于该特征向量

完成特定任务输出,例如序列注释、变异分型;仅编码器/解码器的简化设计,可根据任务需求减少冗余计算,更适配长基因组序列的高效处理。然而主流的Transformer模型计算复杂度高,尤其是在单核苷酸的分辨率下难以应用于超长序列中,因而催生了新的模型架构,旨在用卷积等其他机制替代注意力机制,以减少计算复杂度。HyenaDNA^[25]用隐式长卷积替代Transformer的自注意力机制,通过小型神经网络参数化卷积核,结合快速傅里叶变换实现长序列建模。Mamba^[29]基于选择性状态空间模型(SSM),将序列转化为状态空间中的线性更新过程,通过选择性门控聚焦序列关键信息(如基因组变异位点、功能元件),仅让核心特征参与状态更新,既突破了

Transformer自注意力的效率瓶颈,又避免了RNN的长序列记忆衰减。尽管Hyena与Mamba在长序列处理效率上实现了关键突破,但二者仍聚焦于单一维度的序列建模,难以同时满足基因组研究对多模态功能预测、单碱基分辨率输出及跨层级调控机制解析的综合需求。以AlphaGenome^[30]为主要代表的基因组学专用模型采用U-Net与Transformer混合架构,经下采样压缩、长程依赖建模、上采样还原实现单碱基分辨率输出,融合多任务学习与预训练、知识蒸馏双阶段训练,并通过序列并行处理与跳跃连接,平衡了超长基因组序列的处理效率与局部特征精度。代表性DNA大模型预训练信息如表1所示。

表 1 DNA大模型预训练信息

Table 1 The pretraining information of the DNA large models

Model	Data	Model size	Base model	Pretraining task	Training time
AlphaGenome	Human and mouse reference genomes, 5 930 human genome tracks, 1 128 mouse genome tracks	Not mentioned	U-Net-inspired backbone + Transformer blocks	supervised multimodal sequence-to-function prediction	Not mentioned
Caduceus	Human reference genome	1.9 million	Mamba	MLM	Not mentioned
DNABERT	Human reference genome	90 million	BERT	MLM	25 days on 8 Nvidia 2080Ti GPUs
DNABERT-2	Human reference genome, genome data of the 135 other species (including fungi, protozoa, mammals, invertebrates, etc.)	117 million	BERT	MLM	2 weeks on 8 Nvidia RTX2080Ti GPUs
Evo	2.7 million prokaryotic and phage genomes	7 billion	StripedHyena	CLM	Stage 1: 2 weeks on 64 Nvidia H100 GPUs; Stage 2: 2 weeks on 128 Nvidia A100 GPUs
Evo2	Over 128,000 species, including bacteria and archaea, humans, plants, etc.	7 billion and 40 billion	StripedHyena2	MLM	Not mentioned
HyenaDNA	Human reference genome	1.6 million - 1 billion	Hyena	CLM	80 minutes on 1 Nvidia A100 GPU
Nucleotide Transformer	Human reference genome, 3 203 human genome, 850 genome data of the other species	50 million - 2.5 billion	BERT	MLM	28 days on 128 Nvidia A100 GPUs

MLM: masked language modeling; CLM: causal language modeling.

2.3 深度学习在基因组序列注释中的研究进展

在利用深度学习研究基因组序列的早期,卷积神经网络(CNN)、图神经网络(GNN)等经典深度学习方法被广泛应用于基因组分析,CNN擅长捕捉DNA序列的局部基序特征,GNN则可对基因组的拓扑结构与互动关系进行建模,二者成为基因元件识别、简单序列分类等基础基因组任务的核心工具。Akita^[31]作为3D基因组建模的代表性模型,采用CNN与RNN结合的混合架构,同时创新性地引入对称注意力机制适配染色质接触图的对称结构,能够将一维DNA序列特征精准转化为二维基因组接触频率图谱,仅依靠DNA序列即可快速预测序列变异对基因组3D折叠的调控效应。

随着注意力机制在深度学习领域的快速发展,基于

Transformer架构、专门用于DNA序列分析的DNABERT^[13]应运而生,开启了基因组预训练基础模型的研究新阶段。不同于早期一个模型仅能完成一种任务的局限,通用型预训练模型通过在海量未标注基因组数据上完成自监督预训练学习序列特征,经简单微调后即可拓展应用于多种基因组学任务。DNABERT凭借对基因组序列的深度表征能力,在启动子识别、转录因子结合位点预测、DNA序列分类、非编码RNA识别等任务中均能实现优异性能。

在DNABERT的基础上,后续发展的Nucleotide Transformer^[14]和DNABERT-2^[24]进一步提升了模型性能与跨物种泛化能力,通过整合多物种基因组数据进行预训练,使其能够学习到更通用的物种间序列表示,不仅可完

成DNABERT支持的所有基础基因组任务,还能在跨物种基因元件注释、比较基因组学的序列保守性分析、多物种调控序列预测等任务中发挥作用,同时为宏基因组学中的微生物物种分类、微生物功能基因注释提供了高效的序列表征工具。

针对传统Transformer架构难以处理基因组长距离依赖的问题,Enformer^[32]、HyenaDNA^[25]等模型通过创新长序列建模机制实现了技术突破。Enformer融合扩张卷积与多头自注意力机制,HyenaDNA则基于隐式长卷积与快速傅里叶变换打造高效长序列建模能力,从而能够处理百kb级甚至Mb级的DNA长序列,直接从DNA序列精准预测染色质可及性、全基因组基因表达水平、增强子-启动子远程调控效应、表达数量性状位点(eQTL)预测等依赖长程序列互作的核心任务,为解析基因组远程调控机制提供了关键工具。

Caduceus^[33]作为新一代性能与效率兼具的长程DNA基础模型,基于Mamba架构进行创新,打造了BiMamba双向建模模块与内嵌DNA反向互补等变性的MambaDNA核心模块,以较低的计算复杂度实现百万bp级长序列的高效建模,同时天然适配DNA双链的生物学特性。该模型经微调后,可完成调控序列活性预测、转录起始信号预测等长程DNA预测任务,在变异效应预测中尤其擅长长距离SNP对基因表达的影响分析。

在此基础上,基因组深度学习模型进一步向生成式与多任务统一建模方向发展。Evo模型^[26]基于原核生物和噬菌体DNA数据训练,成功设计出经实验验证的转座子系统与CRISPR-Cas系统;其升级版Evo2^[27]在原始数据集基础上纳入人类和植物等复杂基因组数据,进一步实现生成完整线粒体基因组、简单细菌基因组以及酵母染色体的能力。这些生成式DNA基础模型突破了传统预测型模型的局限,为合成生物学的基因序列设计提供核心支撑。而AlphaGenome^[30]则尝试打造全功能的基因组学统一模型,在单一模型中实现多任务协同建模。它以1 Mb DNA序列为输入,精准预测基因表达水平(含RNA-seq、CAGE等输出类型)、全基因组染色质状态(涵盖可及性、组蛋白修饰等),以及序列变异的多维度效应(包括剪接调控、增强子-基因互作等11种模态),其性能在26项变异效应预测任务中有25项达到或超越SOTA水平,为解析基因组调控密码提供了强大工具。

除上述各模型自身提供的基准比较外,第三方研究者在不同基因组和遗传任务中对各DNA模型进行了基准测试。有研究^[34]对DNABERT-2、Nucleotide Transformer 2、HyenaDNA、Caduceus-Ph和GROVER进

行了系统性评估,涵盖序列分类、基因表达预测、变异效应量化和拓扑关联域(TAD)识别,结果显示通用DNA基础模型在致病性变异识别上表现出竞争力,但在基因表达预测和因果QTL鉴定方面不如专用模型。DNALONGBENCH^[35]专门针对长程DNA依赖关系(最长可达100万碱基对)建立了基准数据集,包含增强子-靶基因互作、eQTL、3D基因组组织、调控序列活性和转录起始信号五个关键基因组任务,得出当前基础模型(HyenaDNA、Caduceus-Ph和Caduceus-PS)在捕捉长程基因组依赖关系方面仍落后于专用模型。一些分析表明模型总体排名不稳定^[36-37],在不同任务类别间剧烈变化;模型规模带来的提升有限且不一致;架构与预训练数据的对齐往往比参数规模更为关键。不过现有基因组基准评测体系数据、评测方法、可解释链路仍然缺乏统一的标准与自动化流程,且所有评测均为计算机模拟数据,缺少体内实验验证,模型真实生物效果存在偏差。

2.4 深度学习在基因组序列注释中的局限性

现有模型能够有效处理的上下文长度仍然有限,而基因组中存在许多跨越更长距离的调控作用,因此对于这类基因组的计算将受到限制,需要更好的模型架构或编码方式。

其次,基因调控的动态性和特异性对基因组大模型的准确预测构成显著挑战。由于当前模型多基于短读长测序的参考基因组(在端粒、着丝粒等复杂区域存在间隙)和有限细胞类型数据预训练,导致模型难以捕捉细胞类型特异的调控全景。为突破此瓶颈,一方面可通过采用长读长技术构建无间隙T2T-CHM13参考基因组^[38]提升参考序列质量;另一方面,为克服生物样本的复杂异质性,研究者们开发了基于单细胞多组学数据的新模型,直接在单细胞分辨率上解析动态调控程序,从而系统性应对上述挑战^[15]。

3 基于深度学习策略的基因组结合其他组学技术的研究方法

3.1 深度学习在基因组中结合其他组学技术的研究进展

深度学习在基因组研究中,正通过融合多组学技术突破单一数据建模的局限。长程互作的建模难题尚未完全破解,例如增强子与启动子的远距离互作难以精准捕捉。由于模型的知识学习深度和广度依赖输入数据类型,而这类相互作用受三维基因组结构动态影响,当前单一的数据输入模式导致模型习得的生物学知识片面、不完整。基于此,一些模型尝试增加其他组学数据的输入。EpiGePT^[39]通过引入转录因子表达谱增强模型对细

胞类型的感知能力,并融入HiChIP等染色质三维互作数据引导训练,实现了调控元件-启动子互作等三维基因组特征的更高效预测。

而在表观遗传修饰研究领域,作为连接基因型与表型的关键桥梁,表观遗传的相关分析也因多组学融合的深度学习方法迎来新突破。深度学习通过学习序列特征与表观组状态的关联,已能建立以DNA序列为主要输入的甲基化及染色质可及性预测模型,为辅助和补充BS-seq、ATAC-seq等实验提供了新途径。DeepCpG模型^[40]耦合CNN与RNN,分别用于提取CpG岛局部特征及相邻位点甲基化依赖关系,在单细胞分辨率下实现了对DNA甲基化状态的高精度预测和有效插补;DeepAccess^[41]基于CNN解析转录因子基序,在多个细胞系中预测ATAC-seq信号,可辅助识别细胞类型特异性开放区域。此外,表观遗传预测类深度学习模型(如DeepSEA^[42])在变异效应预测(variant effect prediction, VEP)中亦显示出巨大潜力,能够推断非编码变异如何通过改变序列基序影响染色质状态,从而助力解析GWAS信号的生物学意义。CpGPT^[43]作为基于Transformer的基础模型,进一步推动了甲基化分析的边界。该模型在超过10万个样本上预训练,能融合DNA序列、位置和表观遗传信息,实现甲基化状态的零样本插补、跨平台转换及样本特异性解读。其在死亡率预测和生物学年龄估计任务中表现卓越,凸显了基础模型在跨任务迁移中的潜力。

3.2 深度学习在基因组学中结合其他组学技术的局限性

在多组学数据本身层面,异质性与质量缺陷造成了融合壁垒。基因组(DNA序列)、表观组(甲基化、染色质可及性)、转录组(基因表达)、三维基因组(HiChIP/Hi-C)等组学数据,在检测平台、量化尺度、样本批次、时空分辨率上差异极大,缺乏通用的标准化方法,直接融合易引入数据偏倚。并且部分组学数据获取成本高、覆盖度不足,可用的高质量训练数据量少。表观组、转录组数据易受实验处理、样本状态影响,噪声干扰也会降低模型学习的有效性。

其次,多组学数据的时空匹配性较差。现有多组学数据多为静态快照,且不同组学数据常来自不同实验条件、时间节点的样本,难以实现真正的时空耦合融合,模型学到的关系易与真实生物学场景脱节。不过,近年来空间转录组学(spatial transcriptomics)技术的快速发展为上述“时空脱节”问题提供了直接的解决方案。Nicheformer^[44]、Novae^[45]等一系列深度学习模型,分别在Transformer架构和图注意力网络(GAT)的基础上,被开发用于空间转录组数据的分析与整合,从而学习到细胞加周边微环境的空间表征。

最后,多组学特征的融合策略缺乏针对性。多数模型只是采用简单拼接、浅层融合的方式整合多组学特征(如将DNA序列嵌入与甲基化数据直接拼接输入模型),未深入挖掘组学间层级化、因果性的生物学关联。部分模型的融合模块为通用架构,未针对基因组学的特有规律(如DNA反向互补、密码子阅读框等)做定制化设计,导致融合效率低、特征利用率差。

4 小结与展望

深度学习框架已在人类基因组研究中表现出卓越的性能与成果。这些进展不仅节省了实验成本,更加速了人类基因组关键信息的发现。深度学习模型通过整合大量数据,正推动研究范式从单一的基因组分析向系统生物学层面跃迁,旨在构建更为完整的生物系统模型。在这一趋势下,“虚拟细胞”概念的提出与发展尤为引人注目^[46]。它基于深度学习整合了多模态数据,能模拟真实细胞的结构、功能及动态行为,在疾病机制解析、药物研发和个性化医疗等领域有广阔前景。目前研究者已首次在计算机中成功模拟了细菌JCVI-syn3A复制DNA并分裂为两个子细胞的完整过程^[47]。

然而,要实现这一宏大目标,当前深度学习方法的性能提升仍面临诸多挑战。其中数据维度单一性是主要的限制因素。从数据集层面来看,现有的深度学习方法大多依赖于信息量有限的数据集。例如:(1)DNA大模型主要基于单一参考基因组数据进行训练,忽视了群体的遗传多样性(如泛基因组^[48-50]),且多基于人类、小鼠等模式生物的数据集训练,存在物种或组织偏倚,模型难以跨物种迁移,对非模式生物的分析效果大幅下降,也无法有效捕捉物种间的进化保守性调控规律;(2)DNA变异检测依赖于短读长测序所产生的数据集,由于短读长测序技术读长较短,难以准确跨越重复区域和大片段结构变异区域,致使基因组组装不完整、不准确,进而导致基于这些不完整、不准确组装的参考基因组进行变异检测、功能分析时,准确性受限的连锁问题。其次,模型的不可解释性以及高昂的计算资源消耗也为深度学习的进一步广泛应用设置了障碍。

为突破上述限制,未来需聚焦三方面:其一,攻克“黑箱”困境,提升模型可解释性。例如,融合基因组大模型与大语言模型的推理能力,生成可被生物学家理解的可解释因果链(如BioReason^[51]);其二,纳入新一代技术,突破数据维度与偏倚的限制。这包括从单参考基因组转向泛基因组,例如DeepGene^[52]将泛基因组图去环序列化后分词输入基因组LLM完成预训练,充分捕获人群遗传变

异以缓解单线性参考基因组带来的表征偏倚。在此基础上进一步融合GNN提取图拓扑特征并嵌入Transformer注意力层,实现泛基因组图拓扑与DNA序列上下文的联合建模。深化长读长测序在大片段结构变异等复杂变异解析中的应用,并充分利用纳米孔单分子“序列+表观”同步捕获的技术优势^[53];同时,转向基于真实体内样本的时空多组学数据,并继续探索最优的物种采样策略和分类学多样性,以平衡泛化能力与特定物种任务的性能。其三,开发高效、可扩展的数据整合与模型训练算法,以应对日益增长的数据规模和计算成本。这些努力将共同为解码基因组复杂功能、并最终实现虚拟细胞的宏大愿景奠定基石,使深度学习在人类基因组研究和更广泛的生物医学领域发挥出真正的变革性作用。

本文中提及的代表性工具见附表1(详见网络资源附件)。

* * *

作者贡献声明 李郁葱负责论文构思和初稿写作,屠凯岭负责论文构思、经费获取和审读与编辑写作,谢丹负责经费获取、提供资源和监督指导。所有作者已经同意将文章提交给本刊,且对将要发表版本进行最终定稿,并同意对工作的所有方面负责。

Author Contribution Li Yucong is responsible for conceptualization and writing--original draft. Tu Kailing is responsible for conceptualization, funding acquisition, and writing--review and editing. Xie Dan is responsible for funding acquisition, resources, and supervision. All authors consented to the submission of the article to the Journal. All authors approved the final version to be published and agreed to take responsibility for all aspects of the work.

利益冲突 本文作者谢丹是本刊编委会青年编委。该文在编辑评审过程中所有流程严格按照期刊政策进行,且未经其本人经手处理。除此之外,所有作者均声明不存在利益冲突。

Declaration of Conflicting Interests Xie Dan is a member of the Junior Editorial Board of the journal. All processes involved in the editing and reviewing of this article were carried out in strict compliance with the journal's policies and there was no inappropriate personal involvement by the authors. Other than this, all authors declare no competing interests.

参 考 文 献

- [1] McKenna A, Hanna M, Banks E, *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data[J]. *Genome Res*, 2010, 20(9): 1297. DOI: 10.1101/gr.107524.110.
- [2] Boeva V, Popova T, Bleakley K, *et al.* Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data[J]. *Bioinformatics*, 2012, 28(3): 423. DOI: 10.1093/bioinformatics/btr670.
- [3] Ye K, Schulz M H, Long Q, *et al.* Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads[J]. *Bioinformatics*, 2009, 25(21): 2865. DOI: 10.1093/bioinformatics/btp394.
- [4] Uffelmann E, Huang Q Q, Munung N S, *et al.* Genome-wide association studies[J/OL]. *Nat Rev Methods Primers*, 2021, 1: 59. <https://www.nature.com/articles/s43586-021-00056-9>. DOI: 10.1038/s43586-021-00056-9.
- [5] Campagna M P, Xavier A, Lechner-Scott J, *et al.* Epigenome-wide association studies: current knowledge, strategies and recommendations [J/OL]. *Clin Epigenet*, 2021, 13: 214. <https://link.springer.com/article/10.1186/s13148-021-01200-8>.
- [6] Belton J M, McCord R P, Gibcus J H, *et al.* Hi-C: a comprehensive technique to capture the conformation of genomes[J]. *Methods*, 2012, 58(3): 268. DOI: 10.1016/j.ymeth.2012.05.001.
- [7] Zhang Y, Liu T, Meyer C A, *et al.* Model-based analysis of ChIP-Seq (MACS)[J/OL]. *Genome Biol*, 2008, 9(9): R137. <https://link.springer.com/article/10.1186/gb-2008-9-9-r137>.
- [8] LeCun Y, Bengio Y, Hinton G. Deep learning[J]. *Nature*, 2015, 521(7553): 436. DOI: 10.1038/nature14539.
- [9] Poplin R, Chang P C, Alexander D, *et al.* A universal SNP and small-indel variant caller using deep neural networks[J]. *Nat Biotechnol*, 2018, 36(10): 983. DOI: 10.1038/nbt.4235.
- [10] Luo R, Sedlazeck F J, Lam T W, *et al.* A multi-task convolutional deep neural network for variant calling in single molecule sequencing[J/OL]. *Nat Commun*, 2019, 10(1): 998. <https://www.nature.com/articles/s41467-019-09025-z>. DOI: 10.1038/s41467-019-09025-z.
- [11] Luo R, Wong C L, Wong Y S, *et al.* Exploring the limit of using a deep neural network on pileup data for germline variant calling[J]. *Nat Mach Intell*, 2020, 2(4): 220. DOI: 10.1038/s42256-020-0167-4.
- [12] Zheng Z, Li S, Su J, *et al.* Symphonizing pileup and full-alignment for deep learning-based long-read variant calling[J]. *Nat Comput Sci*, 2022, 2(12): 797. DOI: 10.1038/s43588-022-00387-x.
- [13] Ji Y R, Zhou Z H, Liu H, *et al.* DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome[J]. *Bioinformatics*, 2021, 37(15): 2112. DOI: 10.1093/bioinformatics/btab083.
- [14] Dalla-Torre H, Gonzalez L, Mendoza-Revilla J, *et al.* Nucleotide Transformer: building and evaluating robust foundation models for human genomics[J]. *Nat Methods*, 2025, 22(2): 287. DOI: 10.1038/s41592-024-02523-z.
- [15] Hao M, Gong J, Zeng X, *et al.* Large-scale foundation model on single-cell transcriptomics[J]. *Nat Methods*, 2024, 21(8): 1481. DOI: 10.1038/s41592-024-02305-7.
- [16] Sedlazeck F J, Rescheneder P, Smolka M, *et al.* Accurate detection of complex structural variations using single-molecule sequencing[J]. *Nat Methods*, 2018, 15(6): 461. DOI: 10.1038/s41592-018-0001-7.
- [17] Jiang T, Liu S, Cao S, *et al.* Structural variant detection from long-read sequencing data with cuteSV[J]. *Methods Mol Biol*, 2022, 2493: 137. DOI: 10.1007/978-1-0716-2293-3_9.
- [18] Lin J, Wang S, Audano P A, *et al.* SVision: a deep learning approach to resolve complex structural variants[J]. *Nat Methods*, 2022, 19(10): 1230. DOI: 10.1038/s41592-022-01609-w.
- [19] Wang S, Xu T, Zhang P, *et al.* Population-level structural variant characterization using pangenome graphs[J]. *Nat Genet*, 2026, 58(3): 664. DOI: 10.1038/s41588-026-02538-6.
- [20] Elrick H, Sauer C M, Espejo V J, *et al.* SAVANA: reliable analysis of somatic structural variants and copy number aberrations using long-read sequencing[J]. *Nat Methods*, 2025, 22(7): 1436. DOI: 10.1038/s41592-025-02708-0.
- [21] Keskus A G, Bryant A, Ahmad T, *et al.* Severus detects somatic structural variation and complex rearrangements in cancer genomes using long-read sequencing[J]. *Nat Biotechnol*, 2026, 44(2): 247. DOI: 10.1038/s41587-025-02618-8.
- [22] Vaswani A, Shazeer N, Parmar N, *et al.* Attention is all you need[PP/OL]. V7. arXiv(2023-08-02)[2026-07-09]. <https://arxiv.org/abs/1706.03762>. DOI: 10.48550/arXiv.1706.03762.
- [23] Devlin J, Chang M W, Lee K, *et al.* BERT: pre-training of deep bidirectional Transformers for language understanding[PP/OL]. V2. arXiv(2019-05-24)[2026-07-09]. <https://arxiv.org/abs/1810.04805>. DOI: 10.48550/arXiv.1810.04805.
- [24] Zhou Z H, Ji Y R, Liu W J, *et al.* DNABERT-2: efficient foundation model and benchmark for multi-species genome[PP/OL]. V2. arXiv(2024-03-

- 18][2026-07-09]. <https://arxiv.org/abs/2306.15006>. DOI: 10.48550/arXiv.2306.15006.
- [25] Nguyen E, Poli M, Faizi M, *et al.* HyenaDNA: long-range genomic sequence modeling at single nucleotide resolution[PP/OL]. V2. arXiv(2023-11-14)[2026-07-09]. <https://arxiv.org/abs/2306.15794>. DOI: 10.48550/arXiv.2306.15794.
- [26] Nguyen E, Poli M, Durrant M G, *et al.* Sequence modeling and design from molecular to genome scale with Evo[J/OL]. Science, 2024, 386(6723): eado9336. <https://www.science.org/doi/10.1126/science.ad9336>.
- [27] Brixi G, Durrant M G, Ku J, *et al.* Genome modelling and design across all domains of life with Evo 2[J]. Nature, 2026, 652(8112): 1349. DOI: 10.1038/s41586-026-10176-5.
- [28] Radford A, Narasimhan K, Salimans T, *et al.* Improving language understanding by generative pre-training[R]. San Francisco: OpenAI, 2018.
- [29] Gu A, Dao T. Mamba: linear-time sequence modeling with selective state spaces[PP/OL]. V2. arXiv(2024-05-31)[2026-07-09]. <https://arxiv.org/abs/2312.00752>. DOI: 10.48550/arXiv.2312.00752.
- [30] Avsec Ž, Latysheva N, Cheng J, *et al.* Advancing regulatory variant effect prediction with AlphaGenome[J]. Nature, 2026, 649(8099): 1206. DOI: 10.1038/s41586-025-10014-0.
- [31] Fudenberg G, Kelley D R, Pollard K S. Predicting 3D genome folding from DNA sequence with Akita[J]. Nat Methods, 2020, 17(11): 1111. DOI: 10.1038/s41592-020-0958-x.
- [32] Avsec Ž, Agarwal V, Visentin D, *et al.* Effective gene expression prediction from sequence by integrating long-range interactions[J]. Nat Methods, 2021, 18(10): 1196. DOI: 10.1038/s41592-021-01252-x.
- [33] Schiff Y, Kao C H, Gokaslan A, *et al.* Caduceus: bi-directional equivariant long-range DNA sequence modeling[PP/OL]. V2. arXiv(2024-06-05)[2026-07-09]. <https://arxiv.org/abs/2403.03234>. DOI: 10.48550/arXiv.2403.03234.
- [34] Feng H, Wu L, Zhao B, *et al.* Benchmarking DNA foundation models for genomic and genetic tasks[J/OL]. Nat Commun, 2025, 16(1): 10780. <https://pmc.ncbi.nlm.nih.gov/articles/PMC12663285/>. DOI: 10.1038/s41467-025-65823-8.
- [35] Cheng W, Song Z, Zhang Y, *et al.* DNALONGBENCH: a benchmark suite for long-range DNA prediction tasks[J/OL]. Nat Commun, 2025, 16(1): 10108. <https://pmc.ncbi.nlm.nih.gov/articles/PMC12627797/>. DOI: 10.1038/s41467-025-65077-4.
- [36] Ledneva D, Nuridinov M, Kuznetsov D. GENE: Why genomic models are hard to compare[PP/OL]. V3. arXiv(2026-06-11)[2026-07-09]. <https://arxiv.org/abs/2606.04525>. DOI: 10.48550/arXiv.2606.04525.
- [37] Yang H, Cole J, Li Y, *et al.* Omnigenbench: a modular platform for reproducible genomic foundation models benchmarking[PP/OL]. V1. arXiv(2025-05-20)[2026-07-09]. <https://arxiv.org/abs/2505.14402>. DOI: 10.48550/arXiv.2505.14402.
- [38] Nurk S, Koren S, Rhie A, *et al.* The complete sequence of a human genome[J]. Science, 2022, 376(6588): 44. DOI: 10.1126/science.abj6987.
- [39] Gao Z, Liu Q, Zeng W, *et al.* EpiGePT: a pretrained transformer-based language model for context-specific human epigenomics[J/OL]. Genome Biol, 2024, 25(1): 310. <https://link.springer.com/article/10.1186/s13059-024-03449-7>.
- [40] Angermueller C, Lee H J, Reik W, *et al.* DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning[J/OL]. Genome Biol, 2017, 18(1): 67. <https://link.springer.com/article/10.1186/s13059-017-1189-z>.
- [41] Hammelman J, Gifford D K. Discovering differential genome sequence activity with interpretable and efficient deep learning[J/OL]. PLoS Comput Biol, 2021, 17(8): e1009282. <https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1009282>.
- [42] Zhou J, Troyanskaya O. Predicting effects of noncoding variants with deep learning-based sequence model[J]. Nat Methods, 2015, 12(10): 931. DOI: 10.1038/nmeth.3547.
- [43] Page L, Singh R, Jenei A, *et al.* CpGPT: a foundation model for DNA methylation[PP/OL]. V3. BioRxiv(2025-08-28)[2026-07-09]. <https://www.biorxiv.org/content/10.1101/2024.10.24.619766v4>. DOI: 10.1101/2024.10.24.619766.
- [44] Tejada-Lapueta A, Schaar AC, Gutgesell R, *et al.* Nicheformer: a foundation model for single-cell and spatial omics[J]. Nat Methods, 2025, 22(12): 2525. DOI: 10.1038/s41592-025-02814-z.
- [45] Blampey Q, Benkirane H, Bercovici N, *et al.* Novae: a graph-based foundation model for spatial transcriptomics data[J]. Nat Methods, 2025, 22(12): 2539. DOI: 10.1038/s41592-025-02899-6.
- [46] Bunne C, Roohani Y, Rosen Y, *et al.* How to build the virtual cell with artificial intelligence: priorities and opportunities[J]. Cell, 2024, 187(25): 7045. DOI: 10.1016/j.cell.2024.11.015.
- [47] Thornburg Z R, Maytin A, Kwon J, *et al.* Bringing the genetically minimal cell to life on a computer in 4D[J]. Cell, 2026, 189(9): 2582. DOI: 10.1016/j.cell.2026.02.009.
- [48] Liao WW, Asri M, Ebler J, *et al.* A draft human pangenome reference[J]. Nature, 2023, 617(7960): 312. DOI: 10.1038/s41586-023-05896-x.
- [49] Hickey G, Monlong J, Ebler J, *et al.* Pangenome graph construction from genome alignments with Minigraph-Cactus[J]. Nat Biotechnol, 2024, 42(4): 663. DOI: 10.1038/s41587-023-01793-w.
- [50] Schloissnig S, Pani S, Ebler J, *et al.* Structural variation in 1, 019 diverse humans based on long-read sequencing[J]. Nature, 2025, 644(8076): 442. DOI: 10.1038/s41586-025-09290-7.
- [51] Fallahpour A, Magnuson A, Gupta P, *et al.* BioReason: incentivizing multimodal biological reasoning within a DNA-LLM model[PP/OL]. V2. arXiv(2025-10-24)[2026-07-09]. <https://arxiv.org/abs/2505.23579>. DOI: 10.48550/arXiv.2505.23579.
- [52] Zhang X, Yang M, Yin X, *et al.* DeepGene: an efficient foundation model for genomics based on pan-genome graph transformer[J]. IEEE Trans Comput Biol Bioinform, 2025, 22(6): 3143. DOI: 10.1109/TCBBIO.2025.3614354.
- [53] Shipony Z, Marinov G K, Swaffer M P, *et al.* Long-range single-molecule mapping of chromatin accessibility in eukaryotes[J]. Nat Methods, 2020, 17(3): 319. DOI: 10.1038/s41592-019-0730-2.

(2025-11-03收稿, 2026-07-09修回)

编辑 姜 恬



开放获取 本文使用遵循知识共享署名—非商业性使用4.0国际许可协议(CC BY-NC 4.0), 详细信息请访问

<https://creativecommons.org/licenses/by-nc/4.0/>。

OPEN ACCESS This article is licensed for use under Creative Commons Attribution-NonCommercial 4.0 International license (CC BY-NC 4.0). For more information, visit <https://creativecommons.org/licenses/by-nc/4.0/>.

© 2026《四川大学学报(医学版)》编辑部

Editorial Office of Journal of Sichuan University (Medical Sciences)