Awesome-LLMs-meet-genomes

Awesome-LLMs-meet-genomes is a collection of state-of-the-art, novel, exciting LLMs methods on genomes. It contains papers, codes, datasets, evaluations, and analyses. Any additional information about LLMs for bioinformatics is welcome, and we are glad to add you to the contributor list here. Any problems, please contact yangchengyjs@163.com. If you find this repository useful to your research or work, it is really appreciated to star this repository. ✨

---

language none Stars 40 Fork 2 visitors 4202

Visitor counts

0 0 0 4 4 7 8

Table of Content

🔔 News

- 🧬 ✔ [2024/09] **Benchmarks for classification of genomic sequences** link.
- 💥 [2024/08] Some real-world experience in training LLMs link.
- 💥 [2024/08] Three ways of Fine-tuning link.
- 💥 [2024/08] Visualisation of the Transformer Principle link.
- ✹ [2024/08] The Cultivation Method of Large Language Models: A Path to Success link.
- 📖 [2024/08] Large Language Models: From Theory to Practice link.

Important Survey Papers

Year	Title	Venue	Paper	Code
2025.03	**Language modelling techniques for analysing the impact of human genetic variation**	arXiv	Link	-
2025.03	**Biological Sequence with Language Model Prompting: A Survey**	arXiv	Link	-
2025.01	**Large Language Models for Bioinformatics**	arXiv	Link	-
2024.09	**Genomic Language Models: Opportunities and Challenges**	arXiv	Link	-
2024.07	**Scientific Large Language Models: A Survey on Biological & Chemical Domains**	arXiv	Link	link
2024.01	**Large language models in bioinformatics: applications and perspectives**	arXiv	Link	-
2023.11	**To Transformers and Beyond: Large Language Models for the Genome**	arXiv	Link	-
2023.01	**Applications of transformer-based language models in bioinformatics: a survey**	Bioinformatics Advances	Link	-

Genomic Large Language Models (Gene-LLMs)

Generic Base Models

Year	Title	Venue	Paper	Code
2025.03	**HydraRNA: a hybrid architecture based full-length RNA language model**	bioXiv	link	link
2025.03	**Pre-training Genomic Language Model with Variants for Better Modeling Functional Genomics**	bioXiv	link	link
2025.03	**Enhancing DNA Foundation Models to Address Masking Inefficiencies**	arXiv	link	-

Year	Title	Venue	Paper	Code
2025.02	**HybriDNA: A Hybrid Transformer-Mamba2 Long-Range DNA Language Model**	arXiv	link	-
2025.02	**GENERator: A Long-Context Generative Genomic Foundation Model**	arXiv	link	link
2025.02	**Omni-DNA: A Unified Genomic Foundation Model for Cross-Modal and Multi-Task Learning**	arXiv	link	link
2025.01	**MutBERT: Probabilistic Genome Representation Improves Genomics Foundation Models**	BioXiv	link	link
2025.01	**GENA-LM: a family of open-source foundational DNA language models for long sequences**	Nucleic Acids Research	link	link
2024.12	**EpiGePT: a pretrained transformer-based language model for context-specific human epigenomics**	Genome Biology	link	link
2024.11	**VIRALpre: Genomic Foundation Model Embedding Fused with K-mer Feature for Virus Identification**	bioRxiv	link	-
2024.11	**BEACON: Benchmark for Comprehensive RNA Tasks and Language Models**	NeurIPS'24	link	link
2024.11	**DNA Language Models for RNA Analyses**	ICLR'25 Conference Submission	link	-
2024.10	**Character-level Tokenizations as Powerful Inductive Biases for RNA Foundational Models**	NeurIPS'24	link	link
2024.10	**Revisiting K-mer Profile for Effective and Scalable Genome Representation Learning**	NeurIPS'24	link	link
2024.10	**A long-context language model for deciphering and generating bacteriophage genomes**	Nature Communications	link	link
2024.10	**Revisiting Convolution Architecture in the Realm of DNA Foundation Models**	ICLR'25 Conference Submission	link	-
2024.10	**Hyperbolic Genome Embeddings**	ICLR'25 Conference Submission	link	-
2024.10	**dnaGrinder: a lightweight and high-capacity genomic foundation model**	ICLR'25 Conference Submission	link	-

Year	Title	Venue	Paper	Code
2024.10	**DNABERT-S: Pioneering Species Differentiation with Species-Aware DNA Embeddings**	ICLR'25 Conference Submission	link	-
2024.10	**Long-range gene expression prediction with token alignment of large language model**	arXiv	link	-
2024.09	**A Comparison of Tokenization Impact in Attention Based and State Space Genomic Language Models**	bioRxiv	link	-
2024.09	**Designing realistic regulatory DNA with autoregressive language models**	Genome Research	link	-
2024.08	**Understanding the Natural Language of DNA using Encoder-Decoder Foundation Models with Byte-level Precision**	Bioinformatics Advances	link	link
2024.08	**Unlocking Efficiency: Adaptive Masking for Gene Transformer Models**	ECAI'24	link	link
2024.07	**Genomics-FM: Universal Foundation Model for Versatile and Data-Efficient Functional Genomic Analysis**	bioRxiv	link	link
2024.07 ✨✨✨	**VQDNA: Unleashing the Power of Vector Quantization for Multi-Species Genomic Sequence Modeling**	ICML'24	link	link
2024.07	**OmniGenome: Aligning RNA Sequences with Secondary Structures in Genomic Foundation Models**	arXiv	link	link
2024.07	**Scorpio : Enhancing Embeddings to Improve Downstream Analysis of DNA sequences**	bioRxiv	link	link
2024.07	**DNA language model GROVER learns sequence context in the human genome (可用于蛋白质-DNA结合预测任务)**	Nature Machine Intelligence	link	link tutorials
2024.05	**Are Genomic Language Models All You Need? Exploring Genomic Language Models on Protein Downstream Tasks**	bioRxiv	link	link
2024.05	**GeneAgent: Self-verification Language Agent for Gene Set Knowledge Discovery using Domain Databases**	arXiv	link	-
2024.05	**DeepGene: An Efficient Foundation Model for Genomics based on Pan-genome Graph Transformer**	bioRxiv	link	link

Year	Title	Venue	Paper	Code
2024.05	**Self-Distillation Improves DNA Sequence Inference Databases**	arXiv	link	link
2024.04	**Effect of tokenization on transformers for biological sequences**	Bioinformatics	link	link
2024.04	**DNABERT-2: Efficient Foundation Model and Benchmark For Multi-Species Genome**	ICLR'24	link	link
2024.02	**Exploring Genomic Large Language Models: Bridging the Gap between Natural Language and Gene Sequences**	bioRxiv	link	link data
2024.02	**Sequence modeling and design from molecular to genome scale with Evo**	bioRxiv	link	link
2024.01	**ProkBERT family: genomic language models for microbiome applications**	Frontiers in Microbiology	Link	link
2023.09	**The Nucleotide Transformer: Building and Evaluating Robust Foundation Models for Human Genomics**	bioRxiv	link	link
2023.08	**DNAGPT: A Generalized Pre-trained Tool for Versatile DNA Sequence Analysis Tasks**	bioRxiv	link	link
2023.07	**EpiGePT: a Pretrained Transformer model for epigenomics**	bioRxiv	link	link
2023.07	**GeneMask: Fast Pretraining of Gene Sequences to Enable Few-Shot Learning**	ECAI'23	link	link
2023.06	**Transfer learning enables predictions in network biology**	nature	link	link
2023.06	**GENA-LM: A Family of Open-Source Foundational DNA Language Models for Long Sequences**	bioRxiv	link	link
2023.06	**HyenaDNA: long-range genomic sequence modeling at single nucleotide resolution**	NIPS'23	link	link
2023.01	**The Nucleotide Transformer: Building and Evaluating Robust Foundation Models for Human Genomics**	bioRxiv	link	link
2023.01	**Species-aware DNA language modeling**	bioRxiv	link	link
2022.08	**MoDNA: motif-oriented pre-training for DNA language model**	BCB'22	link	link

Year	Title	Venue	Paper	Code
2021.02	**DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome**	Bioinformatics	link	link

Downstream Tasks

Gene Pathogenicity Prediction

Time	Title	Venue	Paper	Code
2024.06	**PathoLM: Identifying pathogenicity from the DNA sequence through the Genome Foundation Model**	arXiv	link	-
2024.06	**Gene Pathogenicity Prediction using Genomic Foundation Models**	AAAI'24 Spring Symposium on Clinical Foundation Models	link	-

Retrieval-Augmented Generation

Time	Title	Venue	Paper	Code
2024.06	**GeneRAG: Enhancing Large Language Models with Gene-Related Task by Retrieval-Augmented Generation**	bioRxiv	link	link

Function Prediction

Time	Title	Venue	Paper	Code
2024.07	**FGBERT: Function-Driven Pre-trained Gene Language Model for Metagenomics**	arXiv	link	-
2023.07	**PLPMpro: Enhancing promoter sequence prediction with prompt-learning based pre-trained language model**	CIBM	link	-
2021.10	**Effective gene expression prediction from sequence by integrating long-range interactions**	Nature Methods	link	link

Perturbation

Time	Title	Venue	Paper	Code
2024.08	**Scouter: a transcriptional response predictor for unseen genetic perturbtions with LLM embeddings**	pypi	link	link
2024.07	**Enhancing generative perturbation models with LLM-informed gene embeddings**	ICLR'24 Workshop	link	-

Time	Title	Venue	Paper	Code
2024.03	**A genome-scale deep learning model to predict gene expression changes of genetic perturbations from multiplex biological networks**	arXiv	link	link

Variants and Evolution Prediction

Time	Title	Venue	Paper	Code
2025.02	**A SNP Foundation Model: Application in Whole-Genome Haplotype Phasing and Genotype Imputation**	BioRxiv	link	-
2025.01	**A DNA language model based on multispecies alignment predicts the effects of genome-wide variants**	Nature Biotechnology	link	link
2024.11	**Leveraging genomic deep learning models for non-coding variant effect prediction**	ArXiv	link	-
2024.04	**Species-aware DNA language models capture regulatory elements and their evolution**	Genome Biology	link	link
2023.10	**GenSLMs: Genome-scale language models reveal SARS-CoV-2 evolutionary dynamics**	The International Journal of High Performance Computing Applications	link	link
2023.08	**DNA language models are powerful predictors of genome-wide variant effects**	PNAS	link	link

Fine-tuning for Genomes and proteins

Time	Title	Venue	Paper	Code
2024.09	**Fine-tuning sequence-to-expression models on personal genome and transcriptome data**	bioRxiv	link	link
2024.08	**Enhancing recognition and interpretation of functional phenotypic sequences through fine-tuning pre-trained genomic models**	Journal of Translational Medicine	link	link
2024.08	**Fine-tuning protein language models boosts predictions across diverse tasks**	Nature Communications	link	link
2024.02	**Efficient and Scalable Fine-Tune of Language Models for Genome Understanding**	arXiv	link	link

Time	Title	Venue	Paper	Code
2023.11	**Parameter-Efficient Fine-Tune on Open Pre-trained Transformers for Genomic Sequence**	NeurIPS'23 Workshop GenBio	link	-
2024.01	**ViraLM: Empowering Virus Discovery through the Genome Foundation Model**	bioRxiv	link	link

Interaction Prediction

Time	Title	Venue	Paper	Code
2024.08	**Large-Scale Multi-omic Biosequence Transformers for Modeling Peptide-Nucleotide Interactions**	arXiv	link	link
2024.04	**Genomic language model predicts protein co-regulation and function**	nature communications	link	link
2024.01	**Gene-associated Disease Discovery Powered by Large Language Models**	arXiv	link	-

Identification of Transcription Factor Binding Sites

Time	Title	Venue	Paper	Code
2025.03	**deepTFBS: Improving within- and cross-species prediction of transcription factor binding using deep multi-task and transfer learning**	BioRxiv	link	link
2024.10	**DNA breathing integration with deep learning foundational model advances genome-wide binding prediction of human transcription factors**	Nucleic Acids Research	link	link
2024.08	**BertSNR: an interpretable deep learning framework for single-nucleotide resolution identification of transcription factor binding sites based on DNA language model**	Bioinformatics	link	link
2024.05	**BERT-TFBS: a novel BERT-based model for predicting transcription factor binding sites by transfer learning**	Briefings in Bioinformatics	link	link
2024.01	**Multiomics-integrated deep language model enables in silico genome-wide detection of transcription factor binding site in unexplored biosamples**	Bioinformatics	link	-

Origins of Replication Rite Prediction

Time	Title	Venue	Paper	Code

Time	Title	Venue	Paper	Code
2024.01	**PLANNER: a multi-scale deep language model for the origins of replication site prediction**	IEEE Journal of Biomedical and Health Informatics	link	-

DNA-binding Protein Prediction

Time	Title	Venue	Paper	Code
2024.09	**Improving prediction performance of general protein language model by domain-adaptive pretraining on DNA-binding protein**	Nature Communications	link	link
2024.07	**Prediction of Protein-DNA Binding Sites Based on Protein Language Model and Deep Learning**	International Conference on Intelligent Computing	link	-
2024.03 ✨✨✨	**EquiPNAS: improved protein–nucleic acid binding site prediction using protein-language-model-informed equivariant deep graph neural networks**	Nucleic Acids Research	link	link
2024.01	**Predictive Recognition of DNA-binding Proteins Based on Pre-trained Language Model BERT**	Journal of Bioinformatics and Computational Biology	link	-
2024.01	**Protein–DNA binding sites prediction based on pre-trained protein language model and contrastive learning**	Briefings in Bioinformatics	link	link
2022.09	**Improving language model of human genome for DNA–protein binding prediction based on task-specific pre-training**	Interdisciplinary Sciences: Computational Life Sciences	link	link

RNA Prediction

Time	Title	Venue	Paper	Code
2024.07 ✨✨✨	**Single-sequence protein-RNA complex structure prediction by geometric attention-enabled pairing of biological language models**	bioRxiv	link	link
2024.05	**RNAErnie: Multi-purpose RNA language modelling with motif-aware pretraining and type-guided fine-tuning**	Nature Machine Intelligence	link	link

Time	Title	Venue	Paper	Code
2024.02	**RiNALMo: General-Purpose RNA Language Models Can Generalize Well on Structure Prediction Tasks**	arXiv	link	link
2023.10	**Multiple sequence alignment-based RNA language model and its application to structural inference**	Nucleic Acids Research	link	link
2023.07	**Uni-RNA: Universal Pre-trained Models Revolutionize RNA Research**	bioRxiv	link	-
2023.06	**Prediction of Multiple Types of RNA Modifications via Biological Language Model**	TCBB	link	link
2023.02	**Self-supervised learning on millions of pre-mRNA sequences improves sequence-based RNA splicing prediction**	bioRxiv	link	link

Sequence Modeling

Time	Title	Venue	Paper	Code
2024.11	**Unveiling Protein-DNA Interdependency: Harnessing Unified Multimodal Sequence Modeling, Understanding and Generation**	-	-	link
2024.09	**Toward Understanding BERT-Like Pre-Training for DNA Foundation Models**	arXiv	link	-
2024.08	**LitGene: a transformer-based model that uses contrastive learning to integrate textual information into gene representations**	bioRxiv	link	link
2024.08	**BiRNA-BERT allows efficient RNA language modeling with adaptive tokenization**	bioRxiv	link	link
2024.07 ✨✨✨	**VQDNA: Unleashing the Power of Vector Quantization for Multi-Species Genomic Sequence Modeling**	ICML'24	link	link
2024.06 💥💥💥	**Caduceus: Bi-Directional Equivariant Long-Range DNA Sequence Modeling**	ICML’24	link	link
2024.06	**Contrastive pre-training for sequence based genomics models**	bioRxiv	link	link
2024.05	**Dirichlet Flow Matching with Applications to DNA Sequence Design**	ICML’24	link	link
2024.05 🏋️🏋️	**Self-Distillation Improves DNA Sequence Inference**	arXiv	link	link

Time	Title	Venue	Paper	Code
2024.05 🏋 🏋	**Accurate and efficient protein embedding using multi-teacher distillation learning**	arXiv	link	link
2024.04	**Effect of tokenization on transformers for biological sequences**	Bioinformatics	link	link
2024.04	**A Sparse and Wide Neural Network Model for DNA Sequences**	SRNN	link	link
2024.03	**Self-supervised learning for DNA sequences with circular dilated convolutional networks**	Neural Networks	link	link
2024.01	**ProtHyena: A fast and efficient foundation protein language model at single amino acid Resolution**	bioRxiv	link	link
2023.06	**HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution**	NeurIPS'23	link	link

Basics of Sequence Modeling

Time	Title	Venue	Paper	Code
2025.02	**Linear Attention for Efficient Bidirectional Sequence Modeling**	arXiv	link	link
2024.10	**LongMamba: Enhancing Mamba's Long-Context Capabilities via Training-Free Receptive Field Enlargement**	ICLR'25 Conference Submission	link	-
2024.09	**Reparameterized Multi-Resolution Convolutions for Long Sequence Modelling**	arXiv	link	-
2024.08	**SE(3)-Hyena Operator for Scalable Equivariant Learning**	arXiv	link	-
2024.04	**LongVQ: Long Sequence Modeling with Vector Quantization on Structured Memory**	IJCAI'24	link	-
2024.02	**Transformer-VQ: Linear-Time Transformers via Vector Quantization**	ICLR'24	link	-
2024.01	**Scavenging Hyena: Distilling Transformers into Long Convolution Models**	arXiv	link	-

Tokenization

Time	Title	Venue	Paper	Code
2024.11	**Enhancing Large Language Models through Adaptive Tokenizers**	NeurIPS'24	link	-

Time	Title	Venue	Paper	Code
2024.11	**Theoretical Analysis of Byte-Pair Encoding**	arXiv	link	-
2024.10	**Model Decides How to Tokenize: Adaptive DNA Sequence Tokenization with MxDNA**	NeurIPS'24	link	link
2024.09	**BPE Gets Picky: Efficient Vocabulary Refinement During Tokenizer Training**	EMNLP'24	link	link
2024.09	**A Comparison of Tokenization Impact in Attention Based and State Space Genomic Language Models**	bioRxiv	link	-
2024.04	**Scaffold-BPE: Enhancing Byte Pair Encoding for Large Language Models with Simple and Effective Scaffold Token Removal**	arXiv	link	link
2024.04	**Effect of tokenization on transformers for biological sequences**	Bioinformatics	link	link
2024.02	**Tokenization Is More Than Compression**	arXiv	link	-
2023.10	**Toward Understanding BERT-Like Pre-Training for DNA Foundation Models**	arXiv	link	-

Quantization

Time	Title	Venue	Paper	Code
2024.06	**Low-Rank Quantization-Aware Training for LLMs**	arXiv	link	link

Fine-tuning

Time	Title	Venue	Paper	Code
2024.07	**LoRA+: Efficient Low Rank Adaptation of Large Models**	ICML'24	link	link
2021.10	**LoRA: Low-Rank Adaptation of Large Language Models**	arXiv	link	link
2024.07	**DoRA: Weight-Decomposed Low-Rank Adaptation**	ICML'24	link	link
2024.07	**Accurate LoRA-Finetuning Quantization of LLMs via Information Retention**	ICML'24	link	link
2024.05 🏋️🏋️	**Self-Distillation Bridges Distribution Gap in Language Model Fine-Tuning**	ACL'24	link	link

Reducing Knowledge Hallucination

Time	Title	Venue	Paper	Code
2024.06	**Coarse-to-Fine Highlighting: Reducing Knowledge Hallucination in Large Language Models**	ICML'24	link	link

Data processing

1、从 FASTA 文件中加载并查询基因组序列.
2、DNABERT2 Fine-Tuning for DHS Specificity Prediction.
3、Scaling-Laws-of-Genomic.
4、Deafness-mutation-sites.
5、DNABERT-2_CNN_BiLSTM.
6、1_Train_HG.
7、dbtk-dnabert.
8、DNABERT2_Tokenizer.
9、处理序列的R脚本.

Other Related Awesome Repository

01. Awesome-LLM-Learning
02. Scientific-LLM-Survey (Biological & Chemical Domains)
03. LLM-FineTuning-Large-Language-Models
04. Awesome-llms-fine-tuning (Explore a comprehensive collection of resources, tutorials, papers, tools, and best practices for fine-tuning Large Language Models (LLMs))
05. Awesome-LLM4RS-Papers
06. LLM4Rec-Awesome-Papers (A list of awesome papers and resources of recommender system on large language model (LLM))
07. Awesome-Code-LLM (A curated list of language modeling researches for code and related datasets)

Contributors

(back to top)