Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Structure-aware transfer of genomic foundation representations across population and cross-clade classification

**Xiaoyuan Liu**
Harbin Institute of Technology
**Tianyi Zhao**
zty2009@hit.edu.cn

Harbin Institute of Technology

**Research Article**

**Keywords:** genomic foundation model; ancestry classification; taxonomic classification; sequence representation

**Posted Date:** November 16th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11722538/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** The authors declare no competing interests.

*GigaByte*, 2026, 1–13

doi: xx.xxxx/xxxx
**Manuscript in Preparation**
**Paper**

PAPER

# Structure-aware transfer of genomic foundation representations across population and cross-clade classification

Xiaoyuan Liu[ORCID1] and Tianyi Zhao[ORCID1,*]

[1]Faculty of Computing, Harbin Institute of Technology, Harbin, China

*zty2009@hit.edu.cn

## Abstract

Genomic foundation models provide transferable DNA sequence representations, but their application to sample-level classification requires effective integration of sequencing reads. We present EvoGR (Evo 2-based Genomic Representation), a read-level transfer framework using a shared frozen Evo 2 encoder and Permutation-invariant Attention Aggregation (PAA). Mean pooling converts variable-length single-stranded reads into fixed-dimensional embeddings, while PAA integrates variable-sized, unordered read sets. EvoGR addresses two complementary tasks: population classification using AIM-enriched reads and four-class genome-origin classification involving human, non-human animal, plant, and microorganism sequences. Inference requires only single-stranded reads, without reference genomes, SNP coordinates, or explicit REF/ALT representations. EvoGR achieved competitive performance across both tasks and maintained strong classification performance on real human sequencing data, demonstrating transfer from simulated to real sequencing inputs. Comparisons across genomic foundation models identified Evo 2 as the strongest encoder under the evaluated settings. Ablation experiments demonstrated the contribution of PAA, while read-length sensitivity analyses showed that longer reads generally improved classification performance, with diminishing gains at greater lengths. These findings establish EvoGR as a unified framework for read-level genomic representation transfer across distinct biological classification tasks.

**Key words**: genomic foundation model; ancestry classification; taxonomic classification; sequence representation

## Introduction

With advances in self-supervised learning and large-scale pretraining, genomic foundation models (GFMs) have emerged as an important research direction in genomics. By pretraining on large-scale genomic sequences using self-supervised objectives, these models learn contextual representations of DNA sequences that can be transferred to a range of downstream genomic tasks [1, 2, 3, 4]. Early approaches largely adopted Transformer architectures from natural language processing, treating DNA sequences as biological languages with underlying structural patterns and using objectives such as masked language modeling to capture contextual relationships among nucleotides. Representative models include DNABERT and its successor DNABERT-2, as well as the Nucleotide Transformer [5, 6]. DNABERT-2 introduced byte pair encoding (BPE) for DNA tokenization, improving the efficiency of sequence modeling across multiple species, whereas the Nucleotide Transformer was pretrained on genomic datasets encompassing human and diverse non-human species, extending the taxonomic breadth of learned DNA representations. In parallel, alternative sequence modeling architectures have been developed to address the distinct characteristics of genomic sequences, particularly their length, strand orientation, and contextual dependencies. For example,

**Compiled on:** November 13, 2026.
Draft manuscript prepared by the author.

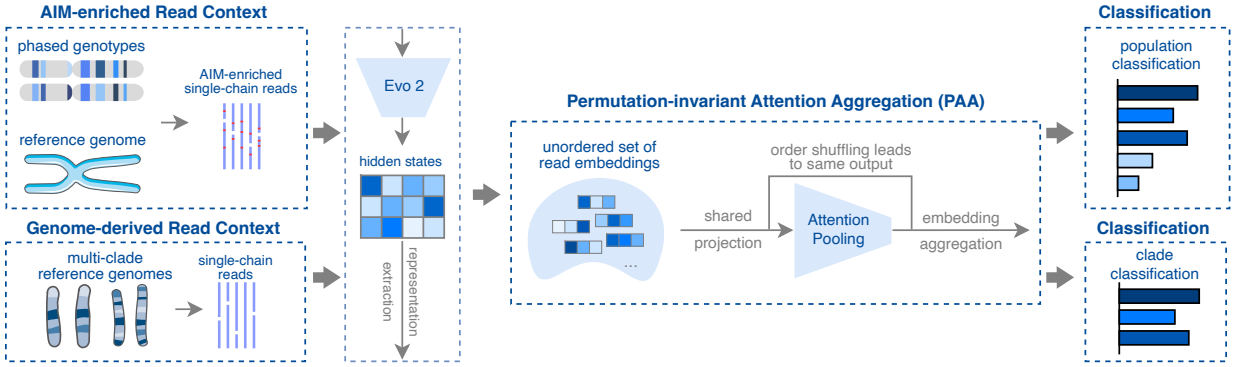



**Figure 1.** Overview of the EvoGR framework.

HyenaDNA employs autoregressive modeling based on the Hyena operator to support long-range sequence modeling at single-nucleotide resolution [7]. Caduceus combines bidirectional state-space modeling with reverse-complement equivariance to capture strand-related relationships in DNA sequences [8], while GROVER explores genome-specific sequence representations through genomic vocabulary learning and next-k-mer prediction [9]. More recently, advances in pretraining scale and model capacity have led to genomic foundation models such as Evo 2, which leverage large-scale genomic datasets spanning diverse domains of life and incorporate long-context sequence modeling, extending the application of GFMs to cross-species sequence analysis, genomic function prediction, and biological sequence generation [10]. Collectively, these developments demonstrate the potential of genomic foundation models to learn sequence representations at different levels of granularity and across varying contextual ranges from large-scale unlabeled DNA sequences, providing a general representation basis for genomic classification, genetic variant analysis, and other downstream biological applications. Recent systematic benchmarking studies of DNA foundation models have further highlighted the influence of model architecture, pretraining data, and representation extraction strategies on downstream task performance.

Despite substantial advances in DNA sequence representation learning, the general-purpose representations acquired during pretraining do not necessarily transfer effectively to all downstream biological tasks. Existing approaches typically adapt pretrained models either by freezing the encoder and extracting its hidden representations or by updating the entire model or a subset of its parameters through fine-tuning. These transfer strategies differ in their capacity for task adaptation, computational requirements, and dependence on training data. Moreover, the effectiveness of pretrained representations is influenced by the prediction units and the way biological information is organized in downstream tasks [11, 12]. In sequence-set classification, pretrained models typically produce contextual representations at the nucleotide or token level. These local representations must be further processed to obtain fixed-dimensional features suitable for sample-level prediction. Accordingly, feature extraction and aggregation serve as important steps linking pretrained sequence encoding to downstream classification [13, 14]. Mean pooling and max pooling provide computationally efficient means of converting local representations into fixed-dimensional vectors. However, mean pooling assigns equal weights to all elements, whereas max pooling retains only the strongest feature responses, and neither explicitly models the relative contributions of individual sequence segments to the final prediction. Attention-based aggregation offers an alternative by assigning data-dependent weights to input representations, thereby enabling their adaptive integration. The suitability of these aggregation strategies, however, depends on the organization of the input data and the biological information relevant to the downstream task. In particular, samples comprising multiple independent sequencing reads present a distinct representation challenge: reads generally lack a predefined spatial order or continuous positional relationship, while both read length and read count may vary across samples and sequencing conditions. Extracting informative sample-level representations from such unordered, variable-sized read sets is therefore an important consideration in transferring genomic foundation models to sequencing-based classification tasks. Furthermore, different downstream tasks involve genetic information at distinct biological scales and with different distributional characteristics. Consequently, the pretrained encoder alone may not provide representations that are equally effective across different classification objectives. Further investigation is therefore warranted into the relationship between pretrained genomic representations and downstream aggregation mechanisms, as well as the suitability of different aggregation strategies for distinct biological classification tasks [15, 16, 17].

In population genetics, demographic processes such as historical migration, genetic drift, gene flow, and admixture have shaped the genetic structure of human populations, resulting in differences in allele frequencies and patterns of genetic variation across the genome. Conventional population classification approaches typically rely on SNP genotypes, allele frequencies, or ancestry-informative markers (AIMs), combined with statistical genetics or machine learning methods for ancestry inference. With the growing availability of sequencing data, direct extraction of population-related genetic information from DNA sequencing reads has become an increasingly relevant research direction. Population classification provides a valuable setting for evaluating genomic foundation models because genetic differences among populations are generally reflected in subtle variations distributed across multiple genomic regions rather than in distinct sequence compositions. Models must therefore identify and integrate population-associated information from multiple independent reads to construct discriminative individual-level representations. Population classification thus provides a representative task for assessing the ability of genomic foundation models to capture fine-scale genetic variation within a species [18, 19, 20, 21].

Genomic sequences from different biological clades exhibit broader patterns of genetic divergence shaped by evolutionary history, phylogenetic relationships, and differences in genome composition. Identifying the biological origin of DNA sequences is an important task in genomic classification, biodiversity assessment, and metagenomic analysis. Conventional approaches primarily rely on sequence alignment, k-mer matching, and reference database searches to assign taxonomic origins by comparing query sequences with known references. Recent advances in deep learning and genomic foundation models have introduced alternative approaches based on learned sequence representations. Through pretraining on large-scale genomic datasets spanning multiple species, these models can capture sequence features and contextual patterns associated with different biological groups, providing a basis for cross-clade classification. In

**Table 1.** Performance comparison of population classification using AIM-enriched single-stranded sequencing reads.

| Method | MCC | AUPRC | AUROC | F1 |
|---|---|---|---|---|
| EvoGR | 0.9637 ± 0.0037 | 0.9828 ± 0.0071 | 0.9878 ± 0.0064 | 0.9647 ± 0.0037 |
| Neural ADMIXTURE [28] | 0.9573 ± 0.0098 | 0.9532 ± 0.0004 | 0.9563 ± 0.0061 | 0.9564 ± 0.0074 |
| AncestryGeni [29] | 0.9339 ± 0.0080 | 0.9393 ± 0.0087 | 0.9378 ± 0.0073 | 0.9285 ± 0.0040 |
| SALAI-Net [30] | 0.9133 ± 0.0032 | 0.9278 ± 0.0115 | 0.8936 ± 0.0086 | 0.9127 ± 0.0087 |
| ntRoot [31] | 0.9210 ± 0.0032 | 0.9113 ± 0.0131 | 0.9029 ± 0.0164 | 0.9112 ± 0.0129 |

**Table 2.** Performance comparison of four-class genome-origin classification using variable-length single-stranded sequencing reads.

| Method | MCC | AUPRC | AUROC | F1 |
|---|---|---|---|---|
| EvoGR | 0.9520 ± 0.0098 | 0.9540 ± 0.0028 | 0.9635 ± 0.0043 | 0.9632 ± 0.0062 |
| DNABERT-S [32] | 0.9418 ± 0.0051 | 0.9289 ± 0.0053 | 0.9214 ± 0.0035 | 0.9205 ± 0.0055 |
| BERTax [33] | 0.9361 ± 0.0051 | 0.9289 ± 0.0016 | 0.9324 ± 0.0023 | 0.9585 ± 0.0039 |
| DeepTaxa [34] | 0.9251 ± 0.0037 | 0.9281 ± 0.0022 | 0.9392 ± 0.0066 | 0.9443 ± 0.0041 |
| BarcodeBERT [35] | 0.8960 ± 0.0024 | 0.8930 ± 0.0042 | 0.9012 ± 0.0063 | 0.8791 ± 0.0077 |

this study, genome-origin classification is formulated as a four-class task involving human, non-human animal, plant, and microorganism sequences. Unlike population classification, which focuses on subtle genetic differences within a species, this task evaluates the ability of models to distinguish sequence features across broader biological groups. Genome-origin classification therefore provides a complementary setting for assessing the transferability of genomic representations across taxonomic clades [22, 23, 24].

Although population classification and genome-origin classification address distinct biological objectives, both can be formulated as sample-level prediction tasks based on sets of independent sequencing reads. Read length and read count vary across samples and sequencing conditions, and individual reads generally lack a predefined ordering. Accordingly, sequencing reads can be represented as unordered sets of variable size. For genomic foundation models, token-level representations from reads of different lengths must first be converted into fixed-dimensional read-level embeddings, which are then aggregated to obtain sample-level representations. Permutation-invariant aggregation methods provide a means of integrating read-level information independently of input order, while attention-based mechanisms can assign adaptive weights according to the characteristics of individual reads. A unified framework for read-level representation and aggregation therefore provides a common methodological basis for transferring genomic foundation models across distinct biological classification tasks [25, 26, 27].

Building on this rationale, we propose EvoGR (Evo 2-based Genomic Representation), a genomic foundation model transfer framework designed for sets of single-stranded sequencing reads. EvoGR employs a frozen Evo 2 model as a shared encoder, applies mean pooling to convert token-level representations of variable-length reads into fixed-dimensional read-level embeddings, and uses a unified Permutation-invariant Attention Aggregation (PAA) architecture to integrate variable numbers of reads into sample-level representations for classification. The framework is evaluated on two complementary tasks: population classification using AIM-enriched reads and four-class genome-origin classification involving human, non-human animal, plant, and microorganism sequences. At inference, EvoGR requires only single-stranded DNA reads, without access to a reference genome, SNP coordinates, or explicit REF/ALT allele pairs. We evaluate the framework using both simulated and real sequencing data, compare alternative genomic foundation model encoders, and conduct ablation and read-length sensitivity analyses to investigate the effects of encoder selection and read-level aggregation on classification performance.

## Methods

### Overview of EvoGR

In this study, we propose EvoGR (Evo 2-based Genomic Representation), a genomic foundation model transfer framework for learning sample-level representations from single-stranded DNA sequencing reads and performing genomic classification across different biological levels. The overall architecture is illustrated in Fig. 1. EvoGR employs a frozen Evo 2 model as a shared sequence encoder for both population classification and genome-origin classification. Each sequencing read is independently processed by Evo 2 to obtain its token-level hidden states, which are subsequently converted into fixed-dimensional read-level embeddings through mean pooling. The embeddings from the same sample are then organized as an unordered set and passed to a Permutation-invariant Attention Aggregation (PAA) module. Through attention-weighted aggregation, PAA integrates sequence information across individual reads to generate a sample-level representation. Finally, the aggregated representation is passed to the corresponding downstream classifier to predict either the population category of an individual or the genomic origin of the input sample.

EvoGR comprises two task-specific branches that differ in data construction and classification objectives. For population classification, AIM-enriched single-stranded sequencing reads are generated using phased VCF data from the 1000 Genomes Project, a reference genome, and ancestry-informative markers (AIMs). These reads provide sequence-level information associated with population-specific genetic variation. For genome-origin classification, variable-length single-stranded reads are generated by randomly extracting DNA sequences from reference genomes representing different biological groups. The resulting classification task distinguishes among human, non-human animal, plant, and microorganism sequences. Both branches use the same Evo 2 encoding procedure and PAA architecture, while their downstream classification networks are trained separately. At inference, EvoGR operates directly on single-stranded DNA reads without requiring a reference genome, SNP coordinates, or explicit REF/ALT allele pairs.

## Construction of Single-stranded Sequencing Read Datasets

***Population Read Generation***

The population classification dataset was constructed using phased VCF data from the 1000 Genomes Project. First, individuals and their corresponding population labels were determined according to the predefined classification scheme. SNP loci included in the AIM panel were then extracted from the corresponding variant data. By combining the reference genome sequence with the allele information in the phased VCF, the two haplotypes of each individual were reconstructed separately, yielding DNA sequences containing individual-specific genetic variation.

Based on the genomic coordinates of the selected AIM loci, sequence contexts surrounding these loci were extracted from the reconstructed haplotypes to generate AIM-enriched single-stranded sequencing reads. A variable number of reads was generated for each individual, with read lengths sampled according to a predefined length distribution. Through this procedure, each individual was represented by a collection of independent single-stranded reads enriched for AIM-associated sequence information, rather than by explicit REF/ALT strand representations or diploid variant feature matrices.

The resulting population classification dataset was defined as:

$$\mathcal{D}_P = \{(\mathcal{R}_i, y_i)\}_{i=1}^{M_P}, \tag{1}$$

where $M_P$ denotes the number of population samples, $y_i$ represents the population label of the $i$-th individual, and $\mathcal{R}_i$ denotes its corresponding read set:

$$\mathcal{R}_i = \left\{r_{i1}, r_{i2}, \ldots, r_{iN_i}\right\}, \tag{2}$$

where $N_i$ is the number of reads associated with the $i$-th sample. As the number of reads may vary across individuals, this representation accommodates sequencing inputs of variable sizes. Notably, the reference genome, phased VCF data, and AIM panel are used exclusively for offline training data construction and are not required as model inputs during inference.

***Genome-origin Read Generation***

The genome-origin classification dataset was constructed from reference genomes representing four broad biological categories: human, non-human animal, plant, and microorganism. Reference genomes with well-defined taxonomic assignments were selected for each category, followed by the assignment of samples to training, validation, and test sets according to the predefined data partitioning strategy. DNA sequences of specified lengths were then randomly extracted from the selected reference genomes to simulate short-read sequencing inputs.

To evaluate the model's ability to process variable-length sequences, datasets were generated using different read lengths. Each sample consisted of a variable number of independent single-stranded reads, without imposing an artificial ordering among them. Each read retained its corresponding genomic origin, while the sample-level label was determined by its assigned genome-origin category.

The genome-origin classification dataset was defined as:

$$\mathcal{D}_C = \{(\mathcal{R}_i, y_i)\}_{i=1}^{M_C}, \tag{3}$$

where $M_C$ denotes the number of genome-origin classification samples, $y_i$ represents the corresponding genome-origin label, and $\mathcal{R}_i$ denotes the input set of independent single-stranded reads. The four classification categories are Human, Non-human Animal, Plant, and Microorganism.

With this data construction strategy, both population classification and genome-origin classification use variable-length and variable-sized sets of single-stranded sequencing reads, providing a unified read-level input interface for EvoGR.

## Read-level Representation Extraction Using Evo 2

EvoGR employs a pretrained Evo 2 model as its shared sequence encoder. All pretrained parameters are frozen during downstream training to preserve the general genomic representations acquired during pretraining while reducing the computational cost of model optimization. For each input read, the DNA sequence is first tokenized according to the Evo 2 tokenizer and subsequently processed by the encoder to obtain its hidden representations.

Let the $j$-th read from the $i$-th sample be represented as:

$$r_{ij} = (x_1, x_2, \ldots, x_{L_{ij}}), \tag{4}$$

where $L_{ij}$ denotes the number of valid tokens in the read. The Evo 2 encoder transforms the input sequence into a token-level hidden-state matrix:

$$H_{ij} = \mathrm{Evo2}(r_{ij}) \in \mathbb{R}^{L_{ij} \times d}, \tag{5}$$

where $d$ denotes the dimensionality of the hidden representations.

Because input reads may differ in length, their token-level representations have varying numbers of tokens. To obtain fixed-dimensional read-level features, mean pooling is applied to the hidden states within each read:

$$h_{ij} = \frac{1}{L_{ij}} \sum_{t=1}^{L_{ij}} H_{ij,t}, \tag{6}$$

where $H_{ij,t}$ denotes the hidden vector corresponding to the $t$-th token of the $j$-th read, and $h_{ij} \in \mathbb{R}^d$ represents the resulting read-level embedding.

This operation converts reads of different lengths into vectors of identical dimensionality, thereby removing sequence-length variation from the input dimensions of subsequent set aggregation. Importantly, mean pooling at this stage operates exclusively on token-level representations within individual reads and is distinct from PAA, which aggregates multiple reads at the sample level.

The collection of read-level representations for the $i$-th sample is defined as:

$$\mathcal{H}_i = \left\{ h_{i1}, h_{i2}, \ldots, h_{iN_i} \right\}, \tag{7}$$

where $N_i$ denotes the number of reads in the sample. This set serves as the input to the PAA module.

## Permutation-invariant Attention Aggregation

To integrate sequence information across multiple reads from the same sample, EvoGR employs a Permutation-invariant Attention Aggregation (PAA) module that transforms a variable number of read-level embeddings into a fixed-dimensional sample-level representation. PAA consists of three main operations: shared feature projection, attention weight computation, and weighted feature aggregation. The resulting aggregation is independent of the ordering of the input reads.

**Shared Projection.** First, each read-level embedding is passed through a shared feature projection layer to map the original read-level features into a common attention representation space:

$$u_{ij} = f_\theta(h_{ij}), \tag{8}$$

where $f_\theta(\cdot)$ denotes a learnable projection function shared across all reads, $\theta$ represents its parameters, and $u_{ij}$ denotes the projected read representation.

**Attention Pooling.** Following feature projection, an attention scoring function is applied to estimate the contribution of each read to the sample-level representation:

$$e_{ij} = g_\phi(u_{ij}), \tag{9}$$

where $g_\phi(\cdot)$ denotes a learnable attention scoring function parameterized by $\phi$, and $e_{ij}$ represents the unnormalized attention score of the $j$-th read. The scores are then normalized across all reads within the same sample using the softmax function:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{N_i} \exp(e_{ik})}, \tag{10}$$

where $\alpha_{ij}$ denotes the attention weight assigned to the $j$-th read in the current sample, satisfying:

$$\sum_{j=1}^{N_i} \alpha_{ij} = 1. \tag{11}$$

**Embedding Aggregation.** Finally, the projected read-level representations are combined through an attention-weighted sum to obtain the sample-level representation:

$$z_i = \sum_{j=1}^{N_i} \alpha_{ij} u_{ij}, \tag{12}$$

where $z_i$ denotes the fixed-dimensional representation of the $i$-th sample after PAA aggregation. Regardless of the number of reads in a sample, PAA produces an output of identical dimensionality, allowing the model to accommodate read sets of varying sizes.

**Permutation Invariance.** PAA is permutation-invariant, meaning that reordering the reads within a sample does not change the resulting aggregated representation. Since all reads are processed using the same feature projection and attention scoring functions, and the final aggregation is performed through a weighted sum, the following equality holds for any permutation $\pi$:

$$\mathrm{PAA}\left(\{h_{i1}, \ldots, h_{iN_i}\}\right) = \mathrm{PAA}\left(\{h_{i\pi(1)}, \ldots, h_{i\pi(N_i)}\}\right). \tag{13}$$

This property enables PAA to aggregate variable-sized, unordered read sets into sample-level representations without introducing artificial positional encodings or dependencies on read ordering.

## Downstream Classification Tasks

### *Population Classification*

For population classification, the sample-level representation generated by PAA is used to predict the predefined population category of each individual. The aggregated representation $z_i$ of the $i$-th sample is passed to the downstream classification network to obtain the

**Table 3.** Population classification performance on real human sequencing data.

| Method | MCC | AUPRC | AUROC | F1 |
|---|---|---|---|---|
| EvoGR | 0.9617 | 0.9892 | 0.9717 | 0.9652 |
| ntRoot [31] | 0.9412 | 0.9353 | 0.9394 | 0.9241 |
| AncestryGeni [29] | 0.9275 | 0.9380 | 0.9241 | 0.9516 |
| iAdmix [36] | 0.9015 | 0.9105 | 0.9067 | 0.8932 |
| LASER [37] | 0.8977 | 0.8686 | 0.8909 | 0.9068 |

predicted population probabilities:

$$\hat{y}_i^P = \text{Softmax}\left(F_{\psi_P}(z_i)\right), \tag{14}$$

where $F_{\psi_P}(\cdot)$ denotes the population classification network, $\psi_P$ represents its learnable parameters, and $\hat{y}_i^P$ denotes the predicted probability distribution over the population categories.

The population classification network is optimized using the population labels in the training data, enabling the model to identify sequence features associated with differences in population genetic backgrounds from AIM-enriched single-stranded reads. The model ultimately outputs a predicted population category for each individual.

***Genome-origin Classification***

For genome-origin classification, the same Evo 2 encoder and PAA architecture are used to generate sample-level representations, which are then passed to a separate classification network:

$$\hat{y}_i^C = \text{Softmax}\left(F_{\psi_C}(z_i)\right), \tag{15}$$

where $F_{\psi_C}(\cdot)$ denotes the genome-origin classification network, $\psi_C$ represents its learnable parameters, and $\hat{y}_i^C$ denotes the predicted probability distribution over the four genome-origin categories.

The classification labels comprise Human, Non-human Animal, Plant, and Microorganism. By aggregating sequence representations from multiple independent reads, the model constructs a sample-level representation of genomic origin and performs four-class classification.

Although the two classification tasks employ the same PAA architecture, their corresponding training datasets, classification labels, and downstream classification networks are optimized independently.

## Results

### Benchmark Performance of EvoGR Across Population and Genome-origin Classification

To evaluate the performance of EvoGR across different genomic classification tasks, we established separate benchmark datasets for population classification and genome-origin classification and compared the proposed framework with representative existing methods for each task. Since some population classification methods operate on SNP genotypes or genetic variation profiles rather than raw DNA sequences, we designed the comparisons to use the same individuals, SNP loci, and data partitions wherever applicable. Methods based on genetic variation were evaluated using the corresponding SNP data, whereas EvoGR used the reference sequences and corresponding phased VCF data to reconstruct the two haplotypes and generate AIM-enriched single-stranded sequencing reads as model inputs. For genome-origin classification, all methods were evaluated using the same test samples, class labels, and data partitions to ensure a consistent basis for comparison. All experiments were conducted using five-fold cross-validation.

Performance was assessed using the Matthews correlation coefficient (MCC), macro-averaged area under the precision-recall curve (Macro-AUPRC), macro-averaged area under the receiver operating characteristic curve (Macro-AUROC), and macro-F1 score. The results for population classification and genome-origin classification are presented in Tables 1 and 2, respectively. The population classification benchmark included representative ancestry inference methods, namely Neural ADMIXTURE, AncestryGeni, SALAI-Net, and ntRoot.

For population classification, EvoGR achieved the best performance across all four evaluation metrics, outperforming the other methods included in the comparison. Neural ADMIXTURE showed performance relatively close to that of EvoGR, whereas AncestryGeni, SALAI-Net, and ntRoot exhibited varying performance across the evaluation metrics. The consistent performance of EvoGR across all four metrics indicates that its sequence representations capture information associated with population-level genetic differences. The subsequent application of Permutation-invariant Attention Aggregation (PAA) further integrates genetic information across multiple AIM-enriched reads into a discriminative sample-level representation. For genome-origin classification, EvoGR achieved the best results in terms of MCC, Macro-AUPRC, and Macro-AUROC, while BERTax obtained a higher Macro-F1 score. The remaining methods also showed differences across the evaluation metrics, indicating variations in their ability to distinguish among the four genome-origin categories. Overall, EvoGR demonstrated strong performance in terms of classification correlation and class discrimination, as reflected by its leading results on three of the four metrics, while maintaining competitive Macro-F1 performance. These findings indicate that EvoGR can extract discriminative representations from variable-length single-stranded sequencing reads and distinguish among human, non-human animal, plant, and microorganism sequences.

Taken together, the results from both classification tasks demonstrate that EvoGR supports sample-level representation learning and classification across distinct biological levels using a shared frozen Evo 2 encoder and a common Permutation-invariant Attention Aggregation (PAA) architecture. The findings suggest that pretrained genomic representations, combined with permutation-invariant

**Table 4.** Performance comparison of different genomic foundation models on population classification.

| Genomic Foundation Model | MCC | AUPRC | AUROC | F1 |
|---|---|---|---|---|
| Evo 2 [10] | 0.9637 | 0.9828 | 0.9878 | 0.9647 |
| NTv3 [6] | 0.9524 | 0.9504 | 0.9495 | 0.9587 |
| DNABERT-2 [5] | 0.9346 | 0.9438 | 0.9269 | 0.9342 |
| HyenaDNA [7] | 0.9282 | 0.9136 | 0.9091 | 0.9121 |
| GROVER [9] | 0.9321 | 0.9297 | 0.9128 | 0.9273 |
| Caduceus [8] | 0.9073 | 0.8992 | 0.9135 | 0.8939 |

**Table 5.** Performance comparison of different genomic foundation models on four-class genome-origin classification.

| Genomic Foundation Model | MCC | AUPRC | AUROC | F1 |
|---|---|---|---|---|
| Evo 2 [10] | 0.9520 | 0.9540 | 0.9635 | 0.9632 |
| NTv3 [6] | 0.9278 | 0.9263 | 0.9383 | 0.9203 |
| DNABERT-2 [5] | 0.9162 | 0.9211 | 0.9337 | 0.9275 |
| HyenaDNA [7] | 0.9073 | 0.9164 | 0.8931 | 0.8772 |
| GROVER [9] | 0.8860 | 0.8998 | 0.8721 | 0.8842 |
| Caduceus [8] | 0.8976 | 0.8643 | 0.8935 | 0.8742 |

attention-based aggregation, can capture information relevant to both population-level genetic variation and genome-origin differences across broad taxonomic groups. This provides an empirical basis for applying genomic foundation model representations to downstream classification tasks involving distinct biological input structures.

## Evaluation on Real Human Sequencing Data

To further assess the performance of EvoGR on real sequencing data, we assembled an independent test set comprising real human sequencing samples and selected representative ancestry inference methods that utilize DNA sequences or sequencing data for comparison, including ntRoot, AncestryGeni, iAdmix, and LASER [31, 29, 36, 37]. For EvoGR, the training data were generated by reconstructing haplotypes from a predefined SNP panel and the reference genome, followed by the extraction of AIM-enriched single-stranded sequencing reads. The resulting model was then evaluated directly on the real human sequencing data, which served as an independent test set. This design allowed us to examine the transfer of the model from simulated sequencing data to real sequencing data, where the data sources and generation procedures differ between training and evaluation. All comparison methods were evaluated on the same test samples using the same population labels. Performance was assessed using MCC, AUPRC, AUROC, and F1, with the results summarized in Table 3.

EvoGR achieved the best performance across all four evaluation metrics, demonstrating consistent classification performance on real human sequencing data. Among the comparison methods, ntRoot showed performance closest to that of EvoGR in terms of MCC, AUPRC, and AUROC, whereas AncestryGeni showed competitive performance in terms of F1. By comparison, iAdmix and LASER yielded lower overall performance, although their results varied across the individual metrics. These findings indicate that the sequence representations learned from simulated sequencing data by EvoGR retain population-discriminative information when applied to real human sequencing data. Together with the results from the simulated-data benchmark, these observations support the use of Evo 2-derived sequence representations in combination with Permutation-invariant Attention Aggregation to integrate population-related genetic information across single-stranded sequencing reads from different data sources. This provides a practical approach to population classification directly from sequencing data.

## Comparative Evaluation of Different Genomic Foundation Models

To investigate the influence of pretrained genomic representations on downstream classification performance and assess the suitability of Evo 2 as a shared encoder, we evaluated several representative genomic foundation models, including Nucleotide Transformer v3 (NTv3), DNABERT-2, HyenaDNA, GROVER, and Caduceus. These models encompass different architectural designs and pretraining strategies, including Transformer-based masked language modeling, Hyena-based long-range sequence modeling, and bidirectional sequence modeling with state-space architectures [6, 5, 7, 9, 8]. To ensure a fair comparison, all models were evaluated using identical training and test set partitions, with the downstream classification architecture, training configuration, and evaluation metrics kept consistent. The only variable was the genomic foundation model used to extract sequence representations. For both classification tasks, the same Permutation-invariant Attention Aggregation (PAA) architecture was applied to aggregate read-level representations, allowing the representation capabilities of different pretrained encoders to be assessed within a common downstream adaptation framework. The results for population classification and genome-origin classification are presented in Tables 4 and 5, respectively.

For genome-origin classification, Evo 2 achieved the best performance across all four evaluation metrics, outperforming the other genomic foundation models. NTv3 yielded results closest to those of Evo 2 across the evaluated metrics, while DNABERT-2 also showed competitive performance on selected metrics. In comparison, HyenaDNA, GROVER, and Caduceus exhibited varying levels of performance across the metrics, with lower overall results than Evo 2. These findings indicate that the sequence representations learned by different genomic foundation models vary in their discriminative capacity for cross-clade genome-origin classification. Among the evaluated encoders, Evo 2 provided representations that were particularly effective in distinguishing the four genome-origin categories. A similar pattern was observed in population classification, where Evo 2 again achieved the best performance across all four metrics, while NTv3 consistently ranked among the strongest alternative encoders. DNABERT-2 and GROVER showed comparable overall performance, al-

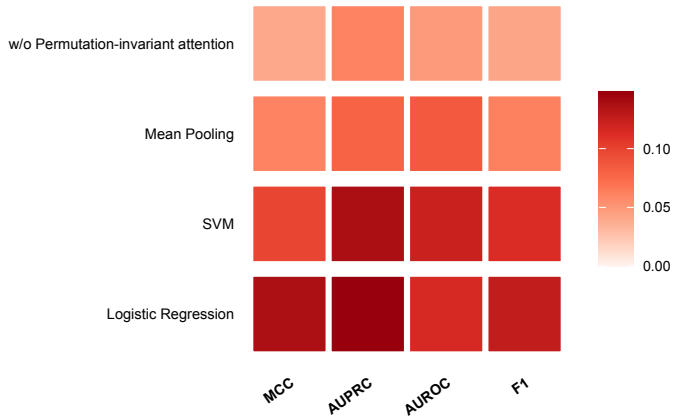



**Figure 2. Ablation analysis of downstream aggregation and classification strategies for population classification.** Color intensity represents the absolute decrease in each evaluation metric relative to the complete EvoGR model, with darker red indicating a greater performance decrease.

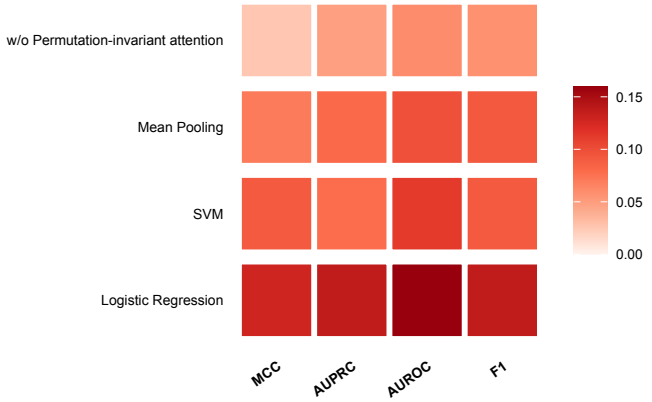



**Figure 3. Ablation analysis of downstream aggregation and classification strategies for genome-origin classification.** Color intensity represents the absolute decrease in each evaluation metric relative to the complete EvoGR model, with darker red indicating a greater performance decrease.

though their relative rankings differed across evaluation metrics. HyenaDNA and Caduceus yielded comparatively lower classification performance. The differences in model rankings between the two tasks suggest that the effectiveness of pretrained genomic representations depends, at least in part, on the downstream classification objective and the characteristics of the underlying genetic information.

Taken together, the results demonstrate that Evo 2 provides consistently effective representations across two classification tasks operating at distinct biological levels, supporting the identification of both population-level genetic variation and genome-origin differences across broad taxonomic groups. The competitive performance of NTv3 across both tasks further indicates the potential of representations learned through multi-species pretraining for transfer across different genomic classification settings. Overall, the observed differences among foundation models highlight the importance of encoder selection for downstream classification. Under the same PAA architecture, Evo 2 provided effective sequence representations for both tasks, supporting its use as the shared encoder in EvoGR.

## Ablation Analysis of Downstream Architectures

To assess the contribution of the downstream aggregation architecture to the overall performance of EvoGR and examine the effectiveness of Permutation-invariant Attention Aggregation (PAA) relative to alternative approaches, we conducted ablation experiments for both population classification and genome-origin classification. To isolate the effects of downstream adaptation from those of pretrained representations, all experiments used the same read-level embeddings extracted by the frozen Evo 2 encoder, with identical data partitions and evaluation metrics. Only the downstream aggregation or classification strategy was varied. Specifically, we evaluated a model without the PAA module and compared the full EvoGR architecture with alternative approaches based on mean pooling, support vector machines (SVMs), and logistic regression. The results are presented as absolute decreases in each evaluation metric relative to the full EvoGR model and visualized as heatmaps, where darker colors indicate larger performance decreases. The ablation results for population classification and genome-origin classification are shown in Figures 2 and 3, respectively.

For population classification, removing the PAA module resulted in decreases across all evaluation metrics, indicating that read-level embeddings alone are insufficient to fully integrate the population-related genetic information contained in different reads. Replacing PAA with mean pooling led to further performance degradation, suggesting that assigning equal weights to all read representations does not adequately account for their varying contributions to population discrimination. The SVM- and logistic regression-based alternatives also showed substantial performance decreases, with logistic regression exhibiting the largest reductions across the evaluated metrics. These results indicate that neither simple aggregation of Evo 2-derived representations nor conventional linear classification fully reproduces the performance of the complete model. In contrast, PAA adaptively assigns attention weights according to the characteristics of

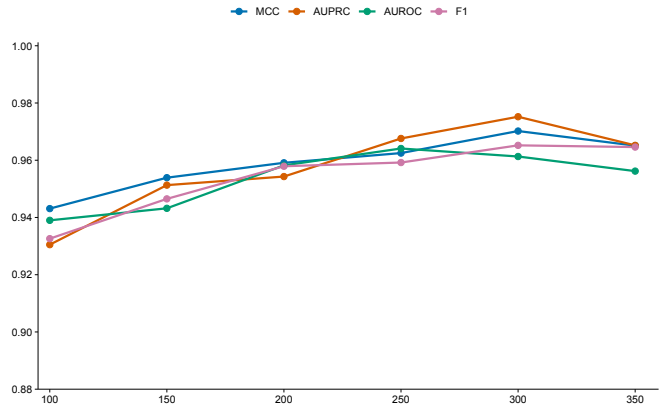



**Figure 4. Population classification performance across different sequencing read lengths.** The x-axis represents the input read length (bp), and the y-axis represents the corresponding classification metric values. The colored curves show the performance trends of MCC, AUPRC, AUROC, and F1 across different read lengths.

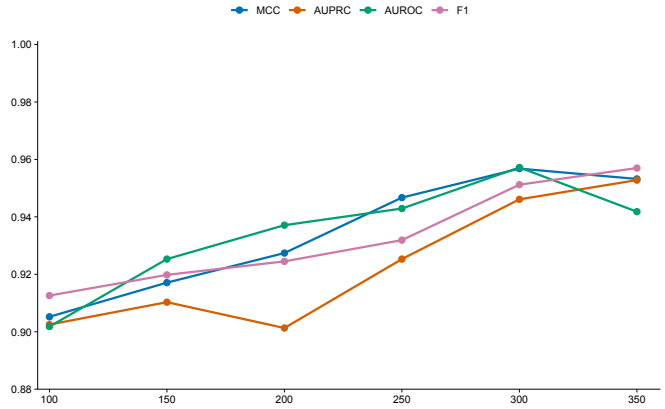



**Figure 5. Classification performance of genome-origin classification across different sequencing read lengths.** The x-axis represents the input read length (bp), and the y-axis represents the corresponding classification metric values. The colored curves show the performance trends of MCC, AUPRC, AUROC, and F1 across different read lengths.

individual read representations and combines them into a sample-level representation, allowing genetic information distributed across AIM-enriched reads to be integrated more effectively.

A similar pattern was observed in genome-origin classification. Removing PAA resulted in decreases across all evaluation metrics, while replacing it with mean pooling led to further performance losses. This suggests that equal-weight aggregation does not fully capture the discriminative information carried by individual reads. The SVM- and logistic regression-based alternatives also showed substantial decreases across the evaluated metrics, with logistic regression again exhibiting the largest reductions. These findings indicate that simple feature aggregation and conventional classifiers do not fully exploit the discriminative information encoded in pretrained sequence representations when samples are represented as variable-sized read sets. By learning read-specific attention weights and adaptively aggregating their representations, PAA provides a more effective sample-level representation for distinguishing among human, non-human animal, plant, and microorganism sequences.

Taken together, the ablation results demonstrate that removing PAA or replacing the complete downstream architecture with simpler aggregation or conventional classification methods leads to varying degrees of performance degradation in both tasks. Mean pooling does not account for differences in the contributions of individual reads, while the conventional classifiers evaluated here do not achieve the performance of the full model when applied to the same read-set representations. These findings demonstrate that downstream aggregation has a measurable influence on classification performance, independently of the pretrained Evo 2 representations. By applying attention-weighted aggregation to unordered read sets, PAA provides an effective means of integrating sequence information across reads and supports its use in both population and genome-origin classification.

## Sensitivity Analysis of Variable-length Sequencing Reads

To investigate the effect of input sequence length on EvoGR classification performance, we conducted read-length sensitivity analyses for both population classification and genome-origin classification. Specifically, read lengths were varied across 100, 150, 200, 250, 300, and 350 bp, while the sample partitions, model architecture, and other experimental settings were kept unchanged. Classification performance at each read length was evaluated using MCC, AUPRC, AUROC, and F1 to examine the model's response to different sequence context lengths. The results for population classification and genome-origin classification are presented in Figures 4 and 5, respectively.

For population classification, all four evaluation metrics showed an overall improvement as read length increased, indicating that longer sequence contexts provided additional information relevant to population-level genetic differences. MCC, AUPRC, and F1 generally

improved with increasing read length before approaching a plateau at longer lengths. AUROC, in contrast, already showed relatively strong performance at intermediate read lengths, with only modest fluctuations thereafter. These patterns suggest that extending the read length can provide additional local sequence context for population discrimination, although the resulting gains do not increase indefinitely. Once the read length reaches a certain range, the additional sequence context contributes progressively less to classification performance.

A similar overall trend was observed in genome-origin classification. Increasing read length generally improved the evaluation metrics, suggesting that longer sequence contexts provide additional discriminative information for distinguishing among genome-origin categories. MCC, AUPRC, and F1 showed overall improvements with increasing read length and achieved stronger performance at longer lengths. AUROC also increased initially but fluctuated at the longer read lengths, with no consistent improvement from further increases in sequence length. The different responses of individual metrics indicate that the effect of sequence context length on classification performance varies depending on the evaluation criterion. Overall, longer reads improved the discrimination of sequences derived from human, non-human animal, plant, and microorganism genomes, although the additional gains became limited as read length increased.

Taken together, these results identify read length as an important factor influencing EvoGR classification performance. Increasing the amount of sequence context generally benefits both classification tasks, although the magnitude of improvement varies across evaluation metrics. At shorter read lengths, the limited sequence context may constrain the amount of task-relevant information available to the model. As read length increases, additional local sequence features become available, while the resulting performance gains gradually diminish at longer lengths. This pattern is consistent with the role of contextual sequence representations in genomic foundation models. Overall, EvoGR accommodates sequencing reads of different lengths and maintains stable classification performance across the evaluated range, while the results indicate that longer input sequences do not necessarily lead to continuous improvements in predictive performance.

## Discussion

In this study, we present EvoGR, a read-level genomic representation transfer framework built upon a frozen Evo 2 encoder. By combining the shared encoder with a unified Permutation-invariant Attention Aggregation (PAA) architecture, EvoGR integrates variable-length and variable-sized sets of single-stranded sequencing reads into sample-level representations for population and genome-origin classification. The experimental results demonstrate strong classification performance across both tasks, with the model also maintaining good performance on real human sequencing data. Comparisons across different genomic foundation models further show that encoder selection influences downstream classification performance. Moreover, replacing PAA with simpler aggregation methods or conventional classifiers, while keeping the Evo 2 representations unchanged, resulted in performance degradation. These findings indicate that downstream aggregation contributes to classification performance alongside the pretrained sequence representations. By learning attention weights across individual reads, PAA provides a means of integrating genetic information distributed across an unordered and variable-sized read set.

Population classification and genome-origin classification capture genetic differences at distinct biological levels and therefore provide complementary settings for evaluating genomic foundation models. Population classification focuses on fine-scale genetic variation within a species, with discriminative information distributed across multiple genomic regions. In contrast, genome-origin classification examines sequence differences across broader biological groups and assesses the ability of pretrained representations to transfer across taxonomic clades. Although the two tasks have different classification objectives, both require the integration of information from multiple independent reads. The use of a shared Evo 2 encoder and PAA architecture across both tasks provides a unified framework for examining the transferability of genomic representations across different levels of biological variation.

In this study, genome-origin classification was formulated as a four-class task comprising human, non-human animal, plant, and microorganism sequences. This coarse-grained classification scheme provides a common evaluation setting for distinguishing among major biological groups with distinct genomic compositions and evolutionary backgrounds. However, substantial genetic diversity exists within each category, and closely related species may exhibit considerable sequence similarity. Consequently, the current results primarily reflect classification performance at a broad taxonomic level and do not establish reliable genus- or species-level identification. Future work could extend the framework to datasets with finer taxonomic annotations, covering family-, genus-, and species-level classification, and investigate hierarchical classification strategies to support predictions across multiple taxonomic levels.

For population classification, EvoGR extracts population-associated genetic information directly from single-stranded sequencing reads, providing an alternative approach to classification from sequencing data. However, human population structure is continuous and shaped by historical migration, genetic drift, gene flow, and admixture. Individuals with mixed ancestry may not be adequately represented by a single predefined population label. Accordingly, the predictions of the current model should be interpreted as genetic similarity to the population categories represented in the training data, rather than as estimates of an individual's full ancestry proportions or local ancestry composition. The diversity and representativeness of the reference populations included in the training data also influence the model's applicability. Future extensions could incorporate continuous ancestry labels, uncertainty estimation, and open-set recognition to improve performance on individuals with complex ancestry backgrounds and populations not adequately represented during training.

EvoGR may also have potential applications in biodiversity surveys, environmental DNA (eDNA) analysis, and the preliminary screening of sequences with unknown origins. By classifying sequencing reads according to their genomic origins, the framework could provide initial information to support subsequent sequence annotation and taxonomic analysis. However, the current model aggregates reads at the sample level and produces a single classification result for each read set. It is therefore not directly designed to resolve complex environmental samples containing DNA from multiple biological sources. Future extensions could combine read-level classification with multi-label prediction and the aggregation of individual classification results to support the analysis of mixed-community composition. For sequences originating from uncharacterized species or taxa poorly represented in reference databases, open-set recognition and subsequent taxonomic validation would also be necessary to improve the reliability of source identification.

Genomic foundation models differ in their architectures, pretraining objectives, tokenization schemes, and training data, all of which can influence the resulting sequence representations and their transferability to downstream tasks. By replacing the pretrained encoder while keeping the downstream PAA architecture unchanged, this study provides an initial assessment of the compatibility of PAA with different genomic foundation models. Evo 2 achieved the strongest overall performance under the evaluated experimental settings. Nevertheless, differences in representation dimensions, contextual features, and information distributions across encoders may affect their com-

patibility with downstream aggregation mechanisms. Further work is therefore needed to investigate encoder-specific adaptation strategies and parameter-efficient fine-tuning approaches. Overall, EvoGR combines pretrained genomic representations with permutation-invariant read-set aggregation to support single-stranded sequencing read classification across distinct biological levels. This framework provides a basis for extending genomic foundation model representations to genomic classification and other sequencing-based biological applications.

## Availability of source code and requirements

- **Project name:** EvoGR
- **Project home page:** https://github.com/qiankunzizairen/EvoGR
- **Operating system:** Linux (x86_64)
- **Programming language:** Python
- **Other requirements:** Python 3.10 or later
- **License:** MIT License

## Data Availability

The data used in this study were primarily obtained from publicly available databases and controlled-access data resources.

The 1000 Genomes high-coverage dataset is publicly available through the official 1000 Genomes Project repository [38]. In this study, we used the high-coverage sequencing and variant data provided by this project. Specifically, phased VCF data from 3,202 individuals were obtained from: https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/20220422_3202_phased_SNV_INDEL_SV/

The sequence index file containing sample information and sequencing data records for 2,504 unrelated individuals was obtained from: https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/1000G_2504_high_coverage.sequence.index

The list of phased VCF files actually used in this study has been archived in the Open Science Framework (OSF) repository (https://doi.org/10.17605/OSF.IO/H53J4). The file `phased_VCF_filenames.txt` records the names of all phased VCF files used in the analyses.

The variant data from the China 100,000 Genome Project were obtained from a controlled-access resource [39] and are archived in the Genome Sequence Archive for Human (GSA-Human) under accession number HRA004204. Access to these data requires approval through the GSA-Human data-access portal (https://ngdc.cncb.ac.cn/gsa-human/) by the corresponding Data Access Committee (DAC; accession number: HDAC002403). The DAC can be contacted through Wang Yadong (email: `ydwang@hit.edu.cn`). Due to data usage restrictions, the authors are unable to redistribute the original data. All analyses involving these data were performed under authorized access, and the raw data are not publicly available.

In addition, some ancestry informative marker (AIM) loci were derived from publicly available materials released by the Shenzhen municipal government (https://www.sz.gov.cn/attachment/0/753/753270/8308806.pdf). We used all loci listed in Section A.2 of this document, comprising a total of 56 loci.

The reference genome sequences used for the cross-clade analysis were obtained through the NCBI Datasets interface [40]. The accession numbers, assembly versions, and corresponding species list of all reference genomes used in this study have been archived in the OSF repository (https://doi.org/10.17605/OSF.IO/H53J4) as the file `gcf_accessions.zip`.

## Declarations

### List of abbreviations

**AIM**, ancestry-informative marker; **ALT**, alternative allele; **AUPRC**, area under the precision–recall curve; **AUROC**, area under the receiver operating characteristic curve; **MCC**, Matthews correlation coefficient; **REF**, reference allele; **SNP**, single-nucleotide polymorphism.

### Ethical approval

This study used publicly available reference data and controlled-access genomic data. No new human or animal samples were collected. Controlled-access data were analyzed in accordance with the applicable data-use requirements.

### Funding

The project was supported by the National Key Research and Development Program of China (Grant No. 2022YFC3321103).

### Competing interests

The authors declare no competing interests.

## References


1. Lin A, Xie B, Ye C, Wang C, Chen D, Wang E, et al. Genos: a human-centric genomic foundation model. GigaScience 2025 Jan;14:giaf132. https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giaf132/8296738.
2. Consens ME, Dufault C, Wainberg M, Forster D, Karimzadeh M, Goodarzi H, et al. Transformers and genome language models. Nature Machine Intelligence 2025 Mar;7(3):346–362. https://www.nature.com/articles/s42256-025-01007-9.
3. Feng H, Wu L, Zhao B, Huff C, Zhang J, Wu J, et al. Benchmarking DNA foundation models for genomic and genetic tasks. Nature Communications 2025 Nov;16(1):10780. https://www.nature.com/articles/s41467-025-65823-8.
4. Veiner M, Supek F. The DNA dialect: a comprehensive guide to pretrained genomic language models. Molecular Systems Biology 2026 Jan;22(3):309–332. https://link.springer.com/10.1038/s44320-025-00184-4.
5. Zhou Z, Ji Y, Li W, Dutta P, Davuluri R, Liu H, DNABERT-2: Efficient Foundation Model and Benchmark For Multi-Species Genome. arXiv; 2023. https://arxiv.org/abs/2306.15006.
6. Boshar S, Evans B, Tang Z, Adel Y, Bornachot Y, Sidbon S, et al., A multispecies foundation model for genome function prediction, gene annotation and sequence design; 2025. http://biorxiv.org/lookup/doi/10.64898/2025.12.22.695963.
7. Nguyen E, Poli M, Faizi M, Thomas A, Wornow M, Birch-Sykes C, et al. HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution. In: Advances in Neural Information Processing Systems 36 New Orleans, Louisiana, USA: Neural Information Processing Systems Foundation, Inc. (NeurIPS); 2023. p. 43177–43201. http://www.proceedings.com/075280-1872.html.
8. Schiff Y, Kao CH, Gokaslan A, et al. Caduceus: bi-directional equivariant long-range DNA sequence modeling. In: Proceedings of the 41st International Conference on Machine Learning, vol. 235 of Proceedings of Machine Learning Research PMLR; 2024. p. 43632–43648.
9. Sanabria M, Hirsch J, Joubert PM, Poetsch AR. DNA language model GROVER learns sequence context in the human genome. Nature Machine Intelligence 2024 Jul;6(8):911–923. https://www.nature.com/articles/s42256-024-00872-0.
10. Brixi G, et al. Genome modeling and design across all domains of life with Evo 2. Nature 2025;.
11. Khan SA, Martínez-de Morentin X, Alsabbagh AR, Maillo A, Lagani V, Gomez-Cabrero D, et al. Multimodal foundation transformer models for multiscale genomics. Nature Methods 2026 Feb;23(2):299–311. https://www.nature.com/articles/s41592-025-02918-6.
12. De Almeida BP, Dalla-Torre H, Richard G, Blum C, Hexemer L, Gélard M, et al. Annotating the genome at single-nucleotide resolution with DNA foundation models. Nature Methods 2025 Nov;22(11):2301–2315. https://www.nature.com/articles/s41592-025-02881-2.
13. Vishniakov K, Viswanathan K, Medvedev A, Kanithi P, Pimentel M, Rajan R, et al. Tokenization to Transfer: Do Genomic Foundation Models Learn Good Representations? In: Vondrick C, Hariharan B, Raffel C, Pinto L, Yang D, Faust A, editors. International Conference on Learning Representations, vol. 2026; 2026. p. 25386–25406. https://proceedings.iclr.cc/paper_files/paper/2026/file/2b09bb02b90584e2be94ff3ae09289bc-Paper-Conference.pdf.
14. Wu W, Song X, Wen Y, Lin Q, Zhou Z, Hu JYC, et al. Genome-Factory: A Library for Tuning, Deploying, and Interpreting Genomic Foundation Models. In: Zhang T, Dudik M, Jaggi M, Agarwal A, Li S, Schuurmans D, et al., editors. Proceedings of the 43rd International Conference on Machine Learning, vol. 306 of Proceedings of Machine Learning Research PMLR; 2026. p. 136774–136796. https://proceedings.mlr.press/v306/wu26aw.html.
15. Benegas G, Ye C, Albors C, Li JC, Song YS. Genomic language models: opportunities and challenges. Trends in Genetics 2025 Apr;41(4):286–302. https://linkinghub.elsevier.com/retrieve/pii/S0168952524002956.
16. Baghbanzadeh M, Mann BT, Crandall KA, Rahnavard A. seqLens: Optimizing Language Models for Genomic Predictions. Molecular Biology and Evolution 2026 Jul;43(7):msag139. https://academic.oup.com/mbe/article/doi/10.1093/molbev/msag139/8715817.
17. Tang Z, Somia N, Yu Y, Koo PK. Evaluating the representational power of pre-trained DNA language models for regulatory genomics. Genome Biology 2025 Jul;26(1):203. https://genomebiology.biomedcentral.com/articles/10.1186/s13059-025-03674-8.
18. Xu L, He W, Tai S, Huang X, Qin M, Liao X, et al. VCF2Dis: an ultra-fast and efficient tool to calculate pairwise genetic distance and construct population phylogeny from VCF files. GigaScience 2025 Jan;14:giaf032. https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giaf032/8106438.
19. Shokrof M, Abuelanin M, Brown CT, Mansour TA. The Great Genotyper: a graph-based method for population genotyping of small and structural variants. GigaScience 2025 Jan;14:giaf112. https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giaf112/8271468.
20. Chen X, Wang H, Broce I, Dale A, Yu B, Zhou LY, et al. Old vs. new local ancestry inference in HCHS/SOL: a comparative study. Human Molecular Genetics 2025 Aug;34(16):1405–1418. https://academic.oup.com/hmg/article/34/16/1405/8158506.
21. Moorjani P, Hellenthal G. Methods for Assessing Population Relationships and History Using Genomic Data. Annual Review of Genomics and Human Genetics 2023 Aug;24(1):305–332. https://www.annualreviews.org/doi/10.1146/annurev-genom-111422-025117.
22. Bağcı C, Negri T, Buena-Atienza E, Gross C, Ossowski S, Ziemert N. Ultra-deep long-read metagenomics captures diverse taxonomic and biosynthetic potential of soil microbes. GigaScience 2025 Jan;14:giaf135. https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giaf135/8300805.
23. Oskolkov N, Jin C, Clinton SL, Guinet B, Wijnands F, Johnson E, et al. Improving taxonomic inference from ancient environmental metagenomes by masking microbial-like regions in reference genomes. GigaScience 2025 Jan;14:giaf108. https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giaf108/8271467.
24. Silva JM, Martins I, Almeida JR. HYMET: a hybrid metagenomic pipeline for accurate and efficient taxonomic classification. GigaScience 2026 Jan;15:giag024. https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giag024/8503491.
25. Rahman MA, Rangwala H. IDMIL: an alignment-free Interpretable Deep Multiple Instance Learning (MIL) for predicting disease from whole-metagenomic data. Bioinformatics 2020 Jul;36(Supplement_1):i39–i47. https://academic.oup.com/bioinformatics/article/36/Supplement_1/i39/5870478.
26. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biology 2019 Nov;20(1):257. https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1891-0.
27. Flagel L, Brandvain Y, Schrider DR. The Unreasonable Effectiveness of Convolutional Neural Networks in Population Genetic Inference. Molecular Biology and Evolution 2019 Feb;36(2):220–238. https://academic.oup.com/mbe/article/36/2/220/5229930.
28. Dominguez Mantes A, Mas Montserrat D, Bustamante CD, et al. Neural ADMIXTURE for rapid genomic clustering. Nature Computational Science 2023;3:621–629.

29. Elhaik E, Behnamian S, Howe M, et al. AncestryGeni: a novel genetic ancestry classification pipeline for small and noisy sequence data. Bioinformatics 2025;41:btaf391.
30. Oriol Sabat B, Mas Montserrat D, Giro-i Nieto X, Ioannidis AG. SALAI-Net: species-agnostic local ancestry inference network. Bioinformatics 2022;38:ii27–ii33.
31. Warren RL, Coombe L, Wong J, et al. ntRoot: computational inference of human ancestry at scale from genomic data. Bioinformatics Advances 2024;5:vbaf287.
32. Zhou Z, Wu W, Ho H, et al. DNABERT-S: pioneering species differentiation with species-aware DNA embeddings. Bioinformatics 2025;41:i255–i264.
33. Mock F, Kretschmer F, Kriese A, et al. Taxonomic classification of DNA sequences beyond sequence similarity using deep neural networks. Proceedings of the National Academy of Sciences USA 2022;119:e2122636119.
34. Salah R, AbdElaal KR, Ghonaim L, et al. DeepTaxa: a hybrid CNN-BERT framework for 16S rRNA taxonomic classification. Bioinformatics Advances 2026;6:vbag166.
35. Millan Arias P, Sadjadi N, Safari M, et al. BarcodeBERT: transformers for biodiversity analyses. Bioinformatics Advances 2026;6:vbag054.
36. Bansal V, Libiger O. Fast individual ancestry inference from DNA sequence data leveraging allele frequencies for multiple populations. BMC Bioinformatics 2015 Dec;16(1):4. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-014-0418-7.
37. The FUSION Study, Wang C, Zhan X, Bragg-Gresham J, Kang HM, Stambolian D, et al. Ancestry estimation and control of population stratification for sequence-based association studies. Nature Genetics 2014 Apr;46(4):409–415. https://www.nature.com/articles/ng.2924.
38. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature 2015;526:68–74.
39. Jiang T, Guo H, Liu Y, et al. A comprehensive genetic variant reference for the Chinese population. Science Bulletin 2024;69(24):3820–3825.
40. O'Leary NA, et al. NCBI Datasets: a modern approach to biological data access. Nucleic Acids Research 2023;51:D36–D43.