SEARCHING FOR PHENOTYPIC NEEDLES IN GENOMIC HAYSTACKS: DNA LANGUAGE MODELS FOR SEX PREDICTION

Alla Chepurova[1] **Yuri Kuratov**[1] **Polina Belokopytova**[2] **Mikhail Burtsev**[3] **Veniamin Fishman**[1,2]
[1]AIRI, Moscow, Russia
[2]Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
[3]London Institute for Mathematical Sciences, London, UK
{Chepurova, Kuratov}@airi.net, belokopytova@bionet.nsc.ru
mb@lims.ac.uk, Fishman@airi.net

ABSTRACT

In this study, we explore fine-tuning of Genomic Language Models (GLM) to predict phenotypic traits directly from genomic sequence, without prior knowledge about causative loci or molecular mechanisms linking genotype to phenotype. As a case study, we focus on sex prediction, a well-defined genomic feature associated with the presence of the Y chromosome in most mammals. We adapt a pre-trained GENA-LM model for trait prediction by introducing a sequence chunk classification component with cross-attention, enabling the model to process larger genomic contexts. Training and evaluation on human and mouse genomes demonstrate that the model does not require high-quality reference genome assembly and converges even when the fraction of genomic signal associated with phenotype is below 1%. Prediction accuracy improves with increased sequencing depth, highlighting the scalability of GLMs for genome-wide tasks. Ablation studies demonstrate that the model relies on the Y chromosome for sex prediction, that aligns with real biological principles. Our findings highlight the applicability of GLMs for trait prediction in long and fragmented genomic data.

1 INTRODUCTION

Identifying genomic regions that contribute to phenotypic traits, whether qualitative or quantitative, is a fundamental challenge in genomics. Traditional approaches, such as quantitative trait loci (QTL) mapping and genome-wide association studies (GWAS), primarily depend on alignment-based methods and genome annotations to identify individual genomic variants and find associations between variants and phenotypes. While these methods have provided valuable insights (Xu et al., 2024; Korte & Farlow, 2013), their dependence on alignment quality and reference annotations can lead to the omission of critical information, particularly from structurally variable or unaligned genomic regions that may play significant roles in trait expression. Moreover, these approaches are limited in their applicability to species with poorly annotated genomes and are unsuitable for interspecies analyses, as they require a common reference genome. This is a crucial constraint, given that interspecies trait variance is much higher than intraspecies variance, presenting potential for identifying causative loci more efficiently.

Addressing these limitations requires novel methods that move beyond traditional techniques. Genomic Language Models (GLMs), such as DNABERT (Ji et al., 2021; Zhou et al., 2023), Nucleotide Transformer (Dalla-Torre et al., 2023), GENA-LM (Fishman et al., 2025), AIDO.DNA (Ellington et al., 2024), offer a promising alternative by treating genomic sequences as language-like data. GLMs process raw genomic sequences directly, eliminating the need for alignment and enabling trait prediction from fragmented or unaligned genomes. Built on transformer-based architectures, these models capture both local and global sequence dependencies, learning hierarchical, context-aware representations without explicit feature engineering. This approach enhances scalability, interpretability, and the ability to uncover deeper biological patterns through pretraining on large

genomic datasets. GLMs, particularly those from the GENA-LM family, can handle long input sequences (up to 36,000 base pairs), providing rich contextual understanding across extensive genomic regions. Due to their flexibility to be fine-tuned for diverse biological tasks, including multispecies trait prediction, GLMs could be a powerful tool for overcoming current limitations in genome-wide trait analysis.

This study investigates the ability of fine-tuned GLMs to predict phenotypic traits from unaligned genome chunks, without relying on positional information or high-quality reference genomes. As a case study, we examine sex determination — a well-defined genomic trait linked to the Y chromosome in most mammals. While sex determination is simpler than polygenic traits such as height or disease susceptibility, it remains challenging for GLMs. The relevant signal spans an entire chromosome but constitutes only $\sim 1\%$ of the human and mouse genome, making detection in fragmented input non-trivial. This challenge is akin to searching a "needle in a haystack", where the model has to to find sparse, relevant information within long contexts, a known limitation of language models (Ivgi et al., 2023; Liu et al., 2024; Kuratov et al., 2024). To evaluate GLMs' capability in trait prediction, we pose the following research questions:

Trait Detection in Unaligned Genomic Data: Can GLMs learn and predict the presence or absence of genomic sequences associated with sex, a well-defined phenotypic trait, using simulated unaligned and unannotated genome chunks as input?

Scalability of Predictions: How does GLM performance change as the number of input genome chunks increases, simulating varying sequencing depth and data availability?

These research questions provide a structured framework to evaluate the robustness and scalability of GLMs in trait prediction tasks, using sex prediction as a case study and paving the way for broader applications in genomic analysis.

2 Materials and Methods

2.1 Dataset and Sampling

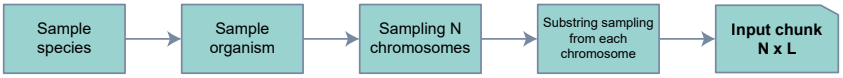


Figure 1: **Procedure simulating uniform sampling from a sequenced genome**. The process begins with the random selection of a species, followed by the sampling of an organism from the dataset. Then, *N* chromosomes are sampled with replacement, and random genomic sequence of length *L* are extracted from the sampled chromosomes to form the input data ($N \times L$ bp).

The dataset for experiments contained genomic data from *Homo sapiens* and *Mus musculus*, with data preparation details provided in Appendix A. Each species were represented by multiple samples: either genomes of different individuals (in case of human data) or different lineages (in case of mice). For both species, available samples were split into training, validation, and test sets.

We implemented a multi-stage data sampling procedure for both training and inference: (1) Species sampling — a species was randomly selected; (2) Sample selection — a sample was randomly chosen from the selected species; (3) Chromosome sampling — N chromosomes were drawn with replacement, weighted by chromosome length; and (4) Sequence sampling — from each sampled chromosome, random substrings of length L base pairs were extracted. Each iteration of such procedure resulted N genomic chunks of total length $N \times L$ base pairs to ensure uniform genome coverage. Figure 1 illustrates the procedure.

2.2 Model Architecture, Training and Inference

Sex prediction was framed as a binary classification task (male = 0, female = 1) using the pre-trained GENA-LM model (`gena-lm-bert-base-t2t`) as the backbone, along with its tokenizer (details in Appendix B). Since the backbone model effectively encodes only ∼4.5 kb of genomic data,

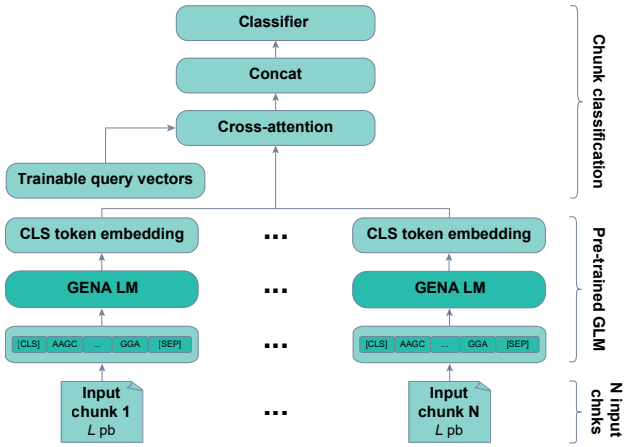


Figure 2: **GENA-LM architecture adaptation for trait prediction tasks**. This architecture allows the model to process up to 49.2 kb of genomic data while preserving critical information across multiple chunks.

which is insufficient for whole-genome tasks where relevant signals are sparse ($\approx 1\%$), we introduced a multichunk classification component. It processes N input chunks, representing each through CLS token embeddings outputted by the backbone model. To handle dependencies across chunks, cross-attention is applied between chunks embeddings and four trainable query vectors, producing four output vectors that are concatenated and passed through a classification layer with a linear transformation and sigmoid activation for binary classification. This architecture extends input capacity to $N \times L$ bp, while preserving relevant genomic information distributed across multiple chunks. Figure 2 illustrates the overall framework.

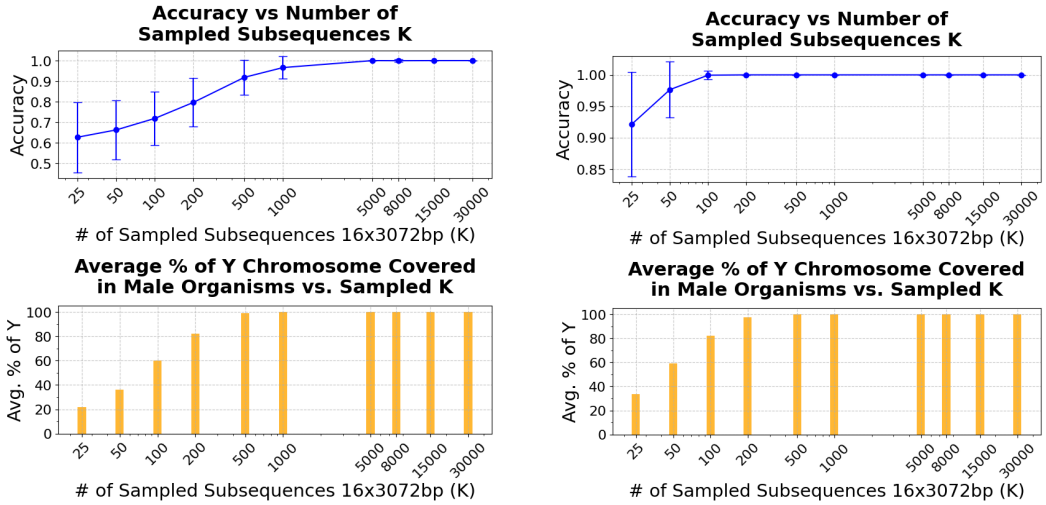
The model was trained with input of $N = 16$ chunks of $L = 3,072$ bp each, with total 49.2 kb per input. Parameters of backbone GENA-LM model were also fine-tuned. Training ran for up to 100,000 steps with a batch size of 128, using 4 NVIDIA A100 GPUs. The AdamW optimizer (Loshchilov & Hutter, 2019) was applied with a constant learning rate of 1×10^{-5}, 1,000-step warm-up, and early stopping after 50 validations without improvement. Model selection was based on the highest ROC-AUC score, with evaluation on every 100 steps. To ensure that the model learned to accurately identify sex-related signals while avoiding overfitting to non-relevant regions, Y chromosome sampling was enforced in male samples during evaluation. In this way, at least one of $N = 16$ sampled chunks was taken from the Y chromosome for male organisms during validation.

Model quality and robustness was tested on varying total number of sampled chunks K (from 50 to 30,000), simulating different levels of sequencing depth and data availability. For each K: (1) a classification threshold was set as the median female class probability ratio (output $P > 0.5$) in validation, (2) based on the calculated threshold, accuracy was evaluated on the test set, and (3) sampling was repeated 200 times to compute mean and standard deviation. This iterative approach assessed performance across diverse sequencing coverage levels, ensuring robustness to real-world variations in sequencing data. Further details are provided in Appendix C, with Figures 4 and 5 illustrating the threshold selection and inference process.

3 Results and Discussion

This study proposed an adaptation of the pre-trained GENA-LM model for trait prediction using a cross-attention mechanism aggregating information from multiple genomic chunks. We evaluated its performance on sex prediction task using human and mouse genomic data.

Scalability of Predictions. Figures 3a– 3b present the evaluation results of the multi-species model, stratified by species - *Homo sapiens* and *Mus musculus* - across different sampling sizes K. As the figures show, the quality of prediction improves with an increasing number of sampled chunks



(a) **Human**, varying number of chunks K. (b) **Mouse**, varying number of chunks K.

Figure 3: The multi-species model accurately identifies sex in both species, requiring different chunk numbers to achieve higher quality due to species-specific Y chromosome proportions in the input.

for both species, demonstrating the model's ability to aggregate genomic signals across multiple chunks. This trend also highlights the importance of sequencing depth in trait prediction and provides insights into the sequencing requirements for achieving desired accuracy. Such scalability demonstrates the potential of GLMs to handle varying data availability scenarios.

Trait Detection in Unaligned Genomic Data. Series of ablation studies (Figures 6a–6b, Appendix D) confirmed that exclusion of the Y chromosome significantly impaired the model's ability to predict gender, indicating that the model correctly identifies and relies on gender-associated regions located on the Y chromosome. In contrast, exclusion of the X chromosome had no significant impact. These findings demonstrate that GLMs can effectively predict the presence of gender-specific genomic signals using unaligned and unannotated genome chunks. The ablation studies underscored the biological relevance of predictions, showing that the Y chromosome is the primary contributor to the model's gender-detection capability, while the X chromosome has little to no impact on predictions.

4 Conclusions and Future Works

Our study highlights the potential of GLMs for trait prediction in genomic analysis. By addressing challenges in unaligned genomic data, these models enable genome-wide tasks without requiring high-quality reference genomes. Their scalability and robustness provide a strong foundation for expanding GLM applications to more complex traits and diverse species. Future research could explore the prediction of more complex genomic traits, such as features of *Canis familiaris*, cataloged in the Dog10K Consortium dataset (Meadows et al., 2023), and extending GLMs to larger multi-species datasets to improve accuracy and generalization across phylogenetic boundaries. Another promising direction is integrating advanced architectures like the Recurrent Memory Transformer (RMT) (Bulatov et al., 2022; 2023), which can aggregate information from extremely long input sequences, making it particularly suitable for genome-wide tasks. The successful use of RMT with GENA-LM for chromatin accessibility prediction (Fishman et al., 2025) suggests its potential for genome-wide trait prediction requiring large-scale input aggregation. Additionally, applying interpretability methods (Sundararajan et al., 2017; Scott et al., 2017) to trait prediction could uncover novel genomic regions associated with specific traits, providing biological insights and expanding the utility of GLMs in functional genomics.

REFERENCES

Aydar Bulatov, Yury Kuratov, and Mikhail Burtsev. Recurrent memory transformer. *Advances in Neural Information Processing Systems*, 35:11079–11091, 2022.

Aydar Bulatov, Yuri Kuratov, Yermek Kapushev, and Mikhail S Burtsev. Scaling transformer to 1m tokens and beyond with rmt. *arXiv preprint arXiv:2304.11062*, 2023.

1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68, 2015.

Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Hassan Sirelkhatim, Guillaume Richard, Marcin Skwark, Karim Beguir, Marie Lopez, and Thomas Pierrot. The Nucleotide Transformer: Building and Evaluating Robust Foundation Models for Human Genomics. *bioRxiv*, pp. 2023.01.11.523679v3, January 2023.

Caleb N. Ellington, Ning Sun, Nicholas Ho, Tianhua Tao, Sazan Mahbub, Dian Li, Yonghao Zhuang, Hongyi Wang, Le Song, and Eric P. Xing. Accurate and general dna representations emerge from genome foundation models at scale. In *NeurIPS 2024 Workshop on AI for New Drug Modalities*. bioRxiv, 2024. doi: 10.1101/2024.12.01.625444. URL https://www.biorxiv.org/content/10.1101/2024.12.01.625444v1.

Veniamin Fishman, Yuri Kuratov, Aleksei Shmelev, Maxim Petrov, Dmitry Penzar, Denis Shepelin, Nikolay Chekanov, Olga Kardymon, and Mikhail Burtsev. Gena-lm: a family of open-source foundational dna language models for long sequences. *Nucleic Acids Research*, 53(2):gkae1310, 2025.

Maor Ivgi, Uri Shaham, and Jonathan Berant. Efficient long-text understanding with short-text models. *Transactions of the Association for Computational Linguistics*, 11:284–299, 03 2023. ISSN 2307-387X. doi: 10.1162/tacl_a_00547. URL https://doi.org/10.1162/tacl_a_00547.

Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics*, 37(15):2112–2120, 02 2021. ISSN 1367-4803.

Konrad J Karczewski, Laurent C Francioli, Grace Tiao, Beryl B Cummings, Jessica Alföldi, Qingbo Wang, Ryan L Collins, Kristen M Laricchia, Andrea Ganna, Daniel P Birnbaum, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581(7809): 434–443, 2020.

Arthur Korte and Ashley Farlow. The advantages and limitations of trait analysis with gwas: a review. *Plant methods*, 9:1–9, 2013.

Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Dmitry Sorokin, Artyom Sorokin, and Mikhail Burtsev. In search of needles in a 10m haystack: Recurrent memory finds what llms miss. *arXiv preprint arXiv:2402.10790*, 2024.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Jennifer RS Meadows, Jeffrey M Kidd, Guo-Dong Wang, Heidi G Parker, Peter Z Schall, Matteo Bianchi, Matthew J Christmas, Katia Bougiouri, Reuben M Buckley, Christophe Hitte, et al. Genome sequencing of 2000 canids by the dog10k consortium advances the understanding of demography, genome function and architecture. *Genome biology*, 24(1):187, 2023.

M Scott, Lee Su-In, et al. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30:4765–4774, 2017.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

Qing Xu, Jianhua Jiang, Chunyu Jing, Changmin Hu, Mengyuan Zhang, Xinru Li, Jiaming Shen, Mei Hai, Ying Zhang, Dezheng Wang, et al. Genome-wide association mapping of quantitative trait loci for chalkiness-related traits in rice (oryza sativa l.). *Frontiers in Genetics*, 15:1423648, 2024.

Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana Davuluri, and Han Liu. DNABERT-2: Efficient Foundation Model and Benchmark For Multi-Species Genome. *arXiv*, June 2023.

A DATA PREPARATION

The dataset for *Homo sapiens* organisms was prepared through the following steps:

- Multivcf files containing phased genomic variants were downloaded from the 1000 Genomes Project database[1];
- For each sample, two DNA sequences in FASTA format were created for each chromosome based on a VCF file with phased variants and hg38 reference. This was accomplished using the bcftools consensus tool with the -H 1pIu or -H 2pIu and -s SampleName options;
- To include SNPs on the Y chromosome, a specific VCF file from the 1000 Genomes Project was used[2].

The dataset for *Mus musculus* was prepared through the following steps:

- Downloading multivcf files with phased genomic variants[3];
- For each mouse line, one sequence per chromosome was generated using only the alternate allele based on mm10 reference. Male mice included the Y chromosome, while non-male samples included only the X chromosome. The bcftools consensus tool was used with the -H A and -s SampleName options, where SampleName corresponds to the specific mouse line.

The resulting datasets for both species were divided into training, validation, and testing subsets. The statistics of these splits are provided in Table 1.

Table 1: The distribution of numbers of organisms in train, validation and test datasets

Species	**Train size**	**Valid size**	**Test size**
Homo sapiens	35	10	24
Mus musculus	16	8	10

B DETAILS OF PRE-TRAINED GLM

The tokenizer used in the experiments was the one implemented in the original GENA-LM paper, based on Byte-Pair Encoding (BPE). It has a dictionary size of 32,000 and an initial character-level vocabulary consisting of ['A', 'T', 'G', 'C', 'N']. This tokenizer was trained on the human T2T v2

[1] http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/20201028_3202_phased/

[2] 20201028_CCDG_14151_B01_GRM_WGS_2020-08-05_chrY.recalibrated_variants.vcf.gz

[3] https://ftp.ebi.ac.uk/pub/databases/mousegenomes/REL-1505-SNPs_Indels/

genome assembly from NCBI (acc. GCFn_009914755.1) and multispecies data from ENSEMBL release 108[4], with equal sampling across species. It also incorporates special tokens: CLS, SEP, PAD, UNK, and MASK. Additionally, for preprocessing, any sequence with more than 10 consecutive 'N' characters is replaced by a single '–' token.

The backbone model, `gena-lm-bert-base-t2t`[5], was pre-trained using a masked language modeling (MLM) objective, where 15% of the tokens were randomly masked. Pre-training was performed on the latest T2T human genome assembly, with data augmentation through mutations sampled from the 1000 Genomes Project (Consortium et al., 2015) and the gnomAD dataset (Karczewski et al., 2020). The pre-training procedure involved 2,100,000 iterations, with a batch size of 256 and a sequence length of 512 tokens.

The model configuration for `gena-lm-bert-base-t2t` used in experiments was as follows:

- Maximum sequence length: 512 tokens
- 12 layers
- 12 attention heads
- Hidden size: 768
- Vocabulary size: 32,000
- Pre-layer normalization

C INFERENCE

To assess the quality and robustness of the model, we simulated scenarios of varying sequencing depths and coverage. Specifically, we sampled different numbers of chunks K (16 x 3,072 bp each) per organism, ranging from 50 to 30,000 chunks. For each sampling size N:

- The classification threshold was calculated on a validation subset as the median ratio of chunks with a predicted probability of the organism being female greater than 0.5, averaged across all organisms in the validation set;
- Accuracy with the calculated threshold was then evaluated on the test dataset;
- Sampling was repeated 200 times to compute the mean and standard deviation for each N.

This iterative approach allowed us to evaluate model performance across a wide range of coverage levels, ensuring robustness to real-world variations in sequencing data. The overall process of threshold choosing is displayed on Figure 4. The process of inference is depicted on Figure 5.

D ABLATION STUDIES

To interpret the model's decision-making process and ensure its predictions are based on relevant genomic regions, we conducted a series of ablation studies. In the first ablation study, we excluded the Y chromosome during inference for the model to assess whether predictions relied on Y-specific genomic features. In a second study, we excluded the X chromosome to determine whether the absence of X chromosome data influenced the model's predictions. The results are presented on Figures 6a– 6b.

[4] `https://ftp.ensembl.org/pub/release-108/`

[5] `https://huggingface.co/AIRI-Institute/gena-lm-bert-base-t2t`

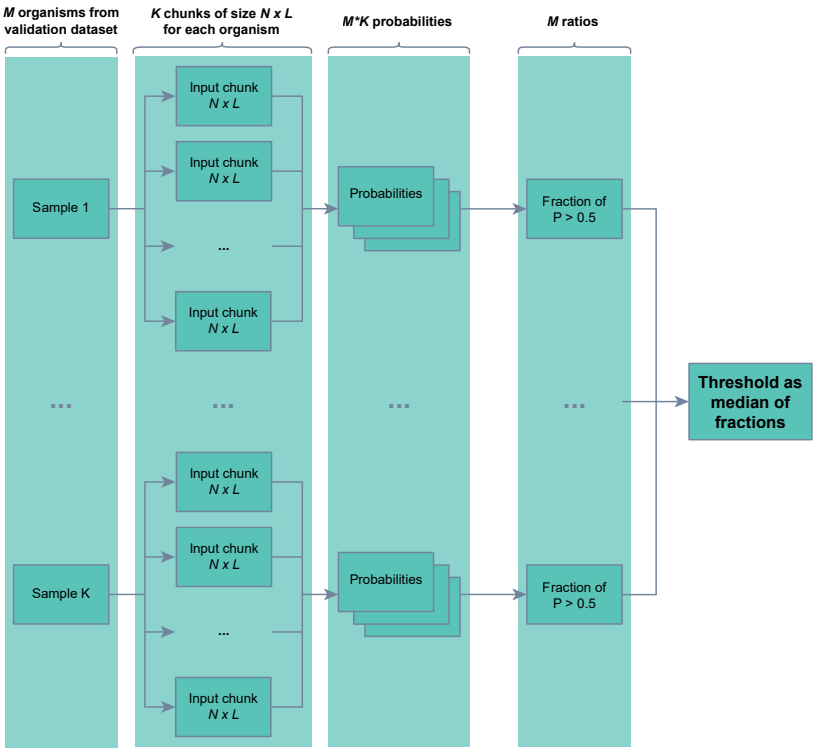


Figure 4: **Determining the threshold using the validation dataset**. For each organism in the validation dataset, K input chunks of size 16 * 3072 were generated. The model outputs probabilities for each chunk, and these probabilities were thresholded at 0.5 to calculate the fraction of chunks classified as positive (i.e., $P > 0.5$) or in other words female for each organism. The final threshold for classification was defined as the median of these ratios across all organisms in the validation dataset.

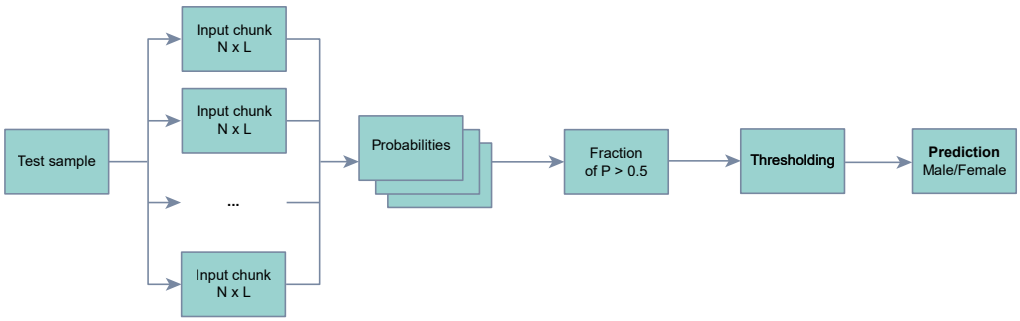
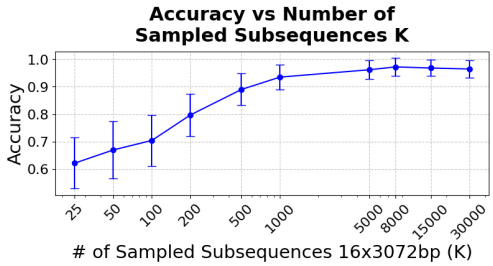
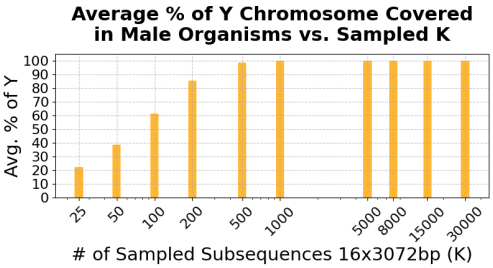
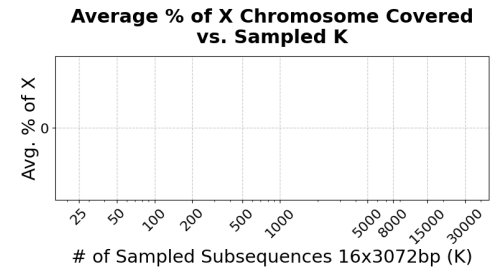


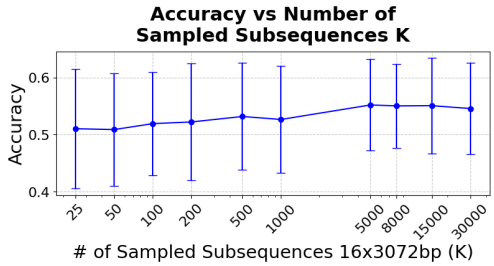
Figure 5: **The inference process for the test dataset**. For each test sample organism, K input chunks of size 16*3072 were generated. The model computed probabilities for each chunk, and these probabilities were thresholded at 0.5 to determine the proportion of chunks classified as female ($P > 0.5$). This proportion was compared to a predefined threshold derived from the validation dataset. The final output label (1 for female, 0 for male) was determined based on this comparison.

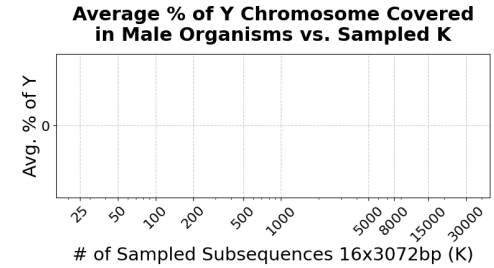


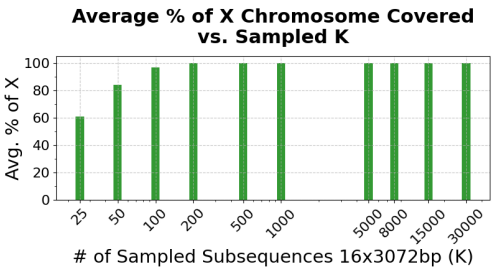




(a) Performance of the model trained on human and mouse genomic data over various numbers of chunks N inferred on *Homo Sapiens* organisms in absence of X chromosome.







(b) Performance of the model trained on human and mouse genomic data over various numbers of chunks N inferred on *Homo Sapiens* organisms in absence of Y chromosome.

Figure 6: Performance of the multi-species model in absence of X or Y chromosome.