A SNP Foundation Model: Application in Whole-Genome Haplotype Phasing and Genotype Imputation

Augix Guohua Xu[2], Yu Xu[4], Yiming Xing[5], Jianbo Yang[6,7*], Yinqi Bai[3,4*], Kun Tang[1*]

1 Research Center For Life Sciences Computing, Zhejiang Lab, Hangzhou, China
2 Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany
3 BGI Research, Hangzhou 310030, China
4 BGI Research, Sanya 572025, China
5 College of Computer Science and Technology, Zhejiang University of Technology
6 Zhejiang Cancer Hospital, Hangzhou Institute of Medicine (HIM), Chinese Academy of Sciences, Hangzhou, 310022, Zhejiang, China
7 Department of Pancreatic Surgery, Xiangya Hospital, Central South University

* corresponding authors: Kun Tang tang.kun@zhejianglab.org
Yinqi Bai baiyinqi@genomics.cn
Jianbo Yang yangjianbo@him.cas.cn

Abstract

Artificial intelligence foundation models are revolutionizing biomedical research. Human genetic diversity data, predominantly in the form of single nucleotide polymorphisms (SNPs), constitutes the largest and fastest growing portions of biobank databases. However, SNP data has been a missing piece in the Bio-AI domain. Here, we present SNPBag, the first SNP foundation model for genome-scale SNP analyses. Built on a large language model (LLM), SNPBag aims to address haplotype phasing and genotype imputation simultaneously. Using data from the 1000 Genome Project, we pre-trained SNPBag on overlapping sequences by masking genotypes. For haplotype phasing task, our model achieved a state-of-the-art (SOTA) 1% switch error rate, outperforming BEAGLE5.2, EAGLE2, and SHAPEIT4. In genotype imputation for the Illumina Omni2.5 SNP array, the model reached an accuracy of 96.88%, comparable to the best available methods. Additionally, when fine-tuned for ancestry inference, the model achieved 97% accuracy in predicting the super-populations of 72 individuals from the 1000 Genome Project. This study demonstrates the great potential of SNPBag in representing and inferring global human genetic diversity. Future research will focus on training SNPBag with large-scale Genome-Wise Association Study (GWAS) data to predict phenotypes from whole-genome genotypes, opening new avenues for understanding human genetics and disease.

Introduction

In recent times, the biomedical research arena has witnessed a paradigm shift due to the advent of artificial intelligence (AI) foundation models, exerting a profound impact across diverse fields, from genomics to proteomics. These models have ushered in a new era of comprehensive omics-scale investigations, revolutionizing our understanding of biological systems. In proteomics, the AlphaFold series has been a game-changer, accurately predicting protein structures and substantially advancing drug discoveries(Jumper *et al.*, 2021). ESM2 and ProteinMPNN have proven to be invaluable in protein-related research and engineering (Dauparas *et al.*, 2022; Lin *et al.*, 2023). In genomics, Nucleotide Transformer (Dalla-Torre *et al.*, 2024) and EVO (Nguyen *et al.*, 2024) process nucleic acid sequences, offering valuable insights into gene essentiality prediction and genomic element analysis. In single cell transcriptomics, Geneformer, scGPT and scBERT have been developed for cell type annotation, gene network inference and perturbation response prediction (Yang *et al.*, 2022; Theodoris *et al.*, 2023; Cui *et al.*, 2024). Meanwhile, UCE has been utilized for cross-species cell alignment and the discovery of new cell types (Rosen *et al.*, 2023).

Human genetic diversity, encompassing all inherited genomic variations, underlies the wide array of human genetic traits, from physical appearance and cognitive abilities to disease predisposition. Since the inception of the Human Genome Project (International Human Genome Sequencing Consortium *et al.*, 2001), a key goal of human genomic research has been to understand the extent and architecture of human genetic diversity. The aim is to predict an individual's trait or disease risks based on their whole-genome variation profile, to the extent determined by the traits' heritability. Single nucleotide polymorphisms (SNPs) are the predominant form of genetic variation and have become the foundation of human genomic research. Over the past two decades, SNP data and related studies have grown exponentially. Large-scale initiatives like the international HapMap project (The International HapMap Consortium, 2003) and the international 1000 Genomes Project (The 1000 Genomes Project Consortium *et al.*, 2015) have been launched. Global biobanks, such as the UK Biobank (Bahcall, 2018), the Million Veteran Program and the All of Us Research Program in the US (The All of Us Research Program Genomics Investigators *et al.*, 2024), China Kadoorie Biobank (Chen *et al.*, 2011), Biobank Japan (Tanaka, Nagata and Takemoto, 2022), deCODE Genetics in Iceland (Chadwick, 1999), the Estonian Biobank (Leitsalu *et al.*, 2015), FinnGen in Finland (Kurki *et al.*, 2023), and Lifelines in the Netherlands (Scholtens *et al.*, 2015), have been established. These biobanks collect large cohorts and survey whole-genome variation at the population scale. Genome-wide association studies (GWAS) have been crucial in deciphering genotype-phenotype relationships. Over the past two decades, GWAS have increased exponentially. Currently, around 7000 publications have reported on approximately 110,000 GWAS studies of about 15,000 different traits and disease liabilities, examining SNPs across millions of individual genomes (Sollis *et al.*, 2023). Given the scale and significance of SNP data in biomedical research, a pressing question is whether AI foundation models can be developed for genome-wide SNP data, similar to

those in proteomics and transcriptomics, to handle essential analysis tasks. In the SNP data analysis pipeline, there are multiple challenges, including SNP calling, genotype imputation, haplotype phasing, population stratification, ancestry inference, and, most importantly, developing prediction models for inferring an individual's genetic predisposition to specific diseases or phenotypes. Traditionally, these tasks have been addressed separately, each with its own computational methods. However, an attention-based foundation model trained on extensive human SNP data could capture marker dependencies across millions of genomes, allowing specific computational tasks to be accomplished via fine-tuning or transfer learning.

Hereby we present SNPBag, the first foundation model specifically designed for genome-scale SNP data. In this study, we focus on solving two critical tasks: haplotype phasing and genotype imputation. Haplotype phasing separates the maternal and paternal copies of allele arrays, while genotype imputation predicts unobserved genotypes at SNP sites. Both tasks are crucial for enhancing the statistical power of GWAS. Existing methods for haplotype phasing and genotype imputation, such as Eagle (Loh *et al.*, 2016), Beagle (Browning *et al.*, 2021), Shapeit (Delaneau *et al.*, 2019), Impute5 (Rubinacci, Delaneau and Marchini, 2020), and Minimac (Howie *et al.*, 2012) are based on the hidden Markov model (HMM) and rely on linkage disequilibrium (LD) between SNPs. In recent years, neural network approaches, including autoencoders (Chen and Shi, 2019; Dias *et al.*, 2022; Song *et al.*, 2022), convolutional neural networks (Naito *et al.*, 2021), and recurrent neural networks (Kojima *et al.*, 2020), have been applied to genotype imputation. Mowlaei et al. proposed STI, a Transformer-based imputation method, which showed high accuracy in yeast genome and human structural variations (Mowlaei *et al.*, 2023). However, there remains a lack of neural network-based methods for haplotype phasing.

SNPBag is based on the Longformer architecture (Beltagy, Peters and Cohan, 2020) and uses genotypes as input. We pre-trained it on overlapping sequences by masking a proportion of genotypes. Haplotype phasing and genotype imputation from SNP arrays are implemented as fine-tuning tasks. We also show that individual genomes can be accurately assigned to their ancestry by fine-tuning a classifier. In the future, we plan to train SNPBag on large-scale GWAS data to evaluate its utility in predicting phenotypes based on whole-genome genotypes. We welcome collaborations with researchers who can provide data and computing resources to further explore the potential of SNPBag in advancing our understanding of human genetics and disease.

Results

Data Preparation

All analyses utilized data from 1000 Genome Project Phase 3, which encompasses 3,202 individuals from 26 diverse populations. This dataset was chosen for its comprehensive representation of global genetic variation. The 3,202 individuals were randomly assigned

to a training group (97%) and a validation group (3%), maintaining a balanced representation of all 26 populations within both groups. In the validation group, children from trio families were excluded, leaving 72 individuals. Our model was trained on the training group and its predictive accuracy was evaluated on the validation group (see Methods: Data Partitioning).

The 1000 Genome Project contains approximately 85 million SNP sites. To reduce expenses, we applied a filter of minor allele frequency (MAF) ≥ 3%, led to ~10 million SNP sites across the human genome (see Methods: Data Pre-processing). Following this, the remaining SNP sites were grouped into 2,000 tilling windows. Each window contained 10,240 consecutive genetic loci spanning approximately 3 million base pairs of a chromosomal region. In each window, genotype and haplotype sequences of length 10,240 were prepared for all individuals. Genotype sequences comprised values of 0s, 1s, and 2s, indicting the number of derived alleles. Haplotype sequences comprised values of 0s and 1s, indicting ancestral and derived alleles (see Methods: Haplotype and Genotype Sequences).

Pre-training for Foundation Model

The abundance of genotype sequences reveals insights into the relationships between SNP sites, such as linkage disequilibrium. These relationships can be learned during model pre-training to create a SNP foundation model. Through fine-tuning, the knowledge about these relationships can be transferred to various downstream tasks including haplotype phasing, genotype imputation, and ancestry inference.

Our pre-training utilized a modified Bidirectional Encoder Representations from Transformers (BERT) architecture, incorporating the Longformer attention to efficiently process long sequences. In each training step, 15% genotype values were randomly masked in each genotype sequence. The model was trained to reconstruct these masked values (Figure 1). The entire training process was executed in parallel for each of the 23 human chromosomes (chr1-22 and chrX). The progress of the training was tracked using several metrics, including the overall prediction accuracy (see Methods: Pre-training for Foundation Model). The learning curve showed a steady increase in prediction accuracy over the training epochs. The trends were consistent between training and validation, suggesting the generalizability of the model (Supplementary Figure 1). Ultimately, the pre-training reached 97% accuracy across all chromosomes in unseen individuals (validation group).

In the context of SNP data processing, having a pre-trained SNP foundation model that can be directly applied across diverse scenarios of missing genotype problems is highly desirable. These scenarios include low-coverage genome sequencing, which can lead to varying degrees of genotype dropout, as well as the use of various mass-produced or customer-designed SNP chips that included only a subset of the entire SNP spectrum (Rubinacci *et al.*, 2021; Marino *et al.*, 2022). Our study demonstrated that, despite being trained at a fixed masking ratio of 15%, the pre-trained model is capable of genotype

imputation across a broad spectrum of masking ratios. Specifically, when 15%, 30%, 50%, and 75% of SNPs were randomly masked, the imputation accuracy reached 98.2%, 97.9%, 97.4%, and 95.2% respectively. These results underscore the model's potential; however, the accuracy declines notably when a large proportion of genotypes is missing. Moving forward, exploring alternative pre-training strategies, such as higher or variable masking ratios, may help improve pre-training accuracies tailored to different experimental scenarios.

Fine-tuning for Haplotype Phasing

Haplotype phasing aims to determine the specific arrangements of alleles across multiple loci (such as SNP sites) on each of the two chromosomes inherited from maternal and paternal lines. Notably, human populations exhibit a finite set of common allelic arrangements on chromosomes, known as haplotype blocks, which significantly shape the observed genotype sequences. Consequently, the knowledge learned in the SNP foundation model can be fine-tuned in haplotype phasing.

To predict haplotype sequences from genotypes, we firstly constructed the switch sequences. These switches represent transitions in allelic state (ancestral or derived) between consecutive heterozygous sites. In fine-tuning, we took the encoder from the foundation model to encode input genotype sequences and trained the decoder to predict the target switch sequences. Each prediction of switch sequence corresponds to a unique solution of haplotype phasing (see Methods: Fine-tuning for Haplotype Phasing).

The benchmark result on chromosome 22 showed that our fine-tuned model achieved the state-of-the-art (SOTA) performance in haplotype phasing. On average, 28,014 heterozygous sites were phased per individual in the validation dataset. The mean switch error rate is only 1.1% in (Figure 3A), outperforming leading conventional approaches including BEAGLE5.2 (Browning *et al.*, 2021)(1.3%), EAGLE2 (Loh *et al.*, 2016) (1.4%) and SHAPEIT4 (Delaneau *et al.*, 2019) (1.4%), all of which utilize the 1000 Genome Project as reference panel. Notably, when these conventional methods operate without a reference panel, their phasing error rate increase substantially to 2.6% for BEAGLE5.2, 3.0% for EAGLE2, and 2.6% for SHAPEIT4. In terms of the total computational time per individual for chromosome 22, our model was comparable to other approaches (Figure 3B). In this comparison, SNPbag uses one GPU (Nvidia RTX 3090) and the other methods uses one CPU.

Fine-tuning for Genotype Imputation

Low-density SNP arrays are cost-effective solutions for analyzing genetic variants. It is important to develop genotype imputation method on array data. However, direct use of our foundation model on input sequences with large fractions of missing genotypes result in substantial decreases in prediction accuracy (See Figure 2).

To address this, we fine-tuned our foundation model with training data tailored to specific SNP array designs. For demonstration, we focused on the Illumina Omni2.5 SNP array and chromosome 22. Among the 153,600 SNP sites on chromosome 22 (MAF ≥ 3%) from 1000 Genome Project, only 23,416 sites (15%) are included in the Omni2.5 array. We prepared the fine-tune datasets by masking the sites missing from the design of the array, and then trained the model to predict these sites. Upon evaluation, the model effectively inferred the missing ~ 85% of SNP sites (approximately 130,000), reached performance comparable to the best available imputation methods (Figure 4A). Most current imputation methods require a high-quality phasing step. Accordingly, we tested all combinations of leading phasing (SHAPEIT4, BEAGLE5.2 and EAGLE2) and imputation methods (IMPUTE5, MINIMAC4 and BEAGLE5.2) (Rubinacci *et al.*, 2021; Howie *et al.*, 2012; Browning *et al.*, 2021). In this benchmark, our method achieved accuracy of 96.88%, outperforming all methods based on SHAPEIT phasing (96.68%, 96.56% and 96.49% respectively; Figure 4A). The best performance (97.15% accuracy) was obtained by the combination of BEAGLE5.2 phasing and BEAGLE5.2 imputation, which was 0.27% higher than our method. However, the performance of our method can be further improved with more computational resources. In term of computational time cost, our method was the second fastest (Figure 4B).

Fine-tuning for Ancestry Inference

Once trained on global genome-scale SNP data, the model effectively "memorizes" the underlying haplotype architectures related to each genomic region across intercontinental groups. Leveraging this, various inference tasks can be conducted, such as inferring ancestry, analyzing population structure, and performing ancestry painting. Here, we demonstrate the principle using simple ancestry inference. All individuals in the 1000 Genome Project belong to 5 super-populations: East Asian (EAS), South Asian (AAS), African (AFR), European (EUR), and American (AMR). We fine-tuned the model to predict these super-population labels for each individual and tested its classification performance on a holdout test panel. We fine-tuned the model on 30 overlapping genotype segments from chromosome 22 (see Methods: Fine-tuning for Ancestry Inference). The model predicted the ancestry for each of the 30 segments, generating a fractional ancestry composition for each individual (Figure 5, left panel). When averaging the ancestry composition across the whole chromosome, the predicted ancestry aligned well with the true labels, achieving an overall accuracy of 97% in the test set of 72 individuals. Notably, SAS and AMR have more mixed ancestry compositions, consistent with their admixed nature. In particular, AMR individuals have a high degree of admixture from American, European, and African ancestries. Interestingly, the 2 AMR individuals misclassified as European are from Puerto Rico, likely due to their substantial European ancestry (Figure 5).

Discussion

In this study, we introduced SNPBag, the first transformer-based SNP foundation model for genome-scale analysis. Pre-trained on ~10 million SNPs (MAF ≥ 3%) across 26 global populations, it can perform multiple downstream tasks such as haplotype phasing, genotype imputation, and ancestry inference. Despite being the first attempt to apply the attention mechanism and pretraining-finetuning framework to phasing and imputation, SNPBag achieved top-tier performance, on par with the best conventional methods refined over decades. However, SNPBag can be further improved. For instance, we could explore variable masking ratios during pre-training, increase model dimensions, or expand the attention windows in the Longformer architecture to enhance performance, especially in genotype imputation.

In the long term, foundation model-based phasing/imputation methods offer several advantages over conventional approaches. First, conventional methods rely on a reference panel of dense SNPs with known phases, and their performance depends on the panel's quality and size (Marino *et al.*, 2022). In contrast, SNP foundation models don't need a reference panel during computation. On the contrast, like large language models using vast training data, SNP foundation models can leverage nearly unlimited pre-training data. After training on millions of whole-genome sequencing datasets, foundation models are likely to outperform all HMM-based methods, as they can better capture complex genetic patterns without the constraints of a fixed reference panel. Second, conventional methods assume homogeneous ancestry within the sample panel and the reference panel, requiring ancestry annotation and alignment beforehand (Marino *et al.*, 2022). SNPBag, trained on a diverse global population, can be applied to any sample without such preconditions. Additionally, the inference of foundation models can be easily scaled up, with individual inferences independent of others, unlike in conventional HMM-based methods. This allows for enhanced efficiency in processing large-scale data once the pipeline is further optimized.

Furthermore, SNPBag is not merely a phasing/imputation tool. As a foundation model, it can be pre-trained on large-scale data and then fine-tuned for various downstream inferences. We postulate that SNP foundation models like SNPBag can be trained on biobank-scale datasets such as UKB, CHB, and other large cohorts. Through one-time training on hundreds of thousands of genomes, we can gain a highly detailed understanding of global genomic diversity. Subsequently, a complete downstream pipeline of SNP analyses can be implemented via fine-tuning or transfer learning with minimal additional computational costs. These analyses may include SNP calling (Poplin *et al.*, 2018), phasing/imputation, ancestry painting (Chen *et al.*, 2024), population structure analysis (Pritchard, Stephens and Donnelly, 2000), genomic GPS (Elhaik *et al.*, 2014), and distant relatedness calculation (Li *et al.*, 2014), etc. In practice, such a model can be easily deployed in fields like medication, personal genomics, and forensics, offering all-in-one AI-based solutions.

Last, the ultimate goal of large-scale SNP mapping is to predict and interpret heritable phenotypes from genotype data. Nearly 20 years of GWAS studies have identified a large number of loci linked to complex traits or disease risks (Sollis *et al.*, 2023). However, each associated locus contributes only a small part to the phenotypic effect (Loos, 2020). To achieve a comprehensive prediction using genome-wide SNPs, various multivariate analysis approaches have been proposed, such as polygenic scores (PGS), polygenic risk scores (PRS), and the GTCA model (Abdellaoui *et al.*, 2023; Yang *et al.*, 2023). Yet, these multivariate predictors also have not been considered highly satisfactory, mainly because they often account for low heritability, resulting in limited prediction power (Abdellaoui *et al.*, 2023; Yang *et al.*, 2023). A crucial factor often overlooked is the high-level epistatic interactions among genetic loci. Given the highly interactive nature of biological systems, these interactions are to be expected. A foundation SNP model offers a solid basis for studying the inter-loci nonlinear interactions on a genome-wide scale. Additionally, a phenotype predictor based on a foundation model is straightforward and can account for complex genome-environment interactions. In conclusion, genotype-phenotype prediction using foundation models shows promise in effectively utilizing large-scale GWAS data.

Methods

Data Partitioning

We used the 1000 Genome Project Phase 3 data for all the analyses. This dataset includes 3,202 individuals from 26 populations. We divided these individuals into training and validation group. Specifically, we randomly selected 3% of individuals from each population (96 individuals) to form the validation group. The remaining 97% of individuals formed the training group. Furthermore, individuals from trio families were excluded, leaving 72 individuals in the validation group. These groups were referred to as the training and validation datasets, respectively.

Data Pre-processing

To make the raw data compatible with deep learning models, we converted it into haplotype matrices. First, we downloaded the phased SNP data in VCF format from the 1000 Genomes Project. This whole-genome dataset contains approximately 85 million SNP sites. To reduce computational demands, we excluded SNP sites with a minor allele frequency (MAF) below 3%. This filtering reduced the number of SNP sites to approximately 10 million. Using these SNPs, haplotype matrices were created for each chromosome (chr1–22 and chrX). Each matrix's rows represent individual chromosomes, and columns represent SNP sites, with values of 0s and 1s indicating ancestral and derived alleles, respectively. These matrices were then used to generate genotype and haplotype sequences for model training.

Haplotype and Genotype Sequences

The 10 million SNP sites (MAF ≥ 3%) were divided into 2,000 overlapping windows, each containing 10,240 consecutive SNP sites. Each window corresponded to a chromosomal region of roughly 3 million base pairs, with adjacent windows overlapping by 5,120 sites. These windows covered almost all chromosomes except Y chromosome. A haplotype sequence represents an array of alleles (0s for ancestral alleles, 1s for derived alleles) for a given chromosome within a given window. A genotype sequence represents the genotype values (0, 1, or 2) of an individual within a single window, where the values indicate the number of derived alleles. The haplotype and genotype sequences served as input and target sequences for model training (described below).

Pre-training for Foundation Model

To learn the dependencies among SNP sites, we employed a BERT-based model. Genotype sequences were masked following standard BERT training protocols. For each genotype sequence, 15% of values were replaced with a mask token (3), a random genotype value (0, 1, or 2), or the original value. These masked sequences were paired with their complete SNP ID sequences to form input sequences, each of length 10,240. With the input sequences, the model predicted the masked genotype values. A cross-entropy loss function was used to compare predictions with original genotype values as targets.

The architecture of the model comprised 8 layers of Longformer encoders, a variant of transformer encoder with 512 dimensions, 8 attention heads and an attention window size of 64. A multilayer perceptron (MLP) served as the decoder. Pre-training was monitored using metrics including overall genotype prediction accuracy and sensitivity in detecting heterozygous sites. To reduce the cost, training was manually halted when most chromosomes reached 95% sensitivity. Further improvements were feasible with additional computational resources.

Fine-tuning for Haplotype Phasing

To assess the model's utility in haplotype phasing, we fine-tuned the pre-trained model with a phasing dataset. In the phasing dataset, we constructed switch sequences as follows. Each genotype sequence is the sum of a paternal and maternal haplotype sequence. From 1000 Genome Project phased data, we selected the haplotype sequence with a derived allele at the 1st heterozygous site. On this haplotype sequence, we then infer, for each heterozygous site, if there is a switch of state from previous to current heterozygous site (ancestral to derived allele or derived to ancestral allele). The model needs to take genotype sequences as input and predict the switches for these heterozygous sites. The phasing task doesn't need to make predictions about the homozygous sites. The switch sequences predicted from the model can be translated into haplotype sequences directly. Children from the trios of the 1000 Genome Project were excluded from this analysis.

Fine-tuning for Genotype Imputation

We focused on the Illumina Omni2.5 SNP array and chromosome 22. Among the 153,600 SNP sites on chromosome 22 (MAF ≥ 3%) from 1000 genome project, only 23,416 sites (15%) are included in the Omni2.5 array. During genotype imputation we aimed to infer the remaining 85% sites. The fine-tune dataset is still from the 1000 Genome Project. It is almost the same as the pre-training dataset except the masking was done on sites missing from the design of Omni2.5 array. The split of training and validation groups kept the same.

Fine-tuning for Ancestry Inference

The 1000 Genome Project's individuals are categorized into five super-populations: East Asian (EAS), South Asian (SAS), African (AFR), European (EUR), and American (AMR). We fine-tuned the foundation model's decoder to predict these super-populations. Each individual's 30 overlapping genotype sequences on chromosome 22 were taken as model input, and predictions were made for each of these sequences. The model was trained on a training set and validated using 72 individuals in validation set, excluding children from trio families. Ancestry fractions were calculated for each individual's sequences, and the super-population with the highest predicted fraction was assigned, which was then compared to the 1000 Genome Project's annotations.

Acknowledgement

This study was supported by High-performance Computing Platform of YZBSTCACC（YaZhou Bay Science and Technology City Advanced Computing Center）

References

Abdellaoui, A. *et al.* (2023) '15 years of GWAS discovery: Realizing the promise', *The American Journal of Human Genetics*, 110(2), pp. 179–194. Available at: https://doi.org/10.1016/j.ajhg.2022.12.011.

Bahcall, O.G. (2018) 'UK Biobank — a new era in genomic medicine', *Nature Reviews Genetics*, 19(12), pp. 737–737. Available at: https://doi.org/10.1038/s41576-018-0065-3.

Beltagy, I., Peters, M.E. and Cohan, A. (2020) 'Longformer: The Long-Document Transformer'. arXiv. Available at: https://doi.org/10.48550/arXiv.2004.05150.

Browning, B.L. *et al.* (2021) 'Fast two-stage phasing of large-scale sequence data', *American Journal of Human Genetics*, 108(10), pp. 1880–1890. Available at: https://doi.org/10.1016/j.ajhg.2021.08.005.

Chadwick, R. (1999) 'The Icelandic database---do modern times need modern sagas?', *BMJ*, 319(7207), pp. 441–444. Available at: https://doi.org/10.1136/bmj.319.7207.441.

Chen, J. and Shi, X. (2019) 'Sparse Convolutional Denoising Autoencoders for Genotype Imputation', *Genes*, 10(9), p. 652. Available at: https://doi.org/10.3390/genes10090652.

Chen, S. *et al.* (2024) 'AncestryPainter 2.0: Visualizing Ancestry Composition and Admixture History Graph', *Genome Biology and Evolution*, 16(11), p. evae249. Available at: https://doi.org/10.1093/gbe/evae249.

Chen, Z. *et al.* (2011) 'China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up', *International Journal of Epidemiology*, 40(6), pp. 1652–1666. Available at: https://doi.org/10.1093/ije/dyr120.

Cui, H. *et al.* (2024) 'scGPT: toward building a foundation model for single-cell multi-omics using generative AI', *Nature Methods*, 21(8), pp. 1470–1480. Available at: https://doi.org/10.1038/s41592-024-02201-0.

Dalla-Torre, H. *et al.* (2024) 'Nucleotide Transformer: building and evaluating robust foundation models for human genomics', *Nature Methods* [Preprint]. Available at: https://doi.org/10.1038/s41592-024-02523-z.

Dauparas, J. *et al.* (2022) 'Robust deep learning–based protein sequence design using ProteinMPNN', *Science*, 378(6615), pp. 49–56. Available at: https://doi.org/10.1126/science.add2187.

Delaneau, O. *et al.* (2019) 'Accurate, scalable and integrative haplotype estimation', *Nature Communications*, 10(1), p. 5436. Available at: https://doi.org/10.1038/s41467-019-13225-y.

Dias, R. *et al.* (2022) 'Rapid, Reference-Free human genotype imputation with denoising autoencoders', *eLife*, 11, p. e75600. Available at: https://doi.org/10.7554/eLife.75600.

Elhaik, E. *et al.* (2014) 'Geographic population structure analysis of worldwide human populations infers their biogeographical origins', *Nature Communications*, 5(1), p. 3513. Available at: https://doi.org/10.1038/ncomms4513.

Howie, B. *et al.* (2012) 'Fast and accurate genotype imputation in genome-wide association studies through pre-phasing', *Nature Genetics*, 44(8), pp. 955–959. Available at: https://doi.org/10.1038/ng.2354.

International Human Genome Sequencing Consortium *et al.* (2001) 'Initial sequencing and analysis of the human genome', *Nature*, 409(6822), pp. 860–921. Available at:

https://doi.org/10.1038/35057062.
Jumper, J. *et al.* (2021) 'Highly accurate protein structure prediction with AlphaFold', *Nature*, 596(7873), pp. 583–589. Available at: https://doi.org/10.1038/s41586-021-03819-2.
Kojima, K. *et al.* (2020) 'A genotype imputation method for de-identified haplotype reference information by using recurrent neural network', *PLOS Computational Biology*, 16(10), p. e1008207. Available at: https://doi.org/10.1371/journal.pcbi.1008207.
Kurki, M.I. *et al.* (2023) 'FinnGen provides genetic insights from a well-phenotyped isolated population', *Nature*, 613(7944), pp. 508–518. Available at: https://doi.org/10.1038/s41586-022-05473-8.
Leitsalu, L. *et al.* (2015) 'Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu', *International Journal of Epidemiology*, 44(4), pp. 1137–1147. Available at: https://doi.org/10.1093/ije/dyt268.
Li, H. *et al.* (2014) 'Relationship Estimation from Whole-Genome Sequence Data', *PLOS Genetics*, 10(1), p. e1004144. Available at: https://doi.org/10.1371/journal.pgen.1004144.
Lin, Z. *et al.* (2023) 'Evolutionary-scale prediction of atomic-level protein structure with a language model', *Science*, 379(6637), pp. 1123–1130. Available at: https://doi.org/10.1126/science.ade2574.
Loh, P.-R. *et al.* (2016) 'Reference-based phasing using the Haplotype Reference Consortium panel', *Nature Genetics*, 48(11), pp. 1443–1448. Available at: https://doi.org/10.1038/ng.3679.
Loos, R.J.F. (2020) '15 years of genome-wide association studies and no signs of slowing down', *Nature Communications*, 11(1), p. 5900. Available at: https://doi.org/10.1038/s41467-020-19653-5.
Marino, A.D. *et al.* (2022) 'A comparative analysis of current phasing and imputation software', *PLOS ONE*, 17(10), p. e0260177. Available at: https://doi.org/10.1371/journal.pone.0260177.
Mowlaei, M.E. *et al.* (2023) 'Split-Transformer Impute (STI): Genotype Imputation Using a Transformer-Based Model'. bioRxiv, p. 2023.03.05.531190. Available at: https://doi.org/10.1101/2023.03.05.531190.
Naito, T. *et al.* (2021) 'A deep learning method for HLA imputation and trans-ethnic MHC fine-mapping of type 1 diabetes', *Nature Communications*, 12(1), p. 1639. Available at: https://doi.org/10.1038/s41467-021-21975-x.
Nguyen, E. *et al.* (2024) 'Sequence modeling and design from molecular to genome scale with Evo', *Science*, 386(6723), p. eado9336. Available at: https://doi.org/10.1126/science.ado9336.
Poplin, R. *et al.* (2018) 'A universal SNP and small-indel variant caller using deep neural networks', *Nature Biotechnology*, 36(10), pp. 983–987. Available at: https://doi.org/10.1038/nbt.4235.
Pritchard, J.K., Stephens, M. and Donnelly, P. (2000) 'Inference of Population Structure Using Multilocus Genotype Data', *Genetics*, 155(2), pp. 945–959. Available at: https://doi.org/10.1093/genetics/155.2.945.
Rosen, Y. *et al.* (2023) 'Universal Cell Embeddings: A Foundation Model for Cell Biology'. Cell Biology. Available at: https://doi.org/10.1101/2023.11.28.568918.

Rubinacci, S. *et al.* (2021) 'Efficient phasing and imputation of low-coverage sequencing data using large reference panels', *Nature Genetics*, 53(1), pp. 120–126. Available at: https://doi.org/10.1038/s41588-020-00756-0.

Rubinacci, S., Delaneau, O. and Marchini, J. (2020) 'Genotype imputation using the Positional Burrows Wheeler Transform', *PLOS Genetics*, 16(11), p. e1009049. Available at: https://doi.org/10.1371/journal.pgen.1009049.

Scholtens, S. *et al.* (2015) 'Cohort Profile: LifeLines, a three-generation cohort study and biobank', *International Journal of Epidemiology*, 44(4), pp. 1172–1180. Available at: https://doi.org/10.1093/ije/dyu229.

Sollis, E. *et al.* (2023) 'The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource', *Nucleic Acids Research*, 51(D1), pp. D977–D985. Available at: https://doi.org/10.1093/nar/gkac1010.

Song, M. *et al.* (2022) 'An autoencoder-based deep learning method for genotype imputation', *Frontiers in Artificial Intelligence*, 5. Available at: https://doi.org/10.3389/frai.2022.1028978.

Tanaka, T., Nagata, Y. and Takemoto, A. (2022) 'Integrating biomedical and clinical data with BioBank Japan', *Nature Cardiovascular Research*, 1(7), pp. 597–598. Available at: https://doi.org/10.1038/s44161-022-00105-w.

The 1000 Genomes Project Consortium *et al.* (2015) 'A global reference for human genetic variation', *Nature*, 526(7571), pp. 68–74. Available at: https://doi.org/10.1038/nature15393.

The All of Us Research Program Genomics Investigators *et al.* (2024) 'Genomic data in the All of Us Research Program', *Nature*, 627(8003), pp. 340–346. Available at: https://doi.org/10.1038/s41586-023-06957-x.

The International HapMap Consortium (2003) 'The International HapMap Project', *Nature*, 426(6968), pp. 789–796. Available at: https://doi.org/10.1038/nature02168.

Theodoris, C.V. *et al.* (2023) 'Transfer learning enables predictions in network biology', *Nature*, 618(7965), pp. 616–624. Available at: https://doi.org/10.1038/s41586-023-06139-9.

Yang, F. *et al.* (2022) 'scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data', *Nature Machine Intelligence*, 4(10), pp. 852–866. Available at: https://doi.org/10.1038/s42256-022-00534-z.

Yang, X. *et al.* (2023) 'Polygenic scores in cancer', *Nature Reviews Cancer*, 23(9), pp. 619–630. Available at: https://doi.org/10.1038/s41568-023-00599-x.

Figures

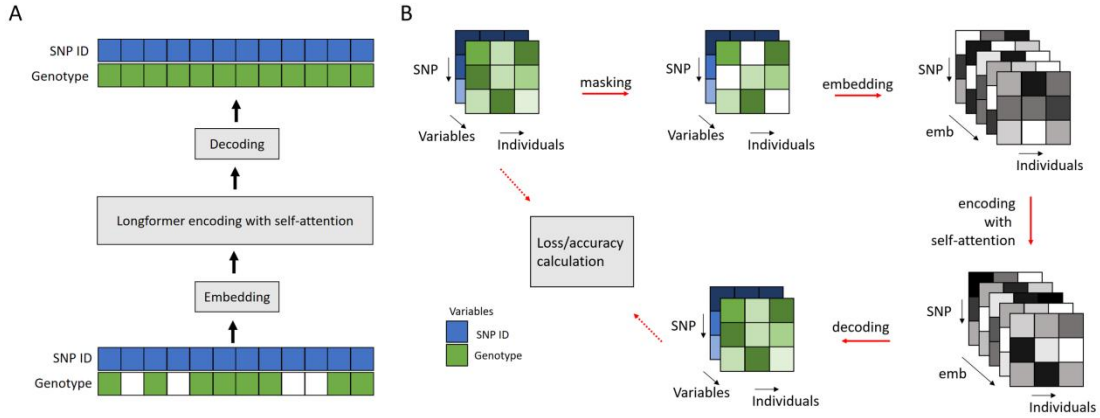


Figure 1: Model architecture. A) an overview. In pre-training, 15% of the values in genotype sequences (green squares) were masked (as white squares). The SNP ID and genotype sequences were aligned, combined and encoded with longformer self-attention. On the top, genotype values were decoded from the encoder output. B) In the detailed model architecture, a single batch of input data had 3 dimensions (SNP sites, individuals and variables). The variables were either SNP ID or genotype values. During masking, genotype values were masked and depicted as white squares. During embedding, the dimension variables were converted to the dimension emb to fit with the attention-based encoder. During decoding, the masked genotype values were recovered. Finally, the recovered genotype values were compared with the original values for calculating loss and accuracy.

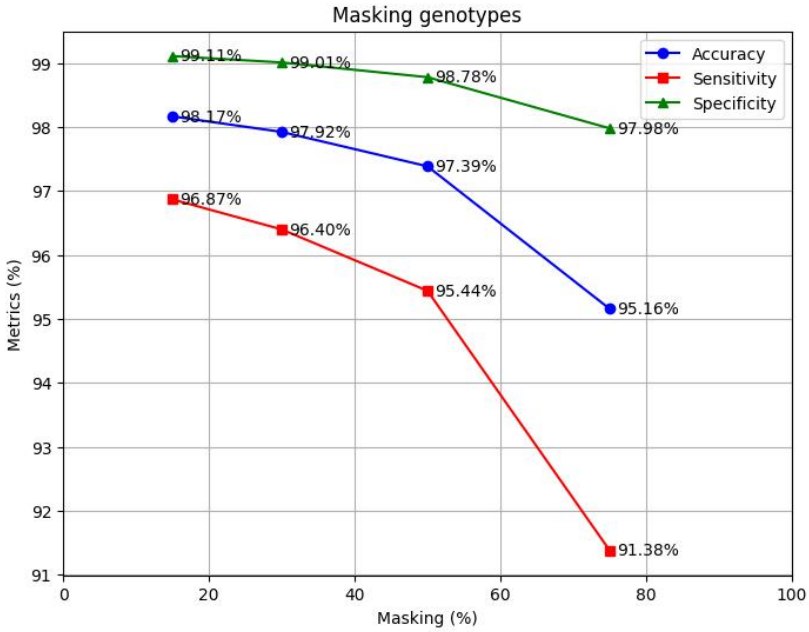


Figure 2: Pre-trained model. Direct inference of the pre-trained model on input sequences with 15%, 30%, 50%, and 75% of missing (masked) genotype values. Blue: overall accuracy in predicting genotypes. Red and green: sensitivity and specificity in predicting the heterozygous genotypes.

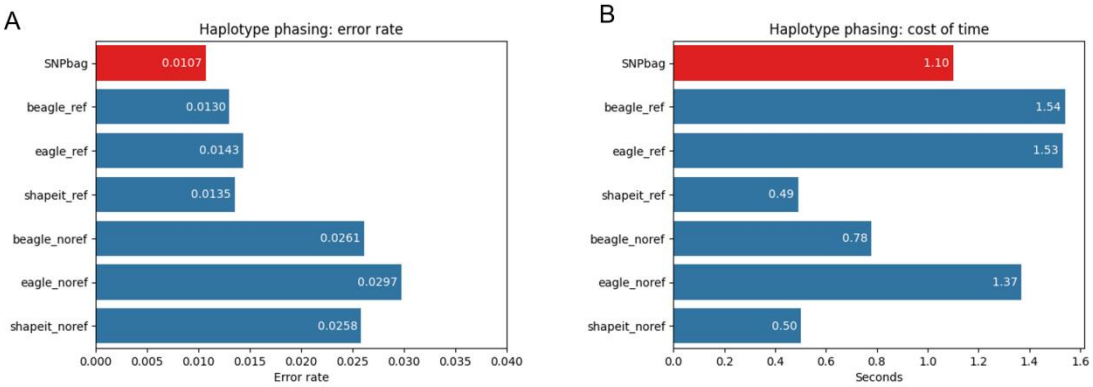


Figure 3: Haplotype phasing. The labels on the left of each bar are the names of the methods. Labels with '_ref' represent methods using a reference panel. Labels with 'noref' represent methods using no reference panel. The numbers on the bars are the actual values. The bars of our method (SNPBag) are colored as red. A) Bars represent the error rates in predicting switches. Error rate 0.0107 correspond to ~1% error. B) Bars represent the cost of time (seconds) for predicting switches for about 28,000 heterozygous sites on chr22 for 1 individual.

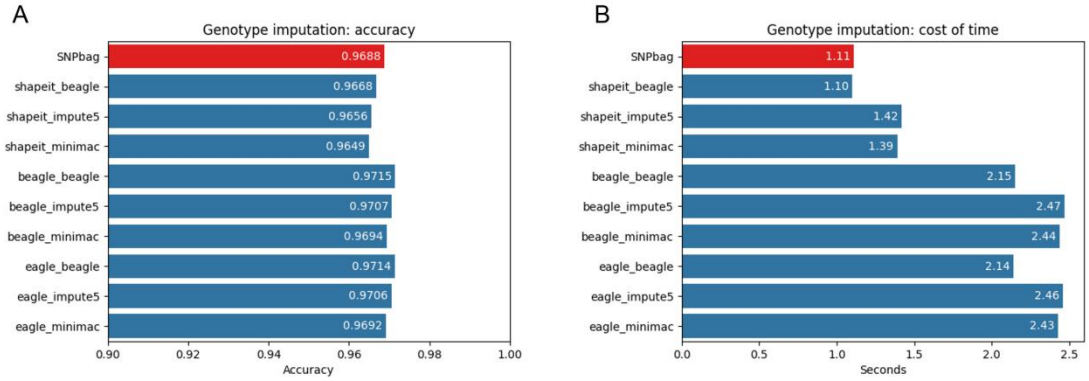


Figure 4: Genotype imputation. The labels on the left of bars indicate the combinations of phasing and imputation methods, joined by underscores, with phasing on left, and imputation on right. The bars of our method (SNPBag) are colored as red. A) Bars represent the overall prediction accuracy. The numbers on the bars indicate the actual values. SNPBag reached accuracy 0.9688, or 96.88%. B) Bars represent the cost of time (seconds) for predicting genotypes of about 130,000 sites on chr22 for 1 individual.

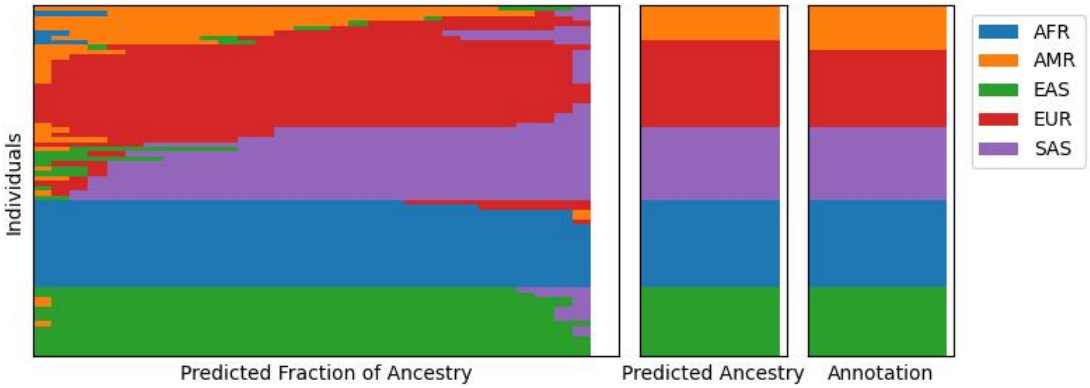
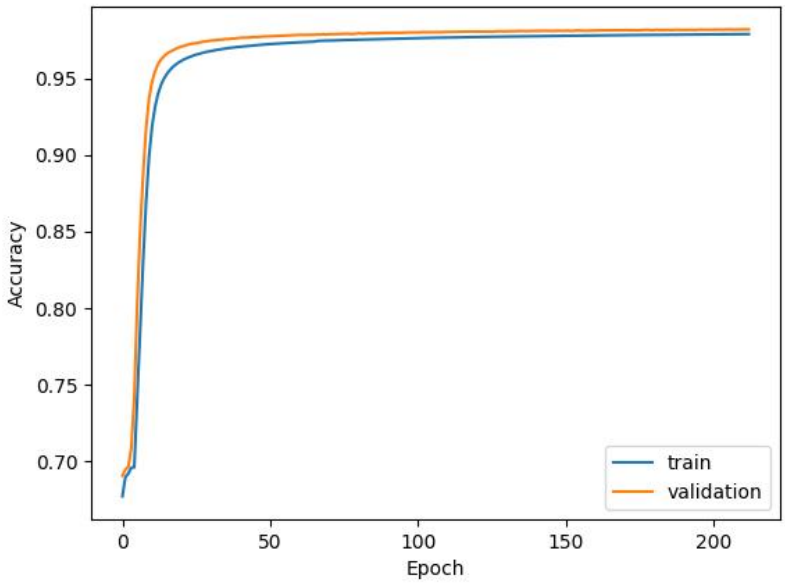


Figure 5: Ancestry inference. Colors represent 5 super-populations. For each individual in validation group, the fractions of chr22 were assigned to 1 of the 5 super-populations and painted as horizontal bars on the left panel. Middle panel showed the ancestries of the most assigned fractions for each individual on the left panel. Right panel showed the actual annotation from 1000 Genome Project for these individuals.

Supplementary Figures



Supplementary Figure 1: Learning Curve. The figure depicts the progression of model accuracy during pre-training. As the model updates its weights to minimize the training loss, the training accuracy (blue) increases steadily over epochs. The validation accuracy (orange) follows a similar trajectory, suggesting effective generalization. Both curves show a rapid initial improvement before gradually converging to a stable accuracy above 97%.