SNPBag: A Foundation Model for Multitask Genome-Scale SNP Analysis

Augix Guohua Xu[2], Yu Xu[3, 4], Yiming Xing[5], Pengchao Luo[6,8], Jianbo Yang[6,7*], Yinqi Bai[3,4*], Kun Tang[1*]

1 Research Center for Life Sciences Computing, Zhejiang Lab, Hangzhou, China
2 Leipzig University, Germany
3 BGI Research, Sanya 572025, China
4 BGI Research, Hangzhou 310030, China
5 College of Computer Science and Technology (College of software), 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
8 Hangzhou Institute for Advanced Study, University of Chinese Academy of sciences, Hangzhou, 310024 Zhejiang, China
* Corresponding authors: Kun Tang tang.kun@zhejianglab.org
Yinqi Bai baiyinqi@genomics.cn
Jianbo Yang yangjianbo@him.cas.cn

Abstract

Foundation models in artificial intelligence are revolutionizing biomedical research, yet single nucleotide polymorphism (SNP) data, critical for advancing biobank studies and decoding human genetic diversity, remain underexplored. We introduce SNPBag, a transformer-based foundation model that redefines genome-scale SNP analysis. Pre-trained on 1 million simulated genomes using 0.8 billion parameters, it captures evolutionary signatures of 6 million SNPs, encoding linkage disequilibrium and haplotype structures with high fidelity. This paradigm-shifting model supports versatile tasks. In genotype imputation, it matches leading algorithms in performance and, when fine-tuned, achieves state-of-the-art (SOTA) accuracy. In haplotype phasing, it outperforms non-reference methods and rivaling the best reference-based method with a 72-fold speedup. Notably, it compresses 6 million SNPs per individual into a 0.75 MB embedding, enabling efficient storage, transfer and downstream applications. In particular, SNPBag embeddings facilitate rapid ancestry inference across global populations and detection of genetic relationships up to 12th-degree relatives. To summarize, SNPBag establishes a scalable, self-sufficient, multitasking AI framework, poised to transform SNP data analysis and unlock the growing potential of biobank resources.

Introduction

Recent advances in artificial intelligence (AI) foundation models have transformed biomedical research, enabling comprehensive omics-scale analyses and reshaping our understanding of complex biological systems. In proteomics, the AlphaFold series accurately predicts protein structures and accelerates drug discovery (Jumper et al. 2021), while models such as ESM2 and ProteinMPNN facilitate protein design and engineering (Dauparas et al. 2022; Lin et al. 2023). In transcriptomics, particularly single-cell transcriptomics, models including Geneformer, scGPT, and scBERT support cell type annotation, gene regulatory network inference, and perturbation response prediction (Yang et al. 2022; Theodoris et al. 2023; Cui et al. 2024). In genomics, models such as Nucleotide Transformer (Dalla-Torre et al. 2024) and EVO (Nguyen et al. 2024) process nucleic acid sequences to predict gene essentiality and annotate genomic elements.

Despite these advances, no foundation model has been specifically developed for population-scale genetic diversity in humans or agriculturally important species, although this domain is intensively studied using conventional approaches. Single nucleotide polymorphisms (SNPs), the most prevalent form of genetic variation, underpin modern genomics. Landmark initiatives, including the International HapMap Project (The International HapMap Consortium 2003) and the 1000 Genomes Project (International Human Genome Sequencing Consortium et al. 2001), catalog human genetic variation globally. Large-scale biobanks—including the UK Biobank (Bahcall 2018), the Million Veteran Program and All of Us Research Program (The All of Us Research Program Genomics Investigators et al. 2024), the China Kadoorie Biobank (Chen et al. 2011), Biobank Japan (Tanaka et al. 2022), deCODE Genetics (Chadwick 1999), the Estonian Biobank (Leitsalu et al. 2015), FinnGen (Kurki et al. 2023), and Lifelines (Scholtens et al. 2015)—provide extensive population-scale genomic datasets, enabling systematic surveys of genome-wide variation and forming a rich substrate for AI model development.

Genome-wide association studies (GWAS) have leveraged these resources to link genotype and phenotype, with over 110,000 studies covering ~15,000 traits and millions of genomes (Sollis et al. 2023). Beyond GWAS, systematic SNP analyses require SNP calling, genotype imputation, haplotype phasing, population stratification, and relatedness inference. Leading bioinformatic platforms and software include GATK and DeepVariant for variant calling (McKenna et al. 2010; Poplin et al. 2018); IMPUTE, BEAGLE, MINIMAC, and EAGLE for imputation (Fuchsberger et al. 2015; Loh et al. 2016; Rubinacci et al. 2020; Browning et al. 2021); BEAGLE, SHAPEIT, and EAGLE for haplotype phasing (Fuchsberger et al. 2015; Rubinacci et al. 2020; Browning et al. 2021; Hofmeister et al. 2023); PLINK PCA and nonlinear methods (t-SNE, UMAP) for population stratification (Purcell et al. 2007; Diaz-Papkovich et al. 2019; Gaspar and Breen 2019); and KING, ERSA, PRIMUS, and Bonsai for relatedness inference (Huff et al. 2011; Ramstetter et al. 2017; Jewett et al. 2021). While effective, these tools face

limitations in scalability, computational efficiency, reference panel dependence, and task-specific fragmentation. Recently, neural network approaches have also emerged (Chen and Shi 2019; Kojima et al. 2020; Naito et al. 2021; Dias et al. 2022; Song et al. 2022; Mowlaei et al. 2023; Huang et al. 2024). However, they primarily provide end-to-end solutions towards specific tasks and share similar characteristics and limitations with conventional methods, further complicating the selection of the most appropriate model. Combining the top-performing tools into a continuously updated 'best-practice' pipeline remains a long-term challenge. In contrast, attention-based foundation models offer a potential paradigm shift, enabling scalable pre-training on population-scale data and multitask inference within a unified framework, without reliance on reference panels.

Here, we present SNPBag, the first foundation model specifically designed for genome-scale SNP data to our knowledge. SNPBag effectively captures SNP sites and inter-site dependencies, enabling broad-range support for all the tasks mentioned above through its base and fine-tuned models, including genotype imputation, haplotype phasing, genome embedding, population classification, and distant relatedness inference. When benchmarked against current state-of-the-art tools, SNPBag demonstrated comparable or superior performance across these tasks, with substantial reductions in running time, achieving 10- to 100-fold speedups. Notably, leveraging the generalization ability of our foundation model, SNPBag was trained on 1 million synthetic genomes while being tested on *bona fide* datasets, thereby mitigating public concerns regarding personal privacy leakage. Its high accuracy in relatedness inference underscores its potential utility in clinical and legal applications. Furthermore, SNPBag's ability to achieve a high compression rate of genomic data highlights its potential for future DNA storage applications. We have made the model inference publicly available on GitHub (https://github.com/augix/SNPBag) and invite the research community to explore additional application scenarios.

Results

Base model pre-training

SNPBag, a model leveraging Bidirectional Encoder Representations from Transformers (BERT), employs an attention mechanism to capture complex dependencies among SNP sites, generating contextualized representations of input genotypes. These representations enable the decoding of diverse features for individual genomes, as outlined in the SNPBag workflow (Fig. 1a). To address limited access to biobank-scale individual genomes, we simulated 1 million phased genomes using publicly available data (Fig. 1a, top). These synthetic genomes were used to pre-train a base model and fine-tune task-specific models (vertical arrows, Fig. 1a). During inference (horizontal arrows, Fig. 1a), input data of varying formats were processed as sequences by trained models to perform tasks such as genotype imputation, haplotype phasing, and genome embedding generation. The resulting compressed genome

embeddings enabled downstream applications, including population classification, relatedness inference, and potentially phenotype prediction.

For pre-training (Fig. 1b, c), we masked 85–99% of genotypes randomly and paired them with corresponding SNP identifiers to form input sequences. These sequences were projected into a 512-dimensional space using linear embedding layers, feeding into a BERT encoder with 16 transformer layers, each with 16 attention heads. The encoder's output was processed by a multilayer perceptron (MLP) to predict genotype values at masked positions, with model weights optimized using cross-entropy loss between predicted and actual genotypes. By swapping decoders, the base model was fine-tuned for tasks such as haplotype prediction (brown squares, Fig. 1b, c). Using this architecture, we trained two base models on distinct datasets: a chromosome 22 (chr22) model with 38 million parameters for benchmarking imputation and phasing, and a whole-genome model with 836 million parameters for broader applications. To prevent data leakage, the chr22 model was pre-trained on 1 million genotype sequences synthesized from 3,202 samples of the 1000 Genomes Project phase 3 (1kG) and evaluated on Human Genome Diversity Project (HGDP) samples. The whole-genome model, designed for public use, was pre-trained on 1 million synthetic genomes from the HAPNEST project (Wharrie et al. 2023), incorporating 1kG and HGDP samples as ancestors to maximize genetic diversity.

SNP sites were restricted to single-nucleotide polymorphisms shared between 1kG and HGDP datasets with a minor allele frequency (MAF) >1% (based on 1kG data). The whole-genome model was trained on 6,030,514 SNPs organized into consecutive contigs of 81,920 sites, while the chr22 model used a single 81,920-site contig, covering ~92% of chromosome 22.

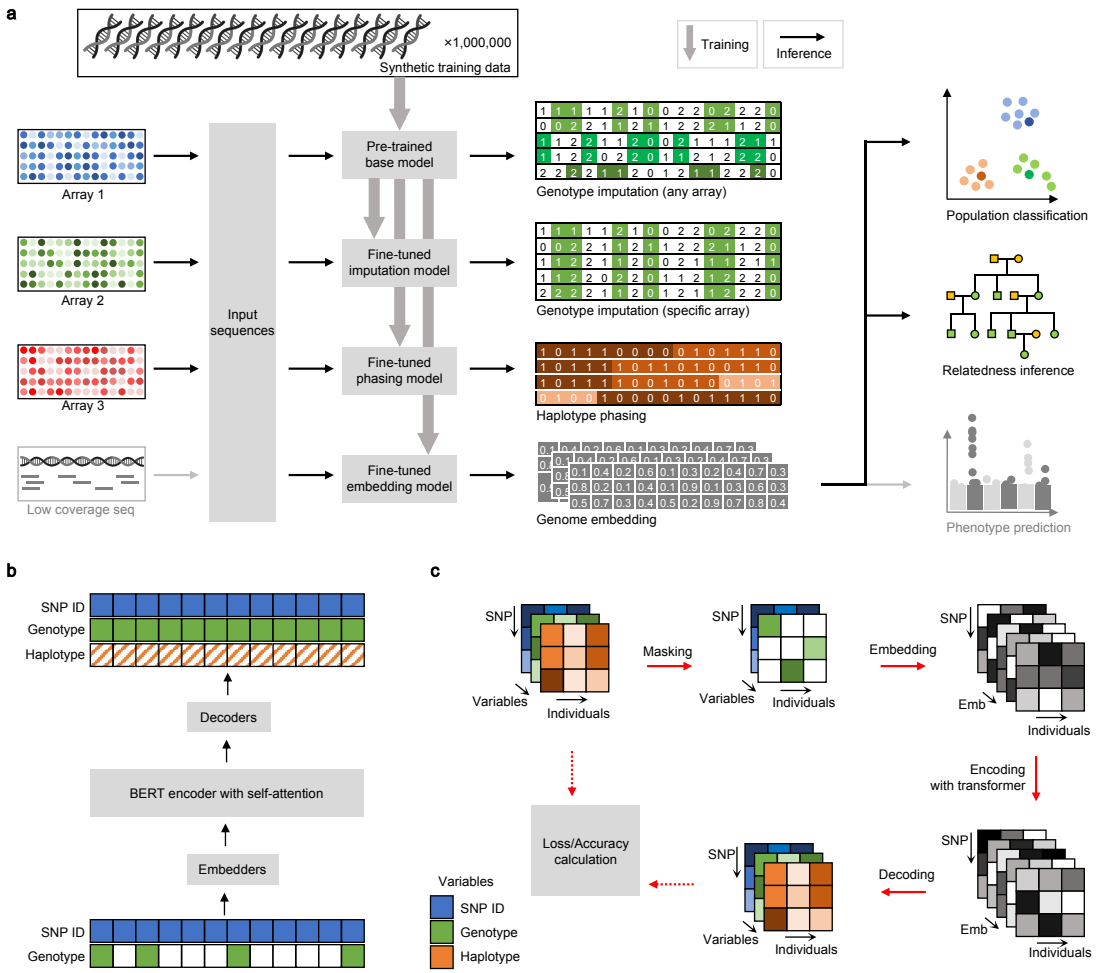

Fig. 1 | Model architecture. **a,** Overview of SNPBag, a multitasking foundation model for efficient genome-scale SNP analysis. One million synthetic genome datasets (top) were used for pre-training and fine-tuning across diverse tasks (vertical arrows). The base model (also in panel b), pre-trained via masked genotype prediction, functions directly as a baseline genotype imputer for both commonly used and custom-designed SNP arrays, and potentially for raw genomic data (e.g., low-coverage sequencing data). Fine-tuning the base model further enhances imputation accuracy for specific SNP arrays. Haplotype phasing is achieved by fine-tuning the base model with a modified decoding head to predict haplotype switch states. Additionally, training an embedding head enables SNPBag to compress genome-wide SNPs into low-dimensional genome embeddings, supporting downstream tasks such as population classification and relatedness inference. **b,** The base model adopts a Bidirectional Encoder Representations from Transformers (BERT) architecture. During pre-training, portions of genotype sequence values (green squares) are masked (white squares), combined with their corresponding SNP IDs, and encoded by the BERT model. Genotype values were decoded in the output. This structure is also adapted for fine-tuning of haplotype phasing (brown) by replacing the output head to a phase decoder. **c,** Detailed model architecture. A single batch of input data—with three dimensions (SNP sites, individuals, variables)—is combined into tokens, where the "variables" dimension is converted to an "emb" dimension to match the attention-based encoder. During decoding, either masked genotype values (green) or haplotype phase (brown) are recovered from output tokens and compared to original values to compute loss and accuracy.

Genotype imputation

Genotype imputation infers untyped variants, enhancing data completeness for GWAS, reducing sequencing needs, and aligning diverse array platforms for broader analyses. The imputation relies on the linkage disequilibrium (LD) structure, where adjacent SNPs are inherited together. To benchmark SNPBag's imputation, we tested 216 HGDP individuals on the Illumina Omni2.5 array on chromosome 22, masking non-array SNPs to impute the full SNP set. Traditional methods, combining phasing (SHAPEIT4, BEAGLE5.2, EAGLE2) and imputation (IMPUTE5, MINIMAC4, BEAGLE5.2) steps with or without using the 1KG reference panel, were compared (Fuchsberger et al. 2015; Loh et al. 2016; Delaneau et al. 2019; Rubinacci et al. 2020; Browning et al. 2021).

The SNPBag base model scored 96.78% imputation accuracy (concordance rate), ranking 11th among all the 20 tests in comparison (Fig. 2a). To optimize imputation performance, the base model was fine-tuned to create the SNPBag omni model by using the masking pattern of the Omni2.5 array instead of random masking. Scaling law experiments with sample sizes of 1 million, 100 thousand, 10 thousand, and 2 thousand revealed diminishing accuracy gains with smaller datasets, suggesting over-fitting, particularly for samples below 100 thousand (Supplementary Fig. 1a, b). Fine-tuning with 1 million RESHAPE (Cavinato et al. 2024) synthetic individuals yielded a state-of-the-art (SOTA) accuracy of 97.16% after approximately one epoch, surpassing the best traditional method (BEAGLE5.2+IMPUTE5+reference panel, 97.10%) with a statistically significant difference (paired t-test *p-value* 2.7×10^{-4}, Fig. 2a). Across minor allele frequency (MAF) intervals with comparable SNP counts, the fine-tuned model (SNPBag omni) consistently outperformed other methods, exceeding even the average accuracy of reference-based methods (REF) by 0.1–0.7% across intervals (Fig. 2b). Furthermore, the fine-tuned model also showed best performance in all continental groups (red boxes in Fig. 2c). Population-specific accuracy was highest in Americans (AMR, 98.3%), remaining relatively high among all the out-of-African (OOA) groups (96.2 - 97.8%), but much lower in Africans (AFR, 94.2%) with higher inter-individual variance (Fig. 2c, Supplementary Fig. 2). The overall continental profile remains similar among all methods (Fig. 2c). The low imputation accuracy rates in Africans reflect the high diversity and low inter-site LD which both lead to higher uncertainty in prediction (The 1000 Genomes Project Consortium et al. 2015). Notably, SNPBag's inference time on chromosome 22 averaged ~0.09 seconds with one NVIDIA A800 GPU, 13–20 times faster than traditional methods (1 Intel® Xeon® Platinum 8458P CPU, Fig. 2d). Overall, SNPBag's reference-free imputation and ability to handle arbitrary arrays highlight its potential for efficient, scalable genomic analysis.

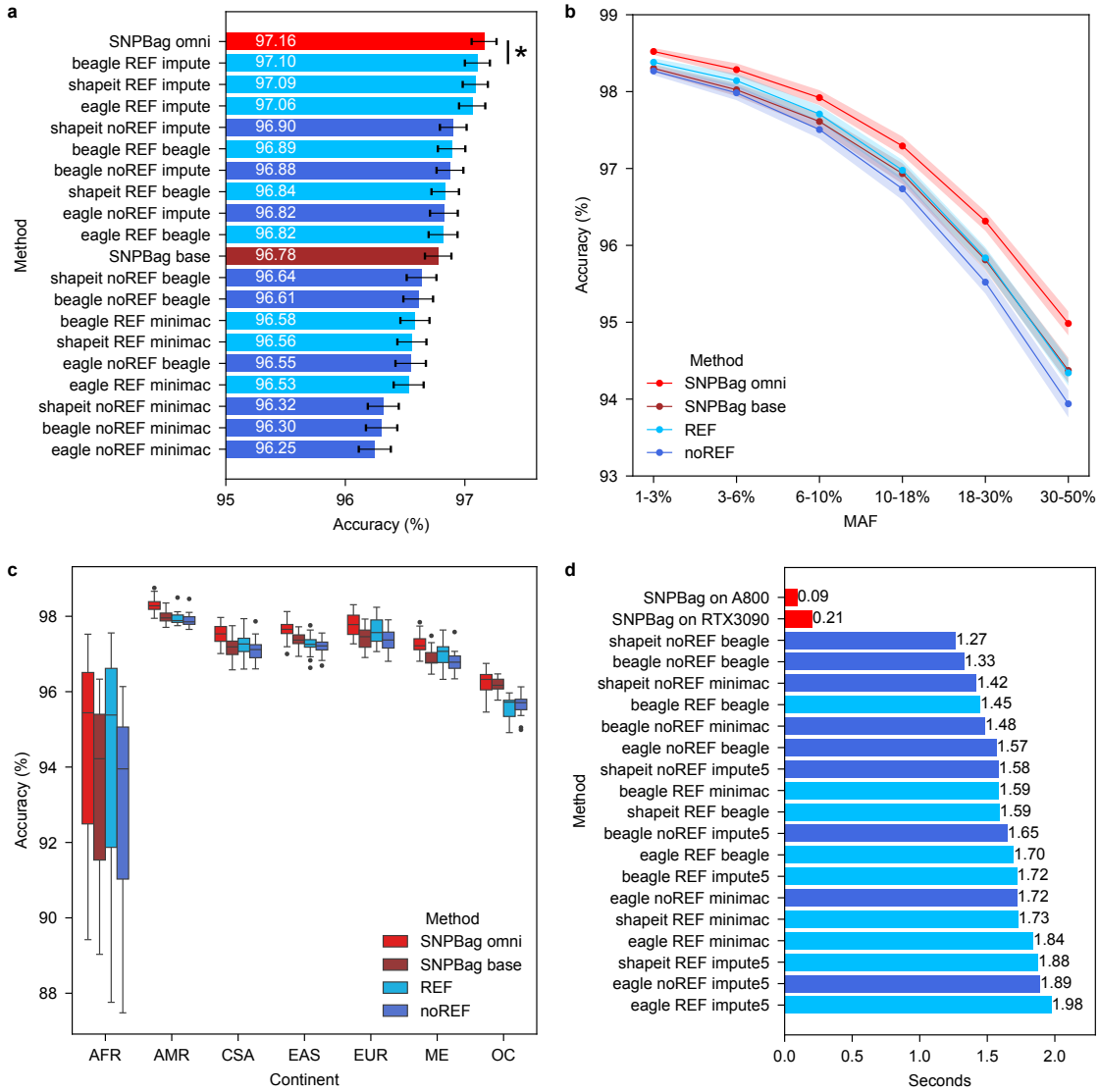

Fig. 2 | Genotype imputation benchmarking. a,d, Performance comparisons of SNPBag fine-tuned model (red) and base model (brown) versus traditional methods with reference panel (light blue), or without reference panel (dark blue) for imputation accuracy (a) and inference speed (d). **b,** Imputation accuracy of different methods across bins of minor allele frequency, with roughly same numbers of SNPs within each bin. REF and noREF represent the average of traditional methods with and without the reference panel. **c,** Boxplot of imputation accuracy by methods and continental groups. Continental groups included Africans (AFR), Americans (AMR), Central and South Asians (CSA), East Asians (EAS), Europeans (EUR), Middle East populations (ME) and Oceanians (OC). Standard errors of mean estimation were plotted as error bars in panel a and as shadows in panel b.

Haplotype phasing

Haplotype phasing assigns alleles to maternal and paternal chromosomes, forming distinct haplotypes that enhance disease association detection and clarify inheritance patterns. These haplotype blocks are fundamental to the diploid human genome. SNPBag's base model, pre-trained to capture haplotype distributions, was fine-tuned

for phasing on the full SNP set using a new MLP head to predict switch states between consecutive heterozygous sites. We trained on 1 million phased genomes from RESHAPE, tested with 26 physically phased HGDP individuals on chromosome 22. Scaling law tests showed that smaller training sample sizes (2 thousand and 10 thousand compared to 1 million) reduced accuracy and caused earlier over-fitting (Supplementary Fig. 1c, d). With 1 million individuals, SNPBag achieved a 3.07% switch error rate, outperforming the traditional methods without reference panel (6.27%–7.04%) but slightly falling below methods using the reference panel (2.54-2.70%, Fig. 3a). SNPBag's phasing accuracy lags behind traditional methods, likely because its training data's 'true phase', derived from traditional phasing results (The 1000 Genomes Project Consortium et al. 2015), limits its performance ceiling. Across continental groups, SNPBag phasing performance exhibits strong consistency with the average of reference-panel methods, with highest errors in Africans (5.0% and 3.9%), followed by Oceanians (3.3% and 2.8%), and generally low errors in other groups (1.3% ~ 2.4% and 1.0 ~ 2.3%) respectively (Fig. 3b). High African errors likely stem from elevated haplotype diversity, while suboptimal Oceanian phasing may result from unrepresented haplotypes in training data or reference panels. Notably, SNPBag's reference-free phasing, at ~0.09 seconds with a NVIDIA A800 GPU, was 11–91 times faster than traditional methods (1.0–8.2 seconds, with an Intel® Xeon® Platinum 8458P CPU), offering efficiency and flexibility (Fig. 3c). Future phasing accuracy gains are expected with true haplotype training data (Snyder et al. 2015).

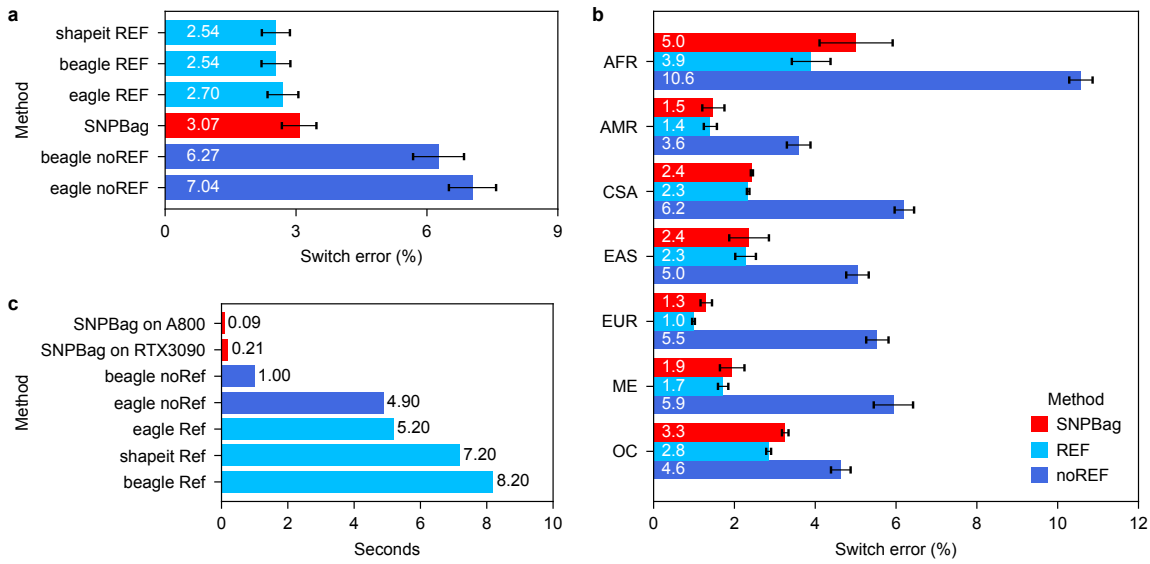

Fig. 3 | Haplotype phasing benchmarking. a, Comparisons of switch error rates between SNPBag (red) and traditional methods with reference panel (light blue) or without reference panel (dark blue). **b,** Phasing performance of different methods across 7 continental groups. **c,** Comparisons of inference speed between SNPBag (red) and traditional methods with reference panel (light blue) or without reference panel (dark blue). Standard errors of mean estimation were plotted as error bars in panel a and b.

Genome embedding

Whole-genome analyses, such as GWAS and population structure inference, pose computational challenges due to numerous loci. Representing genomes in low-dimensional spaces reduces complexity and enables efficient downstream analysis. We first pre-trained SNPBag on genome-wide SNPs using 1 million HAPNEST-synthesized genomes from 1KG and HGDP panels (Wharrie et al. 2023). Pre-training processed each of the 22 autosomal chromosomes in parallel, completing 0.2 epoch (200k individuals), resulted in a base model reaching 97.7% accuracy in genotype imputation on Omni2.5 array (Supplementary Fig. 3a). For genome embedding, we divided the genome into 2,934 consecutive contigs, each containing 2,048 SNPs. Two sequential MLPs compressed each contig's output into a 128-length embedding, with a third MLP decoder recovering the original contig sequence at 95% accuracy (Supplementary Fig. 3b), improvable with further training. Notably, the whole-genome embedding occupies only ~0.75 megabytes (751,104 bytes, or 2,934×128 dim×2 byte/dim), enabling efficient storage, transfer, encryption, and large-scale analysis (Supplementary Fig. 3c). This compact representation enhances the scalability and practicality of genomic analyses, supporting diverse applications without sacrificing accuracy. Here we introduce its application in population classification and relatedness inference.

Population classification

Genome embeddings, capturing high-resolution SNP distributions, enable population structure analysis critical for GWAS correction and forensic ethnicity inference (Diaz-Papkovich et al. 2019; Aalbers et al. 2020). While conventional PCA-based methods have limited resolution for sub-continental groups, machine learning approaches excel in resolving fine-scale structures (Diaz-Papkovich et al. 2019; Gaspar and Breen 2019). We generated embeddings for all 3202 1KG samples, visualized via UMAP, revealing distinct sub-continental clusters (Fig. 4a). Notably, Japanese are separated from Chinese Han (CHB, CHS) and Dai-Kinh (CDX, KHV) clusters; Europeans form Finnish, North-Western (CEU, GBR), and Southern (TSI, IBS) clusters; American groups (PUR, CLM, MXL, PEL) are distinguishable; South Asian populations (PJL, STU, BEB) cluster closely, apart from Gujarati (GIH); and African populations (GWD, LWK, MSL) are distinct, except for overlapping Yorubans (YRI) / Esans (ESN), and African Caribbean (ACB) / African Americans (ASW) (Fig. 4a).

By training a lightweight neural net with genome embeddings of all 1KG samples, we achieved an F1 score of 0.998 for the classification of five continental super-populations (EUR, EAS, SAS, AMR, AFR) (Supplementary Fig. 4a). For the classification of 19 sub-populations, our model attained an average F1 score of 0.861, surpassing Gaspar & Breen's benchmark SVM and GTM models (F1=0.8) (Gaspar and Breen 2019), excelling in sub-continental discrimination (Fig. 4b). Results for all 26 populations aligned with UMAP distributions (Supplementary Fig. 4b).

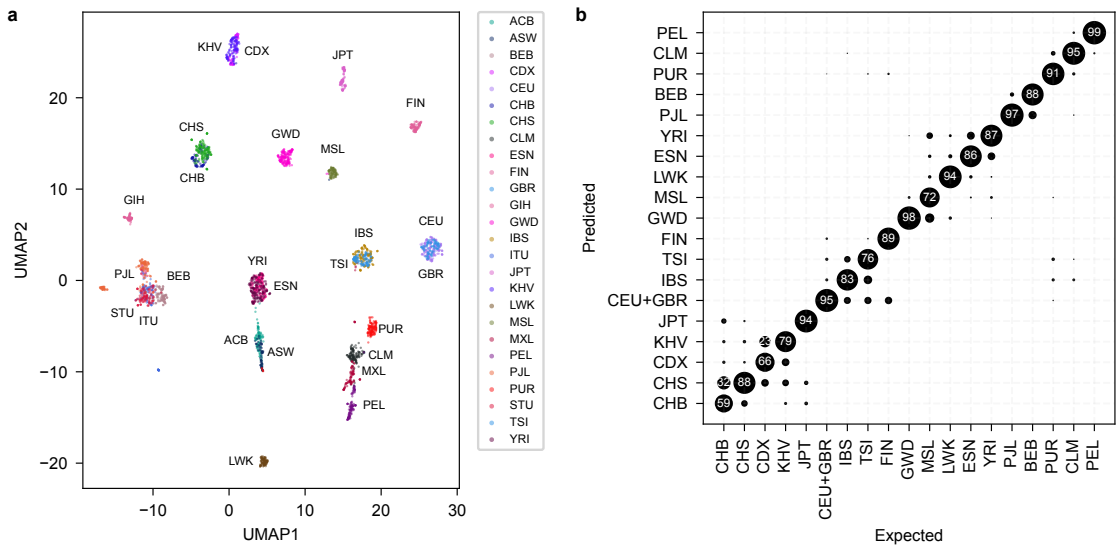

Fig. 4 | Population classification across 1000 Genomes Project (1KG) populations. a, UMAP visualization showing clustering of genome embeddings among 26 global populations, including 7 African (AFR) groups: Yoruba in Ibadan, Nigeria (YRI); Luhya in Webuye, Kenya (LWK); Mende in Sierra Leone (MSL); Esan in Nigeria (ESN); African Ancestry in Southwest USA (ASW); African Caribbean in Barbados (ACB); Gambian in The Gambia (GWD); 5 East Asian (EAS) groups: Han Chinese in Beijing (CHB); Japanese in Tokyo (JPT); Southern Han Chinese (CHS); Chinese Dai in Xishuangbanna (CDX); Kinh in Ho Chi Minh City, Vietnam (KHV); 5 European (EUR) groups: British in England and Scotland (GBR); Utah Residents with Northern and Western European Ancestry (CEU); Toscani in Italy (TSI); Finnish in Finland (FIN); Iberians in Spain (IBS); 5 South Asian (SAS) groups: Gujarati Indian in Houston (GIH); Punjabi from Lahore, Pakistan (PJL); Bengali in Bangladesh (BEB); Indian Telugu in the UK (ITU); Sri Lankan Tamil in the UK (STU); and 4 American (AMR) groups: Colombian in Medellín (CLM); Mexican Ancestry in Los Angeles (MXL); Puerto Rican in Puerto Rico (PUR); Peruvian in Lima (PEL). **b,** Confusion matrix of classification results for 19 populations, with values shown as percentages of individuals in expected classes.

Relatedness inference

Inferring relatedness is critical for case-control studies, where cryptic relatedness can introduce false associations (Ramstetter et al. 2017), and for improving rare disease variant detection through shared ancestry (Lancaster et al. 2024). It also supports forensic and direct-to-consumer kinship analysis (Jewett et al. 2021). Distant relatedness relies on detecting long identical-by-descent (IBD) haplotype segments (Gusev et al. 2009).

SNPBag's genome embeddings, preserving extensive haplotype information, should enable robust distant relatedness inference. We developed an MLP-based relatedness estimator using whole-genome embeddings. Trained on simulated deep pedigrees (up to 9th cousins) derived from 1KG founders, the estimator recalled relationships up to 12 degrees (Fig. 5a). At a high relatedness detection power of 0.98-1.00, the recall rates

for exact matched degrees (D_0) ranged from 0.96-1.00 for close relatives (degrees 1–2), 0.34–0.82 for intermediate (degrees 3–6), and 0.24-0.27 for distant (degrees 7–9) relatives (Fig. 5b). Beyond degree 10, exact match accuracy was low (0–0.09). However, allowing errors of 1–3 degrees (defined as D_1, D_2 and D_3 respectively) yielded high accuracy across degrees 1 – 12, with the accuracy remaining above 0.75 up to the 8th, 10th and 11th degree relationships for D_1, D_2 and D_3 respectively (Fig. 5b). Overall accuracy is comparable to the well-established methods for pairwise relatedness inference (Ramstetter et al. 2017). When comparing to the industry gold standard 23andMe's Bonsai package (Jewett et al. 2021), SNPBag's accuracy profile shows high consistency with Bonsai's DRUID method (Pearson's correlation coefficients: 0.983, 0.975, 0.966, and 0.943 for D_0, D_1, D_2, and D_3, respectively) and aligns with Bonsai's likelihood method for D_0 and D_1 (Pearson's correlations: 0.981 and 0.898, respectively, Supplementary Fig. 5). Notably, SNPBag requires no complex phasing or IBD calculations, resulting in a cost of time less than 0.02 seconds for a single pair of individuals, highlighting high efficiency in applications.

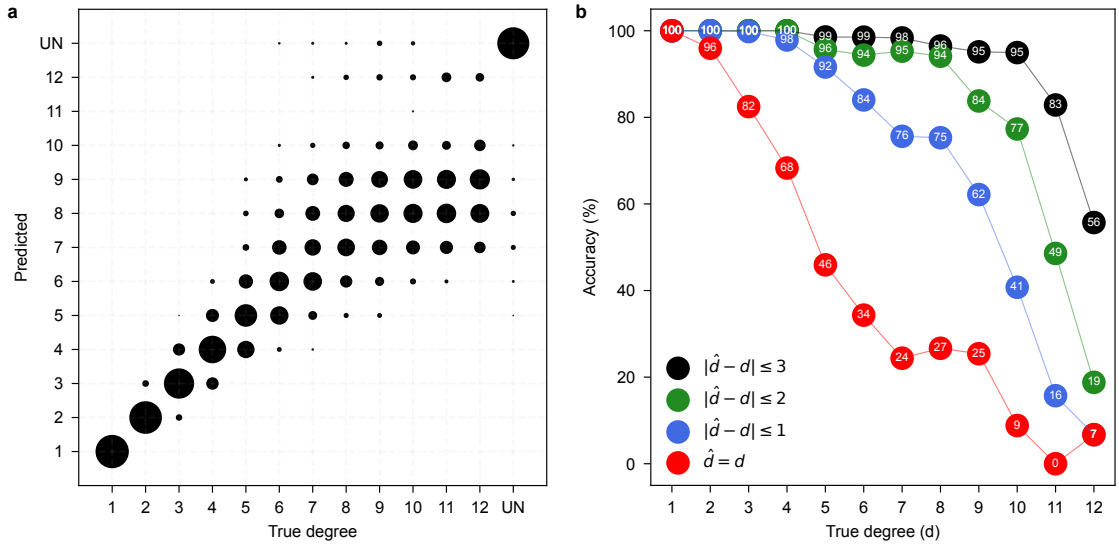

Fig. 5 | Performance of relatedness estimation. a, the bubble plot of the estimated degree of relatedness against the true relationship. Area of the circles represent percentages of pairs in each degree of relatedness. UN: unrelated. **b,** the accuracy of inference was evaluated at four error range allowance: D_0 in red: exact match between the estimation and truth; D_1 in blue: 1-degree mismatch allowed; D_2 in green: up to 2 degrees of mismatch allowed; and D_3 in black: up to 3 degrees of mismatch allowed. The numbers in the curves indicate accuracy in percentage.

Discussion

In this study, we present SNPBag, to our knowledge the first AI model that unified genome-scale analysis tasks like imputation, phasing, population classification, and relatedness inference within a transformer-based foundation model. While still in its early development, SNPBag delivered rivalling capabilities compared to top traditional methods, with competitive accuracies, superior inference efficiency and reference panel

independence. These results highlight foundation model's potential to redefine the paradigm for genome-scale SNP analysis.

SNPBag operates through a fundamentally distinct mechanism compared to traditional methods. The base model, through pre-training on global genomic data, effectively learns inter-site dependencies, underlying haplotype architectures, and even cryptic demographic structures inherent in the training data. In essence, base model pre-training involves learning a condensed, reformulated representation of the human genomic diversity—capturing both its totality and granular details—a process like how large language model (LLMs) internalize patterns in vast text corpora. Fine-tuning for specific tasks primarily involves extracting preexisting knowledge or patterns through representational realignment and refinement. The model inference is simply layer-wise propagation, guided by learned weights, to yield a result based on pre-stored knowledge.

This has several important implications: First, theoretically, the base model can leverage data of unbounded scale—such as the population-level datasets (now in the millions) available or planned (e.g., The Human Genome Project II) (Liu et al. 2024), across global biobanks. Scaling laws ensure performance improvements as data scale increases, suggesting that foundation-model-based inference may ultimately outperform all traditional methods when trained on sufficiently large empirical datasets, analogous to the performance improvement in LLM training. Second, foundation model inference—centered on outputting pre-learned knowledge and probabilities, operating with linear complexity—is inherently more efficient than traditional statistical methods, which typically require intensive *ab initio* computations and exhibit non-linear complexity. This accounts for the notable differences in inference latency between SNPBag and traditional approaches in genotype imputation and phasing. Third, traditional methods require multi-individual-level data, such as reference panels or sample cohorts, relying on group-based statistical distributions. In contrast, SNPBag can perform inference on a single-individual's input without any extra data, significantly enhancing efficiency and flexibility. Finally, the foundation model provides a unified framework ideally suited for multi-modal inputs and multi-tasking inference. Existing models can be further enhanced by iterative training on additional data, or extended via fine-tuning for new tasks.

The exponential growth of personal genomic data offers opportunities but poses significant challenges for modern human genomic analysis. Privacy concerns, driven by re-identification risks and stringent regulations, restrict data sharing. This has hindered the global collaborations on large-scale genomic analysis such as GWAS, as traditional imputation and phasing require large shared reference panels. One solution is to synthesize genome data mimicking empirical distributions, but the synthesized data has limited diversity and may mismatch empirical data in details(Cavinato et al. 2024). Notably, SNPBag can address these challenges by confidential training within each biobank, and sharing the model weights or offering online services for third parties' application. A 'Global SNP Model' may be developed by iterative or distilled training across different national bio-banks, providing genomic inference service based on the

most comprehensive and deep understanding of global human diversity. On the other hand, the handling of large-scale raw genomic data—comprising millions of SNPs and their annotations—presents challenges in storage, transfer, alignment, and encryption. SNPBag's compact genome embeddings (0.75MB with further compression potential) facilitate data handling, support multitask inference, and enable simpler and more secure encryption compared to raw data.

Last, the ultimate goal of large-scale SNP mapping is to predict and interpret heritable phenotypes from genotype data. GWAS studies have identified large number of loci linked to complex traits, yet each contributes minimally to phenotypic effects (Loos 2020; Sollis et al. 2023) Multivariate methods like polygenic scores (PGS), polygenic risk scores (PRS), and GTCA models aggregates SNP effects linearly but fails to account for linkage disequilibrium (LD) complexity and revealed limited predictive power due to missing heritability(Abdellaoui et al. 2023; Yang et al. 2023). Efforts have been taken to use artificial neural network-based approaches to capture the non-linear interactions, but at limited SNP sites (Sigurdsson et al. 2023). SNPBag, which encodes whole-genome SNPs, LD and haplotype structures into compact genome embeddings, may serve as an ideal framework for robust phenotype prediction by efficiently modeling high-order epistatic interactions across the genome.

Methods

Base model pre-training

The SNPBag base model was developed using a Bidirectional Encoder Representations from Transformers (BERT) architecture, following standard BERT pre-training protocols. Genotype sequences, encoded as 0, 1, or 2 (representing the number of derived alleles), were randomly masked at 85–99% of positions (replaced with a mask token, 3) and paired with corresponding SNP identifiers to form input sequences of 81,920 tokens. Each token was projected into a 512-dimensional space via linear layers to meet the encoder's input requirements. The encoder consisted 16 stacked layers of bidirectional transformer encoder, each incorporating 16 attention heads with sliding window attention. The sliding window size was set to 128 tokens for the whole-genome model and 256 tokens for the chr22 model. The encoder output was decoded by a multilayer perceptron (MLP) to predict masked genotype values. The model was optimized using cross-entropy loss against the original unmasked genotypes.

Two base models with this architecture were pre-trained: a chromosome 22 (chr22) model and a whole-genome model. The chr22 model was pre-trained on 1 million genotype sequences simulated from 3,202 samples from the 1,000 Genomes Project (1kG). We grouped these 3,202 individuals into 5 super-populations as ancestors (Africans, Admixed Americans, East Asian, European and South Asian). Within each super-population, we simulated the offspring of 8th generation using RESHAPE software. The performance of the chr22 model was evaluated on independent Human

Genome Diversity Project (HGDP) samples. The whole-genome model was pre-trained on 1 million synthetic genomes from the HAPNEST project, incorporating 1kG and HGDP samples as ancestors to enhance diversity. With approximately 836 million parameters, it was evaluated for genome embedding generation, population classification, and distant relatedness inference.

Structural variants were excluded, and only single nucleotide polymorphism sites shared between the 1kG and HGDP datasets with a minor allele frequency (MAF) >1% (based on 1kG Phase 3 data) were selected, resulting in 6,030,514 sites for the whole-genome model. During training, SNPs on the same chromosome were organized into consecutive contigs ranging from 1,024 to 81,920 sites. For the chr22 benchmark model, a single contig of 81,920 sites was used, covering approximately 92% of chromosome 22. The first 6,000 sites and the last 967 sites of chromosome 22 were excluded due to the scarcity of SNP array probes.

Genotype imputation

Imputation accuracy was evaluated using the Illumina Omni2.5 array as a representative SNP array. For chromosome 22, we considered 81,920 SNP sites, of which 16,230 (20%) were present on the Omni2.5 array, with the remaining 80% requiring imputation. Complete genotype sequences from 216 HGDP individuals (4 randomly selected per 54 populations) were used. Non-array sites were masked according to the Omni2.5 array design, and masked sequences were input to several methods: SNPBag base, SNPBag omni, and conventional imputation pipelines. The SNPBag base method applied the pre-trained chr22 model directly for inference. The SNPBag omni method involved fine-tuning the chr22 model on imputation tasks using masking patterns aligned with the Omni2.5 array and simulated offspring genotypes from 1kG samples. Conventional imputation followed pipelines benchmarked by De Marino et al. (Marino et al. 2022), combining phasing (SHAPEIT4, BEAGLE5.2, or EAGLE2, with or without the 1kG reference panel) and imputation (IMPUTE5, MINIMAC4, or BEAGLE5.2). Performance was stratified by MAF bins and continental population groupings. MAF bins were determined by collecting roughly similar numbers of sites in each MAF interval.

Haplotype phasing

The chr22 base model was fine-tuned for haplotype phasing using 1 million phased genotype sequences synthesized via RESHAPE from 1kG samples. Unphased genotypes were input, and a new MLP decoder was trained to predict allelic state switches (ancestral or derived) between adjacent heterozygous sites, optimized using cross-entropy loss against true switch sequences. Note that 2 haplotype sequences of any given individual produce the same switch sequence as truth. Also, any predicted switch sequence can be directly translated into 2 haplotype sequences. Model performance was evaluated using switch error rates on 26 physically phased HGDP genomes, comparing SNPBag against conventional phasing tools (SHAPEIT4,

BEAGLE5.2, or EAGLE2, with or without the 1kG reference panel).

Genome embedding

The whole-genome base model was fine-tuned to generate genome embeddings. Each genome was segmented into 2,934 contigs of 2,048 SNPs each. Contig sequences were randomly masked at 75–85% of sites and processed by the base model's encoder. The encoder's output (2,048 × 512) of each contig was compressed into a 128-dimensional embedding using two sequential MLPs. The 1st MLP compresses sequence dimension from 2,048 to 16. The 2nd compresses model dimension from 512 to 8. This was flattened to obtain a contig embedding of 128 dimensions (16 × 8). A third MLP predicted the 2,048 SNP genotypes from the contig embedding, optimized via cross-entropy loss until achieving 95% genotype recovery accuracy. At last, contig embeddings were concatenated to form a genome embedding (2,934 × 128) for each individual.

Population classification

A lightweight neural network was trained for population classification using genome embeddings from 3,202 1kG individuals and their population labels. The network comprised three MLPs: two sequentially reduced the 2,934 × 128 genome embeddings to a 6 × 6 representation, and a third projected this onto 19 population categories defined by Gaspar and Breen (2019). Performance was assessed via 10 iterations of 5-fold cross-validation, randomly partitioning the 3,202 individuals into training and test sets (4:1 ratio) per iteration. F1 scores and other metrics were computed on test sets and averaged across iterations. We also did the same analyses for all 26 populations present in 1kG data.

Relatedness inference

From 2,504 unrelated 1kG individuals, 14 of 26 populations were selected. Each population was split into training and test sets, with each test set fixed at 20 individuals. Using ped-sim software, we simulated 40 pedigree trees for training and 4 for testing per population. Each tree, spanning 11 generations with 40 individuals, included 20 1kG genomes and 20 synthetic genomes, capturing pairwise relatedness from degree 1 to 19 and unrelated pairs (Supplementary Fig. 6). This yielded 12,320 synthetic genomes, 759,327 training pairs, and 73,345 test pairs. Genome embeddings were generated using SNPBag, and pairs of embeddings were combined for relatedness classification. A lightweight neural network with MLPs for dimensionality reduction and classification was trained for this task.

Acknowledgement

The research was supported by the Project of Sanya Yazhou Bay Science and Technology City, Grant No: SKJC-2024-02-002. It was also supported by High-

performance Computing Platform of YZBSTCACC (YaZhou Bay Science and Technology City Advanced Computing Center).

References

Aalbers, Sanne E., Michael J. Hipp, Scott R. Kennedy, and Bruce S. Weir. 2020. "Analyzing Population Structure for Forensic STR Markers in next Generation Sequencing Data." *Forensic Science International: Genetics* 49 (November): 102364. https://doi.org/10.1016/j.fsigen.2020.102364.

Abdellaoui, Abdel, Loic Yengo, Karin J.H. Verweij, and Peter M. Visscher. 2023. "15 Years of GWAS Discovery: Realizing the Promise." *The American Journal of Human Genetics* 110 (2): 179–94. https://doi.org/10.1016/j.ajhg.2022.12.011.

Bahcall, Orli G. 2018. "UK Biobank — a New Era in Genomic Medicine." *Nature Reviews Genetics* 19 (12): 737–737. https://doi.org/10.1038/s41576-018-0065-3.

Browning, Brian L., Xiaowen Tian, Ying Zhou, and Sharon R. Browning. 2021. "Fast Two-Stage Phasing of Large-Scale Sequence Data." *American Journal of Human Genetics* 108 (10): 1880–90. https://doi.org/10.1016/j.ajhg.2021.08.005.

Cavinato, Théo, Simone Rubinacci, Anna-Sapfo Malaspinas, and Olivier Delaneau. 2024. "A Resampling-Based Approach to Share Reference Panels." *Nature Computational Science* 4 (5): 360–66. https://doi.org/10.1038/s43588-024-00630-7.

Chadwick, R. 1999. "The Icelandic Database---Do Modern Times Need Modern Sagas?" *BMJ* 319 (7207): 441–44. https://doi.org/10.1136/bmj.319.7207.441.

Chen, Junjie, and Xinghua Shi. 2019. "Sparse Convolutional Denoising Autoencoders for Genotype Imputation." *Genes* 10 (9): 652. https://doi.org/10.3390/genes10090652.

Chen, Z., J. Chen, R. Collins, 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): 1652–66. https://doi.org/10.1093/ije/dyr120.

Cui, Haotian, Chloe Wang, Hassaan Maan, et al. 2024. "scGPT: Toward Building a Foundation Model for Single-Cell Multi-Omics Using Generative AI." *Nature Methods* 21 (8): 1470–80. https://doi.org/10.1038/s41592-024-02201-0.

Dalla-Torre, Hugo, Liam Gonzalez, Javier Mendoza-Revilla, et al. 2024. "Nucleotide Transformer: Building and Evaluating Robust Foundation Models for Human Genomics." *Nature Methods*, ahead of print, November 28. https://doi.org/10.1038/s41592-024-02523-z.

Dauparas, J., I. Anishchenko, N. Bennett, et al. 2022. "Robust Deep Learning–Based Protein Sequence Design Using ProteinMPNN." *Science* 378 (6615): 49–56. https://doi.org/10.1126/science.add2187.

Delaneau, Olivier, Jean-François Zagury, Matthew R. Robinson, Jonathan L. Marchini, and Emmanouil T. Dermitzakis. 2019. "Accurate, Scalable and Integrative Haplotype Estimation." *Nature Communications* 10 (1): 5436. https://doi.org/10.1038/s41467-019-13225-y.

Dias, Raquel, Doug Evans, Shang-Fu Chen, et al. 2022. "Rapid, Reference-Free Human Genotype Imputation with Denoising Autoencoders." *eLife* 11 (September): e75600. https://doi.org/10.7554/eLife.75600.

Diaz-Papkovich, Alex, Luke Anderson-Trocmé, Chief Ben-Eghan, and Simon Gravel. 2019. "UMAP Reveals Cryptic Population Structure and Phenotype Heterogeneity in Large Genomic Cohorts." *PLOS Genetics* 15 (11): e1008432. https://doi.org/10.1371/journal.pgen.1008432.

Fuchsberger, Christian, Gonçalo R. Abecasis, and David A. Hinds. 2015. "Minimac2: Faster Genotype Imputation." *Bioinformatics* 31 (5): 782–84. https://doi.org/10.1093/bioinformatics/btu704.

Gaspar, Héléna A., and Gerome Breen. 2019. "Probabilistic Ancestry Maps: A Method to Assess and Visualize Population Substructures in Genetics." *BMC Bioinformatics* 20 (1): 116. https://doi.org/10.1186/s12859-019-2680-1.

Gusev, Alexander, Jennifer K. Lowe, Markus Stoffel, et al. 2009. "Whole Population, Genome-Wide Mapping of Hidden Relatedness." *Genome Research* 19 (2): 318–26. https://doi.org/10.1101/gr.081398.108.

Hofmeister, Robin J., Diogo M. Ribeiro, Simone Rubinacci, and Olivier Delaneau. 2023. "Accurate Rare Variant Phasing of Whole-Genome and Whole-Exome Sequencing Data in the UK Biobank." *Nature Genetics* 55 (7): 1243–49. https://doi.org/10.1038/s41588-023-01415-w.

Huang, Xin, Aigerim Rymbekova, Olga Dolgova, Oscar Lao, and Martin Kuhlwilm. 2024. "Harnessing Deep Learning for Population Genetic Inference." *Nature Reviews Genetics* 25 (1): 61–78. https://doi.org/10.1038/s41576-023-00636-3.

Huff, Chad D., David J. Witherspoon, Tatum S. Simonson, et al. 2011. "Maximum-Likelihood Estimation of Recent Shared Ancestry (ERSA)." *Genome Research* 21 (5): 768–74. https://doi.org/10.1101/gr.115972.110.

International Human Genome Sequencing Consortium, Whitehead Institute for Biomedical Research, Center for Genome Research:, Eric S. Lander, et al. 2001. "Initial Sequencing and Analysis of the Human Genome." *Nature* 409 (6822): 860–921. https://doi.org/10.1038/35057062.

Jewett, Ethan M., Kimberly F. McManus, William A. Freyman, and Adam Auton. 2021. "Bonsai: An Efficient Method for Inferring Large Human Pedigrees from Genotype Data." *The American Journal of Human Genetics* 108 (11): 2052–70. https://doi.org/10.1016/j.ajhg.2021.09.013.

Jumper, John, Richard Evans, Alexander Pritzel, et al. 2021. “Highly Accurate Protein Structure Prediction with AlphaFold.” *Nature* 596 (7873): 583–89. https://doi.org/10.1038/s41586-021-03819-2.

Kojima, Kaname, Shu Tadaka, Fumiki Katsuoka, Gen Tamiya, Masayuki Yamamoto, and Kengo Kinoshita. 2020. “A Genotype Imputation Method for De-Identified Haplotype Reference Information by Using Recurrent Neural Network.” *PLOS Computational Biology* 16 (10): e1008207. https://doi.org/10.1371/journal.pcbi.1008207.

Kurki, Mitja I., Juha Karjalainen, Priit Palta, et al. 2023. “FinnGen Provides Genetic Insights from a Well-Phenotyped Isolated Population.” *Nature* 613 (7944): 508–18. https://doi.org/10.1038/s41586-022-05473-8.

Lancaster, Megan C., Hung-Hsin Chen, M. Benjamin Shoemaker, et al. 2024. “Detection of Distant Relatedness in Biobanks to Identify Undiagnosed Cases of Mendelian Disease as Applied to Long QT Syndrome.” *Nature Communications* 15 (1): 7507. https://doi.org/10.1038/s41467-024-51977-4.

Leitsalu, Liis, Toomas Haller, Tõnu Esko, et al. 2015. “Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu.” *International Journal of Epidemiology* 44 (4): 1137–47. https://doi.org/10.1093/ije/dyt268.

Lin, Zeming, Halil Akin, Roshan Rao, et al. 2023. “Evolutionary-Scale Prediction of Atomic-Level Protein Structure with a Language Model.” *Science* 379 (6637): 1123–30. https://doi.org/10.1126/science.ade2574.

Liu, Weibin, Yan Li, George P. Patrinos, et al. 2024. “The 1% Gift to Humanity: The Human Genome Project II.” *Cell Research* 34 (11): 747–50. https://doi.org/10.1038/s41422-024-01026-y.

Loh, Po-Ru, Petr Danecek, Pier Francesco Palamara, et al. 2016. “Reference-Based Phasing Using the Haplotype Reference Consortium Panel.” *Nature Genetics* 48 (11): 1443–48. https://doi.org/10.1038/ng.3679.

Loos, Ruth J. F. 2020. “15 Years of Genome-Wide Association Studies and No Signs of Slowing Down.” *Nature Communications* 11 (1): 5900. https://doi.org/10.1038/s41467-020-19653-5.

Marino, Adriano De, Abdallah Amr Mahmoud, Madhuchanda Bose, et al. 2022. “A Comparative Analysis of Current Phasing and Imputation Software.” *PLOS ONE* 17 (10): e0260177. https://doi.org/10.1371/journal.pone.0260177.

McKenna, Aaron, Matthew Hanna, Eric Banks, et al. 2010. “The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data.” *Genome Research* 20 (9): 1297–303. https://doi.org/10.1101/gr.107524.110.

Mowlaei, Mohammad Erfan, Chong Li, Junjie Chen, et al. 2023. “Split-Transformer Impute (STI): Genotype Imputation Using a Transformer-Based Model.” Preprint, bioRxiv, March 18. https://doi.org/10.1101/2023.03.05.531190.

Naito, Tatsuhiko, Ken Suzuki, Jun Hirata, et al. 2021. “A Deep Learning Method for HLA Imputation and Trans-Ethnic MHC Fine-Mapping of Type 1 Diabetes.” *Nature Communications* 12 (1): 1639. https://doi.org/10.1038/s41467-021-21975-x.

Nguyen, Eric, Michael Poli, Matthew G. Durrant, et al. 2024. “Sequence Modeling and Design from Molecular to Genome Scale with Evo.” *Science* 386 (6723): eado9336. https://doi.org/10.1126/science.ado9336.

Poplin, Ryan, Pi-Chuan Chang, David Alexander, et al. 2018. “A Universal SNP and Small-Indel Variant Caller Using Deep Neural Networks.” *Nature Biotechnology* 36 (10): 983–87. https://doi.org/10.1038/nbt.4235.

Purcell, Shaun, Benjamin Neale, Kathe Todd-Brown, et al. 2007. “PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses.” *American Journal of Human Genetics* 81 (3): 559–75. https://doi.org/10.1086/519795.

Ramstetter, Monica D, Thomas D Dyer, Donna M Lehman, et al. 2017. “Benchmarking Relatedness Inference Methods with Genome-Wide Data from Thousands of Relatives.” *Genetics* 207 (1): 75–82. https://doi.org/10.1534/genetics.117.1122.

Rubinacci, Simone, Olivier Delaneau, and Jonathan Marchini. 2020. “Genotype Imputation Using the Positional Burrows Wheeler Transform.” *PLOS Genetics* 16 (11): e1009049. https://doi.org/10.1371/journal.pgen.1009049.

Scholtens, Salome, Nynke Smidt, Morris A. Swertz, et al. 2015. “Cohort Profile: LifeLines, a Three-Generation Cohort Study and Biobank.” *International Journal of Epidemiology* 44 (4): 1172–80. https://doi.org/10.1093/ije/dyu229.

Sigurdsson, Arnór I., Ioannis Louloudis, Karina Banasik, et al. 2023. “Deep Integrative Models for Large-Scale Human Genomics.” *Nucleic Acids Research* 51 (12): e67. https://doi.org/10.1093/nar/gkad373.

Snyder, Matthew W., Andrew Adey, Jacob O. Kitzman, and Jay Shendure. 2015. “Haplotype-Resolved Genome Sequencing: Experimental Methods and Applications.” *Nature Reviews Genetics* 16 (6): 344–58. https://doi.org/10.1038/nrg3903.

Sollis, Elliot, Abayomi Mosaku, Ala Abid, et al. 2023. “The NHGRI-EBI GWAS Catalog: Knowledgebase and Deposition Resource.” *Nucleic Acids Research* 51 (D1): D977–85. https://doi.org/10.1093/nar/gkac1010.

Song, Meng, Jonathan Greenbaum, Joseph Luttrell, et al. 2022. “An Autoencoder-Based Deep Learning Method for Genotype Imputation.” *Frontiers in Artificial Intelligence* 5 (November). https://doi.org/10.3389/frai.2022.1028978.

Tanaka, Toshihiro, Yuki Nagata, and Akira Takemoto. 2022. “Integrating Biomedical and Clinical Data with BioBank Japan.” *Nature Cardiovascular Research* 1 (7): 597–98. https://doi.org/10.1038/s44161-022-00105-w.

The 1000 Genomes Project Consortium, Corresponding authors, Adam Auton, et al. 2015. "A Global Reference for Human Genetic Variation." *Nature* 526 (7571): 68–74. https://doi.org/10.1038/nature15393.

The All of Us Research Program Genomics Investigators, Manuscript Writing Group, Alexander G. Bick, et al. 2024. "Genomic Data in the All of Us Research Program." *Nature* 627 (8003): 340–46. https://doi.org/10.1038/s41586-023-06957-x.

The International HapMap Consortium. 2003. "The International HapMap Project." *Nature* 426 (6968): 789–96. https://doi.org/10.1038/nature02168.

Theodoris, Christina V., Ling Xiao, Anant Chopra, et al. 2023. "Transfer Learning Enables Predictions in Network Biology." *Nature* 618 (7965): 7965. https://doi.org/10.1038/s41586-023-06139-9.

Wharrie, Sophie, Zhiyu Yang, Vishnu Raj, et al. 2023. "HAPNEST: Efficient, Large-Scale Generation and Evaluation of Synthetic Datasets for Genotypes and Phenotypes." *Bioinformatics* 39 (9): btad535. https://doi.org/10.1093/bioinformatics/btad535.

Yang, Fan, Wenchuan Wang, Fang Wang, 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): 852–66. https://doi.org/10.1038/s42256-022-00534-z.

Yang, Xin, Siddhartha Kar, Antonis C. Antoniou, and Paul D. P. Pharoah. 2023. "Polygenic Scores in Cancer." *Nature Reviews Cancer* 23 (9): 619–30. https://doi.org/10.1038/s41568-023-00599-x.