BMC Bioinformatics

SOFTWARE

Open Access

A novel simulation tool for low-coverage whole-genome sequencing using multivariate Gaussian mixture models

Kai-Yu Chen[1*], Shang-Fu Chen[1,2], Raquel Dias[3*] and Ali Torkamani[1,2]

*Correspondence:
Kai-Yu Chen
kchen@scripps.edu
Raquel Dias
raquel.dias@ufl.edu
[1]Present address: Department of Integrative Structural and Computational Biology, Scripps Research, La Jolla, CA 92037, USA
[2]Scripps Research Translational Institute, Scripps Research, La Jolla, CA 92037, USA
[3]Department of Microbiology and Cell Science, University of Florida, Gainesville, FL 32611, USA

Abstract

Background Low-coverage whole-genome sequencing (lcWGS) has emerged as a cost-effective and robust approach for population genomic studies. Despite its advantages, publicly available resources for large-scale lcWGS datasets remain limited. To our knowledge, there is yet a bioinformatics tool capable of directly simulating lcWGS datasets from variant call format (VCF) files. To address this gap, we developed a tool called lcSimVCF, which leverages multivariate Gaussian mixture models (MGMMs) to simulate lcWGS genotype likelihood distributions directly from high-coverage whole-genome sequencing (hcWGS) VCF files. In this study, we introduced a tool called lcSimVCF that aims to use 30× hcWGS VCF files to simulate 1× genotype likelihood distributions as a demonstration.

Results We trained an MGMM framework for three genotypes: homozygous reference, heterozygous, and homozygous non-reference, each simulating its respective lcWGS genotype likelihood distribution. Our results demonstrate the robustness of the MGMM approach in capturing genotype likelihood distributions compared to a single Gaussian model (SGM). Simulated lcWGS data exhibited representative patterns in population stratification analyses and showed potential for applications in polygenic risk score (PRS) modeling. Analysis of 81,271,745 imputed single nucleotide polymorphisms (SNPs) revealed strong correlations among the PRS derived from different phenotypes: PRS_{CAD}, PRS_{T2D}, and PRS_{AF}. The correlations demonstrated high coefficients of determination (R^2) of 0.94, 0.87, and 0.85, respectively. Furthermore, we assessed simulation time across various configurations (1-16 CPUs, 1-800 individuals, 100-10,000 variants), finding its capability in simulating 10,000 variants across 800 individuals in under 30 seconds with 16 CPUs.

Conclusions The developed simulation tool has demonstrated its capability in generating large-scale lcWGS datasets. This tool holds significant potential to facilitate the development and evaluation of bioinformatics tools and analytical pipelines that rely on access to extensive lcWGS data resources.

Keywords open source, Python, multivariate Gaussian mixture models, low-coverage whole-genome sequencing, genome-wide association studies (GWAS)

Introduction

Coupled with genotype imputation, lcWGS has emerged as a cost-effective approach, offering an alternative to hcWGS and genotyping arrays for conducting genome-wide association studies (GWAS) [1–5]. lcWGS enables analysis across a wider range of genomic regions than genotyping arrays, allowing for the discovery of novel and rare variants. Furthermore, imputed lcWGS data has found applications in various downstream analysis [6–8]. However, the development of robust inference models for population genetics analyses, such as PRS analysis, population stratification studies for ancestry adjustment, and GWAS, requires large-scale lcWGS datasets. Despite the advantages of lcWGS, generating large-scale lcWGS datasets remains laborious, time-consuming, and storage-intensive because intermediate files such as mapped Binary Alignment Map (BAM) files must be retained. Moreover, unlike hcWGS, public repositories provide only limited lcWGS VCF resources. Converting existing public data into lcWGS VCFs requires several preprocessing stages: downloading BAM, Compressed Reference-oriented Alignment Map (CRAM), or FASTQ files; downsampling; performing quality control measures such as deduplication and base quality score recalibration (BQSR); and finally variant calling. Each step demands substantial computation and additional storage capacity. These constraints underscore the need for simulation methods capable of generating large-scale lcWGS datasets in a user-friendly output format, enabling researchers to mitigate data scarcity and to develop and validate robust models in population genetics. Several simulation tools have been developed to address the limitations of available sequencing data resources and facilitate the evaluation of statistical methods as well as bioinformatics algorithm development [9–15]. Yet, there is currently no tool specifically designed to directly simulate genotype likelihood distributions for the purpose of generating lcWGS VCF files from existing hcWGS VCF files. To address this gap, we propose leveraging MGMMs to simulate lcWGS genotype likelihood distributions. Gaussian mixture models (GMM) have been proposed to simulate probability distributions by representing the overall distribution as a combination of multiple Gaussian components [16]. This allows GMMs to capture complex, multimodal data distributions by adjusting the weights, means, and covariances of the individual Gaussian components. When applied to multivariate data, GMMs become MGMMs, which are recognized as robust statistical methods for clustering and density estimation in various fields. In the fields of bioinformatics and genomics, MGMMs have been utilized for various applications such as capturing gene expression patterns [17], clustering GWAS metadata with high missingness rate [18], estimating population stratification[19], and discovering patterns in human chromatin structure [20]. Here, we leveraged MGMMs to capture heterogeneous genotype likelihood information from lcWGS data and provide a new strategy for lcWGS simulation. This tool can simulate representative lcWGS genotype likelihood distributions that closely resemble ground-truth lcWGS. We demonstrated that our simulated data has potential for downstream analysis and applications, including population stratification analysis [21, 22] and polygenic risk score calculation [23–25].

Methods

Dataset: New York Genome Center 1000 Genomes Project data

The 1000 Genomes Project (1KGP) [26] aimed to provide a comprehensive overview of common human genetic variation by performing whole-genome sequencing (WGS) on a diverse group of individuals representing multiple populations. Previously, the New York Genome Center (NYGC) re-sequenced all 2504 samples at 30× depth of coverage. The re-sequencing was conducted on an Illumina NovaSeq 6000 system with 2 ×150 bp reads, aligned to the GRCh38 reference genome. In our work, we used 36 individuals of African and European ancestry for the purpose of MGMMs training and 100 sequenced sample data containing five global populations: East Asians (EAS), Europeans (EUR), Africans (AFR), South Asians (SAS), and Admixed Americans (AMR). We also used NA12878 WGS data to demonstrate our simulation results.

Downsampling

SAMtools (v.1.10) was used to downsample the selected 30× mapped 1KGP genomes (n = 100) and the sample NA12878 genome randomly at a coverage level of 1×. We performed downsampling using the command "samtools view-T ${ref_genome}-s 1. 0333-bo ${output_ bam} ${input_bam}", where we used the human genome reference genome (v.hg38) as ${ref_genome}. The downsampled BAM files were indexed with command "samtools index ${output_ bam}".

Reference panels

We used haplotypes derived from 3202 individuals across 26 ancestry populations re-sequenced by NYGC at depth of 30× from 1000 Genomes Project phase (GRCh38) as reference panel. We downloaded the reference panel from (http ://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/20201028_3202_phased/CCDG_14151_B01_GRM_WGS_2020-08-05_${chroms}.filtered.shapeit2-duohmm-phased.vcf.gz{,.tbi}). We then extracted biallelic variant sites from the downloaded reference panel using the command "bcftools view-m 2-M 2-v snps -Oz-o ${output_vcf} ${input_vcf}" and indexed the output VCF files by using "tabix-p vcf ${output_vcf}" to create the final reference panel for this study.

Genotype likelihoods calculation and variant extraction

First, to call genotypes on lcWGS and hcWGS NYGC 1KGP sequencing data, we extracted 61,715,567 biallelic variable positions from the NYGC 1KGP reference panel, including the reported reference (REF) and alternative (ALT) alleles. Then, we ran "bcftools mpileup" and "bcftools call" commands on the CRAM/BAM files to extract these biallelic variable positions and performed genotype calling. Genotype likelihoods were calculated by default during the "bcftools mpileup" step and were passed directly to "bcftools call".

Genotype imputation and phasing

We performed genotype imputation and phasing using GLIMPSE v1.1.0 with default settings using NYGC 1KGP as reference panel with a total of 61,715,567 biallelic variant sites. First, we used the GLIMPSE_chunk tool to split the chromosomes into chunks with commands "--window-size 2,000,000" and "-buffer-size 200,000". Next, we applied

the GLIMPSE_phase tool to perform phasing and imputation by refining genotype likelihoods. This was followed by the GLIMPSE_ligate tool to merge imputed chunks into complete chromosome sequences. Lastly, we used GLIMPSE_sample tool to generate haplotype calls.

Genotype likelihood distributions and allele presence probability calculations

A VCF file is a commonly used standard text file format in bioinformatics that stores genetic variations such as SNPs and indels [27], and is widely used in next-generation sequencing analysis. For each genotype derived from hcWGS VCF files, we trained MGMMs to learn noise patterns from the corresponding variant position using lcWGS genotype likelihood information. This process involved extracting genotype and Phred-scaled likelihood information from 30× (ground-truth) and 1× coverage VCF files, respectively. Phred-scaled genotype likelihood is a three-dimensional vector $PL_{i,j}[g_{i,j}]$ that can be converted and normalized to genotype likelihoods as described [28] (1):

$$GL_{i,j} = \frac{10^{-\frac{PL_{i,j}[g_{i,j}]}{10}}}{\sum_{g_{i,j}} 10^{-\frac{PL_{i,j}[g_{i,j}]}{10}}} \tag{1}$$

A genotype likelihood refers to the likelihood of observing the sequencing reads [29] at a specific genomic location. In the absence of available reads, the likelihood distribution remains uniform, but it converges towards a specific genotype as the number of supporting reads for that genotype increases. In this work, the genotype likelihoods for individual i were calculated across filtered variant position *j* at sequencing read D. This is denoted by $g_{ij} \in \{rr, ra, aa\}$, which indicates homozygous reference, heterozygous, homozygous non-reference genotypes. The allele presence probabilities of the REF allele $\Pr(r_{ij})$ and the ALT allele $\Pr(a_{ij})$ were calculated from the given genotypes (2 and 3):

$$\Pr(r_{ij}) = GL(g_{ij} = rr|D) + GL(g_{ij} = ra|D) \tag{2}$$

$$\Pr(a_{ij}) = GL(g_{ij} = aa|D) + GL(g_{ij} = ra|D) \tag{3}$$

To convert the allele presence probability distribution back into genotype likelihoods and express them in dosage format compatible with VCF file fields, we applied the following conversion formula (4–6):

$$GL(g_{ij} = 0|D) = \Pr(r_{ij}) * (1 - \Pr(a_{ij})) \tag{4}$$

$$GL(g_{ij} = 1|D) = \Pr(r_{ij}) * \Pr(a_{ij}) \tag{5}$$

$$GL(g_{ij} = 2|D) = (1 - \Pr(r_{ij})) * \Pr(a_{ij}) \tag{6}$$

Overview of MGMMs and model training

lcWGS data often lacks sufficient sequencing depth for unambiguous genotype determination. To address this limitation, we employed allele presence probability distributions to characterize the inherent uncertainty in lcWGS genotypes. This probabilistic representation captures lcWGS allele presence uncertainty without requiring a genotype call. We captured lcWGS allele presence probability distributions by first converting genotype likelihood information to allele presence probabilities using formulas (2) and (3).

We then used the converted lcWGS allele presence information to train a MGMM with K components for each hcWGS genotype. We defined lcWGS allele presence probability distributions as an $m \times 2$ matrix denoted by X, where m represents the total number of variants across all input sample individuals. We used 36 individuals of AFR and EUR ancestry for model training. Each row $[x_{i1}, x_{i2}]$ in matrix X indicates the presence probabilities for the REF allele and the ALT allele at the i-th variant site, with each allele pair treated as mutually independent observations. Each MGMM component k (k = 1, 2..., K) comprises a two-dimensional multivariate Gaussian component parameterized by a mean vector μ_k and a 2×2 covariance matrix Σ_k. In this study, biallelic variants were used to train the model. The MGMMs simulate the overall distribution of the $m \times 2$ data as a mixture of these k two-dimensional Gaussian components. The probability density function of the MGMMs is expressed as a weighted linear combination of K Gaussian components (7):

$$f(X) = \sum_{k=1}^{K} \pi_k \mathcal{N}(X|\mu_k, \Sigma_k) \tag{7}$$

where:

- X is the two-dimensional allele presence probability vector $[x_{i1}, x_{i2}]$ composed of the REF allele and the ALT allele at the i-th variant site.
- π_k is the weight of component k in the mixture models that satisfy the constraint $\sum_{k=1}^{K} \pi_k = 1$ with the condition $\pi_k \geq 0$.

$\mathcal{N}(X|\mu_k, \Sigma_k)$ is the multivariate Gaussian density with mean μ_k and covariance Σ_k evaluated at X denoted by (8):

$$\mathcal{N}(X|\mu_k, \Sigma_k) = \frac{1}{\sqrt{2\pi^d |\Sigma_k|}} exp\left(-\frac{1}{2}(\mathrm{X} - \mu_k)^T \Sigma_k^{-1} (\mathrm{X} - \mu_k)\right) \tag{8}$$

where:

- $d = 2$ represents the dimensionality of the data (REF allele and ALT allele).
- $\mu_k = [\mu_{REF,k}, \mu_{ALT,k}]^{\mathrm{T}}$ is the mean vector of the k-th Gaussian component.
- Σ_k is a 2×2 covariance matrix of the k-th Gaussian component.
- $|\Sigma_k|$ denotes the determinant of the covariance matrix of the k-th Gaussian component.

The complete parameter set $\theta = (\pi_1, \mu_1, \Sigma_1, \pi_2, \mu_2, \Sigma_2 \ldots, \pi_k, \mu_k, \Sigma_k)$ was estimated using the expectation–maximization (EM) algorithm to obtain maximum likelihood estimates [30]. The EM process begins by initializing the algorithm, where existing sample coordinates are randomly assigned as initial centroids, denoted as θ^0. During the initialization step, the weight π_k for each component is assigned to $1/K$. The EM algorithm involves alternation between two steps: the expectation step (E-step) and the maximization step (M-step). During the E-step, based on the current parameter values θ, the algorithm calculates the conditional probabilities that each allele presence probability set $[x_{i1}, x_{i2}]$ belongs to component k. During the M-step, the parameters θ are updated by maximizing the likelihood using the conditional probabilities derived in the E-step. These two steps alternate until the change in parameter estimates falls below a

predefined tolerance threshold. MGMM models were trained to simulate the allele presence probabilities for the REF and ALT alleles in lcWGS for each genotype class (9):

$$\left(\Pr\left(r_{i,j}\right), \Pr\left(a_{i,j}\right)\right) \sim \sum_{k=1}^{K} \pi_k f\left(\mu_k, \Sigma_k\right) \tag{9}$$

We used the sklearn.mixture package [31] in Python [32] to fit MGMMs. It assessed the two-dimensional allele presence probability distributions of each variant as a multivariate distribution across the dataset. We performed EM optimization on these Gaussian distributions using the built-in algorithm in sklearn.mixture.GMM, with a tolerance of 1×10^{-8} as the convergence threshold. All other hyperparameters were set to their default values for training the MGMMs.

For each genotype class in hcWGS, the corresponding lcWGS allele presence probability distribution was modeled using the maximum likelihood estimate $\ln\left(\widehat{L}\right)$ computed from the observed lcWGS allele presence probability distributions (10):

$$\hat{L} = \prod_{i=1}^{N} \left(\sum_{k=1}^{K} \pi_k \mathcal{N}\left(X|\mu_k, \Sigma_k\right) \right) \tag{10}$$

We utilized the Bayesian information criterion (BIC) for model selection to determine the number of components K for a MGMM for each genotype class [33, 34] (11):

$$BIC \equiv -2 \cdot \ln\left(\widehat{L}\right) + c \cdot \ln\left(N\right) \tag{11}$$

where c represents the number of estimated parameters, N denotes the number of variants.

Overview of data preparation and simulation workflow

The computational workflow of lcSimVCF is shown in Fig. 1a. The simulation pipeline requires a hcWGS VCF file as input to generate the corresponding lcWGS simulation. To prepare the data for model training, we downloaded and downsampled 36 samples from the NYGC re-sequenced dataset, followed by variant calling. Next, we performed a liftover on both hcWGS and lcWGS sequencing data from the GRCh38 to GRCh37 reference genome build. Genotype and genotype likelihood information were extracted from ground-truth hcWGS and lcWGS VCF files, respectively, using the scikit-allel Python module, specifically assessing the "calldata/GT" field for genotypes and "calldata/PL" field for Phred-scaled genotype likelihoods. For each genotype class, an MGMM model was trained to learn patterns of lcWGS allele presence probabilities from hcWGS. The optimal number of components for each model was determined by BIC. The trained MGMMs were subsequently employed to simulate lcWGS allele presence probabilities, which were then further converted to genotype likelihood information as described in Eqs. (4)–(6) for VCF file export (Fig. 1b). To mimic missing genotypes in lcWGS data, we calculated the missing ratio of ground-truth lcWGS compared to 30 × hcWGS VCF files. We noticed that approximately 37% of the variants were absent in lcWGS data. Therefore, 37% of allele presence probability pairs were masked with sentinel values (−1) and recorded as missing genotypes in the output VCF files.

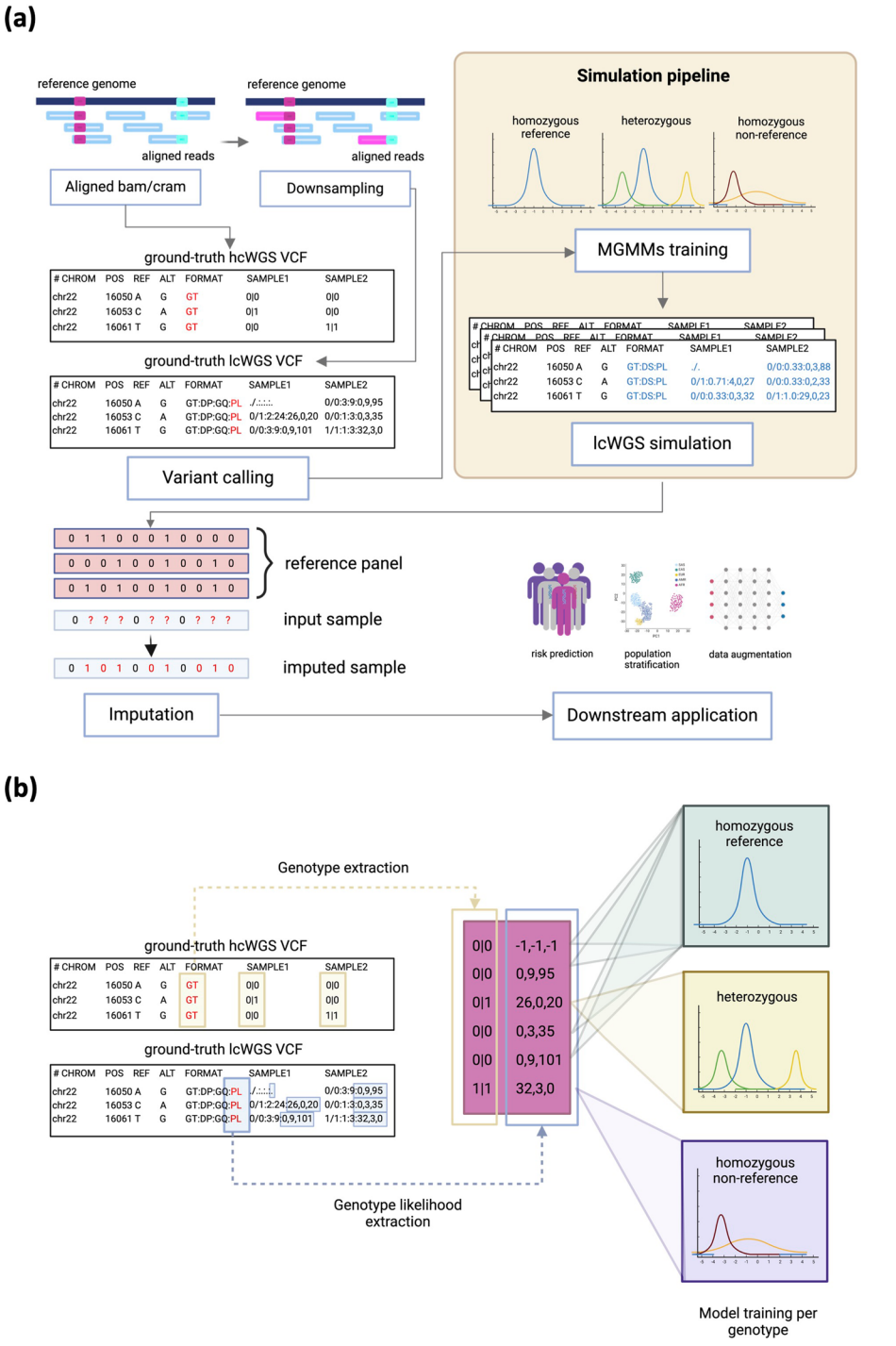

Fig. 1 **a** Overview of data preparation and simulation workflow. Mapped BAM files were downsampled to low-coverage BAM files. Variant calling was conducted on both lcWGS and hcWGS BAM files. Genotypes extracted from hcWGS VCF files and genotype likelihoods extracted from lcWGS VCF files were used to train MGMMs. The trained MGMMs were then used for lcWGS VCF simulation given hcWGS genotype information. The simulated lcWGS VCF files were imputed before downstream applications such as polygenic risk score prediction, population stratification, and data augmentation. At each variant site, alleles were encoded using 0 and 1, where 0 represented the reference allele matching the reference genome and 1 denoted the alternative allele differing from the reference genome. **b** Detailed simulation process using one MGMM per genotype. Genotypes and their associated genotype likelihoods were extracted from hcWGS VCF files and lcWGS files, respectively. Genotype likelihoods were initially represented as Phred-scaled scores, which were then converted into allele presence probabilities. For each genotype class, an MGMM was trained to learn intrinsic patterns within the genotype likelihood distributions from the lcWGS data

Our simulation pipeline transforms input hcWGS VCF files into corresponding lcWGS VCF files while maintaining structural consistency and format compatibility. The simulation framework employs MGMMs to model the relationship between hcWGS genotypes and lcWGS genotype likelihood distributions. Output VCF files contain standardized fields including genotype calls ('FORMAT/GT'), allelic dosage ('FORMAT/DS'), and Phred-scaled genotype likelihoods ('FORMAT/PL'), with results recorded in the 'FORMAT/GT' field. We implemented the simulation tool in a Python3.8 environment.

Results

Comparison of allele presence probability distribution between simulated 1×WGS and ground-truth 1×WGS

The selected NYGC re-sequenced 1KGP data, which included samples for model training data and NA12878, were downloaded as CRAM files from The International Genome Sample Resource (IGSR) via FTP site (https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/). The SAMtools [14] command "samtools views" was used to downsample 30 × CRAM file to 1 × BAM file for the preparation of lcWGS data. We marked and removed duplicative reads [35, 36] using LocusCollector algorithm and Dedup algorithm in Sentieon® Genomics software [37, 38]. Base quality score recalibration (BQSR) was performed with the QualCal algorithm to generate recalibration tables, after which variants were called with the Haplotyper algorithm [39]. During the variant calling step, we applied default parameters for 30x WGS data and specified --emit_mode all, --emit_conf 0 for 1 × WGS. All biallelic variable positions in the 1KGP reference panel were extracted, along with their reported REF and ALT alleles. We then conducted liftover on called variants from human reference genome [40, 41] assembly GRCh38 [42] to human reference genome assembly GRCh37 [43–45] on both 1× WGS and 30× WGS using Picard (Institute 2019) LiftoverVcf. We developed a MGMM for each genotype class based on a randomly selected sample of 100,000 variants from an admixed population of 36 AFR and EUR descents in the NYGC re-sequenced 1KGP dataset (Table S1). SGM has been widely used for density distribution profiling due to its simplicity and effectiveness. Alternatively, MGMMs are often employed as a more flexible approach, where the density distribution is modeled using a combination of multiple Gaussian components [46, 47]. Therefore, we established an SGM as a baseline to determine whether the added complexity of MGMMs provided significant improvement. The mean vectors and covariance matrices that modeled the two-dimensional allele presence probability distributions are shown in Table 1. SGM used a single component to

Table 1 Optimal parameters and component metrics for the SGM

Gaussian component	Mean	Covariance matrix	Weights	Angle	Bhattacharyya distance (BD)
Homozygous reference					
1	[0.9945, 0.2618]	$\begin{bmatrix} 0.0012 & -0.0012 \\ -0.0012 & 0.0095 \end{bmatrix}$	1.0	98.25°	0.156
Heterozygous					
1	[0.7426, 0.7226]	$\begin{bmatrix} 0.1156 & -0.0706 \\ -0.0706 & 0.1195 \end{bmatrix}$	1.0	134.22°	0.0031
Homozygous non-reference					
1	[0.2634, 0.9974]	$\begin{bmatrix} 0.0099 & -0.0007 \\ -0.0007 & 0.0006 \end{bmatrix}$	1.0	175.51°	0.183

capture the two-dimensional allele presence probability distribution. For MGMMs, we used BIC to determine the optimal number of Gaussian components for model selection as shown in Fig. 2. The optimal numbers of Gaussian components for homozygous reference, heterozygous, and homozygous non-reference genotypes were 7, 6, and 4, with BIC values of - 213,054.64, - 185,821.192, and - 207,949.01 respectively (Table 2). We evaluated the simulated allele presence probability distribution on chromosome 22 and on the remaining chromosomes. This evaluation involved calculating the Bhattacharyya distances (BD) between SGM-simulated 1 × WGS and ground-truth 1 × WGS, as well as between MGMM-simulated 1 × WGS and ground-truth 1 × WGS, without masking or excluding missing data points [48]. In chromosome 22, the BD values between SGM-simulated 1 × WGS and ground-truth 1 × WGS for each genotype were 0.156, 0.0031, and 0.183, respectively. In contrast, the BD values between MGMM-simulated 1x WGS and ground-truth 1 × WGS were substantially lower at 0.004, 0.0023, and 0.0046, respectively (Fig. 3; Figs. S1 and S2). These results suggest that MGMM-simulated 1 × WGS data exhibited higher probability distribution similarity to ground-truth 1 × WGS data compared to SGM-simulated 1 × WGS data, demonstrating the superior ability of MGMMs to capture the complex nature of genotype likelihood distributions.

Population stratification

Population genomic analysis allows detection of allele frequency variations [49] driven by evolutionary mechanisms such as natural selection, population migration, and

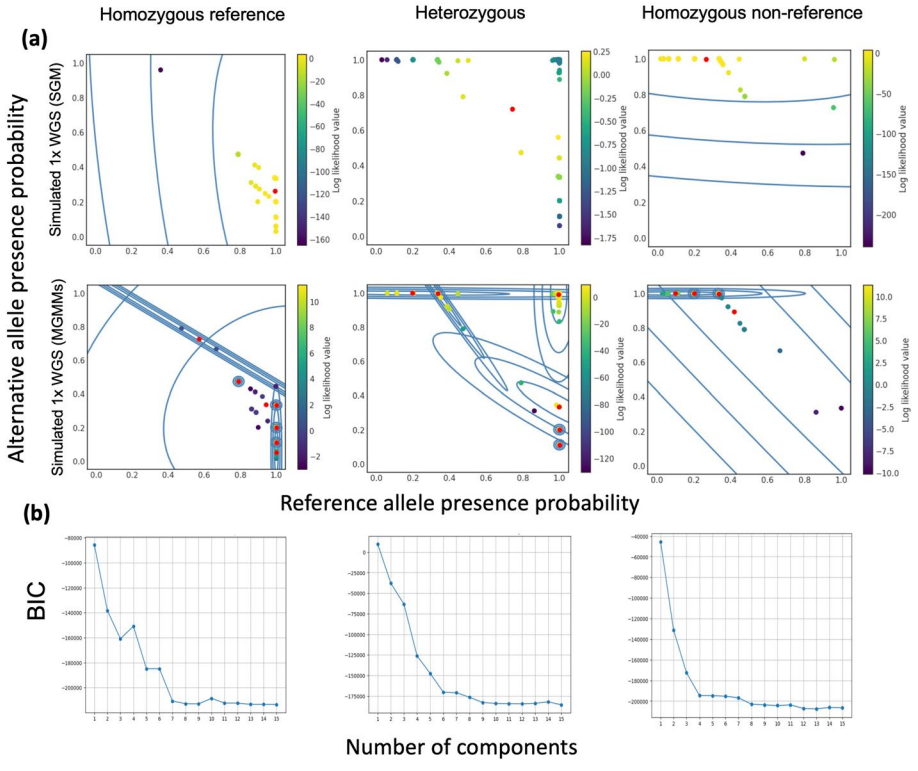

Fig. 2 MGMMs training results for each genotype. **a** 36 samples representing EUR and AFR subpopulations from the 1KGP cohort for training SGMs and MGMMs. We trained models for three genotypes: homozygous reference, heterozygous, and homozygous non-reference. The position and orientation of the Gaussian variance of each cluster is indicated with a blue circle. The cluster centroids are indicated with red dots. **b** The optimal number of components for a MGMM was determined by BIC

Table 2 Optimal parameters and component metrics for the MGMMs

Gaussian component	Mean	Covariance matrix	Weights	Angle	Bhattacharyya distance
Homozygous reference					
1	[0.9999,0.2008]	$\begin{bmatrix} 1.052 \times 10^{-6} & -4.166 \times 10^{-8} \\ -4.166 \times 10^{-8} & 1.033 \times 10^{-6} \end{bmatrix}$	0.296	141.37 °	0.004
2	[0.9993,0.3343]	$\begin{bmatrix} 1.065 \times 10^{-6} & -4.352 \times 10^{-8} \\ -4.352 \times 10^{-8} & 1.029 \times 10^{-6} \end{bmatrix}$	0.532	146.33 °	
3	[0.9999,0.0519]	$\begin{bmatrix} 1.005 \times 10^{-6} & -9.983 \times 10^{-8} \\ -9.983 \times 10^{-8} & 1.764 \times 10^{-4} \end{bmatrix}$	0.037	127.63 °	
4	[0.57, 0.7285]	$\begin{bmatrix} 0.0093 & -0.0059 \\ -0.0059 & 0.0038 \end{bmatrix}$	0.002	90.03 °	
5	[0.7904,0.4734]	$\begin{bmatrix} 1.000 \times 10^{-6} & 6.163 \times 10^{-33} \\ 6.163 \times 10^{-33} & 1.000 \times 10^{-6} \end{bmatrix}$	0.0139	147.36°	
6	[0.9999,0.1118]	$\begin{bmatrix} 1.0037 \times 10^{-6} & -3.3056 \times 10^{-9} \\ -3.3056 \times 10^{-9} & 1.0029 \times 10^{-6} \end{bmatrix}$	0.096	138.39°	
7	[0.94, 0.3365]	$\begin{bmatrix} 0.0032 & 0.0008 \\ 0.0008 & 0.0034 \end{bmatrix}$	0.022	-130.72°	
Heterozygous					
1	[0.9957, 0.9929]	$\begin{bmatrix} 2.02 \times 10^{-5} & -5.23 \times 10^{-7} \\ -5.23 \times 10^{-7} & 2.93 \times 10^{-4} \end{bmatrix}$	0.262	90.11°	0.0023
2	[0.3358, 0.997]	$\begin{bmatrix} 1.6 \times 10^{-4} & -2.39 \times 10^{-4} \\ -2.39 \times 10^{-4} & 3.6 \times 10^{-4} \end{bmatrix}$	0.282	123.67°	
3	[0.9964, 0.3357]	$\begin{bmatrix} 5.25 \times 10^{-4} & -2.9 \times 10^{-4} \\ -2.9 \times 10^{-4} & 2.07 \times 10^{-4} \end{bmatrix}$	0.282	149.17°	
4	[0.196, 0.9998]	$\begin{bmatrix} 2.8 \times 10^{-3} & -2.57 \times 10^{-5} \\ -2.57 \times 10^{-5} & 1.45 \times 10^{-6} \end{bmatrix}$	0.084	179.48°	
5	[0.9999, 0.1118]	$\begin{bmatrix} 1.00 \times 10^{-6} & -5.49 \times 10^{-16} \\ -5.49 \times 10^{-16} & 1.00 \times 10^{-6} \end{bmatrix}$	0.019	138.39°	
6	[0.9999, 0.2007]	$\begin{bmatrix} 1.04 \times 10^{-6} & -2.8 \times 10^{-8} \\ -2.8 \times 10^{-8} & 1.02 \times 10^{-6} \end{bmatrix}$	0.071	141.37°	
Homozygous non-reference					
1	[0.3343, 0.9993]	$\begin{bmatrix} 1.014 \times 10^{-6} & -2.16 \times 10^{-8} \\ -2.16 \times 10^{-8} & 1.032 \times 10^{-6} \end{bmatrix}$	0.549	123.67°	0.0046
2	[0.2007, 0.9999]	$\begin{bmatrix} 1.004 \times 10^{-6} & -5.56 \times 10^{-9} \\ -5.56 \times 10^{-9} & 1.01 \times 10^{-6} \end{bmatrix}$	0.31	128.63°	
3	[0.4191, 0.8942]	$\begin{bmatrix} 0.0264 & -0.029 \\ -0.029 & 0.0334 \end{bmatrix}$	0.027	131.59°	
4	[0.0988, 0.9999]	$\begin{bmatrix} 5.53 \times 10^{-4} & -3.12 \times 10^{-10} \\ -3.12 \times 10^{-10} & 1.000 \times 10^{-6} \end{bmatrix}$	0.114	180°	

genetic drift [50–52]. Over time, these processes lead to accumulated differences in allele frequencies between populations, resulting in distinguishable genetic patterns, a phenomenon known as population stratification [53, 54]. Understanding population stratification is crucial for capturing genetic traits and reducing confounding effects in GWAS [55–57]. The motivation to infer ancestry differences among individuals from various populations [58, 59] or to identify population structure stems from several applications, including population genetics, GWAS [60], and prediction of specific traits for individuals [61–63]. Until recently, stratifying individuals into populations was feasible using only a few markers. However, advancements in technology and decreasing costs now make it possible to use whole genomes for these analyses [64–66]. This allows for highly detailed population analyses based on thousands of markers [67–69]. To demonstrate our simulation tool's performance on population stratification analysis, we followed a protocol similar to that used for sample NA12878. We downloaded selected NYGC

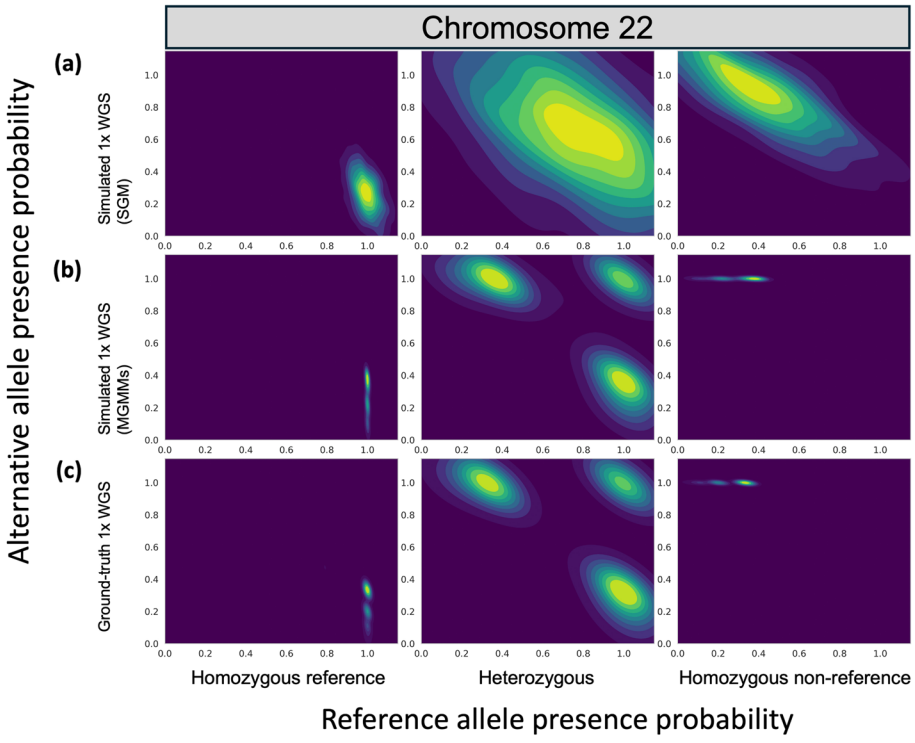

Fig. 3 Evaluation results by genotype on NA12878 chromosome 22. The GMM randomly selected 5000 variants for each genotype class at the corresponding chromosome 22 positions of NA12878 from **a** SGM-simulated data, **b** MGMM-simulated data, and **c** ground-truth lx WGS data

re-sequenced 30 × CRAM files from 100 samples from IGSR, performed downsampling and variant calling using "samtools view -s" and "bcftools mpileup", and conducted coordinate liftover from GRCh38 to GRCh37. For genotype imputation, we used the 1KGP Phase 3 public dataset on GRCh37 as the reference panel, which includes haplotypes across 81,271,745 biallelic variant sites. Both the 30 × VCF files and 1 × VCF files were imputed using default parameters in GLIMPSE v1.1.0. Target samples were excluded from the reference panel during the imputation process. Principal components analysis (PCA) represents the most widely recognized method for population structure identification [70, 71]. PLINK v1.9 [72] was applied for PCA projection, with results visualized using R4.3.0 [73]. To quantitatively evaluate population clustering performance, we employed the adjusted Rand index (ARI) [74], a semi-supervised spectral clustering metric [75]. Our analysis achieved an ARI of 1.0 for post-imputation simulated 1 × WGS with a Euclidean distance of 0.02841 compared to the corresponding 30 × gold standard WGS dataset. PC1(principal component 1) and PC2(principal component 2) captured the two major sources of genetic variation within our dataset. Two-dimensional cluster representation of genetic variation was achieved by plotting PC1 against PC2, where distinct clusters indicated genetic similarity and provided insights into population structure and relationships. Our results suggested that simulated 1 × WGS data can produce population stratification patterns representative of those derived from ground-truth 30 × gold standard WGS dataset following genotype imputation (Fig. 4).

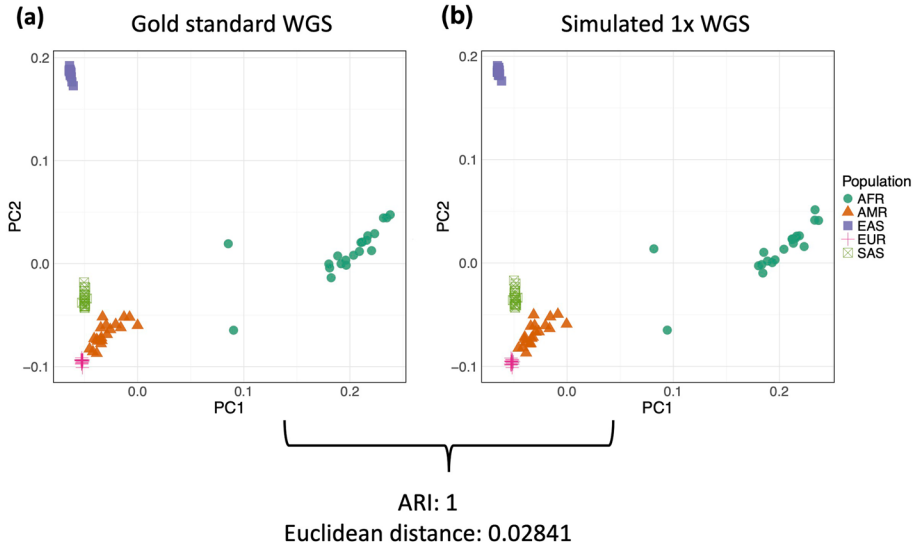

Fig. 4 Comparison of simulated 1 × WGS with 30 × gold standard WGS in population stratification. To demonstrate the utility of our tool, we conducted population stratification analysis on a WGS dataset from the NYGC re-sequenced 1KGP dataset. This dataset comprises genotypes from 100 individuals sampled across five super-ancestry groups. The two panels show the projection of the first two principal components derived from 1KGP samples (n = 100) for **a** ground-truth 30 × gold standard WGS and **b** post-imputation simulated 1 × WGS. Observations are colored according to super-ancestry population

PRS analysis

PRSs are essential for assessing genetic predisposition to complex diseases, as they aggregate data from multiple risk variants pinpointed in GWAS [6, 76–80]. To assess the technical reproducibility of using simulated lcWGS versus 30 × WGS-based data in PRS analysis, we simulated 100 lcWGS samples selected from the 1KGP phase 3 dataset using MGMMs. We calculated PRSs as the sum of the N genome-wide genotype dosage X_i weighted by their corresponding effect sizes $\widehat{\beta}$ derived from GWAS summary statistics (12):

$$PRS = \sum_{i=1}^{N} \hat{\beta}_i \cdot X_i \tag{12}$$

We imputed simulated lcWGS samples through GLIMPSE v1.1.0 [28] using 1KGP phase 3 reference panel. We excluded the target sample from the reference panel during the imputation process. We evaluated the variability of previously published PRSs for three phenotypes: atrial fibrillation (AF) [83], coronary artery disease (CAD) [81], and type 2 diabetes (T2D) [82].These PRSs were obtained through various scoring methodologies detailed in the PGS Catalog [84]. The CAD, T2D, and AF risk scores incorporated 168, 403, and 6,730,542 SNPs, respectively. PRSs derived from simulated lcWGS samples were compared with those derived from corresponding 30 × WGS-based data. Statistical analyses were conducted in R v4.0.0. Across 81,271,745 imputed loci, we found that PRS_{CAD}, PRS_{T2D}, and PRS_{AF} were highly correlated (Fig. 5a–c), with R^2 of 0.94, 0.87, and 0.85, respectively. These results demonstrate that simulated lcWGS has the potential to enable accurate PRS calculation compared to 30 × WGS-based data.

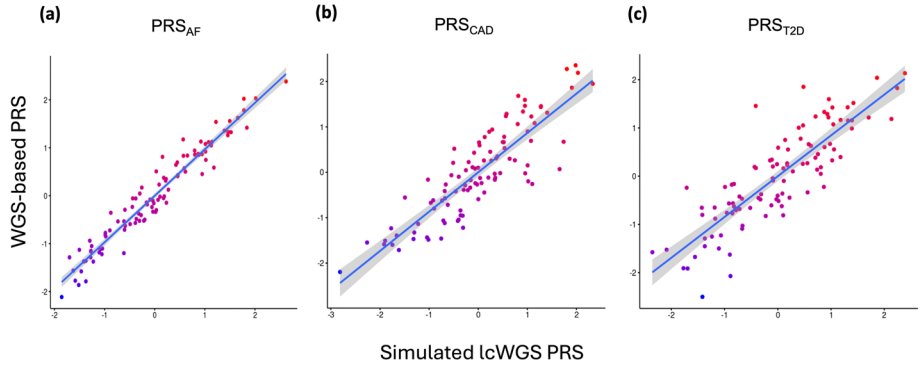

Fig. 5 Correlation of PRSs between the simulated 1×lcWGS post-imputation and the 30×WGS-based data in 100 selected 1KGP samples. After genotype imputation using GLIMPSE v1, we calculated R^2 between PRSs derived from simulated lcWGS data and 30×WGS-based data using 100 selected samples collected from 1KGP phase 3 dataset

Computational performance

We evaluated the total runtime of the simulation tool on a processor with 30 GB RAM using different CPU configurations (1 and 16 cores). We tested simulation performance across varying numbers of variants (100, 500, 1000, 5000, and 10,000) and individuals (1, 10, 50, 100, 200, 400, and 800). Including the simulator's initialization time, we simulated a genomic region of approximately 130 kbp containing 10,000 variants across 800 individuals. The simulation completed in less than 30 seconds using 16 CPUs and in 150 seconds using a single CPU. In addition, we performed whole-genome simulations on sample NA12878 using a per-chromosome approach. We observed that all chromosomes were processed within 1700 seconds using 16 CPUs and 2600 seconds using a single CPU. Since chromosome 1 represents the computational bottleneck due to its size, we conducted scalability testing with expanded RAM (120 GB) using the same CPU configurations. We simulated the entire chromosome 1 with varying sample sizes of 1, 10, 50, and 100 individuals. The chromosome 1 simulation using 16 CPUs demonstrated significant computational efficiency. The process included input data reading, allele presence probability conversion to genotype likelihoods, and VCF file export. The total time complexity of the simulation is $\mathrm{O}(n \cdot v)$, where n is the number of samples and v is the number of variants (Fig. 6a–d).

Discussion

We designed a lcWGS simulation tool lcSimVCF to facilitate development of large-scale machine learning models, PRS calculations, ancestry analysis methods, and other lcWGS-related bioinformatics pipelines. lcWGS has emerged as a cost-effective strategy, serving as an alternative to genotype arrays while preserving whole-genome information [4, 85–87]. However, current methods for generating lcWGS data are time-consuming and laborious. We developed lcSimVCF to enable the simulation of large-scale lcWGS data from hcWGS data. Genotype likelihoods preserve haplotype information in a probabilistic manner [88], which is particularly advantageous for low-coverage whole-genome sequencing data. Due to the inherent noise and uncertainty associated with low-coverage sequencing, employing a probabilistic approach to capture genotype information is more effective compared to non-probabilistic methods that may fail to account for the stochastic nature of such data [86, 89, 90]. MGMMs are a robust probabilistic

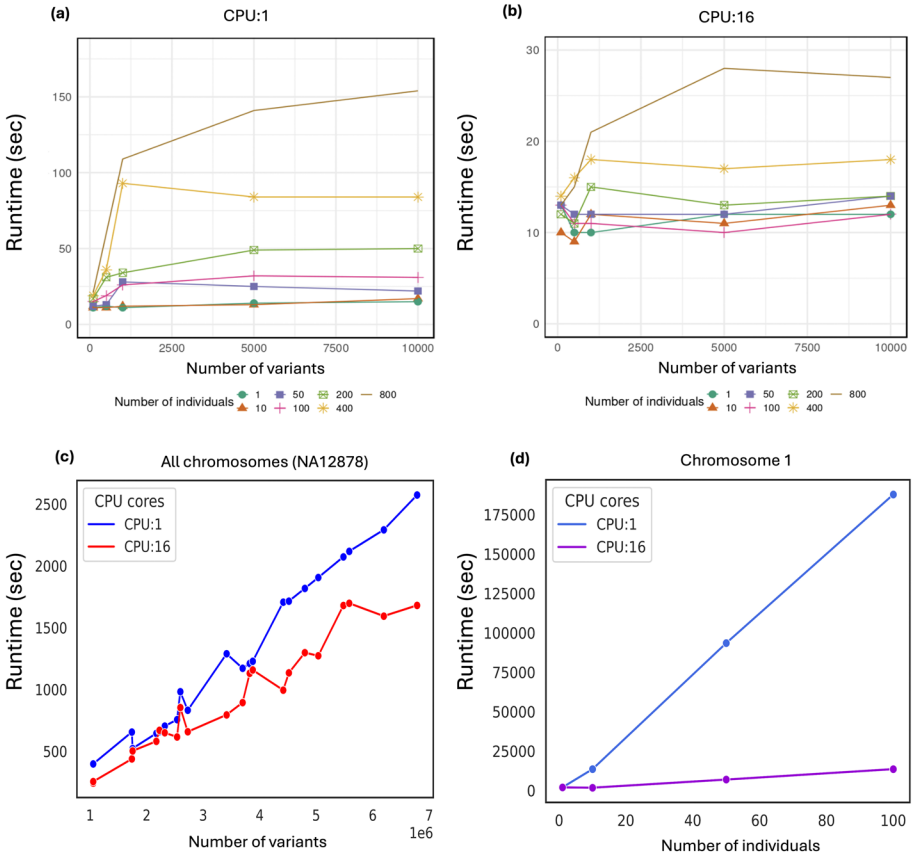

Fig. 6 Runtime assessment. Runtime performance after simulating specified numbers of individuals and variants, across two distinct computational configurations: **a** a single CPU and **b** a 16-CPU system, including initialization time for the simulation process. **c** Runtime across all chromosomes for sample NA12878 conducted on both CPU core environment. **d** Computational scalability testing performed on chromosome 1

modeling approach that can effectively capture complex, multimodal data distributions by representing them as a mixture of multiple Gaussian component distributions. This flexibility renders MGMMs particularly advantageous for density estimation tasks, especially when dealing with heterogeneous datasets exhibiting distinct subpopulations or clusters that cannot be adequately modeled by a single Gaussian distribution. MGMMs have found widespread applications in genomics, including estimating the polygenicity and discoverability of phenotypic traits [91], characterizing the nature of genotype-trait associations [92], clustering gene expression data [17, 93], and identifying key transcriptomic regulators [94, 95]. Therefore, the MGMMs framework presents itself as a suitable methodological approach for estimating genotype likelihoods from lcWGS data.

MGMMs in capturing lcWGS genotype likelihoods

Our findings demonstrate the robustness of MGMMs in estimating genotype likelihoods [95] represented as allele presence probability distributions. MGMMs outperformed the benchmark SGM, reflecting their ability to effectively capture the heterogeneous [96] and potentially multimodal nature of genotype likelihood distributions. During the training of MGMMs, we observed that the homozygous reference genotype required more components than the homozygous non-reference genotype to accurately capture patterns of allele presence probability distributions. This finding suggests that in

homozygous reference genotypes, low coverage and potential sequencing errors may lead to the presence of alternative allele reads, making genotype likelihood calculations less certain. Consequently, additional components are needed to accurately model such noise patterns. Whereas in homozygous non-reference genotypes, even with low coverage, this pattern might be clearer and less affected by noise, requiring fewer components to model.

Linkage disequilibrium is independent of lcWGS genotype likelihood simulation

While variants within linkage disequilibrium (LD) blocks are correlated due to their shared haplotype structures [97], this biological correlation does not translate to dependencies in sequencing noise. Our model is specifically designed to simulate stochastic noise in lcWGS data using hcWGS data as a reference. The noise being modeled reflects technical artifacts, such as sequencing errors and coverage variability, rather than biological processes. Consequently, while LD blocks capture correlations between variants due to inheritance, these correlations do not directly affect the noise characteristics being simulated. Each variant's noise profile primarily depends on local sequencing quality and coverage, which is treated as an independent event in our model. In addition, lcWGS data simulated using this MGMM-based computational tool exhibits utility for applications in population stratification. The simulated lcWGS generated by our developed tool accurately represents the patterns of genetic variation observed among five different superpopulations. It demonstrates concordance with the population structure inferred from ground-truth 30× WGS data.

lcSimVCF shows potential for PRS analysis and reasonable simulation timeframes

PRSs are dynamic estimates that can vary over time due to several key factors, including loci number, genetic architecture, variant weighting, effect size, and underlying genetic model [98–101]. In this analysis, we are able to demonstrate that using lcWGS data simulated from our developed MGMM tool reproduced a significant portion of the variance in PRS results. Furthermore, the simulations can be completed within a reasonable timeframe. Computational efficiency is expected to be enhanced in a multi-core CPU environment, leveraging parallel processing capabilities. Compared to other genetic variant and sequencing data simulators, lcSimVCF stands out as the first tool to offer an efficient and innovative strategy for simulating lcWGS VCF files. It provides a cost-effective approach for obtaining lcWGS SNP markers, which is crucial for enhancing the statistical power of population genetic studies and developing data-greedy models for lcWGS applications. This makes it a valuable resource for researchers.

Limitations and future directions

While the presented simulation approach demonstrates promising capabilities, there are potential limitations. First, the simulated lcWGS VCF files only provide unphased genotype information, even though the tool utilizes genotype likelihoods sampled from the trained MGMMs. To obtain phased genotype data, the application of appropriate imputation or phasing tools would be necessary as a subsequent step [102, 103]. The second limitation is that the trained MGMMs do not account for potential differences in these distributions across varying minor allele frequency (MAF) ranges. The primary objective of lcSimVCF is to simulate allele presence likelihoods and genotype likelihood patterns

in lcWGS data, given a known genotype from hcWGS. As such, the tool is designed to model the probabilistic nature of lcWGS data rather than directly reflect ground-truth information about rare variants. Yet, developing MGMMs that incorporate MAF information could potentially enhance the model's resolution in capturing noisy genotype likelihood distributions for rare variants [104, 105]. This refinement may prove beneficial for accurately simulating genotype likelihoods for variants with MAF < 1%. Third, although the simulated lcWGS data demonstrated high accuracy in PRS prediction, there remains scope for further improvement. Imputation accuracy for rare variants using lcWGS data is currently suboptimal. Leveraging larger, diverse reference panels could potentially enhance imputation performance, particularly for low-frequency variants [28, 106–108]. This limitation could be mitigated by the proposed development of MAF-aware MGMMs. Fourth, the current version of the simulation tool can only generate 1× lcWGS datasets. Future work can involve simulating a broader range of low-coverage data using Bayesian inference techniques. Lastly, lcSimVCF currently does not demonstrate its utility for sex chromosomes and has exclusively focused on biallelic variants. This is primarily due to the additional complexity involved in variant calling and genotype imputation specifically required for sex chromosomes, arising from their unique genomic architecture. Future work will therefore include expanding lcSimVCF's capabilities to accurately model genotype likelihoods for sex chromosomes and developing a robust framework to accommodate multiallelic variants. We aim to improve this tool to simulate various low-coverage levels, enhancing the tool's utility for a broader range of applications. This expansion will further increase the tool's utility across diverse research contexts in population genetics and personalized medicine.

Conclusions

The developed simulation tool demonstrates high reproducibility and efficiency in generating large-scale lcWGS datasets. It accurately simulates lcWGS allele presence probability distributions that closely resemble those observed in empirical lcWGS data. Furthermore, simulated data has shown promising potential for downstream applications, such as population stratification and PRS analysis. Notably, the simulation process can be completed within a reasonable timeframe across various computational environments. This tool holds significant promise in facilitating the development and evaluation of bioinformatics tools that rely on access to extensive lcWGS data resources.

Abbreviations

lcWGS	Low-Coverage Whole-Genome Sequencing
SGM	Single Gaussian Model
MGMMs	Multivariate Gaussian Mixture Models
hcWGS	High-Coverage Whole-Genome Sequencing
PRS	Polygenic Risk Score
GWAS	Genome-Wide Association Studies
1KGP	1000 Genomes Project
UKBB	UK Biobank
VCF	Variant Call Format
EM	Expectation-Maximization
BIC	Bayesian Information Criterion
AFR	Africans
AMR	Admixed Americans
EAS	East Asians
SAS	South Asians
EUR	Europeans
BQSR	Base Quality Score Recalibration
PCA	Principal Component Analysis

PC	Principal Component
ARI	Adjusted Rand Index
SNPs	Single Nucleotide Polymorphisms
CAD	Coronary Artery Disease
T2D	Type 2 Diabetes
PGS	Polygenic Score
AF	Atrial Fibrillation
MAF	Minor Allele Frequency
WGS	Whole-Genome Sequencing
GMM	Gaussian Mixture Model

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-025-06184-3.

Supplementary material 1.

Supplementary material 2.

Supplementary material 3.

Supplementary material 4.

Acknowledgements
We would like to thank Andrew Su's helpful guidance and discussions which made this work possible. We would like to thank J.C. Ducom, Lisa Dong, and the Scripps High Performance Computing service for their support. Fig. 1 was created with BioRender.com.

Author contributions
KC designed the study and developed the simulation pipeline, conducted and generated analysis results presented in the work, created the software, and contributed to the manuscript writing. SC developed PRS calculation pipeline and helped PRS calculation presented in the section of PRS analysis. RD contributed to study design, interpretation of results, and manuscript writing. AT supervised the methodology design and refinement of the work. All authors read and approved the final manuscript.

Funding
This work is supported by 5R01HG010881-03.

Data availability
1KGP [26], http://www.internationalgenome.org/. 1KGP GRCh37 reference panel, https://hgdownload.cse.ucsc.edu/gbdb/hg19/1000Genomes/phase3/. Picard (Institute 2019), https://broadinstitute.github.io/picard/. SAMtools/BCFtools [109], https://www.htslib.org/doc/1.0/bcftools.html. Sentieon software [37], https://www.sentieon.com/. GLIMPSE [28], https://github.com/odelaneau/GLIMPSE/releases/tag/v1.1.0. PLINK [72], https://www.cog-genomics.org/plink/. The simulation tool is an open-source Python package, accessible at https://github.com/chvbs2000/lcWGS_simulation. No datasets were generated or analysed during the current study. Project name: A novel simulation tool for low-coverage whole-genome sequencing using Multivariate Gaussian Mixture Models. Project home page: https://github.com/chvbs2000/lcWGS_simulation. Operating system(s): Platform independent. Programming language: Python. Other requirements: Python packages of scikit allele, pandas, numpy, argparse, and scikit-learn. License: Apache License. Any restrictions to use by non-academics: no restrictions.

Declarations

Ethics approval and consent to participate
No human subjects were used.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 22 July 2024 / Accepted: 5 June 2025
Published online: 25 February 2026

References

1. Tachmazidou I, Süveges D, Min JL, Ritchie GRS, Steinberg J, Walter K, et al. Whole-genome sequencing coupled to imputation discovers genetic signals for anthropometric traits. Am J Hum Genet. 2017;100(6):865–84.
2. Luo Y, De Lange KM, Jostins L, Moutsianas L, Randall J, Kennedy NA, et al. Exploring the genetic architecture of inflammatory bowel disease by whole-genome sequencing identifies association at ADCY7. Nat Genet. 2017;49(2):186–92. https://doi.org/10.1038/ng.3761.

3. Cai N, Bigdeli TB, Kretzschmar W, Lei Y, Liang J, Song L, et al. Sparse whole-genome sequencing identifies two loci for major depressive disorder. Nature. 2015;523(7562):588–91.
4. Gilly A, Southam L, Suveges D, Kuchenbaecker K, Moore R, Melloni GEM, et al. Very low-depth whole-genome sequencing in complex trait association studies. Bioinformatics. 2019;35(15):2555–61.
5. Martin AR, Atkinson EG, Chapman SB, Stevenson A, Stroud RE, Abebe T, et al. Low-coverage sequencing cost-effectively detects known and novel variation in underrepresented populations. Am J Hum Genet. 2021;108(4):656–68.
6. Homburger JR, Neben CL, Mishne G, Zhou AY, Kathiresan S, Khera A V. Low coverage whole genome sequencing enables accurate assessment of common variants and calculation of genome-wide polygenic scores. Genome Med. 2019;11(1).
7. Kim S, Shin JY, Kwon NJ, Kim CU, Kim C, Lee CS, et al. Evaluation of low-pass genome sequencing in polygenic risk score calculation for Parkinson's disease. Hum Genomics. 2021;15(1):1–12. https://doi.org/10.1186/s40246-021-00357-w.
8. Li JH, Mazur CA, Berisa T, Pickrell JK. Low-pass sequencing increases the power of GWAS and decreases measurement error of polygenic risk scores compared to genotyping arrays. Genome Res. 2021;31(4):529–37.
9. Gourlé H, Karlsson-Lindsjö O, Hayer J, Bongcam-Rudloff E. Simulating Illumina metagenomic data with InSilicoSeq. Bioinformatics. 2019;35(3):521–2.
10. Huang W, Li L, Myers JR, Marth GT. ART: a next-generation sequencing read simulator. Bioinformatics. 2012;28(4):593–4.
11. Kuster R, Staton M. Readsynth: short-read simulation for consideration of composition-biases in reduced metagenome sequencing approaches. BMC Bioinformatics. 2024;25(1):1–18. https://doi.org/10.1186/s12859-024-05809-3.
12. Dimitromanolakis A, Xu J, Krol A, Briollais L. sim1000G: a user-friendly genetic variant simulator in R for unrelated individuals and family-based designs. BMC Bioinformatics. 2019;20(1):1–9.
13. Jiang T, Liu S, Cao S, Liu Y, Cui Z, Wang Y, et al. Long-read sequencing settings for efficient structural variation detection based on comprehensive evaluation. BMC Bioinformatics. 2021;22(1):1–17. https://doi.org/10.1186/s12859-021-04422-y.
14. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.
15. Stephens ZD, Hudson ME, Mainzer LS, Taschuk M, Weber MR, Iyer RK. Simulating next-generation sequencing datasets from empirical mutation and sequencing models. PLoS ONE. 2016;11(11):e0167047. https://doi.org/10.1371/journal.pone.0167047.
16. Bouveyron C, Girard S, Schmid C. High-dimensional data clustering. Comput Stat Data Anal. 2007;52(1):502–19.
17. Liu TC, Kalugin PN, Wilding JL, Bodmer WF. GMMchi: gene expression clustering using Gaussian mixture modeling. BMC Bioinformatics. 2022;23(1):1–27. https://doi.org/10.1186/s12859-022-05006-0.
18. McCaw ZR, Aschard H, Julienne H. Fitting Gaussian mixture models on incomplete data. BMC Bioinformatics. 2022;23(1):1–20. https://doi.org/10.1186/s12859-022-04740-9.
19. Budiarto A, Mahesworo B, Hidayat AA, Nurlaila I, Pardamean B (2021) Gaussian mixture model implementation for population stratification estimation from genomics data. Procedia Comput Sci 179(2020):202–10. https://doi.org/10.1016/j.procs.2020.12.026
20. Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, Noble WS (2013) Unsupervised pattern discovery in human chromatin structure through genomic segmentation. In: 2013 ACM Conf Bioinformatics, Comput Biol Biomed Informatics ACM-BCB 2013, vol 9(5), pp 813–4.
21. Price AL, Zaitlen NA, Reich D, Patterson N. New approaches to population stratification in genome-wide association studies. Nat Rev Genet. 2010;11(7):459–63.
22. Hellwege JN, Keaton JM, Giri A, Gao X, Velez Edwards DR, Edwards TL. Population stratification in genetic association studies. Curr Protoc Hum Genet. 2017;95(1):1221–12223.
23. Choi SW, Mak TSH, O'Reilly PF. Tutorial: a guide to performing polygenic risk score analyses. Nat Protoc. 2020;15(9):2759–72.
24. Croucha DJM, Bodmer WF. Polygenic inheritance, GWAS, polygenic risk scores, and the search for functional variants. Proc Natl Acad Sci USA. 2020;117(32):18924–33.
25. Duncan L, Shen H, Gelaye B, Meijsen J, Ressler K, Feldman M, et al. Analysis of polygenic risk score usage and performance in diverse human populations. Nat Commun. 2019;10:1. https://doi.org/10.1038/s41467-019-11112-0.
26. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, et al. A global reference for human genetic variation. Nature. 2015;526:68–74.
27. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27(15):2156.
28. Rubinacci S, Ribeiro DM, Hofmeister RJ, Delaneau O. Efficient phasing and imputation of low-coverage sequencing data using large reference panels. Nat Genet. 2021;53(1):120–6.
29. Delaneau O, Howie B, Cox AJ, Zagury JF, Marchini J. Haplotype estimation using sequencing reads. Am J Hum Genet. 2013;93(4):687–96.
30. Moon TK. The expectatio maximization algorithm. 1996;47–60.
31. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: machine learning in python. J Mach Learn Res. 2011;12:2825–30.
32. Sanner MF. Python: a programming language for software integration and development. J Mol Graph Model. 1999;17(1):57–61.
33. Burnham KP, Anderson DR. Multimodel inference: understanding AIC and BIC in model selection. Sociol Methods Res. 2004;33(2):261–304.
34. Whitt W. Institute of mathematical statistics is collaborating with JSTOR to digitize, preserve, and extend access to The Annals of Statistics. Ann Stat. 1976;4(6):1280–9.
35. Muzzey D, Evans EA, Lieber C. Understanding the basics of NGS: from mechanism to variant calling. Curr Genet Med Rep. 2015;3(4):158–65.
36. Richters MM, Xia H, Campbell KM, Gillanders WE, Griffith OL, Griffith M. Best practices for bioinformatic characterization of neoantigens for clinical utility. Genome Med. 2019;11(1):1–21.
37. Freed D, Aldana R, Weber JA, Edwards JS. The sentieon genomics tools—a fast and accurate solution to variant calling from next-generation sequence data. BioRxiv. 2017. https://doi.org/10.1101/115717.
38. Kendig KI, Baheti S, Bockol MA, Drucker TM, Hart SN, Heldenbrand JR, et al. Sentieon DNASeq variant calling workflow demonstrates strong computational performance and accuracy. Front Genet. 2019;0(JUL):736.

39. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, et al. From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. Curr Protocols Bioinf. 2013. 1–33 p.
40. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Erratum: Initial sequencing and analysis of the human genome: international human genome sequencing consortium (Nature (2001) 409 (860–921)). Nature. 2001;412(6846):565–6.
41. Hattori M. Finishing the euchromatic sequence of the human genome. Tanpakushitsu Kakusan Koso. 2005;50(2):162–8.
42. Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen HC, Kitts PA, et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome Res. 2017;27(5):849–64.
43. Belmont JW, Boudreau A, Leal SM, Hardenbol P, Pasternak S, Wheeler DA, et al. A haplotype map of the human genome. Nature. 2005;437(7063):1299–320.
44. Sudmant PH, Kitzman JO, Antonacci F, Alkan C, Malig M, Tsalenko A, et al. Diversity of human copy number variation and multicopy genes. Science. 2010;330(6004):635–41.
45. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, et al. Mapping and sequencing of structural variation from eight human genomes. Nature. 2008;453(7191):56–64.
46. Yu G, Sapiro G. Statistical compressed sensing of Gaussian mixture models. IEEE Trans Signal Process. 2011;59(12):5842–58.
47. Li H, Hsu W, Lee ML, Wang H. A piecewise Gaussian model for profiling and differentiating retinal vessels. In *Proceedings 2003 International Conference on Image Processing (Cat No03CH37429)*. 2003. p. I–1069.
48. Bhattacharyya A. On a measure of divergence between two multinomial populations. Sankhyā Indian J Stat. 1946;7(4):401–6.
49. Linck E, Battey CJ. Minor allele frequency thresholds strongly affect population structure inference with genomic data sets. Mol Ecol Resour. 2019;19(3):639–47.
50. Simon A, Coop G. The contribution of gene flow, selection, and genetic drift to five thousand years of human allele frequency change. Proc Natl Acad Sci U S A. 2024;121(9):1–12.
51. Barbujani G, Magagni A, Minch E, Cavalli-Sforza L. An apportionment of human DNA diversity. Proc Natl Acad Sci USA. 1997;94(9):4516–9.
52. Hofer T, Ray N, Wegmann D, Excoffier L. Large allele frequency differences between human continental groups are more likely to have occurred by drift during range expansions than by selection. Ann Hum Genet. 2009;73(1):95–108.
53. Barton NH, Hewitt GM. Adaptation, speciation and hybrid zones. Nature. 1989;341(6242):497–503.
54. Chen N, Juric I, Cosgrove EJ, Bowman R, Fitzpatrick JW, Schoech SJ, et al. Allele frequency dynamics in a pedigreed natural population. Proc Natl Acad Sci USA. 2019;116(6):2158–64.
55. Peterson RE, Kuchenbaecker K, Walters RK, Chen CY, Popejoy AB, Periyasamy S, et al. Genome-wide association studies in ancestrally diverse populations: opportunities, methods, pitfalls, and recommendations. Cell. 2019;179(3):589–603. https://doi.org/10.1016/j.cell.2019.08.051.
56. Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D, et al. Genome-wide association studies. Nat Rev Methods Prim. 2021;1(1):467–84. https://doi.org/10.1038/s43586-021-00056-9.
57. Barton N, Hermisson J, Nordborg M. Population genetics: why structure matters. Elife. 2019;8:e45380. https://doi.org/10.7554/eLife.45380.
58. Gravel S. Population genetics models of local ancestry. Genetics. 2012;191(2):607–19.
59. Thornton TA, Bermejo JL. Local and global ancestry inference and applications to genetic association analysis for admixed Populations. Genet Epidemiol. 2014;
60. Freedman ML, Reich D, Penney KL, McDonald GJ, Mignault AA, Patterson N, et al. Assessing the impact of population stratification on genetic association studies. Nat Genet. 2004;36(4):388–93.
61. Marnetto D, Pärna K, Läll K, Molinaro L, Montinaro F, Haller T, et al. Ancestry deconvolution and partial polygenic score can improve susceptibility predictions in recently admixed individuals. Nat Commun. 2020;11(1):1–9. https://doi.org/10.1038/s41467-020-15464-w.
62. Dandine-Roulland C, Bellenguez C, Debette S, Amouyel P, Génin E, Perdry H. Accuracy of heritability estimations in presence of hidden population stratification. Sci Rep. 2016;6(1):1–10. https://doi.org/10.1038/srep26471.
63. Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. Nat Rev Genet. 2018;19(9):581–90. https://doi.org/10.1038/s41576-018-0018-x.
64. Giani AM, Gallo GR, Gianfranceschi L, Formenti G. Long walk to genomics: History and current approaches to genome sequencing and assembly. Comput Struct Biotechnol J. 2020;18:9–19. https://doi.org/10.1016/j.csbj.2019.11.002.
65. Logsdon GA, Vollger MR, Eichler EE. Long-read human genome sequencing and its applications. Nat Rev Genet. 2020;21(10):597–614. https://doi.org/10.1038/s41576-020-0236-x.
66. Hu T, Chitnis N, Monos D, Dinh A. Next-generation sequencing technologies: an overview. Hum Immunol. 2021;82(11):801–11. https://doi.org/10.1016/j.humimm.2021.02.012.
67. Pritchard JK, Rosenberg NA. Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet. 1999;65(1):220–8.
68. Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: dominant markers and null alleles. Mol Ecol Notes. 2007;7(4):574–8.
69. Meirmans PG, Hedrick PW. Assessing population structure: FST and related measures. Mol Ecol Resour. 2011;11(1):5–18.
70. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;38(8):904–9.
71. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLOS Genet. 2006;2(12):1–20. https://doi.org/10.1371/journal.pgen.0020190.
72. Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4(1):13742015–800478. https://doi.org/10.1186/s13742-015-0047-8.
73. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2022.
74. Hubert L, Arabie P. Comparing partitions. J Classif. 1985;2(1):193–218. https://doi.org/10.1007/BF01908075.
75. Yeung KY, Ruzzo WL. Details of the Adjusted Rand index and Clustering algorithms Supplement to the paper "An empirical study on Principal Component Analysis for clustering gene expression data" (to appear in Bioinformatics). In 2001. https://api.semanticscholar.org/CorpusID:16021771

76. Morales J, Welter D, Bowler EH, Cerezo M, Harris LW, McMahon AC, et al. A standardized framework for representation of ancestry data in genomics studies, with application to the NHGRI-EBI GWAS Catalog. Genome Biol. 2018;19(1):1–10.
77. Chatterjee N, Shi J, García-Closas M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. Nat Rev Genet. 2016;17(7):392–406.
78. Claussnitzer M, Cho JH, Collins R, Cox NJ, Dermitzakis ET, Hurles ME, et al. A brief history of human disease genetics. Nature. 2020;577(7789):179–89.
79. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 years of GWAS discovery: biology, function, and translation. Am J Hum Genet. 2017;101(1):5–22. https://doi.org/10.1016/j.ajhg.2017.06.005.
80. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 2017;45(D1):D896-901.
81. Erdmann J, Kessler T, Munoz Venegas L, Schunkert H. A decade of genome-wide association studies for coronary artery disease: the challenges ahead. Cardiovasc Res. 2018;114(9):1241–57. https://doi.org/10.1093/cvr/cvy084.
82. Mahajan A, Taliun D, Thurner M, Robertson NR, Torres JM, Rayner NW, et al. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. Nat Genet. 2018;50(11):1505–13. https://doi.org/10.1038/s41588-018-0241-6.
83. Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat Genet. 2018;50(9):1219.
84. Lambert SA, Wingfield B, Gibson JT, Gil L, Ramachandran S, Yvon F, et al. The Polygenic Score Catalog: new functionality and tools to enable FAIR research. medRxiv [Internet]. 2024 Jan 1;2024.05.29.24307783. http://medrxiv.org/content/early/2024/05/31/2024.05.29.24307783.abstract
85. Martin AR, Atkinson EG, Chapman SB, Stevenson A, Stroud RE, Abebe T, et al. Low-coverage sequencing cost-effectively detects known and novel variation in underrepresented populations. bioRxiv. 2020.
86. Pasaniuc B, Rohland N, McLaren PJ, Garimella K, Zaitlen N, Li H, et al. Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. Nat Genet. 2012;44(6):631–5.
87. Wasik K, Berisa T, Pickrell JK, Li JH, Fraser DJ, King K, et al. Comparing low-pass sequencing and genotyping for trait mapping in pharmacogenetics. BMC Genomics. 2021;22(1):1–7.
88. Crysup B, Woerner AE. A genotype likelihood function for DNA mixtures. Forensic Sci Int Genet. 2022;61:102776.
89. Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet 2011;12(6):443–51. https://doi.org/10.1038/nrg2986
90. Alex Buerkle C, Gompert Z. Population genomics based on low coverage sequencing: how low should we go? Mol Ecol. 2013;22(11):3028–35.
91. Holland D, Frei O, Desikan R, Fan CC, Shadrin AA, Smeland OB, et al. Beyond SNP heritability: polygenicity and discoverability of phenotypes estimated with a univariate Gaussian mixture model. PLoS Genet. 2020;16(5):1–30. https://doi.org/10.1371/journal.pgen.1008612.
92. Au K, Lin R, Foulkes AS. Mixture modelling as an exploratory framework for genotype-trait associations. J R Stat Soc Ser C Appl Stat. 2011;60(3):355–75.
93. McDowell IC, Manandhar D, Vockley CM, Schmid AK, Reddy TE, Engelhardt BE. Clustering gene expression time series data using an infinite Gaussian process mixture model. PLoS Comput Biol. 2018;14(1):1–27.
94. Lawrence ND, Sanguinetti G, Rattray M. Modelling transcriptional regulation using Gaussian processes. NIPS 2006 Proc 19th Int Conf Neural Inf Process Syst. 2006;785–92.
95. Banfield JD, Raftery AE. Model-based Gaussian and non-Gaussian clustering. Biometrics. 1993;49(3):803–21.
96. McLachlan GJ, Lee SX, Rathnayake SI. Finite mixture models. Annu Rev Stat Its Appl. 1988;2019(6):355–78.
97. Wall JD, Pritchard JK. Haplotype blocks and linkage disequilibrium in the human genome. Nat Rev Genet. 2003;4(8):587–97.
98. Gazal S, Finucane HK, Furlotte NA, Loh PR, Palamara PF, Liu X, et al. Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. Nat Genet. 2017;49(10):1421–7. https://doi.org/10.1038/ng.3954.
99. Lambert SA, Abraham G, Inouye M. Towards clinical utility of polygenic risk scores. Hum Mol Genet. 2019;28(R2):R133–42.
100. Yong SY, Raben TG, Lello L, Hsu SDH. Genetic architecture of complex traits and disease risk predictors. Sci Rep. 2020;10(1):1–14. https://doi.org/10.1038/s41598-020-68881-8.
101. Wang Y, Kanai M, Tan T, Kamariza M, Tsuo K, Yuan K, et al. Polygenic prediction across populations is influenced by ancestry, genetic architecture, and methodology. Cell Genomics. 2023;3(10):100408. https://doi.org/10.1016/j.xgen.2023.100408.
102. Peters BA, Kermani BG, Sparks AB, Alferov O, Hong P, Alexeev A, et al. Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. Nature. 2012;487(7406):190–5.
103. Hofmeister RJ, Ribeiro DM, Rubinacci S, Delaneau O. Accurate rare variant phasing of whole-genome and whole-exome sequencing data in the UK Biobank. Nat Genet. 2023;55(7):1243–9.
104. Bomba L, Walter K, Soranzo N. The impact of rare and low-frequency genetic variants in common disease. Genome Biol. 2017;18(1):1–17.
105. Momozawa Y, Mizukami K. Unique roles of rare variants in the genetics of complex diseases in humans. J Hum Genet. 2021;66(1):11–23. https://doi.org/10.1038/s10038-020-00845-2.
106. Howie B, Marchini J, Stephens M. Genotype imputation with thousands of genomes. G3 Genes Genomes Genet. 2011;1(6):457–70.
107. Das S, Abecasis GR, Browning BL. Genotype imputation from large reference panels. Vol. 19, Annual Review of Genomics and Human Genetics. Annual Reviews Inc.; 2018. p. 73–96.
108. Rubinacci S, Hofmeister RJ, Sousa da Mota B, Delaneau O. Imputation of low-coverage sequencing data from 150,119 UK Biobank genomes. Nat Genet. 2023;55(7):1088–90.
109. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021;10(2).

Publisher's Note