Latent generative modeling of long genetic sequences with GANs

Antoine Szatkownik[1], Cyril Furtlehner[1], Guillaume Charpiat[1], Burak Yelmen[1,2,*], Flora Jay[1,*]
[1] Université Paris-Saclay, CNRS, INRIA, LISN, Paris, France
[2] University of Tartu, Institute of Genomics, Tartu, Estonia
[*] These authors contributed equally.
Corresponding author: Antoine Szatkownik <szatkownik@lisn.fr>

Abstract

Synthetic data generation via generative modeling has recently become a prominent research field in genomics, with applications ranging from functional sequence design to high-quality, privacy-preserving artificial in silico genomes. Following a body of work on Artificial Genomes (AGs) created via various generative models trained with raw genomic input, we propose a conceptually different approach to address the issues of scalability and complexity of genomic data generation in very high dimensions. Our method combines dimensionality reduction, achieved by Principal Component Analysis (PCA), and a Generative Adversarial Network (GAN) learning in this reduced space. Using this framework, we generated genomic proxy datasets for very diverse human populations around the world. We compared the quality of AGs generated by our approach with AGs generated by the established models and report improvements in capturing population structure, linkage disequilibrium, and metrics related to privacy leakage. Furthermore, we developed a frugal model with orders of magnitude fewer parameters and comparable performance to larger models. For quality assessment, we also implemented a new evaluation metric based on information theory to measure local haplotypic diversity, showing that generative models yield higher diversity than real genomes. In addition, we addressed the shrinkage issue associated with PCA and generative modeling, examined its relation to the nearest neighbor resemblance metric, and proposed a resolution. Finally, we evaluated the effect of different binarization methods on the quality of the output AGs.

1 Introduction

Growing advances in DNA sequencing have spurred a data surge, elevating machine learning's importance in genomics (Yelmen and Jay, 2023; Korfmann, 2023; Huang et al., 2023). However, a substantial amount of these data, held by companies, institutions or governmental agencies, are tied to privacy concerns impeding their accessibility. In contemporary populations genomics, the unsupervised training of generative neural networks on genomes mainly aims at enhancing visualization by revealing fine-scale population structure (Battey et al., 2021; Meisner and Albrechtsen, 2022; Geleta et al., 2023), inferring evolutionary histories (Wang et al., 2021; Riley et al., 2024), or generating realistic

synthetic data (Yelmen et al., 2021; Perera et al., 2022; Booker et al., 2023; Yelmen et al., 2023; Dang et al., 2023; Zhang et al., 2024; Burnard et al., 2023; Ahronoviz and Gronau, 2024; Szatkownik et al., 2024). For the latter goal, an alternative paradigm consists of simulations based on the resampling of known haplotypes (Su et al., 2011; Tang and Liu, 2019; Dimitromanolakis et al., 2019; Wharrie et al., 2023). Generated synthetic genomes, or artificial genomes (AGs), can potentially become privacy-preserving alternatives to real biobanks and be utilized for genome-wide association studies (GWAS), local ancestry inference, genomic imputation or natural selection scans (Yelmen et al., 2021).

Generative neural networks are a class of deep-learning based algorithms, that once trained, can be used to map noise to a data distribution that should resemble as closely as possible to the real training data distribution. Several models have been developed over the years with varying degrees of success, ranging from variational autoencoders (VAEs) (Kingma and Welling, 2014), generative adversarial networks (GANs) (Goodfellow et al., 2014), Restricted Boltzmann Machines (RBMs) (Smolensky, 1986; Ackley et al., 1985) to diffusion models (Ho et al., 2020) and language models (Zhang et al., 2024). In this paper, we focus on GANs, which consist of a circuit of two neural networks competing with each other where one, the discriminator, is assigned the task of distinguishing fake from real data while the other one, the generator, is simulating real data with synthetic data via a generative process whose purpose is to trick the discriminator.

In this context of generative AI for genomics, Yelmen et al. (2023) tackled the challenging problem of generating very high dimensional genetic data despite small sample sizes by designing and training convolutional Generative Adversarial Networks (GANs), convolutional Variational Autoencoders (VAEs), and conditional Restricted Boltzmann Machines (RBMs). Unlike previous fully connected architectures (Yelmen et al., 2021), the proposed scheme of convolving along the genomic sequences combined with location-specific variables is computationally feasible (Ausmees and Nettelblad, 2022; Yelmen et al., 2023), yet it might still pose challenges when targeting dense and intrinsically complex whole genomes. These difficulties drive the need to seek alternative methods for large genome-wide frameworks.

The recent staggering success of generative AIs for high-resolution image synthesis such as *DALL-E* (Ramesh et al., 2021) or *Stable Diffusion* (Rombach et al., 2022) roots in a two-stage process: learning a good latent representation of the data and then training a generative model in the resulting lower-dimensional subspace. This paradigm sparked growing attention within the ML community, as illustrated by the *NeurIPS Tutorial* set up in 2023, *Latent Diffusion Models: Is the Generative AI Revolution Happening in Latent Space?* which attracted thousands of people (LDM, 2023). Inspired by this, we call *latent generative models*, the association of a compression step followed by generative modeling in that space. This framework enables high-quality synthetic data generation while bypassing massive volumes of computations. Although it has been suggested that these two steps could be learned jointly, results were less prominent (Vahdat et al., 2021; Rombach et al., 2022).

In addition, despite being straightforward architectures to implement, train and optimize, fully-connected (FC) architectures are not practical for direct application to large-scale genetic data. For instance, a simple model consisting of just one dense linear layer with

$n = 1000$ hidden variables and dealing with genetic sequences of L single nucleotide polymorphisms (SNPs), with L potentially of the order of millions (e.g., $L = 1\text{M}$), would have $L \times n = 1\text{B}$ parameters. In a latent generative modeling framework, boiling the important information down to a smaller space bypasses this large computational overhead and renders FC affordable.

To address the high-dimensional generative task, we thus propose to project the genetic data into a lower-dimensional subspace through dimension reduction and to train a GAN within this reduced space. Principal Component Analysis (PCA) was preferred for its easy and faithful reconstruction from Principal Component (PC) space to data space, unlike t-SNE and UMAP, which either lack invertibility or result in poor reconstruction. While autoencoders are also of interest, they are heavily parameterized, leading to long training times, and require careful regularization adjustments. On the contrary, PCA is nonparametric, efficiently computed in minimal time, and widely used in population genetics studies since ancestry-linked genomic variation is well represented in PC space (Tian et al., 2008; Price et al., 2010; Lao et al., 2008; Novembre et al., 2008; Yu et al., 2008; Menozzi et al., 1978; Patterson et al., 2006; Hanotte et al., 2002; Price et al., 2006). In the principal subspace, each new dimension is a linear combination of the input features, therefore the notion of locality along features disappears when replacing the genomic sequences by their PC scores. This makes dense layers or attention layers ideal candidates for the generative network architecture. Importantly, although PCA is linear, the generator output is not, as the GAN can learn non-linearities. This preserves our capacity to capture the complex structure of genomic sequences.

In this study, we introduce three latent generative modeling methods, a PCA Wasserstein GAN with gradient penalty (PCA-WGAN), a chained alternative designed to preserve local information (Glocal-PCA-WGAN), and a frugal version with fewer parameters (Light PCA-WGAN). We showcase the usefulness of our approaches through comparative analyses with previous models on The 1000 Genomes Project Consortium (2010) genetic dataset encompassing diverse human populations and demonstrate improvements on several population genetics summary statistics and metrics related to privacy leakage.

2 Materials and Methods

2.1 Overview

Our general approach for generating realistic genomic sequences consists of (i) reducing the training dataset dimensions through principal component analysis; (ii) training a WGAN on these PC scores instead of the haplotypes; (iii) generating synthetic PC scores by sampling the generator; (iv) inverse-transforming these scores back to the genetic space, yielding synthetic genome samples; (v) assessing the quality of the artificial genomes using multiple descriptive population genetics statistics (**FIG. S1**).

2.2 Dataset description

The 1000 Genomes project is a database of human genome sequences obtained by sampling individuals from 26 populations spread worldwide, with a number of samples for each population ranging from one to roughly a hundred. The 1000 Genomes dataset used in this

study was curated by Yelmen et al. (2023). It consists of 2504 individuals corresponding to 5008 phased haplotypes (*i.e.* for each individual there is one reconstructed haplotype coming from the mother and one from the father), and contains 65,535 contiguous SNPs spanning chr1:534247-81813279 (*i.e.* ∼80 Mega base pairs), within the Omni 2.5 genotyping array framework. The data is formatted in the following way: rows are phased haplotypes; columns are positions of alleles represented by 0 when the corresponding nucleotide is identical with respect to the one in the reference genome (GRCh37) and 1 when it is point-wise mutated. Hence, the dataset is represented as a binary matrix.

2.3 PCA-WGAN

A Generative Adversarial Network (GAN) comprises two competing neural networks, the discriminator and the generator. The discriminator learns to dissociate real data from synthetic ones, while the generator aims at fooling the discriminator in a way that it fails at this classification task. Vanilla GANs face several problems, such as mode collapse, training instability, vanishing gradients, and sensitivity to hyperparameters (Arjovsky et al., 2017; Gulrajani et al., 2017). To account for these challenges, and following its successful application to population genetics (Booker et al., 2023; Yelmen et al., 2023), we used a Wasserstein GAN (Arjovsky et al., 2017) with gradient penalty (Gulrajani et al., 2017) (WGAN-GP) implemented in PyTorch (Paszke et al.). Whereas the discriminator of a vanilla GAN produces a probability of being fake or real, the critic of a WGAN produces an unbounded "realness" score for fake and real samples. The loss of a WGAN-GP, based on the estimation of the Earth Mover's distance by the critic of two distributions on a batch of samples, is as follows :

$$\mathbb{E}_{\tilde{z}\sim\mathbb{P}_{\text{generated}}}\big[C(\tilde{z})\big] - \mathbb{E}_{z\sim\mathbb{P}_{\text{real}}}\big[C(z)\big] + \lambda\mathbb{E}_{\hat{z}\sim\mathbb{P}_{\hat{z}}}\big[(\|\nabla_{\hat{z}}C(\hat{z})\| - 1)^2\big]$$
$$\text{s.t.} \quad \hat{z} = t\tilde{z} + (1-t)z \quad \text{and} \quad t \sim \mathbb{U}[0,1],$$

where C is the critic; $\tilde{z}$ (resp. z) is a sample from the generator modeled distribution $\mathbb{P}_{\text{generated}}$ (resp. from the data distribution $\mathbb{P}_{\text{real}}$). The gradient penalty term, evaluated at a point $\hat{z}$ interpolated between a real and a fake point, is scaled by a constant factor λ and constrains the critic's gradient to be close to 1. This gradient penalty regularization (Gulrajani et al., 2017) is a stabler alternative than weight clipping (Arjovsky et al., 2017), for enforcing the 1-Lipschitzianity property of the critic, a condition required by the Kantorovich-Rubinstein duality (Villani, 2009), otherwise called 1-Wassertein distance ($\mathcal{W}_1$). $\mathbb{U}[0,1]$ is the uniform distribution on $[0,1]$, yielding a distribution $\mathbb{P}_{\hat{z}}$ over $\hat{z}$. Enforcing the critic's gradient to be of unit norm everywhere is intractable as it would require a number of samples exponential in the number of dimensions, instead (Gulrajani et al., 2017) showed that this property is met for an optimal critic along straight lines connecting pairs of points from both real and synthetic distribution.

The PCA-WGAN workflow combines an initial PCA on a binary matrix (5008 samples, 65K SNPs, see Section **2.2**), retaining 4507 components (90% PCs, see **Section 3.1 & S4**) with a generative learning process in this PC space (**FIG.1.A**). To generate new fake binary SNP data, we sampled points in the space learned by the GAN (modeling the PC space), then performed the PCA inverse transformation using eigenvectors computed from real data PCA, and applied a binarization step based on a threshold of 0.5 (see Supplementary text **S4** and Section 2.6 for more details and alternative procedures). All the relevant data and code is publicly available at `https://gitlab.inria.fr/ml_genetics/public/`

artificial_genomes/-/tree/master/1000G_real_genomes for data and https://gitlab.inria.fr/ml_genetics/public/artificial_genomes/-/tree/master/PCA-WGAN for the models.

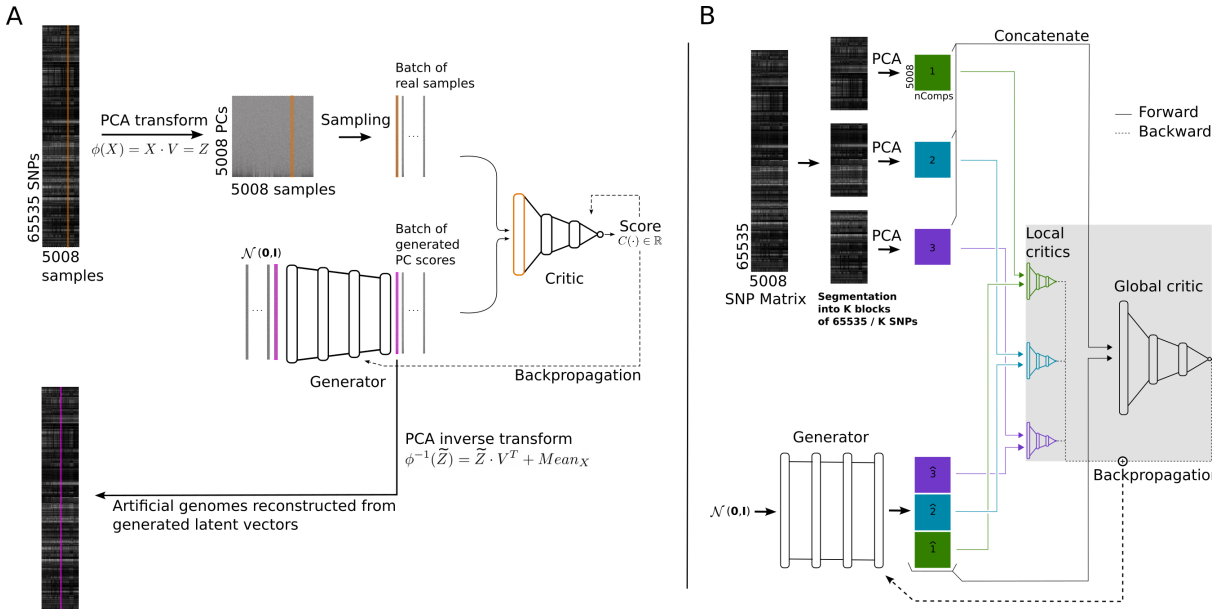

Figure 1: **PCA-WGAN architectures. A.** PCA-WGAN. The SNP matrix of 5008 samples by 65535 SNPs is reduced via PCA to a PC score matrix of 5008 samples by 4507 dimensions (90% PCs). The WGAN learns the PC space distribution. Once the training of the WGAN generator and critic is finished, new PC scores are generated and mapped back to the data space via PCA inverse transform. **B.** Glocal-PCA-WGAN. The SNP matrix is split into even-sized chunks, to which PCA is independently applied. Each local critic is fed with a single PCA chunk (colored block labeled 1, 2, or 3) and a generated PC scores chunk (colored block labeled $\hat{1}$, $\hat{2}$, or $\hat{3}$), while the global critic is fed with the concatenated PCA projections and generated PC scores (i.e. all colored blocks).

2.4 Glocal-PCA-WGAN: Chaining multiple local PCA-WGAN

We designed an improvement of PCA-WGAN, called Glocal-PCA-WGAN (G stands for global), that explicitly leverages both global and local genomic scales. To circumvent difficulties arising from performing dimension reduction in the $\frac{N_{\text{features}}}{N_{\text{samples}}} \gg 1$ regime, our approach segments the SNP matrix into K multiple blocks. The procedure goes as follows: we split the SNP matrix into three successive blocks of equal number of SNPs and applied PCA to each block (i.e., each genomic region) independently of each other (yielding the colored blocks labeled 1, 2 and 3 in **FIG. 1B**). In this architecture, there are multiple local critics and one global critic. Each local critic aims to capture the local information of its assigned block, while the global critic should recover the relations between the blocks, i.e., at a larger genomic scale. The WGAN-GP is trained in the space of the concatenated PCA projections. Once the training is done, we sampled the generator, split the generated PC scores into three parts and inversed transform each chunk with the corresponding PCA (see Appendix **S5** for more details).

Constraints on the choice of block number *K*. Increasing K leads to an increased difficulty for the WGAN as it will not only have to recover long-range interactions between

the blocks but also short-range interactions since increasing K leads to higher risk of breaking interactions within an LD block. Furthermore, if K is high enough such that the blocks have fewer features than samples, then either (i) keeping all principal components amounts to no reduction at all (i.e., the output size of the generator & input size of the global critic will be 65K), or (ii) keeping fewer components leads to substantial reconstruction errors. On the other hand, if K is too small, i.e., the genomic chunks are large, we are compressing too many SNPs into N_{samples} dimensions. This is complicated further by the fact that keeping all PCs yields an input size of $K \times N_{\text{samples}}$ for the global discriminator, which should be kept low. All in all, we want K to be small but not so much so that PCA blocks still capture local information. Hence we chose $K = 3$ to keep the balance between reasonable parameter size and retaining a number of PCs that still yields a nearly optimal reconstruction error.

2.5 Light PCA-WGAN

Finally, we designed a light version of PCA-WGAN. This light model learns to mimic the principal components displaying multi-modal structure, i.e. the first six PCs, and is associated with a counterpart that models the remaining 5002 PCs thanks to a multivariate normal distribution. We trained a neural network (NN) to predict the parameters of the multivariate normal distribution (**FIG. S9**) and coined this model as Light PCA-WGAN with NN. Precisely, the neural network predicts the means and variances of these remaining PCs (low variance) from the first six PCs (high variance). The neural network is optimized on the real training data through the minimization of the Gaussian negative log likelihood loss (Nix and Weigend, 1994).

Sampling. To create novel genomic sequences, we first generated 6-dimensional samples with the Light PCA-WGAN. We then sampled the low variance PC scores conditionally on these high variance PC scores. For this, we ran the 6-dimensional samples through the NN to recover the parameters of a diagonal multivariate normal distribution and sampled this distribution. Finally, we concatenated the high and low variance parts of a sample and inverse-transformed it with PCA.

2.6 Binarization

Latent generative modeling requires projecting synthetic samples from the latent space to the data space once the samples are generated. Given that PCA is not designed for discrete spaces, it assumes continuity for the data space, resulting in inverse transformed samples living in a continuous space. Consequently, a final transformation is necessary to map these samples back into SNP space. We evaluated two approaches for this transformation: a naive binarization method applying a fixed threshold of 0.5 for all sites, and a tailored binarization approach with site-specific thresholds.

For that latter approach, at each site we computed the cumulative distribution of the reconstructed synthetic samples and the allelic frequency in the real data of that site. The threshold was then obtained by running the frequency of the real data through the inverse cumulative (derived by linear interpolation of the empirical cumulative distribution).

2.7 Shrinkage correction

We split the SNP matrix into a train and test sets of equal size, applied PCA to the training set, and used the resulting basis of eigenvectors to project test samples that were not included in the initial PCA. In the high dimension low sample size regime ($\frac{n}{p} \ll 1$ where n is the sample size and p is the number of features), a shrinkage phenomenon appears, as described by Lee et al. (2010, 2014). This phenomenon is related to the asymptotic behavior of the sample eigenvectors, indicating a discrepancy between the population and sample eigenvectors.

Following a correction proposed by Lee et al. (2010) for PCA projection bias, we adjusted each PC score of a shrunk distribution $\mathcal{D}$ (either the synthetic distribution from a generative model or the projected test set) with a multiplication factor equal to the inverse empirical shrinkage factor, $\sqrt{\frac{\mathrm{Var}(PC_{i,\mathcal{D}})}{\mathrm{Var}(PC_{i,Train})}}^{-1}$, where $\mathrm{Var}(PC_{i,\mathcal{D}})$, $\mathrm{Var}(PC_{i,Train})$ are the variances on the i-th PC of the generated distribution in the latent space or the projected test set, and the projected train set respectively.

2.8 Sensitivity of $\mathcal{AA}_{TS}$ with respect to shrinkage

$\mathcal{AA}_{TS}$ (Yale et al., 2020) is a score based on nearest neighbors to assess overfitting for the synthetic data, which can further be used to derive a privacy score by comparing the resemblance of synthetic to train and test datasets.
This score compares the distances between pairs of samples within a distribution, as well as between distributions. Namely, $d_{\mathrm{TT}}(i)$ is the distance of a fixed real sample indexed by i to its nearest neighbor in the real data, while $d_{\mathrm{TS}}(i)$ is the distance of a fixed real sample indexed by i to its nearest neighbor in the synthetic data. The distances $d_{\mathrm{SS}}(i)$ and $d_{\mathrm{ST}}(i)$ are defined similarly. In an ideal scenario of perfect resemblance between the real and synthetic distributions, the nearest neighbor of a real sample would be equally likely to be synthetic as real, and similarly, the nearest neighbor of a synthetic sample would be equally likely to be real as synthetic.

Definition: We assume a training set $T = \{z_i^r, i = 1, \dots N\}$ of size N (r for real), and a synthetic set of same size $S = \{z_i^s, i = 1, \dots N\}$

$$\begin{aligned}\mathcal{AA}_{TS} &\stackrel{\text{def}}{=} \frac{1}{2}(\mathcal{AA}_{truth} + \mathcal{AA}_{syn}) \\ &\stackrel{\text{def}}{=} \frac{1}{2N}\Big(\sum_{i=1}^{N} \mathbb{1}_{\{d_{\mathrm{TS}}(i) > d_{\mathrm{TT}}(i)\}} + \mathbb{1}_{\{d_{\mathrm{ST}}(i) > d_{\mathrm{SS}}(i)\}}\Big)\end{aligned}$$

where

$$d_{\mathrm{TT}}(i) \stackrel{\text{def}}{=} \min_{z^r \in T \setminus \{z_i^r\}} ||z_i^r - z^r||_2, \qquad \forall z_i^r \in T$$

$$d_{\mathrm{TS}}(i) \stackrel{\text{def}}{=} \min_{z^s \in S} ||z_i^r - z^s||_2, \qquad \forall z_i^r \in T$$

$$d_{\mathrm{ST}}(i) \stackrel{\text{def}}{=} \min_{z^r \in T} ||z_i^s - z^r||_2, \qquad \forall z_i^s \in S$$

$$d_{\mathrm{SS}}(i) \stackrel{\text{def}}{=} \min_{z^s \in S \setminus \{z_i^s\}} ||z_i^s - z^s||_2, \qquad \forall z_i^s \in S$$

We further model the effect of shrinkage (see Subsection 2.7) on this indicator, λ being the shrinkage factor.

Proposition:
Let $p_T = \mathcal{N}(0, \mathbb{I}_D)$, $p_S = \mathcal{N}(0, \lambda \mathbb{I}_D)$ with $\lambda \in]0, 1[$ and $D \in \mathbb{N}$ be the latent representation dimension. Let $T = \{\mathrm{z}_i^r\}_{i \in [\![1,N]\!]}$ and $S = \{z_i^s\}_{i \in [\![1,N]\!]}$, where $z_i^r \sim p_T$ and $z_i^s \sim p_S$ $\forall i \in [\![1, N]\!]$. Then

$$\mathcal{AA}_{truth} \xrightarrow[D \to \infty]{} 0$$
$$\mathcal{AA}_{syn} \xrightarrow[D \to \infty]{} 1$$

Our proposition illustrates that in a simple scenario of high-dimensional nested distributions, a generation anomaly often known as mode shrinkage (Stein et al., 2023; Yamaguchi and Fukuda, 2023; Montavon et al., 2016), $\mathcal{AA}_{\mathrm{truth}}$ converges to 0 and $\mathcal{AA}_{\mathrm{syn}}$ converges to 1, *i.e.* $\mathcal{AA}_{\mathcal{TS}}$ converges to 0.5, yet the two distributions are easily distinguishable. Numerical experiments were conducted to confirm this assertion (see Appendix S3.6).

2.9 k-SNP motifs

We investigated the fine-grained structure of synthetic haplotypes. Specifically, we considered non-overlapping windows of size k along a haplotype, aiming to discern how generative models explore through these discrete spaces of sizes 2^k. Inspired by k-mer commonly used in bioinformatics analyses, we dub k-SNP motifs such subsegments. For each motif of size k, with k ranging from 4 to 2048, we counted the number of unique k-SNP motifs within each non-overlapping window of the same size (**FIG.4.A**)
Secondly, we re-evaluated k-SNP motif counts after correcting for potential ancestry biases. As African ancestry contains higher genetic variations, a generative model that creates more samples of African ancestry than contained in the real dataset, might yield higher SNP motif diversity. We generated an excess number of AGs and then downsampled them to match the exact number of samples per continental ancestry group present in the real dataset. To achieve this, we labeled the AGs using a k-NN algorithm. Finally, we counted motifs within each continental ancestry group (**FIG.S12**).
A power-law was fitted on the distribution of k-SNP motif diversity for k ranging from 4 to 2048. To do that, we compiled the motif diversity for each k (the boxes in **FIG.4.A.**) into a single distribution, hence the information relative to k is abstracted away (**FIG.4.C**).

3 Results

As our latent generative modeling approach is based on learning in a reduced space, we first investigated in *Subsection 3.1* whether this smaller space retained enough information regarding the genomic space. In *Subsection 3.2* we describe shrinkage, a phenomenon that might appear both in PCA and in the generative case, albeit stemming from different causes.

Finally, we evaluated the artificial genomes generated with various models trained on a human dataset (The 1000 Genomes Project Consortium, 2010). Previously, Yelmen et al. (2021) investigated the quality of generated data, based on a set of summary statistics coming from population genetics. Here, we present a subset of these metrics and novel ones on the synthetic 65K SNPs data generated by PCA-WGAN, Glocal-PCA-WGAN, Light PCA-WGAN and compare these to previous WGAN and RBM models, which learn directly in the genetic space (Yelmen et al., 2023). The quality assessment of the synthetic data is presented in *Subsections 3.3, 3.4,3.5.*

3.1 Principal components space is a fertile ground for learning

We performed various pre-hoc training experiments to decide how many dimensions (i.e., PCs) should be kept after PCA and to gauge the errors caused by reconstruction. These experiments were performed considering a separate test dataset consisting of real genomes as the best-case generation outcome.

We first wanted to measure the quality of PCA encoding-decoding. Specifically, we aimed to determine whether the data generated in PC space can be reliably mapped back to the original space, assuming the generation step was successful. To check this, we randomly partitioned the dataset into a train and a test set. Both sets follow the same distribution; in that, the same modes appear in both, and they have similar densities. The test set acts as the perfect case where an ideal generative model completely captures the target distribution. Hence, a test set should give us hints on the expected behavior of properly generated data after reconstruction.

We considered the following protocol that does not rely on generated samples. First, we split randomly the dataset ($N_{\text{samples}} = 5008$) into a train and a test set of equal size and performed a PCA on the training set, varying the number of kept components. Then, we projected the test set onto the principal subspace derived from the training set (centered beforehand). Finally, we reconstructed the projected test set back into the original space and binarized it. Binarization was obtained through quantization with a threshold set to 0.5. We computed the mean reconstruction error of the binarized reconstructed test set with respect to the initial test set. The reconstruction error for one sample in the data space is defined as $\frac{1}{L}\sum_{i=1}^{L}|X^i_{\text{reconstructed test}} - X^i_{\text{test}}|$ where L is the number of SNPs, X^i_{test} and
$X^i_{\text{reconstructed test}} \in \{0, 1\}$. The reconstruction error for the training set was zero when keeping 100% of the PCs (**FIG.2.a**), *i.e.*, there is no loss of information, but was equal to 0.06 in average for the test set. Decreasing the number of retained PCs reduced the gap between the train and test errors, indicating limited overfitting when few components were kept. However, because the test reconstruction error was at its minimum when keeping all components, we applied this final procedure for all remaining analyses.

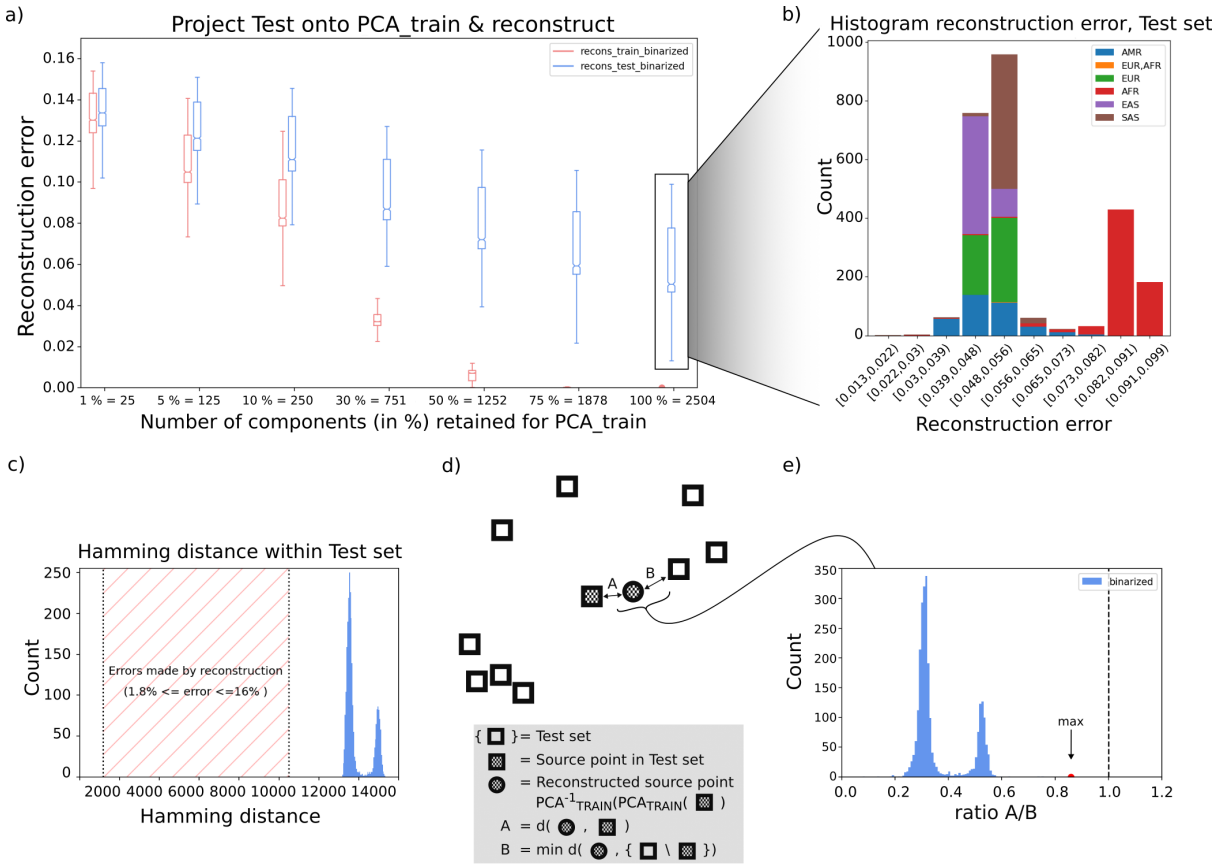

Figure 2: **PCA reconstruction analyses. a)** Reconstruction error as a function of the number of components retained when applying PCA to the training set. The reconstruction error is plotted after reconstructing the training set (red), and after projecting the test set onto the eigensubspace of the training set and reconstructing it (blue). **b)** Histogram of the reconstruction error for the reconstructed test set, keeping all components, colored by superpopulation labels. The superpopulation labels correspond to: AMR: American ancestry, EUR,AFR: European and African ancestry, EUR: European ancestry, AFR: African ancestry, EAS: East Asian ancestry, SAS: South Asian ancestry. **c)** Hamming distance within the test set. **d)** Illustration of how much reconstruction moves data points. **e)** Distribution of the ratio of the distance A between a source point and its reconstruction, by the distance B between the reconstructed source point and its nearest neighbor in the test set minus the source point.

To put this average test reconstruction error value (6%) into context, we computed the distribution of genetic distance separating two individuals (measured via Hamming distance, **FIG. 2.c**). Note that the reconstruction error applied to binary data is exactly the Hamming distance normalized by the sequence length, allowing direct comparisons. We found that the test reconstruction error (6%) is way below the minimal genetic distance in the considered dataset (13200, corresponding to $\sim 20\%$ of errors along the SNP sequence), indicating that the error caused by reconstruction is smaller than the distance between the closest individuals.

We refined this analysis at the individual level by investigating, for each reconstructed point, whether its nearest neighbor is the source point (desired behavior) or another individual of the test set **(FIG. 2.d-e)**. Specifically, we compared the (Hamming) distance

A (between a source point in the test set and its reconstructed point) to the distance B (between the reconstructed source point and its nearest neighbor in the test set excluding the source point). If $B < A$ then the reconstruction step locally distorts the data to the extent that the information at the sample level is lost in the process, *i.e.*, a reconstructed sample can be confused with another test sample through minor alterations. Such an unfaithful reconstruction would discourage to learn the generative model in the PC space. We found that this was not the case **(FIG. 2.e)**, as the distribution of the ratio $\frac{A}{B}$ was considerably lower than 1 (dotted vertical line), with a max value equal to 0.82. This further demonstrated that the information content of the reconstructed source point remains consistent with that of the source point.

Interestingly, the distribution of the reconstruction error appears to be bimodal. We discovered that samples with a higher reconstruction error were of African ancestry while samples from the rest of the world were mixed in the first mode (**FIG. 2.b**). This is likely due to the known higher genetic diversity present on this continent (explainable by human past demographic history) (Tishkoff and Williams, 2002; Campbell and Tishkoff, 2008; The 1000 Genomes Project Consortium et al., 2015).

We further investigated whether the low variance principal axes were relevant when studying the test set (i.e., whether they are better than random directions). For this, we selected n_{Comp} principal axes and constructed the remaining $2504 - n_{\text{Comp}}$ dimensions as random directions, by sampling Gaussian vectors $\mathcal{N}_k(\mathbf{0}, \mathbf{I})$ with $k = 65535$ and orthonormalizing them with respect to the first n_{Comp} axes (Boyd and Vandenberghe, 2018). From **FIG. S2**, we observe that adding random directions to the n_{Comp} first principal axes decreased the reconstruction error as expected, since there is more information when considering more directions, i.e., when the projection space gets larger. However, adding the next principal axes instead of random directions decreased the reconstruction error substantially more. This confirmed that the low variance principal axes contained relevant information for the reconstruction step.

3.2 Shrinkage phenomenon in high dimensional low sample size setting

We performed an optimization-free experiment, similarly to Section 3.1, where the test set, that serves as synthetic set, is used to understand the expected behavior of a distance-based metric, $\mathcal{AA}_{TS}$ (see Method 2.8), on generated samples.

After projecting the test set onto the basis obtained from PCA on a train set, we observed lower variances for PC scores of test samples than train samples (**FIG. S3**). This shrinkage phenomenon, as previously studied in (Lee et al., 2010, 2014), occurred in scenarios with high-dimensional data and low sample sizes. Specifically, it affected the PC scores of test samples not originally included in the PCA performed on a training set.

This phenomenon might lead to significant alterations in distances between the two sets in the PC space. As the density of the projected test set is shrunk with respect to PC scores of the train set (**FIG. S3**), the nearest neighbor of a projected test sample will mostly be a projected test sample rather than a train sample. Similarly, the nearest neighbor of a train sample will mostly be a projected test sample. That is, $\mathcal{AA}_{syn}$ is close to one while

$\mathcal{AA}_{truth}$ is close to zero (see scores for projected and reconstructed test in **FIG. S5**).

To investigate the double impact of shrinkage and space dimension, we computed $\mathcal{AA}$ scores on different slices of the PC space, where shrinkage always exists but can be more or less prominent (**FIG. S3**.) In all cases, we observed that $\mathcal{AA}truth$ and $\mathcal{AA}syn$ converged towards 0 and 1, respectively, quite rapidly, *i.e.*, generally for dimension 40 and higher, with a speed that depends on the expected shrinkage strength (**FIG. S4**).

To further confirm this effect of shrinkage on $\mathcal{AA}_{TS}$, we applied a dilating factor as a correction (see Method 2.7). This correction led to a better behaved $\mathcal{AA}_{TS}$ in PC space, where the different components were closer to the expected 0.5 value, as shown for projected test corrected on **FIG. S5**. It is worth mentioning here that $\mathcal{AA}_{TS}$ in PC space or SNP space without binarization is the same since PCA is just an orthogonal transformation. However, we found that the binarization step with a fixed threshold set to 0.5 was deteriorating the $\mathcal{AA}$ scores (see scores for reconstructed test corrected in **FIG. S5**).

Finally, we reproduced the conditions of the proposition (see Methods 2.8). We constructed two generic multivariate normal distributions $\mathcal{N}(0, \mathbb{I}_D)$ (serving as real data) and $\mathcal{N}(0, \lambda\mathbb{I}_D)$ (serving as synthetic data) with shrinkage factor $\lambda \in]0, 1[$, $D = 200$ and generated 5000 samples from each distributions. We observed convergence of $\mathcal{AA}_{truth}$ to 0 and $\mathcal{AA}_{syn}$ to 1, where the speed of convergence is controlled by λ. For extreme shrinkage, $\lambda = 0$, there is a first regime where both $\mathcal{AA}$ scores are equal to 1 (dark curves), then $\mathcal{AA}_{truth}$ suddenly drops as the number of dimensions increases (see **FIG. S8**).

With these experiments in mind, we will systematically analyze the two $\mathcal{AA}_{TS}$ components separately, as it allows to detect shrinkage on top of other generation anomalies (Yelmen et al., 2023).

3.3 Global patterns of genetic variation: principal component analysis and allelic frequencies

Real data exhibited a strong population structure pattern captured by PC1 to PC4 (**FIG. 3.a**), which was effectively replicated in AGs from all models, as can be seen through the 1D Wasserstein distance between the real and AG projections. Overall the error curve for PCA-WGAN was always below that of the RBM except for PC5, where the model was generating highly concentrated scores, yielding an increased Wasserstein distance. This concentration issue was not encountered in Light PCA-WGAN.

The synthetic datasets followed overall the same trends as the real one, in terms of allelic frequencies (**FIG. 3.d**). For frequencies ≤ 0.2, the correlation score between real and AGs ranged from 0.94 to 0.98. Nonetheless, PCA-WGAN fixed many alleles (20523 $\sim$ 31% of the sites; 15016 for Light PCA-WGAN; 12686 for Glocal-PCA-WGAN; 2672 for WGAN; 55 for RBM; 8 in real), especially the ones which are rare in the original dataset, similar to previous findings from different models (Yelmen et al., 2021, 2023).

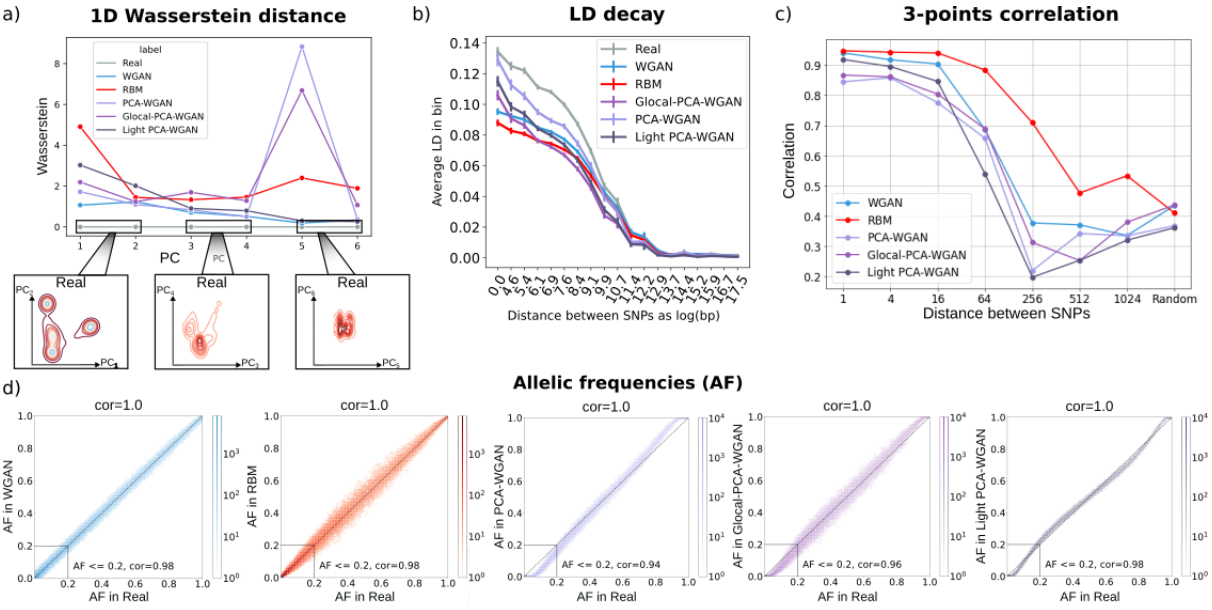

Figure 3: **Population genetics summary statistics. a)** The real and the synthetic data were concatenated into a single dataset, to which PCA is applied. The 1D Wasserstein distance is computed on each PCs. Real data is the grey curve, WGAN from (Yelmen et al., 2023) in blue, RBM from (Yelmen et al., 2023) in red, PCA-WGAN in light purple, Glocal-PCA-WGAN in dark purple and Light PCA-WGAN with NN in grey. The density plot of real data for the first six PCs is detailed in first column of **FIG.S13**. **b)** LD decay approximation. **c)** Three-points correlation separated by a varying amount of SNPs. **d)** 2D Histogram of the allelic frequencies in real data (x-axis) and in AGs produced by the models.

3.4 Local haplotypic structure: zooming on multi-locus interactions

The linkage disequilibrium (LD) curve shows how the correlation between pairs of SNPs decreases as a function of their physical distance (**FIG. 3.b**). PCA-WGAN reproduced these statistics better than all the other models, *i.e.*, the LD curve was the closest to the real curve, notably for short-range interactions. The lower LD observed for Glocal-PCA-WGAN was likely due to the uniform splitting of the SNP matrix (yielding blocks with the same number of SNPs). A potential future improvement in this regard might be to partition into regions that preserve LD blocks (Privé, 2021) (e.g., by splitting at recombination hotspots). Finally, Light PCA-WGAN fell in-between PCA-WGAN and Glocal-PCA-WGAN, and still yielded more accurate short range interactions than WGAN and RBM, even though it has one to two orders of magnitudes fewer parameters (**Table 1**).

In addition to LD, we computed the three-point correlation statistics for SNP triplets separated by *1, 4, 16, 64, 256, 512, 1024* and a random number of SNPs respectively (**FIG. 3.c**). While PCA-WGAN had lower correlations than RBM model in Yelmen et al. (2023), it remained close to the convolutional WGAN model, and demonstrated similar results to both models in the challenging scenario involving SNP triplets separated by random distances. Multiple local critics (Glocal-PCA-WGAN) improved this statistic for almost all SNP distances (except for a distance of 512 SNP) and even marginally surpassed WGAN and RBM benchmark models for SNPs separated by random distances.

Light PCA-WGAN performed better than PCA-WGAN and Glocal variant for short range interactions but its performance declined as the distance between SNPs increases.

After evaluating data quality with summary statistics proposed previously (Yelmen et al., 2021, 2023), we introduced a new one, based on SNP motifs. This allowed us to dissect more closely whether local haplotypic patterns are well preserved, which was previously mostly deduced from the 3-point correlation results. Artificial genomes unanimously exhibited higher motif diversity than in real data (see Method 2.9 and **FIG.4.A.**). For k-SNP motifs with $k \leq 16$, AGs from PCA-WGAN had a diversity lower than other generative models and closer to the real data, which is in line with LD of PCA-WGAN being the closest to real LD for smaller SNP distances. Indeed, highly correlated SNPs tightly constrain how the motif space should be populated. The diversity pattern shown in **FIG.4.A.** was maintained when we focused on each population independently, i.e., AGs had higher diversity than real sequences and those from PCA-WGAN were closer to the real data for $k \leq 16$ (**FIG.S12**).

We further fitted a power-law to the distribution of k-SNP motif diversity for k ranging from 4 to 2048, and retrieved the exponent as a measure of complexity of the data. A lower power-law exponent indicates higher diversity, i.e. more unique motifs. PCA-WGAN had the lowest exponent, which corresponds to a longer tail, while WGAN and RBM displayed a multimodal distribution, where bumps corresponds to particular k-SNP motif sizes (**FIG.4.C.**), i.e., these two latter models produced a diversity that was homogeneous along the haplotype windows. While k-SNP motif hallucination seemed to be a salient feature of generative models (**FIG.4.A.**), PCA-WGAN and its alternative in fact possessed (through the power law exponent) the highest diversity (**FIG.4.C.**). k-SNP motifs of size $k \geq 32$ were likely contributing to this inflated diversity for PCA-WGAN and its Light variant compared to other models.

We further computed Shannon's entropy over each window of size k and compared it to a Bernoulli model having parameters equal to the frequency at each site of the real data. Note that for random k-SNP motifs drawn from a uniform distribution, the entropy will be equal to k. Since a k-SNP motif can take on 2^k possible states, and our real and synthetic datasets provided 5008 samples per window, the distribution might not be well sampled for $k \geq 12$, *i.e.*, when there are more than $2^{12} = 4096$ states. However, rare k-SNP motifs that are missed would not have significantly affected the entropy, as long as they are not too numerous. Entropy calculations showed that a substantial amount of local haplotypic structure was present in real and AGs data compared to random sequences (from the Bernoulli model). Furthermore, k-SNP motifs in AGs were slightly less structured than in real genomes (**FIG. 4.B.**).

3.5 Measuring overfitting in generative models

We evaluated how well the AGs coming from the different models perform in terms of resembling real data (**FIG. 5**). Coincidentally, the $\mathcal{AA}_{TS}$ score of PCA-WGAN and glocal variant showed a behavior similar to the one described for the test set, *i.e.*, $\mathcal{AA}_{syn}$ close to one and $\mathcal{AA}_{truth}$ close to zero (see Subsection 3.2). This is likely due to the exceedingly narrow density of the generated points along the late PCs. However, unlike the shrinkage observed in the projected and reconstructed test set, this effect is most likely driven

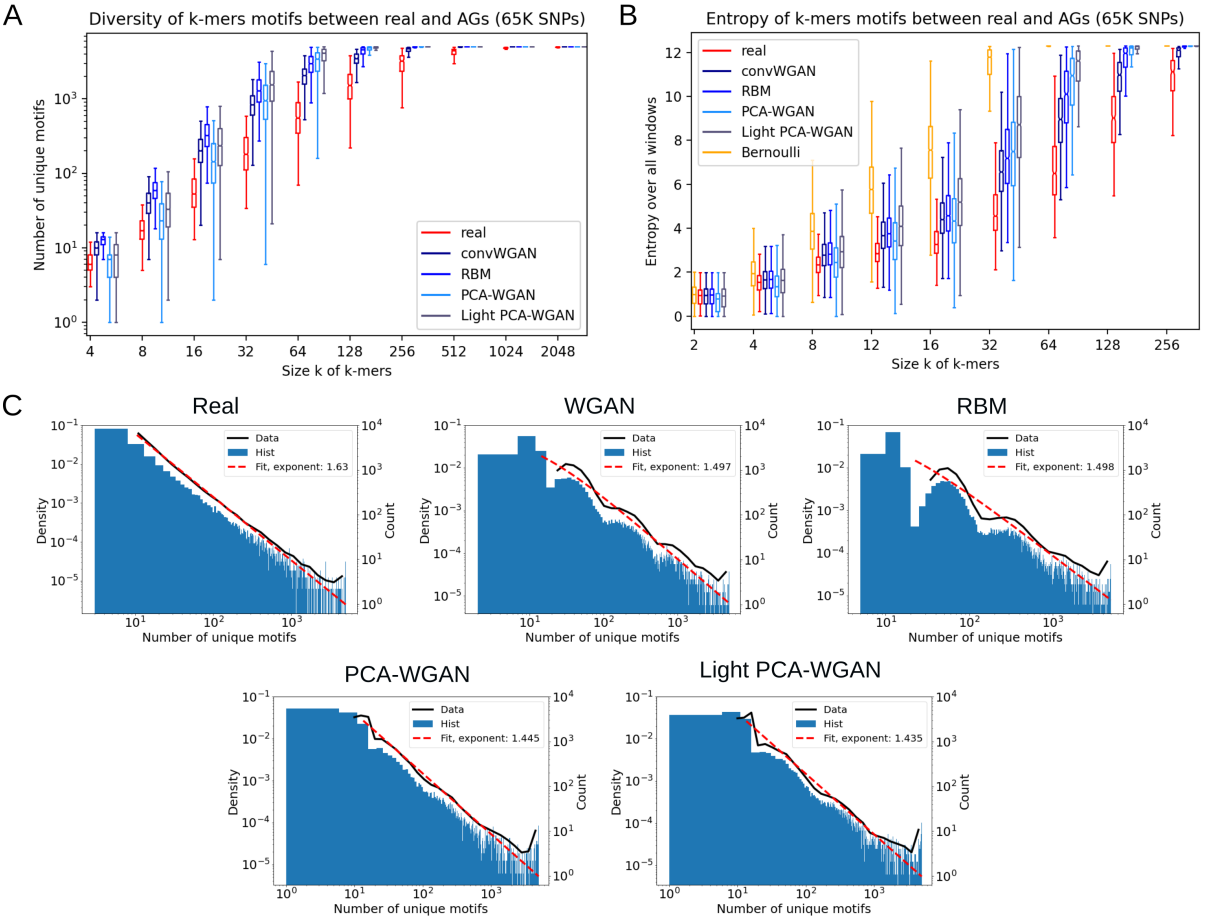

Figure 4: **Diversity & entropy of k-SNP motifs. A.** Distribution of the number of unique motifs over all non-overlapping windows of size k. Diversity of k-SNP motif per window in real data (red), and the synthetic counterpart (blue) for WGAN, RBM (Yelmen et al., 2023), PCA-WGAN and Light PCA-WGAN. **B.** Shannon's entropy computed over all non-overlapping windows of size k. The Bernoulli model has parameters equal to the frequency at each site of the real data. **C.** Histogram of the diversity of k-SNP motifs for multiple k, with the probability density function as black curve and power-law fit as dashed red line. The power-law exponent is $1.63, 1.497, 1.498, 1.445, 1.435$ for Real, WGAN, RBM, PCA-WGAN and Light PCA-WGAN respectively.

by the Wasserstein distance favoring shrunk modeled distributions (see Section 4.3 and FIG.5 in Montavon et al. (2016)). After correcting for the undesired shrinkage effect (see Method 2.7) the score improved not only over the PCA-WGAN but also over the WGAN and RBM from (Yelmen et al., 2023). This enhancement of $\mathcal{AA}_{TS}$ came at the cost of slightly worse LD statistics, staying nonetheless above that of WGAN and RBM for short range interactions (**FIG. S7 Top right**). On the other hand, it also alleviated the high Wasserstein distance to the real AGs on PC5 and the allele fixing issue (**FIG. S7 Top left and Bottom**). Finally, Light PCA-WGAN palliated the difficulties arising from training with Wasserstein loss in high-dimensional space, with $100\times$ fewer parameters than PCA-WGAN (**Table 1**), and achieved the best $\mathcal{AA}_{TS}$ score among all models.

Dissecting the $\mathcal{AA}_{TS}$ of PCA-WGAN generated samples in PCA space, *i.e.* before their transformation into the genomic space, we also found imbalanced $\mathcal{AA}_{syn}$ (close to 1) and

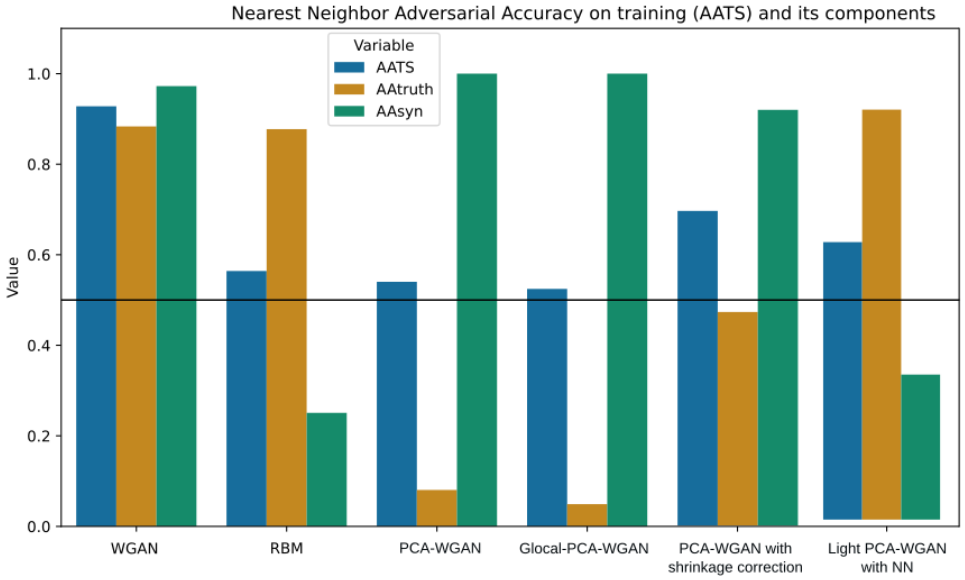

Figure 5: **Nearest neighbor adversarial accuracy.** $\mathcal{AA}_{TS}$ score for AGs was computed in the data space. The black line indicates the optimum value whereas values below the line indicate overfitting and values above the line indicate underfitting.

$\mathcal{AA}_{truth}$ (close to 0) scores. This further confirmed that the shrinkage effect was due to Wasserstein training as previously explained. However, $\mathcal{AA}_{TS}$ of Light PCA-WGAN with NN in PCA space displayed preferable behavior w.r.t resemblance (**FIG. S6**). Similarly to Subsection 3.2, we identified that the binarization step with a fixed threshold set to 0.5 is deteriorating the $\mathcal{AA}_{TS}$ score (**FIG.S6**). In an effort to circumvent both issues of allelic fixation (demonstrated in Subsection 3.3) and poor $\mathcal{AA}_{TS}$, we applied a frequency-based binarization (see Methods 2.6) instead of a hard threshold. This frequency-based binarization yielded allelic frequencies that were identical w.r.t real data, yet the $\mathcal{AA}_{TS}$ behaved badly, i.e., $\mathcal{AA}_{TS} \sim 0.5$, $\mathcal{AA}_{truth} \sim 1$ and $\mathcal{AA}_{syn} \sim 0$. Hence, a fixed threshold set to 0.5 was preferred.

3.6 Computational efficiency comparison

We compared training compute resources between different models (**Table 1**), as well as the number of parameters as a function of sequence length (**FIG.S11**). The CRBM in (Yelmen et al., 2023) consisted of 13 RBMs with 10M parameters each, where each RBM was trained during $\sim$ 10 hours (wall clock time). The training time of the whole set of RBMs depends on the number of available GPUs, and can benefit from parallelization (//), *i.e.*, with a sufficient number of GPUs, the training time can be at best $\sim$ 10 hours. While PCA-WGAN had more parameters than WGAN, its training time remained almost an order of magnitude smaller. Glocal PCA-WGAN had the highest number of parameters, but still was faster to train than WGAN. Light PCA-WGAN with NN had 100$\times$ fewer parameters than PCA-WGAN and 20$\times$ fewer parameters than WGAN, with the fastest training time, on par with Light PCA-WGAN with a multivariate normal distribution (MVN). That last model (see *Appendix S7*) provided similar results on the population genetics summary statistics and $\mathcal{AA}_{\mathcal{TS}}$ metric as Light PCA-WGAN with NN, only requiring 58K parameters, *i.e.*, 1275$\times$ less than PCA-WGAN or 275$\times$ less than WGAN.

Table 1: Training GPU & runtime comparison between models for 65K SNP dataset

Models	**Params**	**GPUs**	**Wall clock**
PCA-WGAN	74M	1-A100-40GB	20 hours
Glocal-PCA-WGAN	140M	1-A100-40GB	2 days
Light PCA-WGAN with MVN	58K	1-A100-40GB	∼2 hours
Light PCA-WGAN with NN	750K	1-A100-40GB	∼2 hours
CRBM	130M	≥1-RTX3090-24GB	13×10 hours (in //)
WGAN	16M	1-A100-40GB	6-8 days

4 Discussion and conclusion

This study tackled scaling generative models for high-dimensional SNP data with PCA-WGAN and its derivative models, and assessed AG quality through population genetics summary statistics, privacy-related scores and information theoretic metrics. To our knowledge, while dimension reduction is routinely used as a ML preprocessing step, its application in generative modeling for genomics is unexplored, making our reduced-space genomic data generation simple and innovative. Moreover, no other studies benchmarked their models on a very high-dimensional tabular dataset, except for models in (Yelmen et al., 2023) and the models presented in this current work. For example, Dang et al. (2023) demonstrated good performances for the 805 SNPs dataset (Colonna et al., 2014; Yelmen et al., 2021) but weaker ones for 10K SNPs, while *DNAGPT* Zhang et al. (2024) demonstrated performances comparable to (Yelmen et al., 2023) for 10K SNPs but did not scale up to 65K SNPs. Despite PCA-WGAN having a drawback in allelic frequencies, it achieved results comparable to recent advances (Yelmen et al., 2021, 2023) and even captured two-point SNP correlations (LD) better. On the other hand, Glocal-PCA-WGAN improved the allele fixing issue since the PCA compression is lighter, but produced worse results in terms of LD. This could potentially be improved by finding optimal splits for blocks with respect to LD (Privé, 2021). Furthermore, Glocal-PCA-WGAN could scale to longer SNP sequences. As we showed that PCA-WGAN on 65K SNPs offered competitive results, we could envision a Glocal-PCA-WGAN processing 3×65K SNPs. In this case, keeping 40% of PCs for each block would lead to the same parameter size as in the current Glocal-PCA-WGAN for 65K SNPs (**FIG.S5**). Finally, we showcased Light PCA-WGAN as a viable frugal alternative to complex sequence modeling architectures, obtaining similar results to large models with order of magnitudes fewer parameters.

As generative modeling becomes more widespread in various domains from natural language processing to image synthesis, defining the desired properties of synthetic data becomes more important (Alaa et al., 2022; van Breugel et al., 2023; van Breugel and van der Schaar, 2023). However, generative models often suffer from multiple shortcomings, one of them, *mode shrinkage*, consists in the concentration of the synthetic data on the modes of the real data without covering the full range (Stein et al., 2023; Yamaguchi and Fukuda, 2023; Montavon et al., 2016). Despite the effectiveness of shrinkage correction and Light PCA-WGAN in addressing the nearest neighbor adversarial accuracy issue, these experiments underscored the significance of distinguishing $\mathcal{AA}_{truth}$ from $\mathcal{AA}_{syn}$ components within $\mathcal{AA}_{TS}$. This decomposition becomes particularly crucial as it reveals instances where the overall adversarial accuracy ($\mathcal{AA}_{TS}$) appears satisfactory, while

individual components $\mathcal{AA}_{truth}$ and $\mathcal{AA}_{syn}$ exhibit poor performance. In addition, we showed that $\mathcal{AA}_{TS}$, being distance-based, may lead to spurious results under certain circumstances, in particular when there are generation anomalies (Yelmen et al., 2023), over- or under-dispersion of samples due to shrinkage (*Proposition 2.8*), mode collapse, or drops in diversity. The adjustment of the PC scores via a dilating factor improved $\mathcal{AA}_{TS}$ at the cost of slightly lowering the quality of AGs is sub-optimal and could be refined (see Method 2.7). Indeed, the non alignment of eigenvectors across train and test datasets induces a dilation in inaccurate directions. Finally, we observed that different set of hyperparameter values for the NN paired with Light PCA-WGAN (Figure S9) yielded substantially different results for $\mathcal{AA}_{TS}$. Instead of manual or grid search for the best set of hyperparameter values (described in Supplementary material S6), AutoML algorithms could be employed in future research.

Analysing the k-SNP motifs along a haplotype revealed simple yet discriminative feature between real and synthetic data. Namely, given a fixed window of size k, AGs contained significantly more motifs than in real data. Despite having their discriminative features, AGs were found useful for multiple tasks (e.g. imputation, selection scan, or ancestry inference) (Yelmen et al., 2021; Montserrat et al., 2019). This raises the question of balancing overall data quality with its utility for specific applications. Currently, there is no universally accepted indicator to effectively evaluate this balance, underscoring the need for further methodological developments. Indeed, classification algorithms that disentangle empirical from simulated data have recently highlighted the unrealistic nature of sequences generated by evolutionary models (Trost et al., 2023). However, there remains a lack of methodology to gauge the practical utility of these recognizable synthetic data.

Overall, PCA-WGAN is a promising novel methodology that combines simple dimensionality reduction and generative modeling to produce diverse and longer artificial genomes, making it a potentially valuable tool for enhancing genome-wide analyses with synthetic data, especially when access to real datasets is limited.

5 Acknowledgements

This work benefited from Inria TAU computing resources and ANR-20-CE45-0010-01 RoDAPoG funding. We thank Michèle Sebag for insightful discussions.

References

Burak Yelmen and Flora Jay. An Overview of Deep Generative Models in Functional and Evolutionary Genomics. *Annual Review of Biomedical Data Science*, 6(1):173–189, August 2023. ISSN 2574-3414, 2574-3414. doi: 10.1146/annurev-biodatasci-020722-115651.

Kevin Korfmann. Deep Learning in Population Genetics. *Genome Biology and Evolution*, 15(2), February 2023. ISSN 0090-5364. doi: 10.1093/gbe/evad008.

Xin Huang, Aigerim Rymbekova, Olga Dolgova, Oscar Lao, and Martin Kuhlwilm. Harnessing deep learning for population genetic inference. *Nature Reviews Genetics*, September 2023. ISSN 1471-0056, 1471-0064. doi: 10.1038/s41576-023-00636-3.

C J Battey, Gabrielle C Coffing, and Andrew D Kern. Visualizing population structure with variational autoencoders. *G3 Genes/Genomes/Genetics*, 11(1):jkaa036, March 2021. ISSN 2160-1836. doi: 10.1093/g3journal/jkaa036.

Jonas Meisner and Anders Albrechtsen. Haplotype and population structure inference using neural networks in whole-genome sequencing data. *Genome Research*, 32(8):1542–1552, August 2022. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.276813.122.

Margarita Geleta, Daniel Mas Montserrat, Xavier Giro-i-Nieto, and Alexander G Ioannidis. Deep Variational Autoencoders for Population Genetics. *bioRxiv : the preprint server for biology*, 2023. doi: 10.1101/2023.09.27.558320.

Zhanpeng Wang, Jiaping Wang, Michael Kourakos, Nhung Hoang, Hyong Hark Lee, Iain Mathieson, and Sara Mathieson. Automatic inference of demographic parameters using generative adversarial networks. *Molecular Ecology Resources*, 21(8):2689–2705, 2021. doi: https://doi.org/10.1111/1755-0998.13386. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.13386`.

Rebecca Riley, Iain Mathieson, and Sara Mathieson. Interpreting generative adversarial networks to infer natural selection from genetic data. *Genetics*, 226(4):iyae024, 02 2024. ISSN 1943-2631. doi: 10.1093/genetics/iyae024. URL `https://doi.org/10.1093/genetics/iyae024`.

Burak Yelmen, Aurélien Decelle, Linda Ongaro, Davide Marnetto, Corentin Tallec, Francesco Montinaro, Cyril Furtlehner, Luca Pagani, and Flora Jay. Creating artificial human genomes using generative neural networks. *PLOS Genetics*, 17(2):e1009303, February 2021. ISSN 1553-7404. doi: 10.1371/journal.pgen.1009303.

Maria Perera, Daniel Mas Montserrat, Mıriam Barrabes, Margarita Geleta, Xavier Giro-i-Nieto, and Alexander G Ioannidis. Generative Moment Matching Networks for Genotype Simulation. *bioRxiv*, 2022. doi: 10.1101/2022.04.14.488350. URL `https://www.biorxiv.org/content/early/2022/04/14/2022.04.14.488350`.

William W Booker, Dylan D Ray, and Daniel R Schrider. This population does not exist: learning the distribution of evolutionary histories with generative adversarial networks. *GENETICS*, 224(2), April 2023. ISSN 1943-2631. doi: 10.1093/genetics/iyad063. URL `http://dx.doi.org/10.1093/genetics/iyad063`.

Burak Yelmen, Aurélien Decelle, Leila Lea Boulos, Antoine Szatkownik, Cyril Furtlehner, Guillaume Charpiat, and Flora Jay. Deep convolutional and conditional neural networks for large-scale genomic data generation. *PLOS Computational Biology*, 19(10):e1011584, October 2023. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1011584.

Meihua Dang, Anji Liu, Xinzhu Wei, and Sriram Sankararaman. Tractable and Expressive Generative Models of Genetic Variation Data. *bioRxiv*, 2023. doi: 10.1101/2023.05.16.541036. URL `https://www.biorxiv.org/content/early/2023/05/18/2023.05.16.541036`.

Daoan Zhang, Weitong Zhang, Bing He, Jianguo Zhang, Chenchen Qin, and Jianhua Yao. DNAGPT: A Generalized Pretrained Tool for Multiple DNA Sequence Analysis Tasks, 2024. URL `https://www.biorxiv.org/content/early/2024/01/04/2023.07.11.548628`.

Callum Burnard, Alban Mancheron, and William Ritchie. Generating realistic artificial Human genomes using adversarial autoencoders, 2023. URL `https://www.biorxiv.org/content/early/2023/12/09/2023.12.08.570767`.

Shaked Ahronoviz and Ilan Gronau. Genome-AC-GAN: Enhancing Synthetic Genotype Generation through Auxiliary Classification. *bioRxiv*, 2024. doi: 10.1101/2024.02.14.580420. URL https://www.biorxiv.org/content/early/2024/02/16/2024.02.14.580420.

Antoine Szatkownik, Cyril Furtlehner, Guillaume Charpiat, Burak Yelmen, and Flora Jay. Towards creating longer genetic sequences with gans: Generation in principal component space. In David A. Knowles and Sara Mostafavi, editors, *Proceedings of the 18th Machine Learning in Computational Biology meeting*, volume 240 of *Proceedings of Machine Learning Research*, pages 110–122. PMLR, 30 Nov–01 Dec 2024. URL https://proceedings.mlr.press/v240/szatkownik24a.html.

Zhan Su, Jonathan Marchini, and Peter Donnelly. HAPGEN2: Simulation of multiple disease SNPs. *Bioinformatics*, 27(16):2304–2305, August 2011. ISSN 1367-4811, 1367-4803. doi: 10.1093/bioinformatics/btr341.

You Tang and Xiaolei Liu. G2P: A Genome-Wide-Association-Study simulation tool for genotype simulation, phenotype simulation and power evaluation. *Bioinformatics*, 35(19):3852–3854, October 2019. ISSN 1367-4803, 1367-4811. doi: 10.1093/bioinformatics/btz126.

Apostolos Dimitromanolakis, Jingxiong Xu, Agnieszka Krol, and Laurent Briollais. sim1000G: A user-friendly genetic variant simulator in R for unrelated individuals and family-based designs. *BMC Bioinformatics*, 20(1):26, December 2019. ISSN 1471-2105. doi: 10.1186/s12859-019-2611-1.

Sophie Wharrie, Zhiyu Yang, Vishnu Raj, Remo Monti, Rahul Gupta, Ying Wang, Alicia Martin, Luke J O'Connor, Samuel Kaski, Pekka Marttinen, Pier Francesco Palamara, Christoph Lippert, and Andrea Ganna. HAPNEST: Efficient, large-scale generation and evaluation of synthetic datasets for genotypes and phenotypes. *Bioinformatics*, 39(9):btad535, September 2023. ISSN 1367-4811. doi: 10.1093/bioinformatics/btad535.

Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. URL https://arxiv.org/abs/1406.2661.

P. Smolensky. Information processing in dynamical systems: Foundations of harmony theory. In *Parallel distributed processing: Explorations in the microstructure of cognition*, pages 194–281–. MIT Press, Cambridge, MA, 1986.

David H. Ackley, Geoffrey E. Hinton, and Terrence J. Sejnowski. A learning algorithm for Boltzmann machines. *Cognitive Science*, 9:147–169, 1985.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. URL https://arxiv.org/abs/2006.11239.

Kristiina Ausmees and Carl Nettelblad. A deep learning framework for characterization of genotype data. *G3 Genes/Genomes/Genetics*, 12(3):jkac020, March 2022. ISSN 2160-1836. doi: 10.1093/g3journal/jkac020.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation, 2021. URL https://arxiv.org/abs/2102.12092.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. URL https://arxiv.org/abs/2112.10752.

Latent diffusion models: Is the generative ai revolution happening in latent space?, 2023. URL https://neurips2023-ldm-tutorial.github.io/.

Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based generative modeling in latent space, 2021. URL https://arxiv.org/abs/2106.05931.

C. Tian, P. K. Gregersen, and M. F. Seldin. Accounting for ancestry: Population substructure and genome-wide association studies. *Human Molecular Genetics*, 17(R2):R143–R150, October 2008. ISSN 0964-6906, 1460-2083. doi: 10.1093/hmg/ddn268.

Alkes L. Price, Noah A. Zaitlen, David Reich, and Nick Patterson. New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, 11(7):459–463, July 2010. ISSN 1471-0056, 1471-0064. doi: 10.1038/nrg2813.

Oscar Lao, Timothy T. Lu, Michael Nothnagel, Olaf Junge, Sandra Freitag-Wolf, Amke Caliebe, Miroslava Balascakova, Jaume Bertranpetit, Laurence A. Bindoff, David Comas, Gunilla Holmlund, Anastasia Kouvatsi, Milan Macek, Isabelle Mollet, Walther Parson, Jukka Palo, Rafal Ploski, Antti Sajantila, Adriano Tagliabracci, Ulrik Gether, Thomas Werge, Fernando Rivadeneira, Albert Hofman, André G. Uitterlinden, Christian Gieger, Heinz-Erich Wichmann, Andreas Rüther, Stefan Schreiber, Christian Becker, Peter Nürnberg, Matthew R. Nelson, Michael Krawczak, and Manfred Kayser. Correlation between Genetic and Geographic Structure in Europe. *Current Biology*, 18(16):1241–1248, August 2008. ISSN 09609822. doi: 10.1016/j.cub.2008.07.049.

John Novembre, Toby Johnson, Katarzyna Bryc, Zoltán Kutalik, Adam R. Boyko, Adam Auton, Amit Indap, Karen S. King, Sven Bergmann, Matthew R. Nelson, Matthew Stephens, and Carlos D. Bustamante. Genes mirror geography within Europe. *Nature*, 456(7218):98–101, November 2008. ISSN 1476-4687. doi: 10.1038/nature07331.

Kai Yu, Zhaoming Wang, Qizhai Li, Sholom Wacholder, David J. Hunter, Robert N. Hoover, Stephen Chanock, and Gilles Thomas. Population Substructure and Control Selection in Genome-Wide Association Studies. *PLoS ONE*, 3(7):e2551, July 2008. ISSN 1932-6203. doi: 10.1371/journal.pone.0002551.

P. Menozzi, A. Piazza, and L. Cavalli-Sforza. Synthetic Maps of Human Gene Frequencies in Europeans. *Science*, 201(4358):786–792, September 1978. doi: 10.1126/science.356262.

Nick Patterson, Alkes L. Price, and David Reich. Population Structure and Eigenanalysis. *PLOS Genetics*, 2(12):e190, December 2006. ISSN 1553-7404. doi: 10.1371/journal.pgen.0020190.

Olivier Hanotte, Daniel G. Bradley, Joel W. Ochieng, Yasmin Verjee, Emmeline W. Hill, and J. Edward O. Rege. African Pastoralism: Genetic Imprints of Origins and Migrations. *Science*, 296(5566):336–339, April 2002. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1069878.

Alkes L. Price, Nick J. Patterson, Robert M. Plenge, Michael E. Weinblatt, Nancy A. Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909, August 2006. ISSN 1546-1718. doi: 10.1038/ng1847.

The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, October 2010. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature09534.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN, December 2017. URL https://arxiv.org/abs/1701.07875.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved Training of Wasserstein GANs, December 2017. URL https://arxiv.org/abs/1704.00028.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. URL https://openreview.net/pdf?id=BJJsrmfCZ.

Cédric Villani. *Optimal Transport*. Springer Berlin Heidelberg, 2009. ISBN 9783540710509. doi: 10.1007/978-3-540-71050-9. URL http://dx.doi.org/10.1007/978-3-540-71050-9.

D.A. Nix and A.S. Weigend. Estimating the mean and variance of the target probability distribution. In *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, volume 1, pages 55–60 vol.1, June 1994. doi: 10.1109/ICNN.1994.374138.

Seunggeun Lee, Fei Zou, and Fred A. Wright. Convergence and prediction of principal component scores in high-dimensional settings. *The Annals of Statistics*, 38(6), December 2010. ISSN 0090-5364. doi: 10.1214/10-AOS821.

Seunggeun Lee, Fei Zou, and Fred A. Wright. Convergence of Sample Eigenvalues, Eigenvectors, and Principal Component Scores for Ultra-High Dimensional Data. *Biometrika*, 101(2):484–490, June 2014. ISSN 0006-3444. doi: 10.1093/biomet/ast064.

Andrew Yale, Saloni Dash, Ritik Dutta, Isabelle Guyon, Adrien Pavao, and Kristin P. Bennett. Generation and evaluation of privacy preserving synthetic health data. *Neurocomputing*, 416: 244–255, November 2020. ISSN 09252312. doi: 10.1016/j.neucom.2019.12.136.

George Stein, Jesse C. Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Leigh Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L. Caterini, J. Eric T. Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models, 2023. URL https://arxiv.org/abs/2306.04675.

Shin'ya Yamaguchi and Takuma Fukuda. On the limitation of diffusion models for synthesizing training datasets, 2023. URL https://arxiv.org/abs/2311.13090.

Grégoire Montavon, Klaus-Robert Müller, and Marco Cuturi. Wasserstein Training of Restricted Boltzmann Machines. 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/728f206c2a01bf572b5940d7d9a8fa4c-Paper.pdf.

Sarah A. Tishkoff and Scott M. Williams. Genetic analysis of African populations: Human evolution and complex disease. *Nature Reviews Genetics*, 3(8):611–621, August 2002. ISSN 1471-0064. doi: 10.1038/nrg865.

Michael C. Campbell and Sarah A. Tishkoff. African Genetic Diversity: Implications for Human Demographic History, Modern Human Origins, and Complex Disease Mapping. *Annual Review of Genomics and Human Genetics*, 9(1):403–433, September 2008. ISSN 1527-8204, 1545-293X. doi: 10.1146/annurev.genom.9.081307.164258.

The 1000 Genomes Project Consortium, Corresponding authors, Adam Auton, Gonçalo R. Abecasis, Steering committee, David M. Altshuler, Richard M. Durbin, Gonçalo R. Abecasis, David R. Bentley, and Chakravarti ... A global reference for human genetic variation. *Nature*, 526(7571):68–74, October 2015. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature15393.

Stephen Boyd and Lieven Vandenberghe. *Introduction to Applied Linear Algebra: Vectors, Matrices, and Least Squares*. Cambridge University Press & Assessment, 1 edition, June 2018. ISBN 978-1-108-58366-4 978-1-316-51896-0. doi: 10.1017/9781108583664.

Florian Privé. Optimal linkage disequilibrium splitting. *Bioinformatics*, 38(1):255–256, December 2021. ISSN 1367-4803, 1367-4811. doi: 10.1093/bioinformatics/btab519.

Vincenza Colonna, Qasim Ayub, Yuan Chen, Luca Pagani, Pierre Luisi, Marc Pybus, Erik Garrison, Yali Xue, Chris Tyler-Smith, and The 1000 Genomes Project Consortium. Human genomic regions with exceptionally high levels of population differentiation identified from 911 whole-genome sequences. *Genome Biology*, 15(6):R88, 2014. ISSN 1465-6906. doi: 10.1186/gb-2014-15-6-r88.

Ahmed M. Alaa, Boris van Breugel, Evgeny Saveliev, and Mihaela van der Schaar. How Faithful is your Synthetic Data? Sample-level Metrics for Evaluating and Auditing Generative Models, July 2022. URL `https://arxiv.org/abs/2102.08921`.

Boris van Breugel, Zhaozhi Qian, and Mihaela van der Schaar. Synthetic data, real errors: How (not) to publish and use synthetic data, July 2023. URL `https://arxiv.org/abs/2305.09235`.

Boris van Breugel and Mihaela van der Schaar. Beyond Privacy: Navigating the Opportunities and Challenges of Synthetic Data, April 2023. URL `https://arxiv.org/abs/2304.03722`.

Daniel Mas Montserrat, Carlos Bustamante, and Alexander Ioannidis. Class-conditional vae-gan for local-ancestry simulation, 2019. URL `https://arxiv.org/abs/1911.13220`.

Johanna Trost, Julia Haag, Dimitri Höhler, Laurent Jacob, Alexandros Stamatakis, and Bastien Boussau. Simulations of Sequence Evolution: How (Un)realistic They Are and Why. *Molecular Biology and Evolution*, 41(1):msad277, 12 2023. ISSN 1537-1719. doi: 10.1093/molbev/msad277. URL `https://doi.org/10.1093/molbev/msad277`.