Exploring the use of Artificial Genomes for Genome-wide Association Studies through the lens of Utility and Privacy

Xinyue Wang[1], **Sitao Min**[1], **Jaideep Vaidya**[1]
[1]**Rutgers University, Newark, NJ**

Abstract

Collaborative Genome-wide association studies (GWAS) have the potential to uncover rare genetic variant-trait associations by leveraging larger datasets and diverse population samples. Despite this potential, privacy concerns and cumbersome review processes for data validation and collaborator selection hinder their broader implementation. Advances in generative models present a possible solution by generating synthetic datasets that closely resemble real genomic data, thus enhancing privacy and expediting the review process. This study assesses the capability of deep generative models to produce artificial genomic data for GWAS applications. We evaluate two state-of-the-art models on real-world datasets, identifying significant limitations in their ability to generate high-quality artificial genomes. Furthermore, we demonstrate that prevailing privacy measures, mainly based on membership inference attacks, are inadequate for providing insightful privacy evaluations. Our findings highlight the critical challenges and suggest future directions for the effective use of artificial genomes in GWAS.

Introduction

Recent advancements in bioinformatics and Next-generation sequencing have highlighted the potential of large-scale collaborative efforts, which entail the joint analysis of datasets distributed across various institutions and organizations. Facilitate collaborative Genome-wide Association Studies (GWAS), a widely used technique to uncover associations between specific genotypes and phenotypes by examining common genetic variants among individuals, can address the key limitations of single-site GWAS with overestimated effects [1] and have demonstrated enhanced capabilities in detecting rare genetic variations and yield accurate findings [2]. However, previous studies mainly adopt the meta-analysis approach by only aggregating summary statistics and, therefore, face challenges such as selection bias and heterogeneity within meta-analyses. The explanatory power of these collaborative GWAS findings for the heritability of common traits remains limited [3], suggesting that future studies may require more extensive collaboration and participant involvement [4]. Nevertheless, the adoption of collaborative GWAS is greatly hindered by the privacy concerns surrounding sharing personal genetic data. The unique nature of genetic data [5], such as DNA sequences, coupled with its implications for not only the individual but also their relatives, demands stringent measures to prevent unauthorized information disclosure in compliance with regulations like GDPR and HIPAA.

To facilitate the collaborative GWAS, researchers (potential collaborators) typically undergo a lengthy institutional review board (IRB) process to exchange raw datasets that include individuals' genomes, demographic data, and disease status. Synthetic data, artificially generated by specific algorithms and models, could expedite the preliminary phases of research (e.g., with less stringent requirements) and address collaboration challenges before full IRB approval (e.g., exploratory results on the synthetic data can be used to identify the valuable researchers and useful data [6]). Additionally, once trained, these models can produce unlimited artificial data, potentially improving the diversity of genomic datasets. The use of synthetic data has been explored in various domains such as vision [7] and has been used in fields such as labor statistics [8] and healthcare [9], their application in collaborative GWAS is still overlooked. Without careful studies and evaluations, adopting artificial genomes can raise critical issues concerning data quality [10] and the potential for privacy breaches [11], especially when employing synthetic data derived from personal information, as highlighted by the controversy over DeepMind's use of NHS patient records. [1]

This paper presents the first thorough evaluation of cutting-edge generative models for genomic data within GWAS, focusing on their capacity to mimic real genomic data while safeguarding original data integrity. Previous evaluations [12] primarily focus on the traditional statistical methods [13] and deep generative models such as WGAN [14], which

[1]BBC News: DeepMind faces legal action over NHS data use, 2021. https://www.bbc.com/news/technology-58761324

are unable to handle the discrete characteristics of genomic data and lack a comprehensive evaluation using only low-order statistics and simple privacy measures. In this work, on the one hand, we examine two specific models, namely CTGAN [15] and DNADiffusion [16], that can handle discrete and high dimensional data, as characterized by genomic data. On the other hand, we employ various metrics, encapsulating low-order statistics, t-SNE [17] visualizations, performance on two GWAS downstream tasks, and five state-of-art membership inference attacks against synthetic data deep generative models to evaluate the fidelity, utility, and privacy of artificial genomes. Through extensive evaluations, our findings indicate that there is still a significant gap in the application of artificial genomes in GWAS, and more sophisticated privacy measures tailored for synthetic genomic data are needed. In terms of utility, while CTGAN shows promise in preserving data distributions, both models struggle to produce high-quality synthetic SNP sequences that maintain phenotype-SNP correlations. Furthermore, we observe discrepancies in model performance across different genomic data groups, raising questions about the practical utility of artificial genomes in GWAS. The privacy evaluation results reveal that, surprisingly, none of the MIAs achieve inference powers better than random guessing on any of the synthetic datasets created by CTGAN and DNADiffusion, raising concerns for the efficiency of existing MIA attacks, hence the need for the development of more discriminative privacy attacks.

Background

SNPs and GWAS. The complete human DNA consists of over 3 billion nucleotide pairs, with 99.9% of them shared among humans. Single Nucleotide Polymorphisms (SNPs) are the most common genetic variants, with two possible alleles : (i) major allele, the most common nucleotide in the population, and (ii) minor allele, the rarer nucleotide. SNPs contain complex information and diverse patterns (e.g., haplotype blocks) and are studied by GWAS, a computational method to identify genomic variants that are statistically associated with a risk for a disease or a particular trait, by screening the entire genome of large numbers of individuals to look for associations between millions of genetic variants. In the genome array, a SNP is encoded by 0, 1, or 2 to indicate the number of minor alleles found in a specific genetic position. The association between each SNP and the phenotype encoded by a binary indicator (for the binary-trait case) is tested through statistical methods such as the χ^2 test and logistic regression.

Datasets We employ two real-world genomic datasets from the 1000 Genomes Project [18] and OpenSNP [19]. For the *1000 Genomes* dataset, we implement standard quality control measures, including the exclusion of SNPs and individuals with missing data, the removal of SNPs with a Minor Allele Frequency (MAF) lower than 0.1, and the elimination of related individuals. While the original data only contains the genomic array, we simulate a phenotype vector using Genetic Complex Trait Analysis (GCTA) [20]. To evaluate the efficacy of DGMs across varying sequence lengths and SNP groups, we randomly select around 500 significant SNPs and around 500 insignificant SNPs (employing a significance threshold of 0.05) for 2,400 individuals using Plink 2.0 [21]. Note that we chose a relatively relaxed significance threshold to guarantee that the size of the significant SNP group was large enough. We maintain the significant-to-insignificant SNP ratio while varying the SNP count per individual to 200, 500, and 1,000. We also execute a train-test split with training set ratios of 0.25, 0.5, and 0.75, altering the seed at each division to produce five distinct splits. For the *OpenSNP* dataset, we adhere to a comparable processing protocol, with two notable exceptions: (1) we utilize binarized eye color as the phenotype instead of a simulated phenotype, and (2) given that the dataset is comprised of 940 individuals, the training set ratios are adjusted to 0.5 and 0.75.

Methods

This section provides an overview of the state-of-the-art models for generating synthetic datasets and the utility and privacy measures to evaluate their performance.

Deep generative models for artificial genomes

Unlike traditional methods, deep generative models learn the underlying data structure in a data-driven manner without needing user-defined mathematical equations or physics simulations. With the neural networks' aptitude to approximate complex non-linear relationships in data, DGMs have demonstrated superior performance in many applications. In this work, we explore two generative models, namely CTGAN (Conditional Tabular Generative Adversarial Net-

work) [15] and DNADiffusion [16].

CTGAN [15] incorporates a mode-specific normalization strategy to handle the multimodal distributions problem in the feature and a conditional vector and novel sampling techniques to handle categorical features. CTGAN was initially proposed for generating tabular data and has achieved good performance in various applications, such as Electronic Health Records (EHR) data [22]. This work is the first to explore the applicability of CTGAN for artificial genome generation, leveraging its ability to handle categorical data to effectively categorize the discrete characteristics of genomic data.

DNADiffusion [16] is a novel approach leveraging diffusion probabilistic models to design cell type-specific DNA regulatory sequences. The backbone of the model is a denoising UNet [23] with two embedding layers for cell label and timestep, respectively. To handle the discrete nature of DNA data, before training, the DNA sequences are converted to numerical forms using a standard technique, one-hot encoding. After training, the model takes in input a cell type label and can generate cell type-specific sequences. Unlike other methods, such as Dirichlet Diffusion [24] DNADiffusion allows training a single model across different cell types and generating cell type-specific sequences without the need for additional models as guidance. In the implementation, we use the phenotype label as the condition to generate phenotype-specific SNP sequences.

For CTGAN, we utilize the default settings from Synthcity [25], including 256 hidden nodes and two layers for both the generator and discriminator. For all the datasets, we train CTGAN for 2000 epochs, although convergence is typically observed after several hundred epochs. For DNADiffusion, we adhere to the implementation described in [16] but limit the diffusion process to 10 steps and training epochs to 1000. Nevertheless, we have explored different diffusion time steps (from 10 to 100) and training epochs (from 1000 to 10000) and have yet to observe significant improvement in the results. Upon completion of training, we generate 5000 synthetic samples.

Evaluation measures

To explore the potential of artificial human genomes in the context of exploratory analysis for GWAS, we select a suite of evaluation metrics, providing insights into utility and privacy.

Utility measures. There are three categories of measurements considered. (1) Two low-order statistics, namely Minor Allele Frequency (MAF) and Heterozygosity, are pivotal in GWAS to offer critical insights into the allele and population statistics in the population. After computing the MAF values of the real (i.e., training set) SNPs and the synthetic SNPs, we first examined the single-SNP level performance by plotting the MAF values at each locus. Then, we employ the relative mean absolute error (RMAE) and the Kolmogorov-Smirnov (K-S) test to evaluate the performance quantitatively on the entire dataset. The RMAE assesses the average deviations in MAF across all generated SNPs from those of actual SNP sequences. The K-S test, a non-parametric method, determines if the MAFs of real SNP sequences and artificial SNPs originate from the same distribution. A larger K-S statistic and correspondingly lower p-value indicate the rejection of the null hypothesis that the MAF values from original and synthetic SNPs are drawn from identical distributions. (2) t-SNE plots [17], widely used for visualizing high-dimensional data in lower-dimensional spaces, are adopted for visualization evaluation of the fidelity and diversity of the synthetic datasets. (3) Two fundamental downstream tasks in GWAS, imputation and association testing, are employed to demonstrate the usefulness of synthetic datasets, aiming to discern performance variations when employing artificial genomic datasets (SNPs and phenotypes) for GWAS. For the imputation task, we adopt the train-synthetic-test-real(TSTR) and train-real-test-real (TRTR) techniques, and measure the performance deviations. Specifically, we introduce missing data in the test dataset, following [26], adjusting the missing rate (θ) to encompass [10%, 20%, 30%], randomly selecting θ of the samples. Within these samples, 5% of SNPs are designated missing, following the Missing Completely At Random (MCAR) mechanism [27]. Once the missing data is introduced, we train an imputation model using the training and synthetic data separately and perform the imputation on the test dataset. We selected the KNN-based imputation model for its proven efficacy [28] and Root Mean Square Error (RMSE) to assess imputation accuracy. Since we aim to assess the comparative performance the DGMs, we measure the RMSE relative to that of training data, i.e., RMSE obtained using the training data and report proportional RMSE (p-RMSE), computed as $|\text{RMSE}_{\text{TSTR}} - \text{RMSE}_{\text{TRTR}}|/\text{RMSE}_{\text{TRTR}}$.

As for the association testing task, we adopt the logistic regression test. The p-values and significance labels (i.e., significant or insignificant) from synthetic datasets and the original dataset (utilizing all samples before the train-test split) are compared with Mean Absolute Error (MAE) and accuracy rate, respectively.

Privacy measures. With the various metrics proposed for evaluating the privacy risks of synthetic datasets, in this work, privacy is examined through the lens of a suite of Membership Inference Attacks (MIAs), for they are prevalent in the context of bioinformatics. MIAs aim to infer whether an individual's data has been used in training a generative model, potentially leading to more invasive privacy breaches such as profiling and feature inference, and have been used in other works [29]. We explore various MIA strategies, typically predicated on the "black-box" scenario where the adversary has access only to the synthetic dataset and a specific target (i.e., record) but may also have a reference dataset that might not align with the training data's distribution. This situation is particularly relevant to bioinformatics, where datasets are often specific to certain demographic groups. Our privacy evaluation encompasses both GAN-specific and more broadly applicable attacks, including:

1. *LOGAN-0* and *LOGAN-D1* [30]. *LOGAN-0* assumes that the adversary has a synthetic dataset produced by the generative model but no access to the model. The synthetic dataset is then used to train a local GAN (we employed CTGAN during the experiments). Once trained, the discriminator from the GAN is used to distinguish the synthetic and real data. *LOGAN-D1*, conversely, assumes the adversary has access to a reference dataset, and then uses it to train a discriminative model for membership inference. Following [11], we implemented a three-layer fully connected neural network as the discriminative model, following the approach in prior work.

2. *GAN-Leaks* [31]. This approach categorizes MIAs against GANs across various scenarios, from fully black-box to white-box settings. We focus on *GAN-Leak 0*, a strict black-box attack. To launch the attack, the adversary generates a set of samples from the generative model, and then estimates the likelihood of a test point belonging to the training set. In particular, the attack strategy involves generating a sample set $S_G^k = \{x_i\}_{i=1}^k$ from the generator G for a test point x^* and a chosen $k \in \mathbb{N}$, utilizing the score $A(x^*, G) = \exp(-\min_{x_i \in S_G^k} L_2(x^*, x_i))$ as an unnormalized surrogate for the probability that the target is generated by the generator. If the score is greater than some predefined threshold, then the target x^* is classified as belong to the training set.

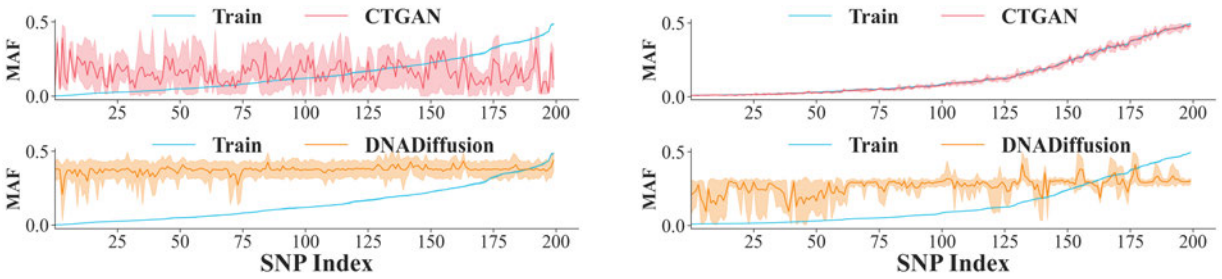
3. Monte Carlo-based MIA (*MC*) [32] Applicable to a wide range of generative models, this attack uses synthetic samples produced by the generative model to perform a Monte Carlo approximation of the model's distribution at each test point. Similar to *GAN-Leak 0*, a set of synthetic samples generated by the target generative model is used to estimate the likelihood of the target record used for training the generator through approximating the distribution of the small neighborhood of the target record.

4. *DOMIAS* [11] A density-based MIA designed for general generative models, comparing density estimates of real and synthetic distributions to identify overfitting in generative models. In particular, given the real data distribution $p_R(X)$ and the synthetic data distribution $p_G(X)$, for a target x^*, with the MIA scoring function $\mathcal{A}_{DOMIAS}(x^*) = f\left(\frac{p_G(x^*)}{p_R(x^*)}\right)$, where $\mathcal{A}$ represents the attack function and $f : \mathbb{R} \to [0, 1]$ is a monotonically increasing function.If the score exceeds some predefined threshold, then the target x^* is classified as belong to the training set. In the implementation, we employ a Gaussian Kernel Density Estimator (KDE) to approximate these distributions.

Results

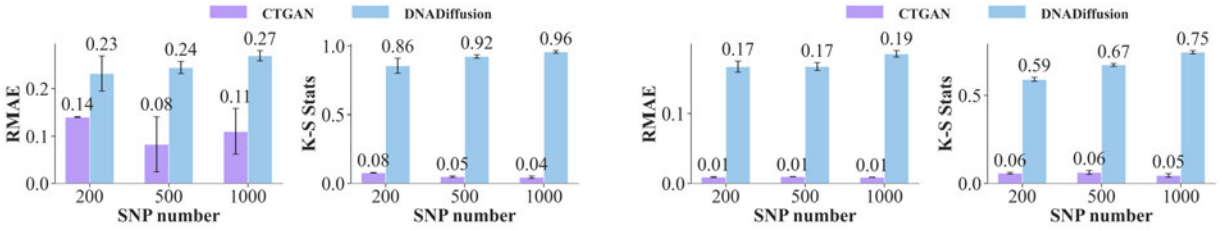
Utility evaluation

In this section, we provide a comprehensive utility evaluation of the performance of CTGAN and DNADiffusion on the two datasets. Firstly, we examine whether artificial SNPs preserve allele and population statistics. For ease of comparison, in Figure 1 and the subsequent figures, we sort the SNPs by their MAF values; nevertheless, during the training and generation, the order of SNPs is randomly shuffled. Figure 1 depicts the MAF values of 200 SNPs from both the training set and artificial genomes generated by CTGAN and DNADiffusion across the *OpenSNP* and



(a) *OpenSNP* (b) *Human1000*

Figure 1: MAF for 200 SNPs generated by CTGAN and DNADiffusion, plotted against the training data, when training ratio = 0.75. Solid lines represent the average values and the bands represent the bounds across five iterations.

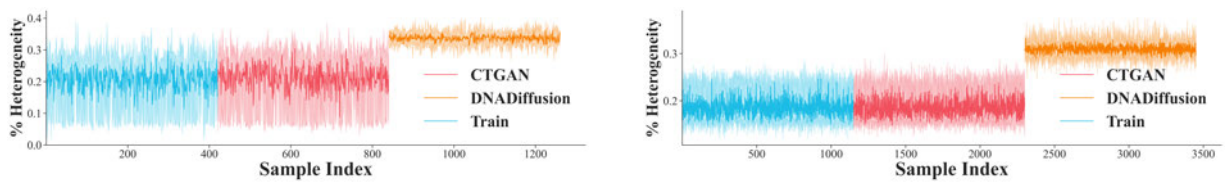


(a) *OpenSNP* (b) *Human1000*

Figure 2: RMAE between the MAF values of all real SNPs and SNPs and K-S test statistics. Average results over five runs with training ratio = 0.75.

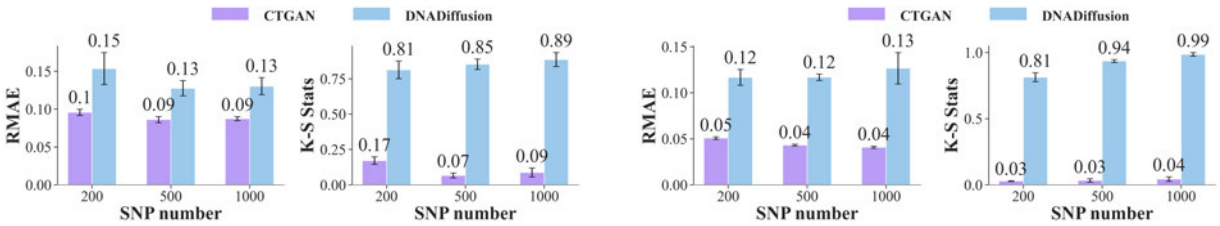
Human1000 datasets. The comparison reveals that CTGAN more accurately replicates allele frequencies than DNADiffusion overall. DNADiffusion exhibits a notable deficiency in preserving the MAF for most SNPs, with a pronounced lack of variation across most loci across both datasets. However, the performance of CTGAN across the two datasets is significantly different, suggesting the intricate nature of genomic data. After the visual comparison on the single-SNP level, the results, presented in Figure 2, demonstrate that for both RMAE and K-S test outcomes, CTGAN surpasses DNADiffusion in generating artificial SNP sequences that more closely resemble real ones in terms of allele frequencies. Similarly, Figure 3 and 4 plot the percentage of heterozygous variants in each individual and the corresponding RMAE and K-S test outcomes, respectively. While DNADiffusion produces more homozygous variants despite discrepancies in minor allele frequencies, indicating a larger diversity in the population, CTGAN produces heterozygosity percentages similar to the real dataset.

Next, we leverage t-SNE plots to visually compare the synthetic dataset generated by CTGAN and DNADiffusion with the original training set. Given genomic arrays' high-dimensional nature, we initially apply Principal Component Analysis (PCA) to both the synthetic and training datasets independently, concentrating on the top 100 principal components (PCs). This reduction is followed by t-SNE analysis on these PCs to visualize the data distributions. The better overlapping of the original and synthetic data points in each t-SNE plot shows the better synchronization of the original and learned distribution. Figure 5 showcases t-SNE plots for the *OpenSNP* and *Human1000* populations across varying training ratios and SNP sequence lengths. CTGAN consistently outperforms DNADiffusion on the *Human1000* dataset by generating data that more closely aligns with the original distribution. However, for the *OpenSNP*



(a) *OpenSNP* (b) *Human1000*

Figure 3: Percentage of heterozygous variants in each sample in the training and generated datasets, using 1000 SNPs.



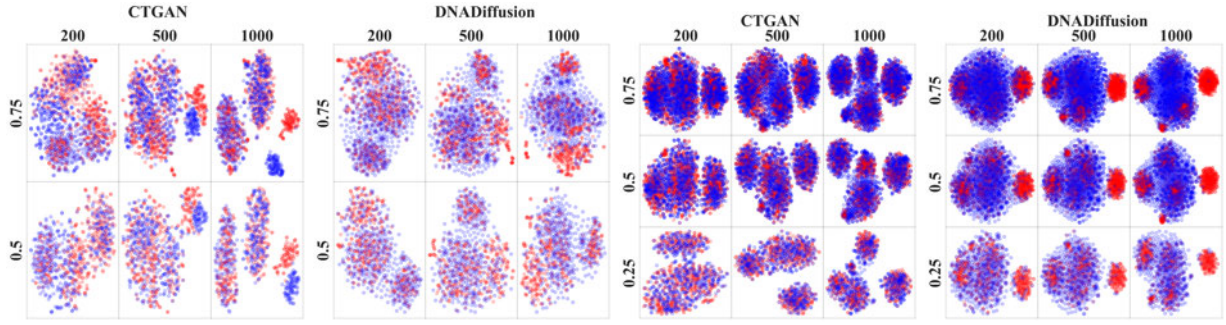
(a) *OpenSNP* (b) *Human1000*

Figure 4: RMAE between the heterozygosity values of real and artificial samples and average K-S test statistics. Results with training ratio = 0.75.

Tasks	θ	*OpenSNP* CTGAN	*OpenSNP* DNADiffusion	*Human1000* CTGAN	*Human1000* DNADiffusion
Imputation p-RMSE	10%	0.169 (0.114)	0.65 (0.063)	0.018 (0.009)	0.593 (0.017)
	20%	0.174 (0.128)	0.653 (0.03)	0.014 (0.005)	0.579 (0.023)
	30%	0.181 (0.138)	0.647 (0.036)	0.013 (0.004)	0.586 (0.024)
LR-test	MAE	0.533 (0.025)	0.490 (0.022)	0.614 (0.02)	0.524 (0.018)
	Accuracy	0.563 (0.038)	0.555 (0.027)	0.662 (0.033)	0.453 (0.036)

Table 1: Downstream tasks performance when SNP number = 500 and training ratio = 0.75. θ represents the missing ratio from the imputation tasks. Standard deviation over five runs is given in the bracket.

dataset, neither CTGAN nor DNADiffusion fully captures the complexity of the real data.

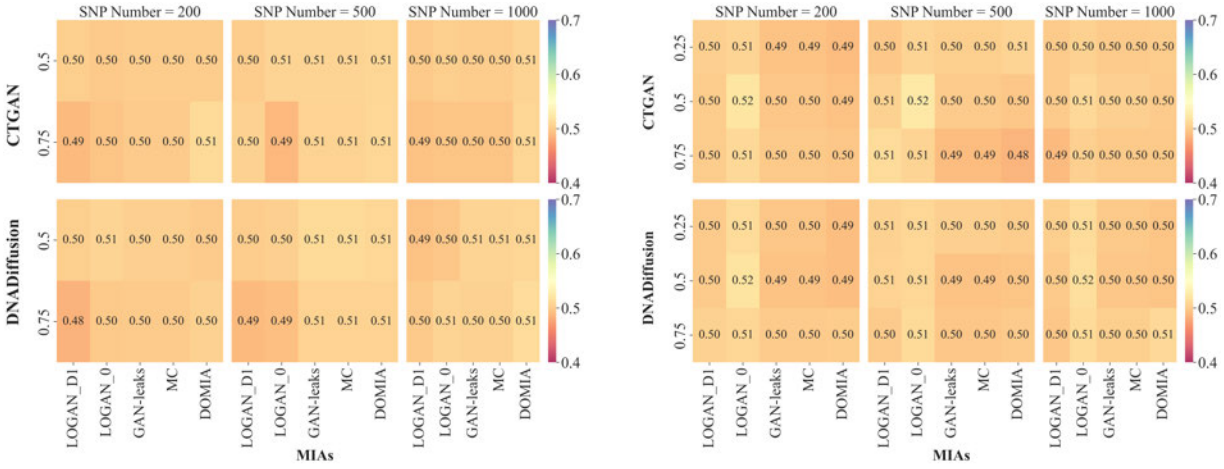


(a) (b) (c) (d)

Figure 5: t-SNE plots across various training ratio (row) and SNP numbers (column). Red: real records. Blue: synthetic records. (a) and (b): *OpenSNP*, (c) and (d): *Human1000*.

Finally, we present the downstream tasks performances of CTGAN and DNADiffusion in Table 1. On the imputation task, we report the average p-RMSE and its standard deviation over five iterations for both CTGAN and DNADiffusion. The results show CTGAN's better performance by achieving similar RMSE to the real data for both *OpenSNP* and *Human1000* datasets. However, discrepancies across datasets are again observed. As for the association testing task, CTGAN again outperforms DNADiffusion. Nevertheless, CTGAN's best-case scenario preserves the significance status for only about 66% of SNPs (on *Human1000*) despite aligning with real data in terms of allele frequencies and heterozygosity.

The above results show that CTGAN consistently generates better artificial SNP sequences than DNA diffusion on both datasets. However, neither model can create SNPs that most resemble the real SNPs regarding the performance of preserving the allele frequencies and downstream tasks. During the experiments, we also observed that while using more training samples improves the overall quality of synthetic artificial, it results in larger MAFs deviant of some SNPs. The significantly different performances of CTGAN on *OpenSNP* and *Human1000*, which may be due to the intricate distributions of the underlying samples, including the population structures and the imbalanced phenotypes, exposes the limitations of these models to model SNP sequences.



(a) *OpenSNP* (b) *Human1000*

Figure 6: Heatmap of the average AUROC of five MIAs

Privacy evaluation

For the privacy evaluation, we implement the five MIAs, discussed in the previous section. Since *DOMIAS* with Gaussian KDE functions require continuous data, we employ PCA first and select the top 100 PCs. While the other MIAs can be implemented in the original space, we employed the same PCA procedure, as no significant performance degradations for these attacks were observed during the experiments. For the evaluation, every sample in the training set is considered a target for each attack. We calculate the scores of each sample to estimate the likelihood of sample inclusion in the training set. We use the median of the scores as the threshold for *GAN-Leaks*, *MC*, and *DOMIAS*. We then compute the accuracy and Area Under the Receiver Operating Characteristic (AUROC) over five iterations. Figure 6 presents the average AUROC and its variability for the *OpenSNP* and *Human1000* datasets. The results suggest a generally low efficacy of MIAs comparable to random guessing across most scenarios. Notably, the performance of specific MIAs, like *LOGAN-0*, varies across datasets and training ratios, indicating inconsistent vulnerability to MIAs. Surprisingly, despite additional information from a reference dataset, attacks like *DOMIAS* and *LOGAN-D1* do not consistently outperform others, with *LOGAN-0* showing superior performance in some cases on *Human1000*.

Several potential reasons exist for the poor performance. On the one hand, all the five attacks except *DOMIA* rely on the distance-based similarity measures and use the distribution of the closest distance between each record and the datasets as the proximity for the closeness of the original and synthetic data distribution, which may not be suitable for the high-dimensional genomic data. *DOMIA*, though directly comparing the distance between the generated and original distribution, can still struggle to obtain accurate approximations of these distributions due to the high dimensionality. On the other hand, these MIAs share the assumption that the generative models overfit may fail to capture the specific overfitting phenomenon [33].

Performance disparity on different groups of SNPs

In practice, a prevalent imbalance exists between SNPs significantly associated with phenotypes and those that are not, in terms of the numbers, the coupled p-values, and the implications in the GWAS. To better understand whether or not the generative models can pick up different signals inherent in these SNPs, we examine the performances separately for significant and insignificant SNP sets. Firstly, Figure 7 and Figure 8 present the RMAE and K-S statistics of the MAF values and t-SNE plots for these SNP groups across *OpenSNP* and *Human1000*. Notably, CTGAN exhibits superior performance on significant SNPs within the *Human1000* populations, whereas its performance inversely correlates with the *OpenSNP* dataset. Conversely, DNADiffusion performs better on insignificant SNPs within the *Human1000* dataset than the significant SNPs. Next, we compare the downstream task performances of two SNP groups. Table 2 shows a distinct performance contradiction between CTGAN and DNADiffusion. CTGAN successfully retains

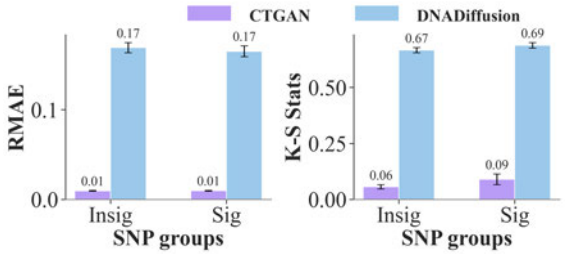


Figure 7: RMAE between the MAF values of real and artificial SNPs and average K-S test statistics on different SNPs groups. Results on *Human1000* with 500 SNPs and training ratio = 0.75.

		CTGAN		DNADiffusion	
Dataset	TrainRatio	*Sig*	*Insig*	*Sig*	*Insig*
OpenSNP	0.5	0.872 (0.033)	0.328 (0.050)	0.450 (0.356)	0.552 (0.339)
	0.75	0.889 (0.037)	0.311 (0.053)	0.156 (0.118)	0.864 (0.136)
Human1000	0.25	0.761 (0.203)	0.403 (0.113)	0.187 (0.099)	0.808 (0.087)
	0.5	0.771 (0.162)	0.444 (0.123)	0.173 (0.023)	0.834 (0.024)
	0.75	0.784 (0.108)	0.499 (0.094)	0.188 (0.118)	0.807 (0.080)

Table 2: LR-test task performance (measured by accuracy rate) on significant SNPs (*Sig*) and insignificant SNPs (*Insig*), when both model are trained with 500 SNPs. Standard deviation over five run is given in the bracket.

the significance status of approximately 80% of artificial SNPs within significant groups but only about 30% within insignificant groups. In contrast, DNADiffusion shows a pronounced improvement in performance on insignificant SNPs compared to significant ones. As for the privacy evaluation, we conduct the same five MIAs, using samples exclusively with significant or insignificant SNPs. Figure 9 shows that the divergent performances of CTGAN and DNADiffusion persist. However, all the attacks fail to achieve higher power than random guessing.

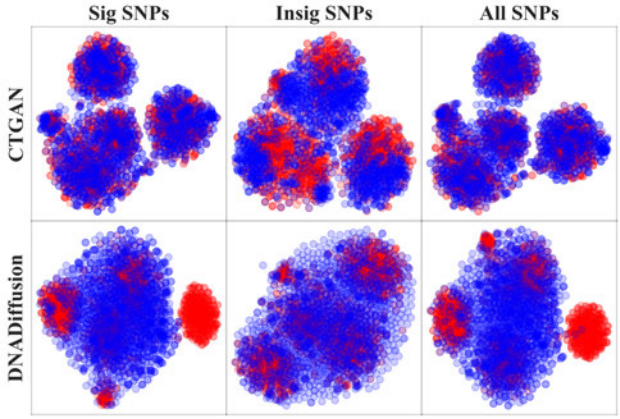


Figure 8: t-SNE plots on significant and insignificant SNPs. Red: real records. Blue: synthetic records. Results on *Human1000* with 500 SNPs and training ratio = 0.75.

Conclusions and discussions

This study conducts a thorough examination of artificial genomic data generation using two state-of-the-art generative algorithms (CTGAN and DNADiffusion) for GWAS applications. The findings highlight considerable limitations in both the utility and privacy of the generated artificial genomes. Specifically, both models struggle to produce artificial genomes that are consistently useful for core GWAS tasks. Furthermore, the performance discrepancy across SNP groups, particularly when these are categorized by their relationship to specific phenotypes, underscores the need for the development of new generative models capable of accurately modeling complex SNP distributions and their

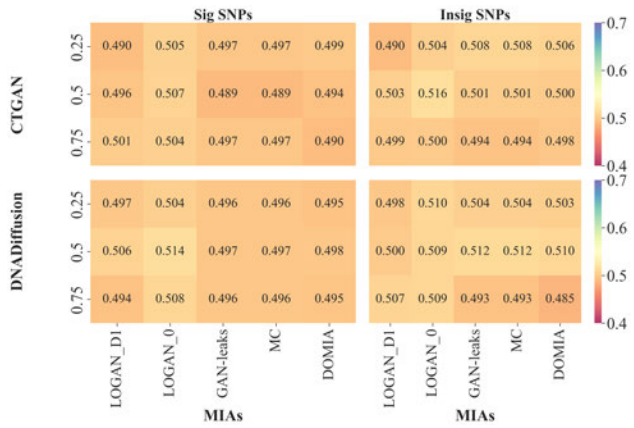


Figure 9: Heatmap of the average AUROC of five MIAs on different sets of SNPs, results on *Human1000*

associations with phenotypes.

In terms of privacy, the inadequacy of the five MIAs is twofold. First, their ineffectiveness across all scenarios demonstrates a fundamental challenge in addressing high-dimensional genomic data. This could be attributed to the inadequate similarity measures used, which fail to offer a meaningful assessment of "closeness," and the challenge of verifying the presumption of overfitting in neural networks. Second, these privacy metrics cannot offer clear insights into the privacy-utility tradeoffs and, more critically, fail to distinguish between the performances of different generative models. To improve the privacy assessment of artificial genomic data, future work may adopt Identity By Descent (IBD) measures, which quantify genetic similarities, as a more nuanced approach to evaluating the "closeness" between artificial and real genomes. Additionally, addressing the limitations related to overrepresentation assumptions might involve a deeper investigation into data-copying issues [33].

Acknowledgments

Research reported in this publication was supported by the National Institutes of Health under award number R35GM134927 and R01LM014520 as well as the National Science Foundation under award number CNS-2333225. The content is solely the responsibility of the authors and does not necessarily represent the official views of the agencies funding the research. We would also like to thank the anonymous reviewers for their comments.

References

1. Kim MS, Patel KP, Teng AK, Berens AJ, Lachance J. Genetic disease risks can be misestimated across global populations. Genome biology. 2018;19:1-14.
2. Lander ES. Initial impact of the sequencing of the human genome. Nature. 2011;470(7333):187-97.
3. Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D. Benefits and limitations of genome-wide association studies. Nature Reviews Genetics. 2019;20(8):467-84.
4. Smeland OB, Frei O, Dale AM, Andreassen OA. The polygenic architecture of schizophrenia—rethinking pathogenesis and nosology. Nature Reviews Neurology. 2020;16(7):366-79.
5. Cooper DN, Smith BA, Cooke HJ, Niemann S, Schmidtke J. An estimate of unique DNA sequence heterozygosity in the human genome. Human genetics. 1985;69:201-5.
6. Vaidya J, Shafiq B, Asani M, Adam N, Jiang X, Ohno-Machado L. A scalable privacy-preserving data generation methodology for exploratory analysis. In: AMIA Annual Symposium Proceedings; 2017. p. 1695.
7. Niemeyer M, Geiger A. Giraffe: Representing scenes as compositional generative neural feature fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021. p. 11453-64.
8. Jarmin RS, Louis TA, Miranda J. Expanding the role of synthetic data at the US Census Bureau. Statistical Journal of the IAOS. 2014;30(2):117-21.
9. Davis P, Lay-Yee R, Pearson J. Using micro-simulation to create a synthesised data set and test policy options:

The case of health service effects under demographic ageing. Health Policy. 2010;97(2-3):267-74.
10. Kodali N, Abernethy J, Hays J, Kira Z. On convergence and stability of gans. arXiv preprint arXiv:170507215. 2017.
11. van Breugel B, Sun H, Qian Z, van der Schaar M. Membership inference attacks against synthetic data through overfitting detection. arXiv preprint arXiv:230212580. 2023.
12. Oprisanu B, Ganev G, De Cristofaro E. On utility and privacy in synthetic genomic data. arXiv preprint arXiv:210203314. 2021.
13. Samani SS, Huang Z, Ayday E, Elliot M, Fellay J, Hubaux JP, et al. Quantifying genomic privacy via inference attack with high-order SNV correlations. In: 2015 IEEE Security and Privacy Workshops. IEEE; 2015. p. 32-40.
14. Killoran N, Lee LJ, Delong A, Duvenaud D, Frey BJ. Generating and designing DNA with deep generative models. arXiv preprint arXiv:171206148. 2017.
15. Xu L, Skoularidou M, Cuesta-Infante A, Veeramachaneni K. Modeling tabular data using conditional gan. Advances in neural information processing systems. 2019;32.
16. Ferreira DaSilva L, Senan S, Patel ZM, Reddy AJ, Gabbita S, Nussbaum Z, et al. DNA-Diffusion: Leveraging Generative Models for Controlling Chromatin Accessibility and Gene Expression via Synthetic Regulatory Elements. bioRxiv. 2024:2024-02.
17. Maaten Lvd, Hinton G. Visualizing data using t-SNE. Journal of machine learning research. 2008;9(Nov):2579-605.
18. Consortium GP, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68.
19. Greshake B, Bayer PE, Rausch H, Reda J. OpenSNP-a crowdsourced web resource for personal genomics. PLOS ONE. 2014;9(3):e89204. Available from: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0089204.
20. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. The American Journal of Human Genetics. 2011;88(1):76-82.
21. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4(1):s13742-015.
22. Sun C, van Soest J, Dumontier M. Generating synthetic personal health data using conditional generative adversarial networks combining with differential privacy. Journal of Biomedical Informatics. 2023:104404.
23. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer; 2015. p. 234-41.
24. Avdeyev P, Shi C, Tan Y, Dudnyk K, Zhou J. Dirichlet Diffusion Score Model for Biological Sequence Generation. arXiv preprint arXiv:230510699. 2023.
25. Qian Z, Cebere BC, van der Schaar M. Synthcity: facilitating innovative use cases of synthetic data in different data modalities; 2023. Available from: https://arxiv.org/abs/2301.07573.
26. Fu W, Wang Y, Wang Y, Li R, Lin R, Jin L. Missing call bias in high-throughput genotyping. BMC genomics. 2009;10(1):1-14.
27. Rubin DB. Inference and missing data. Biometrika. 1976;63(3):581-92.
28. Petrazzini BO, Naya H, Lopez-Bello F, Vazquez G, Spangenberg L. Evaluation of different approaches for missing data imputation on features associated to genomic data. BioData mining. 2021;14(1):1-13.
29. Lyu L, Chen C. A novel attribute reconstruction attack in federated learning. arXiv preprint arXiv:210806910. 2021.
30. Hayes J, Melis L, Danezis G, De Cristofaro E. Logan: Membership inference attacks against generative models. arXiv preprint arXiv:170507663. 2017.
31. Chen D, Yu N, Zhang Y, Fritz M. Gan-leaks: A taxonomy of membership inference attacks against generative models. In: Proceedings of the 2020 ACM SIGSAC conference on computer and communications security; 2020. p. 343-62.
32. Hilprecht B, Härterich M, Bernau D. Monte Carlo and Reconstruction Membership Inference Attacks against Generative Models. Proc Priv Enhancing Technol. 2019;2019(4):232-49.
33. Meehan C, Chaudhuri K, Dasgupta S. A non-parametric test to detect data-copying in generative models. In: International Conference on Artificial Intelligence and Statistics; 2020. .