Population-scale Genomic Data Augmentation Based on Conditional Generative Adversarial Networks

Junjie Chen
Temple University
Philadelphia, PA, USA
junjie.chen2019@temple.edu

Mohammad Erfan Mowlaei
Temple University
Philadelphia, PA, USA
erfan.molaei@gmail.com

Xinghua Shi*
Temple University
Philadelphia, PA, USA
mindyshi@temple.edu

ABSTRACT

Although next generation sequencing technologies have made it possible to quickly generate a large collection of sequences, current genomic data still suffer from small data sizes, imbalances, and biases due to various factors including disease rareness, test affordability, and concerns about privacy and security. In order to address these limitations of genomic data, we develop a Population-scale Genomic Data Augmentation based on Conditional Generative Adversarial Networks (PG-cGAN) to enhance the amount and diversity of genomic data by transforming samples already in the data rather than collecting new samples. Both the generator and discriminator in the PG-CGAN are stacked with convolutional layers to capture the underlying population structure. Our results for augmenting genotypes in human leukocyte antigen (HLA) regions showed that PC-cGAN can generate new genotypes with similar population structure, variant frequency distributions and LD patterns. Since the input for PC-cGAN is the original genomic data without assumptions about prior knowledge, it can be extended to enrich many other types of biomedical data and beyond.

CCS CONCEPTS

• **Applied computing → Computational genomics**.

KEYWORDS

machine learning; deep learning; generative adversarial networks; data augmentation; genomics

ACM Reference Format:
Junjie Chen, Mohammad Erfan Mowlaei, and Xinghua Shi*. 2020. Population-scale Genomic Data Augmentation Based on Conditional Generative Adversarial Networks. In *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB '20), September 21–24, 2020, Virtual Event, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3388440.3412475

1 INTRODUCTION

A large volume of genomic sequences collected at population levels has become quickly available, thanks to the recent progress in sequencing technologies that surpasses the Moore's law. However, human genomic data is still rather limited in several ways that prevent the wide adoption of artificial intelligence and machine learning in solving complex computational challenges in human genetics and genomics. First, although a large collection of human genomes has been sequenced in consortia projects (e.g. the 1000 Genome Project [12], the Genome 10K Project [21], UK Biobank [34]) and clinical settings, it is still costly and sometimes infeasible (e.g. in rare disease studies where samples are limited) to rapidly accumulate sufficient genomes for developing modern AI-empowered data analytics that rely heavily on big data. Second, data imbalance or bias are another key challenge resulting from factors such as racical distribution, disease rareness, and test affordability. The majority of genomes available are from populations with European ancestries and other populations are under-represented in current genomic data. Third, genomic sequences and genetic data are heritable by nature, and thus contain sensitive and private information about individuals and their relatives [23]. Thus, partially due to concerns about privacy and security, human genomic data are not readily accessible or widely shared.

In the meanwhile, biological systems are usually complex with interactions or crosslinks reflecting linear and non-linear relationships among various components and across many layers. Modern machine learning methods, especially deep learning [25], are empowered to reveal such complex interactions by mining massive data sets, and significantly facilitate scientific discovery and biomedical research. Recent studies reported that these interactions in genomic data can be captured using deep learning based methods that outperform traditional bioinformatics methods for predicting the relationship between genotypes and phenotypes [9–11]. Given the constraints on genomic data collection and the demand of large datasets for cutting-edge model development, it becomes crucial for genomic studies to develop data augmentation approaches [35] that can remedy this unfavorable scenario by increasing the amount and diversity of data. Generally speaking, strategies for genomic data augmentation at population levels can be traced back to genome sequence simulation tools extensively used in population genetics, such as SLiM [15] and msprime [20]. Nonetheless, these methods require users to have a thorough understanding of the prior information underlying the generation of population-based genomic data and the parameters used in population genetics models.

Recently, deep learning based data augmentation has been commonly conducted and shown impressive success in various areas including computer vision [3]. Typical data augmentation methods used in computer vision to generate realistic images rely on a particular deep learning framework termed Generative Adversarial Networks (GANs) [14]. A GAN is an adversarial game, in which two neural networks (a generator and a discriminator respectively)

BCB '20, September 21–24, 2020, Virtual Event, USA

ACM ISBN 978-1-4503-7964-9/20/09…$15.00
https://doi.org/10.1145/3388440.3412475

contest with and learn from each other. However, vanilla GANs deployed this way suffer from unstable training. To resolve this problem, many GAN variants are proposed that have significantly improved the quality of data outputs generated by GANs. These new GAN models that show supreme performance include CycleGANs [36], progressively growing GANs [19], deep convolutional GANs (DCGANs)[31], conditional GANs (cGANs) [26], Wasserstein GAN (WGAN) [4], and Boundary Equilibrium GAN (BEGAN) [7]). In general, these new GAN variants differ from vanilla GANs in that they utilize different network architectures, loss functions, evolutionary methods, or additional information. Latest studies have shown that in computer vision and image analysis, data augmented with these GAN variants, together with input original data can be combined together to improve image classifications. For example, Zhu et al. [37] used CycleGAN-based data augmentation for image synthesis for better emotion classification. Frid-Adar et al. [13] employed DCGAN [31] models to augment medical imaging data for improved classification of liver lesions. Alyafi et al. [2] showed that DCGAN [31] models can be used for synthesising photo-realistic breast mass patches with a considerable diversity. Han et al. [16] focused on generating synthetic multi-sequence brain Magnetic Resonance (MR) images using DCGAN [31] and WGAN [4]. Their preliminary validation showed that even an expert physician was unable to accurately distinguish the synthetic images from the real images in a Visual Turing Test. Bailo et al. [5] described how to apply image-to-image translation techniques to medical blood smear data to generate new data samples and meaningfully augment small datasets utilizing cGAN [26] models.

In addition to the success of GAN family models in augmenting imaging data, recent studies have also demonstrated that GAN and its variants can be applied to clinical trial and transcriptomics data. Specifically, Beaulieu-Jones et al. [6] showed that GANs can be trained with differential privacy [1], enabling them to share the synthetic data with others as though they had the original clinical trial (e.g. Systolic Blood Pressure Trial) data. To enhance sparse single-cell RNA-seq data, cscGAN [24] is developed based on Similarity Constraint on GANs (SCGAN) [22] for single-cell transcriptomics analysis. RAN-seq data in single-cell populations augmented with cscGAN generated cells, was reported to improve downstream analyses such as the detection of marker genes and cell type classification [24].

Compared to widely used image datasets, genomic data is significantly smaller, sparser, and is limited with small sample sizes. Therefore, it is an unknown yet challenging task to investigate if GAN and its variants can be applied to meaningfully enhance genomic data with small sets of observations and high dimensions. In addition, we need to develop new measurement metrics to evaluate the performance of data augmentation models for genomic data including genotypes. In computer vision, humans and domain experts can be asked to evaluate the quality of an augmented image by checking if they can distinguish a synthetic image from a real image. This type of evaluation is broadly used in computer vision and image processing. For example, previous studies [8] have reported to measure the quality of images from GAN models by conducting a visual Turing test. In such a test, two humans are asked to distinguish real and artificial images of objects, or two pathological experts are asked to tell a real medical image from a synthetic image for skin lesion classification or liver cancer detection. However, this kind of human expert evaluation is infeasible for genomic data, since humans themselves can not tell the difference between real and synthetic genomic data. Hence, we need to not only develop novel models for data augmentation of genomic data, but also design new metrics to effectively evaluate the performance of data augmented models in genomics.

In this study, we develop a novel Population-scale Genomic data augmentation approach based on Conditional Generative Adversarial Networks (PG-cGAN) utilizing several state-of-the-art strategies in improving GAN models. Specifically, we deploy the generator and discriminator, using stacked convolutional layers, to extract relationships or high correlations within neighboring genomic regions such as Linkage Disequilibrium (LD) patterns. To reflect population structures stratified in human genomic data, we encode population labels as the conditional information on both the generator and discriminator in a cGAN framework. Instead of using visual Turing tests and Fréchet Inception Distance (FID) to measure image qualities of augmented data in computer vision, we utilize frequently-used population and genetics characteristics to evaluate the realness of synthetic human genomes. Specifically, the evaluation metrics include genomic and genetic characteristics such as population stratification using principal component analysis (PCA), minor allele frequency (MAF) distribution, and LD patterns. We then applied the proposed PG-cGAN to a human genomic dataset and demonstrated that PG-cGAN can learn the distribution of real data, and generate high-quality synthetic genotypes.

2 METHODS

We developed a novel population-scale genomic augmentation method, named as PG-cGAN, based on conditional GAN with wasserstein loss function. We then applied the PG-cGAN to model the human leukocyte antigen (HLA) genotypes extracted from the 1000 Genome project [12]. The code and supplementary information of the methods can be accessed at https://github.com/shilab/PG-cGAN.git.

2.1 Data

We evaluated our proposed PG-cGAN model to augment HLA genotypes that covers a 3 Mbp region at chromosome 6p21.31, responsible for the regulation of the immune system in humans [18]. The HLA region is highly polymorphic and heterogeneous across individuals, which means that this region has many different alleles, allowing them to fine-tune the adaptive immune system. We extracted the genotypes of this HLA region from the 1000 Genome Consortium Project [12] with whole genome sequenced for 2,504 individuals from five super-populations worldwide, including European (EUR), East Asian (EAS), Afican (AFR), American (AMR), and South Asian (SAS). **Table S.1** of the supplementary materials summarizes the number of unrelated individuals in each super-population. In this study, we focus on common genetic variants, i.e. single nucleotide polymorphisms (SNPs) with a Minor Allele Frequency (MAF) of at least 5%. We then obtained the genotypes of 7,160 unique SNPs in these 2,504 individuals which will serve as the input to our PG-cGAN model. We converted the genotypes from the original Variant Call Format (VCF) files of the 1000 Genomes

Project as follows: '0' representing the original genotype of '0|0', '1' representing the genotype of '0|1', '2' representing the genotype of '1|0', and '3' representing the genotype of '1|1' respectively. Genotype vectors are then one-hot encoded with 4 channels, which is significantly different from image data in that images usually have 3 color channels and color values are continuous rather than categorical.

2.2 Model description

In this section, we present this novel method termed as Population-scale Genomic data augmentation based on Conditional Generative Adversarial Networks (PG-cGAN). This novel PG-cGAN model is based on conditional GAN (cGAN) framework, with both the generator and the discriminator stacked with convolutional layers with an aim to capture the underlying structures in genomic data, such as LD patterns that describe correlations of genotypes in neighboring genomic regions.

2.2.1 The conditional GAN (cGAN). The cGAN framework [26] was extended from a vanilla GAN by adding auxiliary information as a condition on both the generator and discriminator. In a cGAN model, a generator G and a discriminator D are engaged in a minmax adversarial training process with auxiliary information as condition, as illustrated in **Figure 1 (a)**. In this study, we use population labels as the condition in order to augment genomic data for a particular population. A cGAN can be formulated as:

$$\min_G \max_D E_{(x,y)\sim\mathbb{P}_r}[\log D(x,y)] + E_{z\sim\mathbb{P}_z, y\sim\mathbb{P}_y}[\log(1 - D(G(z,y),y))] \tag{1}$$

where the first term is discriminator loss on the training data and the second term is a generator loss. Both of these two terms are under the condition $y \sim \mathbb{P}_y$. The generator $G(z,y) \rightarrow x$ learns the conditional distribution $\mathbb{P}_g$ over real data distribution $\mathbb{P}_r$ by mapping a random noise vector $z \sim \mathbb{P}_z$ with a given condition y to a sample $x \sim \mathbb{P}_g$. The way to combine the noise z and with a label y is flexible. One widely-used strategy of doing so is through the element-wise multiplication between the noise vector and an embedding of y. $D(x,y) \rightarrow [0,1]$ formulates a discriminator in a cGAN, which accepts an the input x with a condition or label y and predicts the probability under the condition y that x is drawn from the real data distribution rather than from the generative model.

Originally, the optimization of a vanilla GAN or cGAN model is to minimize the Jensen-Shannon divergence [14] between the distributions of real and fake samples. The problem with this type of optimization is the difficulty to achieve a globally optimized solution. For instance, a well-known problem with this optimization goal is that the model will reach a state called mode collapse, where the generator learns to generate only a limited number of samples [4].

2.2.2 Wasserstein loss function. To improve the training stability of cGAN toward an effective model to augment genomic data, we introduce the wasserstein loss function to the cGAN optimization. The wasserstein loss function was first proposed in WGAN [4], attempting to address the instability problems of GANs by replacing the optimization objective with the wasserstein distance. Arjovsky et al. [4] described two main benefits of WGAN compared with vanilla GANs. The first benefit is that WGAN drastically reduces the occurrence of mode collapse phenomena. The second advantage of WGAN is that it can recall the direct relation between the quality of generated data and the loss value of the model, which is a unique and nice property of WGAN.

The Wasserstein distance between $\mathbb{P}_r$ and $\mathbb{P}_g$ is defined as follows:

$$W(\mathbb{P}_r,\mathbb{P}_g) = \inf_{\gamma\in\Pi(\mathbb{P}_r,\mathbb{P}_g)} E_{(x_r,x_g)\sim\gamma}[\|x_r - x_g\|] \tag{2}$$

where $\Pi(\mathbb{P}_r,\mathbb{P}_g)$ is the set of all joint distributions $\gamma(x_r,x_g)$ whose marginals are $\mathbb{P}_r$ and $\mathbb{P}_g$, respectively. In simple terms, $\gamma(x_r,x_g)$, also called a transport plan, represents the amount of "mass" that is needed to transfer from x_r to x_g in order to convert $\mathbb{P}_r$ to $\mathbb{P}_g$. Hence, the wasserstein distance is a cost of an optimal transport plan for this conversion. However, this infimum is hard to achieve. Thus, the authors proposed the wasserstein-1 distance by taking advantage of the Kantorovich-Rubinstein duality [4] in order to use an easier-to-achieve form of Eq. (2) as:

$$W(\mathbb{P}_r,\mathbb{P}_g) = \sup_{\|f\|_L\leq 1} E_{x_r\sim\mathbb{P}_r} f(x_r) - E_{x_g\sim\mathbb{P}_g} f(x_g) \tag{3}$$

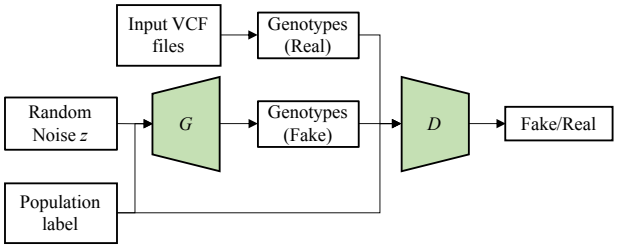
where the supremum is over a set of 1-Lipschitz functions. Basically, Eq. (3) states that all we need to do is to find a regression function f which maximizes the average distance between real and fake samples while maintaining the 1-Lipschitz constraint. In order to satisfy such a constraint, the weights at all layers in a discriminator are clipped to a value no larger than 0.01. Note that the output of a WGAN model is not the log-likelihood of a vanilla GAN. Instead, the output of WGAN is a linear function to measure the realness of generated samples. Thus, WGAN refers to the regression function f as *critic*. However, many studies do not exactly distinguish the terms of discriminator and critic. Hence, in this study, we just use the discriminator except for specific explanations.

2.2.3 The proposed PG-cGAN model. To utilize the advantages of cGAN and WGAN for genomic data augmentation, we propose a novel PG-cGAN based on a cGAN framework incorporated with the wasserstein loss function. Combining Eq. (1) and (3), the objective of optimizing a PG-cGAN model is defined as follows:

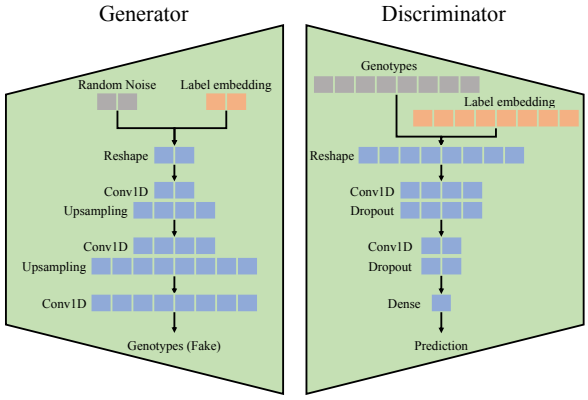
$$\min_G \max_D E_{(x,y)\sim\mathbb{P}_r} D(x,y) - E_{z\sim\mathbb{P}_z, y\sim\mathbb{P}_y} D(G(z,y),y) \tag{4}$$

where D is constricted with weight clipping less than 0.01 in order to satisfy the 1-Lipschitz constraint.

As illustrated in **Figure 1 (a)**, the architecture of PC-cGAN takes input from the original genotypes and population labels as conditions. Both the generator and discriminator are stacked with convolutional layers to capture the local patterns or correlations in neiboring genomic regions (e.g. LD patterns). As shown in **Figure 1 (b)**, our PG-cGAN model employs Convolutional Neural Networks (CNN) for both the generator and discriminator. A noise $z \sim \mathbb{P}_z$ is a 100 dimension vector drawn from a standard Gaussian distribution. Here, x is the one-hot encoded genotype of a sample. Each genotype x belongs to one of the five super-populations (EUR, EAS, AFR, AMR, SAS), where these super-population labels are used as condition information y. The layers and output shapes of the generator and discriminator of PG-cGAN are presented in **Table S.2** of the supplementary materials.



(a) Architecture of conditional GAN on genomics data



(b) Architectures of the generator and discriminator

Figure 1: Architecture of the proposed PG-cGAN model for genomic data augmentation. (a) Architecture of conditional GAN on genomic data, taking genotypes as input and population labels as condition on both generator and discriminator. (b) Specific architectures of the generator and discriminator in PG-cGAN.

3 RESULTS

PG-cGAN mimics the distribution of real genomic data by using stacked convolutional networks to capture local structures in the input data, such that the trained generator can generate genotypes for a specific population similar to the input data. We applied PG-cGAN to generate 2,500 synthetic HLA genotypes with random sizes of five super-populations, with almost the same size of real HLA genotypes collected from the 1000 Genomes Project. In order to effectively evaluate the performance of our models, we utilized several metrics to compare the generated genotypes with real genotypes. These metrics aim to reflect the underlying data distribution and characteristics, as well as those of importance to downstream genomic analysis, including population structures and variant frequency distributions.

3.1 The quality of synthetic HLA genotypes in terms of principal component analysis

Principal component analysis (PCA) [32] is a general dimension reduction method for reducing high-dimensional data, for example, population-level genomic data, to a smaller number of dimensions. PCA plays an important role for many population genetics tasks [28, 32]. For example, it is broadly used for population stratification and provided projected covariates to represent population structures in genome-wide association studies [29] possibly due to different ancestry backgrounds [30].

Therefore, the synthetic HLA genotypes generated from our model must have similar principal components (PCs) compared to the original data, so that these synthetic data can be used in population-level genomics studies. As shown in **Figure 2**, the synthetic HLA genotypes generated by PG-cGAN did have similar PCA distribution and clusters compared with the real HLA genotypes in each super-population. This observation assures us that the synthetic genotypes can depict the population structures underlying these genotypes as encoded in real genotypes.

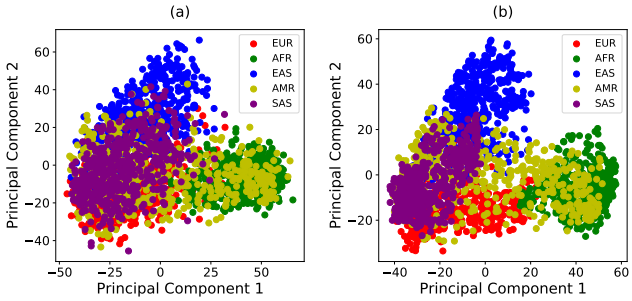


Figure 2: The PCA distribution in a 2-dimension space of real HLA genotypes (left) and synthetic HLA genotypes (right).

3.2 The quality of synthetic HLA genotypes in terms of linkage disequilibrium

Linkage disequilibrium (LD) [33] is the non-random association of alleles at different loci in a given population and reflects the correlation or relationship among nearby genetic variants. LD in human populations is influenced by many evolutionary factors, including selection, the rate of genetic recombination, mutation rate, genetic drift, the system of mating, population structure, and genetic linkage [27]. As a result, LD patterns in a population is a powerful and widely-used signal to decipher relatedness, ancestry, demographic, and evolution of human genomes.

Comparing with the evaluation of genotypes in terms of PCA which investigates the underlying population structure, LD depicts more detailed correlations or relationships between different genomic loci on various chromosomes. **Figure 3** shows the LD patterns measured by pair-wise correlation (i.e. R^2) values of different alleles for real and synthetic HLA genotypes, respectively. We observed that most of the LD blocks lie close to the diagonal line, which means the local alleles have strong LD linkages, while the long-distance alleles have no LD relationship. LD blocks in synthetic HLA genotypes match well with the LD blocks in synthetic HLA genotypes near the diagonal line, which demonstrates that our model can augment genomic data that mimics the original data in terms of reserving LD patterns. Besides, those relatively large LD blocks near the diagonal line, the upper right corner and lower left corner are all clear, which means that our PG-cGAN model not only learns useful local patterns, but also depresses noises in long distances. Our PG-cGAN model achieves such nice performance since it learns LD patterns through stacked convolutional layers in

the generator and discriminator to incorporate the relationships in nearby genomic regions.

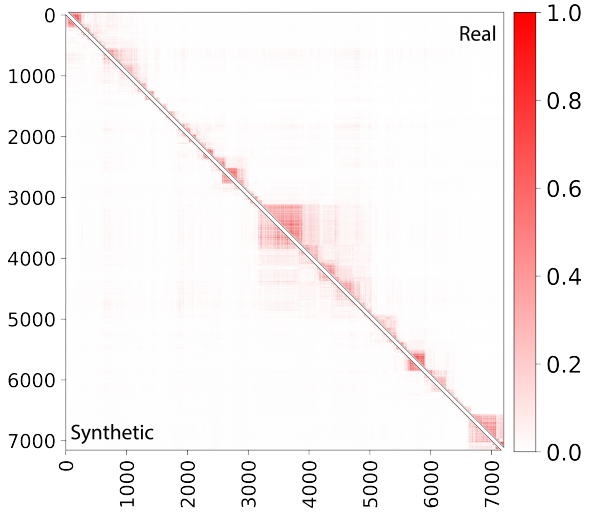


Figure 3: The pairwised R^2 of different loci for real HLA (upper right) and synthetic HLA (lower left). The deeper the red color is, the stronger the LD relationship is in the data.

3.3 The quality of synthetic HLA genotypes in terms of minor allele frequency

Minor Allele Frequency (MAF), i.e., the frequency of minor allele at one locus, is widely utilized in population genetics studies because it depicts the frequency spectrum of genetic variants in a population [17]. Allele frequency can be used to study selection and heritability of genetic regions in a population, and is a critical characteristic in genomic data. We thus compared the MAF distributions of real and synthetic HLA genotypes (**Figure 4**) to evaluate how well our PG-cGAN model captures the allele frequency distributions in genomic data. As shown in **Figure 4 (a)**, the MAF of real and synthetic genotypes align well along the diagonal line, indicating that real and synthetic genotypes have similar MAF distributions. We also created histogram plots (**Figure 4 (b)**) to show that although different bins of real and synthetic data contain slightly different numbers of variants, the MAF distributions are fairly close.

3.4 Conditional data generation for a specific population

It is critical to generate genomic data for a particular population or group based on the real data from that population or group. In this study, we encode this population or group information as a condition in the proposed PG-cGAN model. Conditional generation of genotypes can be particularly useful to increase the number of samples that are hard or infeasible to obtain for a specific population or disease group to address a key challenge of imbalanced and biased data in genomics. **Table S.3** of the supplementary materials shows the number of individuals in each super-population in synthetic HLA genotypes, using the corresponding super-population label as a condition.

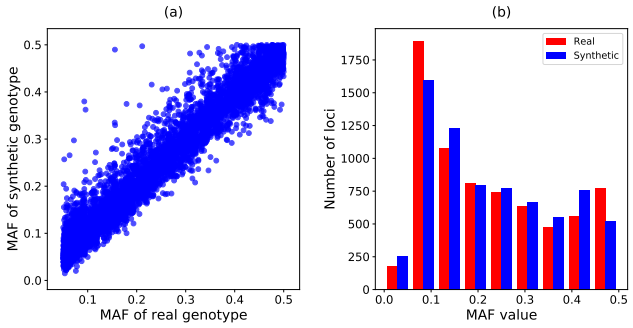


Figure 4: MAF comparison between real and synthetic genotypes. (a) Scatter plot where points lying closer to the diagonal line represent a better match. (b) Histograms of MAF distributions for real (in red) and synthetic (in blue) genotypes.

The quality of the conditional generation of genotypes for each super-population is assessed by using PCA and LD visualization for the five super-populations respectively, as illustrated in **Figure S.2** and **Figure S.3** of the supplementary materials. In the PCA visualization, blue dots represent real HLA genotypes and red dots are the synthetic HLA genotypes in the same super-population. The larger overlap between red and blue areas means a better match between real and synthetic HLA genotypes. Additionally, LD blocks of synthetic (lower left panel) and real (upper right panel) HLA genotypes mirror each other along the diagonal line, indicating that, for each super-population, our PG-cGAN model can accurately learn specific LD patterns in real genomic data. There are slightly different details of LD blocks in each super population, for example, EAS has a larger central LD block, in contrast to the central LD block of AFR which is the smallest and weakest. This difference reflects the different genetic architecture of these super-populations due to evolution and selection in human populations.

We quantitatively measured the difference of two LD matrices by calculating the Mean Squared Error (MSE) between the real and synthetic genotypes in each super-population as shown in **Table. 1**. We observed that PG-cGAN achieves fairly low MSE values on all super-populations. While the lowest MSE was observed in AFR, our PC-cGAN model achieves a similar performance on any of the other four super-populations.

Table 1: Mean squared error of LD values in real and synthetic HLA genotypes. A lower value represents a better match.

Super population	MSE
AMR	0.003059
AFR	0.002380
EAS	0.003525
EUR	0.003694
SAS	0.003620

4 CONCLUSION

In summary, we propose a novel data augmentation method, termed as PG-cGAN, to synthesize and augment genomic data for any specific population or group by transforming existing real genomic data without setting any prior or model parameters. Our results for generating HLA genotypes show that PG-cGAN can generate high quality genotype data of individuals that capture population and correlation structures in real genotypes. The success of PG-cGAN benefits from the stacked convolutional networks in the generator and discriminator, while incorporating population or group labels as conditions in a combined cGAN and WGAN framework.

Although we showed the success of our PG-cGAN model for generating genomic data that is typically sparse and high dimensional, it is still a challenging problem to generate high-dimensional data of high quality using GAN models. In the future, we plan to adopt other emerging techniques to improve data quality in augmenting genomic data, i.e. Boundary Equilibrium GANs (BEGAN) [7].

Note that the generated genomic data are automatically anonymous. Hence, these synthetic genomic data can be widely shared without worries about privacy or security [23]. These synthetic data reflect the population structure and underlying correlations of genetic variants in real genomic data, and can thus allow for augmenting and sharing data for building robust and complex AI and machine learning based models for a wide range of genomics studies.

ACKNOWLEDGMENTS

This work is partially supported by the National Science Foundation of the United States (Award Number: 1750632).

REFERENCES

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. 308–318.

[2] Basel Alyafi, Oliver Diaz, and Robert Martí. 2020. DCGANs for realistic breast mass augmentation in x-ray mammography. In *Medical Imaging 2020: Computer-Aided Diagnosis*, Vol. 11314. International Society for Optics and Photonics, 1131420.

[3] Antreas Antoniou, Amos Storkey, and Harrison Edwards. 2017. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340* (2017).

[4] Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein gan. *arXiv preprint arXiv:1701.07875* (2017).

[5] Oleksandr Bailo, DongShik Ham, and Young Min Shin. 2019. Red blood cell image generation for data augmentation using conditional generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 0–0.

[6] Brett K Beaulieu-Jones, Zhiwei Steven Wu, Chris Williams, Ran Lee, Sanjeev P Bhavnani, James Brian Byrd, and Casey S Greene. 2019. Privacy-preserving generative deep neural networks support clinical data sharing. *Circulation: Cardiovascular Quality and Outcomes* 12, 7 (2019), e005122.

[7] David Berthelot, Thomas Schumm, and Luke Metz. 2017. Began: Boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717* (2017).

[8] Poonam Chaudhari, Himanshu Agrawal, and Ketan Kotecha. 2019. Data augmentation using MG-GAN for improved cancer classification on gene expression data. *Soft Computing* (2019), 1–11.

[9] Junjie Chen and Xinghua Shi. 2019. Sparse Convolutional Denoising Autoencoders for Genotype Imputation. *Genes* 10, 9 (2019), 652.

[10] Junjie Chen and Xinghua Shi. 2019. A Sparse Convolutional Predictor with Denoising Autoencoders for Phenotype Prediction. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. 217–222.

[11] Yifei Chen, Yi Li, Rajiv Narayan, Aravind Subramanian, and Xiaohui Xie. 2016. Gene expression inference with deep learning. *Bioinformatics* 32, 12 (2016), 1832–1839.

[12] 1000 Genomes Project Consortium et al. 2015. A global reference for human genetic variation. *Nature* 526, 7571 (2015), 68–74.

[13] Maayan Frid-Adar, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. 2018. Gan-based data augmentation for improved liver lesion classification. (2018).

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*. 2672–2680.

[15] Benjamin C Haller and Philipp W Messer. 2019. SLiM 3: forward genetic simulations beyond the Wright–Fisher model. *Molecular biology and evolution* 36, 3 (2019), 632–637.

[16] Changhee Han, Hideaki Hayashi, Leonardo Rundo, Ryosuke Araki, Wataru Shimoda, Shinichi Muramatsu, Yujiro Furukawa, Giancarlo Mauri, and Hideki Nakayama. 2018. GAN-based synthetic brain MR image generation. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, 734–738.

[17] Ryan D Hernandez, Lawrence H Uricchio, Kevin Hartman, Chun Ye, Andrew Dahl, and Noah Zaitlen. 2019. Ultra-rare variants drive substantial cis-heritability of human gene expression. *bioRxiv* (2019), 219238.

[18] Jan Hillert. 1994. Human leukocyte antigen studies in multiple sclerosis. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society* 36, S1 (1994), S15–S17.

[19] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2017. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196* (2017).

[20] Jerome Kelleher, Alison M Etheridge, and Gilean McVean. 2016. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS computational biology* 12, 5 (2016).

[21] Klaus-Peter Koepfli, Benedict Paten, Genome 10K Community of Scientists, and Stephen J O'Brien. 2015. The Genome 10K Project: a way forward. *Annu. Rev. Anim. Biosci.* 3, 1 (2015), 57–111.

[22] Xiaoqiang Li, Liangbo Chen, Lu Wang, Pin Wu, and Weiqin Tong. 2018. SC-GAN: Disentangled Representation Learning by Adding Similarity Constraint on Generative Adversarial Nets. *IEEE Access* 7 (2018), 147928–147938.

[23] Jeantine E Lunshof, Ruth Chadwick, Daniel B Vorhaus, and George M Church. 2008. From genetic privacy to open consent. *Nature Reviews Genetics* 9, 5 (2008), 406–411.

[24] Mohamed Marouf, Pierre Machart, Vikas Bansal, Christoph Kilian, Daniel S Magruder, Christian F Krebs, and Stefan Bonn. 2020. Realistic in silico generation and augmentation of single-cell RNA-seq data using generative adversarial networks. *Nature Communications* 11, 1 (2020), 1–12.

[25] Seonwoo Min, Byunghan Lee, and Sungroh Yoon. 2017. Deep learning in bioinformatics. *Briefings in bioinformatics* 18, 5 (2017), 851–869.

[26] Mehdi Mirza and Simon Osindero. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784* (2014).

[27] Magnus Nordborg and Simon Tavaré. 2002. Linkage disequilibrium: what history has to tell us. *TRENDS in Genetics* 18, 2 (2002), 83–90.

[28] John Novembre and Matthew Stephens. 2008. Interpreting principal component analyses of spatial population genetic variation. *Nature genetics* 40, 5 (2008), 646–649.

[29] Nick Patterson, Alkes L Price, and David Reich. 2006. Population structure and eigenanalysis. *PLoS genetics* 2, 12 (2006).

[30] Alkes L Price, Noah A Zaitlen, David Reich, and Nick Patterson. 2010. New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics* 11, 7 (2010), 459–463.

[31] Alec Radford, Luke Metz, and Soumith Chintala. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434* (2015).

[32] David Reich, Alkes L Price, and Nick Patterson. 2008. Principal component analysis of genetic data. *Nature genetics* 40, 5 (2008), 491–492.

[33] David E Reich, Michele Cargill, Stacey Bolk, James Ireland, Pardis C Sabeti, Daniel J Richter, Thomas Lavery, Rose Kouyoumjian, Shelli F Farhadian, Ryk Ward, et al. 2001. Linkage disequilibrium in the human genome. *Nature* 411, 6834 (2001), 199–204.

[34] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. 2015. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine* 12, 3 (2015).

[35] David A Van Dyk and Xiao-Li Meng. 2001. The art of data augmentation. *Journal of Computational and Graphical Statistics* 10, 1 (2001), 1–50.

[36] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*. 2223–2232.

[37] Xinyue Zhu, Yifan Liu, Jiahong Li, Tao Wan, and Zengchang Qin. 2018. Emotion classification with data augmentation using generative adversarial networks. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 349–360.