Class-Conditional VAE-GAN for Local-Ancestry Simulation

Daniel Mas Montserrat[*]
Purdue University

Carlos Bustamante
Stanford University

Alexander Ioannidis
Stanford University

Abstract

Local ancestry inference (LAI) allows identification of the ancestry of all chromosomal segments in admixed individuals, and it is a critical step in the analysis of human genomes with applications from pharmacogenomics and precision medicine to genome-wide association studies. In recent years, many LAI techniques have been developed in both industry [1] and academic research [2]. However, these methods require large training data sets of human genomic sequences from the ancestries of interest. Such reference datasets are usually limited, proprietary, protected by privacy restrictions, or otherwise not accessible to the public. Techniques to generate training samples that resemble real haploid sequences from ancestries of interest can be useful tools in such scenarios, since a generalized model can often be shared, but the unique human sample sequences cannot. In this work we present a class-conditional VAE-GAN to generate synthetic human genomic sequences that can be used to train local ancestry inference (LAI) algorithms. We evaluate the quality of our generated data by comparing the performance of a state-of-the-art LAI method when trained with generated versus real data.

1 Introduction

Human populations all share a common ancient origin in Africa [3], and a common set of variable sites, but correlations between neighboring sites along the genome, which are typically inherited together, differ between sub-populations around the globe [4]. These correlations along the genome, known as linkage, influence polygenic risk scores (PRS) [5], genome-wide association studies (GWAS) [6], and many other aspects of precision medicine. Unfortunately, many of the world's sub-populations have not been included in modern genetic research studies with over 80% of these studies to date including only individuals of European ancestry [7]. This severely restricts the ability to make accurate predictions for the rest of the world's populations [5]. Deconvolving the ancestry of admixed individuals using local-ancestry inference can contribute to filling this gap and to understanding the genetic architecture and associations of non-European ancestries; thus allowing the benefits of medical genetics to accrue to a larger portion of the planet's population.

Many methods for local-ancestry inference exist and are open-source, HAPAA [8], HAPMIX [9] and SABER [10] infer local-ancestry using Hidden Markov Models (HMMs), LAMP [11] uses probability maximization with a sliding window, and RFMix [2] uses random forests within windows. However, these algorithms require accessible training data from each ancestry in order to recognize the respective chromosomal ancestry segments.

A major challenge is that many datasets containing human genomic references are protected by privacy restrictions [12], are proprietary [13, 14], or are otherwise not accessible to the public, particularly datasets for under-served or sensitive populations. Generative models that can be easily shared once trained can be useful in such scenarios. While the datasets with their de-anonymizable

[*] This work was conducted during an internship at Stanford University.

14th Machine Learning in Computational Biology (MLCB 2019), Vancouver, Canada.

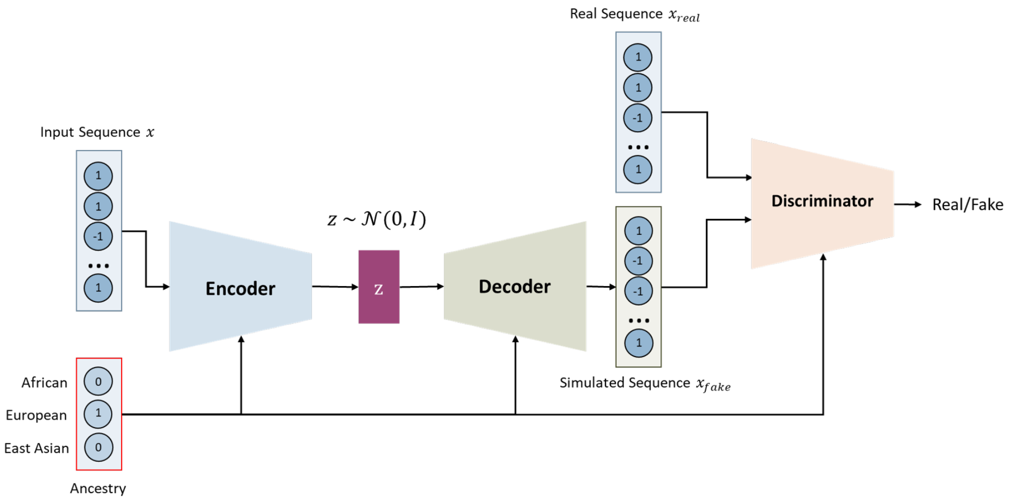

Figure 1: The class-conditional VAE-GAN is composed of 2 parts: (1) A class-conditional VAE consisting of an encoder-decoder pair. The encoder transforms the input sequence x from the ancestry c into an embedded representation z. The decoder transforms the embedding z and ancestry c into a reconstruction of the input sequence, $\tilde{x} = x_{fake}$. (2) A class-conditional GAN consisting of a decoder-discriminator pair. The decoder generates new samples x_{fake} from a Gaussian representation z_x and ancestry c. The discriminator separates between real sequences x_{real} and VAE-GAN generated sequences x_{fake}.

genome-wide sequences remain securely private, models trained on them could be made publicly available.

In recent years, deep learning has proven effective in solving computer vision and natural language processing problems [15]. These methods are being used in the biology, medical and genomics fields [16–19]. Specifically, deep learning-based generative methods have been increasingly popular in recent years. Generative networks such as Variational Autoencoders (VAEs) [20] contain a network that encodes the input data into a lower-dimensional space and a decoder that tries to reconstructs the original input. Generative Adversarial Networks (GANs) [21] have been able to generate samples that resemble the training data. GANs are able to generate realistic data by using two competing networks: a generator that aims to create realistic new samples and a discriminator that classifies between real and generated samples. Many variants and extensions of GANs and VAEs have been presented recently [22–25], including class-conditional VAE-GANs [25].

In this work, we present a class-conditional Variational Autoencoder and Generative Adversarial Network (CVAE-GAN), shown in figure 1, for human genome sequence simulation. The network combines a class-conditional VAE with a class-conditional GAN. The network is able to simulate new single-ancestry sequences that resemble the sequences from the training set. The generated sequences are used to train RFMix.

2 Dataset

We train our neural network using full human genome sequences from the 1000 genomes project [26]. We select a total of 258 single-population individuals from East Asia (EAS), African (AFR) and European (EUR) ancestry. Specifically, we use 83 Han Chinese in Beijing, China (CHB), 88 Yoruba in Ibadan, Nigeria (YRI) and 87 Iberian Population in Spain (IBS).

Additionally, 10 single-individuals per ancestry are used to generate admixed descendants for testing and validation using Wright-Fisher forward simulation over a series of generations. From 30 single-ancestry individuals, a total of 100 admixed individuals are generated with the admixture event occurring 12 generations in their past to create both validation and testing sets. The 258 single-ancestry individuals are used to train RFMix and the class-conditional VAE-GAN, and the 200 admixed individuals of the validation and testing sets are used to evaluate RFMix following training. Throughout we use chromosome 20 of each individual for experiments.

3 Network Architecture

The proposed network splits the genome into fixed-size non-overlapping windows. The individual genomic sites that vary between individuals (single nucleotide polymorphisms, or SNPs) within each window are used as the input for individual class-conditional VAE-GAN's. Each input SNP is encoded as -1 or 1 for the two variants found at that site. Missing input SNPs are modeled by inputting a 0 in the corresponding position. The CVAE-GAN's are composed of three sub-networks: an encoder, a decoder, and a discriminator. Each sub-network is class-conditional (i.e. the ancestry is an additional input of the network). The encoder-decoder pair forms a VAE while the decoder-discriminator pair forms a GAN (figure 1).

The encoder, $q(x; c)$, transforms the input SNPs x from the given the ancestry c (represented with one-hot encoding) into an isotropic Gaussian embedding space z. The network encodes the input sequence to the embedding space by estimating $\mu(x; c)$ and $\log \Sigma(x; c)$. The variance is estimated in a logarithmic form to force $\Sigma(x; c) > 0$. The embedded representation of a sample x from an ancestry c can be sampled from $z_x \sim \mathcal{N}(\mu(x; c),\, \Sigma(x; c))$. The sampling can be performed with the reparametrization trick: $z_x = \mu(x; c) + \Sigma(x; c) \odot \epsilon$, where $\epsilon \sim \mathcal{N}(0,\, I)$ and $\odot$ is an element-wise multiplication. The encoder networks begin with an input linear layer of size $(W + C) \times H$, where W is the window's size, C is the number of ancestries, and H is the size of the hidden layer. Following the first layer, a ReLU non-linearity and batch normalization is used. Then, two linear layers are used with dimensions $H \times J$, where J is the dimension of the embedding space, to estimate $\mu(x; c)$ and $\log \Sigma(x; c)$.

The decoder, with a given ancestry c and embedded representation z_x, tries to reconstruct the input SNPs $\tilde{x} = p(z_x; c)$. In order to obtain training samples for LAI methods, new sequences can be synthetic by selecting the desired ancestry c, sampling a random embedding, $z \sim \mathcal{N}(0,\, I)$, and reconstructing the SNP sequence $x_{new} = p(z; c)$. The decoder networks start with an input layer of size $(J + C) \times H$ followed by a ReLU non-linearity, batch normalization and an output linear layer of size $H \times W$. The discriminator network is trained to distinguish the real samples from the fake samples $\hat{y} = D(x; c)$. The discriminator networks start with an input layer of size $(W + C) \times H$ followed by a ReLU non-linearity, batch normalization and the output linear layer of size $H \times 1$.

The encoder is trained by minimizing the mean square error between the input and reconstructed sequences and the Kullback-Leibler divergence. The encoder loss function is as follows:

$$\mathcal{L}_q(x, c) = ||x - \tilde{x}||_2^2 + \frac{1}{2} \sum_{j}^{J} \mu_j^2 + \Sigma_j - \log \Sigma_j - 1 \tag{1}$$

where x and $\tilde{x}$ are the input and reconstructed sequence respectively, J is the dimension of the embedding space, μ_j is the jth element of $\mu_j(x; c)$ and Σ_j is the jth element of the diagonal of $\Sigma_j(x; c)$. The decoder is trained by minimizing the mean square error of the reconstruction and the adversarial loss:

$$\mathcal{L}_p(x, z, c) = ||x - \tilde{x}||_2^2 + \lambda_1 \log(1 - D(p(z; c))) \tag{2}$$

where $p(z; c)$ is a synthetic sequence from a randomly selected ancestry c and $z \sim \mathcal{N}(0,\, I)$. In our work we select $\lambda_1 = 0.1$. The discriminator is trained using binary cross-entropy with real, x, and synthetic data, $p(z; c)$:

$$\mathcal{L}_D(x, z, c) = -\log(D(x)) - \log(1 - D(p(z; c))) \tag{3}$$

Because the sequence is generated in a windowed approach, a different ancestry can be assigned to each window to simulate an admixed individual. However, in this work we focus on simulating single-ancestry individuals. The network is trained to obtain haploid sequences, but by generating pairs of haploid sequences, diploid chromosomes can be easily simulated. In order to avoid duplicate or very similar individuals, we generate N times the number of desired individuals and compute the pair-wise correlations of the generated sequences. Then, we select the $\frac{1}{N}$ individuals with the lowest average correlation. In this paper we use $N = 2$.

3.1 Location Coordinate Conditional

We present a different approach to make the CVAE-GAN ancestry-conditional. Instead of concatenating a one-hot encoding of the desired ancestry to the network's input, two values representing the longitude and latitude of the location of each ancestry are provided. Using geographic coordinates for ancestry inference has been previously explored in Battey et al [27].

4 Experimental Results

We use the single-ancestry individuals of the training set to train each CVAE-GAN. After training the networks, we generate a total of 100 synthetic samples per ancestry and train RFMix. RFMix is then evaluated with the admixed individuals in the validation set. We select the hyper-parameters of the CVAE-GAN (W, H and J) and the training parameters (learning rate, batch size and epoch) that provide the highest validation accuracy of RFMix. Specifically we select $W = 4000$, $H = 100$, and $J = 10$. In addition, $C = 3$ and $C = 2$ for one-hot and coordinate encoding respectively.

Finally, we compare the testing accuracy of RFMix when trained with real data as opposed to synthetic data generated with the CVAE-GANs. Additionally, we compare the results of including the discriminator and the adversarial loss (CVAE-GAN) with only using a CVAE. We do not compare it with only a generator-discriminator pair, as the encoder and reconstruction loss are important elements to enforce that the generated sequence belong to the desired ancestry.

Table 3 shows RFMix obtains comparable accuracy when trained with real and synthetic data. Accuracy results demonstrate that adding the discriminator and the adversarial loss help the network to learn to simulate human-chromosome sequences that are more similar to the original training data and therefore more useful for training LAI methods, providing a significant increase in accuracy.

Table 1: Accuracy of RFMix [2] trained with real and synthetic data

Method	RFMix Val. Accuracy	RFMix Test Accuracy
1000 Genomes Project Data	95.57%	95.33%
Generated Data (CVAE)	91.81%	91.55%
Generated Data (CVAE-GAN)	95.60%	95.05%
Generated Data (CVAE-GAN+Coord)	95.15%	95.22%

We also analyze how similar the synthetic sequences are to the original sequences used for training. To quantify this we perform an extensive sampling and compute the frequency of synthetic individuals that match some training set sample (real individual) with a 99.9%, 99.99%, 99.999% and 100% threshold. Table 2 shows the number of matches after sampling 10,000 individuals per ancestry.

Table 2: Synthetic individuals (out of 10,000) that have $P\%$ of SNPs matching a training sample.

P	99.9%	99.99%	99.999%	100%
Number of Individuals	2974	266	30	7

5 Conclusions

In this work we present a data generation method using CVAE-GANs. These networks show promising results using real human genomes from the 1000 genomes project. Strong simulation methods allow researchers to infer ancestry using a wide-range of existing ancestry tools without needing to have direct access to human reference data from sensitive populations, or from proprietary or protected databases. Beyond simulation, generative models have the potential to estimate meaningful representations in the embedding space that could be useful for data imputation or reconstruction.

References

[1] E. Y. Durand, C. B. Do, J. L. Mountain, and J. M. Macpherson, “Ancestry Composition: A Novel, Efficient Pipeline for Ancestry Deconvolution,” *bioRxiv*, p. 010512, Oct. 2014.

[2] B. K. Maples, S. Gravel, E. E. Kenny, and C. D. Bustamante, “RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference,” *The American Journal of Human Genetics*, vol. 93, pp. 278–288, August 2013.

[3] M. DeGiorgio, M. Jakobsson, and N. A. Rosenberg, “Out of Africa: modern human origins special feature: explaining worldwide patterns of human genetic variation using a coalescent-based serial founder model of migration outward from Africa.,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, pp. 16057–16062, September 2009.

[4] J. Z. Li, D. M. Absher, H. Tang, A. M. Southwick, A. M. Casto, S. Ramachandran, H. M. Cann, G. S. Barsh, M. Feldman, L. L. Cavalli-Sforza, and R. M. Myers, “Worldwide human relationships inferred from genome-wide patterns of variation,” *Science*, vol. 319, pp. 1100–1104, February 2008.

[5] L. Duncan, H. Shen, B. Gelaye, J. Meijsen, K. Ressler, M. Feldman, R. Peterson, and B. Domingue, “Analysis of polygenic risk score usage and performance in diverse human populations,” *Nature Communications*, vol. 10, pp. 1–9, July 2019.

[6] A. R. Martin, M. Lin, J. M. Granka, J. W. Myrick, X. Liu, A. Sockell, E. G. Atkinson, C. J. Werely, M. Möller, M. S. Sandhu, *et al.*, “An unexpectedly complex architecture for skin pigmentation in africans,” *Cell*, vol. 171, pp. 1340–1353, November 2017.

[7] A. B. Popejoy and S. M. Fullerton, “Genomics is failing on diversity,” *Nature News*, vol. 538, pp. 161–164, October 2016.

[8] A. Sundquist, E. Fratkin, C. B. Do, and S. Batzoglou, “Effect of genetic divergence in identifying ancestral origin using HAPAA,” *Genome research*, vol. 18, p. 676–682, April 2008.

[9] A. L. Price, A. Tandon, N. Patterson, K. C. Barnes, N. Rafaels, I. Ruczinski, T. H. Beaty, R. Mathias, D. Reich, and S. Myers, “Sensitive Detection of Chromosomal Segments of Distinct Ancestry in Admixed Populations,” *PLoS Genetics*, vol. 5, pp. 1–18, June 2009.

[10] H. Tang, M. Coram, P. Wang, X. Zhu, , and N. Risch, “Reconstructing genetic ancestry blocks in admixed individuals,” *The American Journal of Human Genetics*, vol. 79, pp. 1–12, May 2006.

[11] S. Sankararaman, S. Sridhar, G. Kimmel, and E. Halperin, “Estimating local ancestry in admixed populations,” *The American Journal of Human Genetics*, vol. 82, pp. 290–303, February 2008.

[12] G. L. Wojcik, M. Graff, K. K. Nishimura, R. Tao, J. Haessler, C. R. Gignoux, H. M. Highland, Y. M. Patel, E. P. Sorokin, C. L. Avery, G. M. Belbin, S. A. Bien, I. Cheng, S. Cullina, C. J. Hodonsky, Y. Hu, L. M. Huckins, J. Jeff, A. E. Justice, J. M. Kocarnik, U. Lim, B. M. Lin, Y. Lu, S. C. Nelson, S.-S. L. Park, H. Poisner, M. H. Preuss, M. A. Richard, C. Schurmann, V. W. Setiawan, A. Sockell, K. Vahi, M. Verbanck, A. Vishnu, R. W. Walker, K. L. Young, N. Zubair, V. Acuna-Alonso, J. L. Ambite, K. C. Barnes, E. Boerwinkle, E. P. Bottinger, C. D. Bustamante, C. Caberto, S. Canizales-Quinteros, M. P. Conomos, E. Deelman, R. Do, K. Doheny, L. Fernández-Rhodes, M. Fornage, B. Hailu, G. Heiss, B. M. Henn, L. A. Hindorff, R. D. Jackson, C. A. Laurie, C. C. Laurie, Y. Li, D.-Y. Lin, A. Moreno-Estrada, G. Nadkarni, P. J. Norman, L. C. Pooler, A. P. Reiner, J. Romm, C. Sabatti, K. Sandoval, X. Sheng, E. A. Stahl, D. O. Stram, T. A. Thornton, C. L. Wassel, L. R. Wilkens, C. A. Winkler, S. Yoneyama, S. Buyske, C. A. Haiman, C. Kooperberg, L. Le Marchand, R. J. F. Loos, T. C. Matise, K. E. North, U. Peters, E. E. Kenny, and C. S. Carlson, “Genetic analyses of diverse populations improves discovery for complex traits,” *Nature*, vol. 570, pp. 514–518, June 2019.

[13] E. Han, Y. Wang, P. Carbonetto, R. E. Curtis, J. M. Granka, J. Byrnes, K. Noto, A. R. Kermany, N. M. Myres, M. J. Barber, K. A. Rand, S. Song, T. Roman, E. Battat, E. Elyashiv, H. Guturu,

E. L. Hong, K. G. Chahine, and C. A. Ball, "Clustering of 770,000 genomes reveals post-colonial population structure of North America," *Nature Communications*, vol. 8, p. 14238, Feb. 2017.

[14] K. Bryc, E. Y. Durand, J. M. Macpherson, D. Reich, and J. L. Mountain, "The Genetic Ancestry of African Americans, Latinos, and European Americans across the United States," *American journal of human genetics*, vol. 96, pp. 37–53, January 2015.

[15] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[16] D. S. Gareau, J. Correa da Rosa, S. Yagerman, J. A. Carucci, N. Gulati, F. Hueto, J. L. DeFazio, M. Suárez-Fariñas, A. Marghoob, and J. G. Krueger, "Digital imaging biomarkers feed machine learning for melanoma screening," *Experimental dermatology*, vol. 26, pp. 615–618, July 2017.

[17] A. S. Lundervold and A. Lundervold, "An overview of deep learning in medical imaging focusing on mri," *Zeitschrift für Medizinische Physik*, vol. 29, pp. 102–127, May 2019.

[18] X. Li, L. Liu, J. Zhou, and C. Wang, "Heterogeneity analysis and diagnosis of complex diseases based on deep learning method," *Scientific reports*, vol. 8, pp. 1–8, April 2018.

[19] G. Eraslan, Ž. Avsec, J. Gagneur, and F. J. Theis, "Deep learning: new computational modelling techniques for genomics," *Nature Reviews Genetics*, pp. 389–403, April 2019.

[20] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *Proceedings of the International Conference on Learning Representations*, pp. 1–14, April 2014.

[21] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, pp. 2672–2680, 2014.

[22] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," *arXiv preprint arXiv:1701.07875*, 2017.

[23] A. Brock, J. Donahue, and K. Simonyan, "Large scale gan training for high fidelity natural image synthesis," *Proceedings of the International Conference on Learning Representations*, May 2019. New Orleans, LA.

[24] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," *Advances in neural information processing systems*, pp. 3483–3491, 2015.

[25] J. Bao, D. Chen, F. Wen, H. Li, and G. Hua, "Cvae-gan: fine-grained image generation through asymmetric training," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2745–2754, Oct 2017. Venice, Italy.

[26] . G. P. Consortium *et al.*, "A global reference for human genetic variation," *Nature*, vol. 526, no. 7571, p. 68, 2015.

[27] C. J. Battey, P. L. Ralph, and A. D. Kern, "Predicting geographic location from genetic variation with deep neural networks," *BioRxiv*, 2019.

[28] J. Kelleher, A. M. Etheridge, and G. McVean, "Efficient coalescent simulation and genealogical analysis for large sample sizes," *PLoS Computational Biology*, vol. 12, pp. 1–22, May 2016.

[29] S. Gravel, B. M. Henn, R. N. Gutenkunst, A. R. Indap, G. T. Marth, A. G. Clark, F. Yu, R. A. Gibbs, C. D. Bustamante, . G. Project, *et al.*, "Demographic history and rare allele sharing among human populations," *Proceedings of the National Academy of Sciences*, vol. 108, pp. 11983–11988, July 2011.

A Appendix: Results of Out-of-Africa simulations test

A.0.1 Out-of-Africa Data

We use simulated data from an out-of-Africa model created in msprime [28]. This simulation models the origin and spread of humans as a single ancestral population that grew instantaneously into the continent of Africa. The human population within Africa is then modeled as having a constant population size to the present day. However, a small group of individuals are modeled as migrating out of Africa and then splitting: founding the present day European population and the present day East Asian population. Both are modeled as growing exponentially after their separation. The parameters that determine the timing of these events, effective population sizes, and growth rates are presented in Gravel et al. [29].

Following the above out-of-Africa model, we generate three groups of 100 diploid single-ancestry individuals, one group each of African, European and East Asian ancestry. We divide these 300 simulated individuals into training, validation and testing sets with 240, 30 and 30 diploid individuals respectively. The validation and (separately) testing individuals are used to generate admixed descendants using Wright-Fisher forward simulation as follows. From 30 single-ancestry individuals, a total of 100 admixed individuals are created with the admixture event occurring 8 generations in their past, yielding both validation and testing admixed sets. The 240 single-ancestry individuals are used to train RFMix and the class-conditional VAE-GAN, and the 200 admixed individuals of the validation and testing sets are used to evaluate RFMix following training. Throughout we use chromosome 20 of each individual for experiments.

A.0.2 Experimental Results

We use the single-ancestry out-of-Africa individuals of the training set to train each VAE-GAN. After training the networks, we generate a total of 80 synthetic samples per ancestry and train RFMix. The accuracy of RFMix trained using these synthetic samples is then evaluated by using the admixed individuals from the validation set. We select the hyper-parameters of the VAE-GAN (window size, hidden layer size and embedding space) and the training parameters (learning rate, batch size and epoch) that provide the highest validation accuracy of RFMix. Finally, we compare the testing accuracy of RFMix when trained with out-of-Africa data versus when trained with synthetic data generated with the VAE-GANs. Additionally, we compare the results of including the discriminator and the adversarial loss (VAE-GAN) with only using a VAE. We do not compare it with only a generator-discriminator pair, as the encoder and reconstruction loss are important elements to enforce that the generated sequence belong to the desired ancestry.

Table 3 shows that RFMix obtains comparable accuracy when trained with out-of-Africa and synthetic data. Accuracy results show that adding the discriminator and the adversarial loss helps the network to learn to synthesize human-chromosome sequences that are more similar to the original training data and therefore more useful to train LAI methods, providing a significant increase in accuracy.

Table 3: Accuracy of RFMix [2] trained with out-of-Africa and generated data

Method	**RFMix Val. Accuracy**	**RFMix Test Accuracy**
Out-of-Africa Data	97.98%	97.75%
Generated Data (CVAE)	93.21%	93.05%
Generated Data (CVAE-GAN)	97.58%	97.72%