Creating Artificial Human Genomes Using Generative Models

3 4

5

Authors: Burak Yelmen^{1,2,*}, Aurélien Decelle³, Linda Ongaro^{1,2}, Davide Marnetto¹, Corentin Tallec³, Francesco Montinaro^{1,4}, Cyril Furtlehner³, Luca Pagani^{1,5}, Flora Jay^{3,*}

67 Affiliations:

- 8 1 Institute of Genomics, University of Tartu, Estonia
- 9 2 Institute of Molecular and Cell Biology, University of Tartu, Estonia
- 10 3 Laboratoire de Recherche en Informatique, CNRS UMR 8623, Université Paris-Sud,
- 11 Université Paris-Saclay, Orsay, France
- 12 4 Department of Zoology, University of Oxford, UK
- 13 5 APE Lab, Department of Biology, University of Padova, Italy
- 14 *to whom correspondence should be addressed: <u>burakyelmen@gmail.com</u>, <u>flora.jay@lri.fr</u>

15 Abstract

Generative models have shown breakthroughs in a wide spectrum of domains due to recent advancements in machine learning algorithms and increased computational power. Despite these impressive achievements, the ability of generative models to create realistic synthetic data is still under-exploited in genetics and absent from population genetics.

21

Yet a known limitation of this field is the reduced access to many genetic databases 22 due to concerns about violations of individual privacy, although they would provide a 23 rich resource for data mining and integration towards advancing genetic studies. Here 24 25 we demonstrate that we can train deep generative adversarial networks (GANs) and 26 restricted Boltzmann machines (RBMs) to learn the high dimensional distributions of 27 real genomic datasets and create artificial genomes (AGs). Additionally, we ensure 28 none to little privacy loss while generating high quality AGs. To illustrate the promising outcomes of our method, we show that augmenting reference panels with AGs 29 30 improves imputation quality for low frequency alleles. In summary, AGs have the potential to become valuable assets in genetic studies by providing high quality 31 32 anonymous substitutes for private databases.

33 Introduction

34 Availability of genetic data has increased tremendously due to advances in sequencing 35 technologies and reduced costs (Mardis 2017). The vast amount of human genetic data is used in a wide range of fields, from medicine to evolution. Despite the 36 advances, cost is still a limiting factor and more data is always welcomed, especially 37 in population genetics and genome-wide association studies (GWAS) which usually 38 require substantial amounts of samples. Partially related to the costs but also to the 39 research bias toward studying populations of European ancestry, many autochthonous 40 populations are under-represented in genetic databases, diminishing the extent of the 41 resolution in many studies (Cann 2002; Popejoy and Fullerton 2016; Mallick et al. 42 2016; Sirugo et al. 2019). Additionally, a huge portion of the data held by government 43 44 institutions and private companies is considered sensitive and not easily accessible due to privacy issues, exhibiting yet another barrier for scientific work. A class of 45 machine learning methods called generative models might provide a suitable solution 46 to these problems. 47

48

Generative models are used in unsupervised machine learning to discover intrinsic 49 50 properties of data and produce new data points based on those. In the last decade, generative models have been studied and applied in many domains of machine 51 52 learning (Libbrecht and Noble 2015; Zhang et al. 2017; Rolnick and Dyer 2019). There have also been a few applications in the genetics field (Davidsen et al. 2019; Liu et al. 53 2019). Among the various generative approaches, we focus on two of them in this 54 study, generative adversarial networks (GANs) and restricted Boltzmann machines 55 (RBMs). GANs are generative neural networks which are capable of learning complex 56 57 data distributions in a variety of domains (Goodfellow et al. 2014). A GAN consists of two neural networks, a generator and a discriminator, which compete in a zero-sum 58 game (Supplementary Figure 1). During training, the generator produces new 59 instances while the discriminator evaluates their authenticity. The training objective 60 consists in learning the data distribution in a way such that the new instances created 61 by the generator cannot be distinguished from true data by the discriminator. Since 62 63 their first introduction, there have been many successful applications of GANs, ranging from generating high quality, realistic imagery to gap filling in texts (Ledig et al. 2017; 64

Fedus et al. 2018). GANs are currently the state-of-the-art models for generatingrealistic images (Brock et al. 2018).

67

A restricted Boltzmann machine, initially called Harmonium is another generative 68 model which is a type of neural network capable of learning probability distributions 69 through input data (Smolensky 1986; Teh and Hinton 2001). RBMs are two layer neural 70 71 networks consisting of an input (visible) layer and a hidden layer (Supplementary 72 Figure 2). The learning procedure for the RBM consists in maximizing the likelihood 73 function over the visible variables of the model. This procedure is done by adjusting 74 the weights such that the correlations between the visible and hidden variables on both 75 the dataset and sampled configurations from the RBM converge. Then RBM models recreate data in an unsupervised manner through many forward and backward passes 76 77 between these two layers (Gibbs sampling), corresponding to sampling from the learned distribution. The output of the hidden layer goes through an activation function, 78 79 which in return becomes the input for the hidden layer. Although mostly overshadowed by recently introduced approaches such as GANs or Variational Autoencoders 80 (Kingma and Welling 2013), RBMs have been used effectively for different tasks (such 81 as collaborative filtering for recommender systems, image or document classification) 82 and are the main components of deep belief networks (Hinton and Salakhutdinov 2006; 83 Hinton 2007; Larochelle and Bengio 2008). 84

85

Here we propose and compare a prototype GAN model along with an RBM model to 86 87 create Artificial Genomes (AGs) which can mimic real genomes and capture population structure along with other characteristics of real genomes. We envision two main 88 applications of our generative methods: (i) improving the performance of genomic 89 tasks such as imputation, ancestry reconstruction, GWAS studies, by augmenting 90 91 genomic panels with AGs serving as proxies for private datasets, (ii) demonstrating 92 that a proper encoding of the genomic data can be learned and possibly used as a starting input of various inference tasks by combining this encoding with recent neural 93 94 network-based tools for the reconstruction of recombination, demography or selection 95 (Sheehan and Song 2016; Adrion et al. 2019; Flagel et al. 2019).

96 **Results**

97 Reconstructing genome wide population structure:

Initially we created AGs with GAN, RBM, and two simple generative models for 98 comparison: a Bernoulli and a Markov chain model (see Materials & Methods) using 99 100 2504 individuals (5008 haplotypes) from 1000 Genomes data (1000 Genomes Project Consortium et al. 2015), spanning 805 SNPs from all chromosomes which reflect a 101 high proportion of the population structure present in the whole dataset (Colonna et al. 102 2014). Both GAN and RBM models seem to capture a good portion of the population 103 structure present in 1000 Genomes data while the other two models could only 104 produce instances centered around 0 on principal component analysis (PCA) space 105 106 (Figure 1). All major modes, corresponding to African, European and Asian genomes, 107 seem to be well represented in AGs produced by GAN and RBM models. Uniform 108 manifold approximation and projection (UMAP) mapping results also correlate with the 109 performed PCA (Supplementary Figure 3). We additionally checked the distribution of 110 pairwise differences of haploid genomes to see how different AGs are from real 111 genomes (Supplementary Figure 4). Both RBM and GAN models seem to have highly similar distributions to the distribution of pairwise differences of the real genomes within 112 113 themselves. Especially RBM excels at acquiring the real peaks, indicating a high similarity with real genomes. Since GANs and RBMs show an excellent performance 114 115 for this use case, we further explored other characteristics using only these two 116 models.

Figure 1. The six first axes of a PCA applied to real (gray) and artificial genomes (AGs) generated via Bernoulli (green), Markov chain (purple), GAN (blue) and RBM (red) models. There are 5000 haplotypes for each AG dataset and 5008 (2504 genomes) for the real dataset from 1000 Genomes spanning 805 informative SNPs. See Materials & Methods for detailed explanation of the generation procedures.



122

Furthermore, similarly to tSNE and UMAP, RBMs perform a non-linear dimension reduction of the data and provides a suitable representation of a genomic dataset as a by-product based on the non-linear feature space associated to the hidden layer (Supplementary Text). As Diaz-Papkovich et al (Diaz-Papkovich et al. 2019), we found that the RBM representation differs from the linear PCA ones. Here we plot the representation corresponding to the selected RBM model and exhibit its rapid evolution through training (Supplementary Figure 5).

131

Supplementary Figure 5 shows that African, East Asian, and to a lesser extent, European populations stand out on the two first components. The Finnish are slightly isolated from the other European (similar to Peruvian from American) populations on the first components. South Asians are located at the center separated from Europeans, partially overlapping with American populations, and stand out at dimension 5 and higher. Interestingly when screening the hidden node activations, we observed that different populations or groups activate different hidden nodes, each one representing a specific combination of SNPs, thereby confirming that the hidden layer

- 140 provides a meaningful encoding of the data (Supplementary Figure 6).
- 141

142 **Reconstructing local high-density haplotype structure:**

143 To evaluate if high quality artificial dense genome sequences can also be created by the generative models, we applied the GAN and RBM methods to a 10K SNP region 144 145 using (i) the same individuals from 1000 Genomes data and (ii) 1000 Estonian 146 individuals from the high coverage Estonian Biobank (Leitsalu et al. 2015) to generate 147 artificial genomes. PCA results of AGs spanning consecutive 10K SNPs show that both GAN and RBM models can still capture the relatively toned-down population structure 148 149 (Supplementary Figure 7) as well as the overall distribution of pairwise distances 150 (Supplementary Figure 8). Looking at the allele frequency comparison between real 151 and artificial genomes, we see that especially GAN performs poorly for low frequency 152 alleles, due to a lack of representation of these alleles in the AGs (Supplementary 153 Figure 9). On the other hand, the distribution of the distance of real genomes to the 154 closest AG neighbour show that GAN model, although slightly underfitting, outperforms 155 RBM model, for which an excess of small distances points towards slight overfitting 156 (Supplementary Figure 10).

157

Additionally, we performed linkage disequilibrium (LD) analyses comparing artificial 158 and real genomes to assess how successfully the AGs imitate short and long range 159 160 correlations between SNPs. Pairwise LD matrices for real and artificial genomes all 161 show a similar block pattern demonstrating that GAN and RBM accurately captured the overall structure with SNPs having higher linkage in specific regions (Figure 2a). 162 163 However plotting LD as a function of the SNP distance showed that all models capture weaker correlation, with RBM outperforming the GAN model perhaps due to its 164 165 overfitting characteristic (Figure 2b). To further determine the haplotypic integrity of 166 AGs, we performed ChromoPainter (Lawson et al. 2012) and Haplostrips (Marnetto and Huerta-Sánchez 2017) analyses on AGs created using Estonians as the training 167 168 data. It was visually impossible to distinguish the difference between real and artificial 169 genomes in terms of local haplotypic structure with Haplostrips (Supplementary Figure 170 11). However, majority of the AGs produced via GAN model displayed fractured chunks when painted against 1000 Genomes individuals whereas RBM AGs were 171 172 nearly indistinguishable from real genomes (Supplementary Figure 12).

Figure 2. Linkage disequilibrium (LD) analyses on real and artificial Estonian genomes. a) Correlation (r2) matrices of SNPs. Lower triangular parts are SNP pairwise correlation in real genomes and upper triangular parts are SNP pairwise correlation in artificial genomes. b) LD as a function of SNP distance. Pairwise SNP distances were stratified into 50 bins and for each distance bin the correlation was averaged over all pairs of SNPs belonging to the bin.

179





b.



After demonstrating that our models generated realistic AGs according to the 182 described summary statistics, we investigated further whether they respected privacy 183 184 by measuring the extent of overfitting. We calculated two metrics of resemblance and privacy, the nearest neighbour adversarial accuracy (AA_{TS}) and privacy loss presented 185 in a recent study (Yale et al. 2019). AA_{TS} score measures whether two datasets were 186 generated by the same distribution based on the distances between all data points and 187 188 their nearest neighbours in each set. When applied to artificial and real datasets, a 189 score between 0.5 and 1 indicates underfitting, between 0 and 0.5 overfitting (and likely 190 privacy loss), and exactly 0.5 indicates that the datasets are indistinguishable. By using 191 an additional real test set, it is also possible to calculate a privacy loss score that is 192 positive in case of information leakage, negative otherwise, and approximately ranges 193 from -0.5 to 0.5. Computed on our generated data, both scores support haplotypic 194 pairwise difference results confirming the underfitting nature of GAN AGs and slight 195 overfitting nature of RBM AGs with a risk of privacy leakage for the latter (Figure 3a 196 and 3b).

197

198 Since it has been shown in previous studies that imputation scores can be improved 199 using additional population specific reference panels (Gurdasani et al. 2015; Mitt et al. 2017), as a possible future use case, we tried imputing real Estonian genomes using 200 1000 Genomes reference panel and additional artificial reference panels with Impute2 201 software (Howie et al. 2011). Both combined RBM AG and combined GAN AG panels 202 203 outperformed 1000 Genomes panel for the lowest MAF bin (for MAF < 0.05, 0.015 and 204 0.024 improvement respectively) which had 5926 SNPs out of 9230 total (Figure 3b). 205 Also mean info metric over all SNPs was 0.009 and 0.015 higher for combined RBM 206 and GAN panels respectively, compared to the panel with only 1000 Genomes 207 samples. However, aside from the lowest MAF bin, 1000 Genomes panel outperformed both concatenated panels for all the higher bins. This might be a 208 209 manifestation of haplotypic deformities in AGs, that might have disrupted the imputation algorithm. 210

211

Figure 3. a) Nearest neighbour adversarial accuracy (AA_{TS}) of artificial genomes 213 generated from Estonian Biobank. Black line indicates the optimum value whereas 214 values below the line indicate overfitting and values above the line indicate underfitting. 215 b) Privacy loss. Positive values indicate information leakage, hence overfitting. c) 216 Imputation evaluation of three different reference panels based on Impute2 software's 217 info metric. Imputation was performed on 8678 Estonian individuals (which were not 218 used in training of GAN and RBM models) using only 1000 Genomes panel (gray), 219 combined 1000 Genomes and GAN artificial genomes panel (blue) and combined 1000 220 221 Genomes and RBM artificial genomes panel (red). SNPs were divided into 10 strata, from 0.05 to 0.1, after which mean info metric values were calculated for each stratum. 222 223 Bars in the zoomed section show the standard error of mean.





226 Linking genotypes with phenotypes:

227 We then explored the possibility of creating AGs with unphased genotype data and recreating phenotype-genotype associations using generative models. As a proof of 228 concept, we created GAN AGs via training on 1925 Estonian individuals with 5000 229 SNPs using unphased genotypes instead of haplotypes. There was an additional 230 231 column in this dataset representing eye color (blue or brown). This region encompasses rs12913832 SNP which is highly associated with eye color (Han et al. 232 233 2008; Eriksson et al. 2010; Zhang et al. 2013). In our real genomes dataset, nearly 234 96% of the individuals possessing at least one ancestral allele A have brown eye color. 235 The AGs were able to reproduce the same genotypic-phenotypic association with a 236 ratio of 80% with the same allele A. Similarly, 97% of all blue-eyed individuals have the 237 both derived alleles (G) whereas the same percentage is 88% in GAN AGs. There are 238 no blue-eyed individuals who have both ancestral alleles in the real dataset but this number is 9 out 1925 in GAN AGs (Supplementary Table 1). We weren't able to create 239 240 RBM AGs for this dataset.

241 **Discussion**

In this study, we applied generative models to produce artificial genomes and 242 243 evaluated their characteristics. To the best of our knowledge, this is the first application of GAN and RBM models in this context, displaying overall promising applicability. We 244 245 showed that population structure and frequency-based features of real populations can successfully be preserved in AGs created using GAN and RBM models. Furthermore, 246 247 both models can be applied to sparse or dense SNP data given a large enough number of training individuals. Our different trials showed that the minimum required number 248 of individuals for training is highly variable, possibly correlated with the diversity among 249 250 individuals (data not shown). Since haplotype data is more informative, we created 251 haplotypes for the analyses but we also demonstrated that the models can be applied 252 to genotype data too, by simply combining two haplotypes if the training data is not 253 phased (see Materials & Methods). In addition, we showed that it is possible to 254 generate AGs with simple phenotypic traits through genotype data (see Results). Even 255 though there were only two simple classes, blue and brown eye color phenotypes, 256 generative models can be improved in the future to hold the capability to produce 257 artificial datasets combining AGs with multiple phenotypes.

258

259 One major drawback of the proposed models is that, due to computational limitations, 260 they cannot yet be used to create whole artificial genomes but rather snippets or 261 sequential dense chunks. Although parallel computing might be a solution, this might further disrupt the haplotype structure in AGs. Instead, adapting convolutional GANs 262 for AG generation might be another possible solution in the future (Radford et al. 2016). 263 Another problem arose due to rare alleles, especially for the GAN model. We showed 264 265 that nearly half of the alleles become fixed in the GAN AGs in the 10K SNP dataset. 266 whereas RBM AGs seem to capture more of the rare alleles present in real genomes 267 (Supplementary Figure 13). A known issue in GAN training is mode collapse (Salimans 268 et al. 2016). Mode collapse happens when the generator fails to cover the full support 269 of the data distribution. This failure case could explain the inability of GANs to generate rare alleles. For some applications relying on rare alleles, GAN models less sensitive 270 271 to mode dropping would be a promising alternative (Arjovsky et al. 2017; Lucas et al. 2018). 272

274 An important use case for the AGs in the future might be creating publicly available versions of private genome banks. Through enhancements in scientific knowledge and 275 276 technology, genetic data becomes more and more sensitive in terms of privacy. AGs might offer a versatile solution to this delicate issue in the future by protecting the 277 278 anonymity of real individuals. Our results showed that GAN AGs seem to be underfitting while RBM AGs seem to be overfitting based on distribution of minimum 279 280 distance to the closest neighbour (Supplementary Figure 10) and AA_{TS} scores (Figure 281 3a), although this can be investigated further by integrating AA_{TS} scores within our 282 models as a criterion for early stopping in training (before the networks start overfitting). In the context of the privacy issue, GAN AGs have a slight advantage since underfitting 283 284 is preferable. More distant AGs would hypothetically be harder to be traced back to the original genomes. We also tested the sensitivity of the AA_{TS} score and privacy loss 285 286 (Supplementary Figure 14). It appears that both scores are affected very slightly when 287 we add only a few real genomes to the AG dataset from the training set. Therefore, 288 more sensitive measurement techniques should be developed in the future for better assessment of generated AGs. Additionally, even though we did not detect exact 289 290 copies of real genomes in AG sets created either by RBM or GAN models, it is a very 291 complicated task to determine if the generated samples can be traced back to the 292 originals. Reliable measurements need to be developed in the future to assure 293 complete anonymity of AGs to their source.

294

295 Imputation results demonstrated promising outcomes especially for population specific 296 low frequency alleles. However, imputation with both RBM and GAN AGs integrated 297 reference panels showed slight decrease of info metric for higher frequency alleles 298 compared to only 1000 Genomes panel (Figure 3c). We initially speculated that this 299 might be related to the disturbance in haplotypic structure and therefore, tried to filter 300 AGs based on chunk counts from ChromoPainter results, preserving only AGs which 301 are below the average chunk count of real genomes. The reason behind this was to preserve most real-alike AGs with undisturbed chunks. Even with this filtering, slight 302 303 decrease in higher MAF bins was still present (data not shown). Yet results of 304 implementation with AGs for low frequency alleles and without AGs for high frequency 305 ones could be combined to achieve best performance. In terms of imputation, future improved models can become practically very useful, largely for GWAS studies in 306 307 which imputation is a common application to increase resolution. Different generative 308 models such as MaskGAN (Fedus et al. 2018) which demonstrated good results in text 309 gap filling might also be adapted for genetic imputation. RBM is possibly another option 310 to be used as an imputation tool directly by itself, since once the weights have been 311 learned, it is possible to fix a subset of the visible variables and to compute the average 312 values of the unobserved ones by sampling the probability distribution (in fact, it is 313 even easier than sampling entirely new configurations since the fixed subset of 314 variables will accelerate the convergence of the sampling algorithm).

315

316 As an additional feature, training an RBM to model the data distribution gives access 317 to a latent encoding of data points, providing a potentially easier to use representation 318 of data (Supplementary Figure 5). Future works could augment our current GAN model to also provide an encoding mechanism, in the spirit of (Dumoulin et al. 2016 Jun 2), 319 320 (Chen et al. 2016) or (Donahue et al. 2016). These interpretable representations of the 321 data are expected to be more relevant for downstream tasks (Chen et al. 2016) and 322 could be used as a starting point for various population genetics analyses such as 323 demographic and selection inference, or yet unknown tasks. 324

325 Although there are some current limitations, generative models will most likely become

prominent for genetics in the near future with many promising applications. In this work,

we demonstrated the first possible implementations and use of AGs in the forthcoming

328 field which we would like to name artificial genomics.

329 Materials & Methods

330 Data:

We used 2504 individual genomes from 1000 Genomes Project (1000 Genomes 331 Project Consortium 2015) and 1000 individuals from Estonian Biobank (Leitsalu et al. 332 333 2015) to create artificial genomes (AGs). Additional 2000 Estonians were used as a test dataset. Another Estonian dataset consisting of 8678 individuals which were not 334 335 used in training were used for imputation. Analyses were applied to a highly differentiated 805 SNP range selected as a subset from (Colonna et al. 2014) and a 336 dense 10000 SNP range from chromosome 15. We also used a narrowed down 337 version of the same region from chromosome 15 with 5000 SNPs with an additional 338 339 eye color column for unphased genotype data using another 1925 Estonians as 340 training dataset. In this set, 958 of the Estonian samples have brown (encoded as 1) 341 and 967 have blue eyes (encoded as 0). In the data format we used, rows are 342 individuals/haplotypes (instances) and columns are positions/SNPs (features). Each allele at each position is represented either by 0 or 1. In the case of phased data 343 344 (haplotypes), each column is one position whereas in the case of unphased data, each 345 two column corresponds to a single position with alleles from two chromosomes.

346

347 GAN Model:

348 We used python-3.6, Keras 2.2.4 deep learning library with TensorFlow backend (Chollet 2015), pandas 0.23.4 (McKinney 2010) and numpy 1.16.4 (Oliphant 2007) for 349 350 the GAN code. Generator of the GAN model we present consists of an input layer with the size of the latent vector size 600, one hidden layer with size proportional to the 351 number of SNPs as SNP number/1.2 rounded, another hidden layer with size 352 353 proportional to the number of SNPs as SNP number/1.1 rounded and an output layer 354 with the size of the number of SNPs. The latent vector was set with 355 numpy.random.normal function setting the mean of the distribution as 0 and the 356 standard deviation as 1. The discriminator consists of an input layer with the size of 357 the number of SNPs, one hidden layer with size proportional to the number of SNPs as SNP number/2 rounded, another hidden layer with size proportional to the number 358 359 of SNPs as SNP number /3 rounded and an output layer of size 1. All layer outputs except for output layers have LeakyReLU activation functions with leaky_alpha 360 361 parameter 0.01 and L2 regularization parameter 0.0001. The generator output layer

362 activation function is tanh and discriminator output layer activation function is sigmoid. Both discriminator and combined GAN were compiled with Adam optimization 363 364 algorithm with binary cross entropy loss function. We set the discriminator learning rate as 0.0008 and combined GAN learning rate as 0.0001. For 5000 SNP data, the 365 discriminator learning rate was 0.00008 and combined GAN learning rate was 0.00001. 366 Training to test dataset ratio was 3:1. We used batch size of 32 and trained all datasets 367 368 up to 20000 epochs. We stopped training based on coherent PCA results of AGs with 369 real genomes. During each batch, when only the discriminator is trained, we applied 370 smoothing to the real labels (1) by vectoral addition of random uniform distribution via numpy.random.uniform with lower bound 0 and upper bound 0.1. Elements of the 371 372 generated outputs were rounded to 0 or 1 with numpy.rint function.

373

374 **RBM Model:**

375 The RBM was coded in Julia (Bezanson et al. 2017), and all the algorithm for the 376 training has been done by the authors. The part of the algorithm involving linear algebra 377 used the standard package provided by Julia. Two versions of the RBM were 378 considered. In both versions, the visible nodes were encoded using Bernoulli random 379 variables {0,1}, and the size of the visible layer was the same size as the considered 380 input. Two different types of hidden layers were considered. First with a sigmoid activation function (hence having discrete {0,1} hidden variables), second with ReLu 381 (Rectified Linear unit) activations in which case the hidden variables were positive and 382 383 continuous (there are distributed according to a truncated gaussian distribution when 384 conditioning on the values of the visible variables). Results with sigmoid activation function were worse compared to ReLu so we used ReLu for all the analyses 385 386 (Supplementary Figure 15). The number of hidden nodes considered for the experiment was Nh=100 for the 805 SNP dataset and Nh=500 for the 10k one. There 387 388 is no canonical way of fixing the number of hidden nodes, in practice we checked that 389 the number of eigenvalues learnt by the model was smaller than the number of hidden 390 nodes, and that by adding more hidden nodes no improvement were observed during 391 the learning. The learning in general is guite stable, in order to have a smooth learning 392 curve, the learning rate was set between 0.001 and 0.0001 and we used batch size of 393 32. The negative term of the gradient of the likelihood function was approximated using the PCDk method (Brügge et al. 2013), with k=10 and 100 of persistent chains. As a 394

stopping criterion, we looked at when the AA_{TS} score converges to the ideal value of 0.5 when sampling the learned distribution.

397

398 Bernoulli Distribution Model:

We used python-3.6, pandas 0.23.4 and numpy 1.16.4 for the Bernoulli distribution model code. Each allele at a given position was randomly drawn given the derived allele frequency in the real population.

402

403 Markov Chain Model:

We used python-3.6, pandas 0.23.4 and numpy 1.16.4 for the Markov chain model code. Allele at the initial position was set by drawing from a Bernoulli distribution parameterized with the real frequency. Each sequence window consisting of a given number of positions was determined by the probability of the previous sequence window. After the initial position, window size increased incrementally up to the given window size.

410

411 Chromosome Painting:

We compared the haplotype sharing distribution between real and artificial chromosomes through ChromoPainter (Lawson et al. 2012). In detail, we have painted 100 randomly selected "real" and "artificial" Estonians (recipients) against all the 1000 Genome Project phased data (donors). The nuisance parameters -n (348.57) and -M (0.00027), were estimated running 10 iterations of the expectation-maximization algorithm on a subset of 3,800 donor haplotypes.

418

419 Haplostrips:

We used Haplostrips (Marnetto and Huerta-Sánchez 2017) to visualize the haplotype structure of real and artificial genomes. We extracted 500 individuals from each sample set (Real, GAN synthetics, RBM synthetics) and considered them as different populations. Black dots represent derived alleles, white ancestral. The plotted SNPs were filtered for a population specific minor allele frequency >5%; haplotypes were clustered and sorted for distance against the consensus haplotype from the real set. See the application article for further details about the method.

427 Acknowledgements

- 428 This work was supported by the European Union through the European Regional
- 429 Development Fund (Project No. 2014-2020.4.01.16-0024, MOBTT53: LP, DM, BY;
- 430 Project No. 2014-2020.4.01.16-0030: LO, FM); the Estonian Research Council grant
- 431 PUT (PRG243): LP; Laboratoire de Recherche en Informatique: FJ, AD, CT, CF.

432 **References**

- 434 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP,
- 435 Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, et al. 2015. A global
- 436 reference for human genetic variation. Nature. 526(7571):68–74.
- 437 doi:10.1038/nature15393.
- Adrion JR, Galloway JG, Kern AD. 2019. Inferring the landscape of recombination
 using recurrent neural networks. bioRxiv. doi:10.1101/662247.
- 440 Arjovsky M, Chintala S, Bottou L. 2017. Wasserstein generative adversarial
- networks. In: 34th International Conference on Machine Learning, ICML 2017.
- 442 Bezanson J, Edelman A, Karpinski S, Shah VB. 2017. Julia: A fresh approach to 443 numerical computing. SIAM Rev. doi:10.1137/141000671.
- Brock A, Donahue J, Simonyan K. 2018 Sep 28. Large Scale GAN Training for High
 Fidelity Natural Image Synthesis. http://arxiv.org/abs/1809.11096.
- 446 Cann HM. 2002. A Human Genome Diversity Cell Line Panel. Science (80-).
- 447 doi:10.1126/science.296.5566.261b.
- 448 Chen X, Duan Y, Houthooft R, Schulman J, Sutskever I, Abbeel P. 2016. InfoGAN:
- 449 Interpretable representation learning by information maximizing generative450 adversarial nets. In: Advances in Neural Information Processing Systems.
- 451 Chollet F. 2015. Keras: Deep Learning library for Theano and TensorFlow. GitHub
- 451 Chollet F. 2015. Keras: Deep Learning library for Theano and TensorFlow. GitHub
 452 Repos.
- 453 Colonna V, Ayub Q, Chen Y, Pagani L, Luisi P, Pybus M, Garrison E, Xue Y, Tyler-
- 454 Smith C, Abecasis GR, et al. 2014. Human genomic regions with exceptionally high
- 455 levels of population differentiation identified from 911 whole-genome sequences.456 Genome Biol. doi:10.1186/gb-2014-15-6-r88.
- 457 Davidsen K, Olson BJ, DeWitt WS, Feng J, Harkins E, Bradley P, Matsen FA. 2019.
- 458 Deep generative models for T cell receptor protein sequences. Elife. 8.
- doi:10.7554/eLife.46935. https://elifesciences.org/articles/46935.
- Diaz-Papkovich A, Anderson-Trocme L, Gravel S. 2019. Revealing multi-scale
 population structure in large cohorts. bioRxiv. doi:10.1101/423632.
- 462 Donahue J, Krähenbühl P, Darrell T. 2016 May 31. Adversarial Feature Learning.
 463 http://arxiv.org/abs/1605.09782.
- 464 Dumoulin V, Belghazi I, Poole B, Mastropietro O, Lamb A, Arjovsky M, Courville A.
 465 2016 Jun 2. Adversarially Learned Inference. http://arxiv.org/abs/1606.00704.
- 466 Eriksson N, Macpherson JM, Tung JY, Hon LS, Naughton B, Saxonov S, Avey L,
- 467 Wojcicki A, Pe'er I, Mountain J. 2010. Web-based, participant-driven studies yield
- 468 novel genetic associations for common traits. PLoS Genet.
- 469 doi:10.1371/journal.pgen.1000993.
- Fedus W, Goodfellow I, Dai AM. 2018 Jan 23. MaskGAN: Better Text Generation via
 Filling in the_____. http://arxiv.org/abs/1801.07736.
- 472 Flagel L, Brandvain Y, Schrider DR. 2019. The unreasonable effectiveness of
- 473 convolutional neural networks in population genetic inference. Mol Biol Evol.474 doi:10.1093/molbev/msy224.
- 475 Goodfellow I, Pouget-Abadie J, Mirza M. 2014. Generative Adversarial Networks

- 476 (GANs) Tutorial. Neural Inf Process Syst.
- 477 Gurdasani D, Carstensen T, Tekola-Ayele F, Pagani L, Tachmazidou I, Hatzikotoulas
- 478 K, Karthikeyan S, Iles L, Pollard MO, Choudhury A, et al. 2015. The African Genome
- 479 Variation Project shapes medical genetics in Africa. Nature.
- 480 doi:10.1038/nature13997.
- 481 Han J, Kraft P, Nan H, Guo Q, Chen C, Qureshi A, Hankinson SE, Hu FB, Duffy DL,
- 482 Zhen ZZ, et al. 2008. A genome-wide association study identifies novel alleles
- 483 associated with hair color and skin pigmentation. PLoS Genet.
- 484 doi:10.1371/journal.pgen.1000074.
- 485 Hinton GE. 2007. Learning multiple layers of representation. Trends Cogn Sci.
 486 doi:10.1016/j.tics.2007.09.004.
- 487 Hinton GE, Salakhutdinov RR. 2006. Reducing the dimensionality of data with neural
 488 networks. Science (80-). doi:10.1126/science.1127647.
- Howie B, Marchini J, Stephens M. 2011. Genotype imputation with thousands of genomes. G3 Genes, Genomes, Genet. doi:10.1534/g3.111.001198.
- 491 Kingma DP, Welling M. 2013 Dec 20. Auto-Encoding Variational Bayes.
- 492 http://arxiv.org/abs/1312.6114.
- 493 Larochelle H, Bengio Y. 2008. Classification using discriminative restricted boltzmann
- 494 machines. In: Proceedings of the 25th International Conference on Machine495 Learning.
- Lawson DJ, Hellenthal G, Myers S, Falush D. 2012. Inference of population structure using dense haplotype data. PLoS Genet. 8(1):11–17.
- 498 doi:10.1371/journal.pgen.1002453.
- Ledig C, Theis L, Huszár F, Caballero J, Cunningham A, Acosta A, Aitken A, Tejani
 A, Totz J, Wang Z, et al. 2017. Photo-realistic single image super-resolution using a
 generative adversarial network. In: Proceedings 30th IEEE Conference on
- 502 Computer Vision and Pattern Recognition, CVPR 2017.
- Leitsalu L, Haller T, Esko T, Tammesoo ML, Alavere H, Snieder H, Perola M, Ng PC, Mägi R, Milani L, et al. 2015. Cohort profile: Estonian biobank of the Estonian
- 505 genome center, university of Tartu. Int J Epidemiol. doi:10.1093/ije/dyt268.
- Libbrecht MW, Noble WS. 2015. Machine learning applications in genetics and genomics. Nat Rev Genet. doi:10.1038/nrg3920.
- Liu Q, Lv H, Jiang R. 2019. HicGAN infers super resolution Hi-C data with generative
 adversarial networks. In: Bioinformatics. Vol. 35. Oxford University Press. p. i99–
 i107.
- 511 Lucas T, Tallec C, Verbeek J, Ollivier Y. 2018. Mixed batches and symmetric
- 512 discriminators for GAN training. In: 35th International Conference on Machine 513 Learning, ICML 2018.
- 514 Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, Zhao M, Chennagiri N,
- 515 Nordenfelt S, Tandon A, et al. 2016. The Simons Genome Diversity Project: 300
- 516 genomes from 142 diverse populations. Nature. doi:10.1038/nature18964.
- 517 Mardis ER. 2017. DNA sequencing technologies: 2006-2016. Nat Protoc.
- 518 doi:10.1038/nprot.2016.182.
- 519 Marnetto D, Huerta-Sánchez E. 2017. Haplostrips: revealing population structure

- 520 through haplotype visualization. Methods Ecol Evol. 8(10):1389–1392.
- 521 doi:10.1111/2041-210X.12747.
- 522 McKinney W. 2010. Data Structures for Statistical Computing in Python. In:
- 523 Proceedings of the 9th Python in Science Conference.
- 524 Mitt M, Kals M, Pärn K, Gabriel SB, Lander ES, Palotie A, Ripatti S, Morris AP,
- 525 Metspalu A, Esko T, et al. 2017. Improved imputation accuracy of rare and low-
- 526 frequency variants using population-specific high-coverage WGS-based imputation
- reference panel. Eur J Hum Genet. doi:10.1038/ejhg.2017.51.
- 528 Oliphant TE. 2007. SciPy: Open source scientific tools for Python. Comput Sci Eng.
- 529 Popejoy AB, Fullerton SM. 2016. Genomics is failing on diversity. Nature.530 doi:10.1038/538161a.
- 531 Radford A, Metz L, Chintala S. 2016. Unsupervised Representation learning with
- 532 Deep Convolutional GANs. Int Conf Learn Represent. doi:10.1051/0004-
- 533 6361/201527329.
- Rolnick D, Dyer EL. 2019. Generative models and abstractions for large-scale
 neuroanatomy datasets. Curr Opin Neurobiol. doi:10.1016/j.conb.2019.02.005.
- 536 Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A, Chen X. 2016.
- 537 Improved techniques for training GANs. In: Advances in Neural Information
- 538 Processing Systems.
- Sheehan S, Song YS. 2016. Deep Learning for Population Genetic Inference. PLoS
 Comput Biol. 12(3):1–28. doi:10.1371/journal.pcbi.1004845.
- 541 Sirugo G, Williams SM, Tishkoff SA. 2019. The Missing Diversity in Human Genetic 542 Studies. Cell. doi:10.1016/j.cell.2019.02.048.
- 543 Smolensky P. 1986. Information processing in dynamical systems: Foundations of 544 harmony theory. In: Parallel Distributed Processing Explorations in the Microstructure 545 of Cognition.
- Teh YW, Hinton GE. 2001. Rate-coded restricted boltzmann machines for facerecognition. In: Advances in Neural Information Processing Systems.
- Yale A, Dash S, Dutta R, Guyon I, Pavao A, Bennett K. 2019 Apr 24. Privacy
 Preserving Synthetic Health Data. https://hal.inria.fr/hal-02160496/.
- Zhang H, Xu T, Li H, Zhang S, Wang X, Huang X, Metaxas D. 2017. StackGAN: Text
- to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks.
- 552 In: Proceedings of the IEEE International Conference on Computer Vision.
- Zhang M, Song F, Liang L, Nan H, Zhang J, Liu H, Wang LE, Wei Q, Lee JE, Amos
- 554 CI, et al. 2013. Genome-wide association studies identify several new loci associated
- with pigmentation traits and skin cancer risk in European Americans. Hum Mol
- 556 Genet. doi:10.1093/hmg/ddt142.
- 557