Generalization Capability of the Diet Network Model on Genomic Data

Camille Rochefort-Boulanger[1,2,3], Léo Choinière[1,3], Jean-Christophe Grenier[1], Pierre-Luc Carrier[2], Julie Hussin[1,3]
1. Montreal Heart Institute 2. Mila 3. Faculty of medicine - Université de Montréal

1 Introduction

The recent availability of a vast amount of genomic data has raised hopes in the field of personalized medicine, motivating research in risk prediction of complex diseases at the individual level [1–4]. Despite all the efforts made towards risk prediction, building predictive models of complex human phenotypes is still a challenge. While machine learning is a promising approach given its capacity to model interactions between variables, the application of machine learning to genomic data poses an important obstacle, the *fat data problem*: the number of input features is orders of magnitude larger than the number of training examples, which causes models trained on genomic data to overfit.

A novel neural network architecture, the Diet Network (DN) [5], has been proposed to address the *fat data* problem by reducing considerably the number of free parameters arising from the high number of features. This architecture has been developed and experimented on the *1000 Genomes Project* (1000G) [6], a *fat* dataset, and has proven to be effective at determining individuals' population given their Single Nucleotide Polymorphisms (SNPs). However, given the heterogeneity of genomic data collection protocols and the high number of missing data in genomic datasets, it is important to assess if a trained model can generalize well to independent datasets.

To evaluate the generalization capability of the approach, we tested the model on individuals from an alternative dataset, the *Human Genome Diversity Project* (HGDP) [7]. We also evaluated population classifications on previously unseen populations, by training a DN model on non-admixed populations and investigating the model's predictions for admixed individuals.

2 Methods

Data We used the publicly available 1000G genotype data to train DN models. The HGDP dataset [7] was used as test. Autosomal SNPs from 3450 individuals the 1000G dataset come from the Genome-Wide Human SNP Array 6.0 by Affymetrix, phased and imputed using SHAPEIT2 [8]. We harmonized alleles to match the *Reference Human Genome Build 37* (GRCh37). To do so, we used Harmonizer tool [9] and BCFtools [10]. We removed SNPs with a minor allele frequency (MAF) lower than 0.05 and pruned SNPs in linkage disequilibrium (LD) using PLINK *indep-parwise* command (window size $=$ 50 kb, step size $=$ 5, r^2 threshold $=$ 0.5) [11]. The final set of SNPs used for DN training contains 294,427 SNPs (refers herein as 1000G dataset). The HGDP genotype data comes from the Stanford HGDP Dataset, that genotyped 1043 individuals on Illumina HumanHap650Y array [12]. We performed a lift over of positions from GRCh36 to GRCh37 (positions with no coordinates were excluded), we excluded SNPs with MAF$<$0.01 and performed standard quality control steps [5]. The alleles were then harmonized to GRCh37, as described above. SNPs overlapping with 1000G dataset were retained, yielding 256,160 SNPs in HGDP (refers herein as HGDP dataset). A separate dataset was constructed to run RFMix (see below), selecting 15 non-admixed populations of the 26 included in the 1000G dataset within three continental populations (reference populations): European (EUR, N=670), African (AFR, N=783) and East Asian (EAS, N=617). We used 418,230 SNPs pruned as described above, but no MAF filter was applied.

Improvement on Diet Network method We have made modifications to the published DN implementation [5]. These are mainly brought forward to alleviate the training process and add new functionalities. Briefly, biases have been eliminated in the last layer of the auxiliary network, which improved the time per epoch, the amount of learning done per epoch and has stabilized the performance of the model during training. A missing data functionality has been added to allow the DN

to be applied to datasets that do not include all the SNPs used during training. Missing values are set to 0 after the pre-processing of genotypes and the remaining inputs are multiplied by a constant to compensate the missing values. In this work, out of the 294,427 SNPs used for training with the 1000G dataset, $\sim$250K SNPs were found in the HGDP dataset, other SNPs were given to the model as missing values. The mean misclassification error obtained for the improved model trained on 1000G dataset is 4.96 ± 0.19 compared to 7.44 ± 0.45 for the previous published model. [Code can be shared after double-blind review]

Models Three DN models were trained on the 1000G dataset. For all models, the embedding provided to the auxiliary network was the genotypic frequencies computed in populations over which the classification task was done (26, 15 or 3 populations). Inputs to the network were provided using an additive encoding scheme where alternative alleles are counted, such that genotype 2 corresponds to the homozygote with 2 alternative alleles. Hyperparameters proposed in the original publication were maintained for training, and models were trained using 5-fold cross validation. The first model (Model 1) was trained on 1000G dataset with the harmonized set of SNPs and tested on HGDP dataset. The two other models were trained using only non-admixed 1000G populations (15 of the 26 populations). We defined a first model where individuals were classified into 15 populations (15 labels, Model 2) and a second model where individuals from populations within continents (EUR, AFR and EAS) were merged under the same label (3 labels, Model 3). These two models were tested on Americans of African Ancestry in South-West USA (ASW) individuals, an admixed population never seen at training. Models return, for each individual, a vector of class-membership probabilities over populations seen at training. These probabilities are referred to as DN scores.

Population genetics analyses We used the F-statistic (F_{ST}) to measure genetic differentiation between populations. Specifically, we computed pairwise F_{ST} for all HGDP and 1000G population pairs according to [13] using vcftools [14]. F_{ST} values were calculated on common SNPs between 1000G and HGDP dataset. We took the mean F_{ST} computed on all SNPs to get a single F_{ST} value for each population pair. We used RFMix [15], a state-of-the-art local ancestry inference method, to assess proportions of ancestry in individuals from populations of interest. The method looks at haplotypes (obtained by SHAPEIT2) in a reference panel and uses random forest parametrization to infer local ancestry. Inferences are compiled over the genome to produce proportions for each population in the reference panel. We compared these RFMix proportions for ASW individuals with the scores obtained by testing the DN on the same individuals. Furthermore, using results from RFMix on each haplotype, we can compute the number of positions in the genome for which the two alleles of a SNP come from different ancestry populations. From these measures, we computed the proportion of the genome that is ancestrally heterozygous.

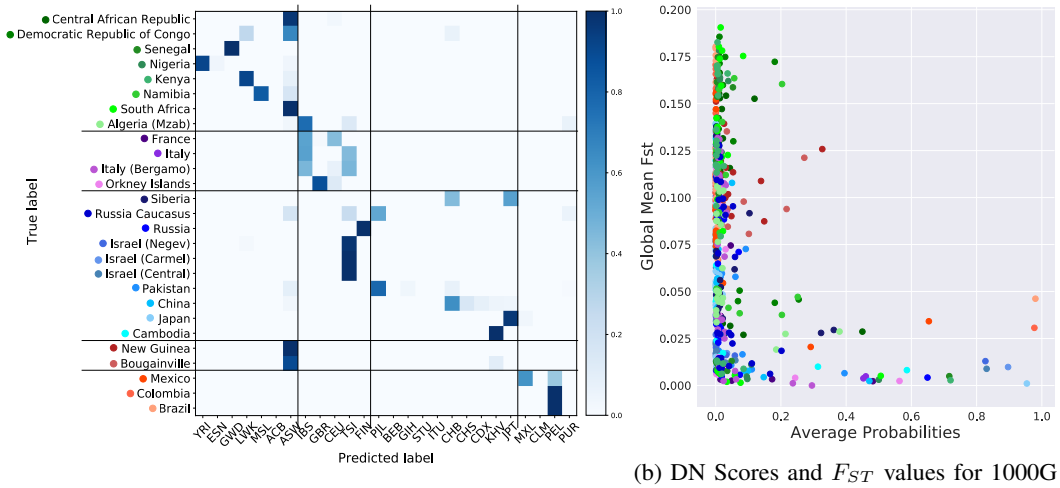
3 Results

In order to assess the generalization capability of the DN approach, we set up an experiment where we trained a DN model on a population stratification task using one dataset coming from specific sample recruitment and genotyping protocols, and tested the trained model on a second dataset generated with different protocols. Here, we compare 1000G dataset (training data), genotyped on an Affymetrix genotyping platform in 2013 to the HGDP dataset (test data) genotyped on an Illumina genotyping platform circa 2007. To make the approach applicable to the HGDP dataset, we trained and tested the model on a set of harmonized SNPs (Model 1, see Methods). This allows us to encode HGDP individuals' SNPs so they could be fed to the network without ambiguity, such that alleles represent nucleotides coming from the same DNA strand in both datasets.

Model 1 was then used to test the 1043 individuals of the HGDP dataset. Figure 1a shows the predictions made by the model for these HGDP individuals. Despite the fact that labels in the HGDP dataset are not equivalent to the ones in the 1000G dataset, reflecting different sampling locations, the confusion matrix shows that the predictions match the true labels at the continental level for most individuals. However, some out-of-continent classifications were made by the model. For example, individuals from Oceania (New Guinea and Bougainville), which are only found in the HGDP dataset, were classified as ASW, a highly admixed population. This classification makes sense since individuals from New Guinea and Bougainville are known to have African ancestors [16].

To further our analysis of the classifications made by the network, we compared the scores returned by the DN with F_{ST}, a statistic measuring population differentiation. We show in Figure 1b that the

DN model returns high scores for 1000G populations that have a low F_{ST} value (low differentiation) with the HGDP population of the individuals being classified.



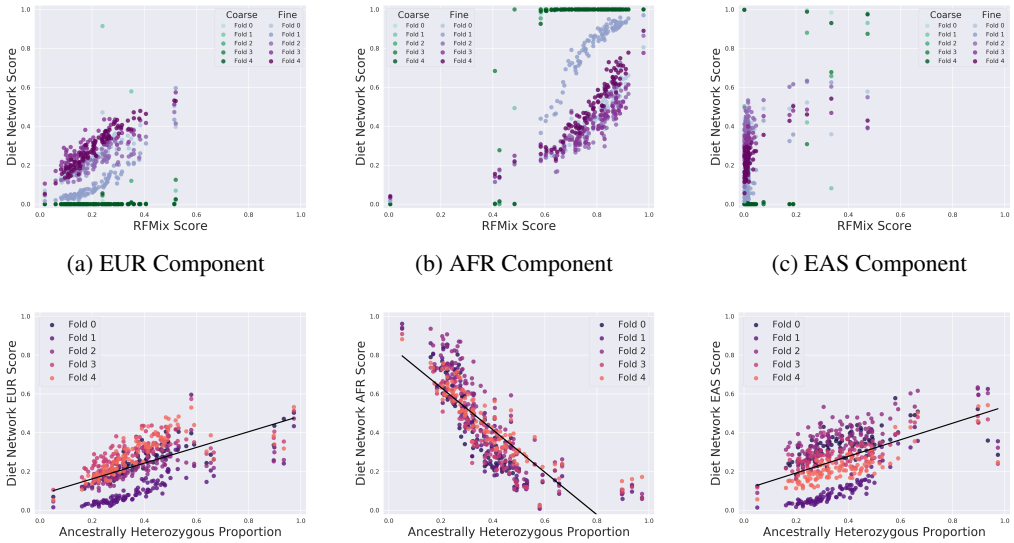
(a) Confusion Matrix

(b) DN Scores and F_{ST} values for 1000G and HGDP population pairs

Figure 1: (a) Confusion matrix of classifications made by Model 1 tested on new individuals from the HGDP dataset. Since the DN model was trained using 5-fold cross validation, scores for HGDP individuals from all folds were averaged to get a single prediction by individual. True labels (y-axis) are not equivalent to the 1000G labels predicted by the network (x-axis). Black lines separate continental populations. (b) Relationship between DN scores and F_{ST} values for HGDP/1000G population pairs. DN scores were averaged on individuals from the same HGDP population. Each HGDP population is represented in the plot by 26 dots (see Fig. 1a for colors associated to HGDP populations). Each dot indicates the F_{ST} value between the given HGDP and a 1000G populations (y-axis) and the DN score for the 1000G population averaged on all individuals of the given HGDP population (x-axis).

To interpret the scores returned by the DN approach, we trained models (Models 2 and 3) on only non-admixed populations of the 1000G dataset. Scores returned by the two models for ASW individuals, which are admixed individuals unseen during training, were compared with ancestry proportions computed by RFMix on EUR, AFR and EAS populations. Figure 2a,2b,2c shows relationships between DN scores for the *fine* (Model 2) and *coarse* (Model 3) models and RFMix ancestry proportions for each of the three reference populations. In the *coarse* model, we observe a bimodal behavior: the DN scores for EUR and AFR are 0 and 1 respectively, unless the EUR proportion exceeds 0.5. This indicates that when trained with labels defined at the continental level, the network does not capture the admixture in the ASW genomes. For the *fine* model on the other hand, a positive correlation between the scores and proportions returned by RFMix is observed, which suggests that this model captures the admixed nature of the genome of ASW individuals. However, the *fine* model shows a clear overestimation of EAS scores when compared to RFMix's proportions (Figure 2c).

To understand this overestimation of the EAS component by the DN, we used RFMix to compute the proportion of the genome where the two chromosomes of an individual are of different ancestry (ancestrally heterozygous regions). We refer to this as the ancestrally heterozygous proportion (AHP), calculated for each ASW individual. These ancestrally heterozygous regions are only observed in admixed populations and therefore represent special cases never seen by the DN model during training. We report in Figure 2d,2e,2f the relationship between DN scores and AHP for ASW individuals. The relationship for EAS component shows increased DN scores for individuals with higher AHP (r^2 = 0.29). This suggests that the overestimation of the EAS component in ASW individuals is partly explained by ancestrally heterozygous regions found in admixed populations. Moreover, Figure 2e and 2d shows that the DN model returns smaller AFR scores and larger EUR scores for individuals with a higher AHP. This reflects the fact that ASW individuals are descendant of African populations that have then been mixed with Europeans.



(a) EUR Component (b) AFR Component (c) EAS Component

(d) EUR Component (e) AFR Component (f) EAS Component

Figure 2: EUR(a), AFR(b) and EAS(c) scores for models trained on the 1000G dataset and tested on ASW compared to ancestry proportions inferred by RFMix. In purple are the scores returned by Model 2 (*fine*), in green by Model 3 (*coarse*). Correlation between *fine* model scores and RFMix proportions, from left to right, are $r^2 = 0.47$, $r^2 = 0.42$ and $r^2 = 0.18$. EUR(d), AFR(e) and EAS(f) scores returned by the *fine* model compared to the ancestrally heterozygous proportion (AHP) from RFMix. From left to right we have $r^2 = 0.34$, $r^2 = 0.41$ and $r^2 = 0.29$.

4 Conclusion

In this work, we showed that the DN approach is applicable to a new dataset. In fact, we have shown that the DN model trained on 1000G dataset can classify individuals of HGDP dataset reasonably well at the continental level. The comparison of DN scores with F_{ST} values shows that the model gives higher scores to 1000G populations that are genetically similar to the given HGDP population. However, HGDP and 1000G populations are relatively similar and good classification at the continental level was expected. To further investigate the generalization capability of the DN, we will test it on populations with peculiar demographic history, such as the French Canadians from Quebec, a population with a founder effect. We have also shown that the DN scores capture intricacies and the composite nature of the ASW genomes, even if it was not trained to recognize admixture. Indeed, scores returned by Model 3 correlate well with ancestry proportions inferred by RFMix ($r^2 > 0.4$), which suggests that a model trained to recognize fine population structure learns a representation which integrates discrete ancestry signal over the genome. On the contrary, when trained with labels defined at the continental level, the DN scores don't capture ancestry composition in the admixed ASW population, which suggests that this DN model only has a superficial understanding of how genetic diversity is mapped to define a population. This is to be expected, as the task performed here was only to differentiate continental populations, which is a relatively easy task. Also, while investigating the DN scores, we noted an overestimation of the EAS component in ASW. We showed that this overestimation can be explained by ancestrally heterozygous regions as reported by RFMix. Despite all those findings, the proposed approach to interpret information leveraged by the network is low resolution. In the next steps, we will be using integrated gradients [17] to interpret information leveraged by the network at the SNP level and try to reveal a more precise picture of what the DN learns.

The DN approach proposes a first attempt in addressing the *fat data* problem in genomics. The fact that the network is able to generalise well on genomic data from independent datasets is promising for predicting more complex phenotypes such as human diseases using data from multiple sources. Furthermore, discerning the features leveraged by the model to make predictions will be crucial to assess the validity of these predictions and could be extended to reveal new insights on complex diseases.

References

[1] N. R. Wray, M. E. Goddard, and P. M. Visscher, "Prediction of individual genetic risk of complex disease," *Current Opinion in Genetics & Development*, vol. 18, no. 3, pp. 257–263, 2008. DOI: https://doi.org/10.1016/j.gde.2008.07.006. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0959437X08000853.

[2] N. Chatterjee, J. Shi, and M. Garcıa-Closas, "Developing and evaluating polygenic risk prediction models for stratified disease prevention," *Nature Reviews Genetics*, vol. 17, 392 EP -, May 2016. [Online]. Available: https://doi.org/10.1038/nrg.2016.27.

[3] P. He, X. Lei, D. Yuan, Z. Zhu, and S. Huang, "Accumulation of minor alleles and risk prediction in schizophrenia," *Scientific Reports*, vol. 7, no. 1, p. 11 661, 2017. DOI: 10.1038/s41598-017-12104-0. [Online]. Available: https://doi.org/10.1038/s41598-017-12104-0.

[4] C. Kooperberg, M. LeBlanc, and V. Obenchain, "Risk prediction using genome-wide association studies," *Genetic epidemiology*, vol. 34, no. 7, pp. 643–652, Nov. 2010. DOI: 10.1002/gepi.20509. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/20842684.

[5] A. Romero, P. L. Carrier, A. Erraqabi, T. Sylvain, A. Auvolat, E. Dejoie, M.-A. Legault, M.-P. Dube, J. Hussin, and Y. Bengio, "Diet networks: Thin parameters for fat genomics," *ICLR*, 2017. DOI: arXiv:1611.09340. [Online]. Available: https://arxiv.org/abs/1611.09340.

[6] T. 1. G. P. Consortium, "A global reference for human genetic variation," *Nature*, 2015. DOI: 10.1038/nature15393.

[7] 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, no. 5866, pp. 1100–1104, 2008, ISSN: 0036-8075. DOI: 10.1126/science.1153717. eprint: https://science.sciencemag.org/content/319/5866/1100.full.pdf. [Online]. Available: https://science.sciencemag.org/content/319/5866/1100.

[8] O. Delaneau, B. Howie, A. J. Cox, J.-F. Zagury, and J. Marchini, "Haplotype estimation using sequencing reads," *The American Journal of Human Genetics*, vol. 93, no. 4, pp. 687–696, 2013, ISSN: 0002-9297. DOI: https://doi.org/10.1016/j.ajhg.2013.09.002. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0002929713004151.

[9] P. Deelen, M. J. Bonder, K. J. van der Velde, H.-J. Westra, E. Winder, D. Hendriksen, L. Franke, and M. A. Swertz, "Genotype harmonizer: Automatic strand alignment and format conversion for genotype data integration," *BMC Research Notes*, vol. 7, no. 1, p. 901, 2014. DOI: 10.1186/1756-0500-7-901. [Online]. Available: https://doi.org/10.1186/1756-0500-7-901.

[10] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1. G. P. D. P. Subgroup, "The sequence alignment/map format and samtools," *Bioinformatics (Oxford, England)*, vol. 25, no. 16, pp. 2078–2079, Aug. 2009. DOI: 10.1093/bioinformatics/btp352. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/19505943.

[11] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, and P. C. Sham, "Plink: A tool set for whole-genome association and population-based linkage analyses," *American journal of human genetics*, vol. 81, no. 3, pp. 559–575, Sep. 2007. DOI: 10.1086/519795. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/17701901.

[12] S. University. (2007). Human genome diversity project, [Online]. Available: http://www.hagsc.org/hgdp/files.htm.

[13] G. Bhatia, N. Patterson, S. Sankararaman, and A. L. Price, "Estimating and interpreting fst: The impact of rare variants," *Genome research*, vol. 23, no. 9, pp. 1514–1521, Sep. 2013. DOI: 10.1101/gr.154831.113. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/23861382.

[14] P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, R. Durbin, and 1. G. P. A. Group, "The variant call format and VCFtools," *Bioinformatics*, vol. 27, no. 15, pp. 2156–2158, Jun. 2011, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btr330. eprint: http://oup.prod.sis.

lan/bioinformatics/article-pdf/27/15/2156/1125001/btr330.pdf. [Online]. Available: https://doi.org/10.1093/bioinformatics/btr330.

[15] B. K. Maples, S. Gravel, E. E. Kenny, and C. D. Bustamante, "Rfmix: A discriminative modeling approach for rapid and robust local-ancestry inference," *American journal of human genetics*, vol. 93, no. 2, pp. 278–288, Aug. 2013. DOI: 10.1016/j.ajhg.2013.06.020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/23910464.

[16] A. Wollstein, O. Lao, C. Becker, S. Brauer, R. J. Trent, P. Nürnberg, M. Stoneking, and M. Kayser, "Demographic history of oceania inferred from genome-wide data," *Current Biology*, vol. 20, no. 22, pp. 1983–1992, 2010. DOI: https://doi.org/10.1016/j.cub.2010.10.040. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0960982210013436.

[17] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," *CoRR*, vol. abs/1703.01365, 2017. arXiv: 1703.01365. [Online]. Available: http://arxiv.org/abs/1703.01365.