Overcoming Data Limitations in Animal Genetics with GAN-based Synthetic Genotypes

Sihan Xie [1], Thierry Tribout [1], Blaise Hanczar [2], Julien Chiquet [3], Eric Barrey [1]

[1] Université Paris-Saclay, AgroParisTech, INRAE, GABI, 78350 Jouy-en-Josas, France [2] IBISC, University Paris-Saclay (Univ. Evry) [3] UMR MIA-Paris, Université Paris-Saclay, AgroParisTech, INRAE

Context

Real genotype data is hard to access due to privacy, cost, and regulatory constraints, posing challenges for animal genetics research. We overcome these issues by employing deep learning models to generate synthetic genotype conditioned on phenotype.

Data

We train our model using a Holstein cow dataset that contains genotype and milk production traits from 93484 cows, featuring 50161 single-nucleotide polymorphism (SNPs) distributed across 29 chromosomes.

Model

We developed a Wasserstein Generative Adversarial Network with Gradient Penalty (WGAN-GP) that leverages a Gumbel-Softmax mechanism to generate discrete synthetic genotype data conditioned on milk production phenotype:

$$\min_G \max_D \underbrace{\mathbb{E}_{x\sim P_r}[D(x|c)] - \mathbb{E}_{z\sim P_z}[D(G(z|c))]}_{\text{Wasserstein distance between real and synthetic distribution}} \underbrace{-\lambda\,\mathbb{E}_{\hat{x}\sim P_{\hat{x}}}\left(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1\right)^2}_{\text{enforce Lipschitz constraint}}$$

where G is a Generator Network and D is a Discriminator Network.

Quantitative Metrics from Deep Learning Literature

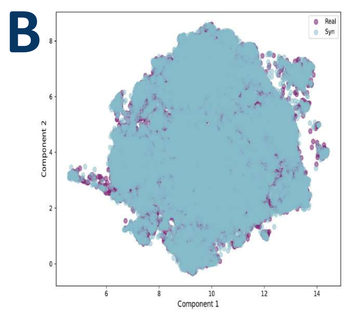
Results for Chromosome 14 (1771 SNPs)

	Precision (%) ↑	Recall (%) ↑	F1 (%) ↑	Correlation Score (%) ↑
Vanilla GAN	80.88 ± 0.23	57.97 ± 0.45	67.53 ± 0.24	72.60 ± 0.13
WGAN-GP	**99.64 ± 0.04**	**99.88 ± 0.02**	**99.76 ± 0.02**	**98.65 ± 0.02**

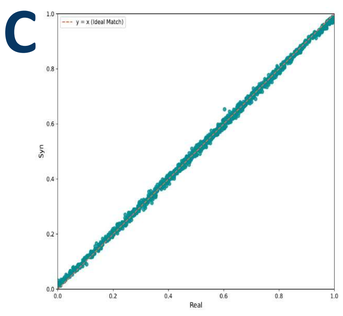
Results for All Chromosomes (50161 SNPs)

	Precision (%) ↑	Recall (%) ↑	F1 (%) ↑	Correlation Score (%) ↑
Vanilla GAN	**100 ± 0.00**	0.00 ± 0.00	0.00 ± 0.00	0.52 ± 0.01
WGAN-GP	92.00 ± 0.16	**99.93 ± 0.01**	**95.80 ± 0.09**	**83.32 ± 0.06**

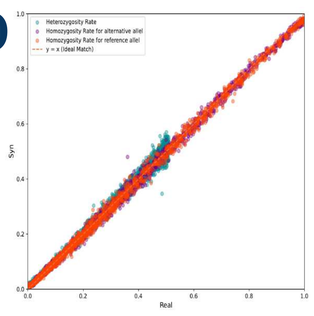
Results for Chromosome 14 (1771 SNPs)



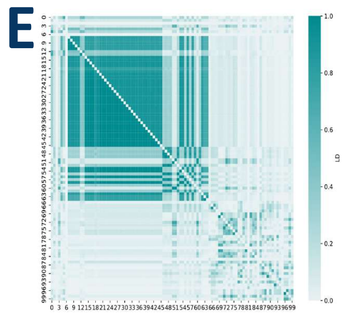
B UMAP Comparaison



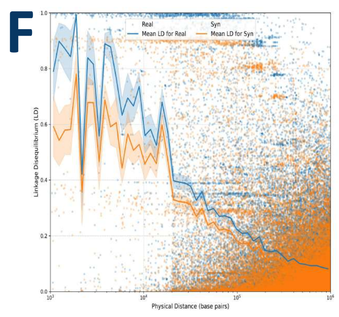
C Allele Frequency Comparaison



D Genotype Frequency Comparaison

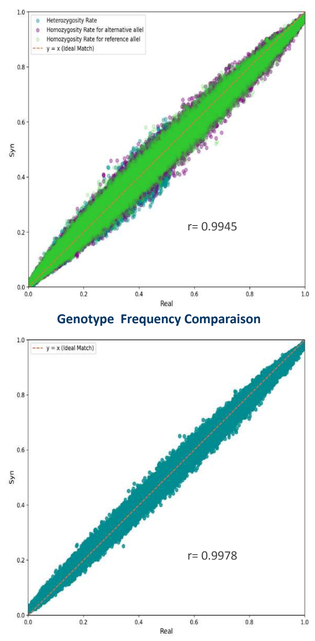


E Linkage Disequilibrium Comparaison

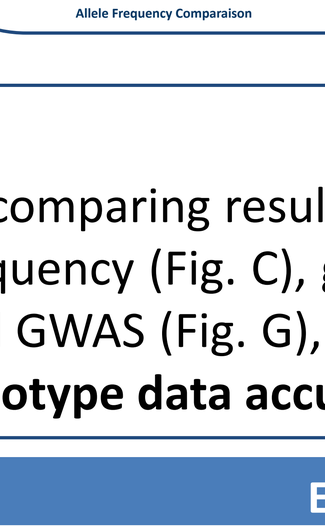


F Linkage Disequilibrium Decay Comparaison

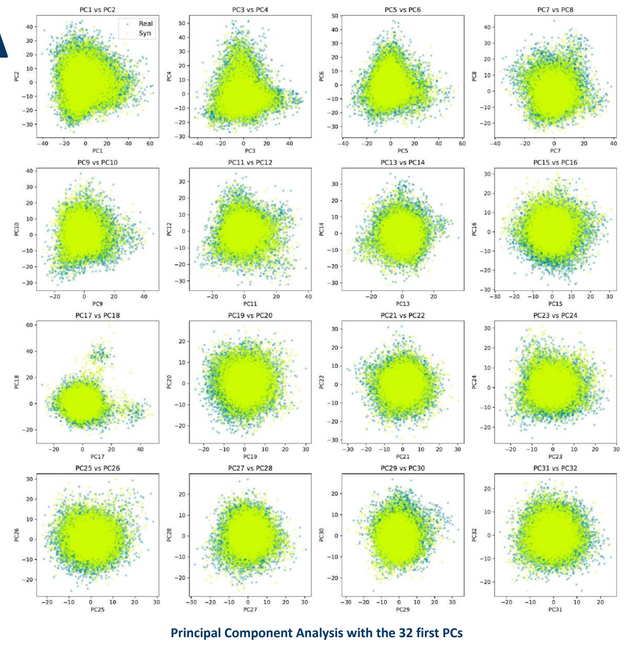
Results for All Chromosomes (50161 SNPs)



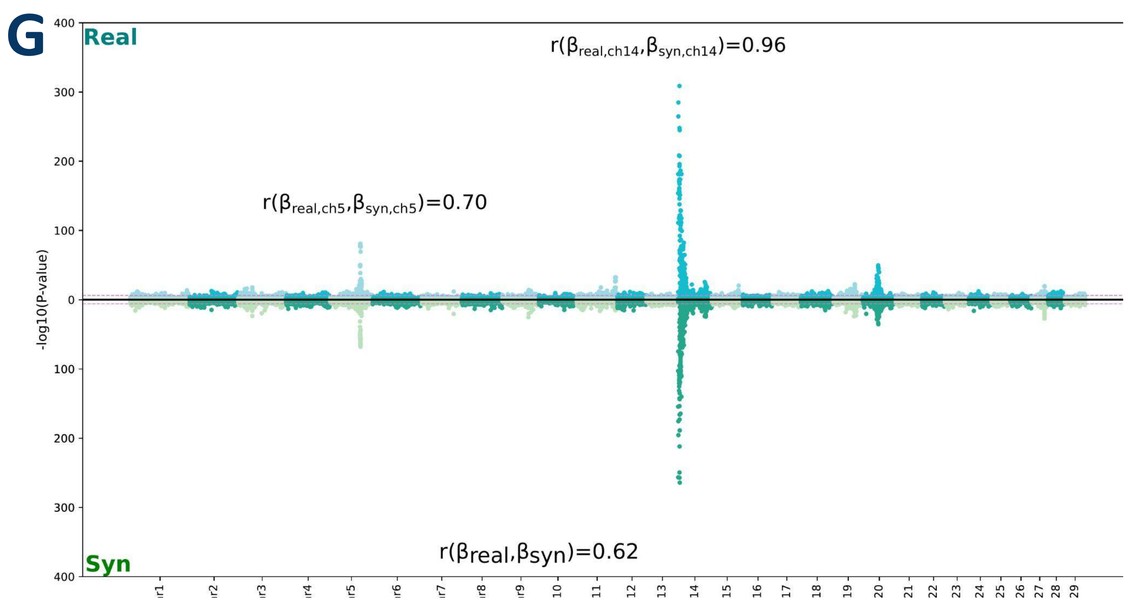
D Genotype Frequency Comparaison



C Allele Frequency Comparaison



A Principal Component Analysis with the 32 first PCs



G Genome-Wide Association Analysis (GWAS) Comparaison

Conclusion

By comparing results for real and synthetic data across several quantitative metrics, PCA (Fig. A), UMAP (Fig. B), allele frequency (Fig. C), genotype frequency (Fig. D), linkage disequilibrium and its decay with physical distance (Figs. E–F), and GWAS (Fig. G), we demonstrate that our **WGAN-GP combined with Gumbel Softmax model generated synthetic genotype data accurately capture key genetic structures while preserving critical genotype–phenotype associations**.

We thank the INRAE GABI GBOS Team for providing and preparing the dataset used in this study.

EAAP Artificial Intelligence For Animal Conference: Zurich, 4-6 June 2025, Switzerland