# DeepSelectGene - Apprentissage profond à partir de données de génotypes et application à la sélection génomique

(01/12/2023 – 01/12/2026)

Doctorante: Sihan XIE (Université Paris-Saclay, AgroParisTech, INRAE, GABI-UMR1313, 78350 Jouy-en-Josas, France)
Presented by Julien CHIQUET (MIA-PS, Université Paris-Saclay, AgroParisTech, INRAE, France)

*Séminaire du Métaprogramme DIGIT-BIO : 12 et 13 décembre 2024, Hôtel Valpré - Ecully*

- ☐ Axe 1 : Décryptage multi-échelle des fonctions du vivant
- ✔ Axe 2 : Prédiction de phénotypes
- ☐ Axe 3 : Transfert et généralisation
- ☐ Axe 4 : Jumeaux numériques

# Contexte et Enjeux

## Technologies de séquençage en sélection génomique

- Données de génotypage haute densité (G)
- Prédiction de phénotypes de descendants (valeur génétique) (P)

## Difficultés et Limites des modèles classiques (GBLUP)

- Gestion de données de grande dimension
- Limités aux interactions GxG, GxP, PxP simples et linéaires

# Thèse de Sihan Xie

## Objectif : nouveaux modèles d'apprentissage profond (DL)

- Adapter les modèles de DL aux données génomiques
- Modéliser des interactions complexes pour mieux prédire
- Gérer le peu d'exemples disponibles (génération de génotypes)

## Encadrement interdisciplinaire

- Eric Barrey (GABI, INRAE)
  *Génétique animale*
- Julien Chiquet (MIA Paris-Saclay)
  *Modèle et Apprentissage statistiques*
- Blaise Hanczar (IBISC, UEVE)
  *DL pour les données génomique*

# Données d'apprentissage

## Génotypage : Single Nucleotide Polymorphism (SNP)

Variation nucléotidique en une position du génome d'un individu vis-à-vis d'une population, codée avec 3 valeurs $\{0, 1, 2\}$.

- *Biologiquement,*
  - 0 = Homozygote pour l'allèle de référence
  - 1 = Hétérozygote
  - 2 = Homozygote pour l'allèle alternatif
- *Statistiquement*, il s'agit d'un comptage de l'allèle alternatif.

# Données d'apprentissage

## Génotypage : Single Nucleotide Polymorphism (SNP)

Variation nucléotidique en une position du génome d'un individu vis-à-vis d'une population, codée avec 3 valeurs $\{0, 1, 2\}$.

- *Biologiquement,*
  - 0 = Homozygote pour l'allèle de référence
  - 1 = Hétérozygote
  - 2 = Homozygote pour l'allèle alternatif
- *Statistiquement*, il s'agit d'un comptage de l'allèle alternatif.

## Les jeux de données de la thèse (génotype + phénotype)

- $\approx 10K$ chevaux : 44K SNPs + 3 phenotypes
- $\approx 100K$ vaches Holstein : 54K SNPs + 33 phenotypes
- $\approx 500K$ humains (UKBioBank) : 800K SNPs et Indels + des dizaines de mesures corporelles

# Objectif 1 : Génération ou simulation artificielle de génotypes

## Motivation : nombre d'exemple de génotypage limité

- Caractère privés, accès limités, répétition à l'envie impossible
- Idée : générer des données de manière confidentielle et simuler des populations spécifiques à des fins de recherche

# Objectif 1 : Génération ou simulation artificielle de génotypes

## Motivation : nombre d'exemple de génotypage limité

- Caractère privés, accès limités, répétition à l'envie impossible
- Idée : générer des données de manière confidentielle et simuler des populations spécifiques à des fins de recherche

## Formalisation du problème

Pour $m$ loci identifiés et deux phénotypes d'intérêt, on note

- $\mathbf{G} = (g_1, g_2, \ldots, g_m)$ le vecteur des SNPs où $g_i \in \{0, 1, 2\}$,
- le sexe **S** (binaire) et la taille **H** (continue).

# Objectif 1 : Génération ou simulation artificielle de génotypes

## Motivation : nombre d'exemple de génotypage limité

- Caractère privés, accès limités, répétition à l'envie impossible
- Idée : générer des données de manière confidentielle et simuler des populations spécifiques à des fins de recherche

## Formalisation du problème

Pour $m$ loci identifiés et deux phénotypes d'intérêt, on note

- $\mathbf{G} = (g_1, g_2, \ldots, g_m)$ le vecteur des SNPs où $g_i \in \{0, 1, 2\}$,
- le sexe $\mathbf{S}$ (binaire) et la taille $\mathbf{H}$ (continue).

Objectif : apprendre $\mathbb{P}(\mathbf{G}|\mathbf{S} = s, \mathbf{H} = h)$ (distribution conjointe multivariée des SNPs conditionnée par $\mathbf{S}$ et $\mathbf{H}$).

Tâche complexe : estimer les fréquences alléliques marginales + les dépendances entre loci + les associations P x G

# Objectif 1 : Génération ou simulation artificielle de génotypes

**Les modèles considérés** :

1. Generative Adversarial Networks (GAN) :

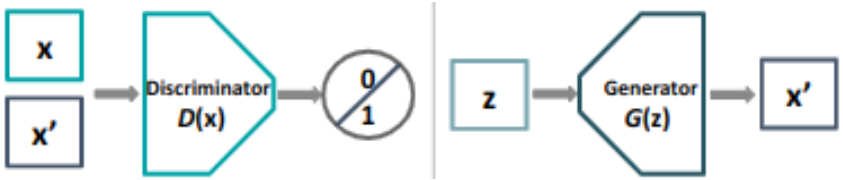



Figure – Architecture des GAN

**Deux réseaux sont entraînés de manière adversariale**

2. Diffusion Model :

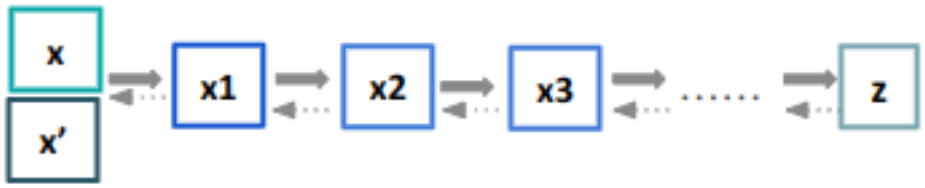



Figure – Architecture de Diffusion Model

**Apprendre à débruiter pour générer**

# Objectif 2 : Prédiction des phénotypes à partir des génotypes

## Formalisation du problème

Pour $m$ loci identifiés, on note

- $\mathbf{G} = (g_1, g_2, \ldots, g_m)$ le vecteur des SNPs où $g_i \in \{0, 1, 2\}$,
- le sexe **S** et diverses covariables environementales**E**.
- **Y**, le phénotype à prédire (lié à la production laitière chez les vaches, à la performance sportive chez les chevaux, à.).

# Objectif 2 : Prédiction des phénotypes à partir des génotypes

## Formalisation du problème

Pour $m$ loci identifiés, on note

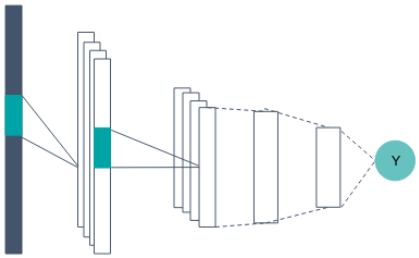
- $\mathbf{G} = (g_1, g_2, \ldots, g_m)$ le vecteur des SNPs où $g_i \in \{0, 1, 2\}$,
- le sexe $\mathbf{S}$ et diverses covariables environementales$\mathbf{E}$.
- $\mathbf{Y}$, le phénotype à prédire (lié à la production laitière chez les vaches, à la performance sportive chez les chevaux, à.).

Objectif : apprendre $P(\mathbf{Y} \mid \mathbf{G}, \mathbf{S}, \mathbf{E})$, la distribution conditionnelle aux génotypes et à l'environnement.

Défis : grande dimension et parcimonie des données SNPs

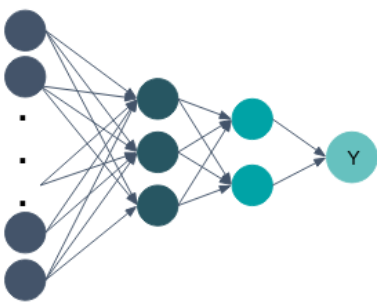
# Objectif 2 : Prédiction des phénotypes à partir des génotypes

**Les modèles considérés** :




**Les CNNs ne sont pas adaptés :**
Déplacer un motif de SNPs dans la séquence change son sens biologique, car les SNPs ne sont pas invariants au décalage.


**Les RNNs ne sont pas adaptés :**
Les SNPs ne sont pas des séries temporelles et ne reflètent pas de dépendances temporelles.




**Les MLP peuvent être adaptés :**
Les MLP traitent les SNP indépendamment.
$\Rightarrow$ De plus, intégrer des connaissances biologiques permettrait de rendre le modèle plus efficace.

# Objectif 3 : Génération de profils génomiques maximisant certains caractères d'intérêt

Motivation : Développer un outil d'aide à la sélection animale basé sur des profils génomiques optimaux.

## Méthodologie

Inspirée des méthodes d'interpolation des GAN pour explorer l'espace latent

- Combiner et figer nos modèles prédictifs et génératifs
- Explorer l'espace latent du générateur pour maximiser une fonction objective basée sur les prédictions du prédicteur.

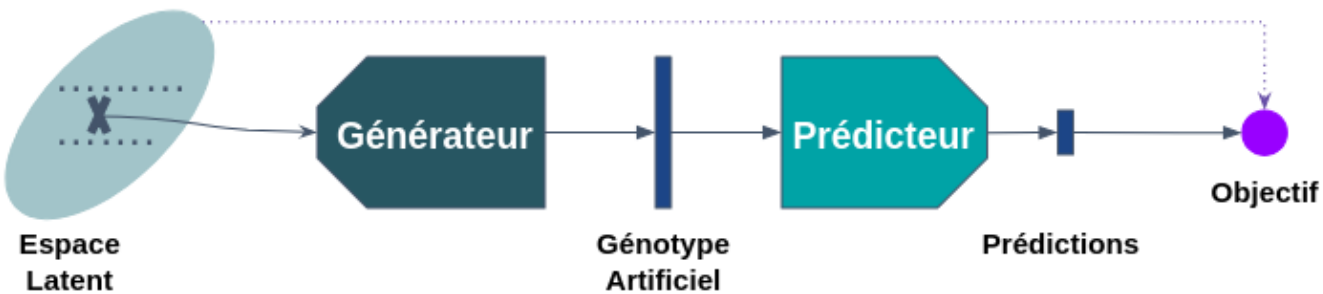
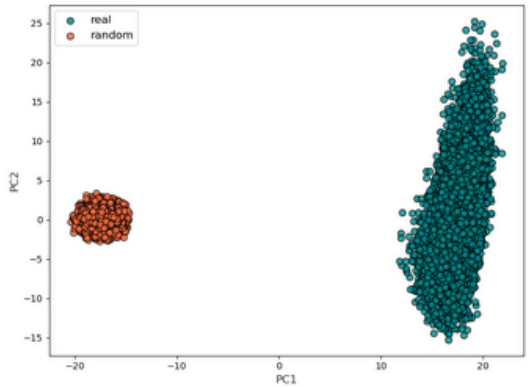


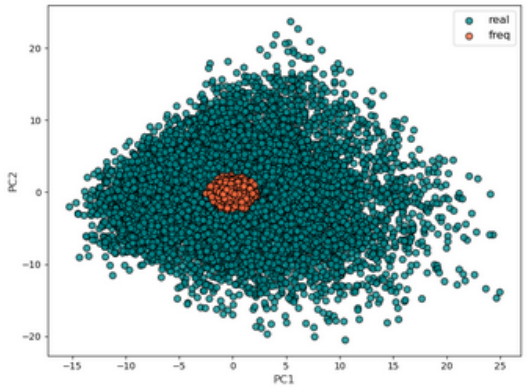

Figure – Schéma de recherche de profils génomiques optimaux

# Résultats obtenus

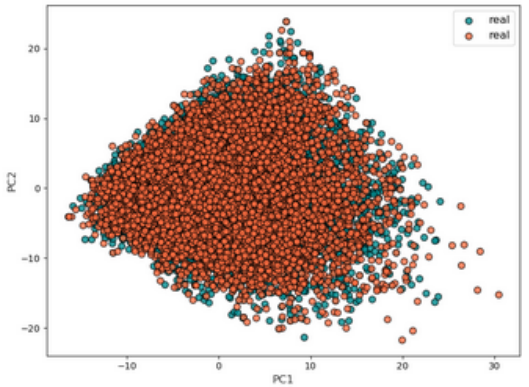
## Génération de génotypes artificiels d'un chromosome chez le bovin (≈ 4000 SNPs)



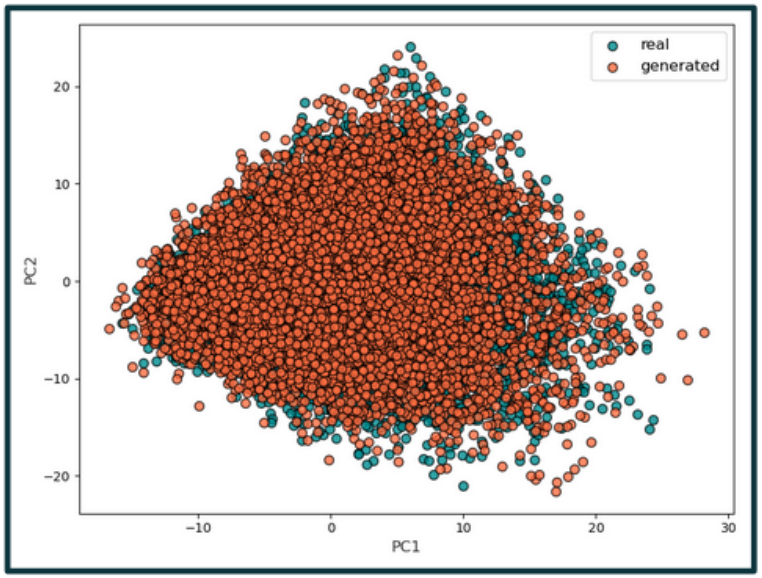

génotypes synthétiques sont totalement aléatoires




génotypes synthétiques échantillonnés en fonction des fréquences alléliques.




les deux populations sont réelles (cas idéal)

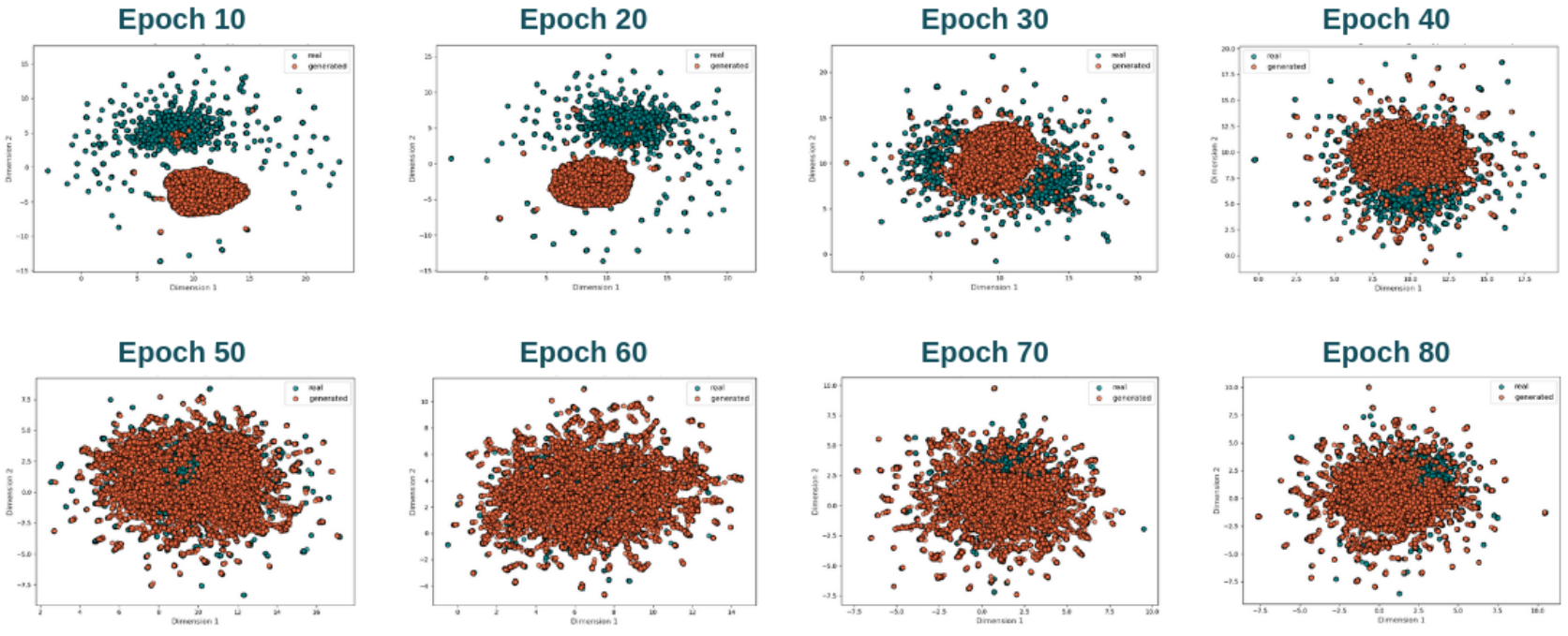



**Nos génotypes synthétiques générés à l'aide d'un modèle basé sur les GANs**

# Résultats obtenus

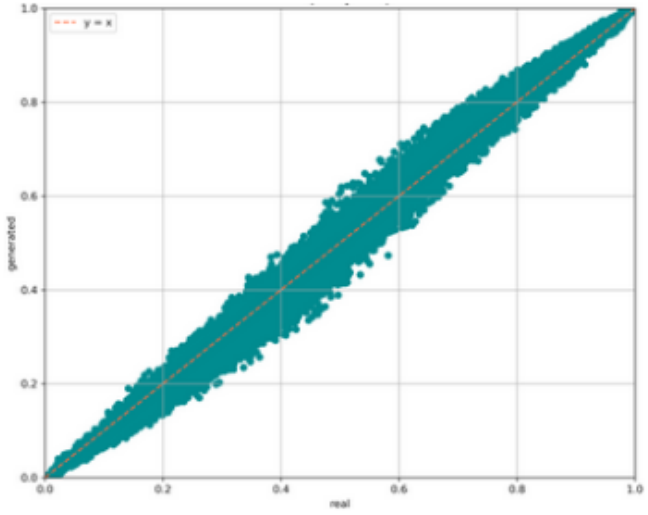
- *Si l'on étend le modèle à l'ensemble des chromosomes chez les bovins (≈ 50K SNPs), le modèle génératif reste-t-il pertinents ?*

⇝ Oui, mais il est nécessaire de concevoir des réseaux de neurones et des méthodes d'entraînement plus sophistiqués.

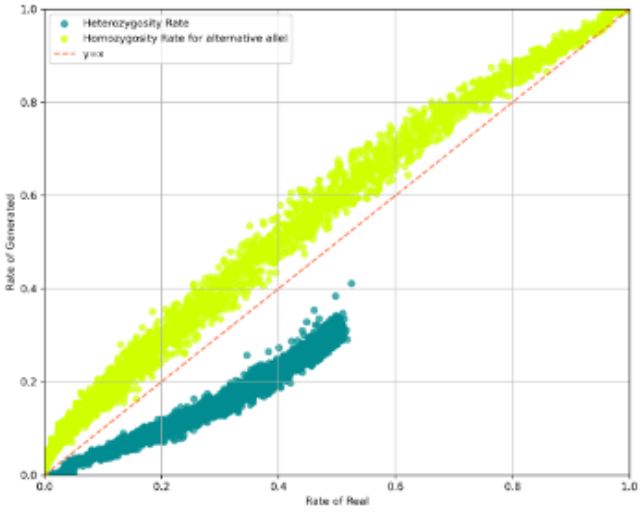


# Résultats obtenus

Au delà d'une comparaison visuelle des 2 distributions, d'autres métriques permettent de vérifier si les génotypes artificiels préservent les propriétés biologiques/génétiques des vrais génotypes




Comparaison des **fréquences alléliques** entre les génotypes réels et les génotypes synthétiques.

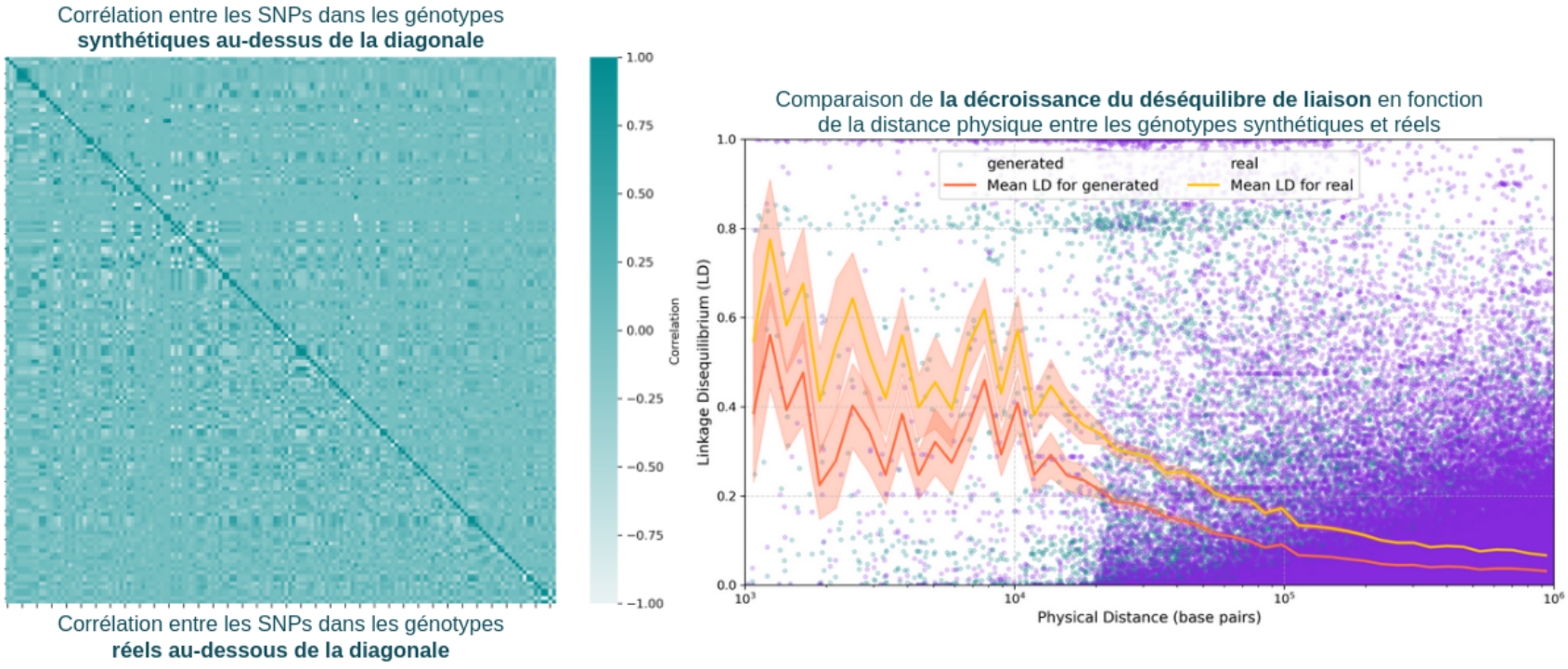



Comparaison des **fréquences génotypiques** entre les génotypes réels et les génotypes synthétiques.

Les génotypes synthétiques ont une fréquence allélique similaire, mais plus d'homozygotie pour les allèles alternatifs et moins d'hétérozygotie.

# Résultats obtenus

Au delà d'une comparaison visuelle des 2 distributions, d'autres métriques permettent de vérifier si les génotypes artificiels préservent les propriétés biologiques/génétiques des vrais génotypes

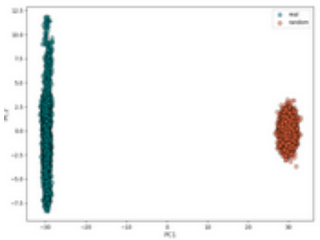



Les génotypes synthétiques semblent préserver la structure génétique type déséquilibre de liaison.

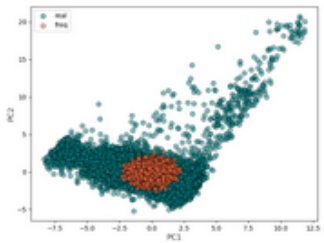
# Résultats obtenus

- Les modèles génératifs possèdent-ils l'universalité applicable à d'autres espèces ?

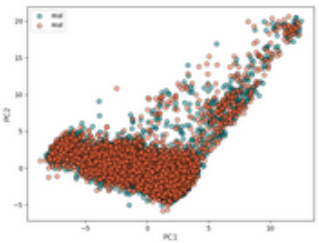
⇝ Pour un jeu de données présentant une plus grande diversité génétique et incluant différentes sous-populations, comme UK Biobank l'apprentissage devient plus complexe.



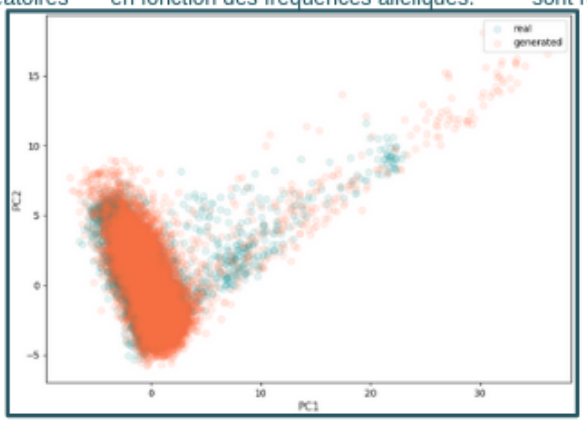
génotypes synthétiques sont totalement aléatoires



génotypes synthétiques échantillonnés en fonction des fréquences alléliques.



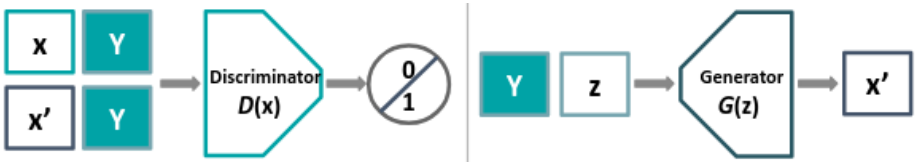
les deux populations sont réelles (cas idéal)



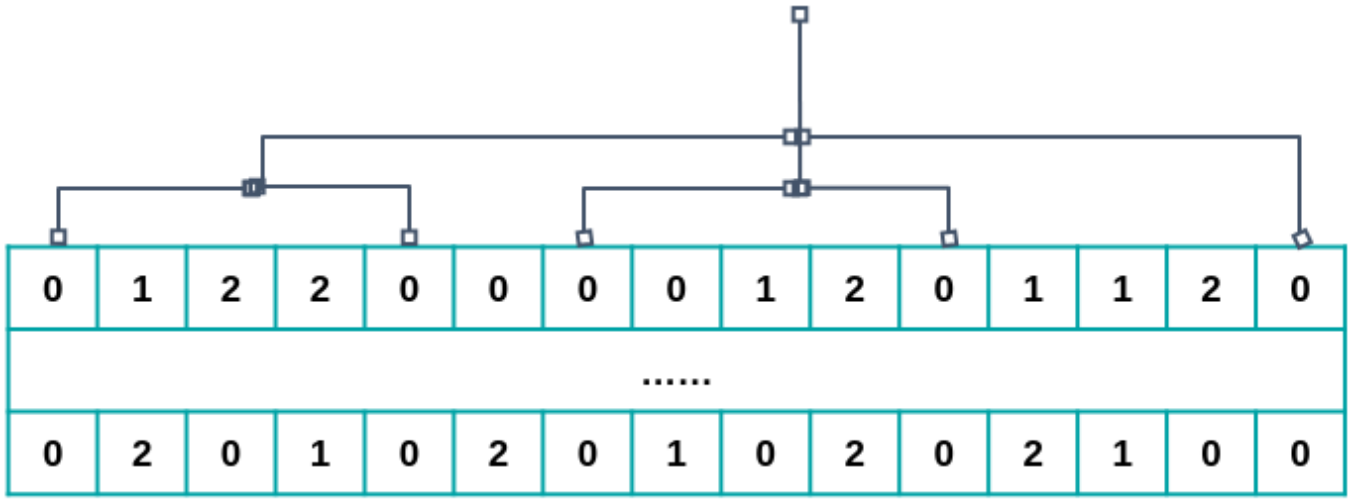

Nos génotypes synthétiques générés à l'aide d'un modèle basé sur les GANs

# Améliorations et développements futurs

1. Développement de modèles génératifs conditionnés par les phénotypes pour générer des paires Gx P (en cours)




2. Intégration du Linkage Disequilibrium dans le modèle génératifs (en cours)




**K-mers clustering basé sur LD pour vectoriser les SNPs**

3. Développement d'un framework d'évaluation pour les génotypes synthétiques (en cours)

# Merci de votre attention