THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE L'UNIVERSITÉ DE MONTPELLIER

En Bioinformatique

École doctorale : Sciences Chimiques et Biologiques pour la Santé

Unité de recherche : Insitut de Génétique Humaine

GENERATION DE GENOMES HUMAINS ARTIFICIELS VRAISEMBLABLES PAR DES AUTOENCODEURS ANTAGONISTES

Présentée par Callum BURNARD
Le 11 décembre 2023

Sous la direction de William RITCHIE et Alban MANCHERON

Devant le jury composé de

Daniel GAUTHERET, Professeur, I2BC	Rapporteur
Denis PUTHIER, Professeur, TAGC	Rapporteur
Magali RICHARD, Chargée de recherche, TIMC-IMAG	Examinatrice
Thérèse COMMES, Professeure des universités, IRMB	Présidente du Jury
William RITCHIE, Directeur de recherches, IGH	Directeur de thèse
Alban MANCHERON, Maître de conférences, LIRMM	Coencadrant de thèse

UNIVERSITÉ DE MONTPELLIER

Génération de génomes humains artificiels vraisemblables par des autoencodeurs antagonistes

Callum BURNARD

Sous la supervision de

William RITCHIE, IGH
Alban MANCHERON, LIRMM

Rapporteurs

Daniel GAUTHERET, I2BC
Denis PUTHIER, TAGC

Examinateurs

Magali RICHARD, TIMC-IMAG
Thérèse COMMES, IRMB

1. Table des matières

1. Table des matières..... 1
2. Remerciements..... 6
3. Abréviations utilisées..... 7
4. Résumé..... 8
5. Abstract..... 9
6. Introduction..... 10
6.1. Histoire de la génomique..... 10
6.1.1. Caractéristiques héréditaires..... 10
6.1.2. Chromosomes en tant que support de l'information biologique..... 11
6.1.3. Séquence de nucléotides..... 12
6.1.4. Séquençage de cette information..... 14
6.1.5. Modèle actuel..... 15
6.1.6. Génomique moderne..... 21
6.2. Bases de données biologiques..... 25
6.2.1. Bases de données existantes..... 25
6.2.1.1. Single Nucleotide Polymorphism Database..... 25
6.2.1.2. HapMap..... 26
6.2.1.3. 1000 Genomes Project..... 27
6.2.1.4. SNPedia..... 28
6.2.1.5. Human Variome project..... 28
6.2.1.6. Genome Aggregation Database..... 29
6.2.1.7. UK Biobank..... 30
6.2.1.8. Autres projets nationaux notables..... 31
6.2.2. Risques associés..... 32
6.2.3. Masquer l'information..... 34
6.2.3.1. Anonymisation..... 34
6.2.3.2. Chiffrement homomorphe..... 41
6.2.3.3. Apprentissage fédéré..... 43
6.2.3.4. Approche par k-mers..... 44
6.2.4. Génération de nouvelles données..... 45
6.2.4.1. Approches statistiques..... 46
6.2.4.2. Approche par intelligence artificielle..... 48
7. État de l'art..... 54
7.1. l'Intelligence Artificielle..... 54

7.1.1. L’intelligence artificielle en général....54
7.1.1.1. Histoire....55
7.1.1.2. Fonctionnement....56
7.1.1.3. Applications....57
7.1.2. Apprentissage automatique....58
7.1.2.1. Histoire....59
7.1.2.2. Fonctionnement....59
7.1.2.2.1. Apprentissage supervisé....63
7.1.2.2.2. Apprentissage non-supervisé....63
7.1.2.2.3. Apprentissage par renforcement....64
7.1.2.3. Applications....66
7.1.3. Les réseaux de neurones....67
7.1.3.1. Histoire....71
7.1.3.2. Forces et faiblesses....71
7.1.3.3. Les réseaux génératifs....72
7.1.3.3.1. Generative Adversarial Network....72
7.1.3.3.2. Restricted Boltzmann Machine....74
7.1.3.3.3. Diffusion stable....76
7.1.4. Questions éthiques autour de l’intelligence artificielle....77
7.1.4.1. Préoccupations générales....77
7.1.4.1.1. Biais dans les jeux de données....77
7.1.4.1.2. Collection de données....78
7.1.4.1.3. Biais dans les algorithmes....78
7.1.4.1.4. Transparence des modèles....79
7.1.4.1.5. Impact sur les emplois....79
7.1.4.1.6. Impact environnemental....80
7.1.4.2. Risques dans le domaine médical....80
7.1.4.2.1. Risques élevés pour certains facteurs....80
7.1.4.2.2. Priorité sur le bien-être du patient....81
7.1.4.2.3. Intégration dans le déroulement des opérations en milieu clinique....81
7.1.4.2.4. Accès équitable....81
7.1.4.2.5. Accélération de risques déjà existants....82
7.1.4.3. Risques liés aux modèles génératifs....82
7.2. Réduction de dimension par intelligence artificielle....85
7.2.1. Sélection de la méthode de réduction de dimensions....85
7.2.1.1. Convolution 1D....86

7.2.1.1.1. Principe de la convolution....86
7.2.1.1.2. Avantages....87
7.2.1.2. Carte autoadaptive....88
7.2.1.2.1. Principe....88
7.2.1.2.2. Forces et faiblesses....88
7.2.1.3. Autoencodeurs....89
7.2.1.3.1. Principe....89
7.2.1.4. Pourquoi la réduction de dimensions....90
8. Contributions....91
8.1. Mon projet....91
8.2. Contributions : Exploration des données....93
8.2.1. Données de mutations dans la population....93
8.2.1.1. Histoire du 1000 Genomes Project....93
8.2.1.2. Diversité ethnique du jeu de données....94
8.2.1.3. Les données....95
8.2.1.4. Utilisation par Yelmen et al. 98
8.2.2. Corrélation entre mutations....99
8.2.2.1. Régulation des recombinaisons méiotiques....99
8.2.2.2. Foyers de recombinaisons....100
8.2.2.3. Découpe du génome....101
9. Résultats....103
9.1. Autoencodage des données de mutations....103
9.1.1. Méthodes de séparation des mutations....103
9.1.1.1. Séparation par fenêtres de taille constante....103
9.1.1.2. Séparation selon les TAD....104
9.1.1.3. Séparation selon les foyers de recombinaison....105
9.1.2. Autoencodage des sections....106
9.1.2.1. Implémentation....106
9.1.2.2. Résultats des autoencodeurs....107
9.1.2.3. Baisses de performance....111
9.1.2.4. Prédictions à améliorer....114
9.1.3. Autoencodeurs variationnels....116
9.1.3.1. Spécificités....116
9.1.3.1.1. Sampling....116
9.1.3.1.2. Reparameterization trick....117
9.1.3.1.3. Divergence de Kullback-Leibler....117

9.1.3.2. Performances comparé aux autoencodeurs déterministes..... 118
9.1.3.3. VAE en tant que générateur..... 120
9.2. Cohérence entre sections..... 131
9.2.1. Pourquoi la cohérence..... 131
9.2.2. Étendeur d'espace latent..... 131
9.2.2.1. Principe..... 131
9.2.2.2. Implémentation de l'étendeur..... 131
9.2.2.3. Résultats..... 132
9.2.2.3.1. Mesures de l'erreur du modèle..... 133
9.2.2.3.2. Biais..... 136
9.2.3. Distance de Wasserstein dans la mesure de l'erreur..... 138
9.2.3.1. Application de ce modèle sur l'espace latent..... 138
9.2.3.2. Architecture..... 140
9.2.3.3. Entraînement..... 141
9.2.3.4. Résultats..... 141
9.2.3.4.1. Deux sections..... 141
9.2.3.4.2. Sections couvrant 1Mb d'ADN..... 142
9.2.3.4.3. Sections non consécutives..... 146
9.3. Évaluation des profils simulés..... 148
9.3.1. Chromopainter..... 148
9.3.1.1. Présentation de l'outil..... 148
9.3.1.2. Application sur population simulée..... 149
9.3.1.3. Ce qu'on cherche à voir..... 149
9.3.1.4. Résultats..... 149
9.3.2. Mesurer le respect de l'anonymat..... 153
10. Discussion..... 157
10.1. Mes contributions..... 157
10.2. Applications..... 158
10.2.1. Découvertes génomiques..... 158
10.2.2. Génération de nouveaux génomes..... 158
10.2.3. Brouillage de cohortes de patients..... 158
10.2.4. Jumeau virtuel..... 159
10.3. Génomique..... 160
10.3.1. Projet 1000 Genomes..... 160
10.3.1.1. Méthodologie..... 160
10.3.1.2. Manque de populations..... 160

10.3.1.3. Haplotypes imparfaits..... 161
10.3.2. Foyers de recombinaison..... 161
10.3.3. Effet mutagène des recombinaisons..... 162
10.4. Perspectives techniques..... 163
10.4.1. Augmenter l'étendue des échantillons..... 163
10.4.2. Super-étendeur basé sur les TAD..... 163
10.4.3. VAEs chevauchants..... 164
10.4.4. Espace latent "étiqueté"..... 164
10.4.5. Réseau de neurones récurrent pour les mutations rares..... 165
10.4.6. De chromosome à chromosome..... 165
10.4.6.1. Principe..... 165
10.4.6.2. Difficultés..... 166
10.4.6.3. Solution optimale..... 166
10.4.7. Spécialisation du modèle sur population spécifique..... 167
10.5. Interprétations..... 168
10.5.1. Comment bien mesurer la vraisemblance d'une population?..... 168
10.5.2. Distance entre réel et simulé..... 168
11. Implémentation..... 170
11.1. Outils utilisés..... 170
11.2. Implémentations spécifiques..... 171
11.3. Disponibilité du code..... 172
12. Références..... 173

2. Remerciements

Je voudrais remercier tout d'abord Daniel GAUTHERET et Denis PUTHIER d'avoir accepté de lire ce présent manuscrit. Leurs avis experts sur la bioinformatique, l'apprentissage automatique et l'analyse de données me tiennent particulièrement à cœur, et j'ai hâte de connaître leurs retours sur le projet.

Merci également à Thérèse COMMES, la présidente de mon jury, et Magali RICHARD d'avoir évalué la présentation orale de mon travail. Cette étape me tenait particulièrement à cœur, et je suis très heureux d'avoir pu la réaliser avec mon jury intégralement en personne pour une discussion très enrichissante.

Je remercie également mes directeurs de thèse, William RITCHIE et Alban MANCHERON, de m'avoir suivi tout au long de ce parcours. Ils m'ont apporté un accompagnement autant scientifique que personnel de haute qualité, et dont je me souviendrai toute ma vie.

Les membres de mon comité de suivi individuel, Grégory BEURIER, Annie CHATEAU, et Charles LECELLIER, m'ont également apporté des retours scientifiques précieux tout au long de ma thèse. Ils ont été vitaux dans le déroulement du projet, et pour cela je les remercie.

Je souhaite remercier la Ligue Contre le Cancer d'avoir fourni les fonds nécessaires à la réalisation du doctorat.

Un merci tout aussi important à la plateforme Génotoul d'avoir fourni les ressources de calcul et de stockage nécessaires aux modèles développés ici.

Je remercie chaleureusement mes collègues de l'Institut de Génétique Humaine, qu'ils soient compagnons doctorants ou non, pour tous les bons moments partagés ensemble. La thèse est un marathon, et je n'aurais pas pu le faire sans vous à mes côtés.

À mes amis du PhD Pub, nos événements mémorables m'ont rappelé chaque mois ce que représente vraiment la communauté scientifique. Un grand merci à vous.

Je n'oublie bien sûr pas mes collègues du MiWoKi, dont les réunions semi-régulières m'ont bien accompagnés au cours du doctorat. Merci à vous pour les discussions scientifiques, et les nombreuses autres aussi.

Merci à toi, Charlotte, pour tout le bonheur et le soutien que tu m'apportes.

3. Abréviations utilisées

- ACP : Analyse en Composantes Principales, méthode de réduction de dimensions simple à mettre en place et très répandue pour la visualisation de données complexes.
- ADN : Acide Désoxyribo-Nucléique, longue molécule contenant l'information biologique des organismes vivants.
- ARN : Acide Ribo-Nucléique, molécule biologique proche de l'ADN.
- CNV : *Copy Number Variation*, variation en nombre de copies, type de modification structurale de l'ADN
- dbSNP : *Single Nucleotide Polymorphism Database*, base de données de mutations ponctuelles chez l'Humain
- DKL : Divergence de Kullback-Leibler, métrique mesurant la différence entre deux distributions.
- GAN : *Generative Adversarial Network*, modèle de réseau de neurones servant à générer de nouveaux échantillons après un entraînement itératif en deux phases.
- GWAS : *Genome-Wide Association Study*, méthode d'investigation des liens entre mutations dans l'ADN et maladies / phénotypes.
- HGP : *Human Genome Project*, premier projet de séquençage du génome humain dans son intégralité.
- IA / AI : Intelligence Artificielle, famille de méthodes visant à créer des comportements considérés comme intelligents dans un automate.
- MNIST : *Modified National Institute of Standards and Technology*, base de données d'images de chiffres manuscrits très utilisées en IA.
- NCBI : *National Center for Biotechnology Information*, centre d'hébergement de plusieurs bases de données publiques en biologie moléculaire aux États-Unis.
- NIH : *National Institutes of Health*, regroupement d'instituts de recherche en biologie-santé aux États-Unis d'Amérique.
- NGS : *Next Generation Sequencing*, technologies de séquençage de seconde génération, apparues sur le marché entre 2005 et 2010.
- RBM : *Restricted Boltzmann Machine*, modèle de réseau de neurones à deux couches servant à générer de nouveaux échantillons.
- RFLP : *Restriction Fragment Length Polymorphism*, technique de biologie moléculaire permettant d'observer les similarités et différences entre les génomes de deux individus.
- SNP : *Single Nucleotide Polymorphism*, mutation dans l'ADN chez l'Humain présent chez au moins 1% de la population.
- TAD : *Topologically Associating Domain*, section d'un génome montrant des liens fonctionnels entre les éléments contenus à l'intérieur.
- VAE : Autoencodeur Variationnel (*Variational Autoencoder* en anglais), modèle de réseau de neurones permettant à la fois une réduction de dimensions et une génération de nouveaux échantillons.
- WGAN : *Generative Adversarial Network using Wasserstein loss*, architecture particulière de GAN, employant une mesure de l'erreur différente pour réduire certains artéfacts pendant l'entraînement.
- 1KG : *1000 Genomes Project*, projet de séquençage international chez l'Humain.

4. Résumé

Un génome humain disponible publiquement représente à la fois une ressource pour la recherche et un risque pour la personne qui a fourni le génome. De nombreuses personnes malveillantes pourraient l'exploiter pour en tirer des informations sur la santé du donneur ou de ses proches. Des travaux récents ont employé des modèles d'intelligence artificielle pour simuler des données génomiques, dans le but de créer des jeux de données synthétiques ayant un intérêt scientifique et qui respectent l'anonymat des patients. Cependant, une difficulté majeure se pose quant à la quantité de données que représente un génome humain dans son entièreté, et donc les ressources requises pour les traiter.

Nous avons mis au point une méthode de réduction de dimensions combinant intelligence artificielle et nos savoirs sur les mécanismes d'association des mutations in vivo. Ainsi, une grande quantité de données pourra être traitée sans nécessiter d'importantes ressources de calcul. Notre découpe du génome s'appuie sur les chances de recombinaison chromosomique, donc est proche des mécanismes de transmission des mutations. Les données sur lesquelles l'entraînement se fait sont issues du *1000 Genomes Project*. Il répertorie plus de 2500 génomes de divers groupes ethniques. Les autoencodeurs variationnels sont une méthode de réduction de dimensions utilisant des réseaux de neurones, servant d'extension au modèle génératif. Les Réseaux Antagonistes Génératifs avec la perte de Wasserstein (WGAN) sont une référence parmi les méthodes de génération sur de nombreux types de données.

Après avoir testé plusieurs stratégies et paramètres différents, la découpe selon les loci de recombinaison chromosomique présente la meilleure combinaison d'exactitude de reconstruction et d'efficacité de compression. Les autoencodeurs entraînés sur ces sections présentent en moyenne une exactitude de reconstruction supérieure à 99% en comprimant le génome haplotypique cent fois (une section de mille mutations sera résumée par dix valeurs).

Le WGAN permet de générer une population simulée vérifiant plusieurs critères essentiels. Cette population présente une bonne diversité, proche de celle trouvée dans le jeu de données de référence. Elle est vraisemblable, dans la mesure où les nouvelles combinaisons de mutations générées ne rompent pas les déséquilibres de liaison trouvées chez l'humain. Elle préserve aussi l'anonymat des donneurs en synthétisant des combinaisons de génomes de référence, qui ne sont pas des copies conformes de celles-ci.

5. Abstract

A publicly available human genome serves as both a valuable resource for researchers and a potential risk to the individual who provided the genome. Many actors with harmful intentions could exploit it to extract information about the donor's health or that of their relatives. Recent efforts have employed artificial intelligence models to simulate genomic data, aiming to create synthetic datasets with scientific merit while preserving patient anonymity. However, a major challenge arises in dealing with the vast amount of data that constitutes a complete human genome and the resources required to process it.

We have developed a dimension reduction method that combines artificial intelligence with our knowledge of in vivo mutation association mechanisms. This approach enables the processing of large amounts of data without significant computational resources. Our genome segmentation follows chromosomal recombination hotspots, closely resembling mutation transmission mechanisms. Training data is sourced from the 1000 Genomes Project, which catalogues over 2500 genomes from diverse ethnic groups. Variational autoencoders, a dimension reduction method utilising neural networks, serve as an extension to the generative model. Wasserstein Generative Adversarial Networks (WGAN) are a benchmark among generation methods for various data types.

After testing various strategies and parameters, segmentation based on chromosomal recombination loci proves to offer the best combination of reconstruction accuracy and compression efficiency. Autoencoders trained on these segments exhibit an average reconstruction accuracy exceeding 99% when compressing the haplotypic genome by a factor of one hundred (a section of a thousand mutations summarised by ten values).

WGAN allows the generation of a simulated population meeting several essential criteria. This population demonstrates good diversity, closely resembling that found in the reference dataset. It is plausible, as newly generated combinations of mutations do not disrupt the linkage disequilibria found in humans. It also preserves donor anonymity by synthesising combinations of reference genomes that are not exact replicas of the originals.

6. Introduction

6.1. Histoire de la génomique

6.1.1. Caractéristiques héréditaires

Depuis la préhistoire, les agriculteurs ont observé avec attention les caractéristiques physiques transmises à travers les générations de leur bétail et de leurs cultures. Cette observation méticuleuse a conduit à la découverte qu'il était possible de sélectionner délibérément les individus porteurs de traits spécifiques afin d'augmenter l'abondance de ces caractéristiques dans la descendance. L'histoire de la génétique trouve ses racines dans ces observations empiriques, mais elle a atteint un tournant décisif grâce aux travaux de Gregor Mendel au cours des années 1856 à 1863.

Gregor Mendel, moine augustinien et botaniste, a entrepris la première étude formelle sur l'hérédité en utilisant le pois comme organisme modèle. Ses expérimentations méticuleuses, consistant en des croisements entre différentes générations de pois (Mendel 1865), ont jeté les bases de la compréhension moderne de la génétique. Les résultats de ses recherches ont abouti à la formulation de cinq lois fondamentales de l'hérédité, également connues sous le nom de lois de Mendel :

- Les caractéristiques sont unitaires et discrètes.
- Ces caractéristiques ont des formes alternatives, chacune héritée de l’un des deux parents.
- Une forme est dominante sur les autres, qui sera visible sur l’individu.
- Les cellules reproductrices sont créées aléatoirement. Des individus contenant les deux formes d’une caractéristique produisent des cellules portant une de ces deux formes à fréquence égale.
- Les caractéristiques se combinent de manière indépendante.

De nos jours, ces unités discrètes d'hérédité sont connues sous le nom de gènes, tandis que leurs différentes versions sont appelées allèles. Cependant, malgré la clarté et la puissance explicative des lois de Mendel, il faut reconnaître les limites inhérentes à ces postulats.

Mendel lui-même, lors de la formulation de ses lois, a exprimé des réserves quant à leur

universalité. Il a averti que l'application de ces principes à tous les organismes et à toutes les caractéristiques pourrait ne pas être justifiée. Les connaissances modernes corroborent ces avertissements en démontrant que ces lois ont des limites dans leur application universelle.

Une autre limite, non anticipée par Mendel mais éclairée par les progrès ultérieurs de la génétique, réside dans la notion d'indépendance entre les caractéristiques héréditaires. Bien que de nombreuses caractéristiques puissent être considérées comme non liées, la recherche du siècle suivant les travaux de Mendel a révélé la présence d'une structure complexe qui porte les gènes, les chromosomes, et qui lie certaines caractéristiques en raison de leur proximité physique.

Ainsi, cette exploration des lois de l'hérédité initiée par Mendel constitue le point de départ de notre compréhension moderne de la génétique. Près de 200 ans plus tard, nous maîtrisons bien mieux les nuances dans les mécanismes de transmission des traits héréditaires, mais toutes ne sont pas encore élucidées.

6.1.2. Chromosomes en tant que support de l’information biologique

Au tournant du XXe siècle, une ère nouvelle s'ouvre dans le domaine de la génétique, avec les travaux révolutionnaires de Sutton et Boveri qui constituent la prochaine étape vers notre compréhension moderne de l'information héréditaire. De manière indépendante, Sutton (Sutton 1902) et Boveri (Boveri 1904) ont proposé que les chromosomes portent l'information héréditaire, conformément aux principes décrits par Mendel. Ces structures distinctes et observables pendant les divisions cellulaires ont été identifiées comme des vecteurs cruciaux de l'hérédité, se transmettant de génération en génération de cellules.

Les premières preuves tangibles de cette proposition ont émergé des expériences menées par Boveri. En étudiant les embryons d'oursins, Boveri a observé que tous les chromosomes étaient nécessaires pour le développement correct de ces organismes. Cette observation a apporté une validation concrète à l'idée que les chromosomes jouent un rôle de support de l’information biologique. De manière similaire, Sutton a réalisé des expériences sur des sauterelles, identifiant des paires de chromosomes provenant de parents différents, renforçant ainsi l'hypothèse de l'importance des chromosomes dans l'hérédité.

Bien que ces propositions aient suscité au départ des controverses, les années suivantes ont apporté des avancées significatives. Les contributions de chercheurs éminents comme Nettie

Stevens sur le chromosome sexuel (Stevens 1905), Thomas Hunt Morgan sur les liaisons entre gènes (Morgan 1915), et Eleanor Carothers sur l'assortiment indépendant des chromosomes (Carothers 1913) ont enrichi la compréhension collective. Morgan, en particulier, a formulé la théorie des recombinaisons génétiques, phénomène où deux gènes liés peuvent se transmettre indépendamment, en changeant physiquement de chromosome. Par la suite, ce processus a été démontré comme ayant lieu pendant la méiose, la série de divisions cellulaires produisant les gamètes, par Barbara McClintock et Harriet Creighton (Creighton and McClintock 1931).

En 1919, Phoebus Levene (Levene 1919) a réalisé une percée majeure en montrant que la molécule portant l'information génétique était composée de sucre, de phosphate et de bases azotées. Il a déterminé l'enchaînement précis de ces composantes, les nommant nucléotides. Cependant, malgré cette découverte fondamentale, l'organisation détaillée de ces nucléotides dans la molécule d'acide nucléique restait floue.

Ces révélations ont suscité la curiosité intense de la communauté scientifique. Soutenus par les progrès simultanés en physique et en chimie au cours de ces décennies, de nombreuses équipes de recherche ont entrepris de raffiner le modèle du support de l'information biologique. Elles ont cherché à comprendre comment les gènes étaient disposés sur les chromosomes, comment l'information génétique était transmise, et quel rôle précis jouaient les nucléotides dans cette transmission.

Ainsi, cette période charnière du début du XXe siècle a marqué une explosion de découvertes et d'innovations dans le domaine de la génétique, ouvrant la voie à une compréhension plus approfondie de la manière dont l'information héréditaire est encodée, transmise et exprimée dans le monde biologique. Ces recherches pionnières ont jeté les bases d'une discipline en constante évolution, avec des implications majeures pour la biologie moléculaire, la médecine et la compréhension globale de la vie elle-même.

6.1.3. Séquence de nucléotides

La résolution de la structure exacte de la molécule portant l'information génétique a marqué un tournant majeur dans l'histoire de la génétique, un accomplissement rendu possible grâce à la cristallographie et à l'analyse par diffraction de rayons X. La molécule en question s'est

révélée être une double hélice. Ce succès scientifique a été le résultat d'un travail collaboratif mené par plusieurs équipes de recherche à travers le monde, mais l'histoire retient principalement les noms de Francis Crick, Rosalind Franklin et James Watson.

L'équipe dirigée par Rosalind Franklin a joué un rôle crucial dans cette entreprise, en utilisant la cristallographie des rayons X pour obtenir des images détaillées de la structure de cette molécule, l'Acide DésoxyriboNucléique (ADN) (Franklin and Gosling 1953). Cependant, cette contribution significative de Franklin a été éclipsée par une controverse ultérieure. En 1952, James Watson et Francis Crick ont utilisé des données provenant du laboratoire de Franklin sans son consentement pour construire leur propre modèle de la structure en double hélice de l'ADN. L'utilisation non autorisée de ces données a suscité des débats éthiques et a jeté une ombre sur la reconnaissance de la contribution de Franklin à cette découverte majeure.

En 1953, Watson et Crick publient leur modèle en double hélice de l'ADN dans la revue Nature (Watson and Crick 1953), un article qui deviendra l'un des plus célèbres dans l'histoire de la biologie. Cependant, l'article original ne fait aucune mention de la contribution de Rosalind Franklin. Cette omission volontaire a contribué à renforcer le récit centré sur Watson et Crick, occultant le rôle crucial de Franklin dans la résolution de la structure de l'ADN (Maddox 2003).

Francis Crick a ensuite énoncé ce qu'il appelait le "Dogme Central de la Biologie Moléculaire" (Crick 1958), qui postule que l'ADN sert de modèle de base pour la transcription de l'Acide RiboNucléique (ARN), qui à son tour peut être directement fonctionnel ou servir de matrice pour la traduction, processus par lequel les protéines sont synthétisées. En 1961, Nirenberg et Matthaei ont apporté une contribution majeure en démontrant que l'information génétique permettant la synthèse des protéines est encodée en triplets de nucléotides, appelés codons (Nirenberg and Matthaei 1961). Les années suivantes ont été marquées par des efforts soutenus pour déterminer les correspondances entre la majorité des codons et leurs acides aminés correspondants.

Cependant, malgré ces avancées, les méthodes utilisées pour analyser la séquence exacte de l'ADN étaient lentes et coûteuses. La recherche de solutions plus efficaces pour décoder l'information génétique a stimulé des innovations dans les années qui ont suivi, ouvrant la voie à des technologies plus rapides et plus rentables pour analyser les séquences d'ADN, un domaine en constante évolution qui continue d'influencer la génétique contemporaine.

6.1.4. Séquençage de cette information

Dans les années 1970, l'émergence des premiers projets de séquençage a constitué un jalon majeur dans l'histoire de la biologie moléculaire. Ces initiatives ont donné naissance à des avancées significatives dans la compréhension de la structure génétique, d'abord avec le séquençage de l'ARN suivi de celui de l'ADN. Ray Wu a été un pionnier en utilisant de courtes séquences d'ADN comme point de départ, initiant la polymérase (complexe biologique servant à répliquer l'ADN dans les cellules) à synthétiser de nouveaux fragments longs avec des nucléotides modifiés et marqués pour identifier précisément la séquence (R. Wu 1970). C'est sur ces bases que Frederick Sanger a développé sa méthode de séquençage en 1977 (Sanger, Nicklen, and Coulson 1977), qui s'est avérée être la première à être popularisée à l'échelle mondiale et utilisée par de nombreuses équipes de recherche.

Simultanément, Allan Maxam et Walter Gilbert ont élaboré une méthode alternative de séquençage basée sur la structure chimique de l'ADN (Maxam and Gilbert 1977). Leur approche, découpant la molécule à différents loci pour identifier les fragments ainsi générés, a aussi été employée, mais était plus complexe à mettre en œuvre par rapport à la méthode de Sanger.

Les années 1980 ont été marquées par la commercialisation des machines basées sur la méthode de séquençage de Sanger, facilitant le processus tant sur le plan de la semi-automatisation que de la pleine automatisation. Cette avancée a ouvert la voie à de nombreux projets de séquençage à grande échelle, adoptant une approche "shotgun" qui consiste à fragmenter aléatoirement une longue molécule d'ADN, séquencer les différents fragments, puis réassembler ces fragments en "contigs" pour reconstruire la séquence d'origine complète. Un exemple remarquable est la publication en 1995 du génome complet de la bactérie Haemophilus influenzae par Craig Venter, Hamilton Smith et leurs collègues (Smith et al. 1995).

Cette méthodologie a également été appliquée au projet emblématique du séquençage du génome humain, le *Human Genome Project* (HGP) (Barnhart 1989). Lancé comme une collaboration internationale entre de nombreux instituts, le HGP avait pour objectif de produire le génome de référence complet de l'Humain. Un génome de référence, assemblage de multiples génomes d'une espèce, vise à établir un consensus pour chaque locus du génome de l'espèce.

Cette ressource sert de base pour faciliter les séquençages ultérieurs d'individus de la même espèce, exprimant les génomes individuels comme ensemble de variations par rapport à la référence. Cependant, il est crucial de noter que le génome de référence n'a pas pour but d'être fonctionnel ; il représente plutôt un outil statistique pour l'étude de la diversité génétique au sein d'une population.

Les premières versions du génome humain de référence ont été publiées en 2001 (Lander et al. 2001), marquant une étape importante dans la compréhension de notre propre code génétique. Le projet a été déclaré "essentiellement complet" en 2003 (Little et al. 2003) avec la publication d'une carte couvrant plus de 90% du génome humain. Au fil des années, de nombreuses versions améliorées de ce génome de référence ont été publiées à mesure que nos connaissances se sont approfondies.

Cependant, ce n'est qu'en 2022 que le premier génome réellement complet, appelé "T2T" (de télomère à télomère) (Nurk et al. 2022), a été publié. Cette réalisation exceptionnelle a permis de capturer l'intégralité des nucléotides composant le génome d'une personne, y compris les régions répétées et difficilement accessibles. Cette avancée marque une nouvelle ère dans la séquençage génomique, offrant une vision plus complète et précise de la complexité de l'information génétique humaine.

6.1.5. Modèle actuel

L’ADN (Acide DésoxyriboNucléique) est le support de l’information biologique. Il est constitué de quatre nucléotides différents : l’adénine, la cytosine, la guanine et la thymine, en général abrégées A, C, G et T respectivement (Leacock 2021). Ces nucléotides s’enchaînent pour créer un long brin, qui s’associe avec un second pour former une double hélice. Chaque nucléotide du premier brin s’associe avec un partenaire spécifique sur le brin opposé : l’adénine avec la thymine, et la cytosine avec la guanine. Ces associations sont dues aux structures chimiques des nucléotides. L’adénine et la thymine possèdent chacune deux sites pouvant créer des liaisons hydrogène, et la cytosine et la guanine en possèdent trois chacune. Ces paires s’associent donc par affinité chimique. On classe également ces nucléotides en deux groupes selon un autre critère de leur structure chimique : les purines et les pyrimidines. Les purines sont composées de deux cycles, alors que les pyrimidines en possèdent une seule. Il devient donc

logique de voir apparaître seulement ces appariements de nucléotides (Guéron and Leroy 1995), indiqués sur la figure 1. Tout autre appariement provoquerait soit un déséquilibre chimique au niveau des liaisons hydrogène, soit un problème de taille de la paire de nucléotides provoquant des variations dans l'épaisseur du double brin d'ADN (ou alors les deux en même temps, lors d'une association A-G ou C-T). Ces mésappariements peuvent se produire, car les processus biologiques de maintien et de lecture de l'ADN ne sont pas parfaits, et seront soit corrigés, soit peuvent donner lieu à des changements lors de la synthèse de nouvelles séquences (G.-M. Li 2008). Ce phénomène arrive le plus souvent autour de régions de faible complexité, c'est-à-dire avec beaucoup de répétitions.

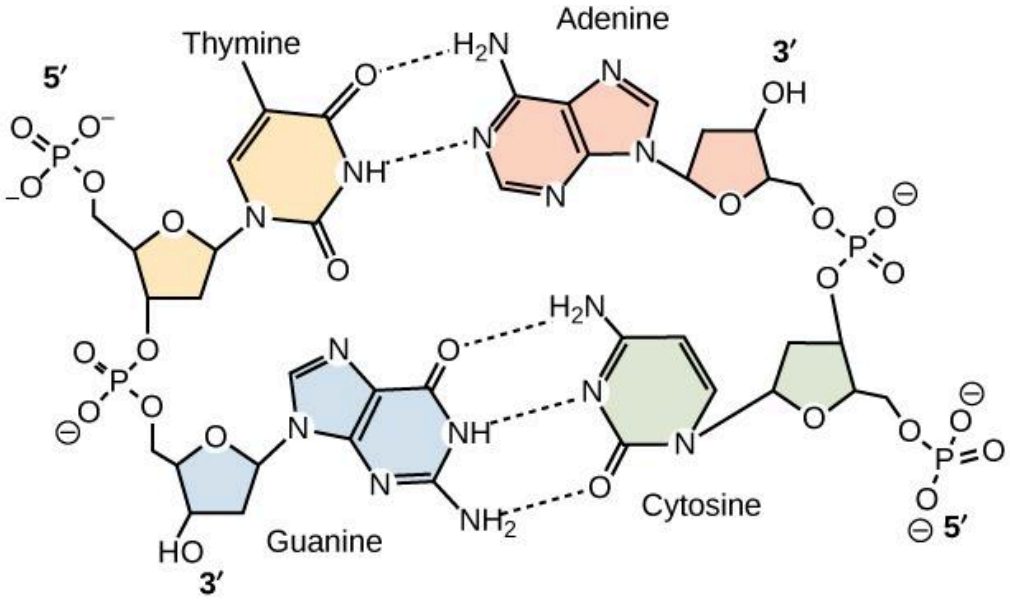


*Figure 1 : Structure chimique des nucléotides de l'ADN. Les pointillés tracés entre nucléotides représentent les liaisons hydrogènes qui maintiennent la structure de double hélice. Les notations **3'** et **5'** dénotent l'orientation des molécules, d'un point de vue chimique.*

On appelle un enchaînement de nucléotides donné une séquence, et l'enchaînement qui s'associe avec pour former un double brin est appelée reverse complément (brin complémentaire lu dans le sens inverse). L'unité de mesure d'une séquence d'ADN et la paire de base, ou bp pour base pair en anglais.

Cet ADN peut être lu par un complexe de protéines appelé ARN polymérase, qui synthétise un brin d'Acide RiboNucléique comportant la séquence complémentaire du brin qu'il

lit (Crick 1958). On appelle ce processus la transcription. L'ARN possède un sucre différent dans sa composition chimique, rendant le brin synthétisé plus fin et plus flexible, ce qui (avec sa petite longueur en bases azotées) permet son exportation en dehors du noyau des cellules, pour une grande partie des molécules synthétisées. Les molécules d'ARN ne contiennent pas de thymine, mais de l'uracile à la place, dont la structure chimique est très proche. Elles peuvent être directement fonctionnelles (par exemple les ARNs de transfert) ou servir de base pour la synthèse de protéines. Les ARNs ayant une fonction directement dans la cellule souvent se replient sur elle-même pour former des structures complexes, se reposant sur les affinités entre nucléotides pour former des tiges et des boucles (Svoboda and Di Cara 2006), illustrées dans la figure 2.

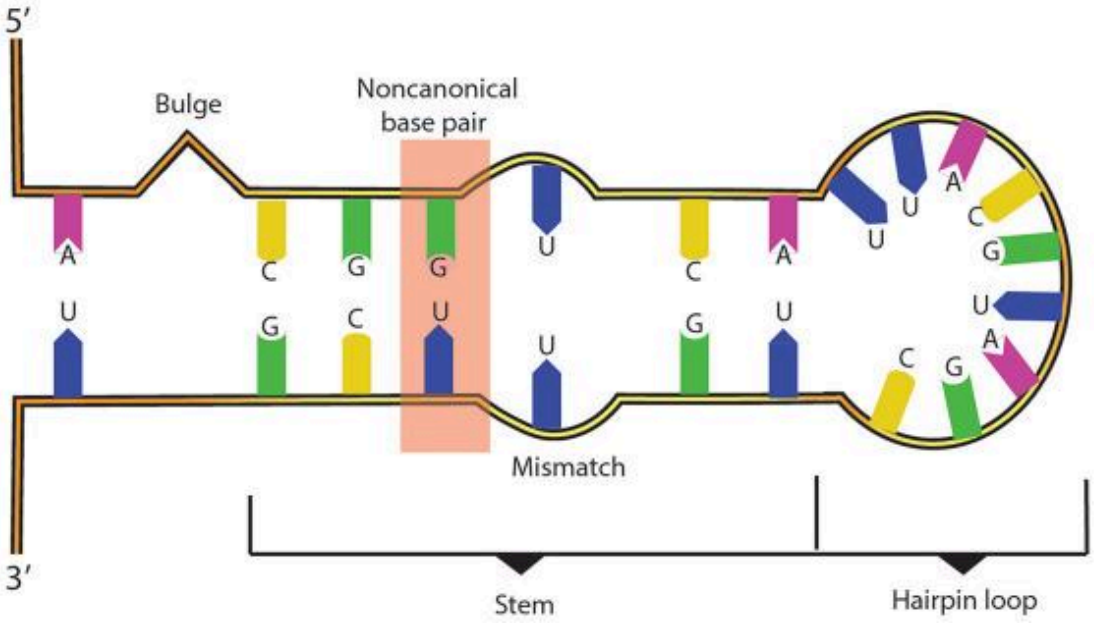


Figure 2 : Structure secondaire d'une molécule d'ARN. Ces molécules forment spontanément des tiges (stem) et boucles (loop), pouvant tolérer quelques mésappariements (mismatch) et gonflements (bulge). Certaines associations de nucléotides non-canoniques sont également supportables chez l'ARN, comme indiqué en rouge.

Si la molécule d'ARN est destinée à la synthèse de protéines, alors elle sera transportée vers un organite dédié (chez les eucaryotes) où la séquence de l'ARN sera lue par un ribosome, pendant un processus appelé traduction (Ramakrishnan 2002). L'enchaînement de triplets de nucléotides, appelés codons, définit la séquence d'acides aminés qui composeront la protéine

synthétisée. Ces acides aminés sont au nombre de 20 chez la plupart des êtres vivants, et, par leurs propriétés physico-chimiques, peuvent s'orienter et s'associer pour former des structures complexes dans l'espace. Ce sont ces structures qui définissent leurs rôles dans l'être vivant, rôle qui peut être structural ou fonctionnel.

Toute modification dans la composition en acides aminés impacte leur structure finale et donc leur fonction dans la cellule, de même pour les ARNs fonctionnels. Comme ces séquences protéiques sont définies par la séquence d'ADN ayant servi à les créer, toute étude de la variation entre individus d'une même espèce se fonde donc sur l'analyse de son génome, c'est-à-dire l'ensemble du matériel génétique de cet individu. Certaines variations peuvent trouver leurs sources dans d'autres éléments, comme l'influence de l'environnement, mais la première étape reste en général l'étude du génome pour déterminer la source de la variation.

Chez l'Humain, comme pour beaucoup d'autres animaux, le génome est organisé en paires de chromosomes (Leacock 2021), chaque membre de ces paires provenant d'un parent différent. Notre génome est donc dit "diploïde" (là où certaines plantes peuvent posséder des sextuplets de chromosomes et sont appelées hexaploïdes, par exemple), et nous transmettons une seule moitié de ces paires de chromosomes à notre descendance. On appelle haplotype (combinaison de "haploïde", état d'une cellule contenant un seul jeu de chromosomes, et "génotype", la composition génétique d'un organisme) le groupe de gènes et leurs variations héritées ensemble d'un seul parent.

L'état des chromosomes dépend de la phase du cycle cellulaire dans laquelle la cellule qui les contient se trouve (Leacock 2021). La phase G1, juste après la division cellulaire, est une phase de croissance où la cellule se fournit en nutriments et organites, et les chromosomes sont composés d'un seul chromatide, c'est-à-dire d'un seul exemplaire de la molécule d'ADN. Ensuite vient la phase S où se déroule la réplication : les deux brins de la molécule d'ADN de chaque chromosome sont séparées et copiées par des complexes de protéines pour créer deux molécules. Les chromosomes sont composés de deux chromatides à la fin de cette réplication. La phase G2 est une nouvelle phase de croissance, cette fois spécialisée pour préparer la mitose, la division cellulaire permettant de créer deux cellules filles identiques à la cellule mère. Enfin, dans la phase M, la cellule se divise, et le cycle reprend en phase G1.

La méiose est un processus de division cellulaire spécial qui permet de créer les gamètes. Elle s'effectue en deux grandes étapes, appelées méiose I et méiose II (Leacock 2021). Dans chacune de ces étapes, plusieurs phases se déroulent, chacune ayant un effet distinct sur l'ADN de la cellule : prophase, métaphase, anaphase et télophase.

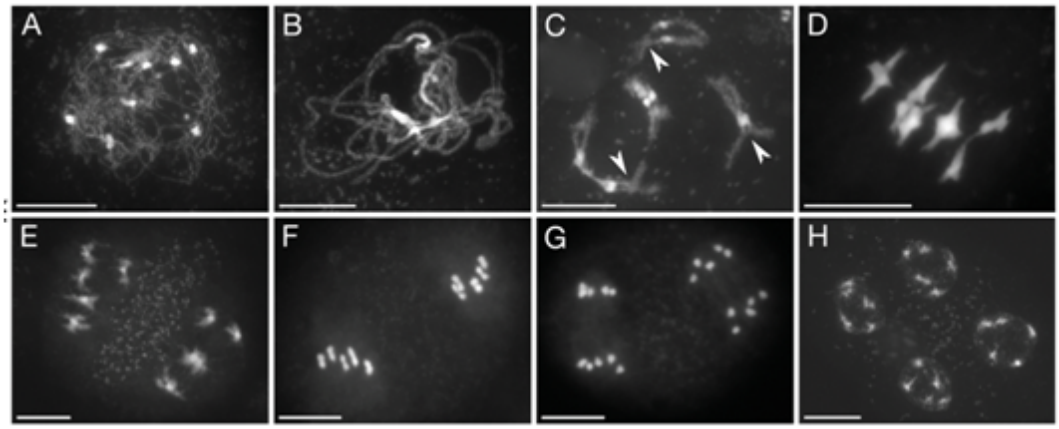


Figure 3 : Étapes de la méiose dans une cellule d'Arabidopsis thaliana vue au microscope, avec teinture de l'ADN. Les étapes représentées sont : prophase I (A-C), métaphase I (D), télophase I (E), métaphase II (F), anaphase II (G) et télophase II (H). Tiré de (Leacock 2021)

Dans la prophase I, l'ADN est organisé sous forme de chromosomes distincts et chaque paire de chromosome s'associe (indiquées par des flèches dans la figure 3). Dans la métaphase I, elles s'alignent au milieu de la cellule puis se séparent pendant l'anaphase I. La télophase I représente l'établissement de deux nouvelles cellules par cloisonnement. Ces étapes mènent à la création de deux cellules haploïdes, avec des chromosomes à deux chromatides.

Pour la méiose II, ces étapes se répètent, mais ce sont les chromatides sœurs de chaque chromosome qui se séparent. Le résultat est donc quatre cellules haploïdes, portant des chromosomes à une seule chromatide.

La prophase I est de loin la phase la plus longue de ce processus. C'est pendant cette étape que de nombreuses cassures double brin viennent mettre en place des chiasmata (pluriel de chiasma, point de contact entre les chromatides de deux chromosomes) entre chromosomes correspondants. Chez l'humain, on observe généralement deux chiasmata par paire de chromosomes (un de chaque côté du centromère), mais le nombre n'est pas complètement fixe

(Hassold et al. 2004). Lors de la réparation de ces cassures, de nombreuses recombinaisons chromosomiques et nouvelles mutations sont acquises de manière permanente. Cette étape est une source extrêmement importante de diversité dans les populations, car elle permet un brassage des chromosomes inégalable dans la nature.

Il existe d'autres sources de mutations pour une cellule, telle que les rayonnements ultraviolets (Bernstein et al. 2002) ou des intercalants de l'ADN (Richards and Rodger 2007), qui peuvent perturber les mécanismes de réparation de l'ADN du noyau, et ainsi causer de nouvelles mutations. Ces événements sont cependant en général limités aux cellules somatiques, et seront donc rarement transmises à la descendance.

Ces variations génomiques peuvent être de natures différentes (Leacock 2021). La plus simple, et la plus commune, est une substitution d'un nucléotide par un autre, qui peut modifier un acide aminé dans la séquence de protéine finale selon le codon qu'elle affecte, dû au phénomène de redondance du code génétique. Selon le site exact de cette substitution, elle peut avoir de grandes conséquences ou absolument aucune. Un nucléotide peut aussi se retrouver ajouté ou retiré à une séquence d'ADN, événement appelé insertion/délétion ou indel. Cette forme de variation est beaucoup plus dangereuse si elle se trouve dans une région qui code pour une protéine, vu qu'elle va ainsi décaler toute la séquence de codons et complètement modifier la séquence d'acides aminés qui en résultera. Un indel peut aussi couvrir plusieurs nucléotides, et la communauté scientifique n'est pas unie dans sa définition de ce qui constitue un court indel ou une modification structurale. Ces mutations peuvent survenir n'importe où dans le génome, mais apparaissent plus souvent dans les régions actives du génome (comme les gènes, qui sont lues fréquemment par l'ARN polymérase).

Les modifications structurales sont des variations génomiques couvrant en général plusieurs dizaines de paires de bases, mais peuvent s'étendre sur plusieurs milliers également. Les insertions et délétions de grandes portions de séquences en constituent une première catégorie. On trouve également des variations de nombre de copies (*Copy Number Variation* ou CNV en anglais), où certaines sections du génome peuvent se dupliquer une ou plusieurs fois. Si cette section contient des éléments codants ou régulatoires, ces événements peuvent avoir de sérieux impacts sur le métabolisme du porteur. Ces CNVs sont cependant différents des éléments répétés (Richard, Kerrest, and Dujon 2008), qui sont de courtes séquences d'ADN (<10

nucléotides : microsatellite, 10-60 nucléotides : minisatellite, >60 : macrosatellite) qui se répètent un grand nombre de fois à la suite dans le génome. Les dernières grandes catégories de variations structurales du génome sont la translocation et l'inversion, qui peuvent se produire en même temps. Une translocation indique un déplacement d'une grande section de la molécule d'ADN, parfois observable à l'échelle du chromosome, et l'inversion indique qu'une section d'ADN est retournée.

On considère qu'une mutation ponctuelle constitue un *Single Nucleotide Polymorphism* (SNP) lorsqu'elle est présente chez au moins 1% des individus dans une population donnée (Keats and Sherman 2013). Certaines sources considèrent uniquement les substitutions ponctuelles comme pouvant qualifier de SNP, alors que d'autres sont plus tolérantes envers d'autres sources de variation génomique.

À l'échelle du chromosome, pour mesurer la distance effective entre deux éléments génétiques (gènes ou variations), on emploie une méthode basée sur l'observation. Le déséquilibre de liaison décrit l'association de ces deux éléments dans une population donnée lorsque cette fréquence d'association est plus forte ou plus faible que la fréquence attendue si ces éléments étaient indépendants. On le mesure en centiMorgan, un centiMorgan représentant 0.01 recombinaisons en moyenne entre deux éléments génétiques à chaque génération (Keats and Sherman 2013) (mesuré sur les gamètes produits par un ou plusieurs individus).

6.1.6. Génomique moderne

Il existe aujourd'hui de nombreuses méthodes pour analyser la séquence d'ADN exacte présente dans le génome d'un organisme, ou de son génome entier. La méthode de Sanger fut la première commercialisée, mais son débit assez lent par rapport à la taille des gènes et génomes de l'Humain ainsi que bon nombre d'autres espèces d'intérêt ont incité des recherches pour trouver des méthodes plus efficaces.

On appelle *Next Generation Sequencing* ou NGS les technologies ayant vu le jour autour de 2005, qui mettent en œuvre une approche de séquençage dite "massivement parallèle" (Mardis 2008). Là où la méthode de Sanger permet de séquencer un fragment d'ADN de quelques centaines de paires de bases avec une grande précision, ces méthodes NGS génèrent énormément

de courtes lectures de séquençage (en général autour de 100 paires de bases) avec un taux d'erreur légèrement plus grand, permettant de couvrir au total de grands fragments d'ADN. On utilise notamment la profondeur de séquençage, qui est le nombre moyen de lectures s'alignant sur une région du génome, pour décrire sa qualité et son niveau de détection de variations dans le génome.

Cette stratégie de séquençage repose sur la disponibilité d'un génome de référence pour l'espèce séquencée pour permettre d'aligner les courtes lectures de séquençage dessus. Il existe des méthodes d'assemblage de génome de novo, mais elles sont beaucoup plus longues à exécuter (Baker 2012).

Ces méthodologies ont permis une grande accélération du rythme de séquençage qu'il est possible de mettre en œuvre. Là où le *Human Genome Project* a mis plusieurs années à établir une première carte du génome humain, en 2008 on estimait pouvoir séquencer un génome humain par jour et par machine (Mardis 2008) (avec une profondeur de séquençage suffisante pour détecter les variations de l'individu séquencé). Le site officiel de la compagnie Illumina, un fabricant de séquenceurs réputé dans le monde, prétend pouvoir séquencer 48 génomes en deux jours sur une machine (Illumina 2017), un chiffre faramineux lorsqu'on se rappelle que cette technologie a vu le jour il y a à peine 50 ans.

Les techniques de séquençage massivement parallèle requièrent une étape d'amplification de la molécule d'ADN pour pouvoir générer la multitude de courtes lectures de séquençage. Depuis une dizaine d'années, des techniques de séquençage sur molécule unique ont vu le jour, également appelées long read pour leur capacité à séquencer une longue molécule en une seule fois. Deux approches principales existent pour cette démarche, commercialisées par Pacific Biosciences et Oxford Nanopore respectivement.

La technologie de PacBio rend un long fragment d'ADN, jusqu'à 25 000 paires de bases, circulaire par un processus biochimique et l'immobilise dans un puits (Eid et al. 2009). La même approche utilisant des nucléotides marqués est employée pour transmettre la séquence de nucléotides à un ordinateur. Vu que la molécule d'ADN est circulaire et seule dans son puits de réaction, elle sera en réalité séquencée de nombreuses fois d'affilée, ce qui permet d'obtenir un taux d'erreur très faible au final.

Oxford Nanopore emploie une approche très différente. La longue molécule d'ADN (ou

d’ARN, depuis peu) passe à travers un pore dans une membrane, ce qui provoque un changement de champ électrique (Branton et al. 2008; Jain et al. 2016). Le pore de la membrane est suffisamment sensible aux objets le traversant que ces changements dépendent de la nature du nucléotide qui passe à travers la membrane. Il est même possible de détecter certaines modifications de ces nucléotides en parallèle (Liu Xu and Seki 2020), comme la méthylation de l’ADN qui intervient dans la régulation de la transcription. Au lancement de cette technologie, les taux d’erreur de la méthode étaient élevés, et elle n’a donc pas connu une adoption rapide. Aujourd’hui, les taux d’erreur sont proches de ce qui est possible avec les technologies NGS, et elles gagnent en popularité d’année en année (Sahlin and Medvedev 2021). Un avantage notable à cette technologie est sa portabilité : là où un séquenceur utilisant des bases marquées fait plusieurs dizaines de centimètres de chaque côté, un séquenceur Nanopore peut faire moins de dix centimètres sur son plus grand côté, et être directement branché sur le port USB d’un ordinateur portable. Ceci permet le déploiement de cette technologie dans des conditions bien plus variées, comme des recherches sur le terrain en milieu tropical.

Les avancées majeures dans le séquençage génomique ont non seulement révolutionné notre compréhension du génome humain, mais ont également ouvert de nouvelles perspectives dans le domaine médical. L'identification des variants génétiques provoquant de nombreuses maladies héritables a permis des progrès significatifs dans le diagnostic, la prévention et le traitement des affections génétiques (Apgar and Sanders 2022).

Certaines maladies héritables résultent de mutations dans un seul gène, phénomène connu sous le nom de monogénisme. Un exemple emblématique est l'anémie falciforme, une maladie génétique du sang causée par une mutation spécifique dans le gène de l'hémoglobine. Ces découvertes ont conduit à la possibilité, ces dernières années, d'utiliser des techniques d'édition génique pour modifier des gènes défectueux chez les patients atteints de ces maladies (Wong 2023). Cependant, il est important de souligner que cette pratique n'est applicable que dans certains cas spécifiques, et son utilisation est soumise à des régulations éthiques rigoureuses.

D'autre part, de nombreuses maladies sont polygéniques, ce qui signifie qu'elles dépendent de variations dans plusieurs loci du génome pour se manifester (Visscher et al. 2021). Parfois, ces affections complexes impliquent également des facteurs environnementaux

(Virolainen et al. 2023). Pour comprendre ces maladies et établir des liens avec leurs sources, les scientifiques réalisent généralement des études d'association à l'échelle du génome, connues sous le nom de *Genome Wide Association Studies* (GWAS) (Hirschhorn and Daly 2005). Ces études mobilisent de grandes cohortes, souvent constituées de plusieurs milliers de personnes, afin de mettre en évidence des liens statistiques entre la présence ou la sévérité d'une maladie et les variations génomiques. Les résultats de ces GWAS permettent d'identifier des marqueurs génétiques associés à des maladies spécifiques (Simón-Sánchez et al. 2009), facilitant ainsi la compréhension des mécanismes sous-jacents.

Une fois ces liens établis, il devient crucial de sauvegarder et partager ces informations avec la communauté scientifique. Les bases de données génomiques, telles que le projet gnomAD (Genome Aggregation Database) (Karczewski et al. 2020), stockent des données provenant de milliers de génomes humains et sont des ressources précieuses pour les chercheurs et les cliniciens. Ces bases de données contribuent à l'expansion des connaissances médicales en facilitant l'accès à des informations génomiques variées, permettant une meilleure compréhension des relations entre les variations génétiques et les maladies.

Les avancées récentes dans le domaine de la génétique et de la génomique médicale ont considérablement influencé notre capacité à comprendre, diagnostiquer et traiter les maladies génétiques. Les pratiques de séquençage et les projets expérimentaux contribuent à élargir nos connaissances et à ouvrir la voie à une médecine plus personnalisée, basée sur une compréhension approfondie des particularités génétiques de chaque individu.

6.2. Bases de données biologiques

6.2.1. Bases de données existantes

6.2.1.1. Single Nucleotide Polymorphism Database

Depuis son lancement en 1998, la *Single Nucleotide Polymorphism Database* (dbSNP) (Stephen T. Sherry, Ward, and Sirotkin 1999; S. T. Sherry et al. 2001) a évolué pour devenir la principale ressource du *National Center for Biotechnology Information* (NCBI) dédiée à l'archivage des mutations génétiques. Initialement conçu pour recueillir exclusivement les mutations ponctuelles, dbSNP a progressivement élargi sa portée pour inclure une diversité de variants génétiques, tels que les insertions/délétions et les microsatellites. Cette expansion a été motivée par la reconnaissance de l'importance de ces différentes formes de variation génétique dans la compréhension des phénotypes complexes et des maladies génétiques.

Jusqu'en 2017, dbSNP acceptait des soumissions pour toutes les espèces, reflétant ainsi la diversité du monde vivant. Cependant, afin de se concentrer sur l'exploration plus approfondie du génome humain et de fournir des données de haute qualité pour la recherche biomédicale, la base de données a récemment restreint sa portée pour se limiter exclusivement à l'espèce humaine.

L'intégration des données dans dbSNP est largement tributaire des contributions des utilisateurs, qu'il s'agisse de chercheurs académiques, d'institutions cliniques ou de laboratoires privés. Ces contributions ont alimenté une base de données qui abrite actuellement une impressionnante quantité de données, dépassant les 2 milliards d'entrées. Cette richesse de données génétiques représente une mine d'information pour un large éventail de projets de recherche en génétique humaine, de l'étude de la variation génétique dans des populations spécifiques à l'identification de variants associés à des maladies complexes.

En fournissant une plateforme acceptant des soumissions de toutes parts et un accès libre aux données à tous les chercheurs dans le monde entier, dbSNP a servi de base à de nombreux projets de recherche en génétique et en génomique. Grâce à cette base de données, ainsi que des méthodologies d’accès facilité à ses données (Kasprzyk et al. 2004; Yuan et al. 2006), beaucoup de mutations ponctuelles ayant un impact direct sur la santé de leurs porteurs ont pu être identifiés (Parker and Borchers 2014; Chin et al. 2008).

Plus récemment, le NCBI a également mis en place ClinVar (Landrum et al. 2014), une base de données recensant spécifiquement ces mutations à effet pathologique confirmé ou suspecté. À son lancement, elle a été peuplée automatiquement en collectant les variants contenus dans dbSNP (et quelques autres bases de données) ayant des informations cliniques. Son établissement et sa popularité font de ClinVar un élément essentiel dans les études de l'impact de variants génétiques sur la santé humaine.

6.2.1.2. HapMap

HapMap représente depuis 2002 une initiative de séquençage d'envergure internationale (Altshuler, Donnelly, and The International HapMap Consortium 2005; Gibbs et al. 2003), dont l'objectif premier réside dans l'approfondissement de notre compréhension des liens entre les combinaisons de mutations génétiques et les maladies courantes, souvent multifactorielles. Initié par une collaboration entre le Canada, la Chine, le Japon, le Nigéria, le Royaume-Uni et les États-Unis, ce projet a réuni ces nations dans un effort conjoint de séquençage au cours des deux premières phases. Par la suite, la troisième phase a élargi cette collaboration pour inclure également le Kenya et l'Italie, enrichissant ainsi la diversité des populations étudiées.

Au cœur de l'initiative HapMap se trouve une exploration minutieuse des associations génétiques en se concentrant spécifiquement sur les mutations portées par un seul haplotype individuel. Ce choix méthodologique vise à élucider les haplotypes les plus fréquents au sein de ces différentes populations, offrant ainsi une perspective unique sur les variations génétiques présentes à l'échelle mondiale.

Cette approche a permis à HapMap de devenir une ressource de premier plan dans le domaine de la génétique humaine, catalysant bon nombre de projets de recherche visant à identifier les variants génétiques associés à des maladies complexes (Hirschhorn and Daly 2005; Simón-Sánchez et al. 2009; Klein et al. 2005). En facilitant l'identification des haplotypes les plus fréquents dans des populations diverses, le projet HapMap a jeté les bases nécessaires pour comprendre la structure génétique de l'humanité et ses implications pour la santé humaine. Il a aussi permis un meilleur traitement de données génétiques, en fournissant des statistiques exploitables sur les variations les plus communément retrouvées ensemble (Y. Li et al. 2010; Barrett et al. 2005).

6.2.1.3. *1000 Genomes Project*

Le *1000 Genomes Project*, une initiative de collaboration d'envergure internationale, était un projet ambitieux visant à séquencer le génome de personnes provenant de diverses régions du monde (The 1000 Genomes Project Consortium 2015). Conçu dans le contexte de la recherche en génomique humaine, ce projet a été mené en parallèle avec le projet HapMap dans le but de générer des données de haute qualité sur les variations génétiques, y compris les liens à longue distance entre les mutations.

La première phase du projet, lancée avec pour objectif initial d'intégrer les données provenant de mille individus, a rapidement évolué pour inclure un nombre plus important de participants. La version finale de la base de données résultante contient ainsi des informations génétiques provenant d'environ 2500 personnes, fournissant une représentation plus exhaustive de la diversité génétique humaine à l'échelle mondiale.

Pour atteindre ses objectifs, le *1000 Genomes Project* a mis en œuvre une combinaison d'approches pour cerner l'abondance et le type des différents variants que contiennent le génome Humain. Les concepteurs du projet ont diversifié à la fois leurs approches techniques, en employant le séquençage d'exome ainsi du séquençage génomique complet, et en diversifiant leurs approches d'échantillonnage, en sélectionnant des participants généraux, ainsi que des "trios" familiaux, c'est-à-dire deux parents et un enfant, pour mieux étudier l'héritage des mutations.

Les données générées par le *1000 Genomes Project* sont utilisées couramment dans la recherche en génomique et en génétique humaine. Par exemple, ces données ont été employées pour étudier le génome mitochondrial chez l'Humain (Ovchinnikov and Kholina 2010), les bases moléculaires du cancer (Ma, Funk, and Price 2010) et l'identification de variants causant des maladies rares (Zawistowski et al. 2010). Tout comme HapMap, le *1000 Genomes Project* a également servi à améliorer nos capacités à analyser les données de séquençage chez l'Humain (Y. Li, Byrnes, and Li 2010).

En fournissant un accès ouvert à ces données précieuses, le *1000 Genomes Project* a stimulé l'innovation et la collaboration dans le domaine de la recherche en génomique, ouvrant ainsi de nouvelles perspectives pour la médecine personnalisée, la prévention des maladies et la

santé publique à l'échelle mondiale.

6.2.1.4. SNPedia

SNPedia est une plateforme collaborative où chaque utilisateur a la possibilité de soumettre des informations concernant les mutations génétiques (Cariaso and Lennon 2012). Ces données proviennent fréquemment de séquençages génétiques privés réalisés par des entreprises privées telles que 23andMe ou MyHeritage. Cette diversité d'origine des données reflète la contribution de multiples sources à la richesse de SNPedia, renforçant ainsi sa pertinence et sa représentativité dans le domaine de la génomique et de la médecine personnalisée. SNPedia a été acquis par MyHeritage en 2019, conjointement avec l'outil Promethease développé par la même équipe. Promethease permet de comparer efficacement ses données personnelles contre celles contenues dans SNPedia.

L’approche unique de SNPedia, qui se présente sous forme de wiki, tente de rendre plus démocratisée l’accès aux connaissances en génétique et génomique. Leur page d’accueil présente des exemples et tutoriels pour aider les utilisateurs à naviguer et comprendre les données. Les pages individuelles des mutations contiennent des informations vulgarisées sur les effets de la mutation, et redirigent les utilisateurs vers des publications académiques permettant d’amener plus de détails. Cette présentation est à saluer pour son originalité et sa transparence, assez unique dans le domaine de la génomique.

Cet outil a aussi été utilisé dans des projets académiques, pour développer des outils d’analyse de tests génétiques (M. J. Li et al. 2016; Folkersen et al. 2020; Lau et al. 2021). Plusieurs de ces outils ont aussi une visée de science ouverte, en essayant de rendre l’information plus accessible aux utilisateurs.

6.2.1.5. Human Variome project

Le Human Variome Project, initié par l'Université de Melbourne en Australie, vise à établir une base de données exhaustive englobant les variations génétiques ayant un impact avéré sur la santé des individus porteurs (Ring, Kwok, and Cotton 2006; Cotton et al. 2008). L'implication du projet dans des initiatives internationales, notamment son affiliation à

l'UNESCO, démontre son engagement à promouvoir une collaboration mondiale dans le domaine de la génomique et de la médecine génétique. Cette collaboration vise à rassembler les connaissances et les ressources provenant de diverses régions du monde, dans le but ultime de mieux comprendre les variations génétiques et leur incidence sur la santé humaine à l'échelle mondiale.

En 2021, une étape majeure a été franchie lorsque le Human Variome Project a fusionné avec la *Human Genome Organisation* (HUGO), consolidant ainsi les données et les efforts dans ce domaine. Cette fusion stratégique a permis de centraliser toutes les ressources disponibles pour faciliter l'accès aux données génétiques et favoriser la collaboration entre les chercheurs et les cliniciens du monde entier. La base de données résultante de cette collaboration conjointe est désormais accessible librement et est hébergée par l'Université de Leiden, offrant ainsi une plateforme robuste pour l'échange et la recherche de données génétiques pertinentes.

Ce projet a eu un impact plus modéré dans la communauté scientifique, mais ses contributions dans la prédiction de maladies à partir de variants génétiques est reconnue (López-Ferrando et al. 2017; Soussi, Leroy, and Taschner 2014; Robinson 2012). Le Human Variome Project et la Human Genome Organisation ont également été salués pour leur rôle central dans la promotion de la recherche collaborative à l’échelle mondiale.

6.2.1.6. Genome Aggregation Database

Le *Genome Aggregation Database* (gnomAD), initialement lancé en 2014 sous l'appellation *Exome Aggregation Consortium* (ExAC) (Karczewski et al. 2017; 2020), représente une entreprise collaborative d'envergure visant à partager des données de séquençage génomique et exomique. Dirigé par deux co-directeurs éminents basés dans le Massachusetts, aux États-Unis, et bénéficiant de l'expertise d'un comité de conseil scientifique composé de chercheurs internationaux, ce projet incarne un effort concerté pour consolider et rendre accessible une vaste gamme de données génétiques.

Les données brutes recueillies par les divers projets intégrés dans cette base de données sont soumises à un processus rigoureux de réanalyse systématique afin de garantir la cohérence et la qualité des informations fournies. Cependant, il est important de noter que les méthodologies et les techniques de séquençage utilisées peuvent varier considérablement d'un

projet à l'autre, entraînant ainsi des différences dans la couverture entre individus et entre différentes régions du génome.

Les données hébergées par le gnomAD ont été utilisées dans un large éventail de projets de recherche en génétique et en génomique, soutenant des études sur des maladies encore peu connues, comme l'autisme ou la schizophrénie (Singh et al. 2022; J. M. Fu et al. 2022). Elles ont également été utilisées pour étudier le SARS-CoV2 (Niemi et al. 2021; Pairo-Castineira et al. 2021).

6.2.1.7. UK Biobank

La UK Biobank, une étude d'envergure lancée par le service hospitalier national au Royaume-Uni, représente une entreprise ambitieuse visant à explorer les liens complexes entre les facteurs génétiques, les habitudes de vie et la santé humaine (Sudlow et al. 2015). Cette étude a mobilisé la participation de près de 500 000 individus, qui ont consenti à remplir des questionnaires détaillés sur leurs habitudes de vie, leur état de santé et les médicaments qu'ils consomment. En outre, des mesures physiologiques telles que la taille, le poids, la pression sanguine et d'autres paramètres ont été recueillies, offrant ainsi une vue complète de la santé et du bien-être des participants.

L'objectif principal de la UK Biobank est de suivre ces participants sur une période de 30 ans, soit de 2006 à 2036, afin de documenter l'apparition et la progression de maladies chroniques et dégénératives. Cette approche longitudinale permettra de mieux comprendre les facteurs de risque associés au développement de maladies telles que le diabète, les maladies cardiovasculaires, le cancer et les troubles neurodégénératifs, ainsi que de développer des stratégies de prévention et de traitement plus efficaces.

Les données recueillies par la UK Biobank ont déjà été employées pour étudier les associations entre maladies, mutations et environnement (Hutton et al. 2018). En particulier, des maladies complexes et multifactorielles comme le diabète, l'arthrite ou la démence (Liu et al. 2023; Hutton et al. 2018; Carrasquilla et al. 2023). En Novembre 2023, la UK Biobank a ouvert l'accès à une grande partie de sa base de données, ce qui favorise l'innovation et la collaboration dans la recherche biomédicale, ouvrant ainsi de nouvelles perspectives pour la prévention, le

diagnostic et le traitement des maladies.

6.2.1.8. Autres projets nationaux notables

Outre la UK Biobank, de nombreux autres projets nationaux de génotypage et de suivi des participants ont émergé à travers le monde, chacun apportant sa propre contribution à la recherche en génomique et en médecine personnalisée. Parmi ceux-ci, le Estonian Genome Project (Metspalu 2002), lancé dès l'an 2000, a été l'un des précurseurs dans ce domaine, bien qu'il ait rencontré des difficultés de financement qui ont restreint son ampleur et son impact potentiel. FinnGen en Finlande vise à comprendre les bases génétiques des maladies courantes dans la population finlandaise (Kurki et al. 2023). La China Kadoorie Biobank a collecté des données auprès de plus de 500 000 participants dans le but d'étudier les facteurs de risque pour les principales maladies chroniques dans la population chinoise (Zhengming Chen et al. 2005). Aux États-Unis, le projet All of Us, lancé par les *National Institutes of Health* (NIH), cherche à rassembler des données de santé, de génétique et de mode de vie provenant d'un million de personnes dans le but de mieux comprendre la santé et la maladie (‘The “All of Us” Research Program’ 2019).

L'entreprise deCODE Genetics, basée en Islande, a réalisé d'importantes avancées dans la cartographie du génome islandais et l'identification de variants génétiques associés à des maladies complexes telles que le cancer et les maladies cardiovasculaires (deCODE 2012). Toutefois, les efforts de deCODE sont davantage axés sur la généalogie et l'utilisation de marqueurs génétiques tels que les microsatellites plutôt que sur le séquençage complet du génome.

Ces initiatives représentent un échantillon diversifié de projets nationaux qui contribuent de manière significative à la recherche en génomique et en médecine personnalisée à l'échelle mondiale. Leur impact se manifeste dans la découverte de variants génétiques associés à des maladies, l'identification de facteurs de risque pour la santé, et la promotion d'une meilleure compréhension des bases génétiques de la maladie. En collaborant et en partageant les données à travers ces projets, la communauté scientifique peut accélérer les progrès vers une médecine plus personnalisée et une meilleure santé pour tous.

L’Union Européenne a initié un projet de partage de ces ressources nationales, la 1+

Million Genomes Initiative (European Genomic Data Infrastructure 2018), qui visait à rendre accessible 1 000 000 de génomes européens en 2022. La disponibilité des données n'a pas encore été annoncée par le comité d'organisation

6.2.2. Risques associés

La présence de ces données détaillées dans ces projets présente cependant des risques pour les personnes qui y ont contribué. Plusieurs projets de réidentification ont eu lieu qui ont montré qu'il est possible de retrouver le nom exact d'une personne dont les données sont disponibles dans des bases de données biologiques (Check Hayden 2013). Ces personnes sont donc légitimes à s'inquiéter sur la sécurité de leur identité ainsi que les maladies auxquelles ils pourraient être susceptibles (Shabani and Marelli 2019).

En parallèle, révéler ces informations génétiques publiquement peut impacter négativement la vie du porteur. En période de manque d'équipement médical, par exemple due à la pandémie de la COVID-19, certains patients porteurs de marqueurs génétiques de maladie chroniques se sont vus placer à un niveau de priorité plus bas que les personnes pour lesquelles cette information n'était pas disponible (Braverman, Shapiro, and Bernstein 2018). Certains sondages ont aussi révélé une timidité dans certains pays à participer à des projets de séquençage (Barlow-Stewart et al. 2018). Si ces pays possèdent des lois obligeant les personnes à divulguer toutes les informations, alors les participants potentiels craignent que des informations sur leur génome révélés par un séquençage leur coûte à l'avenir.

La communauté scientifique en biologie a donc besoin de méthodes pour protéger les identités de ces personnes, si elle souhaite continuer à réaliser ce genre d'étude à grande envergure.

La médecine légale emploie souvent la réidentification de personnes à partir d'échantillons biologiques, dans le cadre de la recherche d'un suspect potentiel ou d'une personne disparue. Les spécialistes dans ce domaine ont établi des registres de SNPs considérés comme utiles dans le contexte de la réidentification (Budowle and van Daal 2008; Novroski and

Cihlar 2022). Ces catégories sont au nombre de cinq, séparées selon le type d'information que la mutation fournit :

- Informations sur l'identité. Ces SNPs sont utilisés pour identifier de manière unique une personne. Pour cela, ils sont idéalement présents à ~50% dans une population donnée, pour apporter un maximum de puissance discriminante.
- Informations sur la lignée. L'information apportée par ces mutations est souvent utilisée lors de cas de personnes disparues, lorsque bien souvent les spécialistes comparent des échantillons d'ADN trouvés sur le terrain à l'ADN de membres de la famille de la personne disparue.
- Informations sur l'ascendance. Avec ces SNPs, il est possible d'inférer des caractéristiques physiques du porteur en se basant sur les moyennes présentes dans certaines populations.
- Informations sur le phénotype. Ces SNPs sont directement liés à des caractéristiques physiques spécifiques, comme la couleur des yeux ou des cheveux. Cependant, beaucoup de caractéristiques physiques chez l'humain sont dites complexes, c'est-à-dire dépendent de plusieurs sites génomiques et sont donc plus difficiles à résoudre.
- Informations pharmacogénétiques. Cette dernière catégorie regroupe les SNP ayant un impact sur le fonctionnement et l'efficacité de médicaments et drogues. Il est surtout pertinent lors de l'établissement de diagnostics et pronostics de patients, mais peut être utile pour les forces de l'ordre pour retrouver une personne qui nécessite un médicament spécifique par exemple.

On pourrait imaginer un score, permettant de caractériser des données génomiques, basé sur ces catégories de SNPs. Si les données partagées contiennent de nombreuses mutations qui sont listées dans ces registres de SNPs, alors le score serait élevé, vu qu'il serait plus facile de réidentifier la personne ayant fourni l'échantillon. Instinctivement, on aurait donc envie d'explorer des méthodes permettant de retirer ces mutations des échantillons présents dans la base de données pour protéger ses participants. Ces méthodes existent, et il est possible d'en implémenter une gamme très variée selon le type de données utilisé et l'objectif précis de l'utilisateur.

6.2.3. Masquer l'information

6.2.3.1. Anonymisation

Une solution envisageable pour faciliter le partage des données sensibles est d'en retirer cette partie sensible. En repérant les éléments qu'on sait révélateurs d'une origine ethnique ou d'une caractéristique physique, et en filtrant un fichier de résultats de séquençage brut pour les retirer, il serait possible d'obtenir des données "anonymisées".

Cette anonymisation peut s'effectuer sous plusieurs formes. On liste en général cinq approches (Lei Xu et al. 2014). Prenons une base de données fictive pour illustrer ces méthodes.

Identifiant	Age	Vaccination	Fièvre	Douleurs	Fatigue	Hospitalisation
ID001	42	Oui	Non	Non	Non	Non
ID002	23	Oui	Non	Non	Non	Non
ID003	56	Oui	Non	Oui	Oui	Non
ID004	40	Oui	Non	Non	Non	Non
ID005	61	Oui	Oui	Oui	Oui	Oui
ID006	61	Oui	Oui	Non	Oui	Non
ID007	33	Oui	Non	Non	Non	Non
ID008	24	Oui	Oui	Non	Oui	Non
ID009	21	Oui	Non	Non	Non	Non
ID010	39	Non	Oui	Oui	Oui	Oui
ID011	62	Non	Oui	Oui	Oui	Oui
ID012	25	Non	Oui	Non	Oui	Non
ID013	40	Non	Non	Oui	Oui	Non
ID014	25	Oui	Non	Non	Non	Non
ID015	30	Oui	Non	Oui	Oui	Non
ID016	46	Oui	Oui	Oui	Oui	Oui
ID017	62	Non	Oui	Oui	Oui	Oui
ID018	30	Oui	Oui	Oui	Oui	Oui
ID019	58	Oui	Non	Oui	Oui	Non
ID020	35	Oui	Non	Non	Non	Non
ID021	68	Non	Oui	Oui	Oui	Oui
ID022	64	Non	Non	Oui	Oui	Non
ID023	44	Oui	Non	Non	Oui	Non
ID024	35	Oui	Non	Non	Non	Non
ID025	31	Oui	Oui	Oui	Non	Non
ID026	40	Non	Oui	Oui	Oui	Oui
ID027	39	Non	Oui	Oui	Oui	Oui
ID028	59	Non	Oui	Oui	Oui	Oui
ID029	74	Oui	Oui	Oui	Oui	Oui
ID030	65	Non	Oui	Oui	Non	Non
ID031	71	Oui	Non	Non	Non	Non
ID032	32	Oui	Non	Non	Non	Non

Figure 4 : Base de données fictive pour illustrer les techniques d'anonymisation.

Elle est représentée dans la figure 4. Pour 32 personnes, cette base de données contient des informations sur leur âge, leur statut vaccinal, les symptômes qu'ils ont développés après infection par une maladie quelconque, et si leur état a nécessité une hospitalisation.

Identifiant	Age	Vaccination	Fièvre	Douleurs	Fatigue	Hospitalisation
ID001	30-49	Oui	Non	Non	Non	Non
ID002	18-29	Oui	Non	Non	Non	Non
ID003	50-69	Oui	Non	Oui	Oui	Non
ID004	30-49	Oui	Non	Non	Non	Non
ID005	50-69	Oui	Oui	Oui	Oui	Oui
ID006	50-69	Oui	Oui	Non	Oui	Non
ID007	30-49	Oui	Non	Non	Non	Non
ID008	18-29	Oui	Oui	Non	Oui	Non
ID009	18-29	Oui	Non	Non	Non	Non
ID010	30-49	Non	Oui	Oui	Oui	Oui
ID011	50-69	Non	Oui	Oui	Oui	Oui
ID012	18-29	Non	Oui	Non	Oui	Non
ID013	30-49	Non	Non	Oui	Oui	Non
ID014	18-29	Oui	Non	Non	Non	Non
ID015	30-49	Oui	Non	Oui	Oui	Non
ID016	30-49	Oui	Oui	Oui	Oui	Oui
ID017	50-69	Non	Oui	Oui	Oui	Oui
ID018	30-49	Oui	Oui	Oui	Oui	Oui
ID019	50-69	Oui	Non	Oui	Oui	Non
ID020	30-49	Oui	Non	Non	Non	Non
ID021	50-69	Non	Oui	Oui	Oui	Oui
ID022	50-69	Non	Non	Oui	Oui	Non
ID023	30-49	Oui	Non	Non	Oui	Non
ID024	30-49	Oui	Non	Non	Non	Non
ID025	30-49	Oui	Oui	Oui	Non	Non
ID026	30-49	Non	Oui	Oui	Oui	Oui
ID027	30-49	Non	Oui	Oui	Oui	Oui
ID028	50-69	Non	Oui	Oui	Oui	Oui
ID029	70+	Oui	Oui	Oui	Oui	Oui
ID030	50-69	Non	Oui	Oui	Non	Non
ID031	70+	Oui	Non	Non	Non	Non
ID032	30-49	Oui	Non	Non	Non	Non

Figure 5 : Généralisation appliquée à la base de données fictive.

La généralisation, illustrée dans la figure 5, consiste à remplacer une valeur spécifique par une valeur moins précise pour ce critère. Ici, on substitue l'âge par une tranche d'âge, qui reste informative mais moins spécifique.

Identifiant	Age	Vaccination	Fièvre	Douleurs	Fatigue
ID001	42	Oui	Non	Non	Non
ID002	23	Oui	Non	Non	Non
ID003	56	Oui	Non	Oui	Oui
ID004	40	Oui	Non	Non	Non
ID005	61	Oui	Oui	Oui	Oui
ID006	61	Oui	Oui	Non	Oui
ID007	33	Oui	Non	Non	Non
ID008	24	Oui	Oui	Non	Oui
ID009	21	Oui	Non	Non	Non
ID010	39	Non	Oui	Oui	Oui
ID011	62	Non	Oui	Oui	Oui
ID012	25	Non	Oui	Non	Oui
ID013	40	Non	Non	Oui	Oui
ID014	25	Oui	Non	Non	Non
ID015	30	Oui	Non	Oui	Oui
ID016	46	Oui	Oui	Oui	Oui
ID017	62	Non	Oui	Oui	Oui
ID018	30	Oui	Oui	Oui	Oui
ID019	58	Oui	Non	Oui	Oui
ID020	35	Oui	Non	Non	Non
ID021	68	Non	Oui	Oui	Oui
ID022	64	Non	Non	Oui	Oui
ID023	44	Oui	Non	Non	Oui
ID024	35	Oui	Non	Non	Non
ID025	31	Oui	Oui	Oui	Non
ID026	40	Non	Oui	Oui	Oui
ID027	39	Non	Oui	Oui	Oui
ID028	59	Non	Oui	Oui	Oui
ID029	74	Oui	Oui	Oui	Oui
ID030	65	Non	Oui	Oui	Non
ID031	71	Oui	Non	Non	Non
ID032	32	Oui	Non	Non	Non

Figure 6 : Suppression d'une valeur dans la base de données fictive.

Lors de la suppression, on retire une partie des données de l'ensemble qui sera partagé. Cette suppression peut concerner une cellule, une variable ou un échantillon, ou plusieurs de ces valeurs, selon l'échelle requise. Un exemple est représenté figure 6, où la base de données n'indique plus si la personne a nécessité une hospitalisation ou non.

Identifiant	Age	Vaccination
ID001	42	Oui
ID002	23	Oui
ID003	56	Oui
ID004	40	Oui
ID005	61	Oui
ID006	61	Oui
ID007	33	Oui
ID008	24	Oui
ID009	21	Oui
ID010	39	Non
ID011	62	Non
ID012	25	Non
ID013	40	Non
ID014	25	Oui
ID015	30	Oui
ID016	46	Oui
ID017	62	Non
ID018	30	Oui
ID019	58	Oui
ID020	35	Oui
ID021	68	Non
ID022	64	Non
ID023	44	Oui
ID024	35	Oui
ID025	31	Oui
ID026	40	Non
ID027	39	Non
ID028	59	Non
ID029	74	Oui
ID030	65	Non
ID031	71	Oui
ID032	32	Oui

Identifiant	Age	Fièvre	Douleurs	Fatigue
ID001	42	Non	Non	Non
ID002	23	Non	Non	Non
ID003	56	Non	Oui	Oui
ID004	40	Non	Non	Non
ID005	61	Oui	Oui	Oui
ID006	61	Oui	Non	Oui
ID007	33	Non	Non	Non
ID008	24	Oui	Non	Oui
ID009	21	Non	Non	Non
ID010	39	Oui	Oui	Oui
ID011	62	Oui	Oui	Oui
ID012	25	Oui	Non	Oui
ID013	40	Non	Oui	Oui
ID014	25	Non	Non	Non
ID015	30	Non	Oui	Oui
ID016	46	Oui	Oui	Oui
ID017	62	Oui	Oui	Oui
ID018	30	Oui	Oui	Oui
ID019	58	Non	Oui	Oui
ID020	35	Non	Non	Non
ID021	68	Oui	Oui	Oui
ID022	64	Non	Oui	Oui
ID023	44	Non	Non	Oui
ID024	35	Non	Non	Non
ID025	31	Oui	Oui	Non
ID026	40	Oui	Oui	Oui
ID027	39	Oui	Oui	Oui
ID028	59	Oui	Oui	Oui
ID029	74	Oui	Oui	Oui
ID030	65	Oui	Oui	Non
ID031	71	Non	Non	Non
ID032	32	Non	Non	Non

Identifiant	Age	Hospitalisation
ID001	42	Non
ID002	23	Non
ID003	56	Non
ID004	40	Non
ID005	61	Oui
ID006	61	Non
ID007	33	Non
ID008	24	Non
ID009	21	Non
ID010	39	Oui
ID011	62	Oui
ID012	25	Non
ID013	40	Non
ID014	25	Non
ID015	30	Non
ID016	46	Oui
ID017	62	Oui
ID018	30	Oui
ID019	58	Non
ID020	35	Non
ID021	68	Oui
ID022	64	Non
ID023	44	Non
ID024	35	Non
ID025	31	Non
ID026	40	Oui
ID027	39	Oui
ID028	59	Oui
ID029	74	Oui
ID030	65	Non
ID031	71	Non
ID032	32	Non

Figure 7 : Anatomisation de la base de données fictive en trois sous-bases de données.

L'anatomisation consiste à séparer les variables contenues dans un jeu de données en plusieurs ensembles. Ainsi, on ne supprime ni ne modifie rien dans ces données, on vient simplement rompre les liaisons entres valeurs. La figure 7 montre la scission de la base de données originale en trois bases distinctes. On peut imaginer que l'accès à chacune de ces bases pourrait être accordé individuellement, ou qu'on retirerait également l'identifiant unique des participants pour briser les liens complètement entre ces informations.

Identifiant	Age	Vaccination	Fièvre	Douleurs	Fatigue	Hospitalisation
ID001	23	Oui	Non	Non	Non	Non
ID002	42	Oui	Non	Non	Non	Non
ID003	40	Oui	Non	Oui	Oui	Non
ID004	56	Oui	Non	Non	Non	Non
ID005	61	Oui	Oui	Oui	Oui	Oui
ID006	61	Oui	Oui	Non	Oui	Non
ID007	24	Oui	Non	Non	Non	Non
ID008	33	Oui	Oui	Non	Oui	Non
ID009	39	Oui	Non	Non	Non	Non
ID010	21	Non	Oui	Oui	Oui	Oui
ID011	25	Non	Oui	Oui	Oui	Oui
ID012	62	Non	Oui	Non	Oui	Non
ID013	25	Non	Non	Oui	Oui	Non
ID014	40	Oui	Non	Non	Non	Non
ID015	46	Oui	Non	Oui	Oui	Non
ID016	30	Oui	Oui	Oui	Oui	Oui
ID017	30	Non	Oui	Oui	Oui	Oui
ID018	62	Oui	Oui	Oui	Oui	Oui
ID019	35	Oui	Non	Oui	Oui	Non
ID020	58	Oui	Non	Non	Non	Non
ID021	64	Non	Oui	Oui	Oui	Oui
ID022	68	Non	Non	Oui	Oui	Non
ID023	35	Oui	Non	Non	Oui	Non
ID024	44	Oui	Non	Non	Non	Non
ID025	40	Oui	Oui	Oui	Non	Non
ID026	31	Non	Oui	Oui	Oui	Oui
ID027	59	Non	Oui	Oui	Oui	Oui
ID028	39	Non	Oui	Oui	Oui	Oui
ID029	65	Oui	Oui	Oui	Oui	Oui
ID030	74	Non	Oui	Oui	Non	Non
ID031	32	Oui	Non	Non	Non	Non
ID032	71	Oui	Non	Non	Non	Non

Figure 8 : Permutation des valeurs d'âge dans la base de données fictive.

La permutation mélange les associations entre valeurs. Cette stratégie peut s'implémenter à plusieurs échelles, par exemple permuter les valeurs d'une base de données établie dans une région spécifique, avant de centraliser les données vers un organisme plus large. Ici, ce sont les

âges des participants qui ont été permutées deux à deux, comme le montre la figure 8.

Identifiant	Age	Vaccination	Fièvre	Douleurs	Fatigue	Hospitalisation
ID001	41	Oui	Non	Non	Non	Non
ID002	23	Oui	Non	Non	Non	Non
ID003	56	Oui	Non	Oui	Oui	Non
ID004	42	Oui	Non	Non	Non	Non
ID005	60	Oui	Oui	Oui	Oui	Oui
ID006	63	Oui	Oui	Non	Oui	Non
ID007	34	Oui	Non	Non	Non	Non
ID008	24	Oui	Oui	Non	Oui	Non
ID009	21	Oui	Non	Non	Non	Non
ID010	37	Non	Oui	Oui	Oui	Oui
ID011	64	Non	Oui	Oui	Oui	Oui
ID012	25	Non	Oui	Non	Oui	Non
ID013	38	Non	Non	Oui	Oui	Non
ID014	24	Oui	Non	Non	Non	Non
ID015	28	Oui	Non	Oui	Oui	Non
ID016	44	Oui	Oui	Oui	Oui	Oui
ID017	64	Non	Oui	Oui	Oui	Oui
ID018	30	Oui	Oui	Oui	Oui	Oui
ID019	60	Oui	Non	Oui	Oui	Non
ID020	37	Oui	Non	Non	Non	Non
ID021	67	Non	Oui	Oui	Oui	Oui
ID022	64	Non	Non	Oui	Oui	Non
ID023	43	Oui	Non	Non	Oui	Non
ID024	34	Oui	Non	Non	Non	Non
ID025	29	Oui	Oui	Oui	Non	Non
ID026	40	Non	Oui	Oui	Oui	Oui
ID027	38	Non	Oui	Oui	Oui	Oui
ID028	60	Non	Oui	Oui	Oui	Oui
ID029	76	Oui	Oui	Oui	Oui	Oui
ID030	64	Non	Oui	Oui	Non	Non
ID031	70	Oui	Non	Non	Non	Non
ID032	31	Oui	Non	Non	Non	Non

Figure 9 : Perturbation des valeurs d'âge dans la base de données fictive.

La perturbation introduit une légère variance pour certains attributs de la base de données, montrée dans la figure 9. On peut ainsi préserver la distribution statistique globale de cette variable mais sans divulguer de données exactes pour celle-ci. Dans notre cas, c'est encore une

fois l'âge qui subit cette modification, où chaque valeur se voit modifiée aléatoirement de 0 à 2 années, positivement ou négativement.

Le processus d'anonymisation est approprié pour des données telles que de la transcriptomique (Ziegenhain and Sandberg 2021), où le taux d'expression relatif de chaque gène est le facteur étudié. Les données génomiques, elles, cherchent justement à employer ces mutations, qui peuvent permettre la réidentification de leur porteur, pour tenter de trouver des liens entre mutations et maladies par exemple. La nature de l'information examinée ici constitue intrinsèquement une menace de divulgation de données personnelles, englobant des renseignements susceptibles de compromettre le droit au secret médical et à la vie privée. Actuellement, il existe peu de solutions garantissant le respect de cette sphère privée.

6.2.3.2. Chiffrement homomorphe

Le chiffrement homomorphe est une méthode qui peut permettre l'analyse de données sensibles par un autre utilisateur (Gentry 2009; Gentry, Sahai, and Waters 2013).

Le chiffrement est un procédé destiné à masquer des données au grand public, en permettant la compréhension à un petit groupe de personnes ayant accès à la clé du chiffrement. Il existe un nombre très important de méthodes de chiffrement différents, mettant en jeu des algorithmes et des clés de natures différentes.

Ici, le terme homomorphe fait référence à la propriété d'homomorphisme mathématique, qui est une relation entre deux structures algébriques de même type qui préserve toutes les opérations et les structures de celles-ci. Les fonctions de chiffrement et de déchiffrement employées dans le chiffrement homomorphe sont des homomorphismes entre les espaces de données originales et de données chiffrées.

Cette famille de méthodes de chiffrement impose une condition supplémentaire sur la transformation cryptographique que subissent les données : le résultat de toute opération réalisée sur les données chiffrées doit, une fois déchiffré, correspondre au résultat de l'application de cette même opération sur les données originales (Kim and Lauter 2015).

Pour des données de séquençage, on pourrait considérer comme exemple un chiffrement

homomorphe très naïf : remplacer les valeurs {A,C,G,T} que peuvent prendre une séquence nucléotidique par quatre autres lettres quelconques. Ainsi, aucune information sur la nature de la séquence (codante, GC%, etc) n'est révélée, mais tous les calculs d'alignement et de conservation entre séquences restent applicables.

La limite de cet exemple intervient lorsqu'on cherche à pousser les analyses plus loin. Si on souhaite savoir à quelle séquence traduite correspond une région conservée entre tous les échantillons, il nous faut connaître la séquence originale pour procéder à la traduction. Les données traduites, elles, sont plus complexes à chiffrer de cette manière. Une substitution d'un acide aminé par un autre implique des distances très différentes selon le résidu original et le nouveau, de par la redondance des codons et de leurs propriétés physico-chimiques. Il faut donc reproduire la matrice de distances entre acides aminés substitués, ce qui peut rompre le chiffrement des données en permettant à un agent externe de les déchiffrer.

Cette limite impacte aussi les outils applicables aux données de séquençage chiffrées, puisque beaucoup de calculs de conservation de séquence sont plus performants lorsqu'ils prennent en compte les séquences protéiques qui résultent des parties codantes du génome.

En outre, le chiffrement homomorphe est applicable lorsque l'utilisateur souhaite examiner ses données sur un serveur de calcul qu'il ne considère pas spécialement digne de confiance, que ce soit en termes d'hébergement ou de mesures de protection mises en place pour prévenir des attaques externes. Il est supposé que ce même utilisateur procédera ensuite à l'interprétation des résultats de ces calculs localement, après les avoir récupérés. Cependant, si une source d'informations centralisée tel qu'un hôpital cherchait à mettre en place cette pratique pour que des chercheurs aient accès plus facilement aux données, cela limiterait énormément l'efficacité possible de l'étude réalisée. "La caractéristique clinique n°X prédit la réponse au traitement avec Y% d'erreur" n'est pas très vendeur comme titre de publication scientifique, et ne permet pas toujours de remonter à la source de cette information, mais il serait compliqué de divulguer plus d'informations sur le jeu de données en question sans commencer à risquer des fuites de données médicales.

Il existe bien évidemment d'autres méthodes de chiffrement homomorphe qui sont applicables dans d'autres domaines (Gentry, Halevi, and Smart 2012; Gentry, Sahai, and Waters 2013; Cheon et al. 2017), et qui ont fait leurs preuves. Leurs fonctionnements exacts sont au-delà

de ce manuscrit, mais ces pratiques sont applicables à des données de séquençage. Cependant, il est très coûteux en ressources de calcul de les appliquer sur des jeux de données pertinents pour la recherche académique en biologie. Certaines techniques hybrides (pour lesquelles les calculs simples sont effectués localement, et les calculs coûteux sur un ou plusieurs serveurs distants) existent (Zhao, Wang, and Tang 2018), mais encore une fois le coût en temps et en ressources de calcul est non-négligeable. Un traitement dans un système totalement homomorphe peut prendre jusqu'à un million de fois plus de temps que le traitement équivalent sur les données brutes (Gentry 2009).

6.2.3.3. Apprentissage fédéré

L'apprentissage fédéré est une stratégie d'entraînement de méthodes d'intelligence artificielle qui permet de respecter les règles de sécurité imposées à certaines structures hébergeueses (Abadi et al. 2016; Abay et al. 2019). Nous discuterons dans cette section des avantages et inconvénients de l'apprentissage fédéré. Le processus d'entraînement de ces méthodes d'intelligence artificielle sera développé plus loin dans la section 7.1.1.2 ("Fonctionnement de l'apprentissage automatique")

Au lieu de créer une grande base de données regroupant tous les échantillons analysés par plusieurs organismes indépendants (par exemple, un ensemble d'hôpitaux dispersés dans un pays), chaque acteur conserve les données localement. Assez souvent, l'apprentissage est dirigé par un organisme central, mais il peut être partiellement ou complètement décentralisé. Une méthode et un modèle initiaux sont choisis, et diffusés à l'ensemble des participants. Ces participants réalisent une optimisation du modèle localement sur leurs données, puis transmettent les paramètres finaux de leurs modèles pour que ces paramètres soient mis en commun, créant un modèle plus robuste que chacun le serait individuellement.

Certains points sont tout de même à prendre en compte lorsque l'on choisit de procéder à un apprentissage fédéré. Il est fort probable que les jeux de données de chaque participant ne contiennent pas le même nombre d'échantillons, ni que ces échantillons proviennent d'une même source. Ceci devient problématique lors de la mise en commun, puisqu'il faut prendre en compte les abondances des diverses sous-populations et leur impact sur les résultats du modèle final. De la même manière, la complexité et la distribution statistique des échantillons dans ces

sous-ensembles de données présenteront de la variabilité (par exemple, pour des données relatives à une maladie infectieuse, les informations issues d'une maison de retraite et des données issues d'une école représentent des sections très différentes de la population totale). En choisissant un type de modèle d'analyse sans prendre en compte ces critères, on risque d'en biaiser les performances. En règle générale, il est de bonne pratique de prévoir que certains participants à cette stratégie puissent avoir de mauvaises intentions, ou simplement être fainéants et ne respectent pas le code de conduite fixé par l'organisation centrale. Dans ces cas, il faut prévoir des stratégies qui permettent de les détecter et de minimiser leur impact sur l'apprentissage.

De manière générale, même s'il est possible de crypter, d'obfusquer, de diverses manières les détails des données que les projets de bioinformatique utilisent, cette pratique restera toujours coûteuse et limitera les stratégies que les chercheurs peuvent employer. ll serait plus aisé pour les équipes de recherche d'avoir accès à des données non-sensibles pouvant être utilisées à loisir pour dresser des nouvelles hypothèses sur leurs thématiques d'intérêt. En génomique, de nombreuses informations clés pour la recherche peuvent également servir à réidentifier la personne ayant fourni ces informations (Shabani and Marelli 2019). Il faut donc trouver un équilibre entre le taux d'information sensible et le taux d'information exploitable. Comment pourrait-on partager des données à la fois informatives mais non-sensibles ?

6.2.3.4. Approche par k-mers

Une méthode qui a fait ses preuves est l'utilisation de *k*-mers, c'est-à-dire des séquences (ici de nucléotides) de taille *k*. Cette technique peut s'utiliser sans alignement, et cherche à identifier de courtes séquences corrélées aux divers groupes d'échantillons dans le jeu de données de base. Elle est particulièrement adaptée pour l'analyse de données métagénomiques (Ren et al. 2017; Wang and Gu 2018), où il peut-être difficile voire impossible d'obtenir des génomes de référence pour procéder à un alignement, et pour la transcriptomique, où les ressources computationnelles nécessaires à l'alignement des lectures de séquençage deviennent extrêmement importantes. Dans ces deux domaines, la démarche la plus classique est de classer les échantillons en groupes (ou de laisser un outil effectuer ce travail automatiquement avec du partitionnement) puis de regarder quels *k*-mers permettent de distinguer les différents groupes,

soit par leur présence ou absence, soit par leur abondance.

Cette méthodologie présente de nombreux avantages clés pour l'anonymisation des données en comparant aux méthodes utilisant l'alignement sur un génome de référence. Tout d'abord, ces *k*-mers ne fournissent pas la même information sur les éléments répétés dans le génome, qui constituent une source importante d'information pouvant identifier une personne (Monckton and Jeffreys 1993), en comparaison aux approches basées sur l'alignement de lectures de séquençage. En général, un jeu de données de *k*-mers couvre uniquement le début et la fin d'une série d'éléments répétés, de par le chevauchement de *k*-mers avec les régions non-répétées voisines, et ne permet pas de savoir combien de répétitions composent cette séquence. Deuxièmement, ils ne permettent pas d'identifier directement certaines mutations pouvant indiquer les origines ethniques d'une personne ou ses caractéristiques physiques. Un *k*-mer muté peut être confondu avec un k-mer constitutif originaire d'une autre section du génome si leur séquence est identique. Enfin, il est possible d'appliquer un pré-traitement sur ces jeux de *k*-mers pour en retirer ceux qui sont jugés superflus pour l'analyse en cours, ce qui filtre encore plus fortement les informations contenues dans le jeu de données et ainsi fait baisser la probabilité qu'on puisse réidentifier une personne en les utilisant.

Ces points sont des forces, mais ne sont pas des garanties. En appliquant des filtres stricts sur les données originales, on modifie la nature de l'information étudiée. De plus, pour des organismes comme l'humain, on prend en général k=31, ce qui équivaut à plus de $4*10^{18}$ *k*-mers possibles et donc le même nombre de variables. Les matrices de *k*-mers peuvent donc occuper un volume important d'espace disque, et rendre leur partage inefficace pour des applications qui requièrent la totalité de ces informations. Les *k*-mers ne représentent donc pas non plus une solution optimale au besoin en données génomiques humaines.

6.2.4. Génération de nouvelles données

Une autre approche est de tenter de générer un nouveau jeu de données, plutôt que de modifier ou protéger un jeu de données issu d'échantillons réels. Différentes approches pour réaliser ces simulations existent, chacune avec leurs forces et faiblesses respectives.

6.2.4.1. Approches statistiques

Les premières approches visaient à simuler des populations. En 2008, (Carvajal-Rodríguez 2008b) en faisait déjà un tour d'horizon pour lister les caractéristiques des différents outils. Parmi ceux-ci, on distingue deux catégories majeures : les outils visant à simuler le passé d'une population donnée, et ceux visant à simuler son avenir.

Parmi les simulateurs coalescents présentés dans cette revue, c'est-à-dire ceux qui simulent les ancêtres d'une population, Seq-Gen (Rambaut 2002) et Evolver (Edgar et al. 2009) présentent le plus de fonctionnalités pour la simulation de l'ADN, par exemple en permettant de prendre en compte les taux de mutations entre différents nucléotides ou acides aminés pour effectuer des prédictions plus réalistes d'un point de vue génomique. Inversement, mlcoalsim (Ramos-Onsins and Mitchell-Olds 2008) se focalise sur la simulation de populations la plus réaliste que possible, en permettant à l'utilisateur d'ajuster les tailles des différentes populations ou leurs échanges notamment. Pour les simulateurs permettant de générer la descendance d'une population (ou plusieurs), GenomePop (Carvajal-Rodríguez 2008a) présente la gamme d'options la plus diversifiée. La seule fonctionnalité notable qu'il lui manque est l'établissement d'arbre d'évolution directement, il faut exporter les séquences des différents individus et réaliser l'arbre via un outil tierce.

Depuis cette revue, deux outils notables ont fait leur apparition dans l'écosystème des simulateurs de population : GenomePop2 (Carvajal-Rodríguez 2017) et GeneEvolve (Tahmasbi and Keller 2016). GenomePop2 est un complément à GenomePop, permettant une meilleure gestion des SNPs dans les individus de ses populations. GeneEvolve permet également une simulation détaillée de SNPs et de gènes, montrant que ce niveau de détail était souhaité de la part de la communauté scientifique.

Certains outils approchent la simulation de matériel génétique en modélisant une séquence d'ADN complète, et non les gènes et leurs allèles. Cette tâche est largement plus complexe, de par l'espace des possibles, et donc complexe à mettre en œuvre. Parmi ces outils, GenRGenS (Ponty, Termier, and Denise 2006) est le premier notable à paraître. Il emploie plusieurs modèles d'analyse de séquences d'ADN comme les modèles de Markov ou des

expressions PROSITE (base de données de motifs conservés dans les protéines). Cette diversité de fonctionnalités lui permet de générer des séquences fonctionnelles de grande taille, en respectant les critères spécifiés par l'utilisateur.

Plus récemment, Simulome (Price and Gibas 2017) vise la création de pseudo-génomes de procaryotes. L'utilisateur peut spécifier des éléments qu'il souhaite obtenir dans ce génome final, comme des gènes ou des opérons (ensemble de gènes et éléments régulateurs chez les procaryotes), et Simulome gère la création des séquences permettant de connecter ces éléments, ainsi que des mutations pouvant affecter ces éléments. Les procaryotes présentent une malléabilité génomique bien plus forte que les eucaryotes, c'est pourquoi ce type d'outil est très intéressant pour les chercheurs étudiant ces organismes. Il présente aussi des applications intéressantes en biologie synthétique, une branche des biotechnologies visant à modifier ou créer des micro-organismes qui peuvent servir dans l'industrie. Une des réussites remarquables de ce domaine est la mise au point de la synthèse de l'insuline humaine (pouvant servir de traitement aux personnes atteintes de diabète) chez la bactérie E. coli en 1979 (Goeddel et al. 1979).

Une autre approche est de simuler des variants sur séquence connue. De manière intéressante, la première implémentation de cette approche était en 1997, avec PAML (Yang 1997). Cet outil peut être utilisé pour générer de nouvelles mutations sur une séquence d'ADN ou de protéine, en prenant en compte les taux de substitutions connus pour une mutation donnée dans son contexte.

Plus récemment, avec l'augmentation de la quantité de données de séquençage disponible, il devient de plus en plus facile d'étudier de larges populations et mutations qui les caractérisent. En employant ces données, il est possible de prendre en compte les taux d'occurrence de ces mutations, ainsi que les liaisons entre mutations pour plus de vraisemblance.

Le premier à utiliser ces données était GWAsimulator (C. Li and Li 2008), qui permet de se baser sur les données du projet HapMap pour recréer le déséquilibre de liaison entre mutations à l'échelle locale. Cette fonctionnalité est appliquée à la création de cohortes de données de GWA, à la fois les populations saines et celles affectées par une maladie au choix de l'utilisateur.

PGsim (Juan et al. 2020) et simuG (Yue and Liti 2019) font partie des projets récents dans ce domaine, publiés en 2019 ou 2020. PGsim emploie un large éventail de bases de données

biologiques pour guider les mutations aléatoires qu'il introduit dans le génome qu'il traite, par exemple si la région code pour une protéine ou non. Il permet également d'intégrer des mutations spécifiques de la part de l'utilisateur en même temps, mais n'intègre pas les liaisons entre mutations. simuG effectue un travail semblable (mais intègre moins de sources de connaissances biologiques externes), sur n'importe quel génome de référence, et permet en plus d'intégrer des variations de plus grande taille, comme des translocations, inversions et duplications de régions entières. Ce processus peut également être guidé par des sites connus pour favoriser ce type de variation, notamment les séquences répétées.

Il est intéressant de noter également SVEngine (Xia et al. 2018) et Xome-Blender (Semeraro, Orlandini, and Magi 2018), qui se spécialisent sur la génération de génomes de cellules cancéreuses. Ces deux outils peuvent notamment générer plusieurs clones d'une même tumeur, tout en intégrant les spécifications de l'utilisateur sur les mutations ponctuelles et structurales que ce génome intègre.

6.2.4.2. Approche par intelligence artificielle

L'équipe de Flora Jay à l'Université Paris-Saclay a publié un article (Yelmen et al. 2021) décrivant le projet qu'ils ont mis en place pour tenter de simuler des données génomiques humaines. Ils ont comparé des stratégies basées sur des statistiques simples à des méthodes basées sur l'intelligence artificielle pour montrer le gain de diversité et de réalisme que l'IA est capable d'amener à ces nouvelles données génomiques. Cette approche vise à entraîner l'IA à reconnaître les corrélations présentes entre mutations chez l'humain, puis de simuler de nouvelles configuration de ces mutations. Une description plus détaillée de ces modèles génératifs viendra dans la section 7.1.3.3.: "Réseaux génératifs".

Deux limites se posent tout de même à ce procédé : la quantité de données nécessaires pour composer un génome humain, et la diversité de simulation attendue pour créer de nouveaux génomes vraisemblables. Leur projet porte sur trois jeux de données de mutations ayant des tailles croissantes : ~800 SNPs sur tout le génome décrites comme reflétant la structure de la population dont elle est issue, ~3 000 SNPs sur tout le chromosome 15 utilisés pour des tests de sélection, et 10 000 SNPs sur une région détaillée du chromosome 15 pour explorer de manière

approfondie les performances de leurs modèles. Pour tous ces jeux de mutations, ~5 000 haplotypes du 1000 Genomes Project ont été utilisés pour détecter les associations entre mutations.

Les résultats qu'ils présentent sont bons : leurs outils basés sur des méthodes d'IA créent de nouvelles données ayant une diversité proche de celle du jeu de données de base. Ils montrent que leurs nouveaux individus sont statistiquement dispersés dans leur population originale.

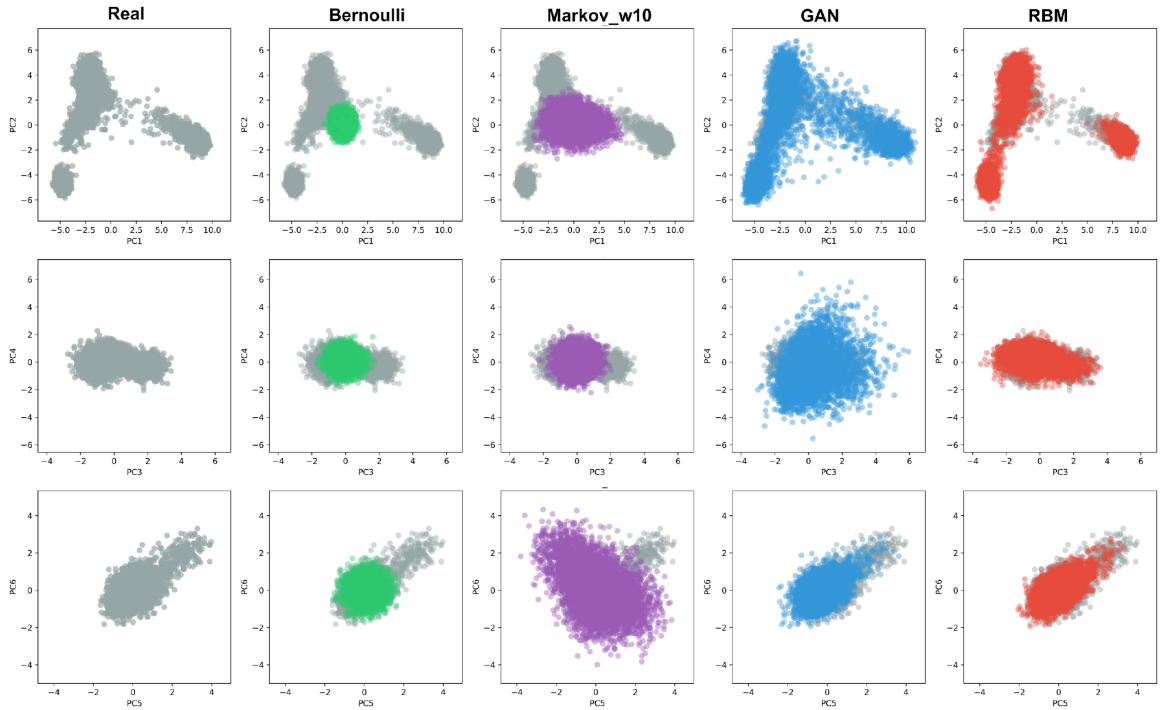


Figure 10 : Projection en Analyse en Composantes Principales (ACP) de la population de référence et celles simulées par diverses méthodes dans (Yelmen et al. 2021). Les auteurs présentent deux méthodes génératives statistiques (vert et violet) et deux méthodes employant l'intelligence artificielle (bleu et rouge) : le Generative Adversarial Network *et la* Restricted Boltzmann Machine.

Les auteurs présentent dans cette figure 10 les diverses populations qu'ils étudient (de référence ou simulée) projetées dans une analyse en composantes principales. Cette technique de visualisation permet de voir en un coup d'œil les caractéristiques de ces diverses simulations en comparaison à la population de référence. Il ne s'agit pas ici d'une métrique statistique montrant une quelconque homogénéité, mais plutôt une indication générale en ce sens.

Les modèles d'intelligence artificielle générative qu'ils emploient sont au nombre de deux : le *Generative Adversarial Network* (GAN) et la *Restricted Boltzmann Machine* (RBM).

Les descriptions détaillées de ces modèles se trouvent dans la section 7.1.3.3, mais en bref, le GAN est un duo de réseau de neurones, un discriminant qui tente de différencier les vrais échantillons des échantillons simulés et un générateur qui tente de tromper le discriminant ; le RBM est un réseau de neurones en deux couches, une visible et une cachée, permettant de générer de nouveaux échantillons en explorant les valeurs que peuvent prendre les neurones de la couche cachée.

Les données qu'ils créent sont aussi réalistes, c'est-à-dire suivent les règles implicites des données fournies en entrée dans les modèles. Plusieurs caractéristiques génomiques sont mesurées pour la population de référence ainsi que pour les simulations réalisées par IA, et ces simulations paraissent vraisemblables.

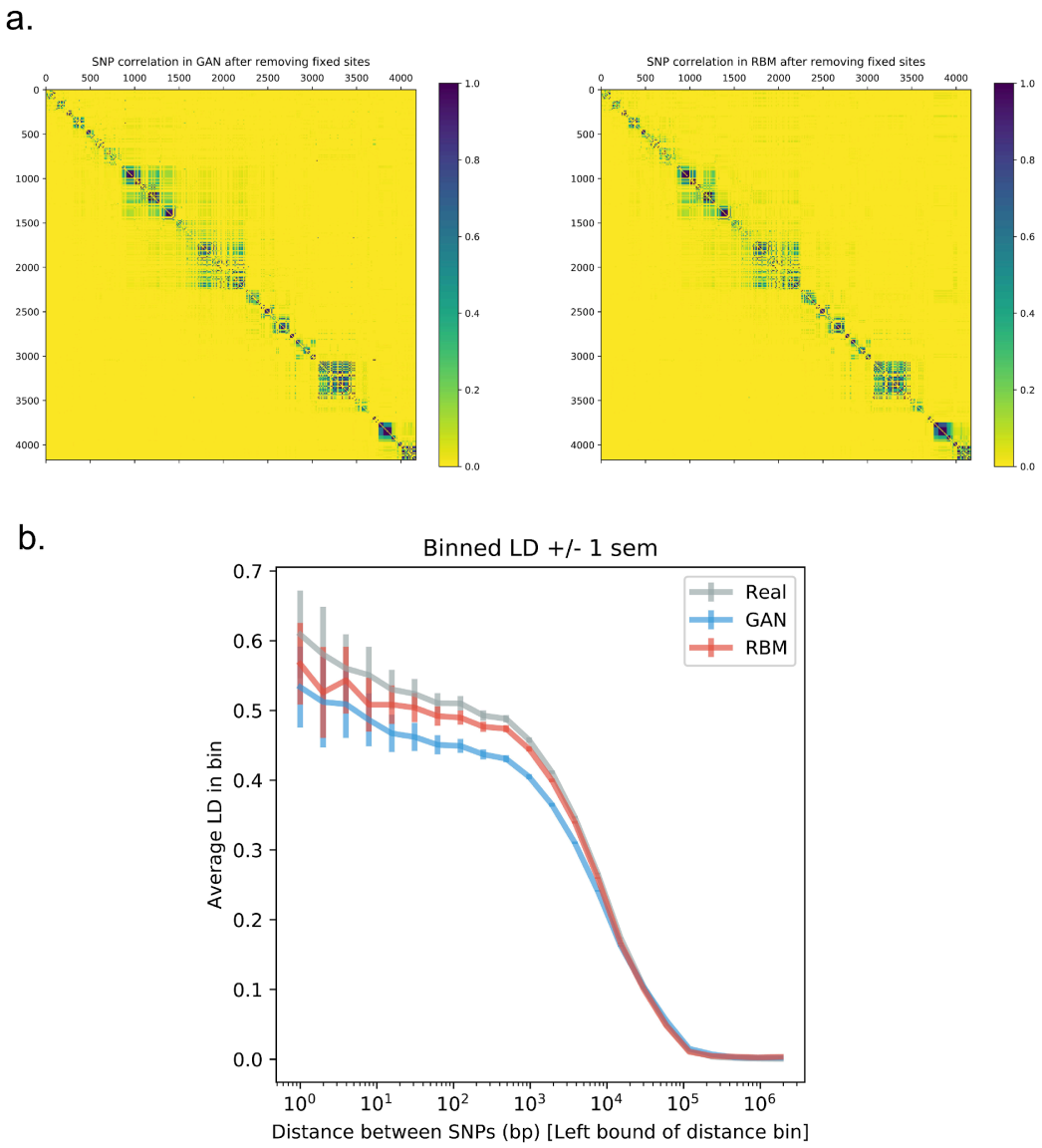


Figure 11 : Mesure du déséquilibre de liaison de la population de référence et celles simulées par IA dans (Yelmen et al. 2021). a. Matrices de corrélation des mutations, la matrice triangulaire inférieure correspond aux données de référence et la supérieure à celles simulées. b. Corrélations entre paires de mutations, organisées en bins de distance.

Les panels de la figure 11 présentent le déséquilibre de liaison entre mutations dans le jeu de données de référence et ces valeurs dans les jeux de données simulées par IA. Les matrices (panel a) présentent la corrélation, exprimée par le r^2, entre paires de mutations, et les courbes (panel b) montrent le déséquilibre de liaison entre ces paires de mutations en fonction de la distance qui les séparent, divisés en bins. Les ressemblances entre ces graphiques pour ces jeux

de données différents soulignent également les similarités de ces populations au niveau des mutations, là où les ACPs ne montrent qu'une vision d'ensemble des données.

Leurs données sont aussi, majoritairement, nouvelles et ne briseraient pas le secret médical des personnes ayant fourni leur génome initialement. Les méthodes utilisées pour les générer sont donc viables en tant que "brouilleur d'identité", qui permettent de reproduire les caractéristiques globales d'un jeu de données génomiques sans faire fuiter d'informations sensibles sur les patients qui la composent.

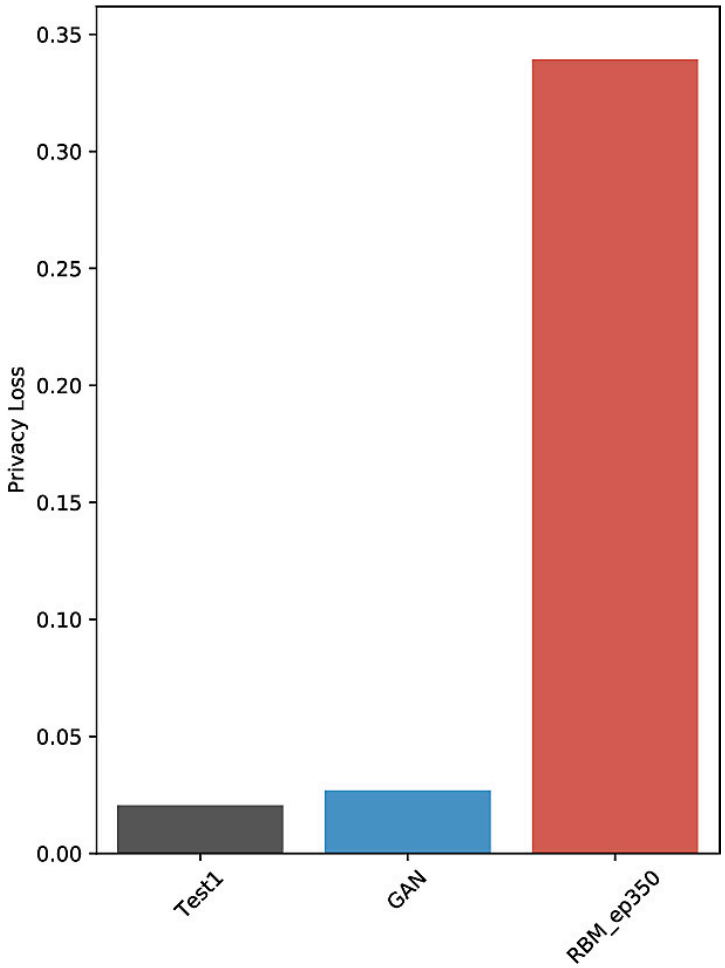


Figure 12 : Mesure de la fuite d'informations causée par les populations simulées par deux méthodes IA dans (Yelmen et al. 2021). La privacy loss compare les distances entre individus simulés ou de référence et leurs voisins les plus proches dans ces deux jeux de données, et mesure à quel point les génomes simulés sont proches des génomes de référence. Test1 est un jeu de données de génomes de références exclus du processus de construction et d'entraînement des modèles.

Dans la figure 12, les auteurs présentent une nouvelle métrique, la *Privacy Loss*, qui

mesure la quantité d'information personnelle qui fuite dans le jeu de données simulé à partir de ces informations personnelles. La formule exacte et l'interprétation détaillée de cette valeur se trouvent dans (Yelmen et al. 2021), mais en bref, elle provient de la comparaison de distances entre individus de référence et/ou simulés et leur plus proches voisins dans le même jeu de données ou dans l'autre. Les simulations du GAN semblent ne laisser fuiter aucune information personnelle du jeu de données original, présentant une valeur de *Privacy Loss* comparable au contrôle négatif. Les données produites par la RBM, elles, semblent contenir énormément d'informations qui pourraient porter atteinte à l'identité des échantillons originaux. Il serait intéressant d'avoir un contrôle positif, par exemple un assemblage de sous-section des données présentes dans le jeu de données original, pour comprendre à quel point le RBM laisse vraiment fuiter les informations personnelles.

D'après ces résultats, il est clair que ces méthodes basées sur l'IA montrent une performance hautement renforcée par rapport aux méthodes statistiques. Tous ces points sont clés dans l'obtention d'une population simulée suffisamment réaliste pour permettre sa réutilisation dans d'autres projets.

La plus grande limite à ces résultats, c'est leur envergure. Le plus grand jeu de données de mutations employé dans (Yelmen et al. 2021) représente moins de 0.1% d'un génome humain entier en termes de paires de bases, et environ 0.2% des mutations identifiées dans (The 1000 Genomes Project Consortium 2015). Cette limite de la quantité de données que représente chaque échantillon n'est pas surprenante, car un génome humain est un objet très complexe.

Le projet de (Yelmen et al. 2021) représente une preuve de concept quant à la génération de nouveaux génomes par IA : ces outils sont bien capables de détecter les motifs intrinsèques d'un jeu de données génomique, et de les reproduire avec diversité et réalisme, sans se contenter de simplement reproduire les échantillons vus dans ce jeu de données. Il faut maintenant se pencher sur la question de la mise à l'échelle, en augmentant le nombre de mutations que ce modèle peut générer, afin de produire un modèle capable de créer des génomes humains complets, ou presque.

7. État de l'art

7.1. l'Intelligence Artificielle

7.1.1. L'intelligence artificielle en général

Le terme intelligence artificielle fait référence à un ensemble de méthodologies où une machine prend des décisions basées sur des données situationnelles pour tenter de créer un résultat optimal (McCorduck 2018).

Cette définition reste très globale et ne permet pas de saisir la profondeur de cette discipline. La question est tellement large qu'il devient difficile, voire impossible, de formuler une définition qui mette tous les utilisateurs d'intelligence artificielle et experts du domaine d'accord. Même inclure le terme "intelligence" est parfois source de controverse, dû aux questionnements philosophiques qui surviennent suite à ce mot (Searle 1980).

Lorsqu'on lit le mot "intelligence", un premier réflexe assez naturel est de se dire que le concepteur cherche à recréer les mécanismes de l'intelligence qu'on peut observer chez l'Humain (ou parfois des colonies d'insectes (Dorigo and Gambardella 1997)). Cette inférence d'idées place donc énormément de préjugés sur ce que l'IA devrait être, et de quelles caractéristiques elle devrait être dotée : mémoire, raisonnement, opinions, bon sens, éthique, philosophie, etc. Cette liste n'est ni exhaustive ni définitive, puisque chaque personne peut avoir une conception différente de ce que représente l'intelligence, mais elle existe très certainement pour la majorité des personnes.

La réalité est que créer une intelligence artificielle qui puisse réaliser toutes ces tâches en simultané relève plutôt de la fiction. Les processus mis en place par les concepteurs de modèles IA tentent de reproduire une compétence bien précise qu'on peut considérer comme intelligente, mais créer des ordinateurs capables de réfléchir tel que le fait un Humain n'est pas l'objectif recherché pour la plupart des projets. Pour prendre l'exemple d'une division posée, enseignée à l'école depuis des décennies, un Humain aura besoin de réfléchir pour résoudre ce problème complexe. Un ordinateur, quant à lui, peut résoudre cette tâche bien plus facilement, mais cette résolution n'implique nullement une réflexion de sa part.

Le nom “Intelligence Artificielle” a été choisi par John McCarthy en 1956, à la base pour mettre d’accord plusieurs groupes de chercheurs en mathématiques sur la nature de la discipline (l’ironie est parfois délicieuse). Dans la foulée, il prépare pour l’été suivant un projet de brainstorming d’environ deux mois avec des collègues dispersés dans les États-Unis pour discuter de la direction que devraient prendre les recherches dans le domaine (McCarthy and Minsky 1955). Quelques principaux sujets évoqués sont le traitement de langages naturels, la résolution de calculs combinatoires, l’abstraction de notions, et la créativité. Ces secteurs sont tous, aujourd’hui encore, d’actualité.

Un phénomène intéressant qu’on peut observer dans le domaine de l’IA est ce qu’on appelle l’effet IA : dès qu’un problème qui paraît complexe est résolu par une méthode basée sur l’intelligence artificielle, alors on considère rapidement cette solution comme triviale. Un bon exemple est la reconnaissance d’écriture manuscrite par IA. Cette tâche a été pendant longtemps un but lointain et difficile à réaliser, jusqu’à ce qu'un outil IA s’impose comme la référence en la matière en 2009 (Angelica and Schmidhuber 2012). Depuis, cette fonctionnalité est intégrée dans de très nombreux objets du quotidien et permettent à n’importe qui, n’importe quand de scanner un texte manuscrit avec une grande précision. Cependant, malgré son omniprésence, cette réalisation ne suscite guère de débats dans les discussions sur l’utilisation de l’IA. C’est bien que cette tâche n’est plus considérée comme nécessitant une intelligence.

Une tendance similaire peut être observée dans le domaine du comportement animal, où persiste une certaine forme d'élitisme humain remontant à l'époque des rois chrétiens au Moyen Âge. Auparavant, des compétences telles que l'utilisation d'outils ou la compréhension des langues étaient considérées comme des indicateurs d'intelligence. Cependant, des expériences multiples ont depuis démontré que ces capacités existent également chez les animaux. Malgré ces constatations, de nombreuses personnes, qu'elles soient issues du monde académique ou non, tentent de redéfinir l'intelligence en utilisant d'autres critères afin de maintenir un sentiment de supériorité (Colbert-White and Kaufman 2023).

7.1.1.1. Histoire

On retrouve depuis l’antiquité des œuvres de fiction et de philosophie décrivant des objets d’habitude inanimés dotés d’intelligence (Talos, un géant de bronze forgé par Héphaïstos dans la

mythologie grecque, en est le premier exemple). Lorsque dans les années 1940 l'ordinateur moderne et programmable est apparu dans les premiers laboratoires, il n'a pas fallu longtemps pour qu'on imagine que cette machine pensante puisse devenir une réalité. La première démonstration d'intelligence artificielle dans un milieu académique fut en 1956 avec le "Logic Theorist" de Newell, Simon et Shaw (Newell and Simon 1956), qui a pu démontrer de nombreuses lois logiques formulées dans Principia Mathematica (Whitehead and Russell 1913), certaines de manières nouvelles que les chercheurs humains n'avaient pas trouvé. Le premier succès commercial de l'IA fut dans les années 1980 avec les systèmes experts (Jackson 1998), des bases de connaissances permettant aux utilisateurs d'extrapoler des informations sur des objets ou des situations à partir de règles pré-établies par des experts humains.

Depuis ces innovations, l'intérêt à la fois scientifique et économique dans l'IA a beaucoup fluctué (McCorduck 2018), notamment dû aux grandes promesses que les chercheurs du domaine ont réalisé quant à son potentiel. Il a fallu attendre les années 1990 lorsque la puissance de calcul des ordinateurs et l'accès à internet ont explosé, permettant à la fois l'accès à de grandes sources de données et le traitement de ces données en temps abordable (Kurzweil 2006). Ces révolutions technologiques ont aussi été accompagnées d'innovations méthodologiques dans le domaine, avec par exemple l'apprentissage automatique et les réseaux de neurones (LeCun, Bengio, and Hinton 2015). Ces méthodes sont développées dans les sections 7.1.2. "Apprentissage automatique" et 7.1.3. "Réseaux de neurones", respectivement.

7.1.1.2. Fonctionnement

Le terme d'intelligence artificielle regroupe donc les méthodes et algorithmes permettant d'appliquer un processus "intelligent" réalisé par un ordinateur à des tâches qui sont d'habitude réservées à des humains. La notion d'intelligence ici comprend la collection d'information, leur synthèse, ainsi que l'inférence et la prise de décision à partir de ces informations. Cette prise de décision repose en général sur une hiérarchisation des tâches et des connaissances, étant donné la situation actuelle dans laquelle se trouve l'agent intelligent.

Les premières approches d'intelligence artificielle reposent sur des règles et lois fournies par des humains, alors que les approches qui ont émergé dans les décennies après reposent sur l'analyse d'information sans contrôle direct par un opérateur humain.

7.1.1.3. Applications

L'outil "Logic Theorist" (Newell and Simon 1956) a reçu un accueil assez tiède de la part de la communauté scientifique spécialisée de l'époque, même s'il représentait une étape majeure dans le domaine de l'intelligence artificielle. Malgré les découvertes réalisées par cette machine, elle n'a pas reçu la reconnaissance que ses créateurs estimaient qu'elle méritait. Cependant, les avancées qu'il a apportées sont nombreuses. Les tâches qu'il a accomplies sont considérées comme intelligentes et créatives, et ce fut la première fois qu'un automate parvint à les réaliser.

Les systèmes experts ont eux connu une grande vague de popularité, lorsque plusieurs grandes entreprises ont vu leur intérêt dans l'automatisation de certaines tâches qui d'habitude nécessitent un apport humain. Leur première implémentation à succès était dans la conception de portes logiques pour processeurs. Plusieurs de ces portes et combinaisons de portes conçues par le système expert étaient plus efficaces que les solutions proposées par des experts humains.

Malgré leur succès initial, leur popularité n'a pas duré très longtemps (McCorduck 2018). Deux blocages majeurs à leur survie au long terme étaient :

1. La nécessité de maîtriser le côté informatique de la base de connaissances. Les systèmes experts étaient beaucoup vendus comme utilisant un langage facile à comprendre par tout le monde, et donc ne nécessitant pas de formation en langages de programmation pour renseigner les règles de logique de la base. Cependant, il fallait tout de même respecter des formulations précises ainsi qu'utiliser des signes de ponctuation spécifiques, ce qui n'est pas naturel pour la plupart des utilisateurs non-informaticiens.
2. Le besoin de l'apport d'experts. Plus le domaine était pointu, plus les experts capables de mettre en place les règles du système étaient difficiles à convaincre pour effectuer le travail fastidieux de lister toutes les règles qui gouvernent leur sujet d'étude, sans parler des éventuelles mises à jours à effectuer sur les informations contenues dans la base de données.

Ces deux applications sont des exemples de tâches basées sur la connaissance et les règles. Une autre application populaire dans ce domaine est l'organisation d'emplois du temps et de plannings, lorsque des tâches doivent s'enchaîner dans un ordre important.

Une approche semblable aux connaissances concrètes est l'utilisation de logique floue, en

tentant d'imiter les prises de décisions d'humains. Là où un ordinateur classique évalue des propositions logiques de manière booléenne (1 ou 0, vrai ou faux), il a été observé que les décisions du quotidien prises par des humains a plus tendance à être basé sur des règles définies de manière approximative. Ce raisonnement vise donc à créer ces ordinateurs pouvant raisonner de manière semblable, en représentant des concepts et en construisant des associations probabilistes.

Un autre domaine d'application de méthodes IA est l'interaction avec le monde physique, lorsque des automates comme des robots doivent faire preuve de comportements raisonnés. En plus de naviguer dans ce monde, ces acteurs sont aussi amenés à interagir avec, sous des formes parfois bien différentes.

Le domaine d'application le plus populaire depuis plusieurs années est l'apprentissage automatique, qui a su s'imposer comme figure de proue du mouvement de l'intelligence artificielle. Sa popularité est telle que ses méthodologies associées sont très souvent appliquées aux domaines cités ci-avant, avec un succès généralement remarquable. C'est avec cette famille de méthodes que je travaille pour ce projet, et donc je vais procéder à une explication détaillée de celle-ci.

7.1.2. Apprentissage automatique

Il existe plusieurs sous-domaines de l'intelligence artificielle, notamment l'apprentissage automatique, qui ont connu un grand essor ces dernières années. Ce domaine repose sur l'ajustement des paramètres de décisions d'une IA réalisé par une heuristique prédéterminée pour tenter de l'optimiser, avec peu voire pas d'intervention humaine une fois le processus d'apprentissage enclenché (LeCun, Bengio, and Hinton 2015; McCorduck 2018).

Dans la section précédente, j'ai décrit des approches basées sur des connaissances décrites explicitement. Même si l'apprentissage automatique a connu un grand essor ces dernières années, cette tâche spécifiquement pose problème dans ce domaine. Il est en général compliqué, voire impossible selon le type de modèle, de spécifier des règles et vérités formelles que l'outil se doit de respecter. Ces approches souhaitent imposer un minimum de contraintes à l'algorithme d'entraînement, et donc on tente de réduire les *a priori* humains que les utilisateurs peuvent ajouter, de peur de biaiser cet entraînement..

7.1.2.1. Histoire

Les premières expériences d'ordinateur qui apprennent d'eux-mêmes datent des années 1960, lorsque Raytheon tente d'implémenter Cybertron pour analyser des signaux sonores (Time Magazine 1961). Sa méthodologie d'apprentissage ressemble beaucoup à ce qu'on appellerait aujourd'hui de l'apprentissage par renforcement. Cette stratégie est développée dans la section 8.2.2.3 : "Apprentissage par renforcement".

L'intérêt scientifique en ces méthodes ne s'était jamais totalement éteint, mais les systèmes experts et leur base de connaissances explicites avaient pris le devant de la scène pendant les années 1980. Dans ce contexte, il a fallu attendre leur amenuisement pour qu'on s'intéresse de plus près à l'apprentissage automatique.

7.1.2.2. Fonctionnement

L'heuristique générale mise en jeu pendant l'apprentissage automatique est de laisser l'algorithme trouver la meilleure solution pour un problème donné, sans que l'utilisateur humain précise les optimisations exactes à effectuer sur les paramètres de décision du modèle. La notion de "meilleure solution" dépend bien sûr de l'objectif fixé par l'utilisateur (qui peut être biaisé par des informations établies *a priori*), et elle est entièrement dépendante sur la quantité et la nature des données disponibles pour l'entraînement de l'outil. Il est admis que la majorité des problèmes à résoudre présenteront des solutions locales que l'algorithme peut préférer à l'exploration du reste de l'espace des solutions, ainsi que des solutions très spécialisées pour le jeu de données étudié lors de l'entraînement mais qui se généralisera assez mal sur un autre jeu de données. Les méthodes mises en place par le concepteur pour tenter d'éviter ces phénomènes sont guidées par des hyperparamètres d'entraînement. Les **hyperparamètres** sont tous les critères de l'heuristique mis en place par l'expert humain en amont de l'entraînement, et qui définiront la recherche des meilleurs **paramètres** du modèle. Le nombre et la nature exacte des hyperparamètres dépend du type de modèle implémenté et des données à traiter.

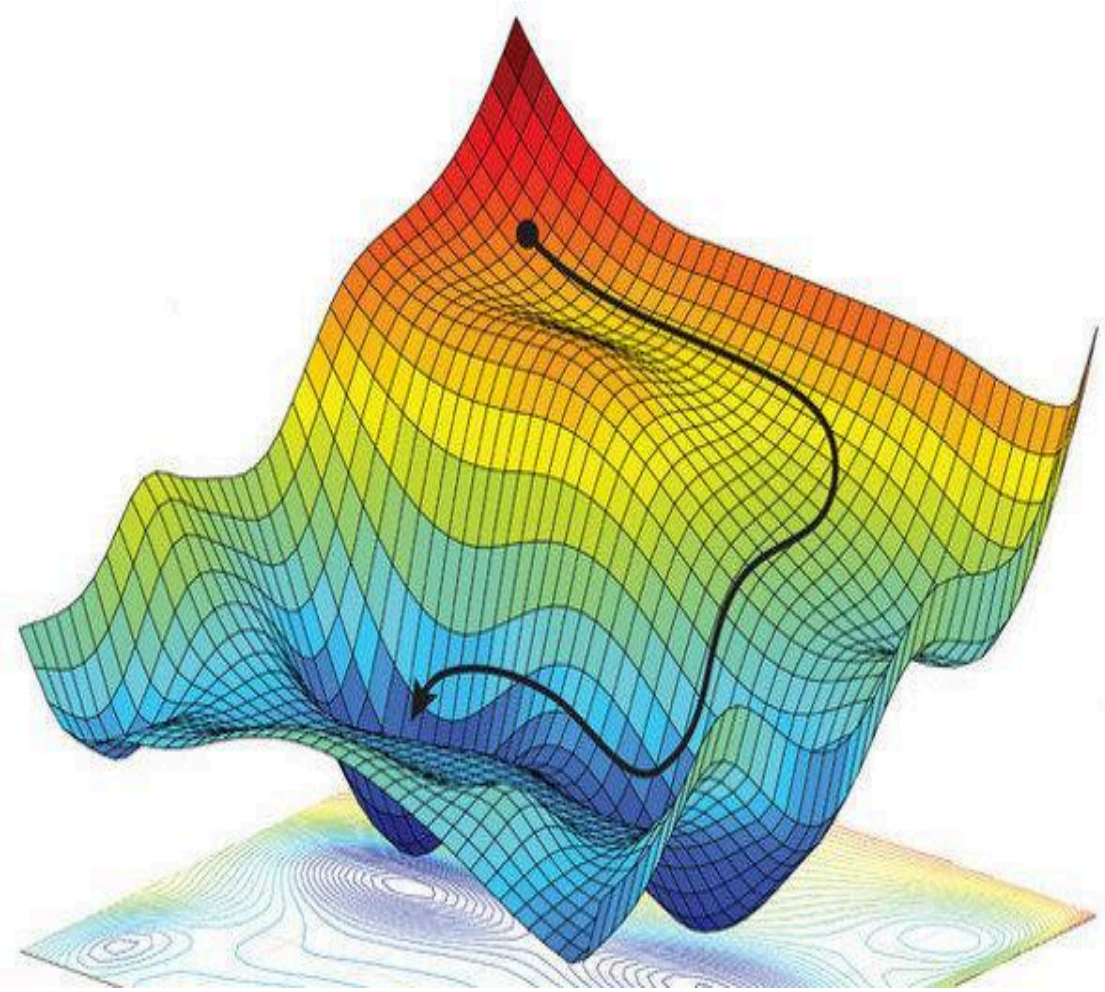


Figure 13 : Visualisation d'un espace de solutions pour un algorithme d'apprentissage automatique. Les dimensions horizontales représentent des paramètres du modèle, et l'axe vertical l'erreur que le modèle produit avec ces paramètres. La flèche indique le "chemin" d'apprentissage que l'on souhaite que le modèle effectue. Tiré de (Amini et al. 2019).

Lorsqu'il est possible mathématiquement, une méthode très efficace pour déterminer ce "chemin", illustré dans la figure 13, à emprunter dans l'espace des solutions est la descente en gradients (Curry 1944). Cette méthode emploie les dérivées premières (ou leurs approximations par échantillonnage local) pour suivre une pente et ainsi identifier un minimum local.

Plusieurs stratégies peuvent être mises en place pour tenter d'optimiser cette recherche, en général basées sur la quantité d'information connue au préalable ainsi que la tâche finale à compléter. Le recuit simulé, nommé en référence à la pratique utilisée en métallurgie, est parmi les plus populaires (Tsallis and Stariolo 1996). Dans ce cadre, l'IA est capable de modifier fortement ses paramètres de décision au début de son entraînement, puis cette capacité diminue

au cours du processus d'apprentissage. Cette logique cherche à permettre à l'IA d'identifier les éléments principaux des décisions à prendre au départ, puis de se spécialiser au sein de ces grandes règles.

Un élément important à déterminer est la méthode de calcul de l'erreur que fait le modèle, vu que celle-ci va déterminer l'apparence du "paysage" d'apprentissage. Celle-ci dépend bien sûr du type de données qu'on cherche à prédire (binaire, catégorielle, numérique, texte, …) ainsi que leur amplitude, mais il faut aussi prendre en compte quelles erreurs on souhaite pénaliser le plus. Il est possible qu'on souhaite tout pénaliser égalitairement, mais bien souvent certaines erreurs sont moins souhaitées que d'autres. Par exemple, si on cherche à entraîner un modèle pour conduire une voiture, il est moins grave de ralentir devant un feu vert que d'accélérer lorsque ce même feu est rouge.

Très souvent, la solution la plus simple est d'utiliser une distance, qui est une application qui fixe une valeur numérique positive à deux ensembles, ce qui représente la longueur entre deux points lorsque ces ensembles correspondent aux coordonnées de ces points dans un espace. Une métrique très répandue est la distance euclidienne (c'est-à-dire la racine carrée de la somme des carrés des différences) pour des données numériques continues, qui en général capte bien la divergence entre les prédictions et la réalité. Il existe aussi de très nombreuses autres fonctions communément utilisées pour mieux capter la distribution des valeurs émises par le modèle, plutôt que chaque valeur individuelle.

Pour des données catégorielles ou binaires, l'exactitude mesure tout simplement le nombre de prédictions correctes divisée par le nombre de prédictions à effectuer. Selon le type de données, il est parfois également nécessaire de fournir d'autres données sur ce que l'utilisateur considère comme un prédiction correcte (nombre de classes à tester, seuil pour attribuer une valeur binaire, etc). Il existe aussi bon nombre de mesures de distances entre vecteurs de valeurs binaires, telle que la distance de Jaccard (Murphy 1996). Pour des données catégorielles, des mesures de dissimilarité existent (Boriah, Chandola, and Kumar 2008), mais sont moins répandues et dépendent de beaucoup de facteurs, comme le nombre de classes différentes ou si certaines classes sont plus proches entre elles que d'autres.

Cette décision revient à l'expert humain qui paramètre le système initial. Le choix doit se faire à la fois en se basant sur des savoirs préétablis et sur l'expérimentation avec le modèle. Il

est très difficile de savoir à l'avance quelle mesure d'erreur donnera la meilleure performance après entraînement, c'est pourquoi il faut souvent s'attendre à réaliser ce processus plusieurs fois.

Il faut également permettre au modèle de réaliser deux comportements clés pendant l'entraînement : l'exploration et l'exploitation. Tout modèle a besoin de réaliser ces étapes pour s'optimiser, mais ils sont contraires en termes d'objectif.

L'exploration consiste à collecter plus d'informations sur les solutions plus ou moins proches de celle considérée comme actuellement la meilleure. Ce comportement permet au modèle de mieux cerner quelles stratégies peuvent fonctionner, et ainsi préciser le chemin qu'il pourrait prendre dans ses futures itérations. Par exemple, un système de recommandations pourrait suggérer des éléments très différents de ce que l'utilisateur a l'habitude d'apprécier. Ceci n'est pas effectué dans le but de réaliser une suggestion optimale, mais plutôt d'en savoir plus sur la réaction de l'utilisateur face à cet élément nouveau.

L'exploitation est à l'opposé : le modèle cherche ici à prédire la solution qu'il considère comme optimale et à s'en rapprocher le plus vite que possible. Il maximise ainsi les gains à court terme, en s'ouvrant à la possibilité de se rapprocher d'un maximum local. Pour le système de recommandation, ce comportement pourrait être représenté par la suggestion d'un élément déjà approuvé dans le passé par l'utilisateur, pour garantir une réponse positive de sa part.

Un élément important à prendre en compte lors de l'établissement d'une stratégie d'apprentissage automatique est la "malédiction de la dimension". Ce terme désigne une série de phénomènes qui surviennent lorsqu'on tente d'analyser des jeux de données comportant de plus en plus de dimensions, sans augmenter le nombre d'échantillons pour illustrer ces dimensions. Ces phénomènes apparaissent aussi dans d'autres domaines qui traitent de grands jeux de données comme la gestion de bases de données.

Bien évidemment, avec la taille du jeu de données, on voit aussi augmenter la puissance et le temps de calcul nécessaires pour le traiter, mais d'autres problèmes surviennent aussi, souvent sans solution directe. Plus le nombre de dimensions augmente, et plus les échantillons contenant des informations deviennent éparse, c'est-à-dire ont de moins en moins de chance de contenir de l'information pertinente pour cette dimension. Ceci entraîne beaucoup de problèmes pour tenter de trouver des individus caractéristiques dans une population, notamment si on cherche à inclure une étape d'échantillonnage. De plus, il devient difficile d'utiliser des mesures

de distances simples comme les distances euclidiennes pour mesurer la similarité entre échantillons, puisque la différence entre les distances maximales et minimales qu'on peut trouver entre échantillons se réduit.

7.1.2.2.1. Apprentissage supervisé

Lors d'un apprentissage supervisé, l'utilisateur spécifie à l'algorithme d'apprentissage automatique un but spécifique qu'il doit atteindre pour chaque échantillon analysé. Ce but spécifique est en général une valeur ou une étiquette rattachée à l'échantillon, attribuée par un expert humain.

Cette méthodologie s'applique bien pour des méthodes de prédiction, où on enseigne au modèle quel facteur prédire à partir des données d'entrée. La prédiction peut être une classification, où l'échantillon porte une étiquette qualitative, ou une régression, où la cible à prédire est numérique. Deux jeux de données classiques illustrent bien ces deux cas d'utilisation. Le MNIST (LeCun, Cortes, and Burges 2011) est un jeu de données d'images de chiffres manuscrits, et sert très souvent de premiers test pour des modèles d'analyse d'écriture. Chaque image porte une étiquette indiquant quel chiffre y figure. De nombreuses méthodes ont été testées dessus depuis sa publication. À l'inverse, il existe de nombreux jeux de données listant des caractéristiques de maisons comme leur superficie ou si elle est équipée de climatisation (OECD 2023). Toutes ces maisons portent aussi le prix de vente associé, qui est une cible très intéressante pour un modèle d'apprentissage automatique.

Ces méthodes se rapprochent de méthodes statistiques, mais un avantage à utiliser l'apprentissage automatique est sa capacité à extraire l'information de données à forte dimensionnalité et de manière non-linéaire.

7.1.2.2.2. Apprentissage non-supervisé

Au contraire, l'apprentissage non-supervisé ne fournit pas un but spécifique au modèle, mais plutôt une métrique d'évaluation qui permet de guider le modèle vers une solution convenable (Kosko 1990). Cette liberté lors de l'entraînement permet de laisser libre cours à la méthode d'apprentissage pour découvrir des associations entre valeurs, là où des approches supervisées peuvent apporter des biais humains dans les résultats finaux.

De nombreuses méthodes de classification et de partitionnement sont non-supervisées et permettent à l'outil de déterminer la répartition finale des objets en classes ou partitions, tout en suivant les règles fixées sur le nombre de ces classes ou leur étendue maximale. Cette caractéristique est très utile notamment dans le partitionnement, qui consiste à réunir les éléments d'un jeu de données ayant une forte ressemblance entre elles. Là encore, il est possible d'employer des méthodes statistiques pour effectuer cette tâche, mais l'apprentissage automatique peut plus facilement séparer des données selon des règles non-linéaires. Par exemple, un partitionnement hiérarchique, qui repose uniquement sur les distances entre objets, se montre efficace sur certains jeux de données, mais ne pourra pas scinder certains groupes là où DBSCAN (Ester, Kriegel, and Xu 1996), qui prend en compte la densité en objets d'une région de l'espace, pourra effectuer un partitionnement plus proche des attentes de l'utilisateur. Il montre aussi des forces pour reconnaître des variations à plusieurs échelles, là où d'autres méthodes peuvent éclipser des variations locales lorsque les variations globales sont très importantes.

Cette approche permet aussi la détection d'anomalies (Barnett and Lewis 1978). La méthode la plus efficace pour implémenter cet objectif est d'avoir un jeu de données hautement travaillé contenant uniquement des échantillons normaux. Après entraînement sur ce jeu de données, il est assez simple pour bon nombre de modèles différents d'identifier si de nouveaux échantillons appartiennent ou non à cette norme préétablie. Il est aussi possible de sélectionner des anomalies dans un jeu de données directement, mais il peut être plus compliqué de le faire efficacement. La réussite de la tâche dépend de la quantité de données normales dans le jeu de données, ainsi que le nombre de classes présentes et les différences entre celles-ci.

7.1.2.2.3. Apprentissage par renforcement

Les stratégies d'apprentissage par renforcement ont un but souvent un peu plus vague, semblable aux stratégies non-supervisées. La grande différence est que l'évaluation de la performance du modèle se fait en décalé (Fu 1970), et donc l'entraînement et l'optimisation se font avec un délai intrinsèque. Les algorithmes reposant sur l'apprentissage par renforcement sont particulièrement adaptés pour la création d'automates de jeu, par exemple aux échecs (Silver et al. 2017). Les conséquences d'un coup donné peuvent se faire sentir plusieurs dizaines de

coups plus tard, c'est pourquoi il est difficile de les évaluer en temps réel.

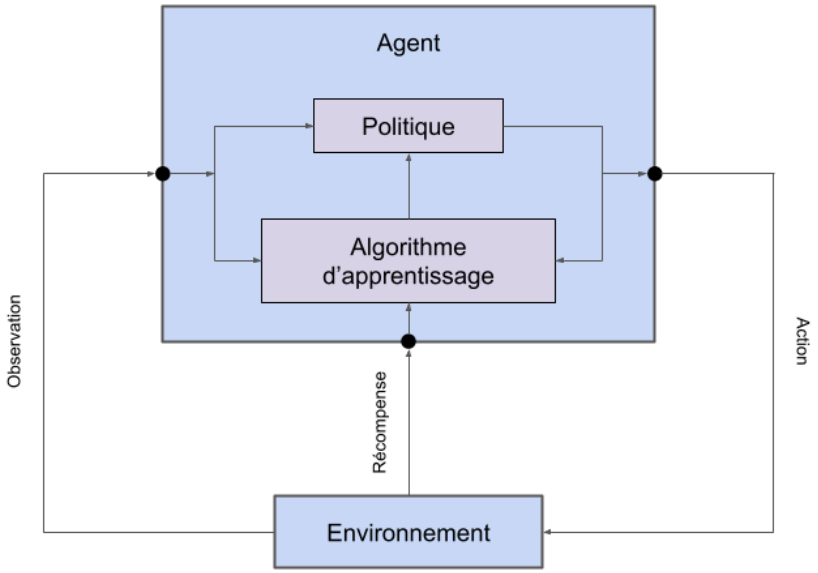


Figure 14 : Schéma de la boucle d'apprentissage d'un système employant l'apprentissage par renforcement.

La figure 14 schématise les différents éléments qui composent une IA employant un apprentissage par renforcement. La politique d'un tel agent représente les heuristiques guidant ses décisions. Elle peut être initialisée aléatoirement ou en se basant sur des informations externes. L'agent commence par interagir avec son environnement pour lui apporter des modifications et collecter des informations dessus, qu'il stocke. Ensuite, en se basant sur l'ensemble des situations qu'il a rencontrées et les résultats des actions qu'il a effectué sur ces situations, il tente de prédire l'issue de la prochaine action qu'il va effectuer sur son environnement. L'agent effectue cette action (ou plusieurs, si cela est permis), et mesure la différence entre sa prédiction et l'effet réel de son interaction, qui impactera ses futures prédictions.

Il y a donc deux phénomènes d'apprentissage qui ont lieu en simultané dans ce système. Les observations que fait l'agent l'informent sur la qualité de ses prédictions, et les récompenses qu'il reçoit lui permettent de se rapprocher de l'action optimale à effectuer dans son environnement. Ces deux fonctions sont essentielles pour atteindre son but final.

7.1.2.3. Applications

L'apprentissage automatique a su s'imposer dans les dernières décennies en employant notamment les progrès des technologies calculatoires à son avantage. De très nombreuses implémentations de ces méthodes existent aujourd'hui. En mathématiques, les analyses de régression et les machines à vecteurs de support font appel à un apprentissage automatique pour approximer des fonctions (Cortes and Vapnik 1995). Les arbres de décision sont une méthode de classification très efficace et employable dans tous les domaines imaginables (X. Wu et al. 2008). Les algorithmes génétiques sont une famille de méthodes permettant de simuler un processus de sélection naturelle sur des fragments de code pour tenter d'en trouver une configuration optimale (Mitchell 1998). Ce processus peut aussi s'appliquer sur des données pour servir d'étape de réduction de dimension, dont les enjeux sont détaillés dans la section 7.2 : "Réduction de dimension par IA". Cette application a déjà fait ses preuves sur des données biologiques complexes (Barriere et al. 2019).

Une utilisation très efficace de l'apprentissage automatique est l'apprentissage ensembliste. Cette approche vise à trouver une solution générale très optimisée en se basant sur de multiples tentatives qui sont individuellement plus faibles et moins précises. Appliquée à l'utilisation d'arbres de décision, on appelle cette pratique le *random forest* (Ho 1995), ou forêt aléatoire en français. Une force de cette approche est la parallélisation, puisque l'entraînement de chaque sous-modèle peut se faire indépendamment.

Récemment, la méthode XGBoost (pour *eXtreme Gradient Boosting*, ou amplification extrême des gradients en français) (T. Chen and Guestrin 2016) a eu un fort impact sur l'écosystème des méthodes d'apprentissage automatique. Elle permet une forte accélération de l'apprentissage, notamment en calculant les dérivées secondes de l'espace dans lequel l'automate apprend pour mieux cerner sa topographie. Elle implémente aussi plusieurs autres fonctionnalités pour accélérer ses calculs, par exemple en approximant des quantiles informatifs de la distribution des valeurs en entrée, en élaguant les branches peu importantes de ses arbres, ou encore en utilisant de la validation croisée sur ses arbres pendant l'entraînement pour mieux évaluer la performance des arbres.

Une des méthodes ayant connu le plus de succès est le réseau de neurone artificiel, de par

sa flexibilité et facilité d'implémentation. Cette approche est tellement polyvalente qu'elle s'est même implantée dans des domaines où précédemment d'autres stratégies d'apprentissage automatique semblaient s'être imposées.

7.1.3. Les réseaux de neurones

Une de ces méthodes d'apprentissage est l'utilisation de réseaux de neurones artificiels qui s'inspirent à la base de la structure du cerveau animal. Chaque neurone reçoit en entrée deux ou plus valeurs, multiplie chacune de ces valeurs par un poids, puis calcule une information unique à transmettre à sa sortie, basée sur une fonction d'activation (McCulloch and Pitts 1943). Ce fonctionnement est illustré dans la figure 15.

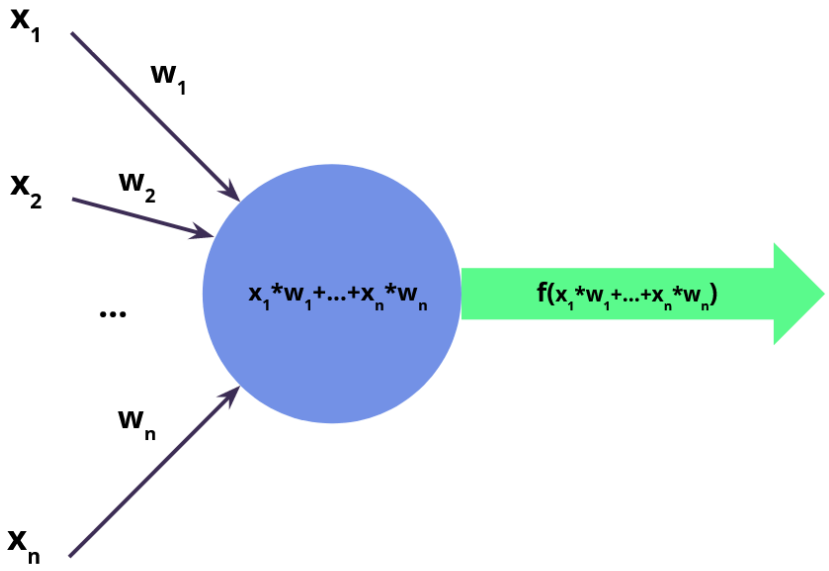


Figure 15 : Représentation schématique d'un neurone artificiel. Les données d'entrée sont notées [x1, …, xn], les poids assignés à chaque connexion [w1, …, wn], et la fonction d'activation f permet de calculer l'information à transmettre.

La fonction d'activation qu'utilise un neurone est un facteur important dans les prédictions que celui-ci peut réaliser. Une des raisons pour lesquelles la méthode n'a pas eu de succès à ses débuts est qu'un neurone basique ne peut effectuer que des prédictions linéaires. En introduisant des fonctions non-linéaires dans ces modèles, cela a permis de réaliser des prédictions non-linéaires, et donc d'ouvrir la méthodologie à un spectre bien plus large d'applications possibles. Le coût en puissance de calcul nécessaire pour entraîner un réseau

entier a également retardé sa popularité (McCorduck 2018).

Même s'il existe en théorie une infinité de fonctions d'activations utilisables pour la sortie d'un neurone, on peut en citer quelques-unes qui sont parmi les plus populaires. Ces graphiques sont tous adaptés de ('Activation Function' 2023) (principalement pour sa charte graphique). L'axe horizontal représente la valeur que la fonction reçoit en entrée, et l'axe vertical la valeur produite par la fonction.

La fonction linéaire vaut f(x)=mx pour tout réel x, avec m étant une constante. Elle est représentée dans le panel A de la figure 16.

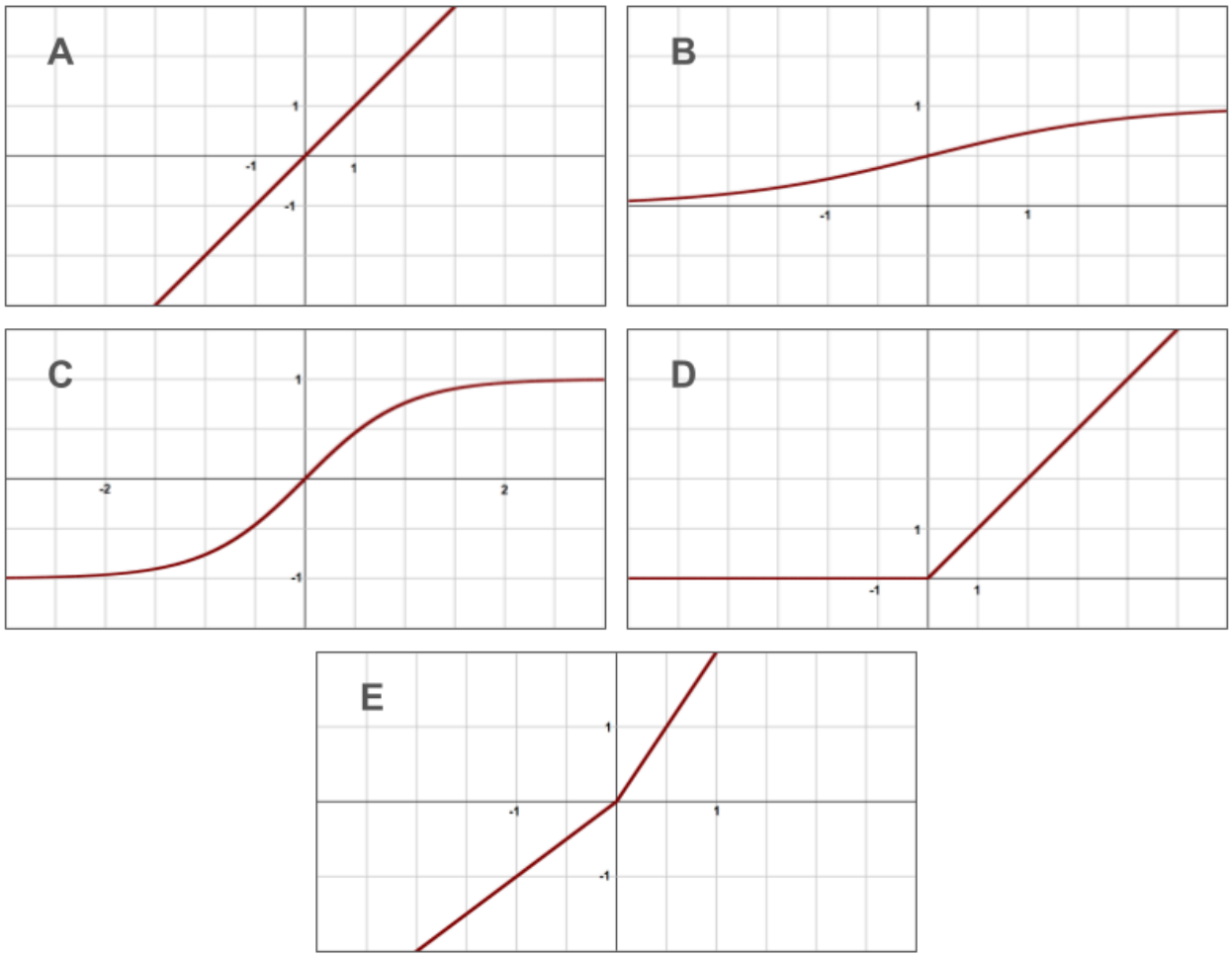


Figure 16 : Fonctions d'activation de neurones communément employés aujourd'hui. Dans l'ordre : linéaire, sigmoïde, TanH, ReLU, LeakyReLU.

Elle est la toute première fonction d'activation utilisée, et elle est très limitée dans ses applications pratiques. Les fonctions non-linéaires offrent un potentiel bien plus développé.

Vu que la dérivabilité de la fonction d'activation est un point clé de son utilisation dans les réseaux de neurones, les prochaines fonctions testées étaient sélectionnées sur cette caractéristique.

La fonction logistique, souvent appelée sigmoïde, vaut $1/(1+e^{-x})$ pour tout réel x. Elle est illustrée dans la figure 16 au panel B. Celle-ci permet de contenir les valeurs dans le réseau entre 0 et 1, ce qui prévient des explosions de gradients. Cependant, elle encourage aussi les valeurs à rester proches de ces extrêmes, ce qui peut augmenter considérablement le temps d'entraînement.

La fonction tangente hyperbolique, souvent abréviée tanh, vaut $(e^{x} - e^{-x}) / (e^{x} + e^{-x})$ pour tout réel x. Elle est illustrée dans la figure 16 au panel C. Très proche de la fonction sigmoïde, celle-ci a la particularité de pouvoir émettre des valeurs négatives, ce qui peut être pertinent, mais souffre aussi de valeurs restant coincés aux extrêmes.

La fonction *Rectified Linear Unit* (Unité Linéaire Rectifiée en français), ou ReLU vaut $f(x) = \max(0,x)$ pour tout réel x. Elle est illustrée dans la figure 16 au panel D. Cette fonction présente de nombreux avantages, comme une aisance de calcul, une indifférence à l'échelle, et un schéma d'activation éparse sur un réseau dans son ensemble (seulement certaines connexions entre neurones seront utilisées, la plupart restent muettes). Ces avantages ont fait d'elle la fonction d'activation la plus populaire, d'après une analyse de 2017, pour des réseaux de neurones profonds (Ramachandran, Zoph, and Le 2017). Cependant, elle peut tout de même causer un problème de gradients qui disparaissent. En effet, lorsqu'une série de neurones se mettent toutes à émettre 0 pendant plusieurs boucles d'entraînement, il peut être impossible de les "réveiller", et ainsi une partie du réseau s'éteint définitivement.

Pour tenter de contrer cela, la fonction *Parametric Rectified Linear Unit*, ou PReLU, remplace la partie négative de la fonction par une variante : elle vaut $f(x) = x$ si $x>=0$ et $f(x) = ax$ si $x<0$, avec $0<a<1$. Elle est illustrée dans la figure 16 au panel E. Très souvent, a prend la valeur 0.01. Cette version est très populaire et porte le nom *Leaky ReLU* (ReLU fuyant en français) de par son comportement proche de ReLU. Elle est aussi extrêmement populaire de nos jours.

Ces neurones sont généralement disposés en couches successives où l'information circule dans une seule direction et, en employant la méthode de la descente en gradients (Curry 1944), il est possible d'optimiser les poids de chaque connexion entre neurones pour minimiser l'erreur

qu'effectue le réseau sur une tâche donnée. Les dérivations impliquées dans ce processus se font en général de manière partielle, c'est-à-dire en prenant en compte un seul paramètre à la fois pour simplifier les calculs. Les gradients calculés sont ensuite multipliés par le taux d'apprentissage, puis soustraits aux paramètres correspondants. Cette étape s'appelle la rétropropagation du gradient, et implémente le théorème de dérivation des fonctions composées de Leibniz.

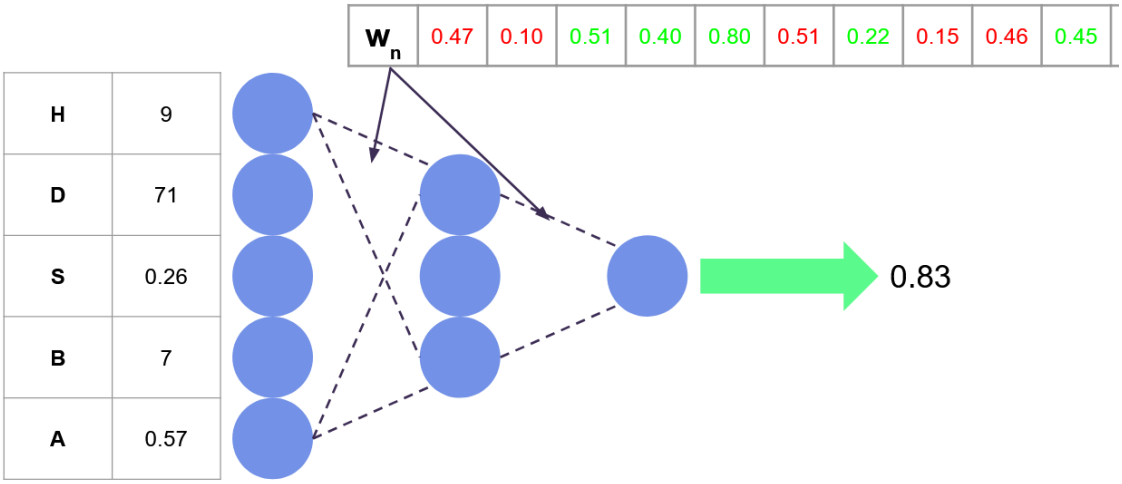


Figure 17 : Représentation schématique d'un réseau de neurones artificiels. À chaque lot d'échantillons qui lui est présenté pendant son processus d'entraînement, le réseau met à jour ses poids pour tenter de se rapprocher des résultats attendus.

La majorité des modèles utilisés de nos jours sont constitués de plusieurs couches de ces neurones, comme schématisé dans la figure 17. Cette succession de couches permet au modèle de représenter des caractéristiques complexes et composées de plusieurs caractéristiques issues de la couche précédente.

Le processus d'enchaînement des couches de neurones ajoute beaucoup plus de possibilités par rapport à ce que peut traiter un modèle à une seule couche, mais aussi de nombreuses complications au système. Plus on multiplie les couches du réseau, plus les dérivées calculées lors de la rétropropagation des gradients deviennent compliquées, et plus le nombre de poids dans le réseau augmente, multipliant aussi le nombre de calculs à effectuer.

Mathématiquement, on peut considérer que le réseau tente d'approximer une fonction qui prend en entrée les différentes données fournies au système et qui calcule la sortie demandée. Tant que les couches qui le composent sont munies de fonctions d'activation non-linéaires, cette

fonction globale pourra l'être aussi.

7.1.3.1. Histoire

La première implémentation de cette méthode était le *Perceptron*, en 1957, qui avait pour tâche de classifier des images (Rosenblatt 1958). Les poids assignés à chaque connexion étaient ajustés mécaniquement par l'action de potentiomètres. Cet outil a originellement créé beaucoup d'espoir en cette méthodologie, mais la technologie disponible ne lui a pas permis de s'épanouir tout de suite. Il a fallu attendre les années 1980 pour que la communauté scientifique y reprenne un véritable intérêt.

7.1.3.2. Forces et faiblesses

L'utilisation de réseaux de neurones artificiels présente bon nombre d'avantages. Le premier de ceux-ci est leur flexibilité : leur modularité permet de façonner un réseau pour presque n'importe quelle tâche, et il existe de nombreuses architectures de modèles publiés ayant fait leurs preuves. Le second avantage est la facilité de la prise en main : il existe plusieurs librairies de code permettant d'implémenter des modèles très rapidement et passer très vite à une phase de production. Les deux plus grandes bibliothèques permettant l'utilisation de ces modèles en python, à savoir Keras (Q. S. Zhu 2023) et PyTorch (Chintala 2016), publiés en 2015 et 2016 respectivement, sont utilisés dans des milliers de projets d'apprentissage automatique et fouille de données (Keras 2023).

Les réseaux de neurones présentent aussi des embûches à leur utilisation, comme toute méthode, mais il est possible de les accommoder pour réduire leur impact. Tout d'abord, comme toutes les méthodes d'apprentissage automatique, il existe pour tout problème au moins une solution optimale que l'on cherche à atteindre, et souvent de nombreuses solutions locales approximatives qui peuvent paraître idéales avec nos données limitées mais présenteront des limites lors de la généralisation du modèle. Ensuite, l'effet boîte noire, c'est-à-dire l'opacité des calculs qui se font dans le réseau de neurones, peut apporter des doutes sur ces modèles, vu qu'il n'existe pas de méthode intrinsèque pour expliquer les résultats qu'ils produisent. Depuis leur montée en popularité, des méthodes *post hoc* permettent d'analyser les réseaux pour trouver cette information, mais ceci rajoute un étape supplémentaire au processus de développement si cette

explication est souhaitée. Le dernier que je listerai ici est le nombre d'hyperparamètres sur lequel l'utilisateur peut jouer. Cela représente à la fois un grand nombre de possibilités pour améliorer les performances du modèle, mais en même temps il peut être difficile de déterminer quels paramètres ajuster, lesquels sont superflus ou même parfois gênants.

7.1.3.3. Les réseaux génératifs

Il existe des modèles de réseaux de neurones qui permettent d'employer leurs fonctions pour créer de nouvelles données. Ces modèles suivent souvent une architecture spécifique qui permet d'utiliser les optimisation mathématiques du réseau de neurones pour améliorer itérativement la vraisemblance des nouveaux échantillons créés.

L'appellation "IA créative" a parfois été employée pour décrire ces modèles, mais ce terme rejoint le phénomène que j'avais évoqué dans la section 7.1.1 : "L'intelligence artificielle en général": on tente d'attribuer à l'intelligence artificielle des caractéristiques comparables à ce qu'effectue un humain. Le résultat est peut-être similaire (ou au minimum dans une catégorie comparable), mais cela n'indique en rien le caractère créatif d'un processus.

7.1.3.3.1. Generative Adversarial Network

Pendant de nombreuses années, la plus populaire de ces méthodes était le GAN : *Generative Adversarial Network* (Goodfellow et al. 2014). En réalité, il s'agit de deux réseaux qui s'entraînent l'un contre l'autre dans un jeu à somme nulle, schématisés dans la figure 18, où toute amélioration dans le score attribué à la performance de l'un cause une baisse du score de l'autre.

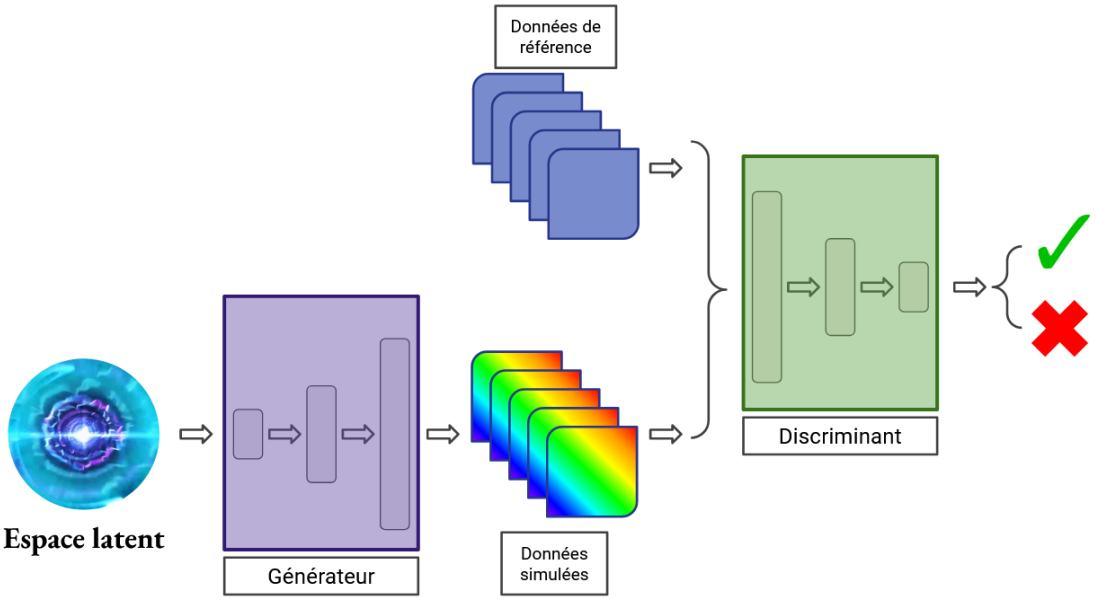


Figure 18 : Diagramme représentant l'architecture d'un GAN.

L'un des deux, appelé discriminant, reçoit en entrée de vraies données ainsi que des données simulées, et tente de différencier les deux, en assignant un score de 1 aux vraies données et un score de 0 aux simulées. L'autre, le générateur, puise sur des données aléatoires pour créer des données qui tentent de tromper le discriminant, en cherchant les échantillons qui donneront un score maximal. Ces deux parties du modèle sont entraînées l'un après l'autre, en boucles successives jusqu'à ce qu'on décide que le générateur crée des échantillons suffisamment réalistes.

Cette condition d'arrêt peut être très compliquée à déterminer. Une pratique habituelle est de faire évaluer les données produites par le générateur par des humains à intervalles réguliers, si on ne dispose pas de méthode calculatoire pour comparer les données simulées aux données de référence.

Le GAN fait partie des réseaux utilisés pour tenter de simuler des données génomiques humaines dans (Yelmen et al. 2021). Le discriminant reçoit en entrée des profils de mutations haplotypiques (présence ou absence de chaque mutation étudiée), à la fois réels (provenant du 1KG) et simulés par le générateur du modèle. Il produit en sortie un nombre en 0 et 1, qui peut être interprété comme la probabilité que cet échantillon soit vrai (déterminé par le discriminant). Le générateur prend en entrée 600 variables aléatoires gaussiennes, qu'il emploie pour générer

des profils de mutations au format identique à ceux tirés du 1KG. Pendant l'entraînement du générateur, on fige les paramètres du discriminant et l'objectif de l'entraînement est de maximiser la sortie du discriminant pour ces échantillons simulés (ils doivent être aussi réalistes que possible).

Ce GAN a montré de très bonnes performances généralement, et est assez simple à implémenter. Le générateur final qui en est produit est aussi plutôt facile à extraire et partager après l'entraînement, ce qui le rend maniable et portatif.

La capacité des GAN à reconnaître et reproduire des motifs implicites dans des jeux de données, mêmes complexes, n'est plus à prouver (Caesar 2017). Cette première application dans la génomique humaine montre que cette capacité se traduit aussi dans ce type de données. Nous pouvons donc espérer utiliser cet outil pour créer des cohortes de patients simulés, basés sur des cohortes de véritables patients. Ainsi, les caractéristiques intrinsèques de cette cohorte seront reproduites dans la simulation, selon les règles que le GAN a repéré.

Lorsque l'équipe de Flora JAY emploie ce modèle dans leur projet (Yelmen et al. 2021), il montre globalement de très bonnes capacités à simuler de nouvelles données génomiques. Notamment, sur les ACP comparant les populations simulées à la population de référence, on voit une bonne capacité d'exploration de la part du GAN. Cette exploration a parfois tendance à déborder dans des zones inexplorées par les autres populations, mais il est compliqué de construire une méthode qui est capable de reproduire une distribution exactement. Il revient à l'utilisateur de s'interroger sur ce qui est plus approprié pour son projet : un système qui explore trop ou trop peu.

7.1.3.3.2. Restricted Boltzmann Machine

Une *Restricted Boltzmann Machine* (Sherrington and Kirkpatrick 1975), ou RBM, est un type de réseau de neurones contenant seulement deux couches, schématisées dans la figure 19 : une visible et une cachée, densément connectées entre elles, mais sans connexions entre les neurones d'une même couche (d'où la notion de "restreinte"). Une autre particularité est que l'information peut circuler soit du visible vers le caché, ou du caché vers le visible. Ces deux étapes se réalisent lors de l'entraînement, l'un après l'autre, pour créer une couche cachée contenant des représentations latentes des données introduites dans la couche visible. Pour

mesurer ses performances, on compare les reconstructions que le modèle effectue lorsque l'information passe de la couche cachée vers la couche visible aux échantillons de référence.

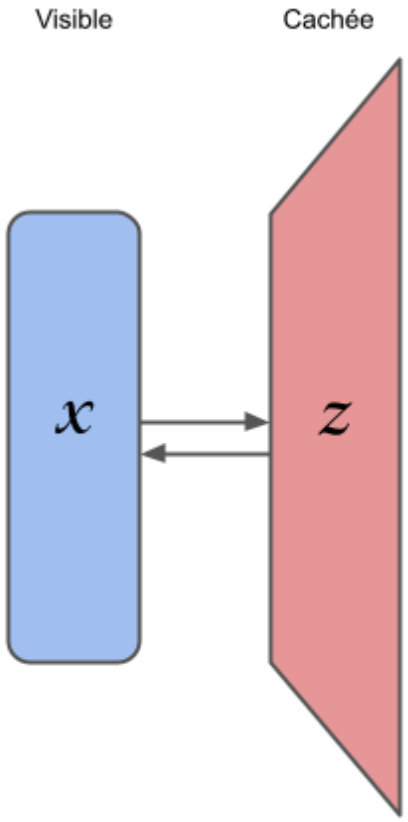


Figure 19 : Schéma de l'architecture d'une RBM.

Ce type de modèle est assez simple à mettre en place, et souvent peu coûteux en mémoire en comparaison aux autres modèles de réseaux de neurones, vu qu'il n'est composé que de deux couches. Il a su faire ses preuves notamment dans la recommandation de vidéos sur plateforme de streaming (Jia et al. 2015).

Dans (Yelmen et al. 2021), ce type de modèle appliqué aux données génomiques donne en général de très bons résultats, souvent meilleurs que le GAN. Cependant, le défaut majeur que ce modèle montre est une grande fuite d'information en comparaison aux données de référence. Cette fuite d'information explique aussi en partie les performances du modèle, vu qu'un jeu de données aussi proche des données de référence aura bien évidemment des caractéristiques plus semblables. À cause de cette fuite d'information, je préfère ne pas employer ce modèle, vu que la question à laquelle je réponds contient une grande part de protection des données médicales de

patients.

7.1.3.3.3. Diffusion stable

La méthode de diffusion stable (Rombach et al. 2022) a connu un grand essor ces quelques dernières années. Sa conception simple mais innovante a su en faire une architecture de modèle incontournable pour les développeurs modernes en génération d'images.

Mathématiquement, une image floutée part d'une image source nette et, pour chacun de ses pixels, on vient moyenner les valeurs de couleurs avec ses pixels voisins. Pour déflouter une image, il faudra donc passer sur ses pixels et leur retirer une partie des valeurs de ses voisins pour inverser ce phénomène. Bien sûr, ce n'est pas aussi direct puisque, après floutage, les voisins directs d'un pixels contiendront aussi de l'information provenant de leurs voisins directs. Les méthodes basées sur l'intelligence artificielle offrent donc une opportunité de réaliser ce défloutage en prenant en compte les données à un spectre légèrement plus large et de manière optimisée, après entraînement. Cet entraînement est d'ailleurs assez simple à mettre en place, puisqu'on peut flouter, à un degré souhaité par l'utilisateur, n'importe quelle banque d'images simplement avec la méthode décrite précédemment pour générer un grand nombre d'échantillons à analyser.

Avec ce processus, on peut donc aisément créer un modèle qui défloute des images de 5% par exemple. Pour une image floutée à un degré connu, on peut donc appliquer ce modèle à de multiples reprises pour réduire son niveau de floutage à zéro.

Un élément important à fournir au modèle en plus des données de l'image que l'on souhaite traiter est une description du contenu de l'image. En effet, une fois floue, il est très facile de confondre des objets présentant une vague ressemblance, mais de nature très différente dans une image nette.

Il faut donc que l'utilisateur spécifie que le modèle identifié dans l'image les éléments sur lesquels il doit se concentrer. Même avec ces indications, il n'est pas garanti que le modèle ne fasse aucune erreur. Ce processus demande donc un peu plus de travail, mais nous disposons de banque d'images étiquetées de plus en plus fournies, ce n'est donc pas impossible.

Maintenant, que se passe-t-il si on présente une image 100% floutée à cet outil ?
En d'autres termes, si nous transmettons une image composée entièrement de bruit à ce modèle,

et lui précisons que l'image contient un chat, l'outil travaillera selon ces spécifications. Au fur et à mesure des défloutages, dans le bruit que nous lui avons transmis, il repérera les portions de l'image qui ressemblent le plus à une truffe, une patte, ou une queue, et créera un chat de toutes pièces.

Cette logique s'est montrée très robuste pour la création de nouvelles images, tant que les utilisateurs sont capables de lui fournir une description suffisamment détaillée de l'image qu'ils souhaitent obtenir. Les résultats sont souvent assez approximatifs, mais cette méthode n'existe que depuis quelques années et pourrait faire ses preuves de manière plus convaincante à l'avenir.

Cette méthode pourrait s'appliquer aussi aux données génomiques que je traite dans ce projet, où le floutage serait cette fois en une seule dimension. Cependant, cette méthode se base notamment sur l'utilisation de la convolution, et comme détaillé dans la section 7.2.1.1 : "Qu'est-ce que la convolution ?", cette approche n'est pas forcément la plus appropriée pour nos données, qui ne suivent pas les mêmes règles que les images. J'ai donc préféré me tenir à des méthodes plus structurées, mais elle reste envisageable pour de nouveaux projets dans ce domaine.

7.1.4. Questions éthiques autour de l'intelligence artificielle

7.1.4.1. Préoccupations générales

7.1.4.1.1. Biais dans les jeux de données

Les jeux de données utilisés pour entraîner un modèle d'intelligence artificielle sont en général de très grande taille, composés de milliers d'échantillons au minimum. Selon le domaine d'application, ces données peuvent être vieilles de plusieurs années voire plusieurs décennies lorsque l'entraînement est lancé, et ainsi ne pas répondre aux critères de validité exigées aujourd'hui (Zhisheng Chen 2023; Belenguer 2022). Un jeu de données récent peut également contenir des manquements, mais avec la montée en popularité de l'intelligence artificielle, de plus en plus de soin est apporté à l'établissement de ces jeux de données. Une IA entraînée sur un jeu de données où il manque certaines catégories de personnes par exemple reproduira de manière intrinsèque ces biais dans les décisions qu'elle prendra, vu qu'elle ne cherche qu'à intégrer les informations contenues dans le jeu de données (Nazer et al. 2023). Il revient à

l’utilisateur de s’assurer que l’information est bien représentée.

7.1.4.1.2. Collection de données

L'entraînement des IAs nécessite une grande quantité de données. La collecte de ces données peut s’effectuer de nombreuses manières, et certaines entreprises peuvent se montrer peu scrupuleuses dans leurs politiques de collecte de données (Zhisheng Chen 2023; Nazer et al. 2023; Belenguer 2022). Il n’est pas un secret qu’aujourd’hui la grande majorité des réseaux sociaux et autres outils en ligne collectent énormément de données sur leurs utilisateurs pour optimiser les publicités qui seront affichées à l’utilisateur, ou les revendre à d’autres partis intéressés. Des cas connus de collecte de données à l’insu d’utilisateurs, comme lors de l’affaire Facebook-Cambridge Analytica (Schneble, Elger, and Shaw 2018), ont mis en lumière les considérations éthiques essentielles à une collecte de données qui respecte les personnes qui y ont contribué.

Le stockage de ces données est également un point important à prendre en compte. Pouvoir les conserver de manière sûre est une contrainte légale de nos jours dans l’Union Européenne (Mondschein and Monda 2019). Cependant, au niveau mondial, il n’est pas toujours garanti que les mêmes mesures de sécurité soient respectées. Un agent malicieux peut ainsi exploiter de multiples voies pour tenter d’accéder aux informations personnelles contenues dans la base de données.

7.1.4.1.3. Biais dans les algorithmes

L’implémentation de systèmes automatisés pour le traitement de données peut mener à des impacts imprévus sur certains groupes sociaux.

Le biais algorithmique est un phénomène connu depuis plusieurs années maintenant. Lorsqu’on confère la responsabilité d’une tâche à un automate, on a tendance à moins remettre en question les erreurs que cet automate peut effectuer (Zhisheng Chen 2023; Belenguer 2022; Varsha 2023). En parallèle, on a tendance à oublier qu’une personne est responsable de l’implémentation de ce comportement erroné, ou au minimum de leur persistance une fois qu’elles sont remarquées. La perception sociale de l’intelligence artificielle comme une révolution technologique accentue ce phénomène, alors que ces méthodologie sont tout autant

sujettes aux biais que les solutions qui existaient précédemment (comme ceux soulignés dans les sections précédentes).

7.1.4.1.4. Transparence des modèles

L'interprétations des décisions réalisées par une IA étant compliquée, il peut être difficile de savoir quel critère dans un échantillon spécifique mène à une décision finale donnée (Zhisheng Chen 2023; Varsha 2023). Ce manque de transparence présente notamment des implications juridiques fortes dans certains cas.

De nombreuses architectures d'IA, notamment les réseaux de neurones, ne permettent pas intrinsèquement de suivre le processus de décisions effectué par le modèle. Il est possible d'analyser ce processus une fois l'entraînement achevé, mais ce n'est pas encore la norme dans le domaine. Cette caractéristique peut cependant se montrer nécessaire lorsque ces décisions sont impliquées dans un procès juridique, où il devient obligatoire de justifier les éléments ayant contribué à une décision donnée. Vu que les modèles IA sont de plus en plus utilisés par les entreprises à travers le monde, il n'est pas impossible que cette interprétabilité devienne un critère requis pour toute utilisation commerciale de l'IA.

7.1.4.1.5. Impact sur les emplois

Depuis la révolution industrielle, l'introduction de solutions automatisées au travail a toujours perturbé les emplois liés aux domaines où ces solutions ont été introduites. Avec la vague de nouvelles avancées permises par le développement de l'intelligence artificielle, de plus en plus de secteurs risquent d'être touchés (Zhisheng Chen 2023) et il est donc fort probable de voir le taux de chômage augmenter dans de nombreux pays. Sans aborder la question de la régulation de l'automatisation au travail, il serait plus éthique de la part des entreprises de préparer une adoption de l'intelligence artificielle en fournissant le moyen à leurs employés de pouvoir se former de nouveau, pour ne pas avoir à effectuer cette transition après avoir perdu leur emploi.

Les biais précédemment cités affectent également le recrutement des employés. Il existe déjà bon nombre de méthodes d'analyser un CV ou une lettre de recrutement par IA pour en déceler les mots clés. Dans un but d'optimisation et de gain de temps, il n'est pas rare que le

département de ressources humaines de grandes entreprises ou des boîtes de recrutement emploient ces méthodes pour trier leurs candidats rapidement, sans se pencher sur les biais que ces outils peuvent avoir envers des minorités ethniques ou les femmes, par exemple (Zhisheng Chen 2023; Belenguer 2022; Varsha 2023).

7.1.4.1.6. Impact environnemental

À cause des besoins en données et puissance de calcul de l'intelligence artificielle, la question de la production d'électricité, nécessaire à ces éléments, devient pertinente (Dhar 2020; OECD 2022). La question cependant s'étend aussi à la production d'électricité et son émission de gaz à effets de serre de manière générale, vu que l'IA n'est pas le seul domaine qui voit sa consommation d'électricité augmenter. Cependant, son accélération massive est une grande source de nouvelles émissions, et le déploiement de grands serveurs de calculs implique aussi d'autres risques pour l'environnement, comme l'utilisation de refroidissement par eau circulante qui peut ensuite impacter les écosystèmes dans laquelle elle est déversée une fois chaude.

7.1.4.2. Risques dans le domaine médical

7.1.4.2.1. Risques élevés pour certains facteurs

Un projet d'intelligence artificielle dans le domaine médical reprend les préoccupations mentionnées précédemment, et exacerbe plusieurs d'entre elles. Il devient impératif de gérer la collection et la gestion des données avec le plus grand soin, vu la nature de celles-ci. Les conséquences potentielles d'une fuite de données médicales peuvent impacter énormément de facettes de la vie des personnes concernées (Nazer et al. 2023; Shabani and Marelli 2019).

Le biais algorithmique est accentué ici, car le domaine médical souffre déjà de ses structures autoritaires (Campbell, Croskerry, and Petrie 2017), où la décision d'une personne haut placée n'est jamais remise en question. L'ajout d'un nouvel agent qui tend à renforcer ce phénomène demande une intégration mesurée pour ne pas porter atteinte aux plus vulnérables. La dimension de l'interprétabilité d'un réseau de neurones devient aussi très importante, au vu de l'impact que peuvent avoir les décisions prises par celle-ci. La possibilité qu'un patient (ou ses proches) amène l'hôpital en justice est à garder à l'esprit.

7.1.4.2.2. Priorité sur le bien-être du patient

Les décisions que réalisent l'IA ne reflètent pas toujours l'opinion d'un professionnel médical. La priorité pour un médecin doit toujours être le bien-être et la santé de son patient, et cette priorité s'exprime par un comportement qui vise à minimiser les risques encourus par le patient pendant son traitement. Une IA ne conçoit pas forcément la notion de risque, et surtout pas les risques que peut encourir une personne à la santé déjà fragile. Si l'IA effectue une prédiction en se basant uniquement sur les données numériques à sa disposition, elle pourrait mettre en danger la vie du patient en suggérant des procédures ou des traitements qu'un expert humain n'envisagerait jamais.

Jusqu'à présent, ce type de risque fait partie des priorités placées sur les modèles IA développés pour un contexte médical (Choudhury and Asan 2020). Le soin apporté à ce critère témoigne de son importance. Il sera nécessaire de conserver cette attention pour les années à venir, pour éviter toute erreur qui ternirait la réputation de l'IA dans le domaine de la santé.

7.1.4.2.3. Intégration dans le déroulement des opérations en milieu clinique

Avec ces conditions en tête, il devient une priorité de mettre en place des modèles qui évaluent les données disponibles et fournissent leurs prédictions de manière à pouvoir collaborer avec des experts humains. De cette manière, l'IA prend la place d'un outil au service des décisions humaines, et ne vise pas à les remplacer (Nazer et al. 2023). Cependant, cette intégration est délicate et nécessite un grand soin dans la formulation de ses tâches et impose de fournir également des métriques de confiance pour caractériser les prédictions réalisées. Ces nouveaux critères peuvent paraître pénibles pour les développeurs, qui sont souvent plus intéressés par la question de la performance de leurs modèles que par le signalement des marges d'erreur avec lesquelles ces modèles opèrent.

7.1.4.2.4. Accès équitable

Un point sensible dans le domaine médical, qui s'étend au-delà de l'utilisation d'intelligence artificielle, est l'accessibilité des dernières avancées et de la médecine expérimentale. En France et en Union Europe, nous disposons de structures permettant un accès plutôt équilibré aux soins (même s'il présente encore des lacunes), mais ce n'est pas le cas

partout dans le monde. Il arrive donc que des citoyens de France ou d'Europe partent à l'étranger pour des séjours à but médical, pour bénéficier des modalités de l'accès aux soins d'autres régions. Les développeurs d'outils IA à but médical se doivent donc de garder en tête les principes de science et de développement ouverts pour permettre l'implémentation de leurs systèmes dans un maximum d'établissements (Nazer et al. 2023).

7.1.4.2.5. Accélération de risques déjà existants

Au-delà des risques cités ci-avant, les performances remarquables de l'intelligence artificielle permettent de rendre plus rapides et plus efficaces des pratiques déjà présentes. Ces applications ne prennent pas plus en compte les notions d'éthique que lorsqu'elles ont été réalisées dans l'utilisation de l'IA. Un exemple qui revient souvent est le *designer baby*, un terme désignant l'édition en détail du génome d'un embryon à implanter dans une mère porteuse.

Les techniques de diagnostic pré-implantatoire, ainsi que les avancées récentes en biologie moléculaire, permettent en théorie de tester un large panel de caractéristiques chez un embryon. Les plus communes sont les maladies génétiques comme la trisomie 21 (Boddupally and Rani Thuraka 2024), mais en théorie tout caractère suffisamment bien documenté peut faire l'objet de la sélection. Ainsi, il est facile d'imaginer un monde où chaque critère est mesuré et sélectionné pour "optimiser" la vie de son enfant, jusqu'à la couleur de sa peau pour lui éviter des discriminations.

Ce monde n'est pas une réalité, mais les outils employant l'intelligence artificielle peuvent permettre à quiconque souhaitant fournir ce service de tisser des prédictions entre critères génétiques et caractéristiques phénotypiques (Ramirez 2018). Ces prédictions ne seraient pas toujours de bonne qualité, notamment pour des caractéristiques très complexes comme la beauté ou l'intelligence, mais pour toutes les raisons citées précédemment, parfois l'argument de l'utilisation de l'IA peut suffire pour convaincre.

7.1.4.3. Risques liés aux modèles génératifs

Ce projet de thèse vise à employer des réseaux de neurones générateurs de nouvelles données. Cette utilisation a connu beaucoup de popularité ces dernières années, après avoir connu une amélioration considérable en termes de performances et de contrôle sur le contenu

généré. Ceci amène donc de nouvelles opportunités pour l'automatisation dans certains domaines, mais aussi de nouvelles problématiques éthiques.

De nombreux artistes travaillant dans le cinéma notamment ont soulevé le risque que cela présente pour leur métier. Dans cet environnement, plusieurs corps de métier dans cette industrie aux États-Unis ont décidé de faire grève en Mai 2023, lié aussi à des questions de salaire et de redevances pour leurs productions. Pendant cette grève les studios Marvel ont publié leur série télévisée *Secret Invasion* sur leur plateforme de streaming Disney + en Juin 2023, utilisant des images synthétisées par IA dans le générique d'ouverture, soulignant les avis des deux partis sur la question (Rogin and Corkery 2023).

Il existe également des craintes relatives à l'imitation d'acteurs via l'intelligence artificielle. Il est possible aujourd'hui d'enseigner à une IA comment reproduire la voix d'une personne spécifiquement, avec un réalisme convaincant. Cette pratique pourrait servir à remplacer des acteurs pour des doublages, ou alors à les imiter sur des plateformes comme les réseaux sociaux pour arnaquer des utilisateurs.

Dans un domaine plus sérieux, le parti Républicain aux États-Unis a publié en Avril 2023 une courte vidéo publicitaire intitulée *"Beat Biden"* ('Beat Biden - YouTube' 2023), qui tente de susciter de la peur sur les potentielles conséquences d'une réélection de Biden en 2024. Pour illustrer ces conséquences, de nombreuses images ont été générées par IA, montrant des images violentes de révoltes, de guerres et de ruines. Lorsque cette vidéo fut diffusée sur les réseaux sociaux et à la télévision, la mention que les images étaient générées par IA était affichée si petite que beaucoup de spectateurs ne l'ont su que lorsque d'autres sources leur ont annoncé (Novak 2023). Cet événement démontre comment les images générées par IA peuvent tromper un œil non-expert.

Très récemment, à la suite de la déclaration de guerre de l'Israël envers le groupe Hamas, le compte Twitter de l'ambassade d'Israël au Panama a partagé une photo montrant un groupe de jeunes dans un aéroport avec un drapeau Israélien (Israel in Panama [@IsraelinPanama] 2023). Le texte accompagnant la photo indique que ce sont de jeunes Israéliens qui rentrent pour s'engager dans l'armée et protéger leurs concitoyens. Très rapidement après la publication de ce billet, de nombreux autres utilisateurs y ont répondu en se focalisant sur certains détails de l'image, comme des visages ou des membres, soulignant le manque de réalisme de ces éléments.

En effet, les déformations sur ces parties de la photo ressemblent énormément à des artéfacts que l'on peut trouver sur des images générées par IA. Ceci remet donc en question la véracité de l'événement, et on peut donc se demander combien de personnes sont réellement rentrées en Israël pour participer à l'effort militaire.

De manière globale, cet outil offre de grandes possibilités pour créer des données de plus en plus vraisemblables, potentiellement pertinentes dans un large éventail de domaines. Un avis répandu dans la communauté scientifique, et que je partage, est qu'un outil n'a pas de propriétés éthiques intrinsèques, mais c'est l'utilisation que chacun en fait qui est sujet à des jugements moraux. Il est important cependant de souligner un phénomène psychologique intéressant : lorsqu'une nouvelle technologie permet de faire quelque chose facilement, alors on arrête de se questionner sur la valeur morale de cette action (Ihde 1990; Ord 2020; MacAskill 2022). Les campagnes politiques malhonnêtes vont-elles devenir banalisées à l'avenir ? Trouvera-t-on des photos retouchées, voire mensongères, sur les réseaux sociaux d'ici quelques années ?

Un élément important qui pourrait être intégré à des modèles génératifs est l'utilisation de *watermarking*, c'est-à-dire intégrer un filigrane dans les créations réalisées par IA directement dans le contenu. Ceci permet de reconnaître instantanément ces images et vidéos. De multiples grandes entreprises se sont engagées à étiqueter les contenus créés par leurs modèles avec ce genre de technologie. Cependant, si le but d'un utilisateur est la désinformation, alors celui-ci cherchera, et très certainement trouvera, une solution pour effacer le filigrane de sa création, par exemple en utilisant d'autres modèles d'intelligence artificielle.

7.2. Réduction de dimension par intelligence artificielle

De nos jours, il devient de plus en plus facile de collecter des volumes massifs de données, sur à peu près n'importe quel sujet. La puissance de calcul dont nous disposons pour les traiter augmente aussi, mais il y a plusieurs enjeux à prendre en compte quand on souhaite attaquer un problème en se reposant simplement sur la puissance de calcul brut : temps de traitement, espace de stockage, coût financier, émissions de CO_2… La liste est longue.

Les méthodes de réduction de dimensions deviennent donc de plus en plus importantes de par leur rôle de prétraitement des données. En effet, comme mentionné dans la section 7.1.2.2, la malédiction de la dimension explique que multiplier les paramètres mesurés cause plus de problèmes que ça n'en résout.

De plus, dans de nombreux cas, les caractéristiques observées directement ne contiennent pas toujours toutes les informations qu'il faut pour analyser les motifs dans les données. Il existe bien souvent des caractéristiques qui dépendent de plusieurs facteurs élémentaires. Par exemple, il est souvent complexe d'assigner un genre musical unique à une chanson, et plusieurs éléments entrent en compte pour tenter de choisir la catégorie efficacement : tempo, tonalité, type d'instruments utilisés, etc. Si une étape de réduction de dimensions permettait de prendre en compte ces dimensions individuelles pour en faire ressortir une notion de genre musical, alors elle serait très utile pour ensuite pouvoir établir un système de recommandation pour une plateforme de streaming de musique.

En biologie, il est de notoriété que les différentes mutations qui nous distinguent n'ont pas toutes le même niveau d'importance, et qu'il existe beaucoup de traits phénotypiques qui dépendent de plusieurs gènes et de leurs allèles. Il paraît donc naturel d'appliquer une étape de réduction de dimensions à des données de mutations pour en extraire les informations essentielles.

7.2.1. Sélection de la méthode de réduction de dimensions

Il existe de très nombreuses méthodes de réduction de dimension, avec chacune ses forces et faiblesses. Dans notre cas, il n'est pas inintéressant de se pencher sur les méthodes basées sur l'apprentissage automatique, et plus particulièrement les réseaux de neurones (Fournier and

Aloise 2019).

Ces méthodes présentent plusieurs avantages très pertinents dans le cadre de mon projet :

- elles traitent mieux les données non-linéaires, là où des méthodes comme l'Analyse en Composantes Principales rencontrent des difficultés,
- elles préservent mieux les structures locales, et pas seulement globales,
- elles permettent d'inférer des caractéristiques implicites dans le jeu de données, (c'est-à-dire de déduire des genres musicaux si on reprend l'exemple précédent),
- elles traitent bien les données bruitées, et peuvent souvent servir d'étape de suppression du bruit en même temps.

Même en restreignant à des méthodes utilisant des réseaux de neurones, il existe plusieurs approches possibles, ayant toutes leurs particularités.

7.2.1.1. Convolution 1D

7.2.1.1.1. Principe de la convolution

En mathématiques, la convolution est une opération qui permet de combiner deux listes de nombres. Elle peut aussi s'appliquer à deux fonctions continues, mais le cas qui nous intéresse ici est l'application à des ensembles discrets. Pour deux listes *a* et *b* chacune de taille *n*, la convolution *(a*b)* vaut :

$$\sum_{i+j=n}^{\forall i,j} a_i \bullet b_j$$

Si les deux listes ne font pas la même taille, alors il y a plusieurs approches possibles. Les deux plus communes sont soit de tronquer la liste la plus grande, soit d'ajouter des éléments nuls à la plus petite liste jusqu'à ce qu'elles aient la même taille.

L'utilisation de la convolution en traitement d'images consiste à appliquer des matrices de taille restreinte sur des échantillons de grande taille et à calculer la valeur de la formule ci-dessus entre ces deux matrices. Ainsi, chaque petite matrice devient un filtre qui permet de détecter des motifs dans l'image, et chaque filtre différent peut détecter un motif particulier, par exemple des

lignes droites horizontales ou verticales comme dans la figure 20. Cette utilisation est devenue très commune en raison de l'augmentation de la résolution des images, qui entraîne une explosion du nombre de dimensions dans le jeu de données.

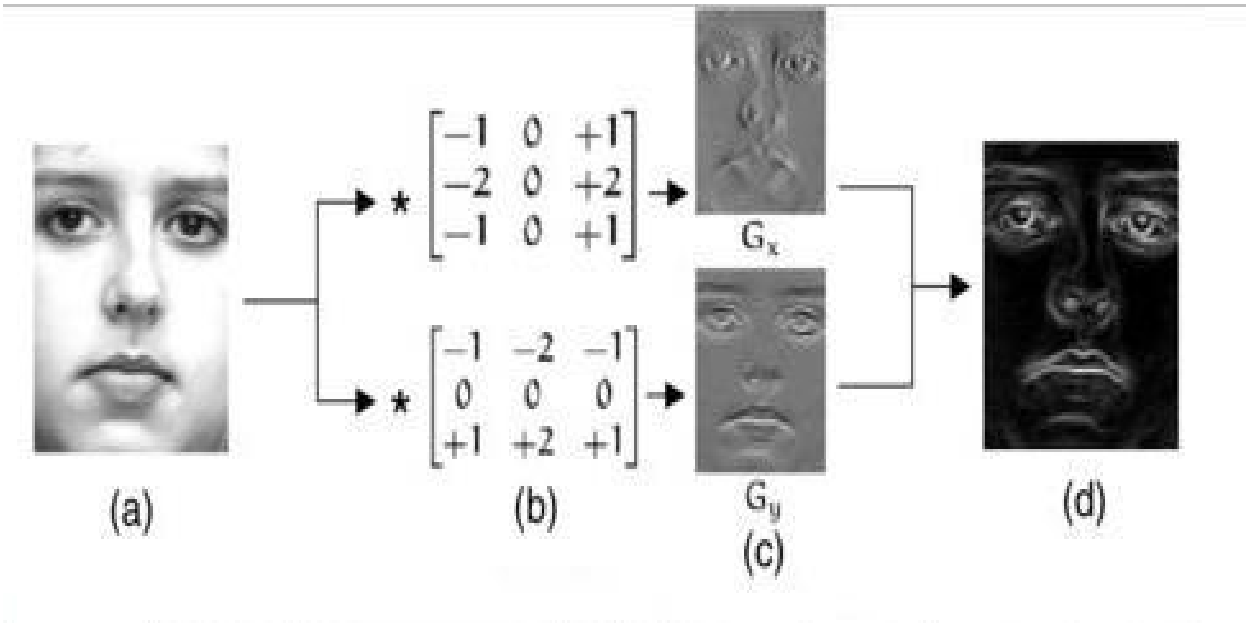


Figure 20 : Résultat de l'application de deux filtres de convolution différents sur une image. Adapté de (Holder and Tapamo 2017).

Il est possible d'appliquer un grand nombre de filtres différents sur les données pour détecter des motifs différents. En effet, la stratégie de l'intelligence artificielle est d'en déployer une grande quantité, initialisés aléatoirement, puis optimisés par apprentissage. Il est possible d'empiler les filtres et de croiser leurs résultats pour détecter des structures plus grandes et plus complexes dans les échantillons. Certains modèles emploient aussi des modèles de convolution "multi-têtes" (Khan and Ahmad 2021), où différentes sections du modèle appliquent des filtres de taille différentes, pour détecter à la fois des motifs dans les détails des données ainsi qu'à grande échelle.

7.2.1.1.2. Avantages

La convolution est très populaire dans l'analyse d'images ou des données sonores, puisque sa méthodologie permet de détecter des motifs n'importe où dans les données. Cette architecture de modèle est bien adaptée à ces types de données à cause de la variance inhérente qu'elles contiennent : les photos d'identité ne sont pas centrées et cadrées exactement de la même

manière, par exemple. L'intensité en couleur d'un pixel à une position exacte dans une telle photo aura donc moins d'impact que le motif à plus grande échelle dont ce pixel fait partie pour un modèle cherchant à effectuer une reconnaissance faciale.

Les données que je traite ici sont stables, la première dimension du jeu de données utilisé correspond toujours à la même mutation. Cette stratégie pourrait donc être "distraite" par des motifs se répétant par coïncidence dans les données, et mal représenter les données initiales.

7.2.1.2. Carte autoadaptive

7.2.1.2.1. Principe

Une carte autoadaptive utilise le réseau de neurones de manière assez particulière. Les neurones sont placés sur un espace souvent en deux dimensions, et contiennent comme paramètres initiaux un nombre de dimensions égale au nombre de dimensions du jeu de données sur lequel la méthode s'applique (Kohonen 1982). Chaque neurone voit son vecteur de paramètres initialisé aléatoirement.

Ensuite, de manière itérative, on présente à la carte un échantillon du jeu de données, et les neurones vont entrer en compétition pour déterminer lequel est le plus proche de l'échantillon (en général en utilisant une mesure de distance euclidienne). Un seul neurone sera choisi après cette compétition, qu'on appelle le gagnant. Ce gagnant et son voisinage (déterminé selon un hyperparamètre) vont ensuite ajuster leurs vecteurs de paramètres pour se rapprocher de l'échantillon en question. Cette étape encourage le réseau à prendre la forme de la distribution des données dans le jeu de base, mais projetée sur l'espace de la carte qui contient les neurones.

Au final, chaque neurone représentera donc une région de l'espace original. Plus une région est représentée dans le jeu de données, plus il y aura de neurones qui s'y agglutineront dans la carte.

7.2.1.2.2. Forces et faiblesses

Ce type de réseau de neurones vise à déformer une distribution pour la simplifier, tout en conservant sa topologie. Cette stratégie peut être intéressante à envisager comme méthode de réduction de dimensions, mais un de ses principaux buts est de créer des cartes visuellement

interprétables. Cette fonctionnalité est très souvent pratique, mais n'est pas requise dans le cadre de ce travail, vu que ce projet ne présente pas de contraintes en termes de visualisation de données. Ceci mène donc à des calculs superflus et pourrait introduire de nouveaux biais dans la réduction de dimension. De plus, il n'y a en général pas de lien direct entre un échantillon dans le jeu de données de base et le réseau de neurones final. La méthode n'est donc pas garantie de fonctionner pour l'utilisation dans ce projet.

7.2.1.3. Autoencodeurs

7.2.1.3.1. Principe

Les autoencodeurs sont une classe de réseaux de neurones ayant pour but la réduction de dimension (Kramer 1992). Un autoencodeur est formé de deux sous parties : un encodeur et un décodeur, schématisés dans la figure 21. L'encodeur cherche à condenser l'information donnée en entrée pour la faire tenir en un nombre très restreint de dimensions, qu'on appelle le goulot ou espace latent, puis le décodeur tente de reconstruire les données d'entrée. La performance du modèle est mesurée par l'erreur dans l'exactitude de reconstruction, c'est-à-dire par le nombre de prédictions incorrectes que fait le décodeur à partir du contenu du goulot.

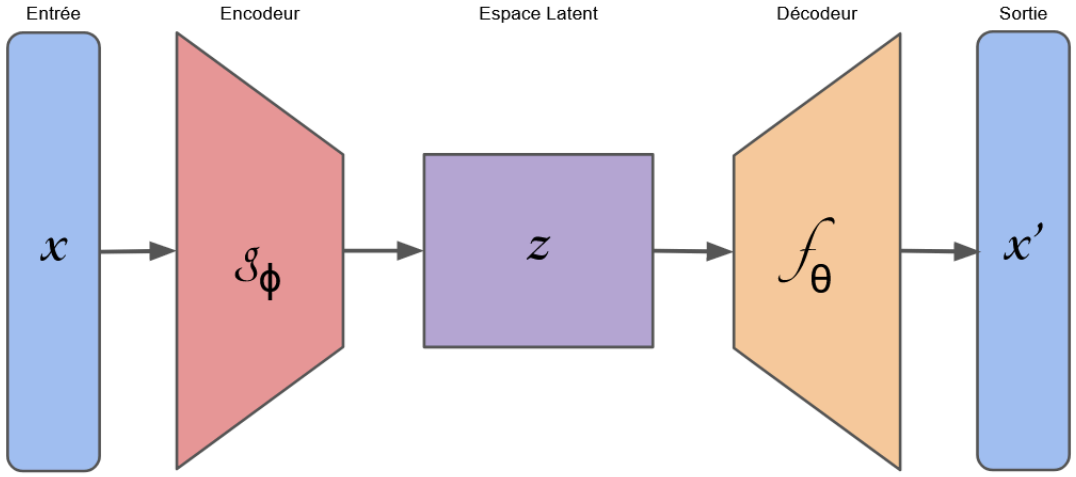


Figure 21 : Représentation schématique de l'architecture d'un autoencodeur. Phi et Theta sont les poids du réseau.

7.2.1.4. Pourquoi la réduction de dimensions

Une contrainte principale sur la simulation de nouveaux génomes humains est la quantité de données à traiter. Comme il peut y avoir au-delà d'un million de mutations sur un seul chromosome, cela entraîne des réseaux de neurones de tailles extraordinaires, et donc des ressources de calcul extrêmement coûteuses.

Les autoencodeurs se présentent comme étant une solution très appropriée pour cette tâche. Étant eux-mêmes des réseaux de neurones, on peut les imaginer comme une extension d'un modèle générateur. L'avantage étant qu'il est plus facile d'optimiser ces sous-parties du réseau total indépendamment, de les entraîner en parallèle, et donc un gain de temps. Le rapport entre le nombre d'échantillons et le nombre de variables à intégrer par section est aussi une puissance : la division en sous-sections permet d'éviter de surcharger le réseau avec une trop grande quantité d'information relativement aux nombre d'échantillons qui portent cette information.

De plus, le décodeur peut être vu comme une forme de générateur. En effet, échantillonner un point dans l'espace latent qui ne résulte pas de la projection d'un vrai génome dans cet espace latent produit un nouveau génome. Il peut cependant être difficile de savoir en avance à quel point ce nouvel échantillon sera distinct de ceux trouvés dans le jeu de données original, et donc si cette fonctionnalité conviendra pour ce projet ou s'il faut intégrer des modèles plus développés.

8. Contributions

8.1. Mon projet

Nous avons maintenant cerné la disponibilité des données génomiques humaines convenables à ce projet, leur volume et leur envergure. Cette ressource est précieuse, et tout projet de séquençage génomique mondial supplémentaire (et compatible) serait le bienvenu pour fournir plus de données.

Bon nombre de modèles d'intelligence artificielle pouvant servir de générateur de nouvelles données sont aussi décrits dans l'introduction de ce manuscrit, ainsi que d'autres servant de méthode de réduction de dimensions. Le nombre important d'approches, de structures et d'implémentations différentes qui existent et qui paraissent régulièrement rendent difficile de tout tester au cours d'un projet de thèse. Même si je rencontre des limites ici, d'autres projets d'un même genre sauront les surmonter ou les contourner à l'avenir, munis d'expérience plus ample ainsi que de nouvelles méthodologies.

Tous les points nécessaires sont maintenant présents pour établir une stratégie pour résoudre le problème énoncé dans l'introduction : mettre à l'échelle une méthode génératrice de données appliquée à des génomes humains.

Les données génomiques se présentent sous format vcf, c'est à dire textuelles indiquant la présence ou absence de variations par rapport à une référence. Cette structure est favorable à l'utilisation de réseaux de neurones, de par la régularité des données et le type binaire, qui est simple à intégrer pour le réseau.

Dans (Yelmen et al. 2021), l'équipe applique directement leur modèle générateur aux données génomiques, indiquées sous forme de variables binaires (présence/absence de chaque SNP pour chaque échantillon). L'approche choisie dans le cadre de mon projet de thèse consiste à mettre en place une méthode de réduction de dimensions avant l'étape de génération de nouvelles données. Le résultat de cette étape est un jeu de données sous format numérique, les valeurs pouvant montrer une grande amplitude de diversité. Chacune des nouvelles dimensions de ce jeu

de données représente un concept abstrait, qui servira au décodeur comme point de départ pour reconstruire les variants détaillés.

Le générateur est donc chargé de créer de nouveaux échantillons dans ce format bien unique, compréhensible seulement par les modèles créés au cours de ce projet. Il faudra ensuite restituer ces données dans un format compréhensible par l'humain avec les décodeurs issus de la méthode de réduction de dimension. Cette génération est traitée dans les sections 9.1.3.3. et 9.2.3.

Après cette phase de génération, il est nécessaire de vérifier la validité de la population synthétique créée vérifiant notamment qu'elle possède plusieurs critères essentiels. Ainsi que l'ont mentionné l'équipe de F Jay dans leur publication (Yelmen et al. 2021), il s'agit de la diversité, la vraisemblance et le respect de l'anonymat.

Il n'existe pas d'outils standards et reconnus par la communauté pour mesurer ces critères directement. Dans la section 9.3, sont proposées des méthodes originales pour vérifier ces conditions dans des populations synthétiques. Des caractéristiques comme le déséquilibre de liaison et la proximité entre les génomes simulés et leurs plus proches voisins dans le jeu de données de référence se présentent comme de bons candidats pour ces mesures, de par leur simplicité de mise en place et leur pertinence.

8.2. Contributions : Exploration des données

8.2.1. Données de mutations dans la population

Pour le cadre de ce projet, j'ai choisi de me limiter aux mutations ponctuelles et aux courts insertions et délétions. Même si les variations structurales plus grandes comportent également une grande quantité d'information biologique pertinente, les standards établis pour les mutations ponctuelles et indels sont mieux établis, et donc plus faciles à communiquer et partager par la suite. Ces variants sont aussi plus simples à étudier, vu que les grands variants structuraux peuvent interagir entre eux, ce qui complique l'information à stocker.

Ce projet nécessite une source de données répondant à plusieurs critères. Les données de mutations doivent provenir de populations du monde entier, pour présenter autant de diversité que possible au modèle. Un réel effort pour s'assurer de la standardisation des méthodes de détection de ces mutations est nécessaire. Enfin, le nombre d'échantillons doit être élevé, plusieurs milliers de génomes séquencés, pour permettre au modèle d'analyser un grand nombre de véritables échantillons.

Pour toutes ces raisons, j'ai choisi d'utiliser les données du *1000 Genomes Project* (1KG). Ici, les mutations issues du réalignement de la Phase 3 du 1KG sur HG38 (Lowy-Gallego et al. 2019) sont étudiées en détail pour décrire leur motifs de distribution. Le fichier au format vcf contenant les mutations ponctuelles et courts indels sur le chromosome 1 a été téléchargé depuis le FTP de la *European Bioinformatics Institute* : http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/release/20190312_biallelic_SNV_and_INDEL/ . Ces données sont également référencées sur la *European Variation Archive* au numéro d'accession PRJEB31735. Ces mutations sont celles qui ont été identifiées sur le chromosome 1 chez environ 2500 personnes venant du monde entier.

8.2.1.1. Histoire du *1000 Genomes Project*

Ce projet a été initié en 2008 pour créer un catalogue de variants génétiques chez l'humain. L'apparition de techniques de seconde génération a beaucoup motivé cette initiative, en réduisant drastiquement le coût des séquençages nécessaires pour construire une telle ressource.
Le projet s'est déroulé en plusieurs phases. D'abord, il y a eu une phase pilote avec un

séquençage en profondeur de six personnes (deux trios familiaux), environ 200 séquençages génomiques à faible profondeur et environ 700 séquençages d'exons. Ces premiers tests ont servi à estimer les besoins du projet en nombre d'échantillons à analyser pour les phases suivantes, et à évaluer les méthodes mises au point pour des études génomiques à grande échelle chez l'humain. Cette phase a tout de même permis à l'équipe de découvrir un très grand nombre de mutations ponctuelles et de variants structuraux, surtout dans les régions codantes du génome, puisque c'est ici que leurs efforts de séquençage se sont concentrés.

La phase suivante comprenait environ 1 000 personnes, de 14 populations différentes à travers le monde. Une partie de ces données sont réutilisées du projet HapMap (Altshuler, Donnelly, and The International HapMap Consortium 2005), le reste sont des séquençages *de novo*, réalisées spécifiquement pour le 1KG. Lors de cette étape, un grand nombre de variants rares ont été identifiés. Une grande partie d'entre eux montrent une spécificité géographique forte, permettant d'inférer leur date d'apparition approximative : plus une mutation est récente, et plus elle sera spécifique à une ou quelques populations. Pendant cette phase, un effort a aussi été porté sur l'établissement d'haplotypes pour chaque génome séquencé, en se basant sur les données du projet HapMap et en les augmentant.

La troisième et dernière phase (il n'y a pas eu de deuxième phase) a voulu donner un spectre plus large au projet. Le nombre de personnes séquencées a atteint plus de 2 500, et le projet compte maintenant 26 populations à travers le monde. Au total, la base de données compte maintenant plus de 88 millions de variants répartis à travers les diverses populations dans le monde. Le projet recense aussi les variants structuraux, et les classifie selon leur nature et leur impact.

8.2.1.2. Diversité ethnique du jeu de données

Les données qui constituent le 1KG sont diploïdes et proviennent de 2 548 personnes d'ethnies variées. Les meneurs du projet ont réalisé un réel effort pour tenter de diversifier autant que possible les sources des données génomiques.

Lors de la première phase, 12 populations ont été échantillonnées. Elles proviennent majoritairement de l'hémisphère Sud. En passant à la troisième phase, ce nombre est passé à 26

au total, mieux réparties à travers le globe. Ces populations et leurs dates d'intégration dans le projet sont indiquées dans la figure 22.

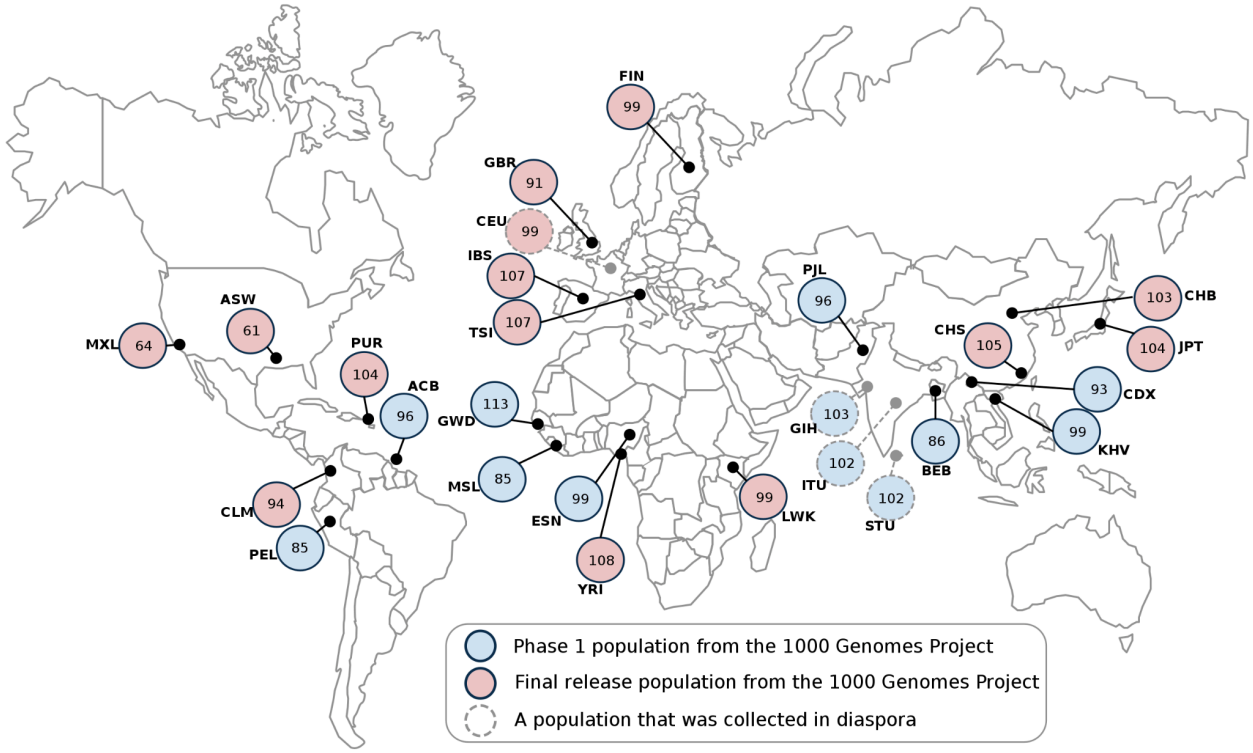


Figure 22 : Carte des populations séquencées pour le projet 1000 Génomes. Tiré de (Oleksyk, Brukhin, and O'Brien 2015)

8.2.1.3. Les données

Pour explorer un peu plus en détail l'organisation des données dans le génome, j'ai choisi de me focaliser sur le chromosome 1, le plus grand (environ 250Mb), pour servir de modèle. Ce jeu de données recense presque 6,2 millions de mutations sur le chromosome 1, à des fréquences très variables.

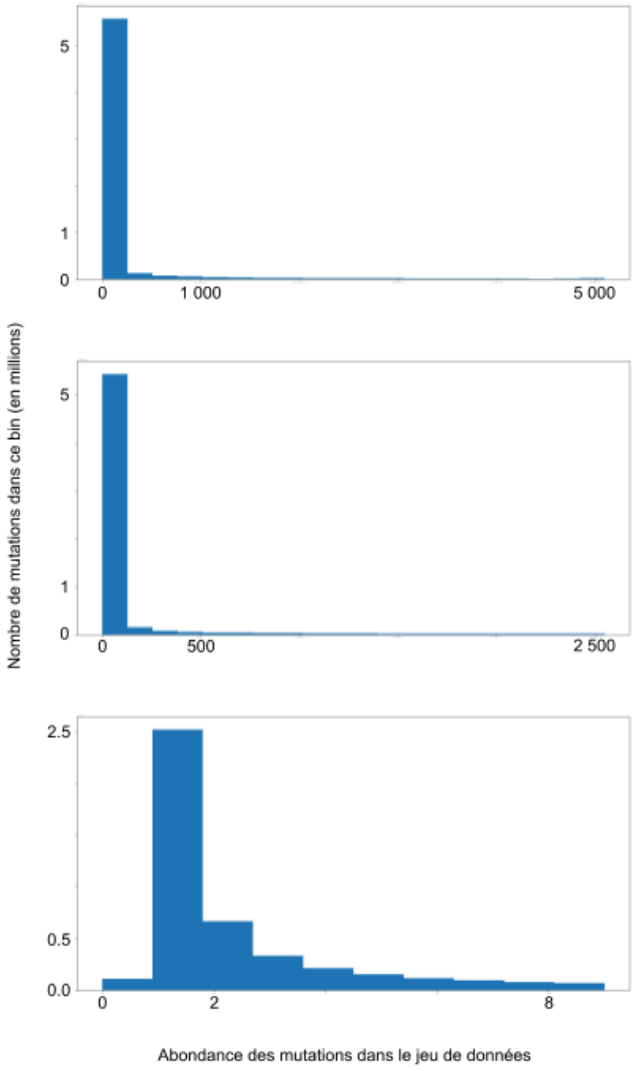


Figure 23 : Taux d'occurrence des mutations sur le chromosome 1. Les mutations sont regroupées en bins de nombre d'apparitions. L'axe horizontal est le nombre d'occurrences de la mutation au total dans le jeu de données, et l'axe vertical le nombre de mutations dans cet intervalle. Le panel A présente la distribution des fréquences brutes, le panel B, ces fréquences après "correction", et le panel C présente un zoom sur les mutations ayant moins de 10 occurrences dans le jeu de données.

Dans la figure 23, panel A, nous pouvons tout de suite observer deux choses:

1. Il y a une très grande quantité de mutations rares, et
2. Certaines mutations ont un taux d'occurrence de plus de 50% (présente dans plus de la moitié des haplotypes du jeu de données, soit toute la partie au-delà de 2500 sur l'axe horizontal).

Le taux d'occurrences de plus de 50% fait référence uniquement à sa présence dans le jeu de données, et non pas les allèles mineurs et majeurs de ces mutations dans la population mondiale. Pour mon projet, qui se porte sur la simulation de profils de mutations, c'est la fréquence de chaque mutation dans le jeu de données qui est intéressante. Une mutation présente à 40% dans la population mondiale mais seulement une occurrence dans le 1KG n'apportera pas d'informations sur l'association de cette mutation à ses voisins. J'ai donc inversé la nature des mutations ayant une fréquence de plus de 50% pour en faire la norme plutôt que la variante. Ces données légèrement transformées sont présentées dans le panel B.

La forte proportion de mutations rares se maintient après cette transformation. Mais il est difficile d'en voir la granularité à grande échelle.

En se focalisant sur les mutations rares (panel C de la figure 23, qui représente uniquement les mutations ayant une occurrence inférieure à 10 dans ce jeu de données), on peut voir que, sur plus de 6 millions de mutations, presque la moitié sont présentes uniquement une seule fois. Comme mentionné précédemment, ce projet de thèse s'intéresse aux variations à la norme et non pas la simulation de séquences d'ADN en entier. Vu que cette stratégie repose sur les associations entre mutations et non pas la position et la nature de celles-ci, je cherche à identifier des liens de corrélation entre mutations. Les mutations présentes une seule fois ne contiennent donc pas d'information sur les liens qu'elles partagent avec les autres mutations contenues dans ce jeu de données. Les mutations ayant une occurrence de deux apportent également peu d'information, notamment en prenant en compte leur nombre important. J'ai donc choisi de filtrer les mutations présentes dans ce jeu de données ayant une occurrence de 2 ou moins pour retirer autant de bruit que possible. Je souhaite tout de même placer ce seuil aussi bas que possible, pour garder autant de détails que possible pour la suite de l'étude.

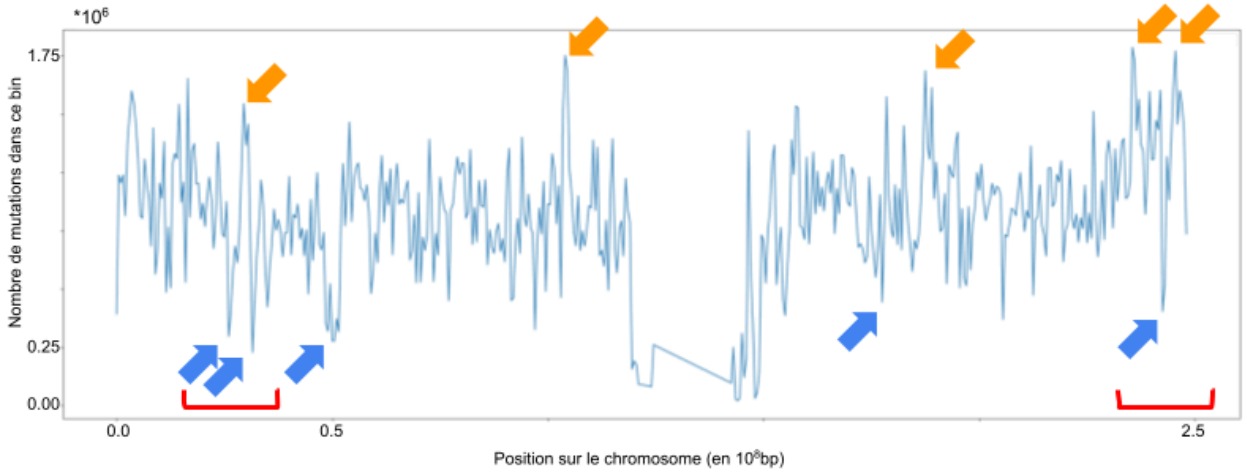


Figure 24 : Occurrences des SNP sur le chromosome 1. Le chromosome est divisé en bins de 500kb, et dans chaque bin, la somme du taux d'occurrence de chaque mutation est calculée. Ces résultats sont présentés ici sur l'axe vertical, avec la position du bin sur le chromosome 1 sur l'axe horizontal. Les flèches oranges indiquent des pics d'enrichissement de mutations, et les bleues, d'appauvrissement.

D'après la figure 24, les taux d'occurrence des mutations ne semblent pas montrer de lien direct avec leur position sur le chromosome, à part pour le centromère, qui ne possède aucune mutation identifiée au cours du projet 1KG. La faible complexité de la séquence de cette région ainsi que son état superenroulé (hétérochromatine constitutive) rend son séquençage très difficile. Seuls certains projets dédiés ont exploré les centromères des chromosomes (Nurk et al. 2022).

Cependant, on peut distinguer des régions qui semblent en contenir plus de mutations que d'autres, avec des "barrières" de faibles mutations entre elles. Cet effet ne présente pas de motif clair, mais le passage de zones très enrichies à très appauvries rapidement (portions indiquées par les crochets rouges) indique qu'il y a très certainement un effet local sur le taux de mutations à ces endroits. Dans la section 8.2.2., nous explorerons une origine potentielle à la variabilité de densité des mutations dans le génome.

8.2.1.4. Utilisation par Yelmen et al.

Dans (Yelmen et al. 2021), l'équipe analyse trois ensembles de mutations tirés de ce jeu de données. Le premier comporte environ 800 mutations échantillonnées dans le génome et définies dans le cadre d'un autre projet comme caractérisant une personne, c'est-à-dire offrant suffisamment de différenciation entre populations pour une réidentification proche du niveau de l'individu. Le deuxième lot en comporte environ 3 300, dispersés à travers le chromosome 15

(leur chromosome d'étude), sur lequel ils ont réalisé des tests de sélection, qui permettent de mesurer si un allèle est sélectionné positivement ou négativement dans une population. Cette application de méthodologie génétique sur des données de synthèse renforce la vraisemblance de celles-ci. Leur dernier ensemble de mutations en contient 10 000, et c'est sur celui-ci qu'ils se penchent pour étudier les liaisons entre mutations, et le degré de fiabilité de leurs modèles sur les reproductions qu'ils effectuent de ces liaisons.

L'équipe a également eu accès à un jeu de données de plusieurs milliers de génomes de la Biobanque Estonienne, qu'ils ont utilisé comme population externe à comparer à leur population de référence.

Quelques détails ne figurent pas dans leur publication, par exemple une méthode de sélection de ces deux derniers lots de mutations, ou un seuil de fréquence basé sur la fréquence dans le jeu de données fourni par le *1000 Genomes Project* (ils utilisent la définition "canonique" de SNP, c'est-à-dire au minimum 1% dans une population). Il serait intéressant de voir les résultats de toutes leurs analyses sur le plus grand ensemble de mutations qu'ils ont employé, pour vérifier que leurs modèles sont capables de maintenir une vraisemblance aussi remarquable sur un jeu de données composé de profils de mutations plus grands.

8.2.2. Corrélation entre mutations

8.2.2.1. Régulation des recombinaisons méiotiques

Lorsque les cassures et recombinaisons se forment lors de la méiose, leurs localisations n'est pas aléatoire. De nombreux facteurs régissent leurs localisations exactes. Cependant, beaucoup de ces facteurs sont complexes et difficiles à déterminer pour chaque individu. Le facteur majeur chez l'humain est l'allèle de PRDM9 que possède l'individu (Baudat et al. 2010). La protéine PRDM9 va reconnaître la séquence d'ADN à cibler pour le chiasma. Cette séquence est toutefois sujette à l'apparition de nouvelles mutations comme toute région du génome, et la dynamique de sélection entre mutations de cette séquence et modifications de la protéine PRDM9 est très intéressante. L'effet final est une variabilité à l'échelle de grandes populations, et donc de la diversité dans le choix définitif des points de recombinaison lors de la méiose.

8.2.2.2. Foyers de recombinaisons

Ces facteurs sont complexes et donc difficiles à modéliser, mais des études de sites de recombinaison au sein de familles permettent de mettre en lumière des sites particulièrement susceptibles de subir une recombinaison (Halldorsson et al. 2019). En particulier, pour arriver à ce niveau de résolution, il faut sélectionner des triplets familiaux (deux parents et un enfant) et réaliser un séquençage génomique en profondeur pour identifier les sites de recombinaison utilisés lors de la création des gamètes qui ont donné lieu à l'enfant.

Les données issues de cette étude sont disponibles sous forme de fichier bedgraph (format texte avec comme colonnes : chromosome, position de début, position de fin, brin, score), indiquant pour chaque position sur le génome le taux de recombinaison à cette position. Les auteurs ont détecté plusieurs facteurs impactant le choix et la position des foyers de recombinaison, comme l'âge de la mère par exemple. Les données utilisées ici sont les moyennes des taux de recombinaison identifiés à chaque locus, établies sur tous les sujets étudiés, peu importe leur sexe ou leur âge. Le taux de recombinaison est exprimé en centiMorgan par megabase d'ADN.

Le centiMorgan (cM) est une unité employée pour mesurer une liaison génétique. Elle est définie comme la distance entre deux positions chromosomiques pour lesquelles on attend un nombre de recombinaisons par génération de 0.01 en moyenne. Elle ne traduit donc pas une distance physique entre ces deux loci, qui sera mesurée plutôt en paires de bases (pb ou bp en anglais).

Exprimer le taux de recombinaison en cM par Mb permet donc de créer une carte détaillée de cette valeur au long d'une section d'ADN (Botstein et al. 1980). Sa valeur vaut la probabilité de recombinaison entre deux sites génomiques A et B divisée par la distance qui les sépare (en Mb).

$$\text{recomb} = \frac{100p(A \times B)}{D_{AB}}$$

Plus cette valeur est forte, et plus il se produit de recombinaisons méiotiques sur ce locus au fil des générations. Les régions ayant un très fort taux de recombinaison représentent donc des

points chauds.

8.2.2.3. Découpe du génome

Ces sites de recombinaison constituent donc des barrières malléables entre sections du chromosome. Leur distribution à travers le chromosome 1 est présentée dans la figure 25. Une mutation située entre deux points chauds présentera une forte corrélation avec les autres mutations dans la même section, mais une faible corrélation avec des mutations dans une section adjacente. Cet effet est le premier facteur qui dicte le taux de corrélation entre mutations, devant la distance qui sépare ces mutations.

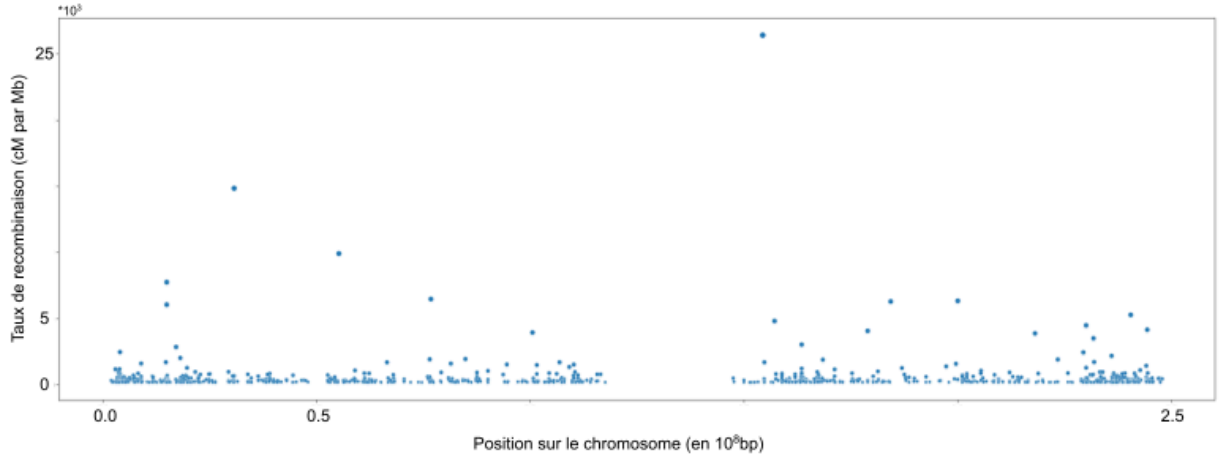


Figure 25 : Points chauds de recombinaison sur le chromosome 1. La position sur l'axe horizontal et la couleur de chaque point dépend de sa position sur le chromosome, et la hauteur du point et sa taille traduisent la "force" du point chaud (en cM par Mb). Les 2000 points chauds les plus forts sont représentés ici.

Dans la section précédente de ce manuscrit, j'ai mis en lumière les motifs de taux de mutation variables le long du chromosome 1. Les foyers de recombinaisons chromosomiques semblent être un bon candidat de séparateurs de ces mutations, mais il est difficile de déterminer le lien entre eux visuellement.

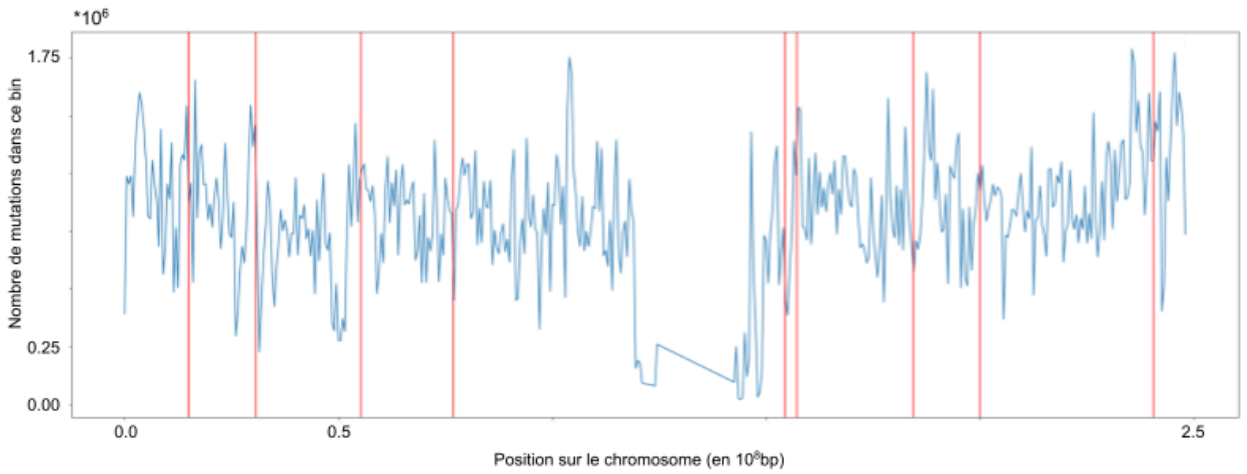


Figure 26 : Occurrences des SNP sur le chromosome 1 en bins de 500kb. Les positions des dix foyers de recombinaison les plus utilisés sur le chromosome 1 identifiés dans (Halldorsson et al. 2019) sont indiquées par des lignes verticales rouges.

D'après la figure 26, les foyers de recombinaison les plus intenses semblent coïncider avec des régions faibles en mutations, ou au minimum ne se situent pas dans des régions très fournies en mutations. Ces données proviennent de populations différentes (islandaise pour l'étude des foyers de recombinaison, mondiale pour le 1KG), donc les foyers de recombinaison les plus utilisés pour ces deux populations ne sont pas forcément les mêmes. La suite de mon travail reposera donc sur la détermination d'un ensemble de foyers de recombinaison montrant à la fois un taux élevé de recombinaisons à ce locus, ainsi qu'une pertinence biologique vis-à-vis des mutations entre ces foyers.

9. Résultats

9.1. Autoencodage des données de mutations

9.1.1. Méthodes de séparation des mutations

Ici, les mutations issues du réalignement de la Phase 3 du *1000 Genomes Project* sur HG38 (Lowy-Gallego et al. 2019) sont divisées en sections selon différents critères. Seules les mutations présentes au minimum trois fois parmi ces ~5000 haplotypes ont été retenues.

9.1.1.1. Séparation par fenêtres de taille constante

Une première approche pour séparer les SNP étudiés consiste à découper le génome en sections linéaires contenant toutes le même nombre de SNP. Cette méthode de séparation est simple et ne prend pas en compte d'*a priori* biologique.

En faisant varier la taille des sections et en observant la distribution des tailles des sections en paires de bases en fonction d'un nombre de mutations fixé, on note principalement sur la figure 27 une diminution de la dispersion pour les valeurs faibles. Cependant, cette première visualisation ne présente pas d'élément permettant de privilégier un nombre de mutations en particulier.

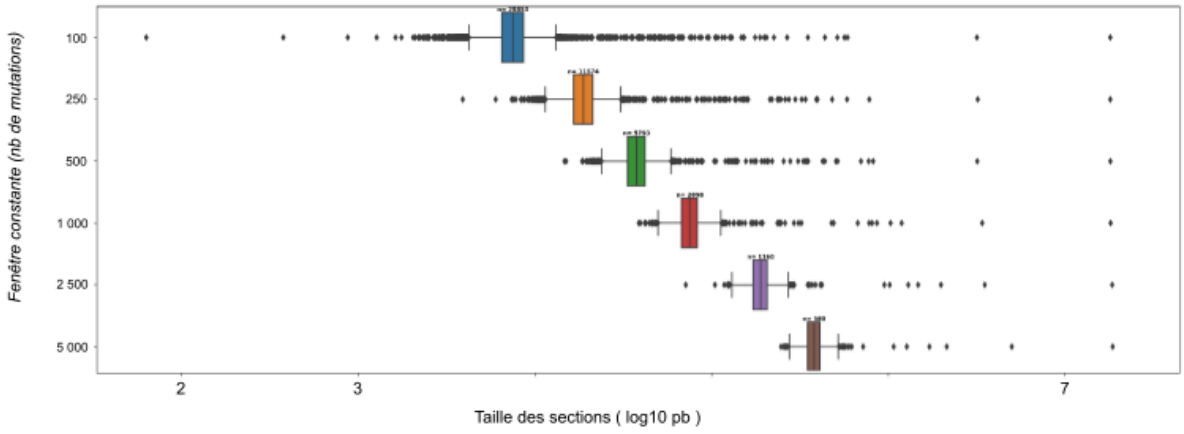


Figure 27 : Distributions des tailles des sous-sections génomiques après découpage linéaire. L'axe horizontal correspond au log10 de la taille en paire de bases de chaque section.

Pour chacune de ces tailles de séparations, des autoencodeurs ont été entraînés sur chaque

section. L'exactitude de reconstruction finale de chaque modèle est utilisée comme paramètre de performance.

9.1.1.2. Séparation selon les TAD

Il est également possible de définir un découpage basé sur des connaissances biologiques. Une organisation importante, et ayant un impact fonctionnel, est le TAD : Topologically Associating Domain. Ces structures de plusieurs mégabases définissent des régions contenant des gènes souvent corégulés et partageant des mécanismes de remodelage chromatinien (Dixon et al. 2012). Il est donc logique de supposer qu'il puisse exister des mécanismes régulant leur héritage générationnel aussi. Bien souvent, ces régions sont définies par des sites de fixation de protéines comme celles faisant partie du complexe cohésine, et donc nécessitent une conservation de leurs séquences pour maintenir leur intégrité. On observe cependant une variation de la location et du nombre de TAD dans des types cellulaires différents (McArthur and Capra 2021). Certains sont relativement spécifiques de certains types cellulaires, tandis que d'autres sont ubiquitaires. En sélectionnant les TAD fortement partagés entre types cellulaires, nous observons dans la figure 28 qu'à partir d'une fréquence d'occurrence de 50% (McArthur and Capra 2021), l'impact sur la distribution des largeurs des sections devient important et entraîne une plus forte variabilité dans la longueur des sections qu'avec le découpage précédent.

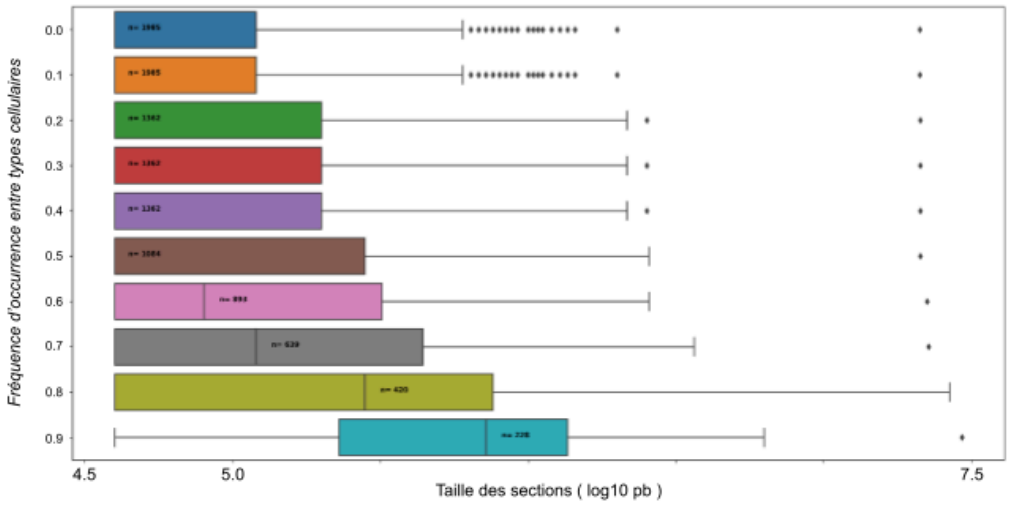


Figure 28 : Distributions des tailles des sous-sections génomiques après découpage basé sur les TAD. L'axe horizontal correspond au log10 de la taille en paire de bases de chaque section. L'axe vertical indique le filtre appliqué au taux de conservation des TAD entre types cellulaires.

Les limites de TAD identifiés comme retrouvés dans 60% des types cellulaires étudiés sont retenus pour la suite de ce travail.

9.1.1.3. Séparation selon les foyers de recombinaison

Une autre organisation consiste à utiliser les points chauds de recombinaison décrits dans la section 8.2.2.2. Ces sites traduisent une organisation de la séquence primaire du génome, là où les TAD portent une information structurale.

Pour sélectionner les points chauds de recombinaison à utiliser en tant que bornes de la segmentation, nous avons fait le choix d'exclure les régions à faible score, puis de réduire chaque région pour n'en garder que le centre. Itérativement, le point ayant le score le plus grand est sélectionné, et tous les autres points dans son voisinage (+/-5kb) sont éliminés, jusqu'à ce que tous les points aient été traités.

Afin de créer des sections de tailles appropriées lors de la séparation des mutations en sections définies par ces points chauds, un minimum et maximum de mutations qu'une section peut contenir est fixé (voir figure 29). Ces valeurs aident à créer des sections de tailles appropriées pour le traitement de ces données. Les trois jeux de valeurs utilisés sont {(100; 1 000), (500; 5 000), (1 000; 10 000)}.

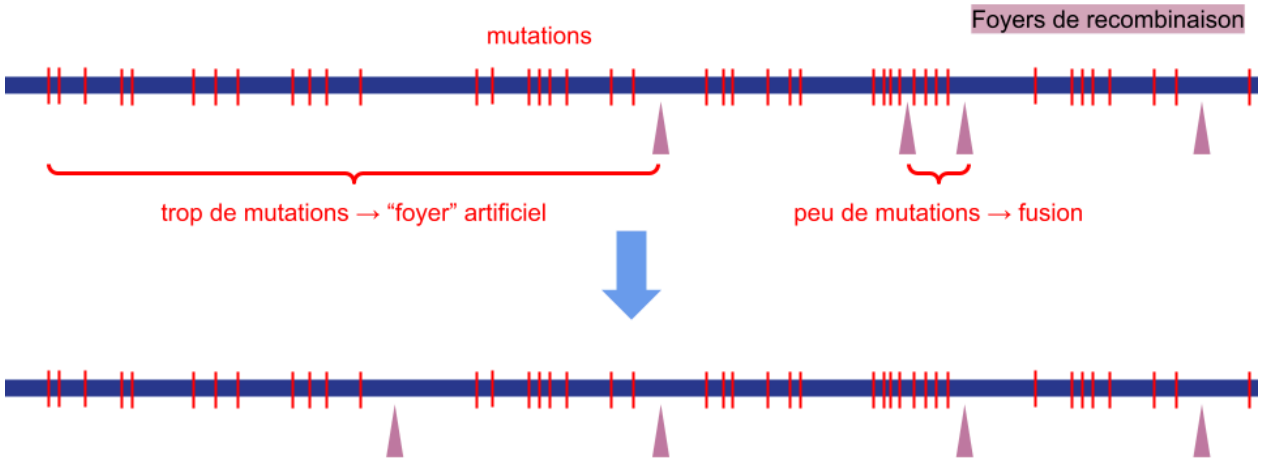


Figure 29 : Schéma du processus de sélection des foyers de recombinaison à utiliser comme délimiteurs entre groupes de mutations.

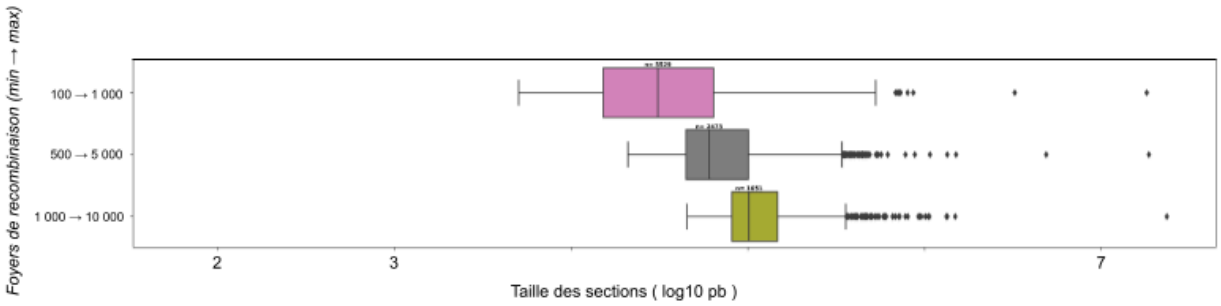


Figure 30 : Distributions des tailles des sous-sections génomiques après découpage basé sur les foyers de recombinaison. L'axe horizontal correspond au log10 de la taille en paire de bases de chaque section.

Comme pour la séparation linéaire, ces trois découpes présentées dans la figure 30 ont été analysées pour mesurer la performance de reconstruction d'un autoencodeur sur les sections produites.

9.1.2. Autoencodage des sections

9.1.2.1. Implémentation

Lors de chaque étape d'entraînement, les données au format vcf (texte, présence ou absence de chaque mutation pour chaque individu, voir figure 31) employées sont séparées en données d'entraînement ou de validation. La séparation est de 80/20.

ID	HG00096.A	HG00096.B		HG00097.A	HG00097.B		HG00099.A	HG00099.B		HG00100.A	HG00100.B	
1.2099402G>A	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.2099427A>G	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.2099538G>A	0	1	1	0	1	1	0	1	0	0	0	0	0	1	0	0	1	0	0	0	0
1.2099594G>A	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.2099609C>T	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.2099610G>A	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.2099617C>T	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.2099662CAG>C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.2099664G>A	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.2099715C>G	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.2099749T>G	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.2099887G>A	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.2099903T>C	0	1	1	0	1	1	1	1	0	0	0	0	0	1	0	0	1	1	1	0	0
1.2099918A>T	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.2099985T>C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 31 : Données aux format vcf haplotypique ayant servi à l'entraînement des autoencodeurs. Chaque ligne contient les informations pour une mutation, et chaque colonne pour une personne. La présence d'une mutation pour un individu est indiquée par un cadre rouge.

Pendant l'entraînement, les performances sur le jeu de données de validation sont

surveillées pour détecter tout phénomène de surapprentissage : si les performances sur le jeu de données d'entraînement s'améliorent mais s'empirent sur les données de validation, alors le réseau se spécialise trop. Lorsque cet événement survient, l'entraînement est stoppé et le modèle revient à son état juste avant la détection de l'événement.

L'entraînement des modèles d'autoencodeur utilise l'algorithme *Adam* comme optimiseur (Kingma and Ba 2017), notamment pour ses capacités à traiter des information très éparses. Il est également efficace en termes de calculs à effectuer, et s'adapte bien à des grands jeux de données. On ne spécifie donc pas de taux d'apprentissage unique au modèle, mais plutôt des hyperparamètres pour guider la mise à jour de cette valeur au cours de l'entraînement.

Les modèles d'autoencodeur utilisés ici implémentent de l'abandon entre les couches d'apprentissage dans l'encodeur et le décodeur, c'est-à-dire qu'à chaque étape de l'entraînement, une partie des neurones sont muets et n'émettent pas d'information, ce qui force l'information à mieux se distribuer entre ces neurones. En combinaison avec la surveillance des performances du réseau sur un jeu de données de validation pendant l'entraînement, cela permet de réduire le risque de tomber sur un phénomène de surapprentissage.

Les modèles créés sont grands et denses, ce qui implique une grande consommation d'espace disque pour les stocker (et pour les partager lorsqu'il le faudra). Il a été démontré qu'un élagage de grands modèles, c'est-à-dire forcer des poids du modèle qui sont proches de zéro à s'éteindre complètement ("couper des branches"), donne lieu à des modèles bien plus pratiques à stocker et partager sans grande perte de performance (M. Zhu and Gupta 2017).

Après ce processus, de nombreuses dimensions de l'espace latent semblent inutiles. Pour vérifier qu'elles ne participent pas à la représentation latente des données génomiques, un deuxième jeu de décodeurs a été entraîné en se basant sur les dimensions latentes que j'ai considéré comme portant de l'information.

9.1.2.2. Résultats des autoencodeurs

Tout d'abord, je me suis penché sur l'utilisation de l'espace latent alloué par l'encodeur. En effet, il est difficile de prédire le taux de compression (ratio entre le nombre de dimensions en entrée et le nombre de dimensions latentes alloué à l'encodeur) maximal qu'on puisse atteindre avec un autoencodeur.

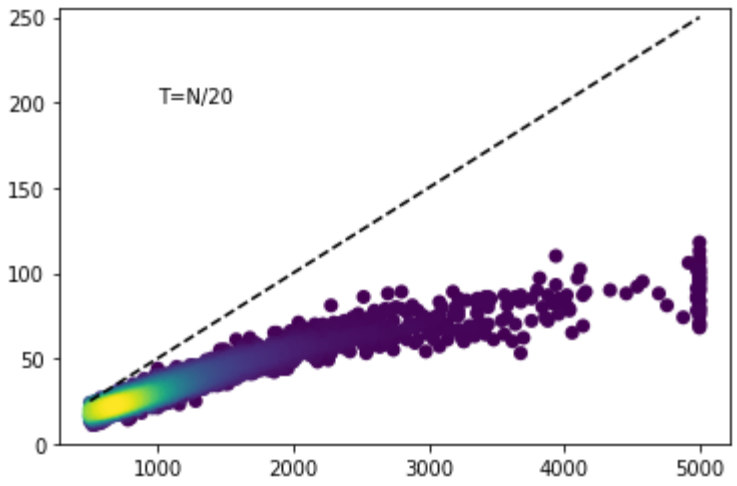


Figure 32 : Taux de compression de mes modèles. L'axe horizontal est le nombre de mutations dans une section, et l'axe horizontal le nombre de dimensions latentes utilisées par l'autoencodeur.

Un taux de compression de 20 fois, illustré par la droite noire en pointillés, a été choisi empiriquement afin de mesurer le nombre de dimensions latentes utilisées. Je considère qu'une dimension latente est considérée comme inutilisée si sa variance est nulle sur les 5000 haplogénomes traités. Il est clair d'après la figure 32 qu'une grande quantité est inutilisée, et que le taux d'utilisation des dimensions latentes n'est pas régulier. Il est donc nécessaire d'implémenter une stratégie pour adapter la taille de cet espace latent après entraînement pour trouver le minimum nécessaire à l'encodage.

Il existe des stratégies implémentées dans certains types d'autoencodeurs pour réduire cette taille après entraînement, mais cela nécessite de multiples entraînements vu qu'il n'est pas possible de déterminer quelles dimensions sont informatives, seulement leur nombre. L'implémentation proposée ici est plus simple: si une dimension présente une variance nulle, alors c'est qu'elle n'apporte aucune information permettant de distinguer entre individus lors de la reconstruction. Afin de valider la pertinence de cette stratégie, une série de décodeurs a été entraînée sur les espaces latents tronqués obtenus après avoir enlevé ces dimensions à variance nulle.

De plus, les connexions entre neurones ayant un poids proche de 0 sont supprimées afin d'alléger davantage le modèle. Ceci permet de réduire autant que possible la taille des espaces latents produits par les autoencodeurs.

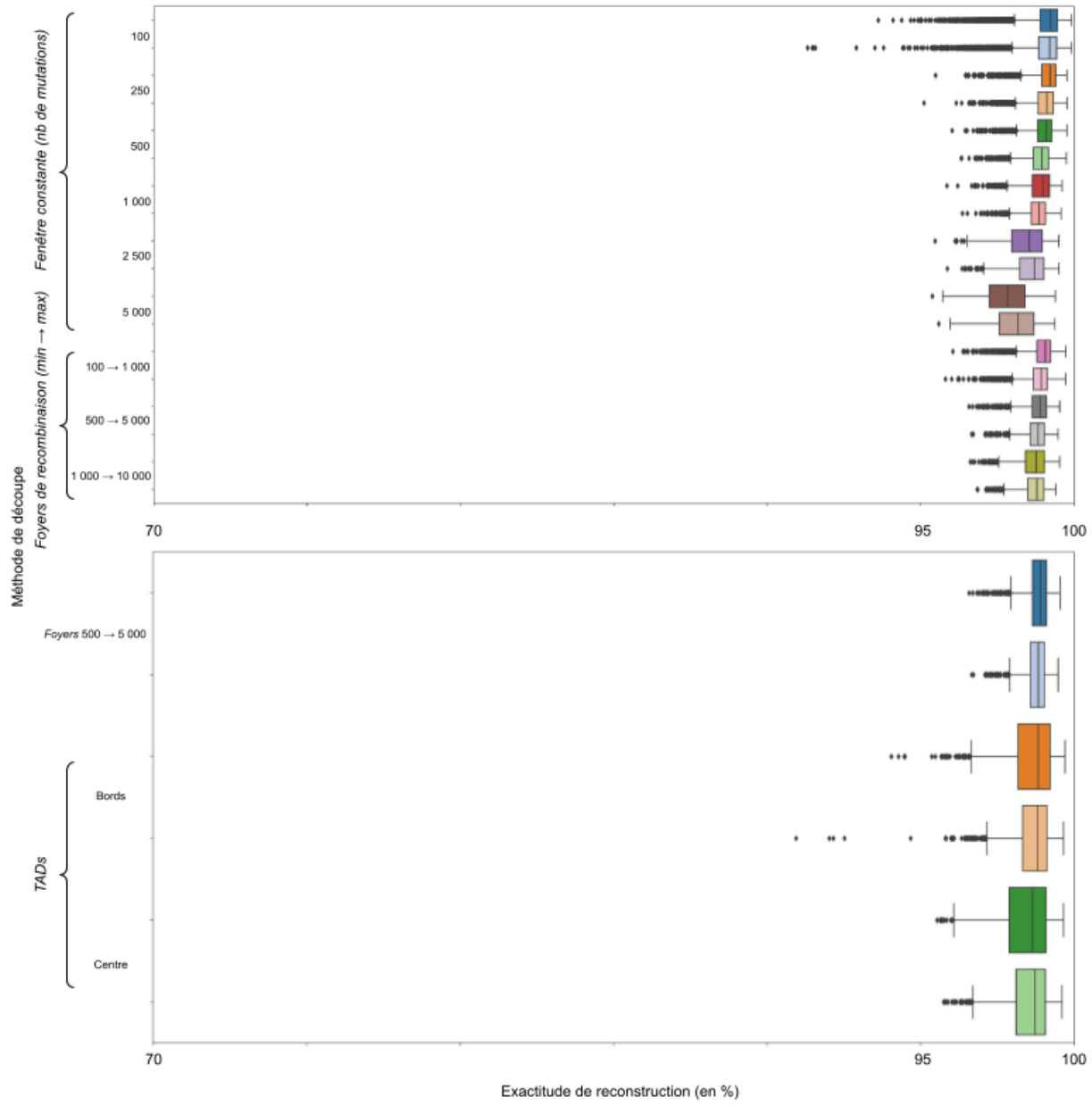


Figure 33 : Distribution des exactitudes finales de chaque autoencodeur selon la méthode de découpage des sous-sections, avant et après réduction de l'espace latent. En couleurs foncées, valeurs après élagage, en couleurs claires, valeurs après réduction de la taille des espaces latents. L'axe horizontal est tronqué à 70% pour mieux illustrer les distributions. Panel A : découpages par fenêtre de nombre de mutations constante allant de 250 (bleu) à 5 000 (marron), puis selon les foyers de recombinaison rose à jaune). Panel B : découpages selon les foyers de recombinaison avec pour minimum 500 mutations et maximum 5 000 (bleu, identique aux distributions en gris dans le panel A) et selon les limites des TAD (orange) ou les centres des TAD (vert).

Comme illustré sur la figure 33, toutes les méthodes de division employées montrent des exactitudes très proches. En comparant les couleurs foncées aux couleurs claires, il est clair que

la réduction de l'espace latent n'a pas de grand impact sur l'exactitude de reconstruction du modèle. Il est donc raisonnable d'appliquer cette stratégie de réduction de l'espace latent.

En comparant les moyennes d'exactitude de reconstruction de ces méthodes, on n'observe pas plus de 2% de différence. C'est donc que ces variations sont très négligeables. La plus grande différence entre ces distributions est la présence d'une "queue" de valeurs extrêmes basses pour une partie des méthodes, notamment lorsque la taille des sections est réduite.

En tant que contrôle interne pour la séparation basée sur les TAD, j'ai aussi effectué une séparation en prenant les centres des TAD, pour retirer l'aspect fonctionnel de cette méthode. Ainsi, on perd la notion de la pertinence biologique des éléments dans le TAD, et on utilise uniquement la distribution des tailles de ceux-ci.

Il ressort également de cette figure que le découpage basé sur les points chauds de recombinaison semble plus efficace que les deux autres stratégies, mais le gain n'est pas très important.

Un élément important à prendre en compte pour le choix du découpage est la taille des sections. Plus les sections contiennent de mutations, et plus on pourra augmenter le niveau de compression que les autoencodeurs appliquent aux données.

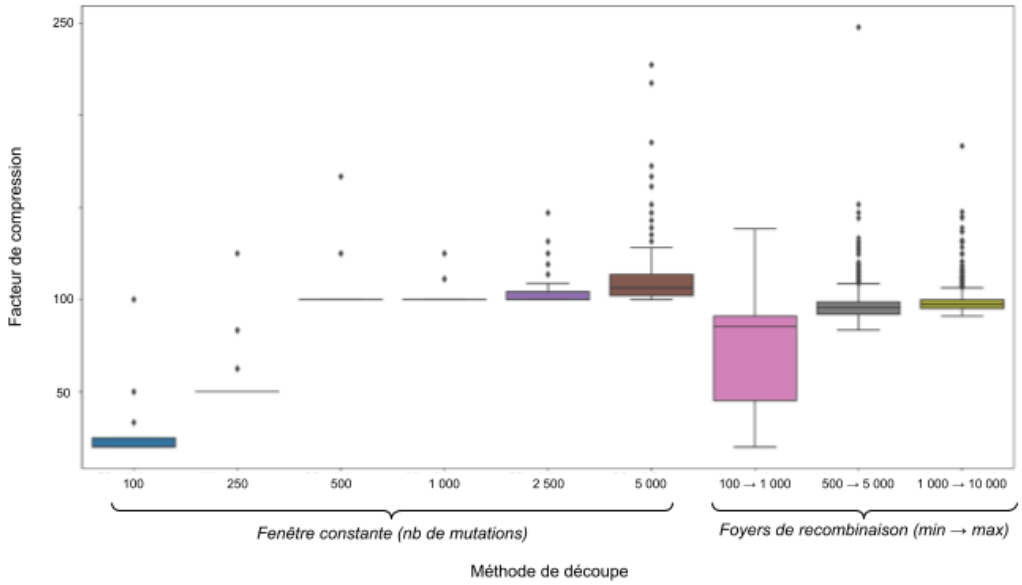


Figure 34 : Distribution des facteurs de compression des sections génomiques pour les différentes méthodes de découpe.

D'après la figure 34, il est clair que les méthodes de découpe autorisant des petites tailles

de section rendent difficile la compression efficace des données génomiques. La méthode donnant le meilleur taux de compression est la fenêtre constante de 5 000 mutations, mais son exactitude de reconstruction baisse de manière remarquable en comparaison aux autres méthodes.

Je choisis donc d'utiliser la méthode de découpage basée sur les foyers de recombinaison, avec pour minimum 500 mutations et maximum 5 000, qui associé à la fois une bonne exactitude de reconstruction et un bon taux de compression.

9.1.2.3. Baisses de performance

Selon notre modèle de corrélation renforcée entre mutations dans une même section de recombinaison, nous pouvons émettre l'hypothèse que ces sections à faible exactitude de reconstruction (trouvés dans la partie basse des distribution présentées au-dessus) contiennent des SNP "mal attribués", c'est-à-dire qui devraient appartenir à une région voisine. Si c'est le cas, alors il existe un ou plusieurs points chauds dans ces sections, alors que les sections montrant une grande exactitude auront tendance à contenir uniquement des mutations cohérentes entre elles, et donc peu de foyers de recombinaison selon notre modèle.

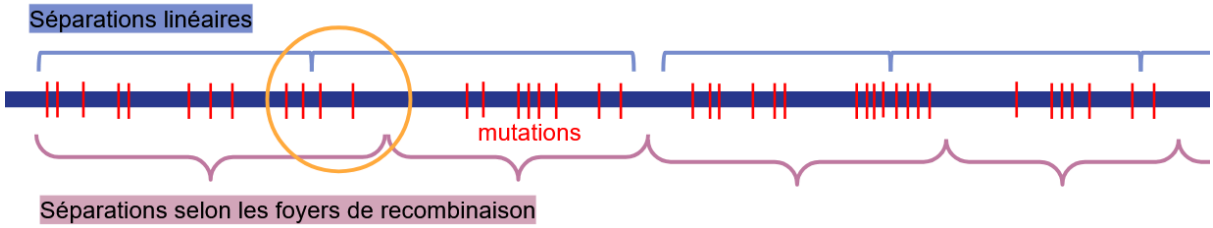


Figure 35 : Schéma illustrant notre hypothèse expliquant la baisse d'exactitude de reconstruction sur certaines sous-sections génomiques selon notre modèle. Les mutations (en rouge) distribuées le long d'un chromosome sont séparées soit par fenêtre constante (bleu, au-dessus), soit en se basant sur les points chauds de recombinaison (violet, au-dessous). Encerclé en orange, un cas où la séparation linéaire rompt l'harmonie entre mutations et sépare un groupe hautement corrélé, alors que l'autre méthode préserve ces liens.

L'idée est de vérifier s'il existe un lien entre baisse d'exactitude de reconstruction et nombre de points chauds de recombinaison. Si c'est le cas, nous pourrons en conclure que ce sont bien ces points chauds qui régulent les corrélations entre mutations. Dans ce but, j'ai séparé les sections en déciles d'exactitude de reconstruction croissante (1er décile = faible exactitude, 10ème = meilleurs). Ces déciles sont illustrés sur les cartes de chaleur dans la figure 36 pour les

méthodes séparant les mutations en fenêtres de taille constante dans une première carte, puis les séparations basées sur les bords et les centres des TAD dans une deuxième carte.

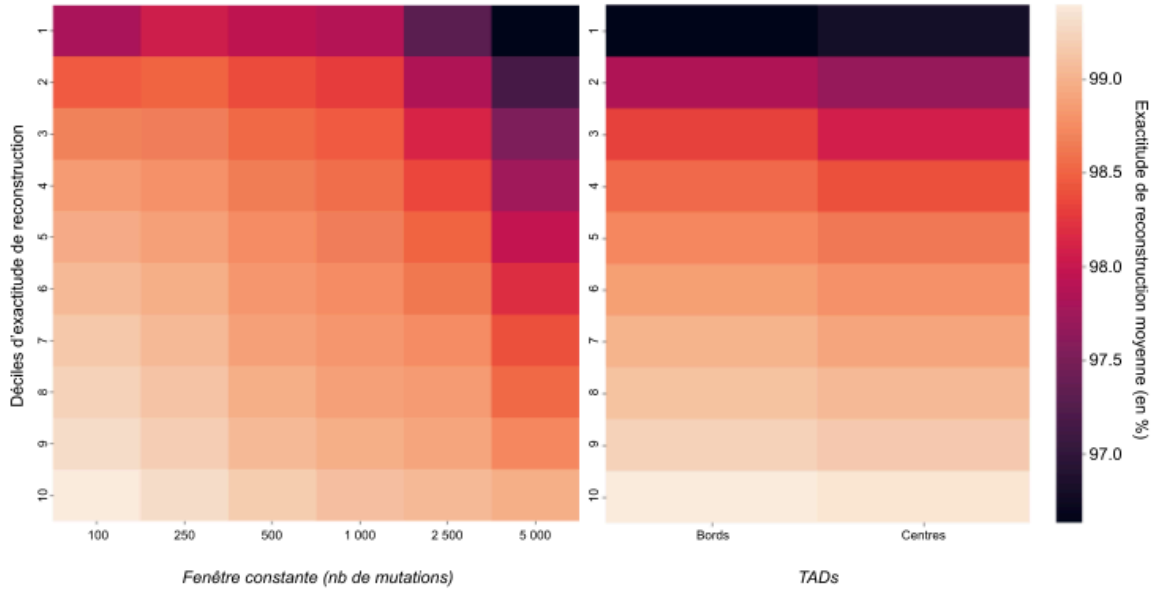


Figure 36 : Carte de chaleur des déciles d'exactitude de reconstruction des sous-sections génomiques selon la taille des fenêtres utilisées.

L'objectif est de déterminer si les déciles de faible exactitude contiennent plus de foyers de recombinaison que celles montrant une meilleure performance. Pour cela, le tri en déciles effectué juste au-dessus est repris, mais la coloration est effectuée selon le nombre de foyers qu'ils contiennent :

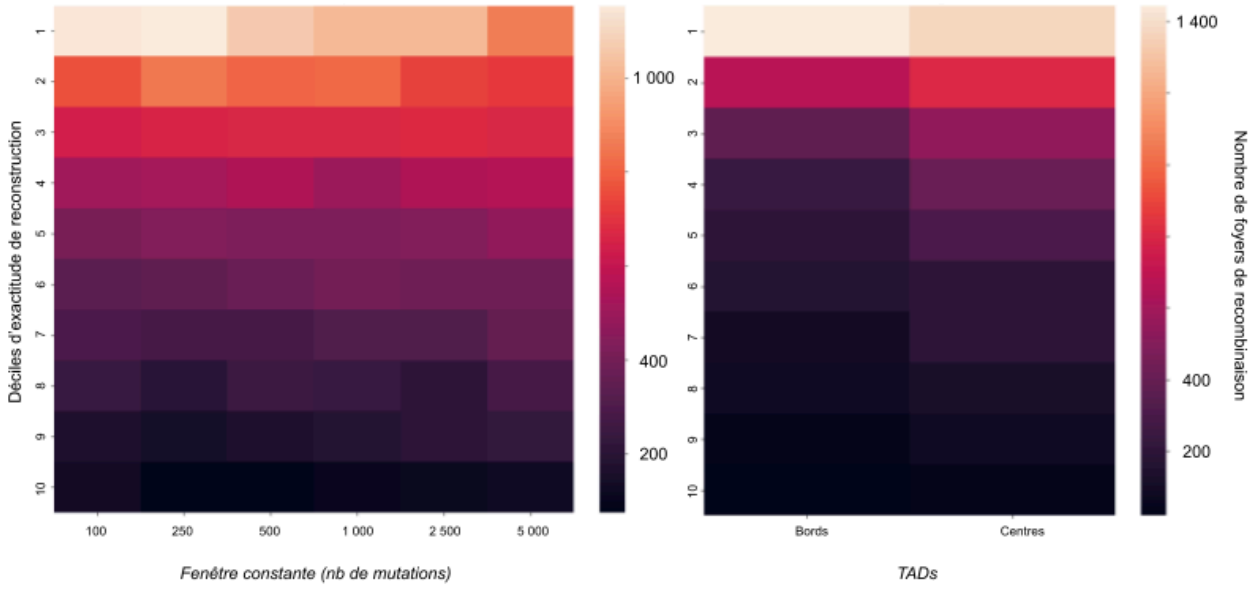


Figure 37 : Carte de chaleur indiquant le nombre de points chauds de recombinaison que contient chaque décile, suivant le tri indiqué dans la carte de chaleur précédente.

Nous observons dans la figure 37 qu'il y a plus de ces foyers dans les déciles de plus faible exactitude. Afin de démontrer un enrichissement statistique pour confirmer que cette distribution inégale est bien corrélée au tri effectué sur les déciles, il est possible d'utiliser la loi de Poisson qui permette d'obtenir la probabilité d'observer le nombre de foyers dans chaque décile si la distribution était aléatoire. Une faible probabilité indique donc un enrichissement statistiquement significatif de foyers dans ce décile. Cette probabilité est représentée sous forme de logarithme décimal négatif, pour traduire une probabilité très faible en valeur entière très grande.

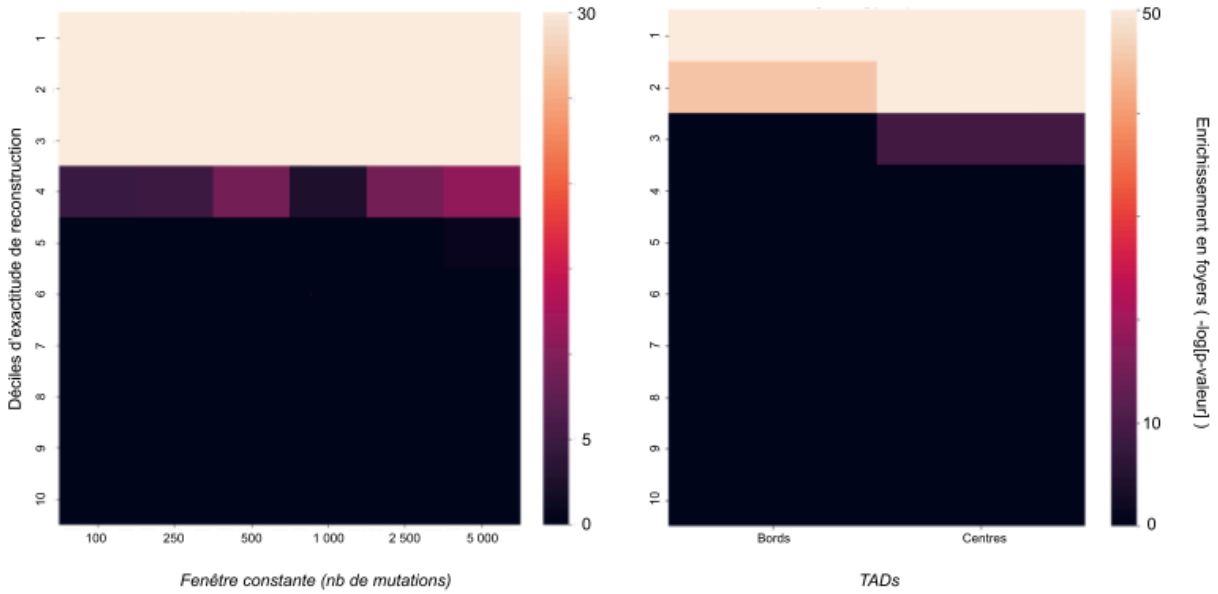


Figure 38 : Carte de chaleur de la significativité statistique de l'enrichissement des déciles d'exactitude de reconstruction, calculée avec une loi de Poisson.

D'après les cartes de chaleur de la figure 38, il existe bien un lien fort entre baisse de l'exactitude et enrichissement en points chauds de recombinaison, donc que la perturbation des liaisons entre mutations par les bords des fenêtres de taille constante correspond à une performance amoindrie. Ainsi, nous pouvons affirmer qu'utiliser ces sites en tant que séparateurs de données génomiques présente un réel gain d'information pour l'application de méthodes de réduction de dimensions.

Cette structuration des données, mise en évidence par une expertise biologique, est donc très pertinente dans le cadre de ce projet. Je peux m'en servir pour améliorer la performance des modèles générés en comparaison à une approche naïve, et ainsi renforcer la stratégie de simulation de nouvelles données.

9.1.2.4. Prédictions à améliorer

Globalement, le taux de reconstructions exactes de ces autoencodeurs est bon, mais lorsqu'on se penche sur les détails, il est clair que les résultats sont loin d'être optimaux.

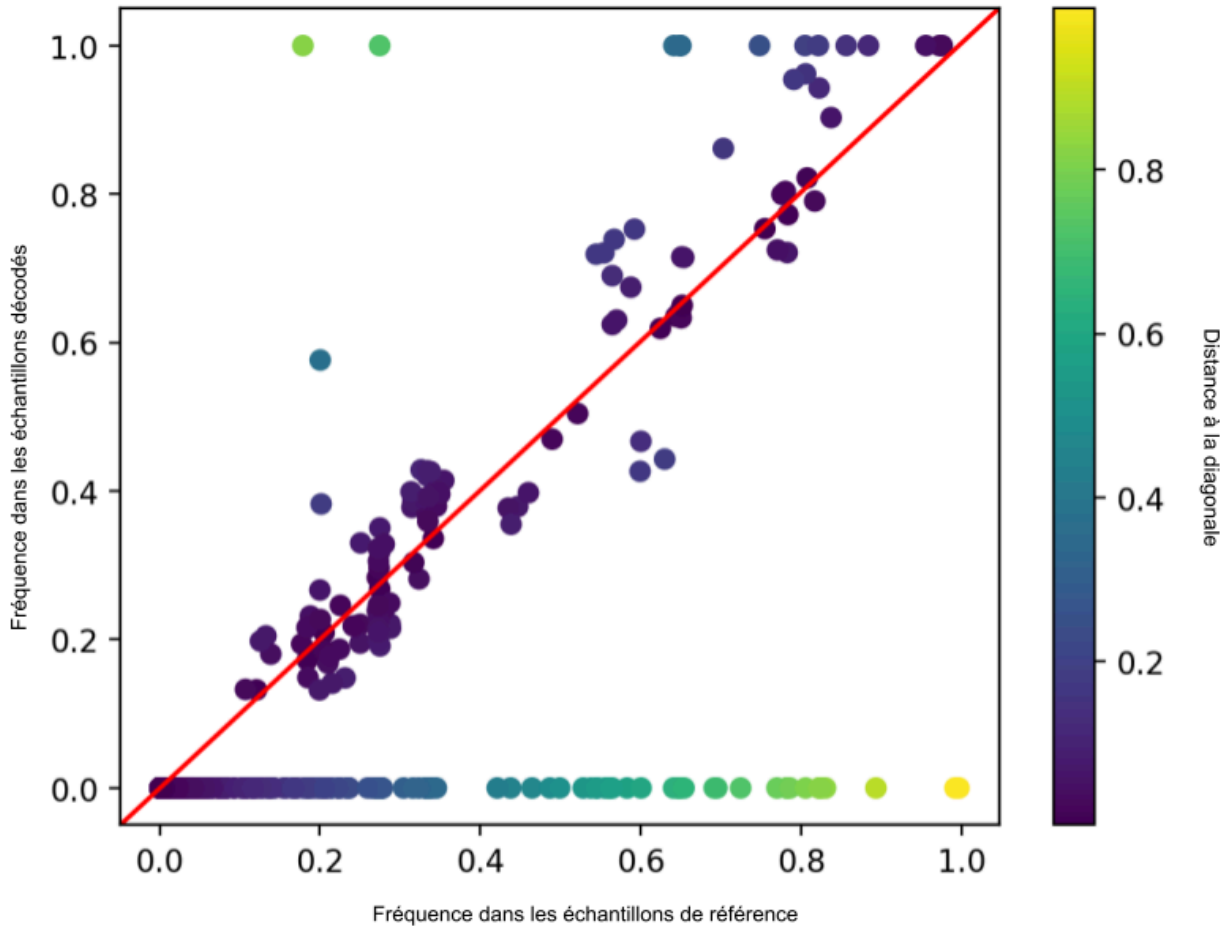


Figure 39 : Nuage de points montrant la divergence des taux d'apparitions des mutations dans le jeu de données de référence et le reconstruction réalisée par l'autoencodeur sur une section. La diagonale est dessinée en rouge, et la couleur des points traduit la distance du point à cette diagonale.

On peut voir sur le nuage de points de la figure 39 que la majorité des mutations sont représentées de manière plutôt bonne, se situant proche de la diagonale. Cependant, pour de nombreuses mutations, le modèle ne parvient pas à introduire de la variance dans ses prédictions. Ce défaut entraîne soit une prédiction toujours positive de la mutation, soit toujours négative.

Une explication possible de ce phénomène est l'absence de continuité entre les points de l'espace latent représentant les données d'apprentissage. En effet, un manque de granularité dans les encodages du modèle peut entraîner des défauts dans la reconstruction des données. Passer à une méthode permettant de mieux représenter cette diversité dans l'espace latent présente un potentiel gain de performances de reconstruction important.

9.1.3. Autoencodeurs variationnels

9.1.3.1. Spécificités

Après discussion avec mes directeurs scientifiques, nous avons décidé de nous orienter vers l'utilisation d'autoencodeurs variationnels. Ils présentent plusieurs avantages qui sont très pertinents pour ce projet :

- Ils permettent de créer des représentations latentes plus diverses qu'un autoencodeur classique.
- Ils traitent mieux les nouveaux échantillons.
- Ils présentent une architecture semblable aux autoencodeurs utilisés jusque-là, et peuvent donc s'insérer dans le projet sans en modifier les autres parties.

Cette méthode repose sur une architecture très semblable aux autoencodeurs présentés plus tôt dans ce manuscrit. Ces deux variantes d'autoencodeurs sont composées d'un élément encodeur, un élément décodeur, et un espace latent entre les deux. La principale différence repose dans la structuration de cet espace latent pour l'autoencodeur variationnel. Comme son nom l'indique, on vient introduire de la variance dans l'encodage des valeurs passées à l'encodeur (Kingma and Welling 2013)

Cette particularité change donc énormément le but à la fois pratique et mathématique du modèle. On ne se contente plus de chercher à optimiser l'exactitude de reconstruction, même si elle entre toujours en compte, il faut aussi contrôler l'aspect de cet espace latent. Mais comment établit-on cette nouvelle architecture d'autoencodeur ?

9.1.3.1.1. Sampling

Dans ce cas de figure, les échantillons en entrée ne sont alors plus représentés par un simple point dans l'espace latent, mais par une distribution gaussienne, comme indiqué dans la figure 40. Cette particularité crée une continuité locale autour de chaque projection dans l'espace latent. L'encodeur prédit donc la moyenne et la variance de cette distribution, et on passe alors au décodeur un point tiré aléatoirement dans cette distribution.

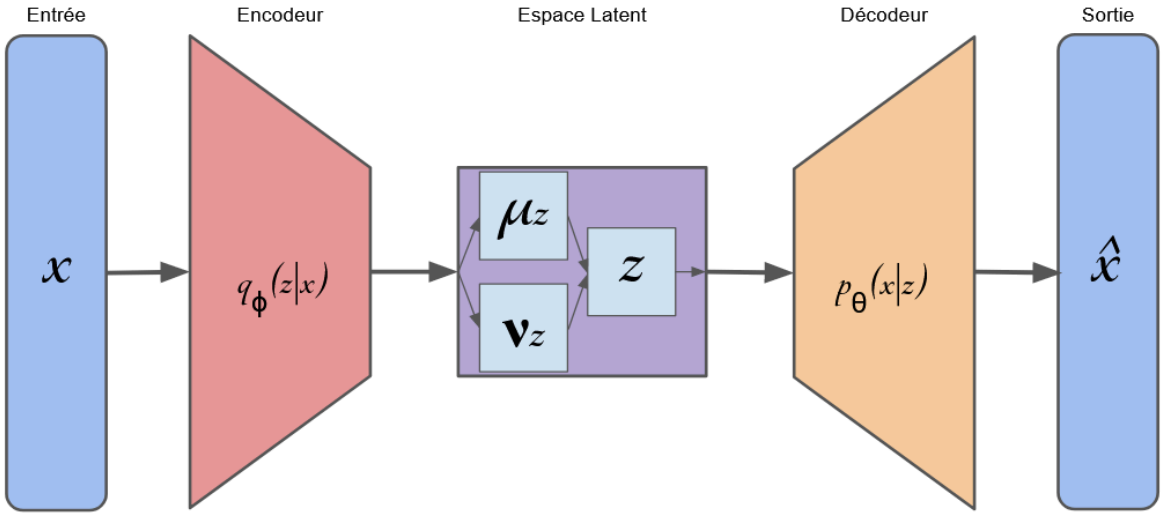


Figure 40 : Représentation schématique de l'architecture d'un autoencodeur variationnel. Phi et Theta sont les poids du réseau.

9.1.3.1.2. Reparameterization trick

L'introduction de variance aléatoire dans le goulot d'un autoencodeur a été théorisé avant d'être implémenté. La communauté scientifique a d'abord pensé que cela serait impossible, puisque le modèle résultant contiendrait de la stochasticité et ne serait pas dérivable. L'astuce de la reparamétrisation a permis d'ajouter cet élément aléatoire sans compromettre l'intégrité mathématique du réseau (Kingma and Welling 2013). Cette méthode introduit l'aléa en dehors du modèle. À la place, le modèle apprend les paramètres qui régissent la loi normale d'où sont tirés les échantillons qui serviront à entraîner le décodeur. Ainsi, il devient possible de calculer les gradients nécessaires à la dérivation du modèle, vu qu'on s'intéresse aux paramètres de distributions de variables aléatoires.

9.1.3.1.3. Divergence de Kullback-Leibler

On vient aussi ajouter un terme supplémentaire à la performance de ce réseau, qui est la divergence de Kullback-Leibler (Kullback and Leibler 1951). Cette distance mesure la différence entre deux distributions, et est utilisée ici pour minimiser la distance entre les distributions projetées dans l'espace latent. La continuité locale s'étend donc à couvrir toutes les projections effectuées par l'encodeur. On cherche à rapprocher autant que possible la distribution des valeurs

que peuvent prendre la projection des échantillons sur une dimension latente à $N(0,1)$. Ceci permet de créer une continuité locale dans l'espace latent. Cette continuité élargit le spectre de compréhension du décodeur et donc augmente les valeurs latentes qu'il peut résoudre pour créer des échantillons cohérents. Au vu de l'objectif de ce travail de thèse, cette propriété est donc une réelle plus value.

Elle n'est pas sans risque pour autant. Lorsqu'on vient contraindre la diversité des représentations que peut créer l'encodeur, celui-ci cherchera à simplifier autant que possible l'information qu'il retient. On voit donc une recherche de mode, c'est-à-dire représenter la diversité continue des échantillons par un ensemble de classes discrètes. Il est possible de diminuer l'importance de ce facteur lors de l'entraînement en le multipliant par un coefficient faible, mais cette pratique est à réaliser avec précaution car elle impacte aussi l'espace latent que crée le modèle.

La divergence KL est employée ici pour pénaliser un réseau de neurones, on s'intéresse donc à sa valeur négative. Sa formule pour deux mesures de probabilité P, Q définies sur R^d et ayant pour densités p,q, respectivement, est donnée dans l'équation (1).

$$-D_{KL}((q_{\boldsymbol{\phi}}(\mathbf{z})||p_{\boldsymbol{\theta}}(\mathbf{z})) = \int q_{\boldsymbol{\theta}}(\mathbf{z}) \left(\log p_{\boldsymbol{\theta}}(\mathbf{z}) - \log q_{\boldsymbol{\theta}}(\mathbf{z})\right) \, d\mathbf{z} \tag{1}$$

Appliquée à deux distributions normales, dont une suit $N(\ 0,\ I\ )$ et l'autre $N(\ \mu,\ \sigma^2\ )$, cette équation devient simplifiable d'après (Kingma and Welling 2013). Sa solution est donnée dans l'équation (2).

$$= \frac{1}{2} \sum_{j=1}^{J} \left(1 + \log((\sigma_j)^2) - (\mu_j)^2 - (\sigma_j)^2\right) \tag{2}$$

Avec J la dimensionnalité de l'espace latent utilisé. Pour tout lecteur intéressé, les développements complets sont disponibles dans (Coste 2022) pour la divergence KL entre deux gaussiennes quelconques et dans (user3658307 2018) pour le cas où une des gaussiennes suit $N(0, I)$.

9.1.3.2. Performances comparé aux autoencodeurs déterministes

Les premiers essais effectués avec les VAE sont concluants : nous retrouvons une bonne

exactitude de reconstruction, comparable, voire meilleure que celles obtenues avec les autoencodeurs déterministes. Ces exactitudes sont représentées dans la figure 41.

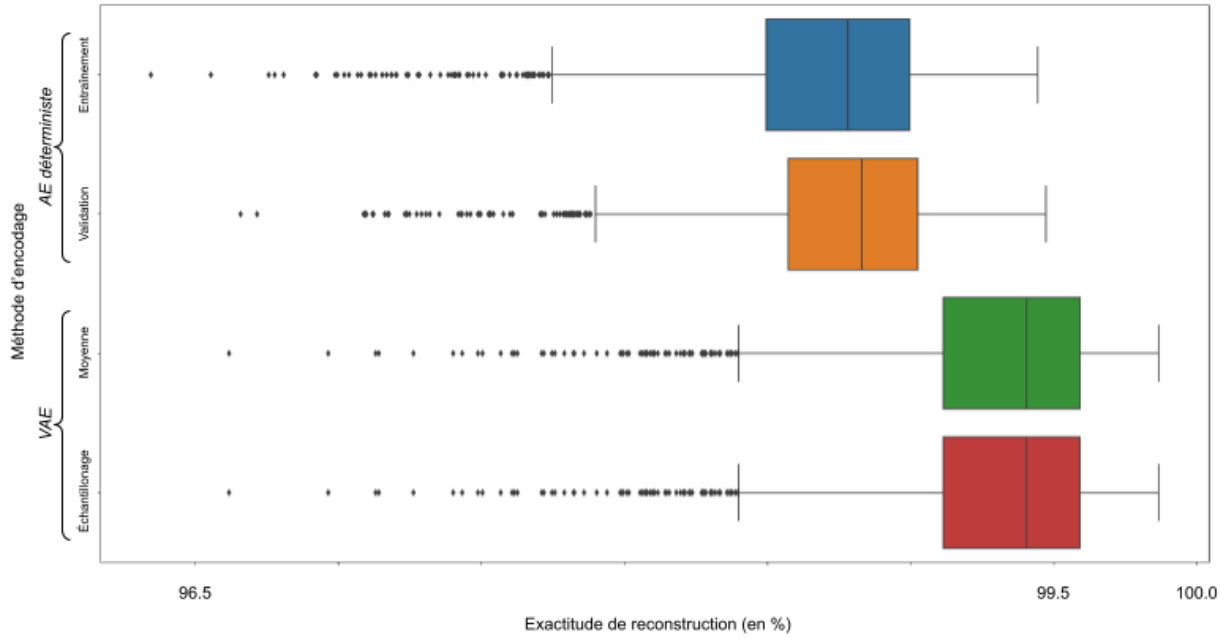


Figure 41 : Distribution des exactitudes de reconstruction des deux approches d'autoencodage. Sont représentés, dans l'ordre, les exactitudes de reconstruction pour le jeu d'entraînement des autoencodeurs classiques, le jeu de validation des autoencodeurs classiques, la moyenne des projections des autoencodeurs variationnels, et un échantillonnage d'individus dans l'espace latent des VAEs

Ce changement de méthode représente aussi un léger gain d'exactitude de reconstruction. Cette architecture de modèle est donc à préférer à la méthode classique pour la réduction de dimension de données génomiques.

Comme avec le modèle déterministe, le taux de restitution de mutations dans les données décodées constitue une métrique intéressante à observer pour évaluer les performances du modèle.

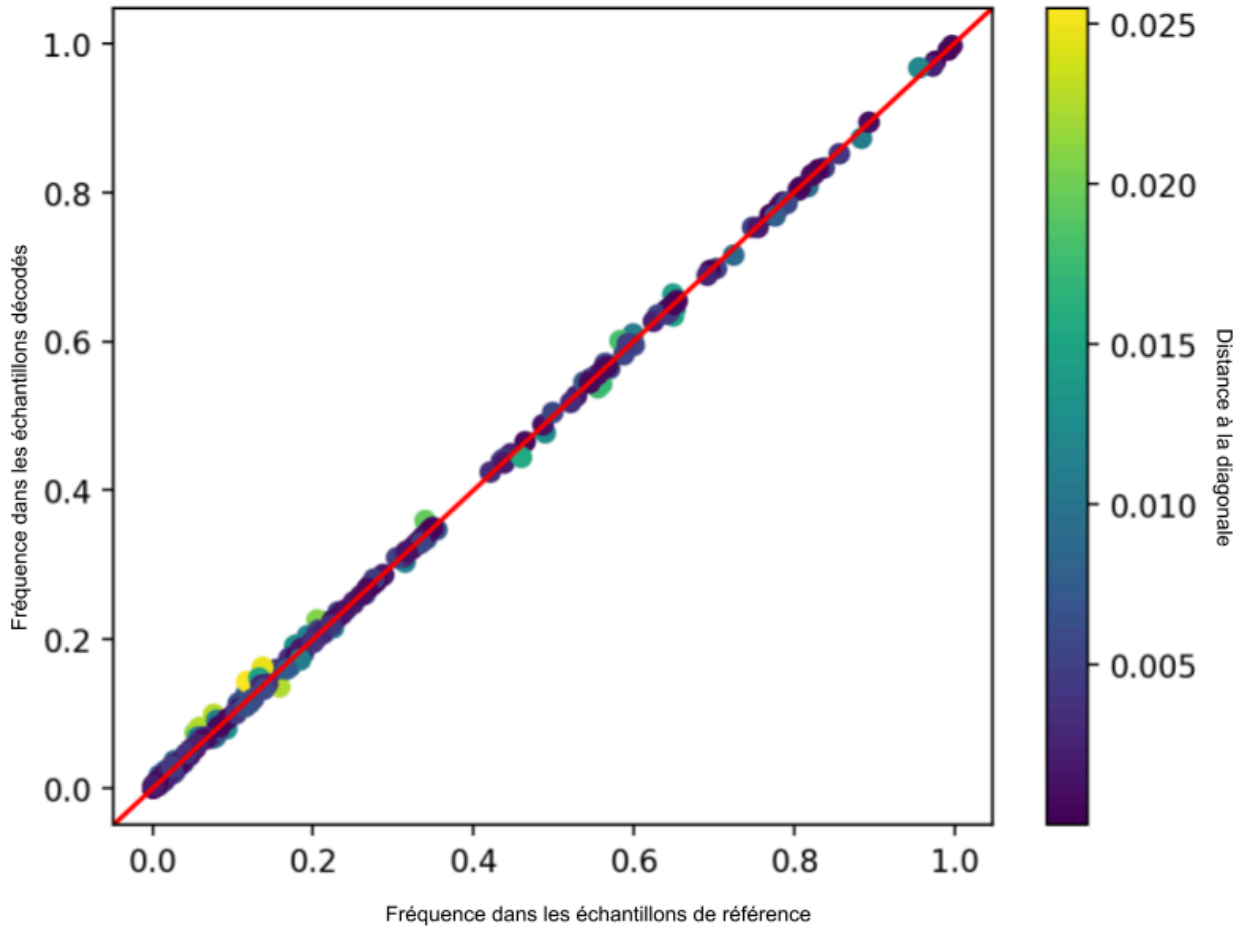


Figure 42 : Nuage de points montrant la divergence des taux d'apparitions des mutations dans le jeu de données de référence et le reconstruction réalisée par l'autoencodeur variationnel sur une section. La diagonale est dessinée en rouge, et la couleur des points traduit la distance du point à cette diagonale.

Il est tout de suite évident d'après la figure 42 que ce modèle montre des performances bien meilleures lors de la reconstruction. Cette stratégie est donc préférable pour la réduction de dimensions de mes données génomiques, même avant de prendre en compte la topologie de l'espace latent.

9.1.3.3. VAE en tant que générateur

Tout seul, un VAE peut servir de générateur. Les projections des individus sur chaque dimension de l'espace latent suivent une distribution proche d'une gaussienne grâce à la contrainte apportée par la divergence KL. On peut donc échantillonner sur cette gaussienne pour créer des points interprétables par le décodeur, mais ne correspondant à aucun individu du jeu de

données d'entraînement.

Il est aussi possible de sélectionner des points à intervalles fixes sur ces distributions pour explorer au maximum l'espace latent. Ce processus est très simple à réaliser et produit des résultats frappants pour des modèles simples à deux dimensions latentes. Un cas d'école pour démontrer ce processus est le jeu de données MNIST (LeCun, Cortes, and Burges 2011), une grande collection d'images de chiffres manuscrits :

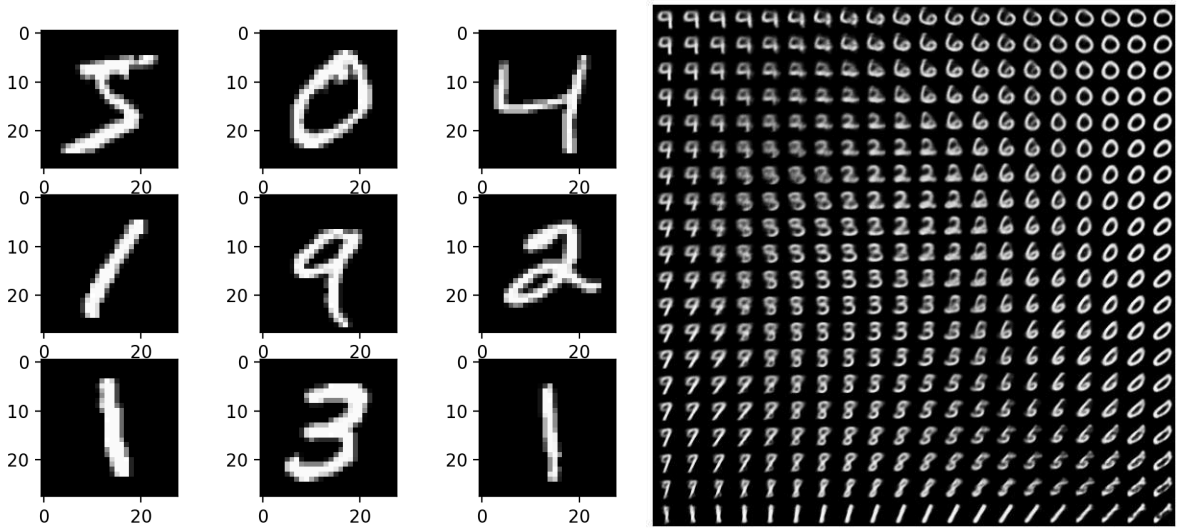


Figure 43 : À gauche, images du jeu de données MNIST. À droite, images issues du parcours de l'espace latent d'un VAE entraîné sur le jeu de données MNIST. Tiré de (TensorFlow 2023)

Dans la figure 43, on a sélectionné 20 points sur chacune des deux dimensions latentes d'un autoencodeur variationnel, et pour chacune des 400 paires de coordonnées, le décodeur produit un "chiffre" correspondant. On peut identifier facilement les régions de l'espace latent contenant certaines classes de ces chiffres. Aux frontières de ces régions, les images produites ne représentent pas un chiffre donné mais plutôt quelque chose qui mélange les chiffres aux alentours.

Il serait intéressant d'observer une telle représentation sur les espaces latents calculés à partir des données génomiques, mais ayant bien plus de dimensions à explorer, il devient difficile d'explorer toutes les combinaisons de points situés à intervalles réguliers. J'ai donc décidé de sélectionner 10 points espacés équitablement sur la distribution gaussienne de chaque dimension latente et de décoder 1000 combinaisons aléatoires de ces points. Ensuite, ces individus simulés et les individus de référence sont projetés dans une Analyse en Composantes Principales.

L'objectif est d'observer la diversité des individus échantillonnés par cette méthode et de la comparer à la diversité des individus de référence, sans avoir recours à une quelconque mesure statistique.

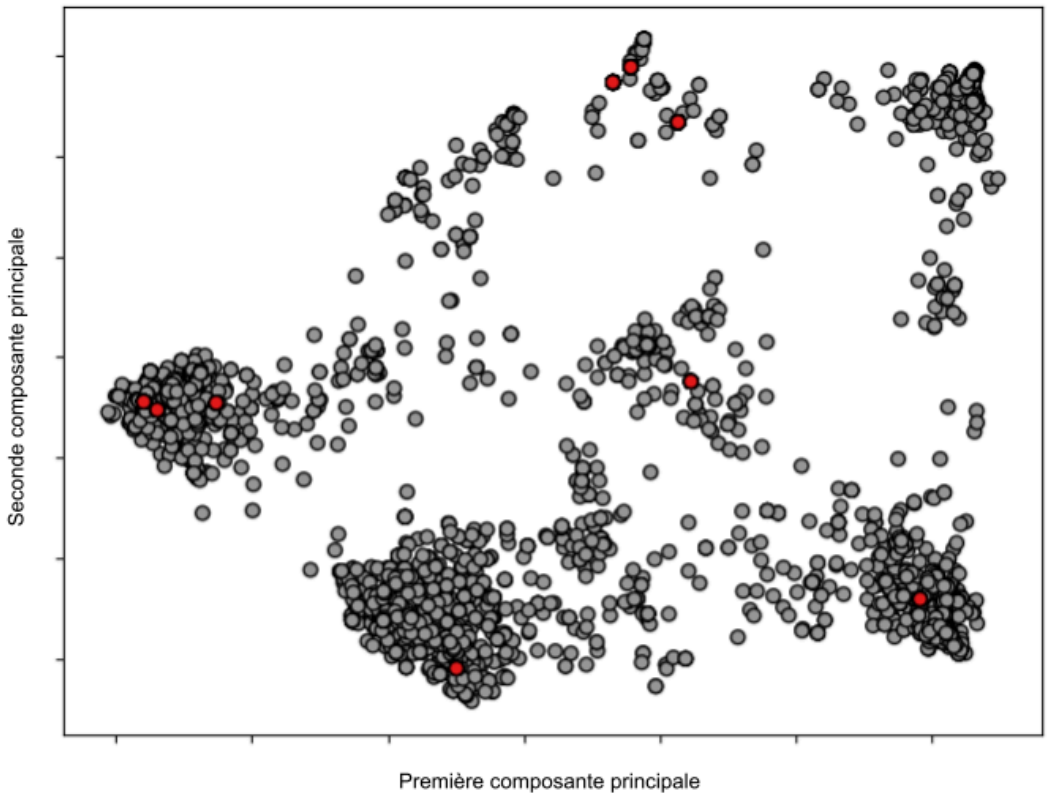


Figure 44 : Projection des génomes de référence (gris) et 1000 simulés par VAE (rouge) (10 points par dimension latente) dans une ACP à deux dimensions.

Il est clair d'après l'ACP dans la figure 44 que cette méthode ne permet pas de parcourir l'entièreté de l'espace latent créé par le VAE. Les 1000 individus simulés projetés se confondent et s'agglutinent en neuf points à travers la distribution. Un parcours plus complet aurait montré plus de diversité autour des modes représentatifs de la distribution complète, de manière à se superposer à la distribution originale autant que possible. J'ai donc poussé les analyses sur ce sujet plus loin, en sélectionnant plus de points sur chaque dimension latente.

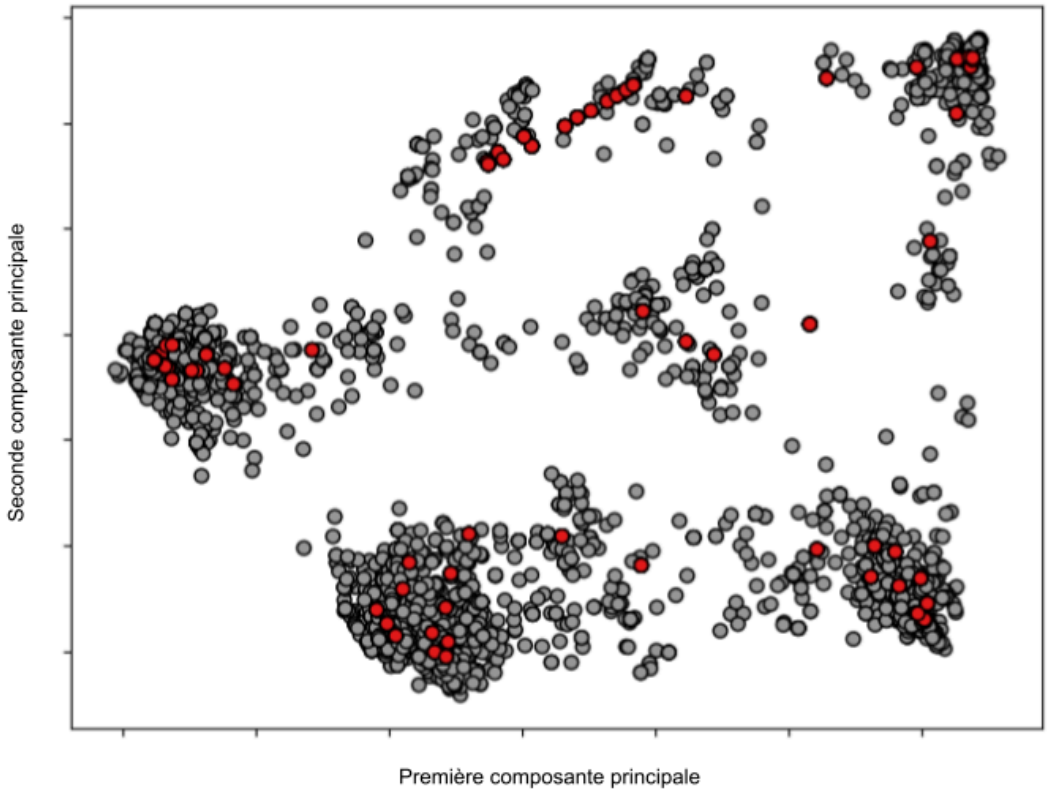


Figure 45 : Projection des génomes de référence (gris) et 1000 simulés par VAE (rouge) (100 points par dimension latente) dans une ACP à deux dimensions.

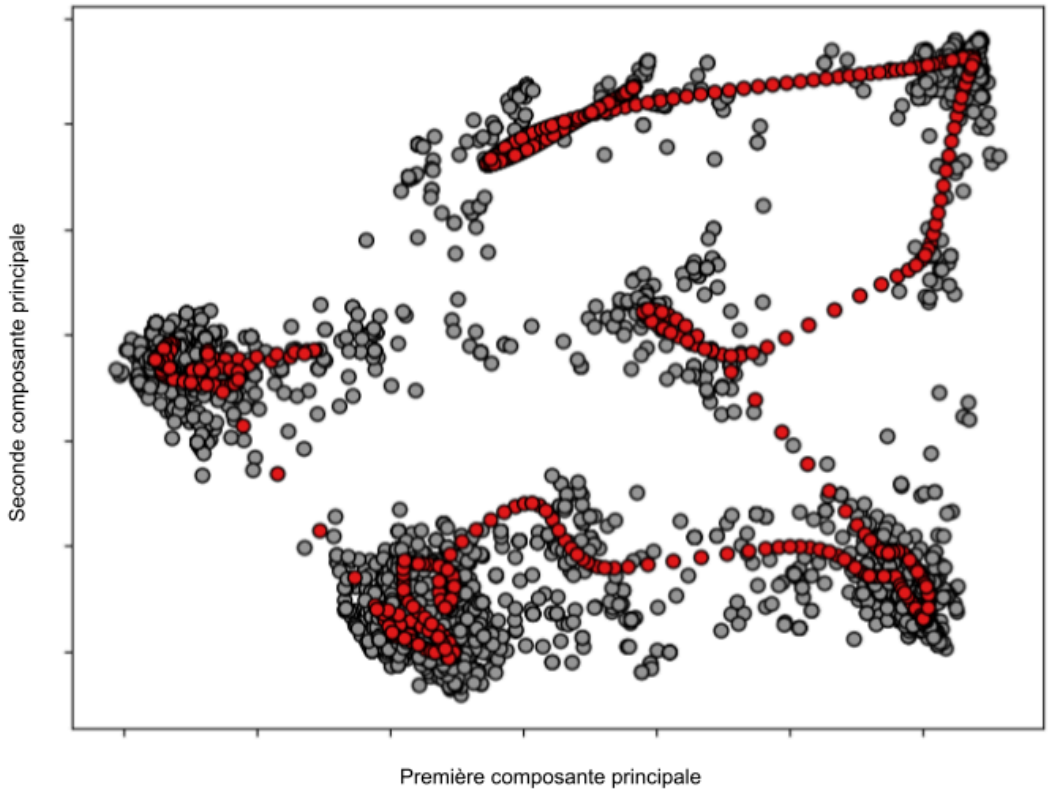


Figure 46 : Projection des génomes de référence (gris) et 1000 simulés par VAE (rouge) (1000 points par dimension latente) dans une ACP à deux dimensions.

En échantillonnant plus intensément dans les dimensions latentes, comme montré dans les figures 45 et 46, il est possible d'observer une meilleure diversité dans les individus générés. Le nombre de combinaisons de points ne semble pas jouer sur la diversité finale. Avec ces informations, l'hypothèse formulée est qu'un petit nombre de dimensions latentes contient énormément d'information, et que les autres ne contiennent que des détails. Cette caractéristique serait difficile à mesurer sans une exploration bien plus approfondie de l'espace latent. Cependant, une étude de l'espace latent du VAE peut informer plus en détails sur ce phénomène.

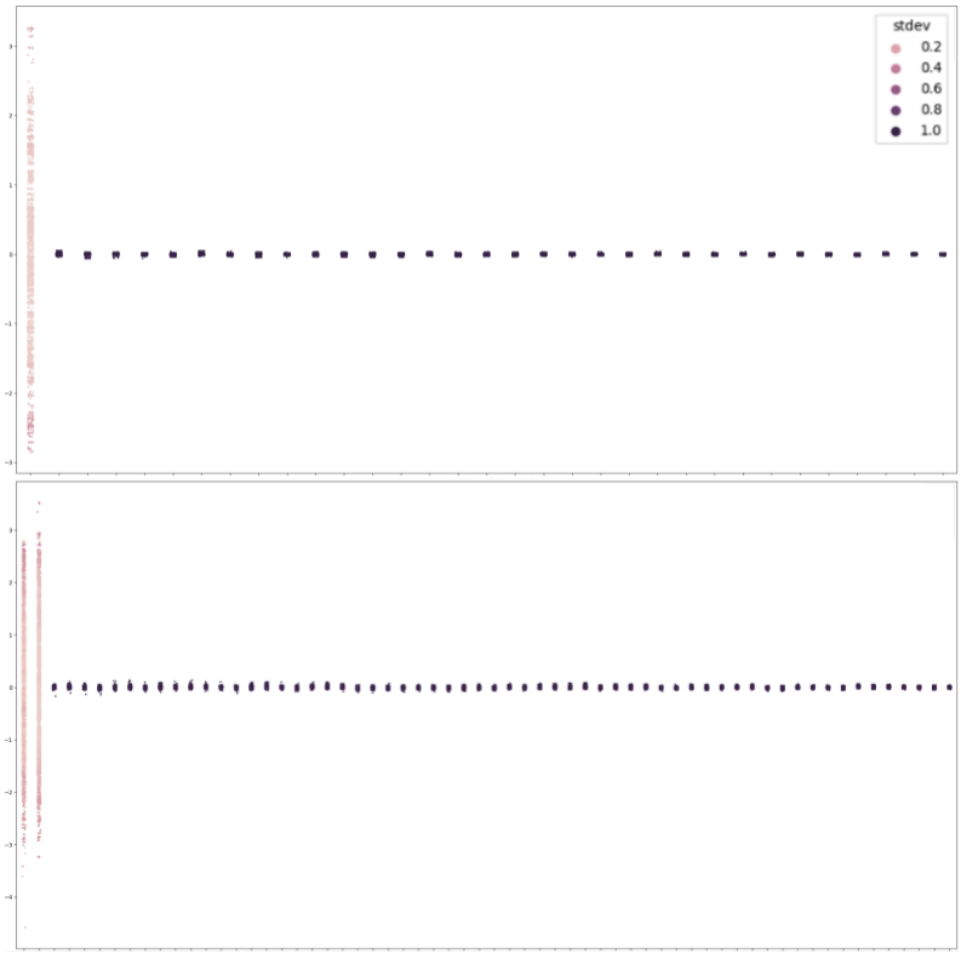


Figure 47 : Diagramme à points des dimensions latentes du VAE. Chaque colonne représente une dimension latente, et la position sur l'axe vertical de chaque point de ceux-ci représente la moyenne de la distribution projetée d'un individu sur cette dimension latente, et la couleur du point, l'écart-type de sa distribution. Deux sections sont présentées ici, la suite dans la figure suivante.

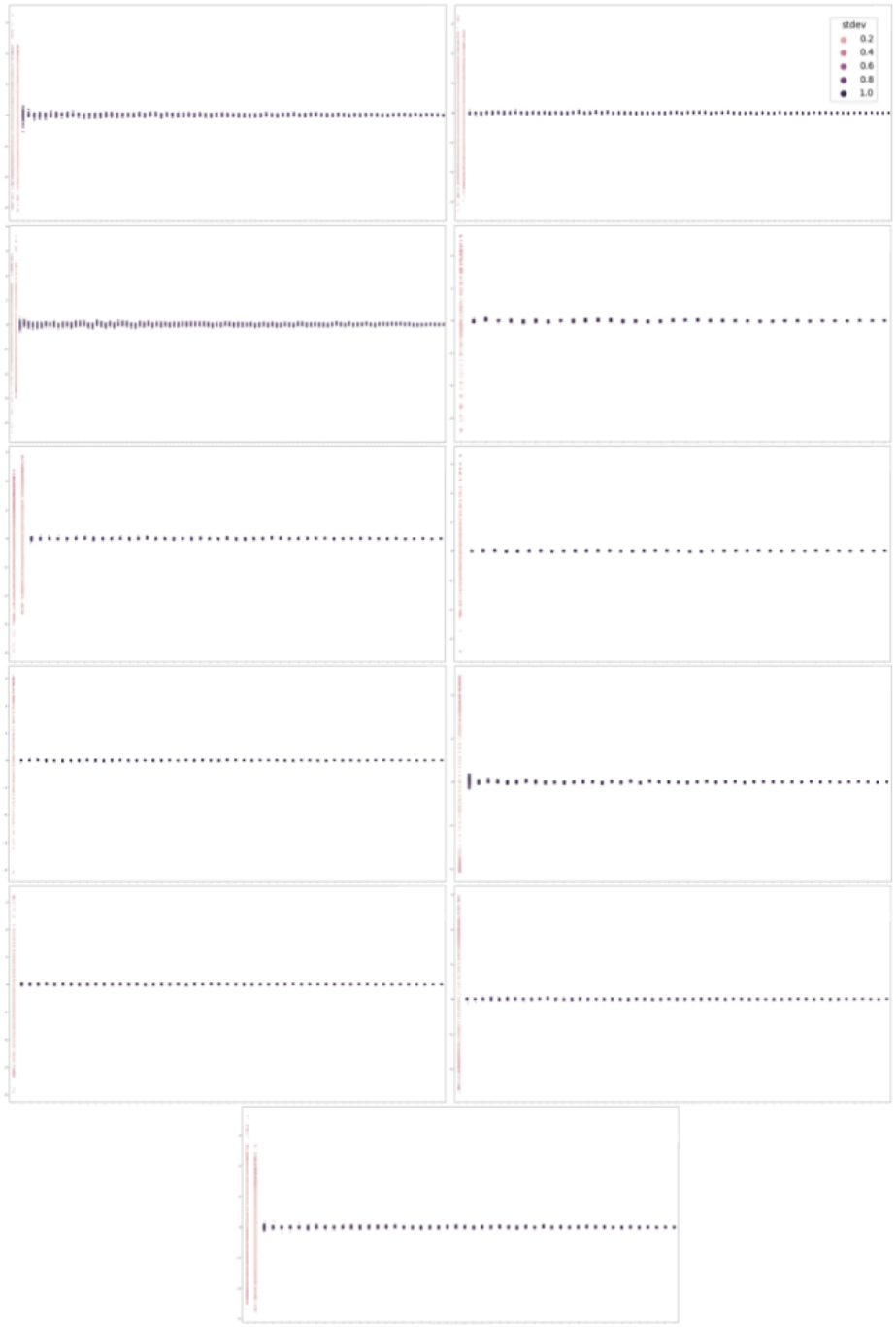


Figure 48 : Diagramme à points des dimensions latentes du VAE. Chaque colonne représente une dimension latente, et la position sur l'axe vertical de chaque point de ceux-ci représente la moyenne de la distribution projetée d'un individu sur cette dimension latente, et la couleur du point, l'écart-type de sa distribution. Le reste des sections analysées sont présentées ici.

D’après les distributions présentées dans les figures 47 et 48, il est clair qu’une dimension possède une amplitude bien plus forte que les autres. Il n’est cependant pas certain pour autant que cette amplitude plus forte traduise plus d’information contenue dans cette dimension.

J’ai testé d’autres configurations du réseau présenté ici pour tenter de modifier cette particularité de l’espace latent, notamment en jouant sur deux paramètres : l’utilisation de l’abandon, et l’impact de la divergence KL sur les dimensions latentes. Pour ces explorations, j’ai également réduit la taille de l’espace latent, vu qu’il n’est pas entièrement utilisé par le VAE pour stocker l’information.

Le modèle original implémente de l’abandon entre les couches cachées du réseau, mais pas entre la dernière couche cachée de l’encodeur et le goulot. J’ai donc testé d’implémenter cette fonctionnalité à cet endroit, pour voir si cela entraîne une plus grande distribution de l’information entre les dimensions de l’espace latent.



Figure 49 : Diagramme à points des dimensions latentes du VAE, après introduction d’abandon juste avant son goulot. Chaque point représente la moyenne de la distribution projetée d’un individu sur cette dimension latente, et la couleur du point, l’écart-type de sa distribution.

La figure 49 montre que ce facteur ne semble pas impacter l’amplitude des dimensions de l’espace latent.

La divergence KL, qui mesure la divergence entre deux distributions de probabilité,

rapproche les distributions des projections encodées autour du centre de la dimension latente. Elle peut donc entraîner un biais dans l'apprentissage, où le modèle sacrifie de l'exactitude de reconstruction pour favoriser un gain d'efficacité dans ce paramètre du calcul de sa performance. En baissant son importance dans ce calcul, le modèle pourrait montrer des comportements différents quant à l'amplitude de ses dimensions latentes. Je l'ai donc divisé par 5 dans la formule de calcul de la perte totale du modèle.

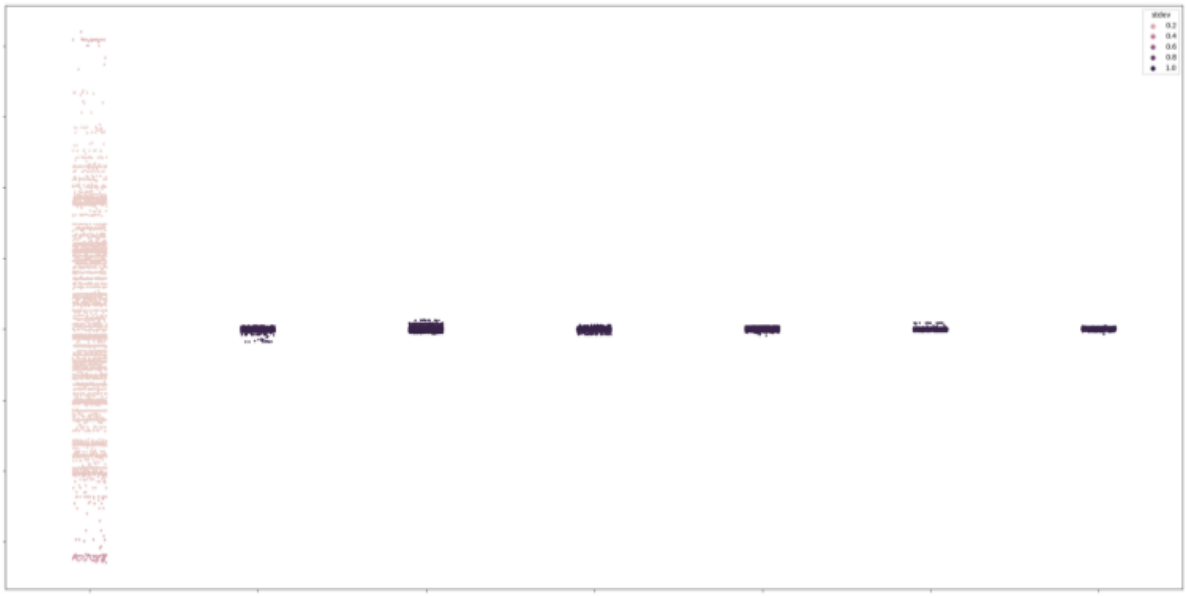


Figure 50 : Diagramme à points des dimensions latentes du VAE, après réduction du poids de la divergence KL dans le calcul de ses performances. Chaque point représente la moyenne de la distribution projetée d'un individu sur cette dimension latente, et la couleur du point, l'écart-type de sa distribution.

Ici aussi, dans la figure 50, cette modification n'impacte pas fortement l'espace latent produit par l'encodeur. J'ai effectué des tests plus poussés avec ce facteur, allant jusqu'à réduire 50 fois son importance dans le calcul de performance du modèle.



Figure 51 : Diagramme à points des dimensions latentes du VAE, après réduction importante du poids de la divergence KL dans le calcul de ses performances. Chaque point représente la moyenne de la distribution projetée d'un individu sur cette dimension latente, et la couleur du point, l'écart-type de sa distribution.

Dans la figure 51, nous pouvons observer, cette fois, un changement d'organisation de l'espace latent : de nombreuses dimensions montrent une variance élevée, et les distributions des individus sur ces dimensions montrent beaucoup plus de variance. Une hypothèse serait que cette augmentation de l'amplitude des dimensions de l'espace latent de mon modèle se traduit par une meilleure distribution de l'information entre ces dimensions.

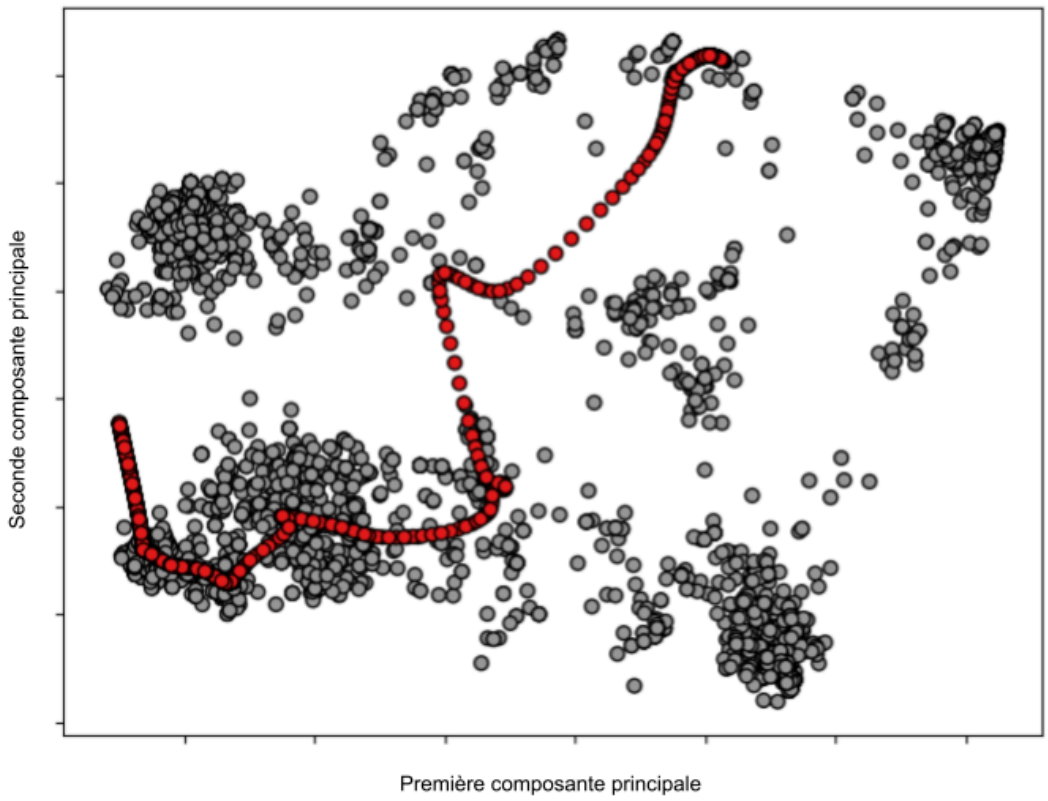


Figure 52 : Projection des génomes de référence (gris) et 1000 simulés par VAE (rouge), après réduction importante du poids de la divergence KL dans le calcul de ses performances, (1000 points par dimension latente) dans une ACP à deux dimensions.

Toutefois, dans la figure 52 nous n'observons pas plus de diversité dans les génomes générés par échantillonnage gaussien de cet espace latent. L'échantillonnage paraît même moins diversifié que précédemment. Ce processus n'a donc pas amélioré la distribution de l'information dans l'espace latent.

9.2. Cohérence entre sections

9.2.1. Pourquoi la cohérence

Les VAE peuvent servir de générateur pour simuler de nouveaux génomes, basés sur les liens de corrélation qu'il a appris sur ses données d'entraînement. Cependant, chaque VAE est appliqué sur une section de manière indépendante. Les génomes simulés seront donc composés de sections indépendantes aussi, et ne montreront aucune cohérence partagée entre sections voisines.

Il est possible de parcourir l'espace latent d'un VAE pour mettre en évidence des liens clairs entre dimension latente et caractéristique finale de l'échantillon produit (Klys, Snell, and Zemel 2018). Dans notre cas, nous avons devant nous des milliers de modèles ayant chacun plusieurs dimensions latentes. Il nous faudrait donc tout parcourir pour tenter de mettre en évidence ces liens, et ensuite déterminer quelles seraient les caractéristiques à sélectionner pour créer des génomes cohérents à grande échelle, pour finalement introduire ces contraintes à nos décodeurs.

9.2.2. Étendeur d'espace latent

9.2.2.1. Principe

Les limites établies entre sections correspondent à des sites où il existe une forte probabilité de recombinaison chromosomique. Ces recombinaisons ne surviennent pas non plus à chaque méiose, et on constate un lien entre l'identité génomique d'une section à une autre. Nous nous sommes donc demandés si un réseau de neurones pourrait prédire cette identité génomique entre régions adjacentes.

En partant de la projection dans l'espace latent d'un individu dans une section génomique donnée, l'idée est de prédire sa projection dans l'espace latent d'une section voisine.

9.2.2.2. Implémentation de l'étendeur

Chaque réseau est composé de trois couches cachées, schématisées dans la figure 52, et leurs architectures exactes dépendent de la taille des espaces latents d'entrée et de sortie de

chacun. Le poids de ces valeurs dans les tailles des couches ont été choisies empiriquement. Les tailles de ces couches ont pour valeurs :

$$C_1 = N_{entrée} + \frac{1}{2} N_{sortie}$$

$$C_2 = N_{entrée} + N_{sortie}$$

$$C_3 = \frac{1}{2} N_{entrée} + N_{sortie}$$

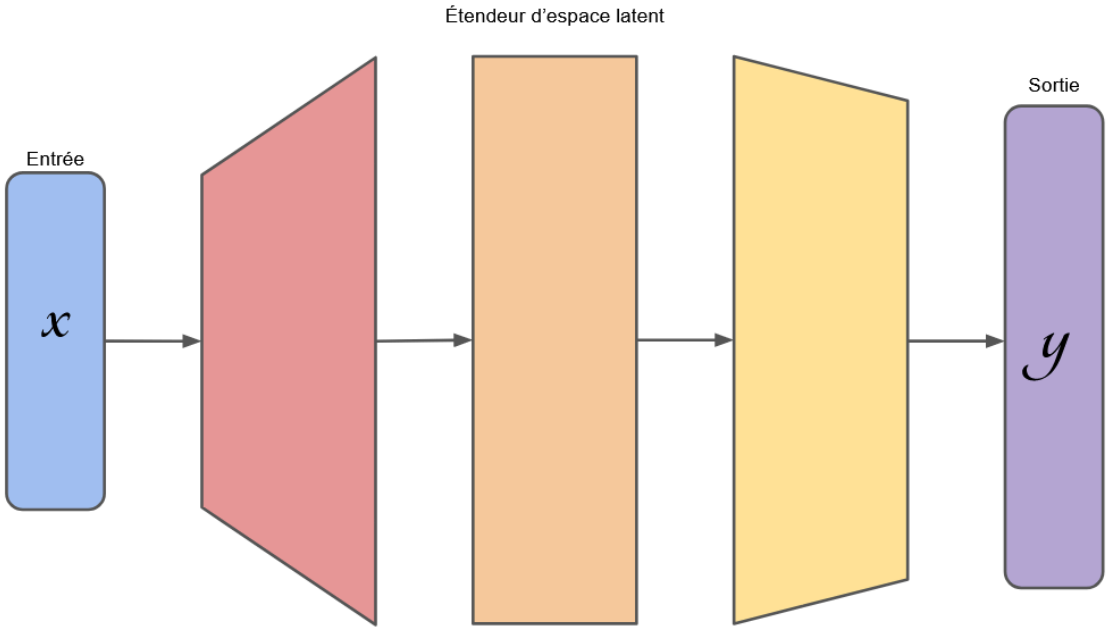


Figure 52 : Schéma de l'architecture de ce modèle créé sur mesure. X représente les représentations latentes de génomes pour une section donnée, et Y ces représentations pour la section suivante.

9.2.2.3. Résultats

Pour mesurer la performance de ces réseaux, il faut trouver une métrique adéquate. Le but final est de se rapprocher autant que possible de la véritable section de génome du jeu de données. Pour l'entraînement du modèle, je mesure la distance du point latent prédit à sa vraie position attendue, mais je mesure aussi l'exactitude de reconstruction entre l'individu prédit décodé et l'individu de référence.

Cette méthode permet de prédire des génomes latents avec une exactitude de reconstruction d'au moins 90%, allant jusqu'à 98% pour certaines sections. Il faut tout de même retenir que cette valeur dépend aussi des capacités de reconstruction du décodeur, qui sont en général bonnes, mais peuvent rajouter de l'erreur sur la qualité de la prédiction.

9.2.2.3.1. Mesures de l'erreur du modèle

Initialement, ce réseau utilisait une distance euclidienne pour mesurer l'erreur réalisée entre la position de l'individu de référence dans l'espace latent ciblé et la prédiction effectuée par le modèle. Ses résultats étaient bons, mais comme pour les échantillons générés avec échantillonnage gaussien de l'espace latent, il y a un biais dans la diversité des échantillons produits. J'ai donc exploré quelques autres mesures pour tenter d'améliorer ce résultat.

L'objectif est donc de trouver une mesure qui permettrait en théorie de rapprocher la prédiction à la position exacte de l'individu de référence. Une possibilité est de calculer le ratio des distances euclidiennes entre l'individu prédit et l'individu de référence et entre l'individu prédit et son autre voisin le plus proche dans le jeu de données de référence. Elle a pour formule :

$$\mathrm{D1} = \frac{\mathrm{D}_{pred \to soi}}{\mathrm{D}_{pred \to autre}}$$

En théorie, cette métrique aide le prédicteur à se positionner plus exactement sur le point recherché lorsqu'il se trouve déjà dans son voisinage. Lors de l'entraînement, j'ai observé qu'il montrait des difficultés à trouver ce voisinage au départ et donc à la place d'améliorer la prédiction, elle s'est empirée. Le prédicteur cherchait à ramener le ratios de ces deux distances aussi proche que possible de 1 en éloignant la prédiction très loin de la distribution des vrais individus.

En modifiant le ratio de distances de la formule ci-dessus, l'impact de la distance entre l'individu prédit et l'individu de référence a été augmenté en l'élevant au carré. La formule devient donc :

$$\mathrm{D2} = \frac{\left(\mathrm{D}_{pred \to soi}\right)^2}{\mathrm{D}_{pred \to autre}}$$

Ces fonctions de mesure de dissimilarité présentent toutefois un risque : leurs valeurs dépendent du nombre d'autres individus dans le jeu de données. Ce lien n'est pas proportionnellement direct, mais cette dépendance interdit de comparer ces mesures entre elles. Pour l'usage qui en est fait ici, cette condition ne pose pas de problème.

Malheureusement, ces changements n'ont pas eu l'effet escompté sur les prédictions, comme on peut le voir en comparant les trois diagrammes suivants. En majorité, mesurer l'erreur de la moyenne des carrés est le meilleur moyen de prédire les caractéristiques d'un individu dans la section suivante. Cette métrique est celle qui donne la meilleure diversité d'échantillons lors de la prédiction.

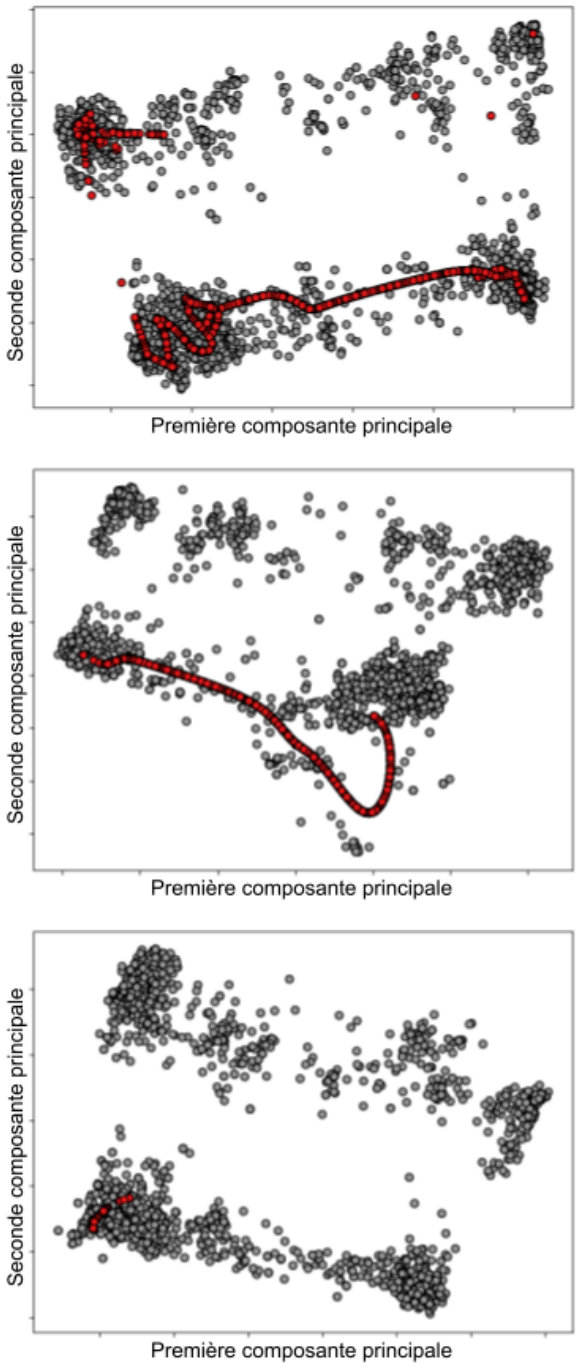


Figure 53 : Projection des génomes de référence et leurs prédictions effectués par l'étendeur d'espace latent utilisant une distance euclidienne (haut), le ratio de deux distances (milieu), et ratio modifié (bas).

Les projections de la figure 53 montrent que les mesures de dissimilarité que j'ai mis en place ne permettent pas de mieux guider les prédictions de ces modèles. Les individus prédits ne

présentent pas une meilleure diversité en utilisant ces mesures en tant que fonction de perte du modèle. Je conserve donc cette mesure des erreurs par la moyenne des écarts au carré pour ce modèle.

9.2.2.3.2. Biais

D'après les projections des individus prédits et ceux de référence dans une ACP, il est clair que le modèle n'arrive pas à prédire le spectre entier de la diversité de la population. Ce phénomène n'est pas surprenant, il serait assez exceptionnel qu'un tel modèle arrive à prédire exactement cette diversité.

Ce phénomène pose toutefois problème si on cherche à appliquer cette méthode de cohérence sur un grand nombre de sections, puisqu'on introduira un biais de plus en plus prononcé sur la population prédite, jusqu'à potentiellement n'avoir plus qu'un seul mode prédit au final.

Ce biais devient clairement apparent lorsqu'on enchaîne plusieurs étendeurs à la suite, c'est-à-dire qu'on essaie de prédire une section en se basant sur des données qui sont elles-mêmes des prédictions. À chaque itération, le spectre de prédiction diminue, jusqu'à ce qu'il devienne impossible de distinguer plusieurs points dans ces projections. Ce phénomène est illustré sur les graphiques suivants, qui comparent les échantillons de référence et ceux obtenus par prédiction par l'étendeur de la section précédente.

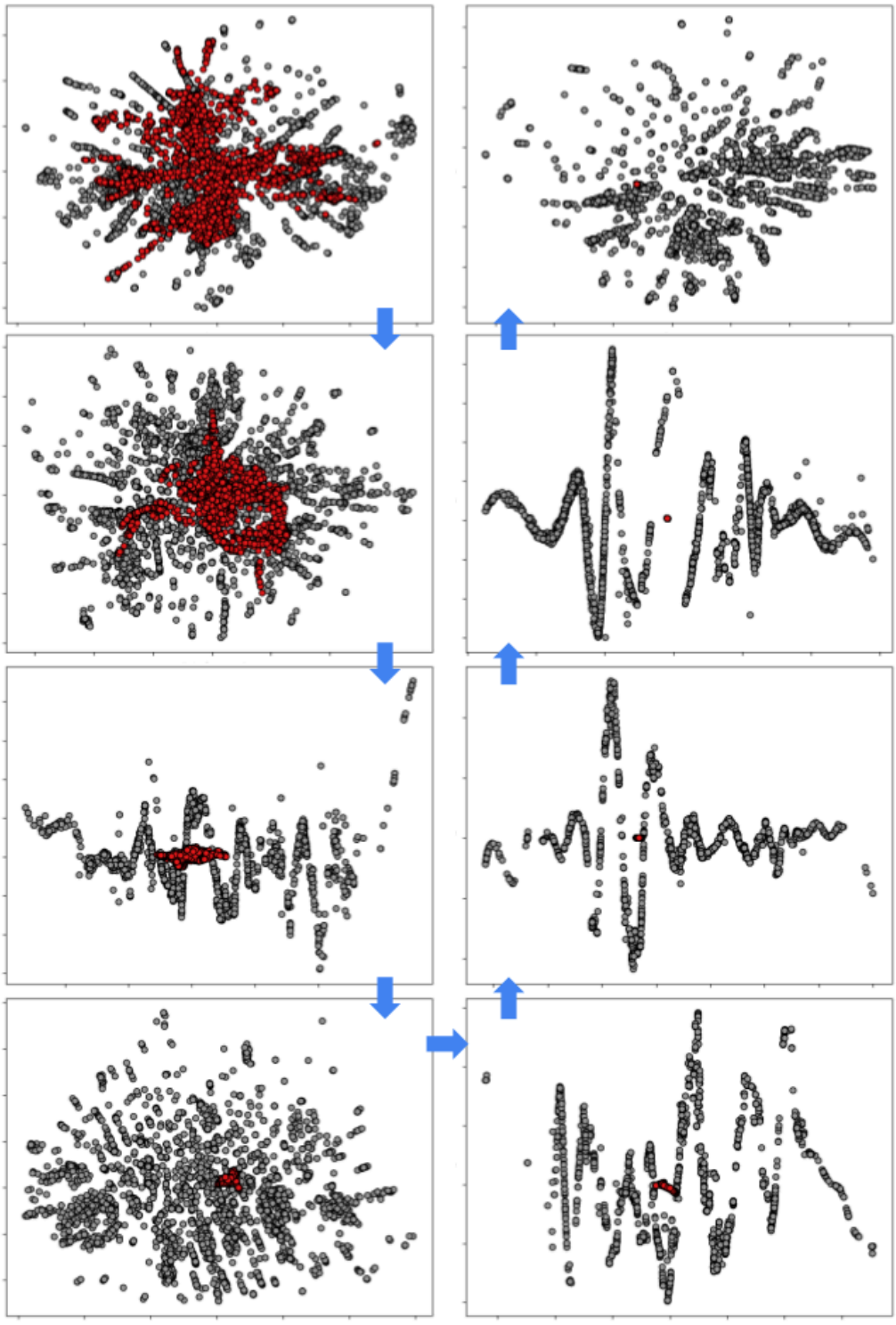


Figure 54 : Projections des individus de référence (gris) et leurs prédictions (rouge) effectués par l'enchaînement d'étendeurs d'espace latent, dans les espaces latents que les modèles tentent de prédire. Les flèches bleues indiquent le cheminement des prédictions.

La figure 54 présente les profils de mutations prédits par enchaînement d'étendeurs. À chaque nouvelle prédiction, ce processus cumule les biais de tous les modèles qui ont servi à cette prédiction. Il est clair que la distribution des échantillons prédits se restreint à chaque étape. Dans les dernières sections montrées ici, il devient même impossible de distinguer plusieurs points.

Cette méthode n'est donc pas viable à grande échelle. Il me faut explorer d'autres moyens d'établir de la cohérence entre sections générées.

9.2.3. Distance de Wasserstein dans la mesure de l'erreur

9.2.3.1. Application de ce modèle sur l'espace latent

Une autre manière de lier les sections encodées entre elles est de faire réaliser le travail de liaison par un réseau de neurones, qui est en théorie capable de traiter le grand nombre de données requis pour trouver ces liaisons.

Le modèle de GAN, décrit dans la section 7.1.3.3. "Réseaux génératifs" , me semble être le plus approprié pour effectuer cette tâche. Cette architecture de modèle a fait ses preuves dans la tâche de simulation de données génomiques dans (Yelmen et al. 2021), là où la RBM protégeait mal les données sensibles des donneurs. La diffusion stable a montré des capacités incroyables dans la génération ciblée d'images, mais prendre en compte une étiquette pour la génération de nouvelles données est un pas supplémentaire et assez complexe. Il serait intéressant d'explorer l'application de cette méthode dans un autre projet.

Les exemples les plus récents de GAN ayant bien fonctionné contiennent des WGAN, c'est-à-dire emploient la perte de Wasserstein (Arjovsky, Chintala, and Bottou 2017). La formule exacte de cette mesure de distance, pour toute paire de mesures de probabilité μ et ν avec p moments, est :

$$W_p(\mu,\nu) = \left(\inf_{\gamma \in \Gamma(\mu,\nu)} E_{(x,y)\sim\gamma} d(x,y)^p\right)^{1/p}$$

Où Γ(μ, ν) est l'ensemble de tous les couplages de μ et ν.

Cette mesure permet d'éviter autant que possible un phénomène appelé *mode collapse*, ou

effondrement de mode. Lorsque ceci se produit, le générateur cesse de tenter de reproduire la distribution entière du jeu de données d'entraînement, et se contente à la place de cibler un mode dans ce jeu de données qu'il a identifié comme donnant un bon score de la part du discriminant. Une fois que le discriminant repère que les échantillons de ce mode ont de fortes chances de provenir de la population simulée, alors il punit ce comportement en leur assignant des scores bas. Par simplicité, le générateur ne cherche pas à modifier l'étendue de sa distribution, mais le recentre sur un autre mode qu'il a repéré dans les données d'entraînement. Il est presque impossible de revenir à une distribution proche de celle du jeu de données de référence à partir de ce moment.

Le WGAN se montre plus efficace en étant plus permissif envers les solutions approximatives du générateur. Cette tolérance vient du gradient du nouveau discriminant, qui est proche de 1 sur la majorité de l'espace, alors que le GAN classique a tendance à adopter un gradient fort entre les distributions et nul sur les autres régions. Ce changement de comportement, illustré dans la figure 55, crée plus de fluidité et de permissivité dans les prédictions du modèle.

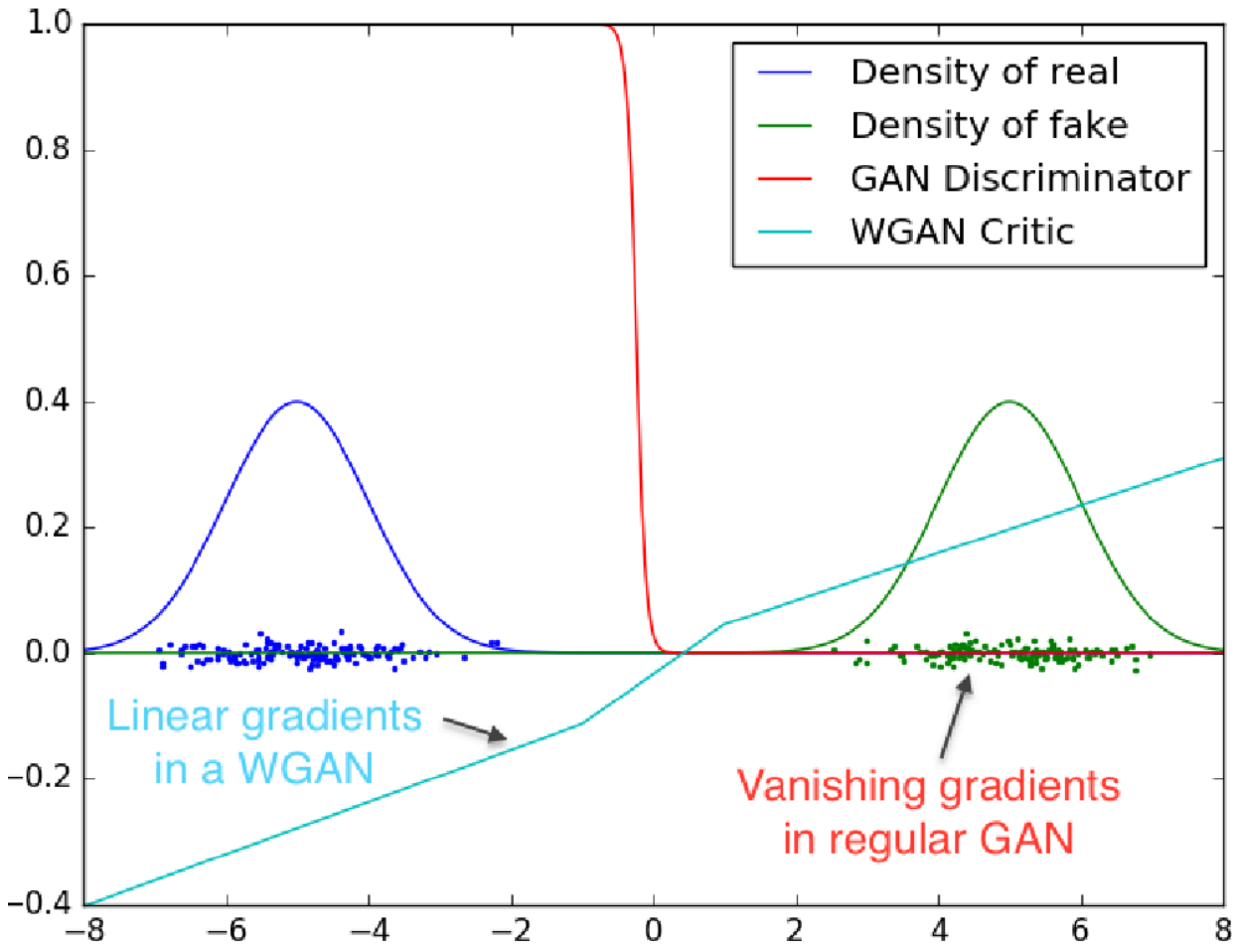


Figure 55 : Visualisation des séparations réalisées par les discriminants de deux modèles. Les points verts et bleus représentent les données simulées et réelles respectivement. Les gradients d'un WGAN (bleu clair) sont comparés à ceux d'un GAN classique (rouge), montrant la différence de l'approche. Tiré de (Arjovsky, Chintala, and Bottou 2017)

Cependant, là où le discriminant d'un GAN classique émet un réel entre 0 et 1 qui peut être interprété comme une probabilité de réalisme, ici, cette partie du modèle émet un score de vraisemblance, qui est un réel positif. Il est donc difficile d'interpréter ce score de manière indépendante. C'est pourquoi on appelle maintenant le réseau qui tente de différencier les échantillons un **critique**.

9.2.3.2. Architecture

Le modèle est composé de deux réseaux, le **critique** et le **générateur**. Le critique est composé de trois couches cachées, ayant pour taille N/10, N/50 et N/100 respectivement, N étant

le nombre de dimensions dans les échantillons du jeu de données d'entrée. Les poids du critique subissent un écrêtement (*clipping* en anglais) pour les limiter. Le générateur est aussi composé de trois couches cachées, ayant pour taille N/10, N/5 et N/2. Pour ces deux modèles, l'abandon est implémenté entre chaque couche cachée.

9.2.3.3. Entraînement

Les modalités d'entraînement d'un WGAN modifient le processus normal pour un GAN. La perte de Wasserstein met en place un critique qui peut être entraîné jusqu'à optimisation, selon ses créateurs. Le critique remplace le discriminant, et son nom diffère car son rôle diffère. Le discriminant émet un nombre compris entre 0 et 1, qui peut être interprété comme la probabilité que l'échantillon analysé soit réel, alors que le critique émet un réel qui peut prendre n'importe quelle valeur. La distance de Wasserstein compare deux distributions directement, donc le critique peut réellement appréhender cette distribution, et non pas se contenter de l'inférer à partir des échantillons vus, comme le fait le discriminant d'un GAN. Dans ce but, à chaque boucle d'entraînement, le critique est entraîné cinq fois alors que le générateur ne l'est qu'une fois. Il est possible de tester d'autres ratios, mais celui-ci constitue une base solide. L'optimiseur utilisé est la propagation par moyenne quadratique.

Le modèle implémente aussi une contrainte sur les poids assignés à ses connexions entre neurones, mais il a été testé avec et sans. Cette contrainte limite l'amplitude que peut prendre les poids du modèle dans les couches cachées du modèle. Cette contrainte rejoint le taux d'apprentissage comme méthode de forcer le modèle à s'adapter de manière stable aux corrections que sa perte lui instruit.

9.2.3.4. Résultats

9.2.3.4.1. Deux sections

Tout d'abord, j'ai voulu implémenter ce modèle sur deux sections voisines pour jauger les capacités du GAN à la tâche qui lui est destinée : reprendre les projections des sections génomiques dans les espaces latents créés par les encodeurs variationnels et simuler de nouveaux échantillons.

Les résultats montrés ici emploient la méthode d'entraînement finale décrite dans la section 9.2.3.4.2. Les principaux résultats de ce manuscrit ont été obtenus en analysant le jeu de données de la section 9.2.3.4.2., mais je trouvais tout de même intéressant de se pencher sur un exercice simple pour commencer à évaluer les performances du WGAN.

Pour mesurer la performance de ce WGAN, la première confirmation de sa performance peut se faire visuellement : est-ce que le modèle parvient à reproduire la distribution des échantillons en entrée?

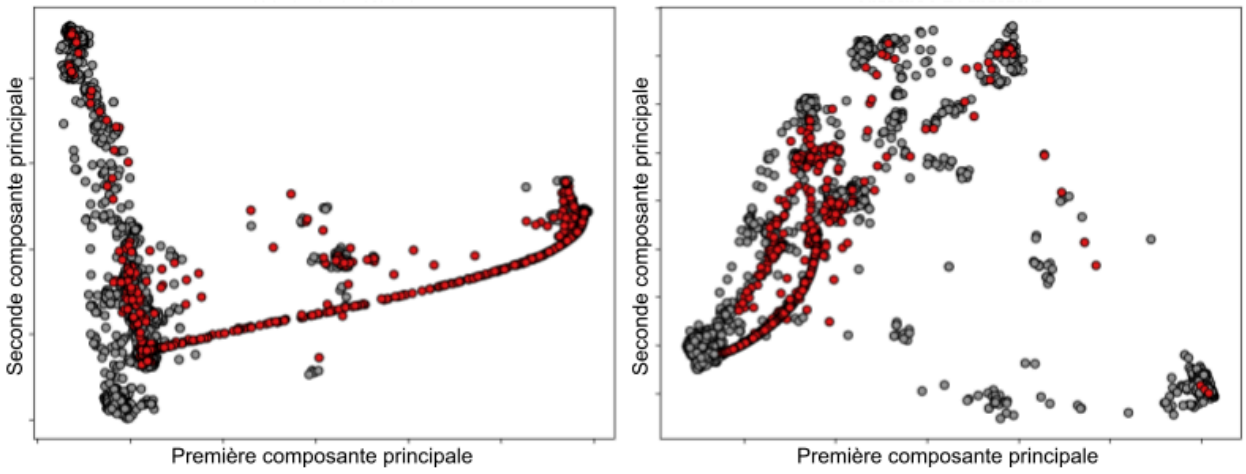


Figure 56 : Représentation des génomes de référence et ceux simulés par les premiers tests de WGAN sur des sections génomiques, projetée sur une ACP.

D'après les projections en ACP de la figure 56, nous pouvons observer que le modèle arrive plutôt bien à reproduire la distribution des échantillons qu'il a vu en entrée. La couverture n'est pas parfaite, mais *a minima*, le réseau semble adapté à la tâche qui lui est présentée.

9.2.3.4.2. Sections couvrant 1Mb d'ADN

Le prochain ensemble de données sur lequel ce modèle a été entraîné s'étend sur un million de paires de bases d'ADN, correspondant à environ 15 000 mutations. Les sections qui la composent sont contiguës.

En passant à une échelle plus grande, de nouveaux risques sont introduits. Avec un nombre de dimensions qui augmente mais pas le nombre d'échantillons, il y a de plus en plus de chances d'assister à un effondrement du générateur, comme évoqué dans la section 9.2.3.1.

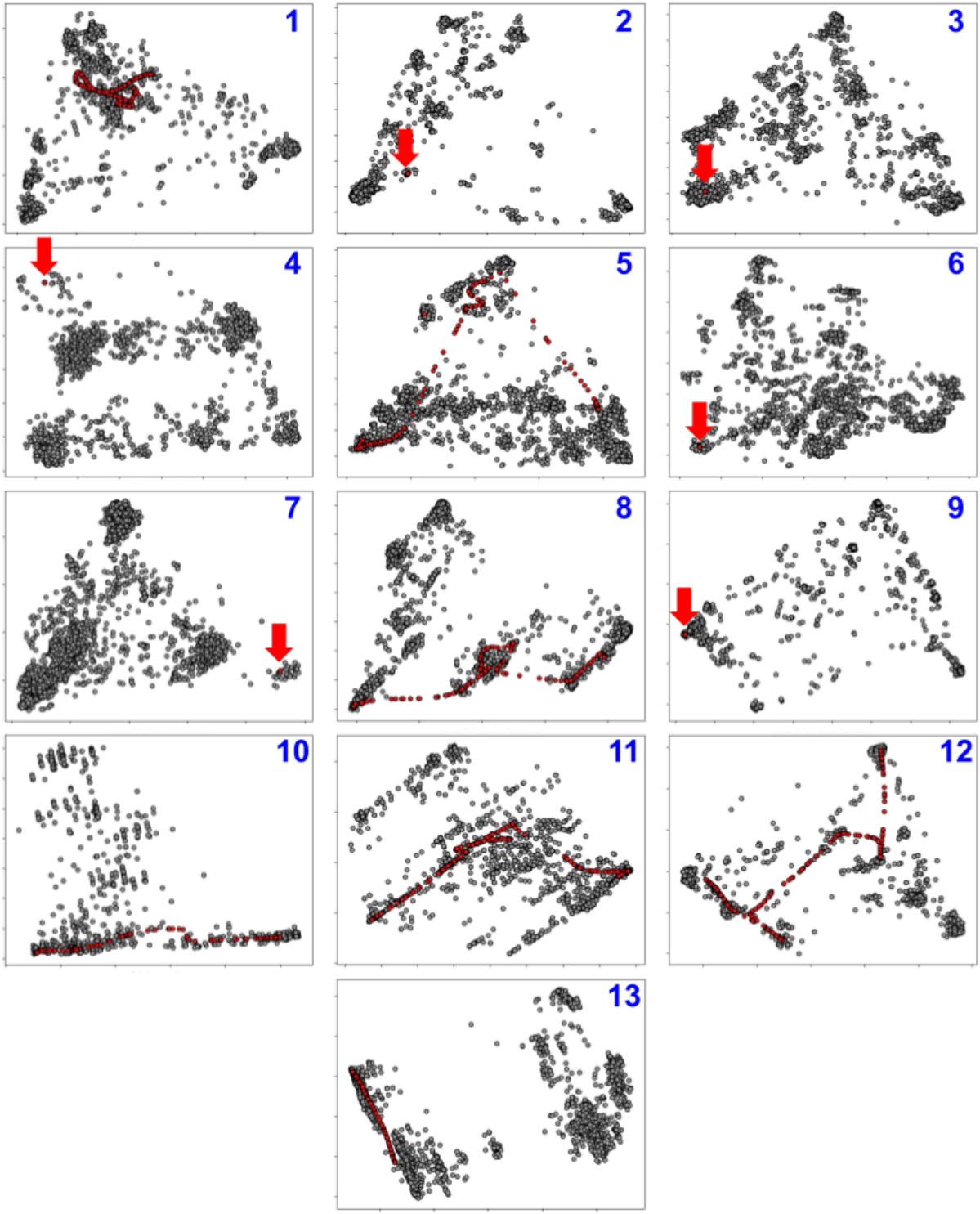


Figure 57 : Représentation des génomes de référence et ceux simulés par WGAN sur toutes les sections génomiques composant le Mb d'ADN, projetée sur une ACP.

Ce phénomène se produit pour plusieurs sections traitées, mais pas toutes. Dans les

projections 2, 3, 4, 6, 7 et 9 de la figure 57, l'ensemble des génomes simulés se confondent en un seul point (indiquées par des flèches rouges). Les autres montrent de la diversité, mais contiennent tout de même de forts biais : les données simulées ne s'étendent pas sur l'intégralité de la distribution, mais se confinent à un "chemin" étroit qui passe par les régions contenant de fortes concentrations d'échantillons. Les distributions de ces génomes encodés simulés semblent suivre les biais observés dans le parcours des espaces latents des VAE détaillés dans la section 9.1.3.3. Nous pouvons donc en conclure que la limite ici semble se porter sur l'espace latent créé par les encodeurs. Cependant, il est légitime de se demander si ces espaces latents biaisés sont plus "simples" à parcourir pour le GAN, vu que leur variance est plus limitée. Ceci pourrait expliquer pourquoi les autres sections ont montré un effondrement total, de par leur complexité plus élevée.

Après plusieurs tentatives en explorant des variations de différents hyperparamètres, aucun modèle ne semble donner de résultats franchement meilleurs. Je me suis donc repenché sur la conception du modèle et j'ai investigué en profondeur les éléments qui le composent en croisant plusieurs sources d'informations. À partir de ces recherches, j'ai voulu tester d'enlever la limite d'amplitude sur les poids du modèle, qui peut être source d'effondrement lors de la prédiction d'échantillons complexes (Gulrajani et al. 2017).

Cette pratique n'est pas recommandée, il est préférable d'avoir une forme de limite sur les poids du modèle. Il existe d'autres méthodes pour agir sur l'intensité des poids dans le modèle que l'écrêtement (Gulrajani et al. 2017), qui mériteraient d'être testées également.

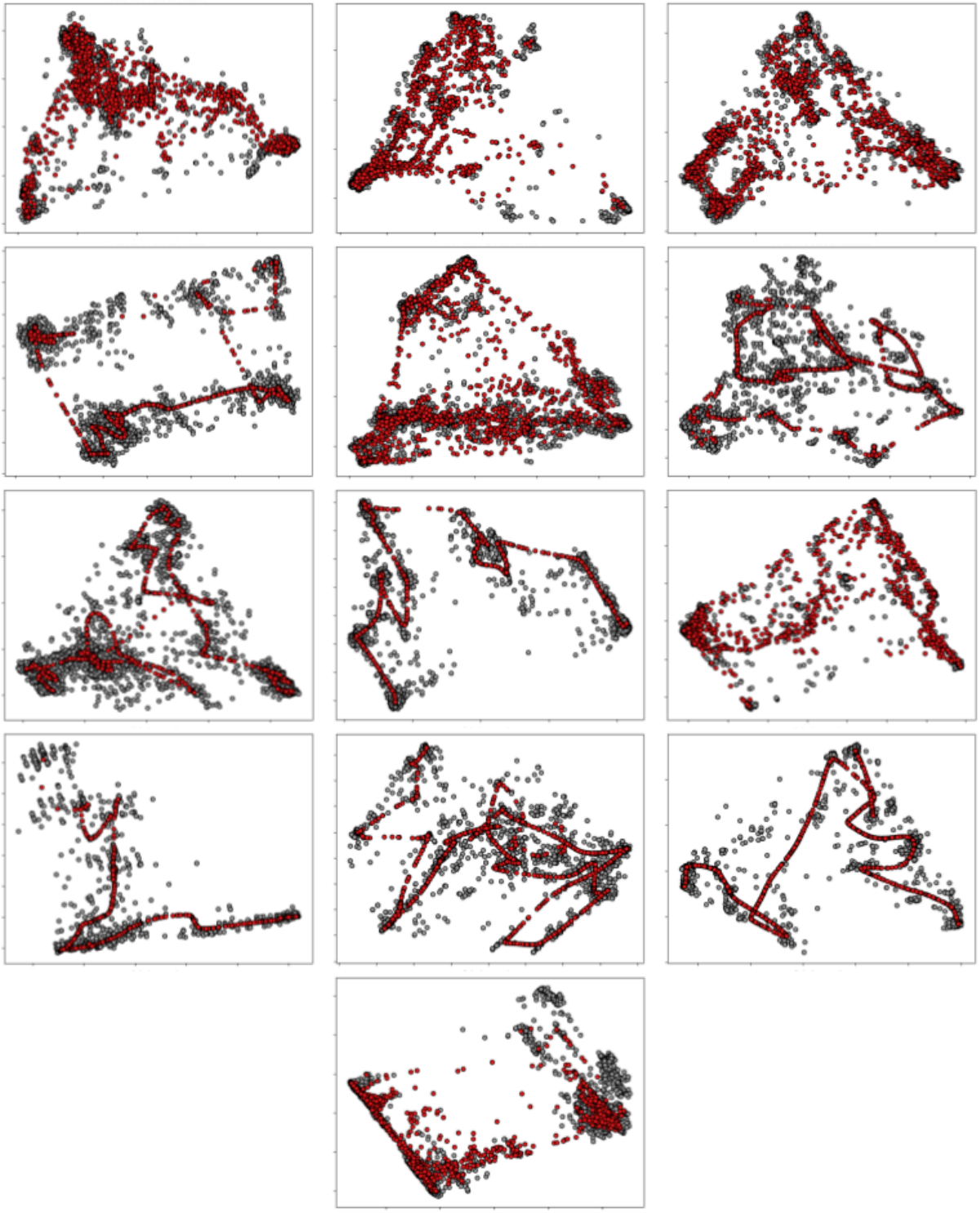


Figure 58 : Représentation des génomes de référence et ceux simulés par WGAN sans limite d'amplitude de poids sur des sections génomiques, projetée sur une ACP.

D'après la figure 58, il est clair que le retrait de ce paramètre aide énormément le modèle.

Il parvient nettement mieux à reproduire les distributions des génomes de références dans les sections où, précédemment, le modèle ne parvenait pas du tout à capturer cette diversité. Les limites de diversité sur certaines sections persistent, très certainement à cause des biais des VAE, mais pas pour toutes les sections où ce phénomène est observé en parcourant l'espace latent du VAE. Nous pouvons clairement conclure que la limite vient de l'espace latent créé par l'encodeur. Il faudrait donc se repencher sur cet élément là pour tenter de recréer une vraie diversité.

9.2.3.4.3. Sections non consécutives

Cette implémentation est effectuée sur une section contigüe de génome, ce qui peut aider le modèle à trouver des liens entre régions. Je me suis demandé si ces performances se maintiennent sur une collection de sections génomiques non-consécutives.

Les sections choisies sont toutes éloignées d'environ 1Mb, pour augmenter les divergences entre mutations dans les différentes sections. Au total, ces données représentent un peu plus de 6 500 mutations. Les résultats de l'application des mêmes méthodes sont illustrées dans la figure 59.

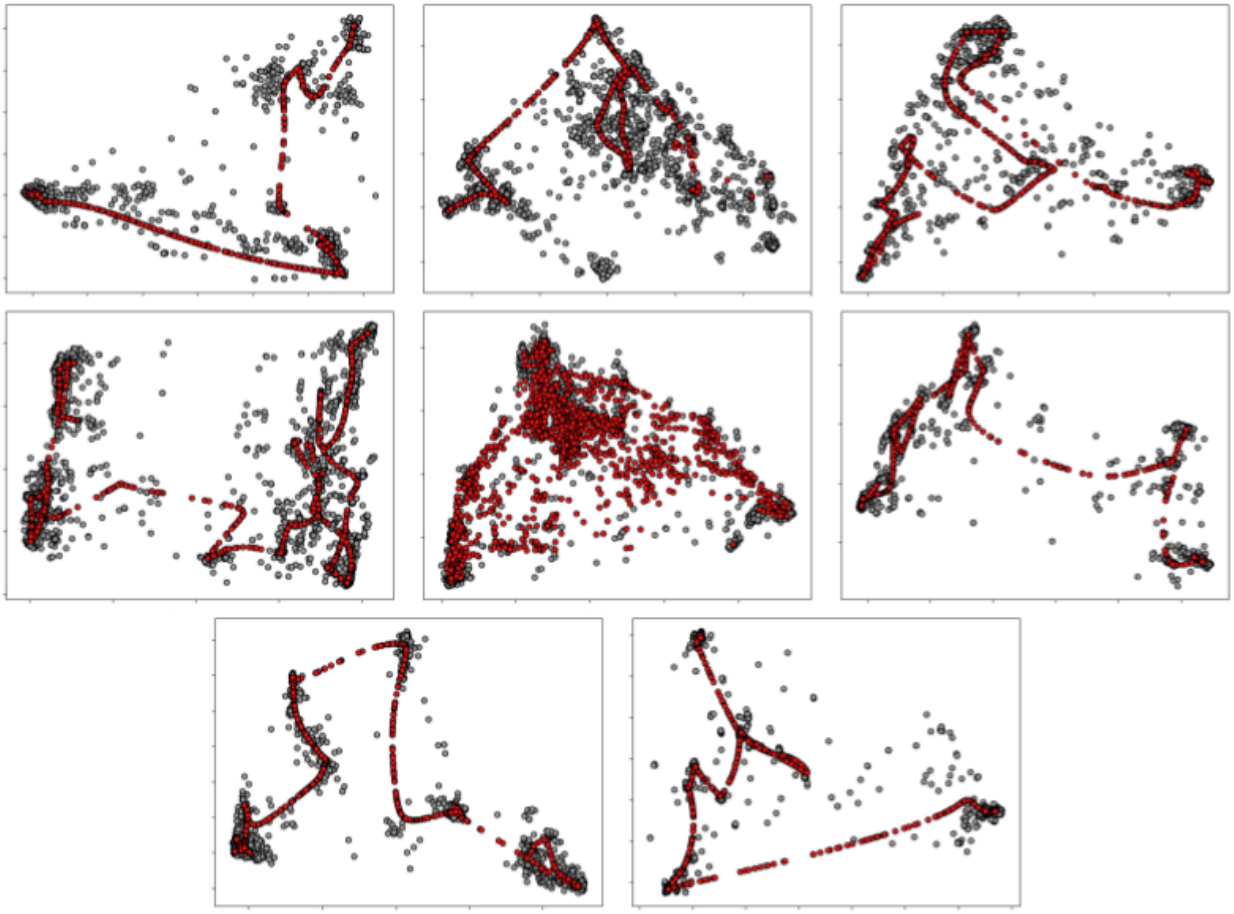


Figure 59 : Représentation des génomes de référence et ceux simulés par WGAN sans limite d'amplitude de poids sur des sections génomiques non consécutives, projetée sur une ACP.

Les performances du modèle se maintiennent plutôt bien, mais pas non plus parfaitement. Nous observons un biais assez fort sur la plupart des sections, ce qui peut traduire de nouveau un biais dans les VAEs utilisés pour la réduction de dimension.

9.3. Évaluation des profils simulés

Dans cette section, une attention particulière est portée sur l'analyse des données génomiques simulées par le WGAN. Plusieurs critères sont à évaluer pour valider l'outil. Pour reprendre les trois essentiels que j'avais mis en lumière pour (Yelmen et al. 2021), il faut que la population simulée soit diverse, réaliste et respecte l'anonymat.

Pour ce qui est de la diversité, la représentation des sections générées par ACP montre que les données génomiques simulées couvrent une distribution semblable à celle des données de référence, avec quelques biais. Je vais maintenant étudier les deux autres caractéristiques : le réalisme des échantillons générés, et leur respect de l'anonymat vis-à-vis des données originales.

9.3.1. Chromopainter

9.3.1.1. Présentation de l'outil

Chromopainter est un outil permettant d'explorer la distribution de SNP au sein de populations ayant connu des périodes d'isolation et de mixité (Lawson et al. 2012). Pour ce faire, il prend en entrée un jeu de données d'individus dont on connaît la population d'origine, appelée **population donneuse**, et une population dont on cherche à connaître les origines potentielles, appelée **population requête**. Pour chaque individu de ces deux populations, il nous faut renseigner la présence ou absence de chaque SNP détecté, ainsi que leurs positions et le déséquilibre de liaison entre eux.

À partir de ces deux populations, l'outil attribue à des sections génomiques de plusieurs SNP d'un génome requête une origine ancestrale dans la population donneuse. Le déséquilibre de liaison entre les différentes positions sert à décider si une mutation (ou plusieurs) a eu lieu dans une section ancestrale ou s'il y a eu recombinaison avec une section d'un autre individu. Ce processus est réalisé par défaut 10 fois pour chaque génome requête.

La sortie finale est sous format texte et présente pour chaque individu soumis dans la population requête les reconstitutions effectuées par l'outil. À chaque position, il indique simplement l'identifiant de l'individu donneur correspondant à la source de cette section.

9.3.1.2. Application sur population simulée

Notamment avec le WGAN, les IA génératives se basent sur les données de référence pour recréer de nouveaux échantillons. Chromopainter présente donc dans le cas présent une opportunité pour lire un peu dans la boîte noire qu'est un réseau de neurones et apercevoir quels mélanges le modèle a effectué pour créer les nouveaux échantillons.

9.3.1.3. Ce qu'on cherche à voir

En introduisant dans l'outil des génomes humains non utilisés dans le processus d'apprentissage, nous pouvons observer l'hétérogénéité "naturelle" chez l'humain. En comparant le nombre de sources qui composent ces individus avec ce même critère pour des génomes simulés, nous pouvons vérifier que le modèle générateur ne crée pas de mélange aberrant lors de la production de nouveaux échantillons.

9.3.1.4. Résultats

Le générateur décrit dans le chapitre précédent a été utilisé pour générer 1 000 profils de mutations couvrant le mégabase d'ADN étudié. Chromopainter a été appliqué sur ces 1 000 profils simulés et 78 profils de référence (trois par sous-population listée dans le projet des 1 000 Génomes). Pour chaque section de génome, dix reconstructions d'héritage sont générées.

Les reconstructions, visibles dans les figures 60 et 61, se présentent sous cette forme :

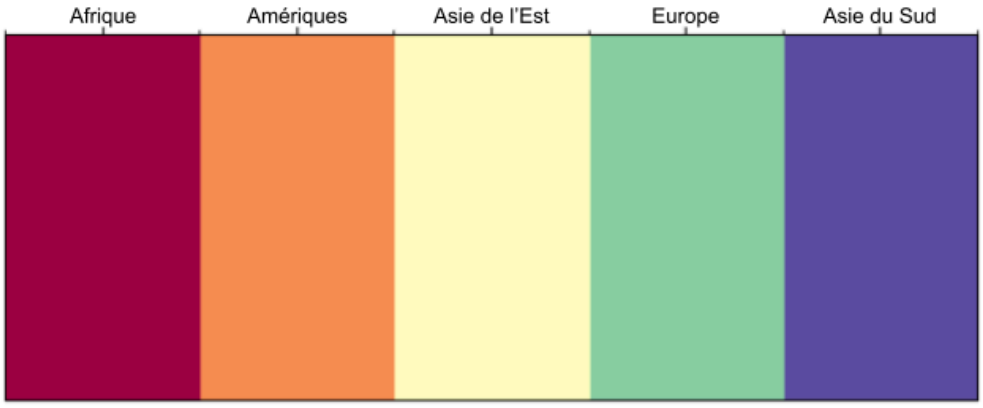


Figure 60 : Échelle de couleur des origines ethniques.

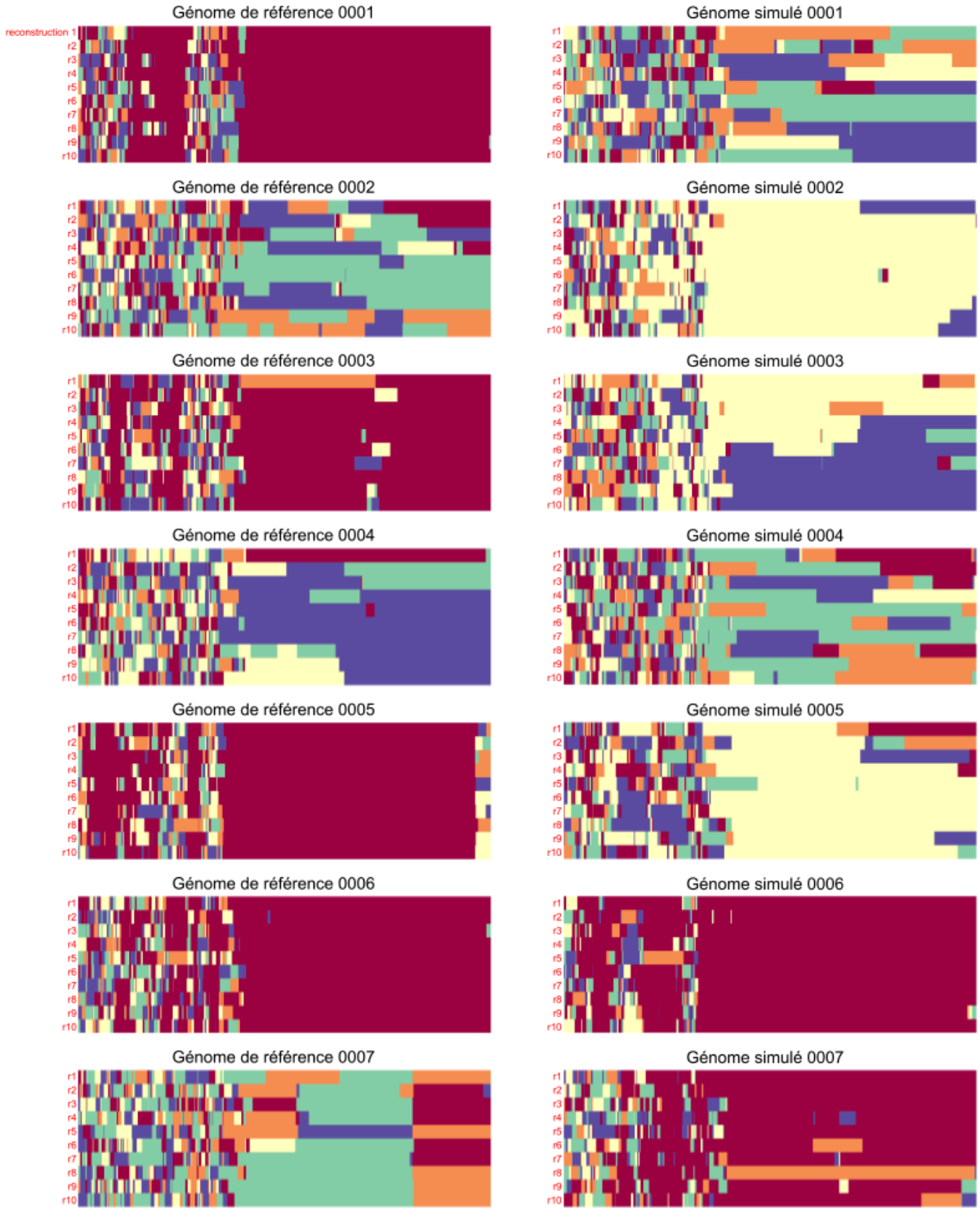


Figure 61 : Héritage génomique déterminé par Chromopainter de profils de mutation de référence ou simulés par WGAN. Chaque ligne représente une reconstruction possible, et chaque position sur l'axe horizontal correspond à une mutation.

On représente ici les dix reconstructions, une par ligne. L'axe horizontal n'est pas clairement représenté, mais est composé de chaque mutation de la section génomique représentée. La couleur à chacune de ces positions symbolise la population de laquelle est tirée cette section. Des grandes plages d'une couleur identique représente donc une série de plusieurs variants tous hérités de la même population (mais peut venir de génomes donneurs différents dans cette population), alors qu'une zone multicolore est d'héritage composite.

Il est difficile de conclure grand-chose à cette échelle, ou du moins la tâche prendrait du temps pour passer en revue tous les génomes simulés analysés de cette manière. J'ai donc choisi de créer des métaprofils à partir de ces reconstructions, pour en avoir une meilleure vue d'ensemble.

Les métaprofils se présentent sous cette forme :

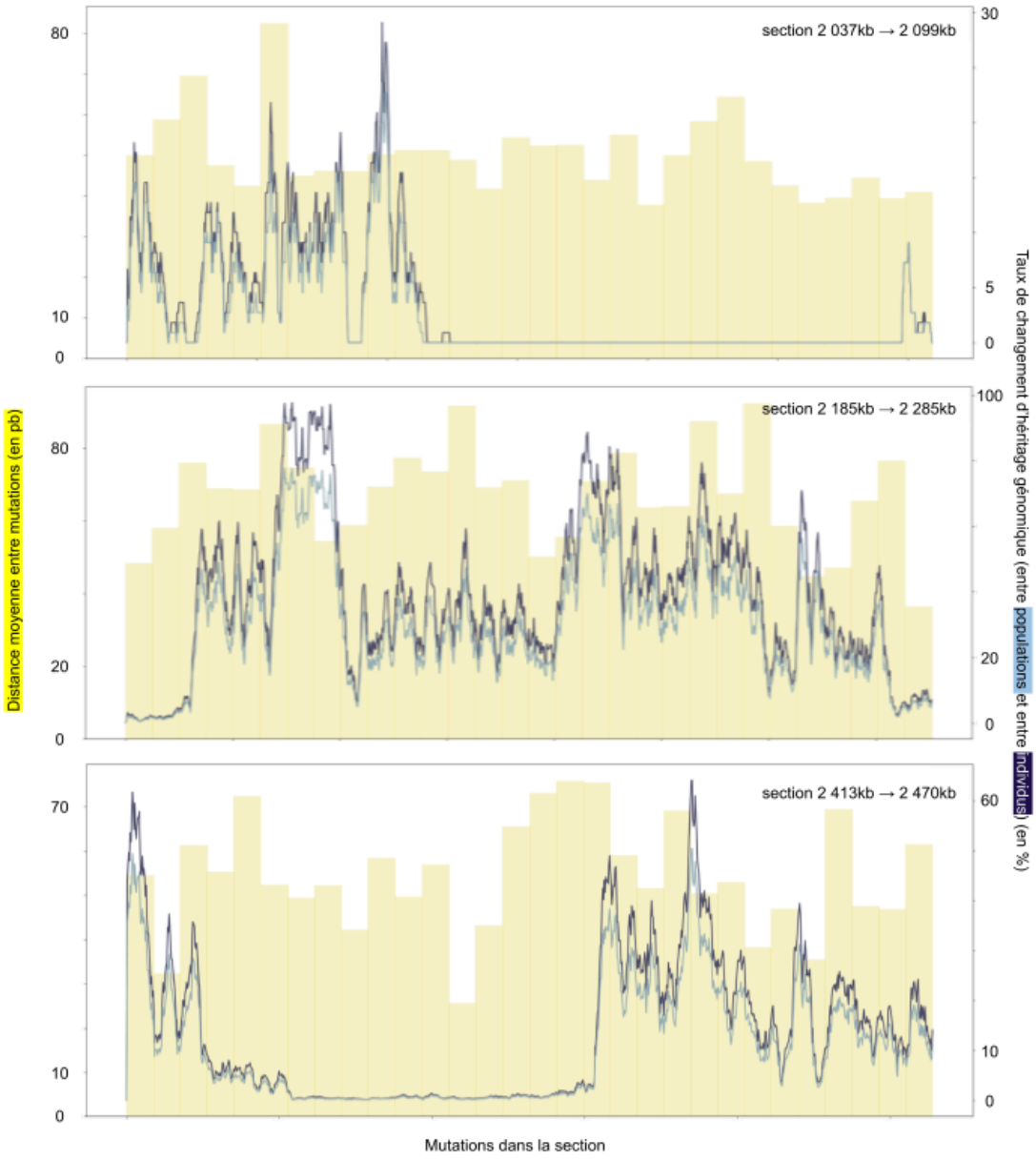


Figure 62 : Métaprofils des changements d'héritage de génomes de référence déterminés par chromopainter (lignes) et distances chromosomiques entre mutations (barres).

Les deux courbes dans la figure 62 représentent le nombre moyen de changements d'héritage le long de la section génomique, avec en gris clair les changements à l'échelle de la population et en bleu foncé les changements à l'échelle de l'individu. Les barres jaunes traduisent la distance en paires de bases entre les mutations.

Cette visualisation montre qu'il n'y a pas de lien entre recombinaisons identifiées par

chromopainter (courbes) et la distance génomique entre SNP (barres jaunes). L'outil traduit bien le déséquilibre de liaison entre ces mutations.

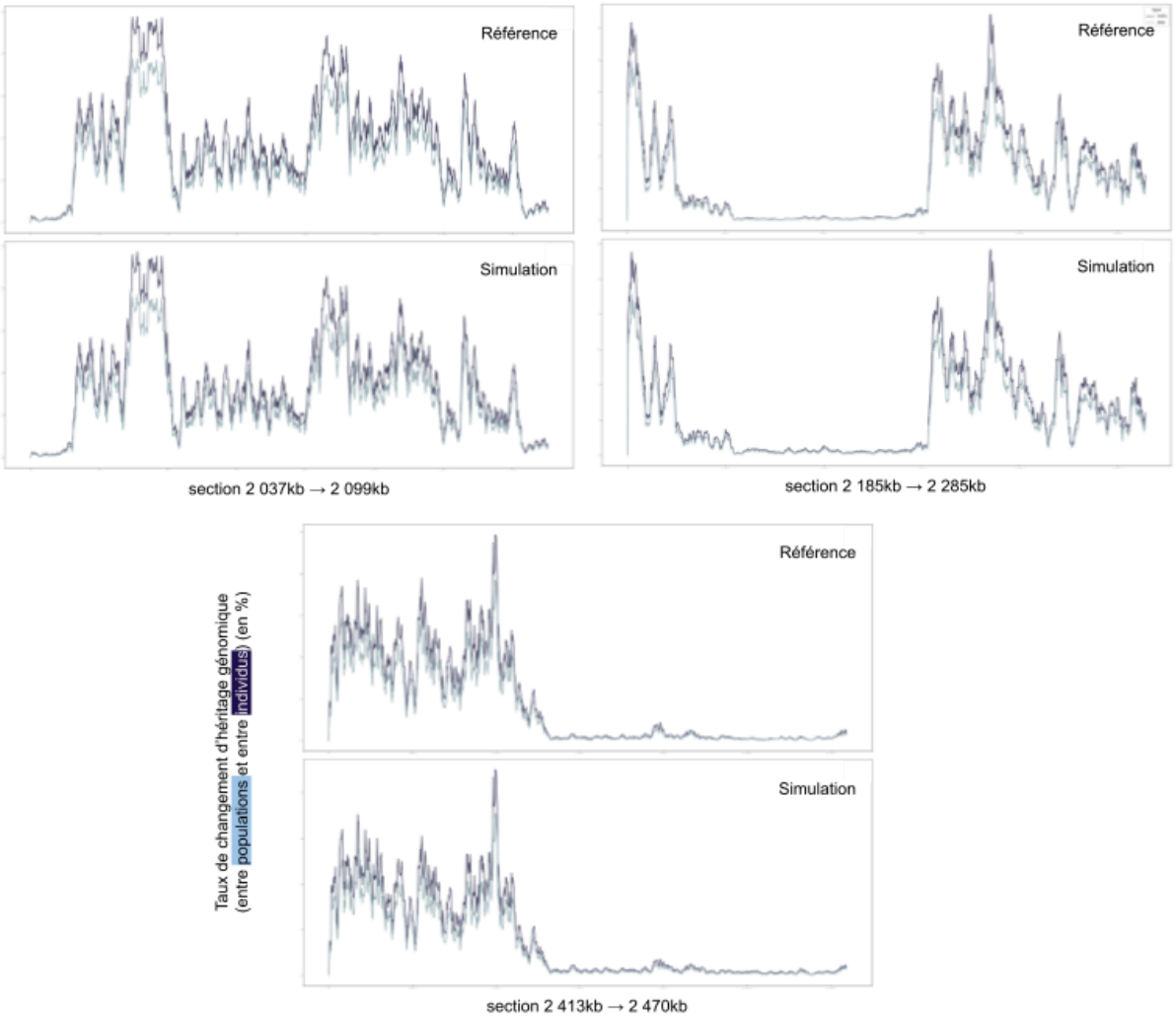


Figure 63 : Métaprofil des changements d'héritage de génomes de référence (haut) et simulés (bas), déterminés par chromopainter.

Les métaprofils présentés dans la figure 63 se ressemblent énormément entre population simulée et de référence. Nous pouvons donc conclure que ces génomes sont réalistes et pourraient survenir par recombinaison d'haplotypes naturellement.

9.3.2. Mesurer le respect de l'anonymat

Un modèle capable de produire des données génomiques humaines est très intéressant, mais il faut vérifier que ces données sont bien des simulations. Un modèle qui se contente de

reproduire les échantillons vus dans le jeu de données d'entraînement n'est pas utile pour créer des données anonymisées.

La distance de Hamming mesure le nombre de remplacements qu'il faut effectuer pour passer d'une séquence de symboles à une autre de même taille (Waggener 1995). Cette métrique est donc très pertinente pour déterminer à quel point les sections génomiques simulées sont proches de celles contenues dans le jeu de données de référence.

Le principe est donc de calculer cette distance de Hamming pour chaque paire individu simulé-individu de référence, puis je sélectionne le plus proche voisin de l'individu simulé.

Une distance nulle signifie que l'individu simulé est une copie exacte d'un individu de référence, tandis qu'une distance trop forte signifie que l'individu simulé ne ressemble à aucun des individus de référence. Ainsi, un individu simulé acceptable est un individu dont la distance reste faible mais non nulle.

Pour faciliter cette interprétation, ce n'est pas la distance en soi qui sera utilisée, mais le rapport entre cette distance et la distance maximale possible.

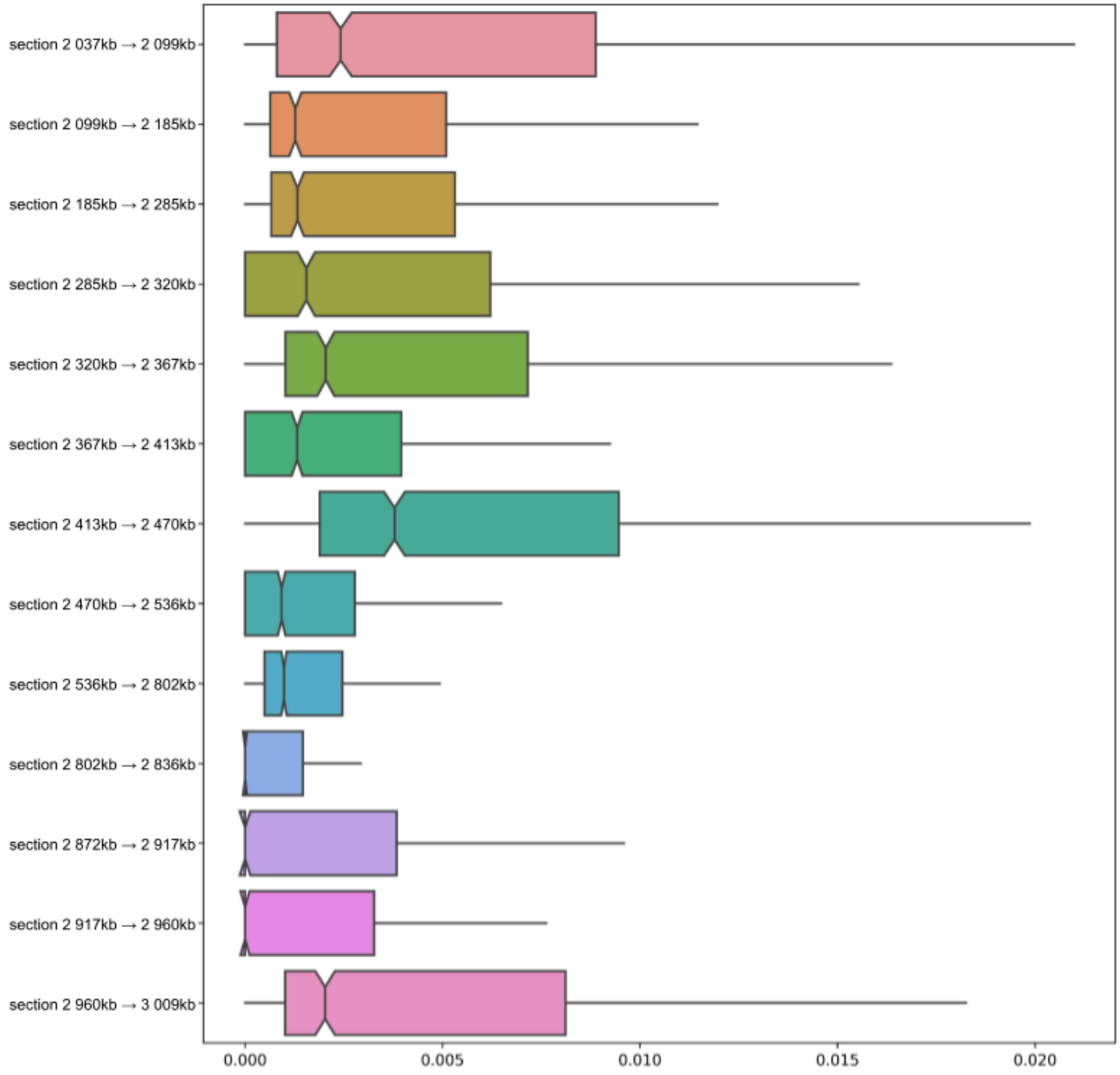


Figure 64 : Boîte à moustaches des distances de Hamming entre chaque individu simulé et son voisin le plus proche dans le jeu de données de référence par section génomique. Les valeurs aberrantes hautes ont été retirées pour une meilleure interprétation visuelle.

Les distributions de la figure 64 montrent qu'il y a bien des sections génomiques pour lesquelles les profils simulés se superposent à des profils de référence. Ceci n'est bien sûr pas souhaitable. Pour certaines sections, ce comportement semble même être la norme, si bien que la médiane de leurs distributions de distances est à 0.0. Cependant, il n'est pas possible de savoir si les profils simulés superposés sur leurs voisins dans une section donnée le sont aussi pour

d’autres sections avec cette figure. J’ai donc appliqué le même processus sur les données totales.

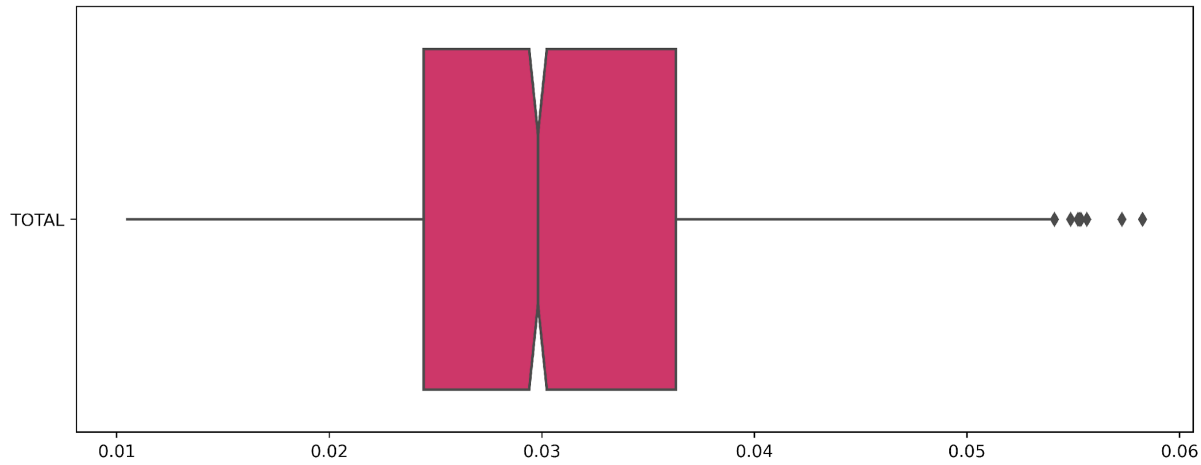


Figure 65 : Boîte à moustaches des distances de Hamming entre chaque individu simulé et son voisin le plus proche dans le jeu de données de référence au total.

D’après la figure 65, la valeur minimale de distance d’un génome simulé à son plus proche voisin dans le jeu de référence est 0.0105, ce qui équivaut à 152 mutations différentes entre les deux sections de génome concernés. Il n’y a donc aucun profil simulé dans ce jeu de données synthétique qui soit une copie conforme d’un profil dans le jeu de données de référence.

0.01 représente une distance d’édition assez faible tout de même. Au-delà de se demander s’il est souhaitable d’avoir une distance d’édition très élevée ou non (c.f. section 10.5.2), je vois une opportunité où il serait possible de l’augmenter. J’ai vu dans le parcours de l’espace latent des autoencodeurs variationnels une recherche de mode, ce qui entraîne un comportement proche du surapprentissage dans le décodeur. Celui-ci aura tendance à recréer des échantillons proches de ceux contenus dans le jeu de données d’entraînement.

En examinant empiriquement, on peut observer un potentiel lien entre la distribution de ces scores basés sur la distance de Hamming et la distribution des génomes simulés créés par le WGAN. Dans les sections pour lesquelles le score minimal vaut 0, et en particulier pour celles où la médiane est très proche de 0, on voit un manque d’exploration de l’espace latent de la part du WGAN. Une hypothèse possible est que cette restriction est plus due à la configuration de l’espace latent par le VAE qu’à un comportement du WGAN. Si cette hypothèse est vraie, alors une amélioration dans la création de l’espace latent par l’autoencodeur produirait aussi une augmentation dans la distance de Hamming observée pour les génomes dans cette section.

10. Discussion

10.1. Mes contributions

À travers ce projet, j'ai pu me former sur le développement et l'utilisation de modèles d'intelligence artificielle sur des données biologiques complexes. En parallèle, les résultats produits contribuent à plusieurs domaines d'analyse très pertinents aujourd'hui.

L'application d'autoencodeurs sur des données génomiques n'est pas une pratique rare, mais ils sont en général utilisés pour leur capacité à traiter des données bruitées ou éparses comme du séquençage en cellule unique (Rashid et al. 2021; J. Xu et al. 2023). Mon emploi en tant qu'étape de réduction de dimensions est novatrice due à la réutilisation des données réduites issues de mes modèles, et non pas uniquement à but de visualisation comme dans des projets déjà existants (Wang and Gu 2018; Karwath et al. 2021).

Les limites du GAN employé par (Yelmen et al. 2021) conduit logiquement au WGAN, comme l'a fait l'équipe à l'origine de cette étude (Yelmen et al. 2023). Mon application ici les utilise en conjonction avec les autoencodeurs variationnels pour augmenter le spectre des données qu'il lui est possible d'analyser, tout en limitant sa consommation de ressources de calcul.

L'utilisation de métaprofils Chromopainter pour étudier la vraisemblance de ma population simulée est nouvelle. Reprendre cet outil afin de vérifier qu'elle possède des caractéristiques comparables à une population de génomes de référence représente une métrique innovante pour l'étude de génomes simulés.

10.2. Applications

10.2.1. Découvertes génomiques

La première étape de ce projet est l'autoencodage de données génomiques de personnes saines. En explorant en profondeur l'espace latent de ces autoencodeur, il est possible de découvrir quelles règles l'encodeur a mis en place pour comprimer l'information génomique, et donc quelles dépendances entre mutations le modèle a identifié. Muni de cette méthode, on peut appliquer ce processus uniquement dans le but de découvrir ces règles, chez des donneurs sains ou malades. Ainsi, la phase d'autoencodage peut servir à mettre en lumière des règles d'association de mutations et donner des nouvelles pistes pour le diagnostic de facteurs de risque de maladie.

Ces autoencodeurs peuvent aussi créer de nouveaux critères à partir des données qui leur sont présentées. On appelle souvent ces critères *metafeature* en anglais, signifiant une caractéristique synthétisée à partir de plusieurs autres caractéristiques élémentaires. Ces nouvelles caractéristiques peuvent se montrer utiles dans la classification de patients par exemple, en offrant de nouvelles possibilités pour les cliniciens d'interpréter des données génomiques.

10.2.2. Génération de nouveaux génomes

Le manque de données génomiques humaines publiques contraint sévèrement les méthodes statistiques applicables sur les études qui se portent sur ces données. Cette nouvelle méthode d'augmentation de données permet de gonfler le nombre d'échantillons à étudier, et donc prêter plus de robustesse à ces travaux.

Cette pratique peut cependant être risquée, vu que notre l'augmentation de données ne fera que retranscrire des biais déjà présents dans le jeu de données de base. Il faudra porter un grand soin à la curation des données de départ ainsi qu'à l'analyse des données simulées.

10.2.3. Brouillage de cohortes de patients

Lorsqu'un patient voit son génome séquencé par un hôpital pour tenter de dépister des

marqueurs génétiques de maladie, en France, ces données reviennent ensuite en général à l'hôpital. Il est par la suite compliqué pour d'autres chercheurs ou cliniciens d'avoir accès à ces données, et le processus administratif peut devenir très long et lourd. Cette difficulté empêche bon nombre de projets de pouvoir se porter sur de grandes cohortes de données de patients, alors que ce sont souvent celles qui contiennent le plus d'information et la plus pertinente.

La méthodologie exposée ici, appliquée à une cohorte de patients partageant tous un critère phénotypique ou pathologique, permet de recréer des génomes qui sont semblables à ceux de cette cohorte en tous points, sans reproduire un génome d'un individu donné. Cette application permet donc de partager cette précieuse source de données sans porter atteinte au secret médical et le droit à la vie privée des personnes qui y ont contribué.

10.2.4. Jumeau virtuel

Le concept de génome synthétique présenté ici se rapproche du concept de jumeau virtuel, employé depuis plusieurs années dans les procédés de conception et réalisation de nombreux domaines, tels que l'aéronautique ou en architecture. Elle peut aussi servir pour le suivi d'objets connectés.

Ce terme fait référence à l'utilisation d'un modèle virtuel visant à reprendre les points essentiels d'un objet physique. Ce jumeau virtuel peut ensuite être utilisé pour de nombreuses fins, comme réaliser des simulations avec. Ces dernières années ont connu une augmentation drastique dans la puissance de calcul et de stockage disponibles, rendant cette solution bien plus facile à implémenter, alors que précédemment, elle était surtout trouvée dans des laboratoires dédiés ou des sections de recherche et développement de grandes entreprises.

On peut donc considérer que cette méthodologie pourrait permettre à un individu de créer un jumeau virtuel brouillé de son génome. On pourrait donc le configurer pour qu'il reproduise certains aspects, comme l'appartenance à une population et la présence d'une mutation pathogène, sans reproduire toutes ses caractéristiques. Ainsi, ce jumeau serait partageable librement sans crainte de la part de l'individu modèle.

Il serait aussi possible de créer des jumeaux virtuels brouillés de cohortes et populations entières. Ceux-ci permettraient de présenter à des utilisateurs potentiels de la cohorte ses caractéristiques globales avant de procéder à une demande officielle d'accès aux données.

10.3. Génomique

10.3.1. Projet 1000 Genomes

10.3.1.1. Méthodologie

Dans le projet des 1000 génomes, les équipes se focalisent sur le séquençage d'exomes : tous les participants à la phase trois ont eu un séquençage génomique complet ainsi qu'un séquençage d'exome. La profondeur moyenne du séquençage génomique complet est d'environ 7, et la profondeur moyenne pour le séquençage d'exome vaut environ 65 ((The 1000 Genomes Project Consortium 2015). Il est donc possible que des régions moins accessibles ou à faible complexité dans les régions non-codantes contiennent des mutations qui n'ont pas été détectées par ce projet, à cause de biais techniques dans le séquençage. La profondeur moyenne de 7 peut survoler de telles mutations, si la variance de cette valeur est forte (elle n'est cependant pas fournie, et la calculer personnellement demanderait beaucoup de temps et de ressources de calcul).

Ces exomes contiennent beaucoup d'information, mais pas l'intégralité de ce qui est pertinent en génomique. En portant un œil plus attentionné sur les parties non-codantes du génome, on pourrait y découvrir des variants ayant des impacts phénotypiques par voie régulatoire ou autre.

10.3.1.2. Manque de populations

Malgré son effort pour séquencer des individus de tous horizons, la phase 3 du 1KG ne parvient pas à capter toutes les populations du monde. Des critiques lui ont été portées pour son manque de données, notamment pour les personnes asiatiques (Lu and Xu 2013). Ce manque crée des vides dans l'espace des génomes que les modèles exposés ici peuvent parcourir.

Il existe aussi des petites populations qui parfois refusent de se faire génotyper, par exemple des communautés aborigènes d'Amérique ou d'Australie (Taitingfong et al. 2020). Ce refus vient souvent du mauvais traitement historique de ces populations par les autorités coloniales installées dans la région. Il y a aussi des craintes que ces données soient utilisées contre eux, par exemple pour leur faire payer plus cher l'accès aux soins si des marqueurs

génétiques de maladie sont mis en évidence.

Visuellement, sur la carte des populations intégrées dans le jeu de données (Oleksyk, Brukhin, and O'Brien 2015), on peut déjà y constater des manquement à l'œil nu. L'article ayant publié cette carte se focalise sur la Russie, et on peut y ajouter également d'autres pays d'Europe du Nord et de l'Est. Des régions telles que le Moyen-Orient et le Maghreb, ou encore l'Indonésie et les îles du Pacifique manquent aussi à l'appel. Ces régions sont voisines de régions incluses dans l'étude, mais présentent tout de même des spécificités.

Celles-ci sont les plus visibles, mais certaines populations peuvent rester invisibles au premier coup d'œil. Par exemple, il y a bien plus de sous-populations échantillonnées dans les pays développés que dans les pays en voie de développement. On pourrait donc attendre une exploration plus approfondie des habitants d'Afrique et d'Amérique du Sud. Notamment, cette réflexion peut aussi s'appliquer aux populations aborigènes des Amériques ou d'Australie qui se sont assimilées à d'autres populations, souvent d'origines caucasiennes.

10.3.1.3. Haplotypes imparfaits

Les génomes présents dans le 1KG sont indiqués sous format haplotypé. Cette information est basée sur d'autres projets de séquençage génomique, tels que HapMap, Omni, et le 1KG lui-même. Ces bases tentent de maintenir des listes des haplotypes les plus fréquents dans les génomes humains à travers le monde. Ces listes ne sont ni exhaustives ni exactes, et donc les données de base sur lesquelles le projet est construit sont approximatives. Nos connaissances en la matière sont bonnes, mais loin d'être parfaites.

10.3.2. Foyers de recombinaison

Les points chauds de recombinaison obtenus de (Halldorsson et al. 2019) sont basés sur les populations locales à laquelle l'équipe de recherche avait accès. Les islandais sont connus pour avoir une forte consanguinité, ce qui en fait une population qui a beaucoup dérivé par rapport aux autres populations du monde. De plus, les cartes de recombinaison ont été moyennés sur de nombreux facteurs comme le sexe et l'âge des participants. Ces foyers sont donc approximatifs, et il serait préférable de prendre en compte les taux de recombinaisons présents chez d'autres populations pour établir des limites entre sections de chromosomes qui représentent

un véritable consensus mondial.

10.3.3. Effet mutagène des recombinaisons

Les recombinaisons effectuées par la cellule lors de la méiose ont un effet mutagène sur le génome de la cellule la contenant, vu qu'elles sont effectuées par le biais de cassures double brin de l'ADN. Avec ce processus, les foyers de recombinaison sélectionnés comme limites de régions de SNP corrélés représentent aussi des foyers de mutations. On trouve donc davantage de mutations autour de celles-ci, ce qui peut brouiller leur effet de frontière.

Cependant, ces mutations ont moins tendance à être fixés dans la population et à présenter des propriétés de marqueurs de populations. L'inquiétude n'est donc pas très grande quant à cet effet.

10.4. Perspectives techniques

10.4.1. Augmenter l'étendue des échantillons

En passant à des profils de mutations à analyser plus grands, il faut prendre en compte plusieurs facteurs comme la malédiction de la dimension, qui pourraient perturber nos attentes.

Les analyses réalisées sur le jeu de données de 1Mb prennent en compte 730 variables latentes pour 5 000 échantillons, et présentent déjà un équilibre tendu. Les résultats qu'il crée sont bons, mais présentent des biais, et il a fallu trouver des paramètres spécifiques pour les obtenir. Il est raisonnable de craindre un effondrement du modèle en tentant de passer à une échelle plus large.

Plusieurs options sont envisageables pour combattre cette éventualité. La première serait de collecter plus d'échantillons, venant de plusieurs bases de données, pour renforcer le jeu de d'entraînement utilisé. Dans ce cas, il faudra veiller à diversifier les nouveaux échantillons, comme l'a fait le 1000 Genomes Project, pour ne pas risquer de biaiser l'apprentissage qu'effectue notre modèle. Une autre approche serait de tenter d'explorer une compression plus grande par les autoencodeurs, pour réduire le nombre de variables présentées au modèle. Cependant, cela ne changerait pas la quantité d'information (sauf s'il y a une perte d'information avec ces nouveaux encodeurs, ce qui n'est pas souhaitable), et donc il n'y a pas de garantie que cela affecterait la difficulté de la tâche présentée. De plus, les autoencodeurs variationnels utilisés ici ne sont pas les plus fiables, et présentent des biais dans leur espace latent. Tenter de leur faire comprimer les données encore plus pourrait renforcer ces biais et ainsi rendre les données inexploitables.

10.4.2. Super-étendeur basé sur les TAD

Pour tenter de pallier le biais d'enchaînement des étendeurs d'espace latent, il avait été imaginé lors de réunions avec mes directeurs d'en enchaîner peu mais intelligemment. En choisissant un "centre" basé sur des données biologiques et en étendant successivement à partir de cette section, on vient limiter le biais induit. Cette limite sera je pense modérée, vu qu'il s'agit toujours de l'enchaînement d'étendeurs, mais tout bénéfice est bon à prendre.

Le centre envisagé pour cette étape serait le centre de TAD, vu que ceux-ci définissent de grandes régions chromosomiques ayant des liens métaboliques prouvés entre éléments à l'intérieur. De cette manière, on vient créer les génomes simulés TAD par TAD, et tout lien de cohérence s'établit maintenant entre ces centres de TAD, en comptant sur les étendeurs pour maintenir l'identité génomique.

10.4.3. VAEs chevauchants

Une option à explorer pour la génération de génomes cohérents serait d'exploiter la stratégie utilisée par des outil de génération d'image : partir du début et continuer jusqu'à la fin (Oord et al. 2016). Ceci présente, de premier abord, d'importants biais : le début d'un chromosome de contient pas plus d'informations génomiques que d'autres, bien au contraire. Il ne paraît pas très logique, en considérant ce critère, d'employer cette technique. Cependant, cette heuristique a fait ses preuves en génération d'images, où le premier pixel en haut à gauche de comporte pas spécialement d'information.

Pour implémenter cette stratégie, j'imagine un ensemble de VAEs où le premier est tout à fait classique, mais tous les autres présentent une différence majeure. Le modèle n+1 chevauche la section de données attribuées au modèle n, mais ne voit pas les vrais données d'entrée pour cette partie. À la place, elle prend en entrée la représentation latente de la section de génome traitée par le modèle n, et doit reconstruire correctement les données de référence à partir de ces données.

Ceci est en quelque sorte un approche semblable au modèle d'étendeur d'espace latent, mais aidé en partie par les données de référence non chevauchantes.

10.4.4. Espace latent "étiqueté"

Pendant le temps de ce projet, l'utilisation de modèles à diffusion stable a explosé, notamment pour la génération d'images (Rombach et al. 2022). Une particularité de ce type de modèle est que son espace latent comporte à la fois des données latentes déterminées par le réseau et des informations données par l'utilisateur (par exemple, une description de l'image à générer).

Ce type d'espace latent aurait pu être implémenté pour cette tâche : en indiquant en entrée

une liste de caractéristiques génomiques (Colonna et al. 2014) qu'on souhaite retrouver dans le génome généré, l'utilisateur bénéficie d'un moyen de contrôler la sortie du réseau.

Ceci permet aussi d'implémenter une forme de continuité entre les sections, puisque cette information fournie par l'utilisateur est invariable.

10.4.5. Réseau de neurones récurrent pour les mutations rares

Les données que tentent de simuler les modèles présentés ici sont uniquement celles prévisibles par un outil d'apprentissage statistique, c'est-à-dire ayant une fréquence suffisamment élevée dans la population pour que le modèle comprenne ses corrélations. Ce seuil retire donc une grande partie des données contenues dans le fichier vcf de base. Beaucoup de ces mutations rares contiennent assez peu d'information, mais c'est aussi dans cette catégorie qu'on peut trouver certains marqueurs de maladie.

Une stratégie que j'aurais aimé mettre en place pour tenter de recréer ces mutations rares est l'utilisation de réseaux de neurones récurrents. Ils ont la particularité de pouvoir prédire une séquence de taille variable plutôt que limitée à une taille fixe, et de même pour les données qu'il reçoit en entrée (Miljanovic 2012).

L'approche aurait été la suivante : le jeu de données d'entrée est composé des mutations fréquentes d'une personne, et l'outil tente de prédire la position des nouvelles mutations, ainsi que leur nature (transition, transversion, petit indel, grand indel, etc). Nous voyons ici une différence assez importante par rapport aux précédents modèles : celui-ci serait conscient de la nature spatiale de l'ADN, alors que les autres y sont aveugles.

Cette partie du projet aurait pu constituer un stage d'étudiant de master par exemple. Malheureusement, pas manque de temps, je n'ai pas pu mettre en place un appel à candidats digne de ce nom.

10.4.6. De chromosome à chromosome

10.4.6.1. Principe

De manière semblable à l'étendeur d'espace latent, nous considérons que la projection dans l'espace latent d'un individu par un autoencodeur représente une sorte d'identité

génomique. S'il est possible de passer de section à section voisine sur un même chromosome, il n'est pas impossible qu'on puisse aussi prédire l'identité génomique d'un individu sur un autre chromosome à partir de ces informations.

Idéalement, il serait souhaitable de prendre en compte autant d'information génomique que possible depuis le chromosome de départ pour donner autant d'information que possible au modèle pour sa prédiction. C'est pourquoi je me suis tourné vers les modèles de convolution, qui permettent de condenser des informations à longue distance très efficacement. En appliquant cette convolution sur toutes les projections latentes d'un individu sur le chromosome 1, j'espère récolter suffisamment d'information pour prédire l'identité latente de cet individu pour une section du chromosome 2.

10.4.6.2. Difficultés

La cohérence à établir entre sections d'un même chromosome est assez intuitive à construire. Parvenir à établir des liens semblables entre chromosomes est une tâche autrement plus difficile.

Les chromosomes que portent une personne montreront certainement un minimum de lien entre eux, puisqu'ils ont plus de chances de provenir de la même ascendance génétique. Mais à ce jour, aucun lien de migration jointe n'a été mis en évidence lors de la méiose.

Une option serait de considérer les deux haplotypes de chaque individu en même temps pour prédire les deux haplotypes sur le chromosome suivant. Ainsi, on prend en compte toute l'information génomique distillée par les autoencodeurs pour reprendre les prédictions.

Une autre option serait d'effectuer 4 prédictions par individu : chr1hapA → chr2hapA, chr1hapA → chr2hapB, chr1hapB → chr2hapB, et chr1hapB → chr2hapB, et ensuite évaluer celle ayant la meilleure performance pour tenter de déterminer quels chromosomes sont hérités du même parent. L'héritage ne sera cependant pas parfait, puisqu'il y a des recombinaisons à chaque méiose.

10.4.6.3. Solution optimale

Idéalement, j'aimerais pouvoir obtenir une identité génomique latente à partir d'un chromosome d'un individu, et que cette identité puisse prédire le profil de mutations de cet

individu sur tous ses chromosomes. Cette tâche est nettement plus difficile à atteindre, à cause de la variance inhérente entre chromosomes.

10.4.7. Spécialisation du modèle sur population spécifique

Un des buts finaux de ce projet est de pouvoir recréer des génomes ressemblant à des génomes de personnes faisant partie de cohortes d'intérêt médical. Cependant, Le WGAN ne permet pas de spécifier au générateur de caractéristiques souhaitées dans l'individu généré. Pour obtenir un générateur capable de créer des patients simulés, il faudrait l'entraîner sur la cohorte que l'on souhaite reproduire.

Ceci amène une nouvelle difficulté : le nombre d'individus que la cohorte contient. Recréer l'ensemble du processus d'entraînement sur un nombre réduit de génomes risque d'introduire de gros biais dans l'apprentissage. Il est même probable que la majorité des cohortes de patients ne contiennent pas suffisamment de génomes pour entraîner un modèle stable.

Une possibilité serait de commencer l'entraînement sur un grand jeu de données génomique, par exemple 1KG, jusqu'à obtenir un modèle capable de créer des génomes vraisemblables, puis de le spécialiser sur les génomes de la cohorte d'intérêt. Vu que le critique voit uniquement les données de référence de son lot d'entraînement, et que celui-ci force le générateur à émettre une distribution proche de celles-ci, un ré-entraînement crée donc un générateur spécialisé dans le dernier jeu de données sur lequel il a été entraîné.

10.5. Interprétations

10.5.1. Comment bien mesurer la vraisemblance d'une population?

Il n'existe pas à ce jour de méthode pour mesurer la vraisemblance d'une population, vu qu'il n'existe pas non plus de population simulée. Les méthodes mises en place ici et dans (Yelmen et al. 2021) permettent de suggérer une vraisemblance, mais ne la prouvent pas.

Mettre au point un test capable de valider ou non le "réalisme" d'un génome nécessite de prendre en compte des éléments structurels et fonctionnels, vu que ce sont ces facteurs qui définiront sa viabilité *in vivo*. Chacun des critères à prendre en compte pour ces éléments devra être évalué individuellement afin de les valider.

10.5.2. Distance entre réel et simulé

Lors de la dernière section d'analyse de ma population simulée (9.3.2 : "Mesurer le respect de l'anonymat"), je mesure la distance entre chaque génome simulé et son plus proche voisin dans le jeu de données de référence. J'ai utilisé la distance de Hamming, mais il existe bien d'autres mesures de distance applicables, ainsi que des mesures synthétiques prenant en compte plusieurs facteurs.

Une question d'interprétation à considérer est : quelle est la distance souhaitée entre un individu simulé et son plus proche voisin de référence ?

Il est tout à fait possible qu'il n'existe pas de réponse objectivement correcte à cette question. D'un côté, une distance trop petite signifie que mes simulations sont presque des copies conformes de mes échantillons de référence. Une distance trop grande indique que mes simulations sont très différentes de leur cible. Je mesure aussi la vraisemblance par d'autres moyens, qui permettent de vérifier que mes génomes simulés sont réalistes, ce qui entre en considération aussi. Mais ensuite vient des questionnements sur l'équilibre entre ces deux : quelle perte de vraisemblance peut-on accepter pour un gain de respect de l'anonymat ? Là encore, la réponse dépend très certainement de l'application spécifique qu'on cherche à réaliser.

Dans (Yelmen et al. 2021), l'équipe utilise une mesure appelée AA_{TS}, qui prend aussi en compte la distance entre voisins dans le jeu de données de référence. Cette approche permet de

“normaliser” la distance entre individu simulé et individu de référence par la distance “attendue” entre deux individus. Cette normalisation est cependant biaisée par la variance locale dans les données de référence, et est établie pour toute une population, et non par individu. Des métriques basées explicitement sur la proximité des données que l’on simule seraient les bienvenues.

11. Implémentation

11.1. Outils utilisés

Tous les modèles de réseaux de neurones décrits dans ce manuscrit, ainsi que la grande majorité des traitements de données associés à leur entraînement et interprétation, ont été implémentés avec le langage Python (version 3.8.10) en utilisant la librairie de gestion de réseaux de neurons TensorFlow (version 2.12.0), avec l'interface Keras (version identique). Les librairies de gestion de données numpy (version 1.24.3) et pandas (version 1.2.4), ainsi que matplotlib (version 3.7.1) et seaborn (version 0.12.2) pour la génération de graphiques ont également été utilisés. Quelques transformations de données ont aussi été effectuées en shell et en R (version 4.3.1).

Les calculs plus lourds, comme la séparation des mutations en sections, ainsi que les entraînements des modèles ont été réalisés sur un serveur dédié utilisant slurm (version 22). L'outil parallel (publication du 22/12/2016) (Tange 2011) a été très utile dans l'entraînement des autoencodeurs et des étendeurs, pour un gain de temps remarquable. Il est téléchargeable à l'adresse https://ftp.gnu.org/gnu/parallel/ .

11.2. Implémentations spécifiques

Les autoencodeurs variationnels nécessitent plusieurs fonctions et classes spécifiques pour leur fonctionnement, qui ne sont pas inclus de base dans TensorFlow. Ces éléments ont été repris depuis les manuels disponibles sur le site de Keras (Chollet 2020) et adaptés pour convenir aux données de ce projet.

Les mesures de dissimilarité employées dans l'entraînement des étendeurs d'espace latent (section 9.2.2.3.1.) ne sont pas implémentées dans les fonctions proposées. Pour les appliquer sur ces modèles, je les ai définis en utilisant les fonctions dédiées de TensorFlow servant à établir des fonctions de perte. Ces fonctions présentent des contraintes particulières, dues aux besoins de calculer la perte du modèle à la volée de nombreuses fois pendant l'entraînement. Cette étape représente donc un apprentissage de ma part ainsi qu'un effort dédié pour implémenter ces mesures.

L'entraînement du GAN utilisant la perte de Wasserstein nécessite d'entraîner le critique plusieurs fois à chaque itération. Ceci implique donc d'utiliser une boucle d'entraînement personnalisée, et non pas la fonction d'entraînement native de TensorFlow. J'ai implémenté cette boucle et inclu plusieurs fonctionnalités supplémentaires, comme la possibilité d'enregistrer des points de sauvegarde pendant l'entraînement.

11.3. Disponibilité du code

Les scripts essentiels aux travaux de ce manuscrit sont disponibles dans un dépôt GitHub à l'adresse https://github.com/callum-b/H2G2 .

Il contient les principaux scripts employés, mais pas tous. J'ai effectué un travail de clarification et documentation de ces scripts afin de les rendre plus accessibles à la communauté scientifique. L'établissement de certains graphiques et transformations de données sont basiques, et ne nécessitent pas de script personnel dédié.

Pour le prétraitement des données, je fournis les étapes de recherche de foyers de recombinaison, de séparation des données selon ces foyers et leur conversion en données haplotypiques. La préparation des fichiers d'entrée de Chromopainter est également présente.

Les trois principaux modèles ayant produit des résultats satisfaisants, à savoir l'autoencodeur, le VAE et le WGAN, disposent de fonctions permettant de créer les modèles. Les fonctions et classes nécessaires à leur établissement, tels que les fonctions de perte ou l'échantillonnage, sont également présentes.

12. Références

Abadi, Martín, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. 'Deep Learning with Differential Privacy'. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 308–18. https://doi.org/10.1145/2976749.2978318.

Abay, Nazmiye Ceren, Yan Zhou, Murat Kantarcioglu, Bhavani Thuraisingham, and Latanya Sweeney. 2019. 'Privacy Preserving Synthetic Data Release Using Deep Learning'. In *Machine Learning and Knowledge Discovery in Databases*, edited by Michele Berlingerio, Francesco Bonchi, Thomas Gärtner, Neil Hurley, and Georgiana Ifrim, 510–26. Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-10925-7_31.

'Activation Function'. 2023. In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Activation_function&oldid=1178730781.

Altshuler, David, Peter Donnelly, and The International HapMap Consortium. 2005. 'A Haplotype Map of the Human Genome'. *Nature* 437 (7063): 1299–1320. https://doi.org/10.1038/nature04226.

Amini, Alexander, Ava Soleimany, Sertac Karaman, and Daniela Rus. 2019. 'Spatial Uncertainty Sampling for End-to-End Control'. arXiv. http://arxiv.org/abs/1805.04829.

Angelica, Amara, and Jürgen Schmidhuber. 2012. 'How Bio-Inspired Deep Learning Keeps Winning Competitions « the Kurzweil Library + Collections'. 2012. https://www.thekurzweillibrary.com/how-bio-inspired-deep-learning-keeps-winning-competitions.

Apgar, Tucker L., and Charles R. Sanders. 2022. 'Compendium of Causative Genes and Their Encoded Proteins for Common Monogenic Disorders'. *Protein Science: A Publication of the Protein Society* 31 (1): 75–91. https://doi.org/10.1002/pro.4183.

Arjovsky, Martin, Soumith Chintala, and Léon Bottou. 2017. 'Wasserstein GAN'. arXiv. http://arxiv.org/abs/1701.07875.

Baker, Monya. 2012. 'De Novo Genome Assembly: What Every Biologist Should Know'. *Nature Methods* 9 (4): 333–37. https://doi.org/10.1038/nmeth.1935.

Barlow-Stewart, K., M. Liepins, A. Doble, and M. Otlowski. 2018. 'How Are Genetic Test Results Being Used by Australian Life Insurers?' *European Journal of Human Genetics* 26 (9): 1248–56. https://doi.org/10.1038/s41431-018-0198-z.

Barnett, Vic, and Toby Lewis. 1978. *Outliers in Statistical Data*. Wiley Series in Probability and Mathematical Statistics. Chichester: Wiley. http://www.gbv.de/dms/hbz/toc/ht000021894.pdf.

Barnhart, Benjamin J. 1989. 'The Department of Energy (DOE) Human Genome Initiative'. *Genomics* 5 (3): 657–60. https://doi.org/10.1016/0888-7543(89)90041-4.

Barrett, J. C., B. Fry, J. Maller, and M. J. Daly. 2005. 'Haploview: Analysis and Visualization of LD and Haplotype Maps'. *Bioinformatics* 21 (2): 263–65. https://doi.org/10.1093/bioinformatics/bth457.

Barriere, Sylvain, Aubin Thomas, Lucile Broseus, Julie Brooke, Claudio Lorenzi, Jean-Philippe Villemin, Gregory Beurier, et al. 2019. 'GECKO Is a Genetic Algorithm to Classify and Explore High Throughput Sequencing Data'. *Communications Biology* 2 (June): 222.

https://doi.org/10.1038/s42003-019-0456-9.
Baudat, F., J. Buard, C. Grey, A. Fledel-Alon, C. Ober, M. Przeworski, G. Coop, and B. de Massy. 2010. 'PRDM9 Is a Major Determinant of Meiotic Recombination Hotspots in Humans and Mice'. *Science (New York, N.Y.)* 327 (5967): 836–40. https://doi.org/10.1126/science.1183439.
'Beat Biden - YouTube'. 2023. 25 April 2023. https://www.youtube.com/watch?v=kLMMxgtxQ1Y.
Belenguer, Lorenzo. 2022. 'AI Bias: Exploring Discriminatory Algorithmic Decision-Making Models and the Application of Possible Machine-Centric Solutions Adapted from the Pharmaceutical Industry'. *Ai and Ethics* 2 (4): 771–87. https://doi.org/10.1007/s43681-022-00138-8.
Bernstein, Carol, Harris Bernstein, Claire M. Payne, and Harinder Garewal. 2002. 'DNA Repair/pro-Apoptotic Dual-Role Proteins in Five Major DNA Repair Pathways: Fail-Safe Protection against Carcinogenesis'. *Mutation Research/Reviews in Mutation Research* 511 (2): 145–78. https://doi.org/10.1016/S1383-5742(02)00009-1.
Boddupally, Kavitha, and Esther Rani Thuraka. 2024. 'Artificial Intelligence for Prenatal Chromosome Analysis'. *Clinica Chimica Acta; International Journal of Clinical Chemistry* 552 (January): 117669. https://doi.org/10.1016/j.cca.2023.117669.
Boriah, Shyam, Varun Chandola, and Vipin Kumar. 2008. 'Similarity Measures for Categorical Data: A Comparative Evaluation'. In *Proceedings of the 2008 SIAM International Conference on Data Mining (SDM)*, 243–54. Proceedings. Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611972788.22.
Botstein, D, R L White, M Skolnick, and R W Davis. 1980. 'Construction of a Genetic Linkage Map in Man Using Restriction Fragment Length Polymorphisms.' *American Journal of Human Genetics* 32 (3): 314–31.
Boveri, Theodor. 1904. *Ergebnisse Über Die Konstitution Der Chromatischen Substanz Des Zellkerns*. Jena: G. Fischer. https://doi.org/10.5962/bhl.title.28064.
Branton, Daniel, David W. Deamer, Andre Marziali, Hagan Bayley, Steven A. Benner, Thomas Butler, Massimiliano Di Ventra, et al. 2008. 'The Potential and Challenges of Nanopore Sequencing'. *Nature Biotechnology* 26 (10): 1146–53. https://doi.org/10.1038/nbt.1495.
Braverman, Genna, Zachary E. Shapiro, and Jonathan A. Bernstein. 2018. 'Ethical Issues in Contemporary Clinical Genetics'. *Mayo Clinic Proceedings: Innovations, Quality & Outcomes* 2 (2): 81–90. https://doi.org/10.1016/j.mayocpiqo.2018.03.005.
Budowle, Bruce, and Angela van Daal. 2008. 'Forensically Relevant SNP Classes'. *BioTechniques* 44 (5): 603–10. https://doi.org/10.2144/000112806.
Caesar, Holger. 2017. 'Really-Awesome-Gan'. https://github.com/nightrome/really-awesome-gan.
Campbell, Samuel G., Pat Croskerry, and David A. Petrie. 2017. 'Cognitive Bias in Health Leaders'. *Healthcare Management Forum* 30 (5): 257–61. https://doi.org/10.1177/0840470417716949.
Cariaso, Michael, and Greg Lennon. 2012. 'SNPedia: A Wiki Supporting Personal Genome Annotation, Interpretation and Analysis'. *Nucleic Acids Research* 40 (Database issue): D1308-1312. https://doi.org/10.1093/nar/gkr798.
Carothers, E. Eleanor. 1913. 'The Mendelian Ratio in Relation to Certain Orthopteran

Chromosomes'. *Journal of Morphology* 24 (4): 487–511. https://doi.org/10.1002/jmor.1050240403.
Carrasquilla, Germán D., Lars Ängquist, Thorkild I. A. Sørensen, Tuomas O. Kilpeläinen, and Ruth J. F. Loos. 2023. 'Child-to-Adult Body Size Change and Risk of Type 2 Diabetes and Cardiovascular Disease'. *Diabetologia*, December. https://doi.org/10.1007/s00125-023-06058-4.
Carvajal-Rodríguez, Antonio. 2008a. 'GENOMEPOP: A Program to Simulate Genomes in Populations'. *BMC Bioinformatics* 9 (1): 223. https://doi.org/10.1186/1471-2105-9-223.
———. 2008b. 'Simulation of Genomes: A Review'. *Current Genomics* 9 (3): 155–59. https://doi.org/10.2174/138920208784340759.
———. 2017. 'GenomePop2: A Program to Simulate SNPs in Populations'. 2017. http://acraaj.webs.uvigo.es/GenomePop2.htm.
Check Hayden, Erika. 2013. 'Privacy Loophole Found in Genetic Databases'. *Nature*, January. https://doi.org/10.1038/nature.2013.12237.
Chen, Tianqi, and Carlos Guestrin. 2016. 'XGBoost: A Scalable Tree Boosting System'. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–94. KDD '16. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785.
Chen, Zhengming, Liming Lee, Junshi Chen, Rory Collins, Fan Wu, Yu Guo, Pamela Linksted, and Richard Peto. 2005. 'Cohort Profile: The Kadoorie Study of Chronic Disease in China (KSCDC)'. *International Journal of Epidemiology* 34 (6): 1243–49. https://doi.org/10.1093/ije/dyi174.
Chen, Zhisheng. 2023. 'Ethics and Discrimination in Artificial Intelligence-Enabled Recruitment Practices'. *Humanities and Social Sciences Communications* 10 (1): 1–12. https://doi.org/10.1057/s41599-023-02079-x.
Cheon, Jung Hee, Andrey Kim, Miran Kim, and Yongsoo Song. 2017. 'Homomorphic Encryption for Arithmetic of Approximate Numbers'. In *Advances in Cryptology – ASIACRYPT 2017*, edited by Tsuyoshi Takagi and Thomas Peyrin, 409–37. Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-70694-8_15.
Chin, Lena J., Elena Ratner, Shuguang Leng, Rihong Zhai, Sunitha Nallur, Imran Babar, Roman-Ulrich Muller, et al. 2008. 'A SNP in a Let-7 microRNA Complementary Site in the KRAS 3′ Untranslated Region Increases Non–Small Cell Lung Cancer Risk'. *Cancer Research* 68 (20): 8535–40. https://doi.org/10.1158/0008-5472.CAN-08-2129.
Chintala, Souminth. 2016. 'Release Alpha-1 Release · Pytorch/Pytorch'. GitHub. 2016. https://github.com/pytorch/pytorch/releases/tag/v0.1.1.
Chollet, François. 2020. 'Keras Documentation: Variational AutoEncoder'. 2020. https://keras.io/examples/generative/vae/.
Choudhury, Avishek, and Onur Asan. 2020. 'Role of Artificial Intelligence in Patient Safety Outcomes: Systematic Literature Review'. *JMIR Medical Informatics* 8 (7): e18599. https://doi.org/10.2196/18599.
Colbert-White, Erin, and Allison Kaufman. 2023. *Historical Perspective on Animal Cognition*. Springer Publishing Company. https://connect.springerpub.com/content/book/978-0-8261-6235-9/chapter/ch01.

Colonna, Vincenza, Qasim Ayub, Yuan Chen, Luca Pagani, Pierre Luisi, Marc Pybus, Erik Garrison, Yali Xue, and Chris Tyler-Smith. 2014. 'Human Genomic Regions with Exceptionally High Levels of Population Differentiation Identified from 911 Whole-Genome Sequences'. *Genome Biology* 15 (6): R88. https://doi.org/10.1186/gb-2014-15-6-r88.

Cortes, Corinna, and Vladimir Vapnik. 1995. 'Support-Vector Networks'. *Machine Learning* 20 (3): 273–97. https://doi.org/10.1007/BF00994018.

Coste, Simon. 2022. 'The Kullback-Leibler Divergence between Gaussians'. 2022. https://scoste.fr/posts/dklgaussian/.

Cotton, Richard G. H., Arleen D. Auerbach, Myles Axton, Carol Isaacson Barash, Samuel F. Berkovic, Anthony J. Brookes, John Burn, et al. 2008. 'The Human Variome Project'. *Science (New York, N.Y.)* 322 (5903): 861–62. https://doi.org/10.1126/science.1167363.

Creighton, Harriet B., and Barbara McClintock. 1931. 'A Correlation of Cytological and Genetical Crossing-Over in Zea Mays'. *Proceedings of the National Academy of Sciences* 17 (8): 492–97. https://doi.org/10.1073/pnas.17.8.492.

Crick, F. H. 1958. 'On Protein Synthesis'. *Symposia of the Society for Experimental Biology* 12: 138–63.

Curry, Haskell B. 1944. 'The Method of Steepest Descent for Non-Linear Minimization Problems'. *Quarterly of Applied Mathematics* 2 (3): 258–61.

deCODE. 2012. 'SCIENCE'. deCODE Genetics. 6 September 2012. https://www.decode.com/research/.

Dhar, Payal. 2020. 'The Carbon Impact of Artificial Intelligence'. *Nature Machine Intelligence* 2 (8): 423–25. https://doi.org/10.1038/s42256-020-0219-9.

Dixon, Jesse R., Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S. Liu, and Bing Ren. 2012. 'Topological Domains in Mammalian Genomes Identified by Analysis of Chromatin Interactions'. *Nature* 485 (7398): 376–80. https://doi.org/10.1038/nature11082.

Dorigo, M., and L.M. Gambardella. 1997. 'Ant Colony System: A Cooperative Learning Approach to the Traveling Salesman Problem'. *IEEE Transactions on Evolutionary Computation* 1 (1): 53–66. https://doi.org/10.1109/4235.585892.

Edgar, Robert, George Asimenos, Serafim Batzoglou, and Arend Sidow. 2009. 'Evolver Home Page'. 2009. http://www.drive5.com/evolver/.

Eid, John, Adrian Fehr, Jeremy Gray, Khai Luong, John Lyle, Geoff Otto, Paul Peluso, et al. 2009. 'Real-Time DNA Sequencing from Single Polymerase Molecules'. *Science* 323 (5910): 133–38. https://doi.org/10.1126/science.1162986.

Ester, Martin, Hans-Peter Kriegel, and Xiaowei Xu. 1996. 'A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise'.

European Genomic Data Infrastructure. 2018. 'Beyond One Million Genomes (B1MG) Project'. 2018. https://b1mg-project.eu/.

Folkersen, Lasse, Oliver Pain, Andrés Ingason, Thomas Werge, Cathryn M. Lewis, and Jehannine Austin. 2020. 'Impute.Me: An Open-Source, Non-Profit Tool for Using Data From Direct-to-Consumer Genetic Testing to Calculate and Interpret Polygenic Risk Scores'. *Frontiers in Genetics* 11. https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2020.00578.

Fournier, Quentin, and Daniel Aloise. 2019. ‘Empirical Comparison between Autoencoders and Traditional Dimensionality Reduction Methods’. In *2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, 211–14. https://doi.org/10.1109/AIKE.2019.00044.

Franklin, Rosalind E., and R. G. Gosling. 1953. ‘Evidence for 2-Chain Helix in Crystalline Structure of Sodium Deoxyribonucleate’. *Nature* 172 (4369): 156–57. https://doi.org/10.1038/172156a0.

Fu, Jack M., F. Kyle Satterstrom, Minshi Peng, Harrison Brand, Ryan L. Collins, Shan Dong, Brie Wamsley, et al. 2022. ‘Rare Coding Variation Provides Insight into the Genetic Architecture and Phenotypic Context of Autism’. *Nature Genetics* 54 (9): 1320–31. https://doi.org/10.1038/s41588-022-01104-0.

Fu, King-Sun. 1970. ‘Learning Control Systems–Review and Outlook’. *IEEE Transactions on Automatic Control* 15 (2): 210–21. https://doi.org/10.1109/TAC.1970.1099405.

Gentry, Craig. 2009. ‘Fully Homomorphic Encryption Using Ideal Lattices’. In *Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing*, 169–78. STOC ’09. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/1536414.1536440.

Gentry, Craig, Shai Halevi, and Nigel P. Smart. 2012. ‘Homomorphic Evaluation of the AES Circuit’. In *Advances in Cryptology – CRYPTO 2012*, edited by Reihaneh Safavi-Naini and Ran Canetti, 7417:850–67. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-32009-5_49.

Gentry, Craig, Amit Sahai, and Brent Waters. 2013. ‘Homomorphic Encryption from Learning with Errors: Conceptually-Simpler, Asymptotically-Faster, Attribute-Based’. In *Advances in Cryptology – CRYPTO 2013*, edited by Ran Canetti and Juan A. Garay, 8042:75–92. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-40041-4_5.

Gibbs, Richard A., John W. Belmont, Paul Hardenbol, Thomas D. Willis, Fuli Yu, Huanming Yang, Lan-Yang Ch’ang, et al. 2003. ‘The International HapMap Project’. *Nature* 426 (6968): 789–96. https://doi.org/10.1038/nature02168.

Goeddel, D V, D G Kleid, F Bolivar, H L Heyneker, D G Yansura, R Crea, T Hirose, A Kraszewski, K Itakura, and A D Riggs. 1979. ‘Expression in Escherichia Coli of Chemically Synthesized Genes for Human Insulin.’ *Proceedings of the National Academy of Sciences of the United States of America* 76 (1): 106–10.

Goodfellow, Ian J., Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. ‘Generative Adversarial Networks’. arXiv. https://doi.org/10.48550/arXiv.1406.2661.

Guéron, Maurice, and Jean-Louis Leroy. 1995. ‘Studies of Base Pair Kinetics by NMR Measurement of Proton Exchange’. In *Methods in Enzymology*, 261:383–413. Nuclear Magnetic Resonance and Nucleic Acids. Academic Press. https://doi.org/10.1016/S0076-6879(95)61018-9.

Gulrajani, Ishaan, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. 2017. ‘Improved Training of Wasserstein GANs’. arXiv. http://arxiv.org/abs/1704.00028.

Halldorsson, Bjarni V., Gunnar Palsson, Olafur A. Stefansson, Hakon Jonsson, Marteinn T. Hardarson, Hannes P. Eggertsson, Bjarni Gunnarsson, et al. 2019. ‘Characterizing

Mutagenic Effects of Recombination through a Sequence-Level Genetic Map'. *Science* 363 (6425): eaau1043. https://doi.org/10.1126/science.aau1043.
Hassold, T., L. Judis, E.R. Chan, S. Schwartz, A. Seftel, and A. Lynn. 2004. 'Cytological Studies of Meiotic Recombination in Human Males'. *Cytogenetic and Genome Research* 107 (3–4): 249–55. https://doi.org/10.1159/000080602.
Hirschhorn, Joel N., and Mark J. Daly. 2005. 'Genome-Wide Association Studies for Common Diseases and Complex Traits'. *Nature Reviews Genetics* 6 (2): 95–108. https://doi.org/10.1038/nrg1521.
Ho, Tin Kam. 1995. 'Random Decision Forests'. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, 1:278–82 vol.1. https://doi.org/10.1109/ICDAR.1995.598994.
Holder, Ross P., and Jules R. Tapamo. 2017. 'Improved Gradient Local Ternary Patterns for Facial Expression Recognition'. *EURASIP Journal on Image and Video Processing* 2017 (1): 42. https://doi.org/10.1186/s13640-017-0190-5.
Hutton, Joseph, Tahzeeb Fatima, Tanya J. Major, Ruth Topless, Lisa K. Stamp, Tony R. Merriman, and Nicola Dalbeth. 2018. 'Mediation Analysis to Understand Genetic Relationships between Habitual Coffee Intake and Gout'. *Arthritis Research & Therapy* 20 (1): 135. https://doi.org/10.1186/s13075-018-1629-5.
Ihde, Don. 1990. 'Technology and the Lifeworld'. *Indiana University Press* (blog). 1990. https://iupress.org/9780253205605/technology-and-the-lifeworld/.
Illumina. 2017. 'Welcome to Immense Discovery Power'. 2017. https://emea.illumina.com/systems/sequencing-platforms/novaseq.html.
Israel in Panama [@IsraelinPanama]. 2023. 'Rumbo a Israel... en estos momentos están saliendo jóvenes israelíes desde Panamá para enlistarse en el ejército y cuidar a nuestros ciudadanos🇮🇱❤️ En nombre de la Embajada de Israel queremos agradecer a todos las organizaciones locales y personas por todo su apoyo en estos… https://t.co/fjiwOewmpt'. Tweet. *Twitter*. https://twitter.com/IsraelinPanama/status/1711535585775530373.
Jackson, Peter. 1998. *Introduction to Expert Systems*. 3rd ed. USA: Addison-Wesley Longman Publishing Co., Inc.
Jain, Miten, Hugh E. Olsen, Benedict Paten, and Mark Akeson. 2016. 'The Oxford Nanopore MinION: Delivery of Nanopore Sequencing to the Genomics Community'. *Genome Biology* 17 (1): 239. https://doi.org/10.1186/s13059-016-1103-0.
Jia, Xiaowei, Aosen Wang, Xiaoyi Li, Guangxu Xun, Wenyao Xu, and Aidong Zhang. 2015. 'Multi-Modal Learning for Video Recommendation Based on Mobile Application Usage'. In *2015 IEEE International Conference on Big Data (Big Data)*, 837–42. Santa Clara, CA, USA: IEEE. https://doi.org/10.1109/BigData.2015.7363830.
Juan, Liran, Yongtian Wang, Jingyi Jiang, Qi Yang, Qinghua Jiang, and Yadong Wang. 2020. 'PGsim: A Comprehensive and Highly Customizable Personal Genome Simulator'. *Frontiers in Bioengineering and Biotechnology* 8. https://www.frontiersin.org/articles/10.3389/fbioe.2020.00028.
Karczewski, Konrad J., Laurent C. Francioli, Grace Tiao, Beryl B. Cummings, Jessica Alföldi, Qingbo Wang, Ryan L. Collins, et al. 2020. 'The Mutational Constraint Spectrum Quantified from Variation in 141,456 Humans'. *Nature* 581 (7809): 434–43. https://doi.org/10.1038/s41586-020-2308-7.

Karczewski, Konrad J., Ben Weisburd, Brett Thomas, Matthew Solomonson, Douglas M. Ruderfer, David Kavanagh, Tymor Hamamsy, et al. 2017. 'The ExAC Browser: Displaying Reference Data Information from over 60 000 Exomes'. *Nucleic Acids Research* 45 (D1): D840–45. https://doi.org/10.1093/nar/gkw971.

Karwath, Andreas, Karina V Bunting, Simrat K Gill, Otilia Tica, Samantha Pendleton, Furqan Aziz, Andrey D Barsky, et al. 2021. 'Redefining β-Blocker Response in Heart Failure Patients with Sinus Rhythm and Atrial Fibrillation: A Machine Learning Cluster Analysis'. *Lancet (London, England)* 398 (10309): 1427–35. https://doi.org/10.1016/S0140-6736(21)01638-X.

Kasprzyk, Arek, Damian Keefe, Damian Smedley, Darin London, William Spooner, Craig Melsopp, Martin Hammond, Philippe Rocca-Serra, Tony Cox, and Ewan Birney. 2004. 'EnsMart: A Generic System for Fast and Flexible Access to Biological Data'. *Genome Research* 14 (1): 160–69. https://doi.org/10.1101/gr.1645104.

Keats, Bronya J. B., and Stephanie L. Sherman. 2013. 'Chapter 13 - Population Genetics'. In *Emery and Rimoin's Principles and Practice of Medical Genetics (Sixth Edition)*, edited by David Rimoin, Reed Pyeritz, and Bruce Korf, 1–12. Oxford: Academic Press. https://doi.org/10.1016/B978-0-12-383834-6.00015-X.

Keras. 2023. 'Keras Documentation: Why Choose Keras?' 2023. https://keras.io/why_keras/.

Khan, Zanobya N., and Jamil Ahmad. 2021. 'Attention Induced Multi-Head Convolutional Neural Network for Human Activity Recognition'. *Applied Soft Computing* 110 (October): 107671. https://doi.org/10.1016/j.asoc.2021.107671.

Kim, Miran, and Kristin Lauter. 2015. 'Private Genome Analysis through Homomorphic Encryption'. *BMC Medical Informatics and Decision Making* 15 (Suppl 5): S3. https://doi.org/10.1186/1472-6947-15-S5-S3.

Kingma, Diederik P., and Jimmy Ba. 2017. 'Adam: A Method for Stochastic Optimization'. arXiv. https://doi.org/10.48550/arXiv.1412.6980.

Kingma, Diederik P., and Max Welling. 2013. 'Auto-Encoding Variational Bayes'. arXiv. http://arxiv.org/abs/1312.6114.

Klein, Robert J., Caroline Zeiss, Emily Y. Chew, Jen-Yue Tsai, Richard S. Sackler, Chad Haynes, Alice K. Henning, et al. 2005. 'Complement Factor H Polymorphism in Age-Related Macular Degeneration'. *Science* 308 (5720): 385–89. https://doi.org/10.1126/science.1109557.

Klys, Jack, Jake Snell, and Richard Zemel. 2018. 'Learning Latent Subspaces in Variational Autoencoders'. arXiv. https://doi.org/10.48550/arXiv.1812.06190.

Kohonen, Teuvo. 1982. 'Self-Organized Formation of Topologically Correct Feature Maps'. *Biological Cybernetics* 43 (1): 59–69. https://doi.org/10.1007/BF00337288.

Kosko, B. 1990. 'Unsupervised Learning in Noise'. *IEEE Transactions on Neural Networks* 1 (1): 44–57. https://doi.org/10.1109/72.80204.

Kramer, M.A. 1992. 'Autoassociative Neural Networks'. *Computers & Chemical Engineering* 16 (4): 313–28. https://doi.org/10.1016/0098-1354(92)80051-A.

Kullback, S., and R. A. Leibler. 1951. 'On Information and Sufficiency'. *The Annals of Mathematical Statistics* 22 (1): 79–86. https://doi.org/10.1214/aoms/1177729694.

Kurki, Mitja I., Juha Karjalainen, Priit Palta, Timo P. Sipilä, Kati Kristiansson, Kati M. Donner, Mary P. Reeve, et al. 2023. 'FinnGen Provides Genetic Insights from a Well-Phenotyped

Isolated Population'. *Nature* 613 (7944): 508–18. https://doi.org/10.1038/s41586-022-05473-8.
Kurzweil, Ray. 2006. *The Singularity Is near: When Humans Transcend Biology*. New York: Penguin Books. http://books.google.com/books?isbn=0143037889.
Lander, Eric S., Lauren M. Linton, Bruce Birren, Chad Nusbaum, Michael C. Zody, Jennifer Baldwin, Keri Devon, et al. 2001. 'Initial Sequencing and Analysis of the Human Genome'. *Nature* 409 (6822): 860–921. https://doi.org/10.1038/35057062.
Landrum, Melissa J., Jennifer M. Lee, George R. Riley, Wonhee Jang, Wendy S. Rubinstein, Deanna M. Church, and Donna R. Maglott. 2014. 'ClinVar: Public Archive of Relationships among Sequence Variation and Human Phenotype'. *Nucleic Acids Research* 42 (Database issue): D980–85. https://doi.org/10.1093/nar/gkt1113.
Lau, W.W., M. Hardt, Y.H. Zhang, M. Freire, and S. Ruhl. 2021. 'The Human Salivary Proteome Wiki: A Community-Driven Research Platform'. *Journal of Dental Research* 100 (13): 1510–19. https://doi.org/10.1177/00220345211014432.
Lawson, Daniel John, Garrett Hellenthal, Simon Myers, and Daniel Falush. 2012. 'Inference of Population Structure Using Dense Haplotype Data'. *PLOS Genetics* 8 (1): e1002453. https://doi.org/10.1371/journal.pgen.1002453.
Leacock, Stefanie. 2021. *Genetics Textbook*. University of Arkansas at Little Rock. https://bio.libretexts.org/@go/page/25718.
LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. 'Deep Learning'. *Nature* 521 (7553): 436–44. https://doi.org/10.1038/nature14539.
LeCun, Yann, Corinna Cortes, and Chris Burges. 2011. 'MNIST Handwritten Digit Database'. 2011. http://yann.lecun.com/exdb/mnist/.
Levene, P.A. 1919. 'The Structure of Yeast Nucleic Acid'. *Journal of Biological Chemistry* 40 (2): 415–24. https://doi.org/10.1016/S0021-9258(18)87254-4.
Li, Chun, and Mingyao Li. 2008. 'GWAsimulator: A Rapid Whole-Genome Simulation Program'. *Bioinformatics* 24 (1): 140–42. https://doi.org/10.1093/bioinformatics/btm549.
Li, Guo-Min. 2008. 'Mechanisms and Functions of DNA Mismatch Repair'. *Cell Research* 18 (1): 85–98. https://doi.org/10.1038/cr.2007.115.
Li, Mulin Jun, Zipeng Liu, Panwen Wang, Maria P. Wong, Matthew R. Nelson, Jean-Pierre A. Kocher, Meredith Yeager, et al. 2016. 'GWASdb v2: An Update Database for Human Genetic Variants Identified by Genome-Wide Association Studies'. *Nucleic Acids Research* 44 (D1): D869–76. https://doi.org/10.1093/nar/gkv1317.
Li, Yun, Andrea E. Byrnes, and Mingyao Li. 2010. 'To Identify Associations with Rare Variants, Just WHaIT: Weighted Haplotype and Imputation-Based Tests'. *American Journal of Human Genetics* 87 (5): 728–35. https://doi.org/10.1016/j.ajhg.2010.10.014.
Li, Yun, Cristen J. Willer, Jun Ding, Paul Scheet, and Gonçalo R. Abecasis. 2010. 'MaCH: Using Sequence and Genotype Data to Estimate Haplotypes and Unobserved Genotypes'. *Genetic Epidemiology* 34 (8): 816–34. https://doi.org/10.1002/gepi.20533.
Little, Julian, Muin J. Khoury, Linda Bradley, Mindy Clyne, Marta Gwinn, Bruce Lin, Mary-Lou Lindegren, and Paula Yoon. 2003. 'The Human Genome Project Is Complete. How Do We Develop a Handle for the Pump?' *American Journal of Epidemiology* 157 (8): 667–73. https://doi.org/10.1093/aje/kwg048.
Liu, Jiahao, Yutong Chen, Xi Lu, Xiaojing Xu, Gabriella Bulloch, Susan Zhu, Zhuoting Zhu, et

al. 2023. ‘The Association between Dietary Iron Intake and Incidence of Dementia in Adults Aged 60 Years or over in the UK Biobank’. *Nutrients* 15 (2): 260. https://doi.org/10.3390/nu15020260.
López-Ferrando, Víctor, Andrea Gazzo, Xavier de la Cruz, Modesto Orozco, and Josep Ll Gelpí. 2017. ‘PMut: A Web-Based Tool for the Annotation of Pathological Variants on Proteins, 2017 Update’. *Nucleic Acids Research* 45 (W1): W222–28. https://doi.org/10.1093/nar/gkx313.
Lowy-Gallego, Ernesto, Susan Fairley, Xiangqun Zheng-Bradley, Magali Ruffier, Laura Clarke, and Paul Flicek. 2019. ‘Variant Calling on the GRCh38 Assembly with the Data from Phase Three of the 1000 Genomes Project’. *Wellcome Open Research* 4 (December): 50. https://doi.org/10.12688/wellcomeopenres.15126.2.
Lu, Dongsheng, and Shuhua Xu. 2013. ‘Principal Component Analysis Reveals the 1000 Genomes Project Does Not Sufficiently Cover the Human Genetic Diversity in Asia’. *Frontiers in Genetics* 4. https://www.frontiersin.org/articles/10.3389/fgene.2013.00127.
Ma, Shuyi, Cory C. Funk, and Nathan D. Price. 2010. ‘Systems Approaches to Molecular Cancer Diagnostics’. *Discovery Medicine* 10 (55): 531–42.
MacAskill, William. 2022. *What We Owe the Future*. Oneworld.
Maddox, Brenda. 2003. ‘The Double Helix and the “Wronged Heroine”’. *Nature* 421 (6921): 407–8. https://doi.org/10.1038/nature01399.
Mardis, Elaine R. 2008. ‘Next-Generation DNA Sequencing Methods’. *Annual Review of Genomics and Human Genetics* 9 (1): 387–402. https://doi.org/10.1146/annurev.genom.9.081307.164359.
Maxam, A M, and W Gilbert. 1977. ‘A New Method for Sequencing DNA.’ *Proceedings of the National Academy of Sciences of the United States of America* 74 (2): 560–64.
McArthur, Evonne, and John A. Capra. 2021. ‘Topologically Associating Domain Boundaries That Are Stable across Diverse Cell Types Are Evolutionarily Constrained and Enriched for Heritability’. *The American Journal of Human Genetics* 108 (2): 269–83. https://doi.org/10.1016/j.ajhg.2021.01.001.
McCarthy, J, and M. L. Minsky. 1955. ‘A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence’. 1955. http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html.
McCorduck, Pamela. 2018. *Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence*. An A K Peters Book. Boca Raton London New York: CRC Press.
McCulloch, Warren S., and Walter Pitts. 1943. ‘A Logical Calculus of the Ideas Immanent in Nervous Activity’. *The Bulletin of Mathematical Biophysics* 5 (4): 115–33. https://doi.org/10.1007/BF02478259.
Mendel, Gregor. 1865. *Experiments in Plant Hybridization*. http://www.esp.org/foundations/genetics/classical/gm-65.pdf.
Metspalu, Andres. 2002. ‘Estonian Genome Project—before the Take-off and Take-Off’. *Bioinformatics* 18 (suppl_2): S152. https://doi.org/10.1093/bioinformatics/18.suppl_2.S152.
Miljanovic, Milos. 2012. ‘Comparative Analysis of Recurrent and Finite Impulse Response Neural Networks in Time Series Prediction’ 3 (1).

Mitchell, Melanie. 1998. *An Introduction to Genetic Algorithms*. The MIT Press. https://doi.org/10.7551/mitpress/3927.001.0001.
Monckton, Darren G, and Alec J Jeffreys. 1993. 'DNA Profiling'. *Current Opinion in Biotechnology* 4 (6): 660–64. https://doi.org/10.1016/0958-1669(93)90046-Y.
Mondschein, Christopher F., and Cosimo Monda. 2019. 'The EU's General Data Protection Regulation (GDPR) in a Research Context'. In *Fundamentals of Clinical Data Science*, edited by Pieter Kubben, Michel Dumontier, and Andre Dekker. Cham (CH): Springer. http://www.ncbi.nlm.nih.gov/books/NBK543521/.
Morgan, Thomas Hunt. 1915. *The Mechanism of Mendelian Heredity*. H. Holt.
Murphy, Allan H. 1996. 'The Finley Affair: A Signal Event in the History of Forecast Verification'. *Weather and Forecasting* 11 (1): 3–20. https://doi.org/10.1175/1520-0434(1996)011<0003:TFAASE>2.0.CO;2.
Nazer, Lama H., Razan Zatarah, Shai Waldrip, Janny Xue Chen Ke, Mira Moukheiber, Ashish K. Khanna, Rachel S. Hicklen, et al. 2023. 'Bias in Artificial Intelligence Algorithms and Recommendations for Mitigation'. *PLOS Digital Health* 2 (6): e0000278. https://doi.org/10.1371/journal.pdig.0000278.
Newell, A., and H. Simon. 1956. 'The Logic Theory Machine–A Complex Information Processing System'. *IRE Transactions on Information Theory* 2 (3): 61–79. https://doi.org/10.1109/TIT.1956.1056797.
Niemi, Mari E. K., Juha Karjalainen, Rachel G. Liao, Benjamin M. Neale, Mark Daly, Andrea Ganna, Gita A. Pathak, et al. 2021. 'Mapping the Human Genetic Architecture of COVID-19'. *Nature* 600 (7889): 472–77. https://doi.org/10.1038/s41586-021-03767-x.
Nirenberg, Marshall W., and J. Heinrich Matthaei. 1961. 'The Dependence of Cell-Free Protein Synthesis in E. Coli upon Naturally Occurring or Synthetic Polyribonucleotides'. *Proceedings of the National Academy of Sciences* 47 (10): 1588–1602. https://doi.org/10.1073/pnas.47.10.1588.
Novak, Matt. 2023. 'GOP Releases First Ever AI-Created Attack Ad Against President Biden'. Forbes. 25 April 2023. https://www.forbes.com/sites/mattnovak/2023/04/25/gop-releases-first-ever-ai-created-attack-ad-against-president-biden/.
Novroski, Nicole M. M., and Jennifer C. Cihlar. 2022. 'Evolution of Single-Nucleotide Polymorphism Use in Forensic Genetics'. *WIREs Forensic Science* 4 (6): e1459. https://doi.org/10.1002/wfs2.1459.
Nurk, Sergey, Sergey Koren, Arang Rhie, Mikko Rautiainen, Andrey V. Bzikadze, Alla Mikheenko, Mitchell R. Vollger, et al. 2022. 'The Complete Sequence of a Human Genome'. *Science* 376 (6588): 44–53. https://doi.org/10.1126/science.abj6987.
OECD. 2022. 'Measuring the Environmental Impacts of Artificial Intelligence Compute and Applications: The AI Footprint'. OECD Digital Economy Papers 341. Vol. 341. OECD Digital Economy Papers. https://doi.org/10.1787/7babf571-en.
———. 2023. *OECD Economic Outlook, Interim Report September 2023: Confronting Inflation and Low Growth*. Paris: Organisation for Economic Co-operation and Development. https://www.oecd-ilibrary.org/economics/oecd-economic-outlook/volume-2023/issue-1_1f628002-en.
Oleksyk, Taras K., Vladimir Brukhin, and Stephen J. O'Brien. 2015. 'The Genome Russia

Project: Closing the Largest Remaining Omission on the World Genome Map'. *GigaScience* 4. https://doi.org/10.1186/s13742-015-0095-0.
Oord, Aaron van den, Nal Kalchbrenner, Oriol Vinyals, Lasse Espeholt, Alex Graves, and Koray Kavukcuoglu. 2016. 'Conditional Image Generation with PixelCNN Decoders'. arXiv. http://arxiv.org/abs/1606.05328.
Ord, Toby. 2020. *The Precipice*. Hachette Books.
Ovchinnikov, Igor V., and Olga I. Kholina. 2010. 'Genome Digging: Insight into the Mitochondrial Genome of Homo'. *PLOS ONE* 5 (12): e14278. https://doi.org/10.1371/journal.pone.0014278.
Pairo-Castineira, Erola, Sara Clohisey, Lucija Klaric, Andrew D. Bretherick, Konrad Rawlik, Dorota Pasko, Susan Walker, et al. 2021. 'Genetic Mechanisms of Critical Illness in COVID-19'. *Nature* 591 (7848): 92–98. https://doi.org/10.1038/s41586-020-03065-y.
Parker, Carol E., and Christoph H. Borchers. 2014. 'Mass Spectrometry Based Biomarker Discovery, Verification, and Validation – Quality Assurance and Control of Protein Biomarker Assays'. *Molecular Oncology*, Biomarker discovery, 8 (4): 840–58. https://doi.org/10.1016/j.molonc.2014.03.006.
Ponty, Yann, Michel Termier, and Alain Denise. 2006. 'GenRGenS: Software for Generating Random Genomic Sequences and Structures'. *Bioinformatics* 22 (12): 1534–35. https://doi.org/10.1093/bioinformatics/btl113.
Price, Adam, and Cynthia Gibas. 2017. 'Simulome: A Genome Sequence and Variant Simulator'. *Bioinformatics* 33 (12): 1876–78. https://doi.org/10.1093/bioinformatics/btx091.
Ramachandran, Prajit, Barret Zoph, and Quoc V. Le. 2017. 'Searching for Activation Functions'. arXiv. http://arxiv.org/abs/1710.05941.
Ramakrishnan, V. 2002. 'Ribosome Structure and the Mechanism of Translation'. *Cell* 108 (4): 557–72. https://doi.org/10.1016/s0092-8674(02)00619-0.
Rambaut, Andrew. 2002. 'Seq-Gen'. 2002. http://tree.bio.ed.ac.uk/software/seqgen/.
Ramirez, Vanessa Bates. 2018. 'Designer Babies, and Their Babies: How AI and Genomics Will Impact Reproduction'. *Singularity Hub* (blog). 14 November 2018. https://singularityhub.com/2018/11/14/designer-babies-and-their-babies-where-ai-and-genomics-could-take-us/.
Ramos-Onsins, Sebastian, and Thomas Mitchell-Olds. 2008. 'Mlcoalsim'. 2008. http://www.ub.edu/softevol/mlcoalsim/.
Rashid, Sabrina, Sohrab Shah, Ziv Bar-Joseph, and Ravi Pandya. 2021. 'Dhaka: Variational Autoencoder for Unmasking Tumor Heterogeneity from Single Cell Genomic Data'. *Bioinformatics* 37 (11): 1535–43. https://doi.org/10.1093/bioinformatics/btz095.
Ren, Jie, Nathan A. Ahlgren, Yang Young Lu, Jed A. Fuhrman, and Fengzhu Sun. 2017. 'VirFinder: A Novel k-Mer Based Tool for Identifying Viral Sequences from Assembled Metagenomic Data'. *Microbiome* 5 (July): 69. https://doi.org/10.1186/s40168-017-0283-5.
Richard, Guy-Franck, Alix Kerrest, and Bernard Dujon. 2008. 'Comparative Genomics and Molecular Dynamics of DNA Repeats in Eukaryotes'. *Microbiology and Molecular Biology Reviews : MMBR* 72 (4): 686–727. https://doi.org/10.1128/MMBR.00011-08.
Richards, Adair D., and Alison Rodger. 2007. 'Synthetic Metallomolecules as Agents for the Control of DNA Structure'. *ChemInform* 38 (25): chin.200725222.

https://doi.org/10.1002/chin.200725222.
Ring, Huijun Z, Pui-Yan Kwok, and Richard GH Cotton. 2006. 'Human Variome Project: An International Collaboration to Catalogue Human Genetic Variation'. *Pharmacogenomics* 7 (7): 969–72. https://doi.org/10.2217/14622416.7.7.969.
Robinson, Peter N. 2012. 'Deep Phenotyping for Precision Medicine'. *Human Mutation* 33 (5): 777–80. https://doi.org/10.1002/humu.22080.
Rogin, Ali, and Andrew Corkery. 2023. 'Why Artificial Intelligence Is a Central Dispute in the Hollywood Strikes'. PBS NewsHour. 2 September 2023. https://www.pbs.org/newshour/show/why-artificial-intelligence-is-a-central-dispute-in-the-hollywood-strikes.
Rombach, Robin, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. 2022. 'High-Resolution Image Synthesis with Latent Diffusion Models'. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10674–85. New Orleans, LA, USA: IEEE. https://doi.org/10.1109/CVPR52688.2022.01042.
Rosenblatt, F. 1958. 'The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain'. *Psychological Review* 65 (6): 386–408. https://doi.org/10.1037/h0042519.
Sahlin, Kristoffer, and Paul Medvedev. 2021. 'Error Correction Enables Use of Oxford Nanopore Technology for Reference-Free Transcriptome Analysis'. *Nature Communications* 12 (1): 2. https://doi.org/10.1038/s41467-020-20340-8.
Sanger, F., S. Nicklen, and A. R. Coulson. 1977. 'DNA Sequencing with Chain-Terminating Inhibitors'. *Proceedings of the National Academy of Sciences of the United States of America* 74 (12): 5463–67.
Schneble, Christophe Olivier, Bernice Simone Elger, and David Shaw. 2018. 'The Cambridge Analytica Affair and Internet‐mediated Research'. *EMBO Reports* 19 (8): e46579. https://doi.org/10.15252/embr.201846579.
Searle, John R. 1980. 'Minds, Brains, and Programs'. *Behavioral and Brain Sciences* 3 (3): 417–24. https://doi.org/10.1017/S0140525X00005756.
Semeraro, Roberto, Valerio Orlandini, and Alberto Magi. 2018. 'Xome-Blender: A Novel Cancer Genome Simulator'. *PLOS ONE* 13 (4): e0194472. https://doi.org/10.1371/journal.pone.0194472.
Shabani, Mahsa, and Luca Marelli. 2019. 'Re-Identifiability of Genomic Data and the GDPR'. *EMBO Reports* 20 (6): e48316. https://doi.org/10.15252/embr.201948316.
Sherrington, David, and Scott Kirkpatrick. 1975. 'Solvable Model of a Spin-Glass'. OpenAIRE - Explore. 1975. https://explore.openaire.eu/search/publication?pid=10.1103%2Fphysrevlett.35.1792.
Sherry, S. T., M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin. 2001. 'dbSNP: The NCBI Database of Genetic Variation'. *Nucleic Acids Research* 29 (1): 308–11. https://doi.org/10.1093/nar/29.1.308.
Sherry, Stephen T., Minghong Ward, and Karl Sirotkin. 1999. 'dbSNP—Database for Single Nucleotide Polymorphisms and Other Classes of Minor Genetic Variation'. *Genome Research* 9 (8): 677–79. https://doi.org/10.1101/gr.9.8.677.
Silver, David, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, et al. 2017. 'Mastering Chess and Shogi by Self-Play with a General

Reinforcement Learning Algorithm'. arXiv. https://doi.org/10.48550/arXiv.1712.01815.
Simón-Sánchez, Javier, Claudia Schulte, Jose M. Bras, Manu Sharma, J. Raphael Gibbs, Daniela Berg, Coro Paisan-Ruiz, et al. 2009. 'Genome-Wide Association Study Reveals Genetic Risk Underlying Parkinson's Disease'. *Nature Genetics* 41 (12): 1308–12. https://doi.org/10.1038/ng.487.
Singh, Tarjinder, Timothy Poterba, David Curtis, Huda Akil, Mariam Al Eissa, Jack D. Barchas, Nicholas Bass, et al. 2022. 'Rare Coding Variants in Ten Genes Confer Substantial Risk for Schizophrenia'. *Nature* 604 (7906): 509–16. https://doi.org/10.1038/s41586-022-04556-w.
Smith, H. O., J. F. Tomb, B. A. Dougherty, R. D. Fleischmann, and J. C. Venter. 1995. 'Frequency and Distribution of DNA Uptake Signal Sequences in the Haemophilus Influenzae Rd Genome'. *Science (New York, N.Y.)* 269 (5223): 538–40. https://doi.org/10.1126/science.7542802.
Soussi, Thierry, Bernard Leroy, and Peter E.M. Taschner. 2014. 'Recommendations for Analyzing and Reporting TP53 Gene Variants in the High-Throughput Sequencing Era'. *Human Mutation* 35 (6): 766–78. https://doi.org/10.1002/humu.22561.
Stevens, N. M. 1905. 'A Study of the Germ Cells of Aphis Rosæ and Aphis Œnotheræ'. *Journal of Experimental Zoology* 2 (3): 313–33. https://doi.org/10.1002/jez.1400020302.
Sudlow, Cathie, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, et al. 2015. 'UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age'. *PLoS Medicine* 12 (3): e1001779. https://doi.org/10.1371/journal.pmed.1001779.
Sutton, Walter S. 1902. 'On the Morphology of the Chromoso Group in Brachystola Magna'. *The Biological Bulletin* 4 (1): 24–39. https://doi.org/10.2307/1535510.
Svoboda, P., and A. Di Cara. 2006. 'Hairpin RNA: A Secondary Structure of Primary Importance'. *Cellular and Molecular Life Sciences: CMLS* 63 (7–8): 901–8. https://doi.org/10.1007/s00018-005-5558-5.
Tahmasbi, Rasool, and Matthew Keller. 2016. 'GeneEvolve: A Fast and Memory Efficient Forward-Time Simulator of Realistic Whole-Genome Sequence and SNP Data'. *Bioinformatics (Oxford, England)* 33 (September). https://doi.org/10.1093/bioinformatics/btw606.
Taitingfong, Riley, Cinnamon S Bloss, Cynthia Triplett, Julie Cakici, Nanibaa' Garrison, Shelley Cole, Julie A Stoner, and Lucila Ohno-Machado. 2020. 'A Systematic Literature Review of Native American and Pacific Islanders' Perspectives on Health Data Privacy in the United States'. *Journal of the American Medical Informatics Association : JAMIA* 27 (12): 1987–98. https://doi.org/10.1093/jamia/ocaa235.
Tange, Ole. 2011. 'GNU Parallel: The Command-Line Power Tool'. *Login*.
TensorFlow. 2023. 'Convolutional Variational Autoencoder'. TensorFlow. 27 July 2023. https://www.tensorflow.org/tutorials/generative/cvae.
The 1000 Genomes Project Consortium. 2015. 'A Global Reference for Human Genetic Variation'. *Nature* 526 (7571): 68–74. https://doi.org/10.1038/nature15393.
'The "All of Us" Research Program'. 2019. *The New England Journal of Medicine* 381 (7): 668–76. https://doi.org/10.1056/NEJMsr1809937.
Time Magazine, (No author specified). 1961. 'Science: The Goof Button'. *Time*, 18 August 1961.

https://content.time.com/time/subscriber/article/0,33009,872691,00.html.
Tsallis, Constantino, and Daniel A. Stariolo. 1996. ‘Generalized Simulated Annealing’. *Physica A: Statistical Mechanics and Its Applications* 233 (1–2): 395–406. https://doi.org/10.1016/S0378-4371(96)00271-3.
user3658307. 2018. ‘Answer to “Deriving the KL Divergence Loss for VAEs”’. *Cross Validated*. https://stats.stackexchange.com/a/370048/262016.
Varsha, PS. 2023. ‘How Can We Manage Biases in Artificial Intelligence Systems – A Systematic Literature Review’. *International Journal of Information Management Data Insights* 3 (1): 100165. https://doi.org/10.1016/j.jjimei.2023.100165.
Virolainen, Samuel J., Andrew VonHandorf, Kenyatta C. M. F. Viel, Matthew T. Weirauch, and Leah C. Kottyan. 2023. ‘Gene–Environment Interactions and Their Impact on Human Health’. *Genes and Immunity* 24 (1): 1–11. https://doi.org/10.1038/s41435-022-00192-6.
Visscher, Peter M., Loic Yengo, Nancy J. Cox, and Naomi R. Wray. 2021. ‘Discovery and Implications of Polygenicity of Common Disease’. *Science (New York, N.Y.)* 373 (6562): 1468–73. https://doi.org/10.1126/science.abi8206.
Waggener, Bill. 1995. *Pulse Code Modulation Techniques*. Springer Science & Business Media.
Wang, Dongfang, and Jin Gu. 2018. ‘VASC: Dimension Reduction and Visualization of Single-Cell RNA-Seq Data by Deep Variational Autoencoder’. *Genomics, Proteomics & Bioinformatics* 16 (5): 320–31. https://doi.org/10.1016/j.gpb.2018.08.003.
Watson, J. D., and F. H. C. Crick. 1953. ‘Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid’. *Nature* 171 (4356): 737–38. https://doi.org/10.1038/171737a0.
Whitehead, Alfred North, and Bertrand Russell. 1913. *Principia Mathematica*. Cambridge University Press.
Wong, Carissa. 2023. ‘UK First to Approve CRISPR Treatment for Diseases: What You Need to Know’. *Nature* 623 (7988): 676–77. https://doi.org/10.1038/d41586-023-03590-6.
Wu, Ray. 1970. ‘Nucleotide Sequence Analysis of DNA: I. Partial Sequence of the Cohesive Ends of Bacteriophage λ and 186 DNA’. *Journal of Molecular Biology* 51 (3): 501–21. https://doi.org/10.1016/0022-2836(70)90004-5.
Wu, Xindong, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, et al. 2008. ‘Top 10 Algorithms in Data Mining’. *Knowledge and Information Systems* 14 (1): 1–37. https://doi.org/10.1007/s10115-007-0114-2.
Xia, Li Charlie, Dongmei Ai, Hojoon Lee, Noemi Andor, Chao Li, Nancy R. Zhang, and Hanlee P. Ji. 2018. ‘SVEngine: An Efficient and Versatile Simulator of Genome Structural Variations with Features of Cancer Clonal Evolution’. *GigaScience* 7 (7): giy081. https://doi.org/10.1093/gigascience/giy081.
Xu, Junlin, Jielin Xu, Yajie Meng, Changcheng Lu, Lijun Cai, Xiangxiang Zeng, Ruth Nussinov, and Feixiong Cheng. 2023. ‘Graph Embedding and Gaussian Mixture Variational Autoencoder Network for End-to-End Analysis of Single-Cell RNA Sequencing Data’. *Cell Reports Methods* 3 (1): 100382. https://doi.org/10.1016/j.crmeth.2022.100382.
Xu, Lei, Chunxiao Jiang, Jian Wang, Jian Yuan, and Yong Ren. 2014. ‘Information Security in Big Data: Privacy and Data Mining’. *IEEE Access* 2: 1149–76. https://doi.org/10.1109/ACCESS.2014.2362522.
Xu, Liu, and Masahide Seki. 2020. ‘Recent Advances in the Detection of Base Modifications

Using the Nanopore Sequencer'. *Journal of Human Genetics* 65 (1): 25–33. https://doi.org/10.1038/s10038-019-0679-0.
Yang, Z. 1997. 'PAML: A Program Package for Phylogenetic Analysis by Maximum Likelihood'. *Computer Applications in the Biosciences: CABIOS* 13 (5): 555–56. https://doi.org/10.1093/bioinformatics/13.5.555.
Yelmen, Burak, Aurélien Decelle, Leila Lea Boulos, Antoine Szatkownik, Cyril Furtlehner, Guillaume Charpiat, and Flora Jay. 2023. 'Deep Convolutional and Conditional Neural Networks for Large-Scale Genomic Data Generation'. Preprint. Genomics. https://doi.org/10.1101/2023.03.07.530442.
Yelmen, Burak, Aurélien Decelle, Linda Ongaro, Davide Marnetto, Corentin Tallec, Francesco Montinaro, Cyril Furtlehner, Luca Pagani, and Flora Jay. 2021. 'Creating Artificial Human Genomes Using Generative Neural Networks'. *PLOS Genetics* 17 (2): e1009303. https://doi.org/10.1371/journal.pgen.1009303.
Yuan, Hsiang-Yu, Jen-Jie Chiou, Wen-Hsien Tseng, Chia-Hung Liu, Chuan-Kun Liu, Yi-Jung Lin, Hui-Hung Wang, Adam Yao, Yuan-Tsong Chen, and Chun-Nan Hsu. 2006. 'FASTSNP: An Always up-to-Date and Extendable Service for SNP Function Analysis and Prioritization'. *Nucleic Acids Research* 34 (suppl_2): W635–41. https://doi.org/10.1093/nar/gkl236.
Yue, Jia-Xing, and Gianni Liti. 2019. 'simuG: A General-Purpose Genome Simulator'. *Bioinformatics* 35 (21): 4442–44. https://doi.org/10.1093/bioinformatics/btz424.
Zawistowski, Matthew, Shyam Gopalakrishnan, Jun Ding, Yun Li, Sara Grimm, and Sebastian Zöllner. 2010. 'Extending Rare-Variant Testing Strategies: Analysis of Noncoding Sequence and Imputed Genotypes'. *American Journal of Human Genetics* 87 (5): 604–17. https://doi.org/10.1016/j.ajhg.2010.10.012.
Zhao, Yongan, Xiaofeng Wang, and Haixu Tang. 2018. 'A Secure Alignment Algorithm for Mapping Short Reads to Human Genome'. *Journal of Computational Biology* 25 (6): 529–40. https://doi.org/10.1089/cmb.2017.0094.
Zhu, Michael, and Suyog Gupta. 2017. 'To Prune, or Not to Prune: Exploring the Efficacy of Pruning for Model Compression'. arXiv. http://arxiv.org/abs/1710.01878.
Zhu, Qianli Scott. 2023. 'Keras Release 2.14.0 · Keras-Team/Keras'. GitHub. 12 September 2023. https://github.com/keras-team/keras/releases/tag/v2.14.0.
Ziegenhain, Christoph, and Rickard Sandberg. 2021. 'BAMboozle Removes Genetic Variation from Human Sequence Data for Open Data Sharing'. *Nature Communications* 12 (1): 6216. https://doi.org/10.1038/s41467-021-26152-8.