A3 2023-2024

Visualisation de données génomiques

Mouhamed SY

61958

Professeur : Dr. Jérémie Sublime

Table des matières

Introduction

L'analyse des données génomiques anciennes constitue une méthode précieuse pour mieux comprendre les migrations humaines qui ont façonné l'histoire. En particulier, l'étude des polymorphismes nucléotidiques (SNPs) issus de l'ADN ancien permet d'identifier des marqueurs génétiques spécifiques qui révèlent les mouvements de populations à travers le temps. Le chromosome Y, qui se transmet uniquement de père en fils sans recombinaison, joue un rôle central dans cette analyse car il permet de suivre les lignées patrilinéaires sur plusieurs générations. Les haplogroupes, qui regroupent des séries d'allèles caractéristiques d'un chromosome, offrent ainsi des informations précieuses sur l'origine géographique des populations et les grandes vagues de migration masculine. Ce projet s'attache à visualiser ces données génétiques issues du chromosome Y afin de mieux comprendre les déplacements de populations anciennes et d'établir des corrélations entre mutations génétiques et régions géographiques. Pour atteindre cet objectif, nous utiliserons des techniques de réduction de dimensionnalité, comme le Multi-Dimensional Scaling (MDS), qui permettent de projeter les données complexes du chromosome Y dans un espace en deux ou trois dimensions. Cela facilitera la représentation des similarités et des dissimilarités entre individus, et mettra en évidence les relations entre les mutations SNP et les zones géographiques d'origine. Il vise ainsi à fournir une visualisation claire et intuitive des mouvements de populations masculines à travers les âges, en s'appuyant sur les données génétiques disponibles et les haplogroupes associés, et en permettant d'obtenir une vue d'ensemble des migrations humaines anciennes dans un contexte scientifique rigoureux.

I. Recherches bibliographiques sur les migrations et les haplogroupes

1. Mouvements migratoires depuis la préhistoire

Depuis des millénaires, les migrations humaines ont façonné les populations et la diversité génétique que nous observons aujourd'hui. Parmi les plus anciens et importants déplacements figure la migration hors d'Afrique, il y a environ 60 000 à 80 000 ans. C'est à ce moment-là qu'Homo sapiens a commencé à peupler d'autres continents, une étape fondamentale dans l'histoire humaine souvent désignée sous le nom de "**Out of Africa**". Ce mouvement massif a permis aux premiers humains modernes de s'établir en Eurasie, en rencontrant des environnements diversifiés auxquels ils se sont adaptés, ce qui a contribué à la richesse génétique des populations actuelles. Les traces de cette migration sont encore visibles dans les marqueurs génétiques, tels que les haplogroupes du chromosome Y et ceux de l'ADN mitochondrial.

Plus tard, il y a environ 5 000 ans, les invasions indo-européennes ont marqué l'Europe et l'Asie. Ces peuples, originaires des steppes eurasiennes, ont diffusé non seulement leurs langues mais aussi leurs gènes, bouleversant les structures sociales et génétiques des régions où ils se sont installés. Ce mouvement est intimement lié à la propagation de l'haplogroupe **R1b** en Europe occidentale et **R1a** en Europe de l'Est et en Asie du Sud.

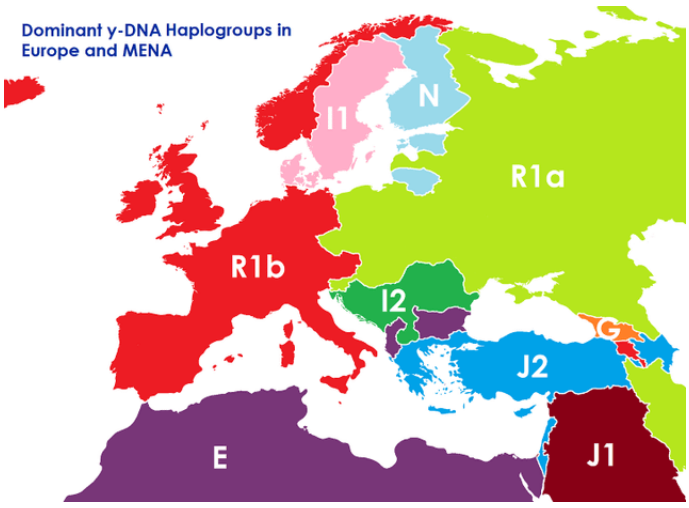

Figure 1 : *Répartition des Haplogroupes du Chromosome Y Dominants en Europe et dans la région MENA (Moyen-Orient et Afrique du Nord)*

Un autre exemple de migration ayant laissé une empreinte génétique est celle des Vikings, qui ont parcouru l'Europe du Nord et de l'Ouest entre le VIIIe et le XIe siècle. En parallèle, les vastes expansions des Mongols sous l'égide de Gengis Khan ont également contribué à la diversité génétique des populations d'Asie et d'Europe. Ces mouvements ont entraîné des échanges génétiques majeurs, modifiant profondément les lignées génomiques dans les régions concernées.

Ces migrations ont joué un rôle déterminant dans la structure génétique des populations actuelles. Elles ont provoqué des mélanges, des adaptations, et sont à l'origine de la grande diversité génétique observable aujourd'hui. Les marqueurs génétiques permettent de retracer ces événements à grande échelle, notamment grâce à l'étude des haplogroupes du chromosome Y et de l'ADN mitochondrial.

2. Haplogroupes associés aux migrations

Les haplogroupes du chromosome Y, en raison de leur transmission exclusivement patrilinéaire, sont des outils précieux pour suivre les migrations masculines à travers l'histoire. Parmi les haplogroupes les plus étudiés, on trouve **R1b**, dominant en Europe de l'Ouest, particulièrement en Irlande, au Pays de Galles et en Espagne. Cet haplogroupe est étroitement lié à l'expansion des populations indo-européennes. Son origine remonte

aux steppes eurasiennes et son expansion est souvent associée aux premières migrations agricoles en Europe. De l'autre côté, l'haplogroupe **R1a** est plus fréquent en Europe de l'Est et en Asie du Sud, illustrant ainsi une diffusion géographique différente mais également liée aux Indo-Européens.

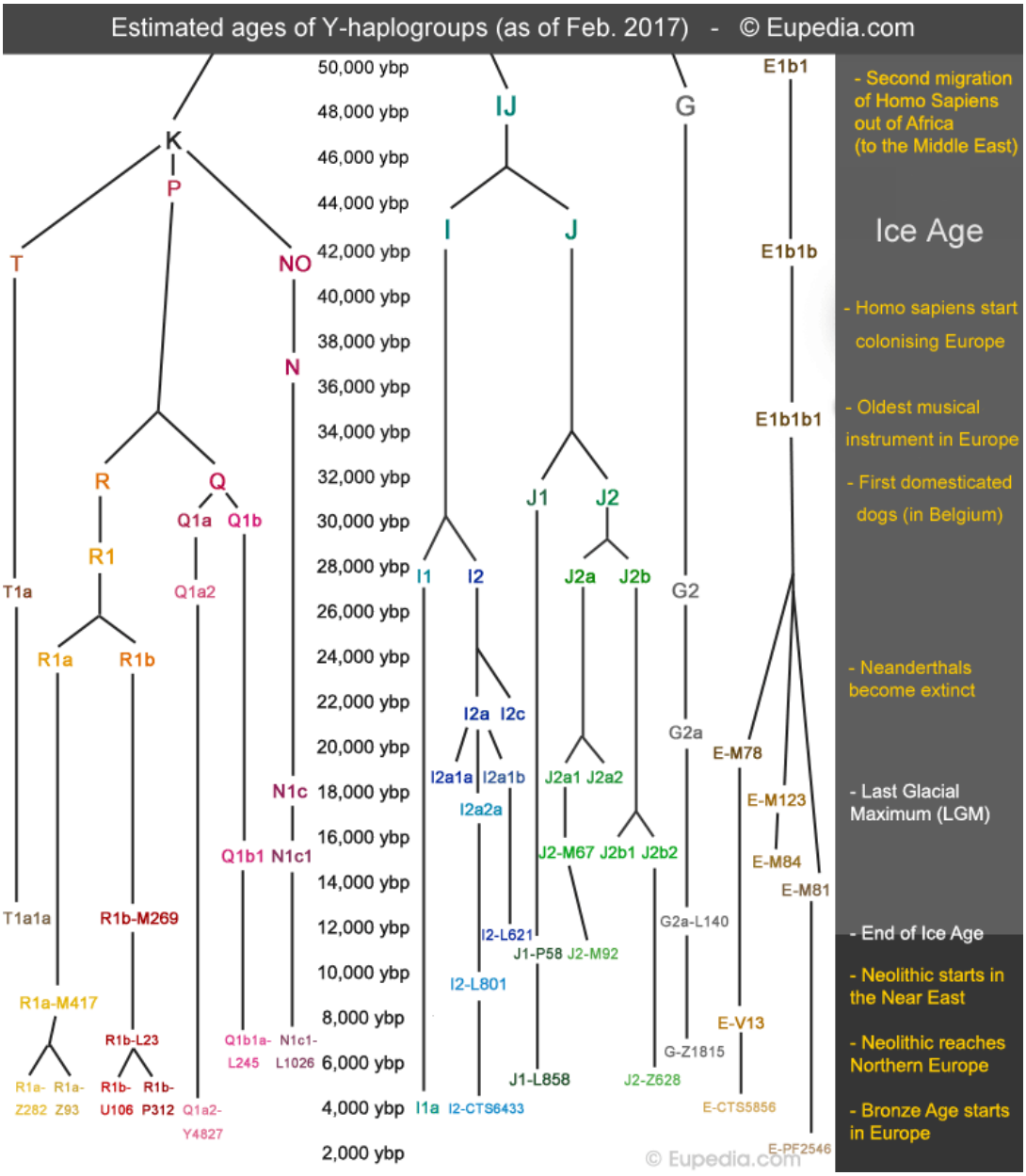

Figure 2 : *Développement chronologique des principaux haplogroupes Y-ADN d'Eurasie occidentale depuis le Paléolithique tardif jusqu'à l'âge du fer*

Un autre haplogroupe important est **E1b1b**, largement répandu dans le bassin méditerranéen. Cet haplogroupe, fréquent en Afrique du Nord et en Europe du Sud, est souvent associé aux migrations néolithiques qui ont accompagné la diffusion de l'agriculture depuis le Proche-Orient vers l'Europe.

L'analyse des haplogroupes de l'ADN mitochondrial permet, quant à elle, de retracer les lignées matrilinéaires. Bien que ces données ne soient pas disponibles dans ce projet, elles offrent néanmoins des informations cruciales pour comprendre les migrations féminines. Par exemple, l'haplogroupe **H** est présent chez environ 40 % des Européens et est associé à la première colonisation de l'Europe par Homo sapiens. De même, en Asie et dans les Amériques, des haplogroupes comme **B** et **C** révèlent des migrations de longue date, notamment à travers le détroit de Béring il y a plus de 15 000 ans.

Les haplogroupes mitochondriaux sont également utiles pour étudier des migrations spécifiques, comme celles dans le Pacifique. Par exemple, l'haplogroupe **B4a** est associé aux premières populations polynésiennes, témoignant de leur migration vers les îles du Pacifique, la dernière grande expansion humaine.

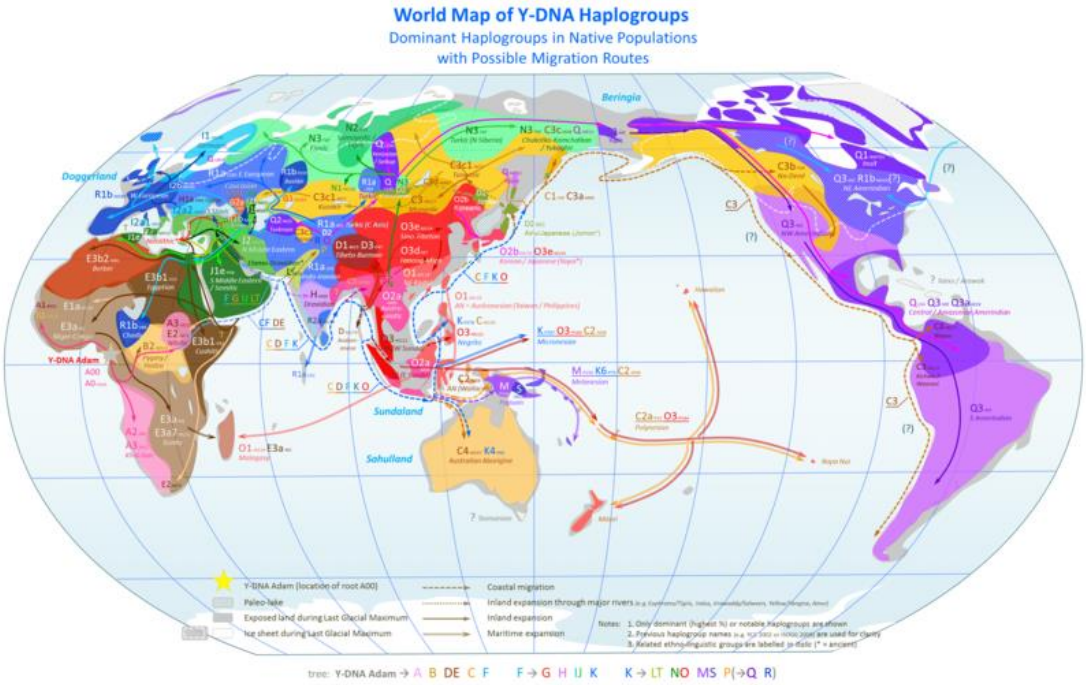

Figure 3 : *Carte de la répartition des principaux haplogroupes Y en fonction des régions géographiques*

II. Données et méthodologie

1. Description des données

Les données génomiques utilisées dans ce projet proviennent de la base de données **Allen Ancient DNA database** (version 54, mars 2023), qui recense 20 503 individus, chacun étant décrit par 597 573 polymorphismes nucléotidiques simples (SNPs). Ce vaste ensemble de données permet d'explorer les trajectoires migratoires des populations humaines anciennes à travers leurs marqueurs génétiques. Cependant, ces individus incluent parfois des doublons, ce qui nécessite une manipulation attentive des données.

Chaque SNP peut prendre l'une des quatre valeurs suivantes : **G**, **A**, **T**, ou **C**. En génomique, pour rendre ces données analysables, un encodage numérique est souvent utilisé. Selon le nucléotide de référence, les SNPs sont traduits de la manière suivante :

2 : les deux allèles correspondent au nucléotide de référence (par exemple, "GG").

1 : un seul allèle correspond (par exemple, "GA").

0 : aucun des deux allèles ne correspond (par exemple, "TT").

3 ou **9** : ces valeurs représentent des données manquantes.

Cet encodage permet de simplifier l'intégration des données génétiques dans des modèles statistiques. Toutefois, plusieurs limites doivent être prises en compte. Premièrement, certaines données peuvent être incomplètes en raison de la méthode de séquençage utilisée ou de l'ancienneté des génomes étudiés, surtout pour des échantillons extrêmement anciens. De plus, bien que le **chromosome Y** soit un outil puissant pour suivre les lignées patrilinéaires, il ne permet d'examiner que la moitié de l'histoire génétique d'un individu, laissant de côté les lignées matrilinéaires qui ne sont pas capturées par ces données.

Un autre point important est l'absence d'informations sur l'**ADN mitochondrial** dans le jeu de données. Cet ADN, transmis uniquement par la lignée maternelle, aurait permis de compléter l'analyse des migrations féminines et d'obtenir une vue d'ensemble plus équilibrée des mouvements de population.

2. Méthodologie

Pour analyser et visualiser les données du chromosome Y, nous avons choisi d'utiliser des techniques de réduction de dimensionnalité. Les données initiales comportant près de 600 000 dimensions (les SNPs), il est impossible de les visualiser directement en deux ou trois dimensions. Nous avons donc opté pour l'algorithme de **Multi-Dimensional Scaling (MDS)**, en nous appuyant sur la méthode **SMACOF** (Scaling by Majorizing a Complicated Function).

Le MDS est une méthode qui permet de projeter des données d'un espace de haute dimension (dans ce cas, les SNPs du chromosome Y) vers un espace de plus faible dimension, tout en conservant autant que possible les distances relatives entre les points. L'objectif est de minimiser la différence entre les distances dans l'espace d'origine et celles dans l'espace projeté. Pour ce faire, une **matrice de dissimilarité** est construite,

permettant de mesurer les relations entre chaque individu selon leurs mutations génétiques.

La méthode **SMACOF** optimise cette projection en minimisant de façon itérative une fonction de **stress**, qui représente l'écart entre les distances dans l'espace initial et celles obtenues dans l'espace réduit. Cette optimisation garantit une meilleure correspondance entre les projections et les distances réelles des données originales, avec chaque itération réduisant progressivement la valeur du stress jusqu'à atteindre une convergence satisfaisante.

Les principales étapes de cette méthodologie incluent :

- Construction de la matrice de dissimilarité : Les distances entre individus sont calculées à partir des SNPs encodés pour construire une matrice reflétant leurs similarités génétiques.

- Implémentation de l'algorithme SMACOF : La fonction de stress est minimisée itérativement, ajustant à chaque étape la projection des données dans un espace à faible dimension (2D ou 3D).

- Visualisation finale : Les individus sont représentés dans cet espace réduit, permettant d'identifier des regroupements selon leurs haplogroupes ou leurs régions géographiques.

Cette approche permet de transformer les données complexes en représentations visuelles compréhensibles, tout en révélant les similarités génétiques entre les individus et les populations à travers le temps.

III. Visualisation des données génomiques

1. Méthodes de réduction de dimensionnalité

La réduction de dimensionnalité est un processus clé pour rendre les données génomiques plus accessibles et interprétables. En génomique, les données, notamment les SNPs, sont souvent de très haute dimension, rendant leur analyse et leur visualisation

difficile. Dans le cadre de ce projet, les données génétiques issues du chromosome Y comprennent des centaines de milliers de SNPs par individu, soit autant de dimensions. Une visualisation directe dans cet espace n'est pas possible, ce qui rend nécessaire l'utilisation de techniques permettant de réduire cette complexité tout en conservant le maximum d'information. L'une des méthodes les plus couramment utilisées pour réduire la dimensionnalité est le Multi-Dimensional Scaling (MDS). Le MDS vise à trouver une représentation des individus dans un espace de plus faible dimension, en préservant autant que possible les distances (ou dissimilarités) initiales entre les individus. Cette méthode est particulièrement utile pour comprendre les relations entre individus dans des jeux de données complexes, tels que ceux comportant des SNPs. Dans le cadre de ce projet, nous utilisons une variante de MDS appelée SMACOF (Scaling by Majorizing a Complicated Function). SMACOF est une méthode itérative qui minimise le stress, une fonction mesurant la différence entre les dissimilarités dans l'espace d'origine et les distances dans l'espace projeté.

2. Implémentation et analyse de SMACOF

Après avoir appliqué l'algorithme SMACOF à notre matrice de dissimilarité, nous obtenons une projection en deux dimensions qui révèle plusieurs informations importantes sur les relations génétiques entre les individus. Voici les principaux résultats :

La projection met en évidence des regroupements d'individus présentant une forte similarité génétique, indiquée par des dissimilarités faibles dans la matrice de départ. Ces regroupements pourraient correspondre à des populations ayant des origines géographiques ou des haplogroupes communs.

En revanche, certains individus se distinguent clairement des autres, ce qui pourrait refléter une divergence génétique significative ou des origines géographiques distinctes.

Le choix d'utiliser le **RMSD (Root Mean Square Deviation)** comme indicateur de qualité de la projection repose sur sa capacité à quantifier l'écart entre les dissimilarités d'origine (dans l'espace génétique à haute dimension) et les distances après projection (dans l'espace réduit). Le **stress normalisé**, qui est couramment utilisé pour évaluer la qualité des projections MDS, ne permet pas de capturer la précision au niveau individuel. Le

RMSD, en revanche, offre un indicateur plus intuitif et granulaire, notamment pour comprendre comment chaque individu est représenté dans l'espace projeté.

Le RMSD est calculé de la manière suivante :

On compare les dissimilarités originales entre les individus (issues de la matrice de dissimilarité) avec les distances calculées dans l'espace projeté.

Un RMSD global est obtenu, reflétant la qualité globale de la projection. Plus le RMSD est faible, mieux les distances originales sont préservées.

Deux projections ont été réalisées : une en deux dimensions et une en trois dimensions. La projection en 2D, illustrée dans la première figure, présente un RMSD global de **4.7948**, qui évalue la qualité de la conservation des distances dans l'espace réduit. Les points sont colorés en fonction de leur RMSD individuel, une mesure qui indique la qualité de la projection pour chaque individu. Les points bleus signalent une bonne qualité de projection, où les distances dans l'espace projeté sont relativement proches des distances d'origine. En revanche, les points rouges et oranges indiquent une distorsion plus importante, ce qui signifie que la projection en 2D n'a pas réussi à bien représenter les distances d'origine pour ces individus. Cela est particulièrement visible pour certains individus situés aux extrêmes du graphique, en haut à droite et en bas, qui semblent être mal représentés en raison de leurs dissimilarités génétiques significatives, difficiles à capturer dans un espace de seulement deux dimensions.

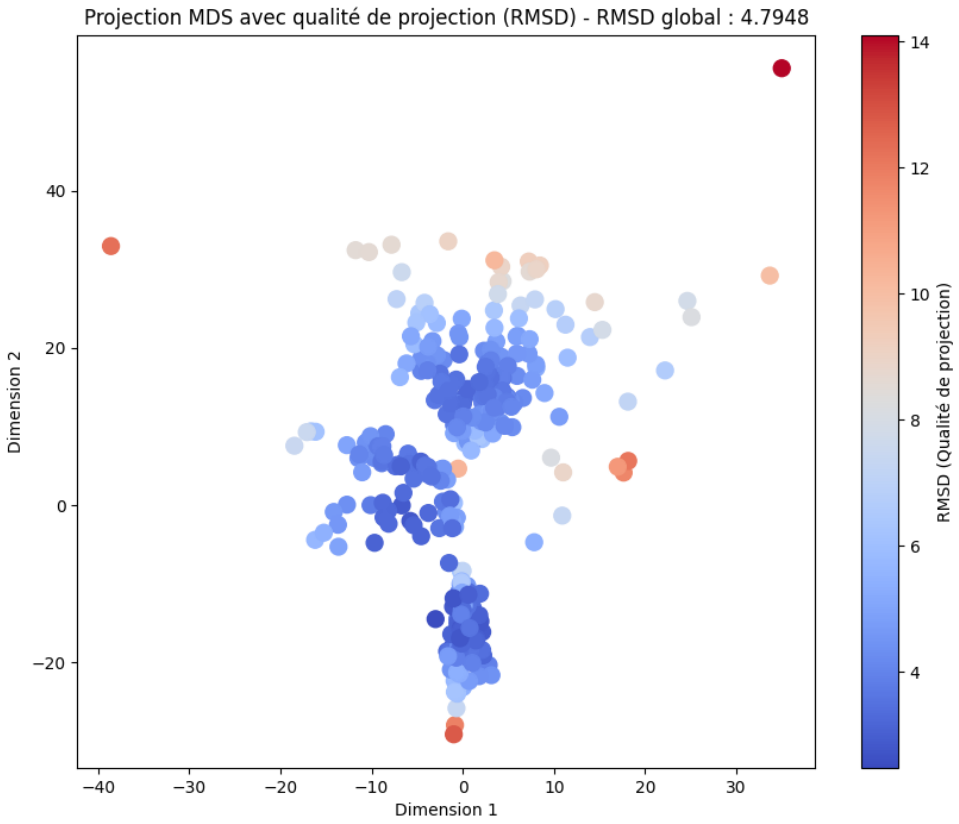

Figure 4 : *Projection MDS en 2 dimensions avec un RMSD global de 4.7948*

Pour améliorer la qualité de cette représentation, une seconde projection en trois dimensions a été réalisée, montrée dans la deuxième figure. L'ajout d'une troisième dimension a permis de réduire le RMSD global à **4.4309**, indiquant que la projection en 3D conserve mieux les distances génétiques que celle en 2D. Bien que la majorité des individus soient toujours bien représentés (comme le montrent les points majoritairement bleus), le nombre d'individus mal projetés, représentés en rouge ou orange, a diminué par rapport à la projection en 2D. Cela signifie que la structure des relations de similarité génétique est mieux préservée dans un espace en trois dimensions. Les individus génétiquement distincts qui étaient mal projetés en 2D bénéficient ainsi d'une représentation plus fidèle en 3D, même si quelques-uns continuent de présenter une distorsion notable.

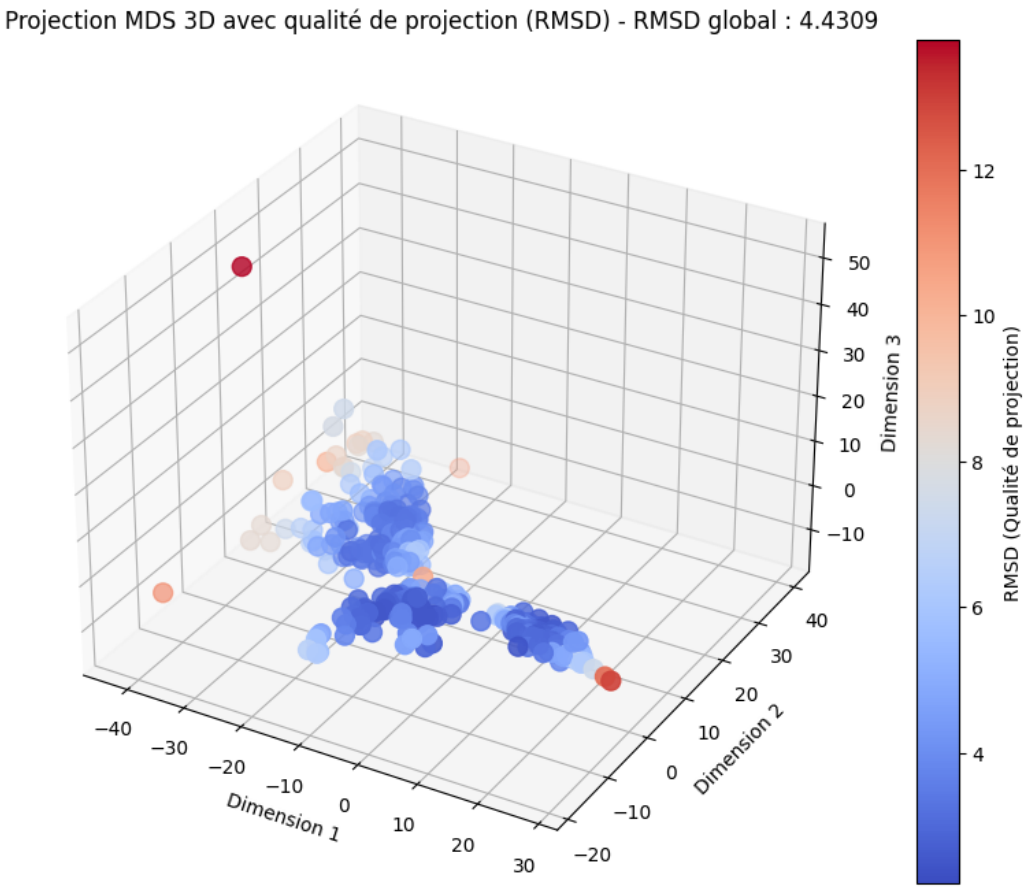

Figure 5 : *Projection MDS en 3 dimensions avec un RMSD global de 4.4309*

L'utilisation du RMSD pour évaluer la qualité des projections a été particulièrement utile dans ce contexte. Contrairement au stress normalisé qui fournit une mesure globale de la qualité de la projection, le RMSD permet d'analyser la précision pour chaque individu, offrant ainsi une évaluation plus fine des points qui sont mal représentés dans l'espace projeté. Cela est essentiel dans l'analyse des données génétiques complexes, où certains individus peuvent présenter des relations dissimilaires difficiles à projeter correctement dans des espaces de faible dimension. En comparant les deux projections, l'ajout d'une troisième dimension a clairement amélioré la qualité de la représentation globale, comme en témoigne la diminution du RMSD. Cependant, certains individus demeurent mal représentés même en 3D, suggérant que leur structure génétique nécessite peut-être plus de dimensions pour être correctement capturée. Ces résultats soulignent l'importance d'adapter le nombre de dimensions lors de la visualisation des relations

complexes dans les données génétiques, pour minimiser la distorsion tout en conservant une interprétation visuelle claire et intuitive des mouvements de populations anciennes.

IV. Projections des résultats et interprétation

1. Projections géographiques

Les projections MDS (Multi-Dimensional Scaling) en 2D et 3D présentées dans les figures visent à visualiser les relations entre les mutations génétiques des individus, en particulier les SNPs du chromosome Y, et leurs zones géographiques d'origine. Chaque point sur ces graphiques représente un individu, et les couleurs indiquent les régions géographiques associées, telles que l'Europe, l'Asie, l'Afrique, et les Amériques. Ces projections permettent d'étudier comment les mutations génétiques s'organisent selon des zones géographiques spécifiques, facilitant ainsi l'identification de corrélations entre certaines mutations et des régions du monde.

Dans la projection MDS en 2 dimensions, les individus sont répartis selon deux axes principaux, ce qui permet d'observer quelques distinctions géographiques notables. Par exemple, les individus de l'Europe du Sud (en orange) et de l'Afrique de l'Est (en rouge) forment des clusters relativement distincts des autres groupes, indiquant une divergence génétique importante pour ces populations. Ces groupes sont géographiquement et génétiquement distincts, ce qui reflète des mutations spécifiques à ces zones. Toutefois, la majorité des points restent concentrés au centre, illustrant une certaine similarité génétique entre les individus de certaines régions. Cela peut être dû en partie à la compression des données lors de la réduction à deux dimensions, qui peut masquer les différences plus subtiles entre certaines zones géographiques. La projection 2D, bien qu'informative, ne parvient pas toujours à capturer pleinement les nuances des relations génétiques entre les individus.

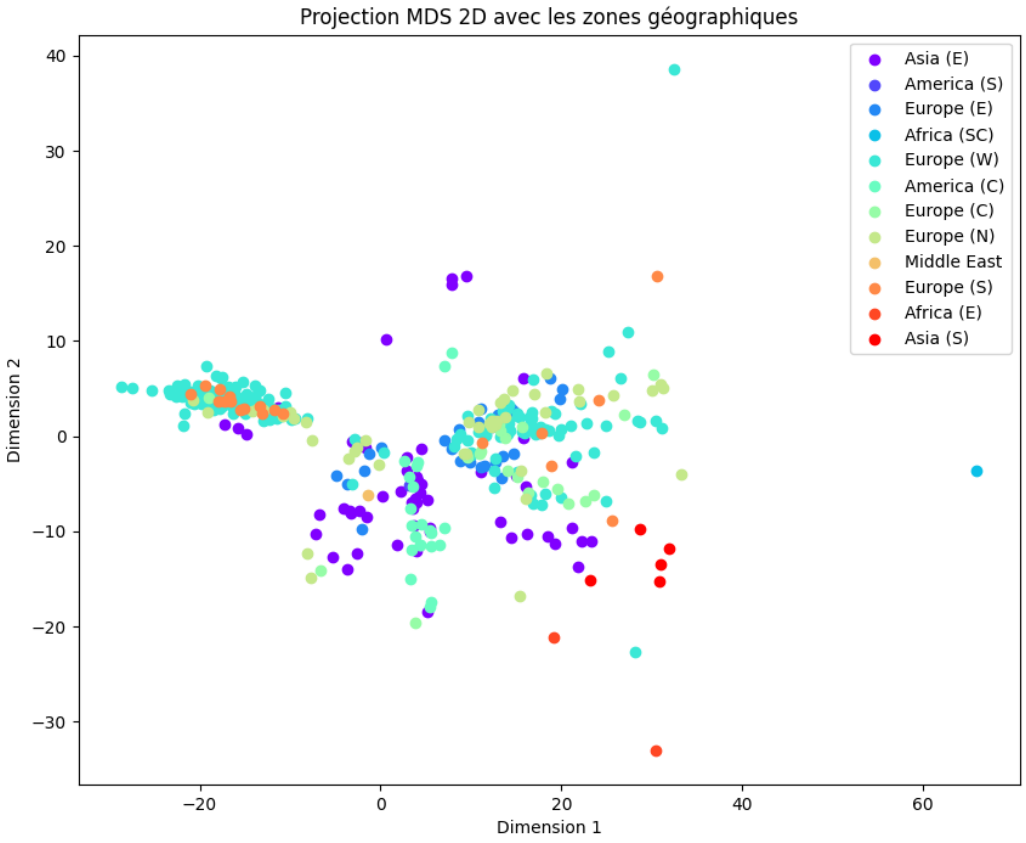

Figure 6 : *Projection MDS en 2 dimensions avec les zones géographiques*

La projection en 3 dimensions, quant à elle, offre une vue plus fine et une meilleure séparation des groupes géographiques. L'ajout d'une dimension supplémentaire permet de mieux représenter la diversité des mutations génétiques et les liens avec les zones géographiques. Par exemple, les individus de l'Asie du Sud (en violet) et de l'Afrique de l'Est (en rouge), déjà distincts en 2D, sont également bien séparés dans l'espace 3D. En outre, d'autres groupes, comme ceux de l'Europe de l'Ouest (en bleu clair) et du Moyen-Orient (en jaune), apparaissent plus clairement différenciés dans la projection en 3D. Cette représentation tridimensionnelle met en lumière des relations géographiques plus complexes et permet une visualisation plus précise des mutations génétiques associées aux régions géographiques, révélant des schémas de diversité génétique plus subtils et des regroupements régionaux plus cohérents.

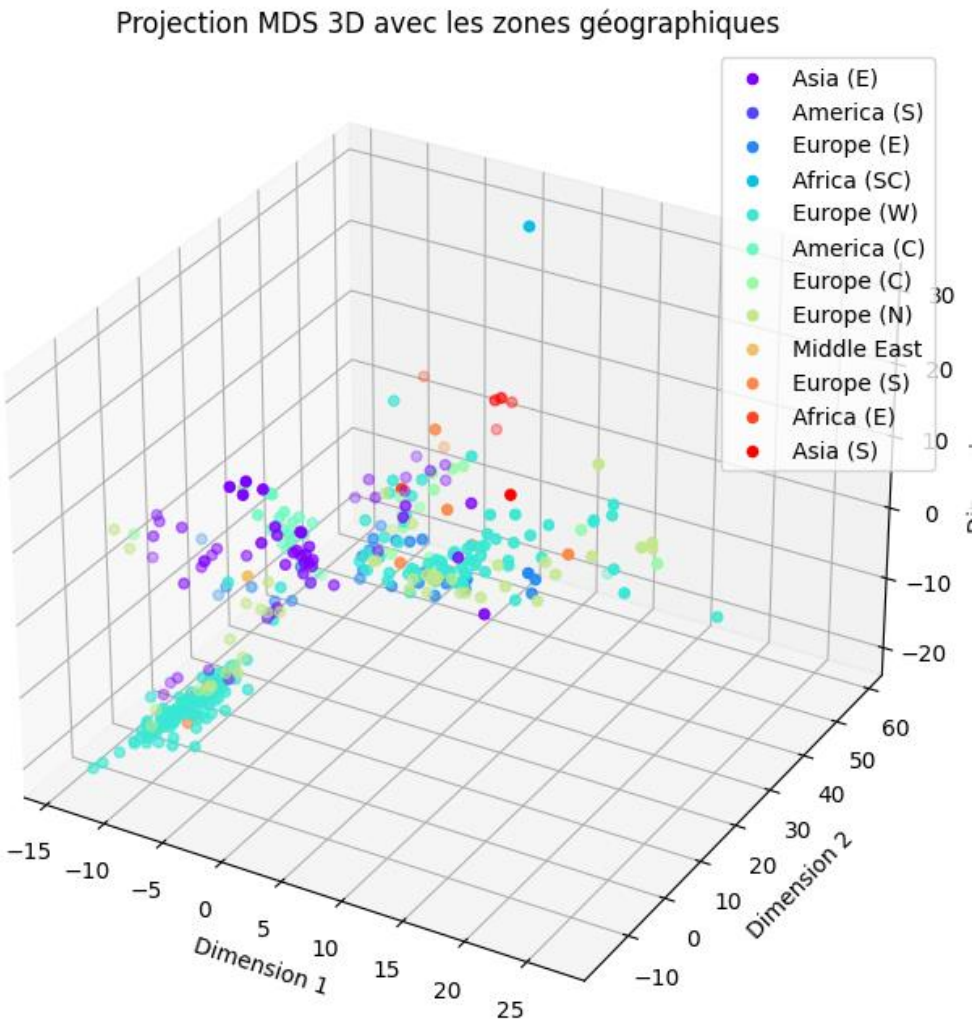

Figure 7 : *Projection MDS en 3 dimensions avec les zones géographiques*

En conclusion, les projections géographiques en 2D et 3D offrent un aperçu visuel des liens entre les mutations génétiques et les zones géographiques des individus. Bien que la projection 2D fournisse une vue d'ensemble, la projection en 3D améliore la compréhension des relations géographiques, révélant des schémas plus distincts et des regroupements plus cohérents entre les haplogroupes et les régions d'origine. Ces résultats soulignent la complexité des migrations humaines et des évolutions génétiques à travers les âges, et démontrent que certaines régions du monde abritent des populations génétiquement distinctes, tandis que d'autres montrent une plus grande diversité génétique, reflet de leur rôle historique dans les échanges migratoires.

2. Projections des haplogroupes Y

Dans cette section, nous analysons les projections MDS (Multi-Dimensional Scaling) en 2D et en 3D des haplogroupes Y, qui permettent de visualiser les relations entre les mutations génétiques du chromosome Y et les individus. Les différentes couleurs sur les graphiques représentent chaque haplogroupe Y, et les positions des points révèlent les relations de similarité génétique entre ces individus. Ces projections sont particulièrement utiles pour comprendre la répartition des haplogroupes Y dans le contexte des migrations humaines passées, en mettant en lumière les regroupements génétiques et les divergences observées entre les individus.

La projection MDS en 2 dimensions montre une vue d'ensemble simplifiée des relations génétiques entre les haplogroupes Y. Sur ce graphique, chaque point coloré correspond à un individu, et les couleurs représentent les haplogroupes auxquels ces individus appartiennent. Bien que cette visualisation permette de distinguer certains haplogroupes, notamment ceux représentés par les couleurs rouges et bleues, une grande partie des points est concentrée au centre du graphique. Cette concentration centrale indique une forte similarité génétique entre ces individus, rendant parfois difficile la distinction claire entre plusieurs haplogroupes. En effet, les mutations partagées entre certains individus créent des regroupements étroits, et dans certains cas, les différences entre haplogroupes ne sont pas clairement visibles dans l'espace réduit en 2D. Cette limitation est en partie due à la compression des données lors de la réduction de la dimensionnalité, où certaines nuances dans les mutations génétiques sont perdues.

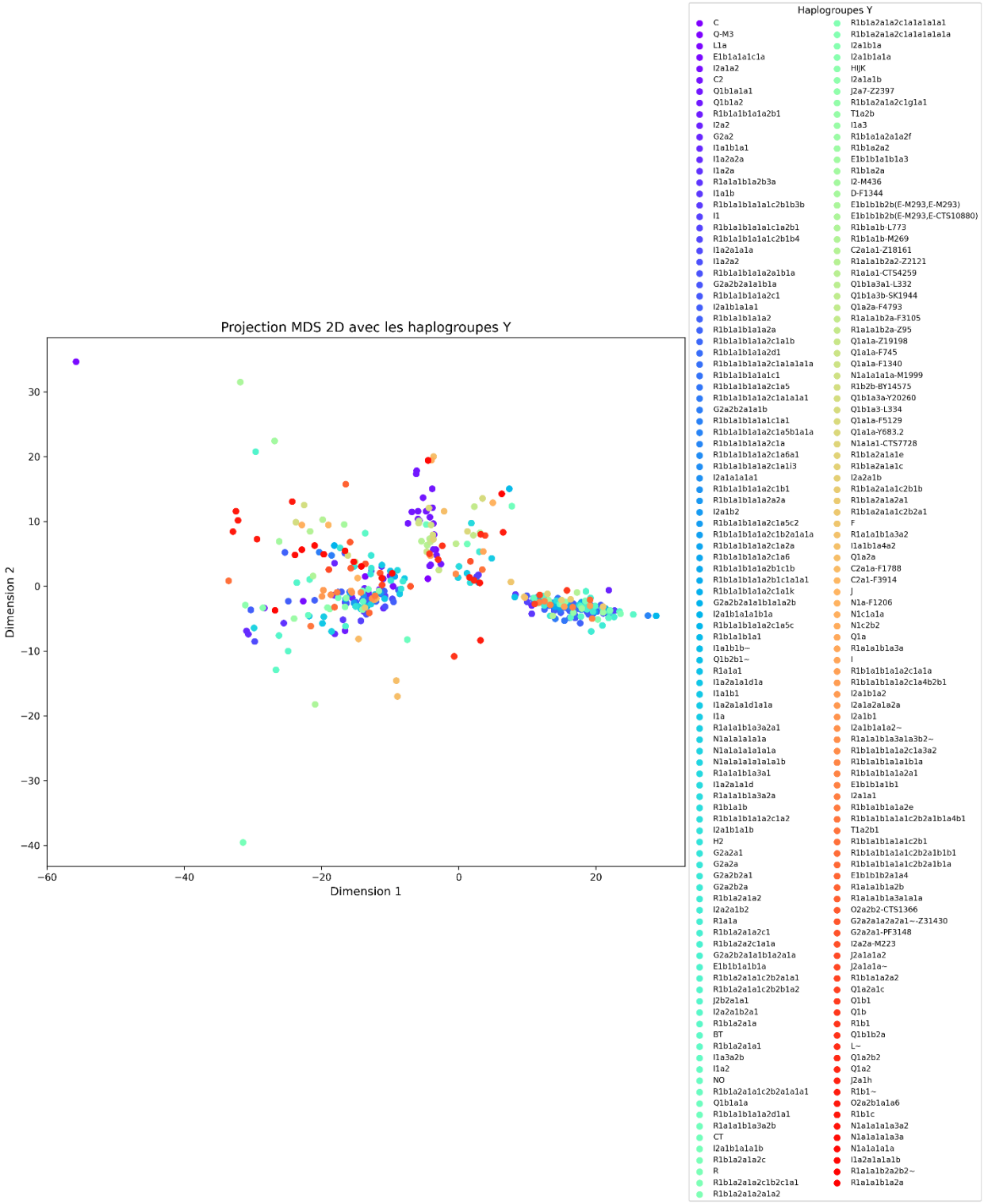

Figure 8 : *Projection MDS en 2D des haplogroupes Y*

La projection MDS en 3 dimensions apporte des améliorations significatives par rapport à la projection en 2D. En ajoutant une dimension supplémentaire, les individus sont mieux séparés dans l'espace, ce qui permet de distinguer plus clairement certains haplogroupes. Ceux qui semblaient se superposer dans la projection 2D apparaissent désormais plus distincts dans l'espace tridimensionnel. Cela permet une meilleure représentation des similarités et des différences génétiques entre les haplogroupes. Par exemple, des haplogroupes qui étaient difficiles à distinguer dans la projection 2D, car trop proches les uns des autres, se révèlent plus précisément en 3D, en particulier les individus appartenant aux groupes rouges et bleus. La répartition des points est également plus homogène, et les chevauchements observés en 2D sont réduits, offrant ainsi une représentation plus claire et plus fidèle des relations génétiques entre les individus.

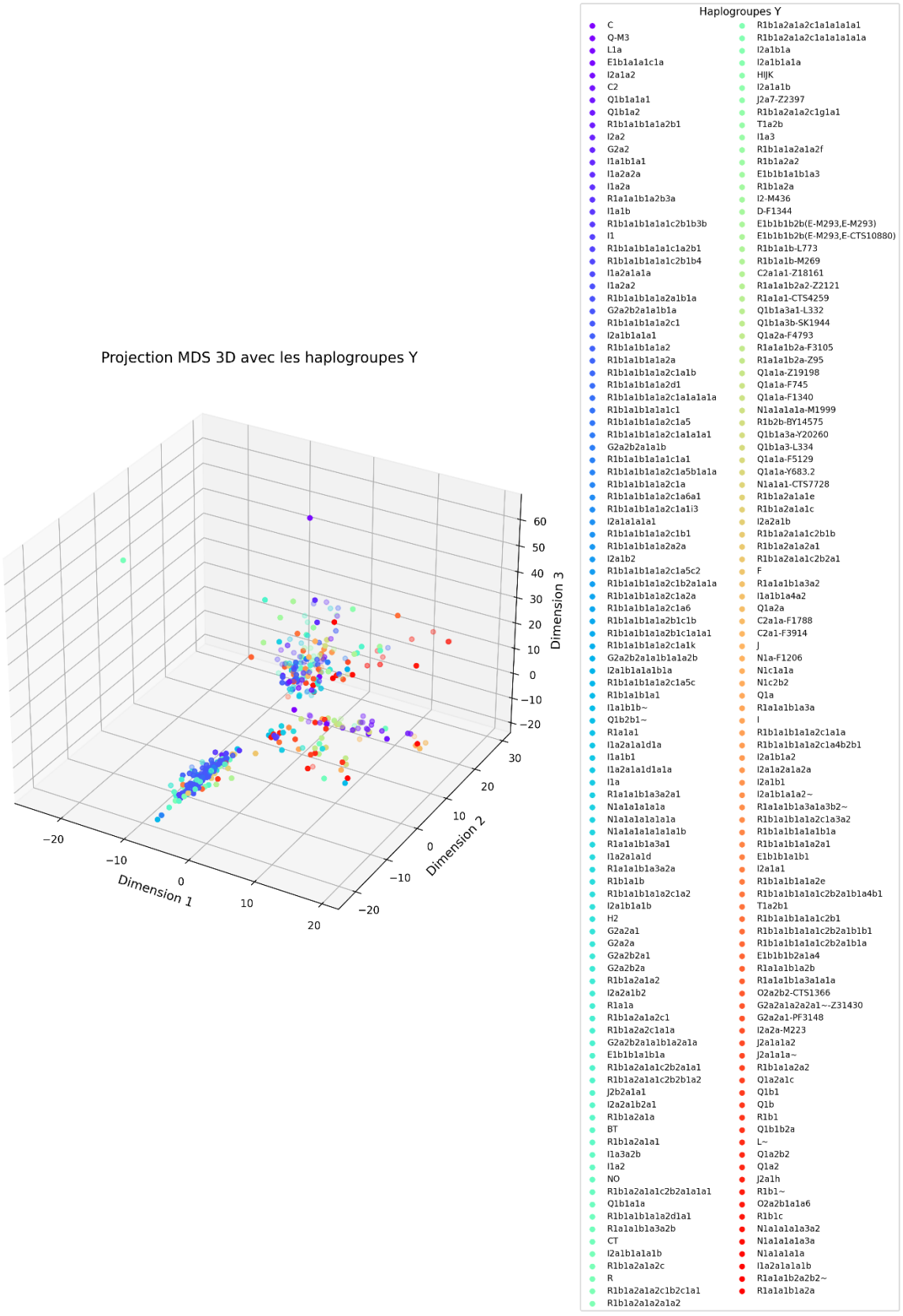

Figure 9 : *Projection MDS en 3D des haplogroupes Y*

Les projections MDS, en 2D et en 3D, révèlent des clusters génétiques correspondant aux différents haplogroupes Y. Ces regroupements sont particulièrement intéressants pour l'étude des migrations humaines, car ils permettent de visualiser comment certaines populations se sont déplacées et ont évolué de manière indépendante. Les individus appartenant à un même haplogroupe partagent des mutations similaires, ce qui reflète souvent des trajectoires migratoires communes ou des périodes d'isolement géographique. Par exemple, les haplogroupes représentés par des points rouges et bleus montrent une forte séparation par rapport à d'autres groupes. Ces distinctions peuvent être associées à des populations qui ont migré ou évolué séparément, accumulant ainsi des mutations spécifiques à leur groupe. Ces mutations peuvent être le résultat d'un isolement géographique, où les groupes sont restés isolés des autres populations pendant de longues périodes, favorisant l'apparition de mutations génétiques distinctes dans leurs lignées patrilinéaires. À l'inverse, les individus qui apparaissent plus groupés ou superposés dans les projections, en particulier au centre du graphique en 2D, montrent des similarités génétiques plus importantes, suggérant que ces populations ont connu un brassage génétique plus récent, potentiellement dû à des migrations ou à des interactions entre différentes populations.

L'analyse des clusters génétiques permet également de relier ces regroupements aux migrations anciennes des populations humaines. Certains haplogroupes distincts observés dans la projection 3D peuvent être associés à des vagues migratoires spécifiques, comme la migration indo-européenne en Europe, l'expansion mongole en Asie, ou encore les migrations des premiers Amérindiens à travers le détroit de Béring. Ces événements migratoires ont laissé des traces génétiques distinctes dans certaines populations, qui apparaissent aujourd'hui sous forme de clusters dans les projections.

Les populations ayant migré vers des régions géographiquement isolées, telles que des îles ou des zones montagneuses, ont souvent évolué de manière indépendante, accumulant des mutations uniques en raison du faible brassage génétique avec d'autres populations. Ces mutations se traduisent aujourd'hui par des haplogroupes distincts, facilement visibles dans les projections MDS. En étudiant ces haplogroupes, il est possible de retracer les trajectoires migratoires de ces populations anciennes et de comprendre comment ces migrations ont influencé la diversité génétique actuelle.

En conclusion, les projections MDS des haplogroupes Y révèlent des informations précieuses sur les relations génétiques entre les individus et les mutations du chromosome Y. La projection 2D fournit une vue simplifiée des regroupements génétiques, mais elle présente des limitations en termes de clarté et de distinction entre certains haplogroupes en raison de la compression des données. En revanche, la projection 3D améliore nettement la clarté et la précision, offrant une meilleure séparation des individus et une représentation plus fidèle des relations génétiques sous-jacentes. Ces résultats soulignent l'importance d'utiliser une analyse tridimensionnelle pour étudier les données génétiques complexes, car cela permet de capturer plus de détails et de révéler des patterns de migration et d'évolution qui ne seraient pas visibles dans une projection en 2D. En conclusion, l'ajout d'une troisième dimension dans l'analyse MDS est essentiel pour mieux comprendre les dynamiques évolutives et les migrations associées aux haplogroupes Y.

V. Limites et perspectives de l'approche génomique

L'analyse génomique, notamment à travers l'étude de l'ADN ancien et des polymorphismes nucléotidiques (SNPs), a permis des avancées significatives dans la compréhension des mouvements de population. Cependant, cette approche présente des limites importantes, qui doivent être examinées pour mieux appréhender ses possibilités et ses contraintes.

L'utilisation de l'ADN ancien pour retracer les migrations humaines comporte plusieurs difficultés. Tout d'abord, la recombinaison génétique complique l'interprétation des relations génétiques entre individus. Les recombinaisons sont des processus naturels qui échangent des segments d'ADN entre les chromosomes parentaux, rendant parfois difficile la distinction claire des lignées ancestrales et récentes. Cette complexité introduit une perte d'information, car les recombinaisons brouillent la transmission linéaire des traits génétiques. De plus, les SNPs manquants sont un autre problème majeur. Dans l'ADN ancien, les échantillons sont souvent dégradés en raison du temps et des conditions environnementales, ce qui entraîne des données incomplètes. Ces SNPs manquants rendent difficile la reconstitution précise des lignées et la compréhension des relations génétiques entre les populations.

En outre, la focalisation sur le chromosome Y et l'absence d'ADN mitochondrial dans certaines bases de données limitent la portée des études. L'ADN mitochondrial, qui se transmet par la lignée maternelle, est essentiel pour étudier les migrations féminines. En l'absence de cette information, seules les lignées patrilinéaires sont analysées, ce qui laisse de côté une partie importante de l'histoire génétique. Cette absence crée un biais, réduisant la capacité à obtenir une vue complète des dynamiques migratoires passées. Par ailleurs, il est également difficile de retracer les mouvements récents des populations en utilisant uniquement l'ADN ancien. Les migrations survenues au cours des derniers siècles, qui ont souvent impliqué des mélanges génétiques massifs, sont plus difficiles à détecter car elles se confondent avec les signaux génétiques plus anciens. Ainsi, l'approche génomique a du mal à capturer les migrations récentes et à les distinguer des vagues migratoires plus anciennes.

Face à ces limites, il est important d'explorer des approches complémentaires pour enrichir l'analyse des migrations humaines. L'archéologie, la linguistique, et les méthodes isotopiques sont des disciplines qui peuvent apporter des informations cruciales pour compléter les études génomiques. Par exemple, les fouilles archéologiques fournissent des preuves matérielles des migrations, telles que des artefacts, des structures et des sépultures, qui peuvent être corrélées avec des découvertes génétiques. De même, l'étude des langues peut révéler les mouvements des populations à travers l'évolution et la diffusion des langues. Les groupes linguistiques qui se sont étendus à travers le monde sont souvent associés à des migrations humaines, et leur analyse peut renforcer les conclusions tirées des études génétiques. Les analyses isotopiques, en examinant les éléments chimiques présents dans les ossements et les dents, permettent d'identifier l'origine géographique des individus et d'en savoir plus sur leur mobilité au cours de leur vie. Ces méthodes offrent une perspective plus large sur les déplacements des populations, venant enrichir les données génétiques.

Pour améliorer les méthodes d'analyse génomique, l'inclusion de l'ADN mitochondrial est une piste cruciale. Cela permettrait d'obtenir une vision plus complète des migrations humaines en intégrant les lignées matrilinéaires, souvent négligées dans les analyses actuelles basées sur le chromosome Y. Par ailleurs, l'utilisation croissante de l'intelligence artificielle (IA) peut révolutionner la façon dont ces données sont analysées. Des

algorithmes d'apprentissage automatique peuvent combiner plusieurs sources de données, telles que les informations génétiques, géographiques, archéologiques et historiques, pour créer des modèles plus sophistiqués et précis des mouvements de population. Par exemple, en intégrant les modèles géographiques avec les données archéologiques, il serait possible de générer des simulations des migrations humaines anciennes et récentes, permettant de mieux comprendre les dynamiques complexes qui ont façonné les populations contemporaines. L'IA peut également aider à repérer des schémas cachés dans les données massives, facilitant ainsi la détection des migrations récentes, qui sont souvent masquées par les événements plus anciens dans les analyses traditionnelles.

En conclusion, bien que l'approche génomique présente des limites notables, elle peut être considérablement enrichie par des méthodes complémentaires et des avancées technologiques. L'intégration de disciplines telles que l'archéologie et la linguistique, combinée à l'utilisation de l'intelligence artificielle, permettra de pallier les manques actuels et de fournir une vision plus complète des migrations humaines. Cela contribuera à améliorer la précision des analyses génétiques et à offrir des perspectives nouvelles sur l'évolution des populations humaines au fil des millénaires.

Conclusion

En conclusion, ce projet a permis de mettre en lumière l'importance de l'analyse génomique, en particulier des SNPs du chromosome Y, pour mieux comprendre les migrations humaines anciennes. À travers l'utilisation d'algorithmes de réduction de dimensionnalité comme le MDS et l'implémentation de la méthode SMACOF, nous avons pu visualiser les relations génétiques complexes entre les individus, tout en identifiant des clusters correspondant à des haplogroupes spécifiques. Ces résultats montrent que l'étude des haplogroupes Y permet de retracer des trajectoires migratoires et de mieux comprendre la diversité génétique actuelle. Toutefois, cette approche présente certaines limites, notamment l'absence d'ADN mitochondrial et la difficulté à capturer les mouvements récents. Les perspectives d'amélioration incluent l'intégration de données archéologiques et linguistiques, ainsi que l'utilisation croissante de l'intelligence artificielle pour combiner plusieurs sources de données. Ces développements permettront une meilleure compréhension des migrations humaines et une approche plus complète de l'évolution des populations à travers le temps.

Bibliographie

Behar, D.M., et al., 2007. The Genographic Project Public Participation Mitochondrial DNA Database. *PLoS Genet*, 3(6), p.e104.

Borg, I. & Groenen, P., 1997. *Modern Multidimensional Scaling: Theory and Applications*. Springer Series in Statistics. Springer.

Cavalli-Sforza, L.L., Menozzi, P. & Piazza, A., 1994. *The History and Geography of Human Genes*. Princeton: Princeton University Press.

Coudray, C., 2004. *Les haplogroupes des populations anciennes*. Disponible à : http://www.didac.ehu.es/antropo/18/18-6/Coudray.pdf [Consulté le 12 septembre 2024].

Eupedia, 2024. *Y-DNA Haplogroups in Europe*. Disponible à : https://www.eupedia.com/europe/cartes_haplogroupes_ADN-Y.shtml [Consulté le 12 septembre 2024].

International Society of Genetic Genealogy (ISOGG), 2024. *ISOGG Y-DNA Haplogroup Tree 2024*. Disponible à : https://isogg.org/ [Consulté le 12 septembre 2024].

Jobling, M.A., Hollox, E.J., Hurles, M.E., Kivisild, T. & Tyler-Smith, C., 2013. *Human Evolutionary Genetics: Origins, Peoples & Disease*. 2e éd. New York: Garland Science.

Kruskal, J.B., 1964. Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 29(2), pp.115–129.

Kruskal, J.B., 1964. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1), pp.1–27.

National Geographic, 2024. *The Genographic Project*. Disponible à : https://www.nationalgeographic.org/projects/genographic [Consulté le 12 septembre 2024].

Oppenheimer, S., 2004. *Out of Eden: The Peopling of the World*. London: Robinson Publishing.

Rajawat, K. & Kumar, S., 2016. Stochastic Multidimensional Scaling. *IEEE*. Disponible à : https://arxiv.org/pdf/1612.07089.

Reich, D., et al., 2023. *Allen Ancient DNA Database*. Harvard Medical School, Reich Lab. Disponible à : https://reich.hms.harvard.edu/datasets [Consulté le 12 septembre 2024].

Underhill, P.A., et al., 2000. Y Chromosome Sequence Variation and the History of Human Populations. *Nature Genetics*, 26(3), pp.358–361.

van Oven, M. & Kayser, M., 2009. Updated Comprehensive Phylogenetic Tree of Global Human Mitochondrial DNA Variation. *Human Mutation*, 30(2), pp.E386–E394.

Wells, S., 2002. *The Journey of Man: A Genetic Odyssey*. Princeton: Princeton University Press.

Kruskal, J.B., 1964. *Nonmetric multidimensional scaling: A numerical method*. Psychometrika, 29(2), pp.115–129.

Kruskal, J.B., 1964. *Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis*. Psychometrika, 29(1), pp.1–27.

Borg, I. & Groenen, P., 1997. *Modern Multidimensional Scaling: Theory and Applications*. Springer Series in Statistics. Springer.

SMACOF Package, 2024. https://cran.r-project.org/web/packages/smacof/smacof.pdf