Paléogénomique de l'évolution des Bovina et l'impact sur la domestication des bovins

Wejden Ben Dhafer

université PARIS-SACLAY

Paléogénomique de l'évolution des *Bovina* et l'imapct sur la domestication des bovins

Paleogenomics of the evolution of Bovina and its impact on cattle domestication

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 577, Structure et Dynamique des Systèmes Vivants (SDSV)
Spécialité de doctorat: Génétique
Graduate School : Sciences de la Vie et Santé. Référent : Faculté des sciences d'Orsay

Thèse préparée dans l'unité de recherche **Institut Jacques Monod (Université de Paris, CNRS)**, sous la direction de **Eva-Maria GEIGL**, directrice de recherche au CNRS, et la co-direction de **Thierry GRANGE**, directeur de recherche au CNRS

Thèse soutenue à Paris-Saclay, le 20 décembre 2021, par

Wejden BEN DHAFER

Composition du Jury

Jean-Luc GUADELLI Docteur, Directeur de Recherche CNRS, Université de Bordeaux	Président
Christiane DENYS Professeure, Muséum National d'Histoire Naturelle	Rapporteure & Examinatrice
Christine KEYSER Professeure, Université de Strasbourg	Rapporteure & Examinatrice
Claude THERMES Docteur, Directeur de Recherche CNRS, Université de Paris-Saclay	Examinateur
Arturo MORALES Professeur, Université Autonome de Madrid	Examinateur
Claude GUINTARD Docteur, ONIRIS, Nantes Atlantique	Examinateur
Eva-Maria GEIGL Docteur, Directrice de Recherche CNRS, Université de Paris	Directrice de thèse
Thierry GRANGE Docteur, Directeur de Recherche CNRS, Université de Paris	Co-Directeur de thèse

THESE DE DOCTORAT
NNT : 2021UPASL107

université PARIS-SACLAY

ÉCOLE DOCTORALE
Structure et dynamique des systèmes vivants (SDSV)

Titre : Paléogénomique de l'évolution des *Bovina* et l'impact sur la domestication des bovins

Mots clés : Paléogénomique, Evolution, Domestication, Dynamiques de populations, *Bos* et *Bison*

Résumé : Les genres *Bos* et *Bison* constituent la sous-tribu des *Bovina* et incluent plusieurs lignées de bovins. Les espèces domestiques et sauvages appartenant à ces deux genres auraient divergé au Pléistocène inférieur. L'histoire génomique évolutive récente de *Bos* et *Bison* ne reflète qu'imparfaitement les liens phylogénétiques anciens car ceux-ci sont éclipsés par les événements récents, en particulier la quasi-extermination aux 19éme et 20éme siècles des bisons américains (*Bison bison*) et bisons européens (*Bison bonasus*), respectivement, ainsi que la disparition au 17éme siècle de l'aurochs, ancêtre des bovins domestiques actuels.

Dans le but de clarifier l'histoire évolutive de ces deux genres, nous avons mis en place une approche paléogénomique permettant de séquencer l'ADN ancien extrait à partir de spécimens fossiles des genres *Bos* et *Bison*. L'objectif était de reconstruire les mitogénomes et les génomes nucléaires de restes fossiles datant du Pléistocène a l'Holocène, témoins des événements évolutifs précédant les événements les plus récents qui ont remodelé la diversité génétique des populations modernes. Nous avons ainsi analysé des restes fossiles du Néolithique jusqu'à l'ère moderne, témoins de l'impact de la domestication sur la structuration génétique des bovins.

Nous avons adapté les méthodes d'analyse de l'ADN aux caractéristiques particulières de l'ADN ancien pour maximiser la récupération de l'ADN ancien à partir des restes fossiles et convertir efficacement les fragments d'ADN contenus dans l'extrait fossile en banques génomiques pour le séquençage de nouvelle génération (NGS). Nous avons aussi optimisé un protocole de capture du génome mitochondrial de bovins anciens pour augmenter son efficacité avec des échantillons mal conservés. L'optimisation méthodologique nous a permis d'obtenir la séquence de 20 mitogénomes complets de bisons européens et bisons des steppes datant du Pléistocène moyen jusqu'aux aux temps modernes et 98 mitogénomes complets d'aurochs et de bovins domestiques eurasiatiques et africains datant du Pléistocène supérieur jusqu'au Moyen Age.

Nous avons reconstruit l'histoire évolutive de *Bos* et *Bison* à travers une analyse phylogénétique qui rassemble 279 mitogénomes anciens et 671 mitogénomes modernes des *Bovina*. Ceci nous a permis d'établir une datation robuste des radiations phylogénétiques et de reconstruire les dynamiques des populations bovines pendant les 50 000 dernières années.

Nous avons identifié un nouvel haplogroupe mitochondrial d'aurochs rassemblant des échantillons du Pléistocène supérieur originaire de l'Europe de l'Ouest. Cet haplogroupe s'est séparé le premier des lignées d'aurochs européens et proche-orientaux, des bovins domestiques eurasiatiques et africains ainsi que des zébus. Ceci met en évidence une dynamique complexe des populations bovines ancestrales au Pléistocène moyen et supérieur ainsi que les liens partagés entre les populations eurasiatiques et du sous continent indien.

L'analyse d'échantillons précédant et suivant le dernier maximum glaciaire, entre 30 000 et 12 000 ans, et les modélisations démographiques utilisant les mitogénomes anciens et actuels des différentes lignées ont permis d'évaluer l'impact de cette glaciation sur la dynamique des tailles effectives des populations d'aurochs. Notre étude permet de réévaluer les causes du goulot d'étranglement précédemment attribué à la domestication des aurochs. Nous avons aussi caractérisé en parallèle les génomes d'une vingtaine de bisons qui permettront de comparer l'évolution de ces deux lignées cousines, l'une ayant été domestiquée et l'autre non. L'ensemble du travail réalisé permet d'établir une base solide pour étudier les changements génomiques associés à la domestication des bovins.

université PARIS-SACLAY

ÉCOLE DOCTORALE
Structure et dynamique des systèmes vivants (SDSV)

Title : Paleogenomics of the evolution of *Bovina* and its impact on cattle domestication

Keywords : Paleogenomics, Evolution, Domestication, Populations dynamics, *Bos* and *Bison*

Abstract : The *Bos* and *Bison* genera constitute the *Bovina* subtribe and include several bovine lineages. Domestic and wild species belonging to these two genera supposedly have diverged in the Lower Pleistocene. The recent genomic evolutionary history of *Bos* and *Bison,* however, only partially reflects their ancient phylogenetic links, since recent events, in particular the near-extermination of American bison (*Bison bison*) and European bison (*Bison bonasus*) in the 19th and 20th centuries, respectively, and the extinction in the 17th century of the aurochs, ancestor of present-day domestic cattle, have blurred our understanding of their phylogenetic relationships. In order to clarify the evolutionary history of these two genera, we implemented a paleogenomic approach, i.e., sequencing ancient DNA extracted from fossil specimens of *Bos* and *Bison,* in order to reconstruct their mitogenomes and nuclear genomes dating from the Pleistocene to the Holocene.

These fossil specimens are direct witnesses of ancient evolutionary events, allowing us to directly investigate the genetic diversity of *Bos* and *Bison* before the most recent events, which reshaped the genetic diversity of modern populations. Furthe more, to investigate the impact of domestication on the genetic structure of cattle, we analyzed fossil remains from the Neolithic to modern times. We adapted DNA analysis methods to the particular characteristics of ancient DNA to maximize the recovery of ancient DNA from fossil remains and to efficiently convert DNA fragments contained in fossil extracts into genomic libraries for next-generation sequencing (NGS). We also optimized a protocol for capturing ancient cattle mitochondrial genomes to increase its efficiency for poorly preserved samples. This methodological optimization allowed us to obtain 20 complete mitogenomes of European bison and steppe bison dating from the Middle Pleistocene to the modern times, and 98 complete mitogenomes of aurochs and Eurasian and African domestic cattle dating from the Upper Pleistocene until the Middle Age. We reconstructed the evolutionary history of *Bos* and *Bison* through phylogenetic analysis integrating 279 ancient mitogenomes and 671 modern mitogenomes from the *Bovina* subtribe. This compilation has allowed us to establish a robust dating of the phylogenetic radiation and to reconstruct the dynamics of bovine populations during the last 50,000 years.

We have identified a new mitochondrial haplogroup of aurochs in a group of Upper Pleistocene samples from Western Europe. This haplogroup diverged early from all European and Near-Eastern aurochs, Eurasian and African domestic cattle, as well as Indian zebus, highlighting a complex dynamic of ancestral *Bos* populations in the Middle and Late Pleistocene, as well as revealing the links shared between the Eurasian and the Indian subcontinent populations. Through analysis of samples preceding and following the last glacial maximum, 30,000 and 12,000 years ago, and through use of demographic models combining ancient and present-day mitogenomes of different lineages, we were able to assess the impact of this glaciation on the effective population sizes of aurochs. Thus, our study reassesses the causes of the bottleneck previously attributed to the domestication of aurochs. In parallel, we have also characterized the genomes of an additional twe ty bison, which will allow us to further compare the evolution of these two cousin lineages, one having been domesticated and the other not. Together, these analyses provide a solid basis for studying and better understanding the genomic changes associated with the domestication of cattle.

Remerciements

« La gratitude est la clé qui ouvre les portes du vrai savoir. »

Omraam Mikhaël Aïvanhov

J'adresse mes premiers remerciements pour **Eva-Maria** et **Thierry** pour m'avoir accueilli au sein de leur laboratoire afin d'effectuer mon stage de M2 et me donner l'opportunité de faire ma thèse. Je vous remercie pour votre assistance par vos conseils, vos encouragements et vos bienveillances. J'ai appris de vous tellement de choses pendant ces années qui vont me servir aussi bien dans ma vie personnelle que professionnelle. Merci pour votre soutien et pour tous les moments inoubliables qu'on avait passés ensemble. Vous êtes une famille.
Je garderai un souvenir ému de mon passage dans votre laboratoire qui m'a tant donné. Grâce à vous, j'aborde une nouvelle étape de ma vie avec une vision tout à fait différente.

Je remercie vivement les membres du Jury, **Pr. Christiane Denys**, **Pr. Christine Keyser**, **Dr. Claude Thermes**, **Dr. Jean-Luc Guadelli**, **Prof. Arturo Morales** et **Dr. Claude Guintard** pour leurs intérêts et le temps accordé pour la lecture de mon manuscrit.

Je remercie **Eve GAZAVE** et **Dominique Rocha** pour leur participation à mon comité de suivi de thèse et leurs suggestions toujours avisées.

Je remercie mes compagnons de route et amis qui ont contribué au maintien d'une bonne humeur au sein du laboratoire.
Jeanne, ma chère amie, merci pour tes câlins aux moments de déception, merci de m'écouter et merci pour les beaux moments qu'on avait passés ensemble au sein et en dehors du laboratoire. Tu es extraordinaire.
Oguzhan, une des plus belles personnes au monde, sympathique et cool. Merci pour tes encouragements à chaque fois qu'on se parle et on se voit.
Caitlin, ce qui est sûr c'est qu'on ne s'ennuie
pas avec toi. Tu es impressionnante. Merci pour ton soutien et ton encouragement.
La petite **Hélène**, souriante et motivée, bon courage.

Une pensée à **Diyendo**, **Andy**, **Nastassia** et **Zoé.**

Je remercie **Olivier Rué**, mon tuteur de la formation «Initiation au traitement des données de génomique» à Roscoff.

J'aimerai remercier tous les chercheurs et les étudiants de l'Institut Jacques Monod ainsi que le personnel technique pour leur accompagnement quotidien en particulier **Laurent**, **Norry**, **Alfred** et **Jonathan**.

Je remercie Mme **Sandrine le Bihan,** assistante de l'école doctorale SDSV, pour la qualité de son travail, son efficacité et ses orientations.

Dédicace

« Soyons reconnaissants aux personnes qui nous donnent du bonheur, elles sont les charmants jardiniers par qui nos âmes sont fleuries. »

Marcel Proust

En termes de connaissance de leurs sacrifices et en témoignage de mes profonds sentiments à leurs égards, je dédie ce travail :

A mon père, pour son soutien moral, son affection et la confiance qu'il m'a accordé.
A ma mère pour ses sacrifices, son amour et ses encouragements.
Je vous aime de tout mon cœur…

A mes sœurs, merci de me soutenir et de m'accepter comme je suis. Vous êtes ma vie.

A mon petit frère, ma source de bonheur. Merci de me transmettre une énergie positive à travers tes sourires et les histoires de ton monde innocent.

A Issam, ma source d'espoir et de motivation. Merci de me soutenir, de me conseiller, de m'encourager et d'être toujours à mes côtés.

A mes chers amis, cousins et cousines, **Afef**, **Hana**, **Sabrine**, **Kaouther**, **Andrea**, **Yossra**, **Kawther**, **Anwar**, **Hana**, **Ayet**, **Safa**, **Samar**, **Myriam**, **Yessmine**, **Molka**, **Sarra**, **Sourour**, **Montassar**, **Youssef**, **Zied, Yessine, Nour, Ahmed, Hounaida, Wided,** qui me soutiennent par leurs petits messages quotidiens.

A ma grande mère, mes tentes et mes oncles, **Khadija, Tarek**, **Hamed**, **Mourad**, **Olfa**, **Najla**, **Salha**, **Saida**, **Lamia**, **Omar**, **Fathi**, **Najet**, **Habiba**, **Aouicha**, je vous aime beaucoup.

Une pensée pour **Marwa** qui m'a soutenu jusqu'à ses derniers jours, **Hamza**, mon cousin et mon ami d'enfance, et pour mes grand-parents. Que vos âmes reposent en paix.

Merci…

Table of Contents

Introduction

L’évolution génétique s’intéresse à l’étude des changements génétiques des organismes au cours du temps et plus précisément à l’étude des changements des fréquences alléliques et génotypiques. Plus les espèces sont apparentées génétiquement plus leurs séquences d’ADN sont proches. Ceci permet aux biologistes d’étudier l’évolution en analysant les modifications qui peuvent s’accumuler au cours du temps, ce qui tend à augmenter la diversité génétique. Plusieurs mécanismes génétiques sont responsables de l’évolution dont chacun a un effet particulier. En effet, les mutations spontanées et la sélection naturelle jouent un rôle crucial dans le processus d’évolution. Mais en plus de ces deux phénomènes évolutifs, interviennent les migrations des êtres humains ou des animaux ce qui peut entraîner certaines modifications génétiques dans les populations due à la dérive génétique ou à l’adaptation aux nouvelles conditions de vie. Ces modifications génétiques peuvent être étudiées grâce aux nouvelles technologies de séquençage couramment appelées NGS pour « Next Generation Sequencing »(Wadman, 2008) et grâce aux approches bio-informatiques.

Ceci permet d’étudier les génomes nucléaires ou mitochondriaux des espèces actuelles ou éteintes et de retracer leurs évolutions. Des traces d’ADN peuvent être préservées dans des tissus anciens, le plus souvent dans les tissus calcifiés (os, dents), les tissus kératinisés (sabots, bois, ongles, cheveux) ou encore dans l’environnement (sols...) (Damgaard et al., 2015; Ellegaard et al., 2020; Gelabert et al., 2021). Il est possible de replacer des espèces éteintes au sein des arbres phylogénétiques des espèces à partir de fragments d’ossements, reconstruire leur évolution, ainsi qu’identifier les migrations, extensions, rétrécissements et remplacements des populations humaines, animales et végétales (Brunel et al., 2020; Grange et al., 2018; Kistler, 2012; Massilani et al., 2016; Mikić, 2015; Perri et al., 2021; van der Valk et al., 2021; Verdugo et al., 2019). Il est ainsi possible d’établir les éventuelles relations entre extinction des espèces et changements climatiques et environnementaux du passé. Ces derniers ont joué un rôle important dans la dispersion et les mouvements des espèces sur la terre. Chaque modification issue d’une sélection naturelle positive confère un avantage adaptatif naturel.

A partir du Néolithique, les mécanismes d’évolution appelés « sélection artificielle » ou bien « sélection dirigée » voient le jour car c’est l’être humain qui sélectionne dans une population animale ou végétale certains individus de façon dépendante de son besoin. En effet, depuis l’apparition de l’agriculture il y a plus de 10.000 ans, les agriculteurs et les éleveurs ont commencé à contrôler la reproduction des végétaux et des animaux et à sélectionner parfois ceux qui possédèrent des traits nouveaux.

Ce processus de domestication amorce l’émergence de la sélection dite artificielle. En effet, la domestication d’une espèce animale ou végétale résulte d’une interaction entre les êtres humains et les espèces animales ou végétales. Elle est considérée comme une sélection orientée d’une façon progressive et elle a plusieurs effets sur les populations d’intérêts car, par exemple, l’isolement d’un nombre limité d’individus conduit à la réduction de la diversité génétique initiale, mais ces individus vont subir une évolution différente de celle du reste de la population initiale et qui est susceptible d’aboutir à des adaptations aux nouvelles conditions environnementales. Ces adaptations sont les conséquences de l’expression de nouveaux traits génotypiques sélectionnés au cours de la domestication.

Comme pour les êtres humains où des données génétiques des restes fossiles ont permis de retracer l'histoire évolutive et de bien comprendre les origines des sociétés humaines, des mitogénomes préservés dans des restes paléontologiques et archéologiques ont permis aussi de retracer une partie de l'évolution des genres *Bos* et *Bison* de la famille des Bovidés qui incluent plusieurs espèces dont certaines ont été domestiquées et d'autres sont restées sauvages.

Dans le travail présenté ici, on vise à comprendre les événements de domestication des bovins qui ont eu lieu durant le Néolithique et à retracer le chemin évolutif au cours de leur domestication en utilisant des approches paléogénomiques. Celles-ci consistent en des méthodes particulières d'extraction et de construction de banques d'ADN ancien, suivies par le NGS. Les séquences obtenues seront ensuite analysées par des méthodes bio-informatiques, de phylogénie et de la génétique de populations. Dans la présente étude, des échantillons anciens du genre *Bos* et *Bison* récupérés de différents sites archéologiques et qui couvrent une large période temporelle ont été analysés en vue de répondre à différentes questions aussi bien techniques que phylogénétiques et phylogéographiques.

Chapitre I: Caractéristique de l’ADN ancien

Appelé aussi ADN fossile, il s’agit d’un ADN extrait de vestiges fossiles, dans la plupart des cas des os. Cet ADN est dégradé, de très petite taille, généralement de 30 à 100 nucléotides et présent en des quantités très faibles. La première séquence d’ADN ancien (ADNa) publiée correspond à un fragment d’ADN mitochondrial (ADNmt) de 229 pb d’une sous-espèce de zèbre, le Quagga ou Equus quagga quagga qui s’est éteinte en 1883 (Higuchi et al., 1984). Ce résultat a ouvert des perspectives aux généticiens et paléogénéticiens et les a encouragés à utiliser cette approche pour comprendre les histoires évolutives des êtres vivants et remonter le temps, grâce à l’ADNa.
Les premières études qui ont suivi se sont avérées pleines d’artéfacts qui ont permis de comprendre les difficultés de l’analyse de l’ADNa. Je vais parler dans le paragraphe suivant des caractéristiques de l’ADNa.

I) La dégradation post-mortem :

Après la mort d’un organisme, les enzymes autolytiques dégradent toutes les composants des cellules. Lors de l’autolyse, les membranes cellulaires sont rompues, et les nucléases et les radicaux libres dégradent les molécules d’ADN. Simultanément, les bactéries symbiotiques et commensales du corps, les bactéries environnementales ainsi que les insectes commencent à décomposer le corps, un processus appelé « putréfaction ». En même temps agissent hydrolyse et oxydation sur les molécules d’ADN. Si quelques cellules de l’organisme échappent aux mécanismes de putréfaction, il est possible de retrouver des traces de son ADN. La majorité des mécanismes enzymatiques responsables de la dégradation de l’ADN ont besoin de l’eau pour être actifs. Si une partie de l’organisme se dessèche avant la dégradation complète de l’ADN, les processus de dégradation s’arrêtent et l’ADN peut être préservé. Au niveau des tissus durs, comme les dents et les os, les molécules d’ADN peuvent échapper à la dégradation complète parce qu’ils sont moins hydratés que les autres tissus. En outre, lors de la dégénérescence des cellules, les molécules d’ADN peuvent se lier à la matrice minérale du tissu dur ce qui peut leur permettre d’échapper à l’action des nucléases.

Après cette première phase de putréfaction et de dégradation enzymatique, l’ADN résiduel sera dégradé plus lentement au cours du temps par des réactions chimiques. Les fragments d’ADN présentent plusieurs lésions telles que des sites abasiques, des cassures simple brin et d’autres lésions qui provoquent des erreurs de codage. Ces erreurs reflètent une dégradation chimique de type oxydation ou hydrolyse. Les bases nucléotidiques, et en particulier les bases puriques A et G, peuvent être dissociées de la chaîne d’ADN par la rupture de la liaison glycosidique entre les bases et le désoxyribose. Le site abasique résultant étant très instable, va subir une réaction de béta-élimination qui dégrade le désoxyribose et conduit à la production d’une lésion simple brin (Figure 1) (A. W. Briggs et al., 2007). Lorsque des coupures simple brin sont présentes sur chaque brin en étant distantes de quelques nucléotides, deux fragments d’ADN de part et d’autre peuvent se séparer et les extrémités de ces molécules seront alors simples brins. Après cartographie des séquences sur le génome, on observe ainsi une fréquence plus élevée de purines à la position adjacente à l’extrémité des fragments séquencés.

L’hydrolyse touche aussi les résidus aminés des bases azotées. Les cytosines sont les bases les plus sensibles à la désamination. L’hydrolyse du groupement amine convertit les cytosines en uraciles. Les uraciles seront remplacés par des thymines pendant la PCR ce qui entraîne une transition C vers T (A. W. Briggs et al., 2007; Hofreiter, 2001). Les uraciles accumulés peuvent inhiber les polymérases (Heyn et al., 2010). La désamination des cytosines est beaucoup plus importante lorsque l’ADN est simple brin car les groupements amines sont plus exposés et réactifs lorsqu’ils ne sont pas impliqués dans des liaisons hydrogènes existant dans l’ADN double-brin. Cela se traduit par un taux de transition C vers T beaucoup plus élevé aux extrémités des fragments d’ADN (A. W. Briggs et al., 2007). Ces deux modifications, dépurination et désamination des cytosines, sont responsables de la signature de l’ADNa permettant de le distinguer de l’ADN moderne. Ainsi après cartographie des séquences sur le génome, on observe un enrichissement des transitions C vers T (ou G vers A sur le brin complémentaire) aux extrémités des fragments et un enrichissement en purines (ou pyrimidines sur le brin complémentaire) aux positions adjacentes proximales. Pour réduire le taux d’erreur dans les séquences, les cytosines désaminées sont éliminées, au moment de la construction des banques d’ADNa, à l’aide de l’enzyme USER (Uracil-Specific Excision Reagent) qui est un mélange de deux enzymes, d’une part de l’uracil DNA glycosylase (UDG) qui catalyse l’excision des uraciles formant ainsi un site abasique tout en laissant le squelette phosphodiester intact, et d’autre part une endonucléase VIII qui enlève le site abasique.

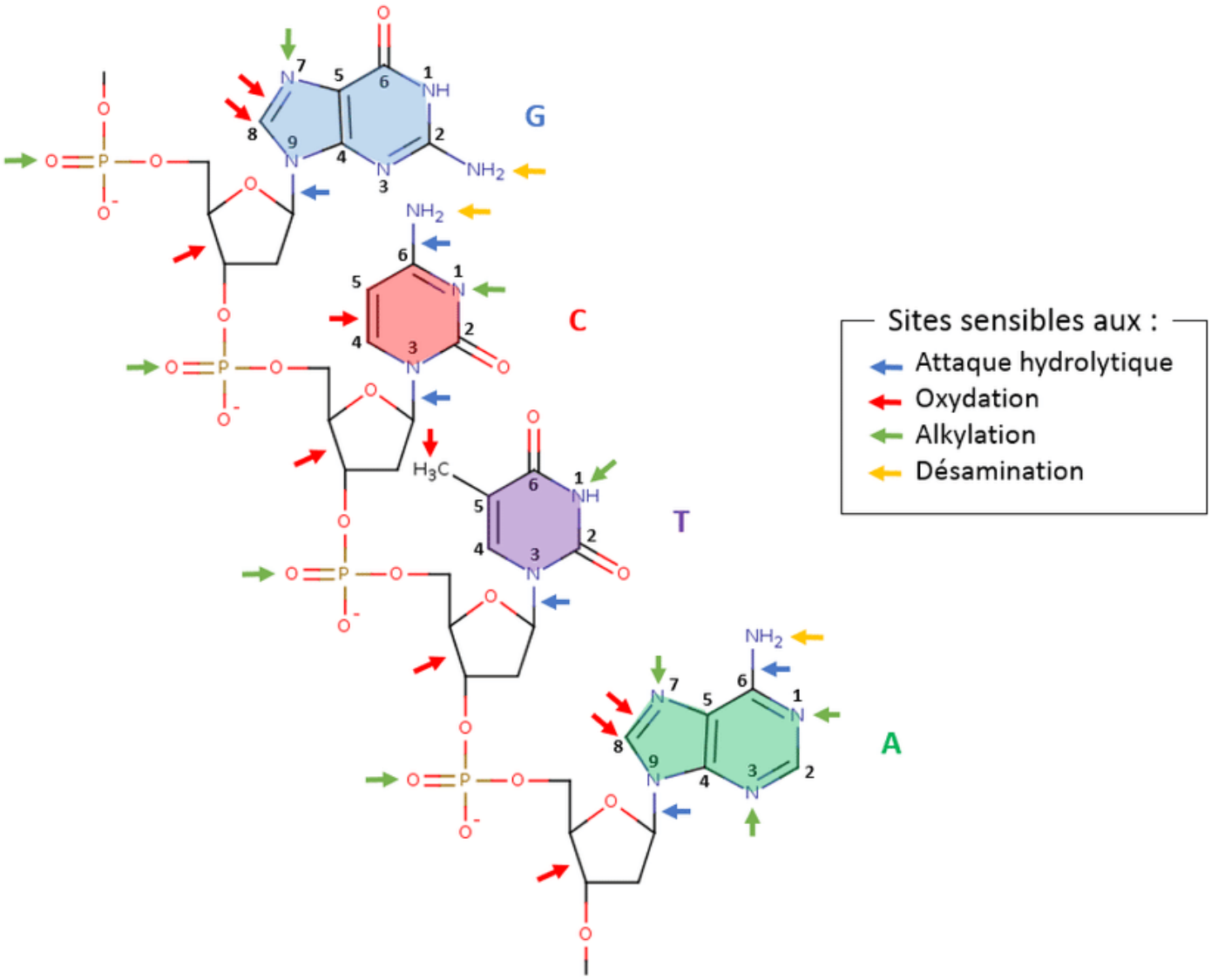


Figure 1: Les dommages des molécules d'ADNa d'après ***(Lindahl, 1993a)*** *(Flèche bleue: hydrolyse de liaison covalente , flèche rouge: oxydation, flèche verte: alkylation, flèche jaune: désamination).*

II) Facteurs influençant la préservation /dégradation de l'ADN pendant l'enfouissement :

La plupart des études ne montrent pas une corrélation stricte entre l'âge des échantillons, la taille des fragments d'ADN et la quantité d'ADN préservé (par exemple, (Sawyer et al., 2012)).

Par contre, l'évolution post-mortem des biomolécules dans les restes organiques dépend de plusieurs facteurs environnementaux. L'étude de ces transformations post-mortem s'appelle la taphonomie, « taphos » signifiant « tombe » en grec. Les facteurs physico-chimiques environnementaux interagissent entre eux de manière complexe pour influencer le devenir de l'ADN (Geigl, 2002). La très grande variabilité des conditions taphonomiques fait qu'il est très difficile d'évaluer solidement la contribution de chaque facteur individuellement. La température est le facteur le plus clairement impliqué. Les basses températures ralentissent les dégradations enzymatiques et chimiques de l'ADN (Smith et al., 2001).

Il a été interpolé qu'une augmentation de 15°C suffirait pour accélérer de 16 fois le taux de dégradation de l'ADN en solution physiologique (Lindahl & Nyberg, 1972). Ainsi, les échantillons provenant de pergélisol, région gelée en permanence, procurent de l'ADNa bien préservé (Keyser-Tracqui et al., 2005). On a ainsi pu analyser de l'ADN de dents de mammouths de plus d'un million d'années provenant du pergélisol (van der Valk et al., 2021). Cet ADN était mieux préservé que dans la plupart d'échantillons de quelques milliers voire centaines d'années provenant des régions tempérées. Dans les grottes, où les variations climatiques sont réduites et les températures relativement stables, une bonne préservation d'ADNa est souvent observée (par ex., (Reich et al., 2010). Par contre, l'obtention de données à partir d'ossements provenant de régions chaudes, voire désertiques, telles que la Syrie ou l'Égypte, est beaucoup moins fréquente (Pruvost et al., 2008). Les régions chaudes et humides, telles que les régions tropicales, semblent encore moins favorables et il n'y a que des très rares données, très peu anciennes produites en provenance de ces régions. Les climats tempérés, continentaux modérés ou méditerranéens sont moins propices à la préservation de l'ADN dans les échantillons anciens que le pergélisol, mais permettent l'obtention de résultats au prix d'efforts plus importants (Pruvost et al., 2008).

Le pH, le potentiel oxydoréducteur, la composition chimique des os et du sol et l'hydrologie sont aussi des facteurs qui pourraient jouer un rôle dans la préservation de l'ADN, mais leur impact exact n'est pas clairement établi et seules des corrélations empiriques sont envisageables. Par exemple, la modélisation de ces processus basée sur une cinétique de premier ordre indique qu'une diminution de pH de 7.4 à 6.4 augmente le taux de dépurination de 3.3 fois et qu'une diminution de la teneur en sel augmente la vitesse de dépurination qui serait 7 fois plus rapide lorsque la concentration en NaCl diminue de 0.1 à 0 M et cela a permis d'interpréter des observations expérimentales (Pruvost, 2007). Après prélèvement, les ossements archéologiques doivent être préservés dans des bonnes conditions et manipulés avec précautions car il a été montré que les os non lavés et fraîchement excavés et non traités contiennent plus d'ADN endogène que les os traités par les procédures standards (Pruvost, 2007).

III) La contamination de l'ADNa :

L'ADNa est dilué dans de l'ADN environnemental. En effet, les extraits d'ADNa peuvent contenir en plus de l'ADN de l'organisme initial l'ADN des microbes et des parasites associés à l'organisme durant sa vie (Côté et al., 2016), l'ADN des organismes qui ont contribué à la décomposition du corps après la mort et l'ADN des organismes du sol qui ont pénétré l'os (Côté et al., 2016; Der Sarkissian et al., 2014; Noonan, 2005; Sampietro et al., 2006). Ceci augmente la complexité de l'analyse de l'ADNa. De plus, à cause de sa faible abondance, l'ADNa peut facilement être contaminé avec de l'ADN moderne et plus le nombre de molécules d'ADNa est faible plus la susceptibilité à la contamination est élevée. Ceci peut produire des résultats erronés ce qui nous oblige de travailler dans un laboratoire de haut confinement.

1) Contamination avant enfouissement :

Après la mort de l'organisme, les bactéries et les micro-organismes se multiplient et participent à sa décomposition et le colonisent (Allentoft et al., 2012; Lindahl, 1993).

Des traces nucléotidiques de ces micro-organismes pourraient être préservées dans l'échantillon fossile mais c'est très difficile à établir clairement car il n'est pas aisé de distinguer les séquences qui pourraient provenir de microorganismes environnementaux anciens ou beaucoup plus récents car la diffusion dans l'ossement de l'ADN contenu dans le sol se produit tout au long de la période d'enfouissement. En effet, la principale source de contamination provient du sédiment dans lequel les restes fossiles sont préservés.

2) Contamination post-fouille :

Du moment du prélèvement des ossements par les fouilleurs jusqu'au laboratoire, l'échantillon ancien peut subir différents types de contamination allant du transfert de cellules de l'épiderme du préleveur jusqu'aux contaminations qui peuvent avoir lieu durant les différentes étapes d'analyse au laboratoire paléogénétique. Quand l'échantillon provient de restes humains, la contamination post-fouille peut biaiser significativement les résultats d'ADNa.
Quatre sortes de contamination peuvent avoir lieu dans le laboratoire.

a) La contamination au laboratoire :

Chaque objet dans le laboratoire peut être contaminé par le manipulateur, c'est pourquoi il est nécessaire de manipuler les échantillons avec des gants javellisés comme il faut javelliser aussi la paillasse et irradier avec des rayons UV les pipettes à courte distance après chaque utilisation (Champlot et al., 2010). L'irradiation entraîne la fragmentation de l'ADN mais elle n'aboutit pas à sa destruction complètement. Bien que les quantités des ces fragments soient minimes, leur présence dans un tube contenant l'extrait d'un fossile qui n'a que peu ou pas d'ADN, peut poser un problème.

b) Les contaminations croisées :

Au moment de la manipulation des extraits fossiles, il est possible de contaminer les échantillons entre eux. Un autre danger est la contamination appelée « Carry over contamination » qui résulte d'un mélange dans le laboratoire de molécules d'ADN amplifiées avec les extraits d'ADNa ce qui conduit à des résultats erronés des analyses expérimentales. En effet, une seule réaction de PCR est suffisante peut produire jusqu'à 10^{13} molécules d'ADN dont une trace peut suffire pour contaminer la PCR suivante. Ces contaminants peuvent facilement être disséminés à travers les laboratoires de biologie moléculaire sous forme d'aérosol. Une micro-goutte d'aérosol dans une réaction de PCR peut contenir plus de molécules d'ADN que dans un gramme de reste fossile (Pruvost et al., 2005).
Lorsque les molécules contaminantes disséminées sur les paillasses correspondent aux molécules ciblées dans l'expérience, le danger est encore plus important. De plus, puisque les molécules contaminantes ne présentent pas de lésions diagénétiques, elles seront amplifiées plus efficacement que les autres molécules endogènes contenues dans l'extrait fossile.

c) Les contaminations des réactifs biologiques :

Le mode de production des réactifs biologiques par les sociétés biotechnologiques pose un problème supplémentaire aux défis de l'analyse de l'ADNa. Les réactifs utilisés pendant l'extraction, la purification et l'amplification d'ADN peuvent contenir de l'ADN actuel d'origine humaine, bactérienne ou animale. Les enzymes sont stabilisées avec des protéines comme l'albumine sérique bovine.

Cette protéine est le plus souvent purifiée à partir de sang bovin et peut ainsi contenir des quantités d'ADN bovin comparable à celles contenues dans certains extraits fossiles.
Ainsi, des séquences anciennes bovines sont souvent contaminées avec celle de bovins modernes surtout les bovins occidentaux à partir desquels les produits ont été produits (Champlot et al., 2010; Leonard et al., 2007). Il a été montré que les résultats obtenus à partir de contrôles négatifs, c'est-à-dire ne contenant pas d'ADN, sont semblables à ceux trouvés dans les publications sur les bovins anciens analysés par PCR ciblée (Champlot et al., 2010). Le même problème a été constaté pour des données porcines à cause de la gélatine d'origine porcine utilisée pour stabiliser la Taq polymérase (Champlot et al., 2010). Il est possible de minimiser ces problèmes en décontaminant les réactifs (Champlot et al., 2010). Malheureusement, ces procédures exigeantes sont rarement appliquées ce qui entache d'incertitudes un très grand nombre de données publiées obtenues par PCR ciblée.

IV) Notion de niches moléculaires :

L'hypothèse a été avancée qu'il existerait des niches moléculaires de préservation à l'intérieur de l'os qui permettent la protection de l'ADN contre la dégradation (Geigl, 2002). Ces niches moléculaires de préservation posséderaient des propriétés biologiques et physico-chimiques particulièrement favorables à la préservation de l'ADN. En effet, alors que la chimie de l'ADN après la mort d'un organisme est encore soumise aux règles de la chimie de la solution aqueuse, dans les fossiles plus anciens, il a été observé que l'ADN peut être préservé sous forme insoluble, et ceci, en raison de son adsorption à la phase minérale et aux substances humiques et autres matières organiques insolubles ce qui ralentirait les réactions de dégradation (Geigl, 2002). Un type de niche de préservation seraient des amas de cristaux d'hydroxyapatite de l'os non-affectés par des changements diagénétiques et auxquels l'ADN est adsorbé (Salamon et al., 2005). Plus l'os est dense et compact, plus sa richesse en niches moléculaires semble élevée. Ainsi l'os pétreux, l'os le plus dense et compact du squelette, est la source la plus favorable pour la préservation de l'ADN (Geigl & Grange, 2018; Pinhasi et al., 2015).

V) Potentiel inhibiteur des échantillons :

Les extraits d'échantillons fossiles contiennent, en plus d'acides nucléiques endogènes et d'acides nucléiques d'origines diverses appelés ainsi des molécules environnementales, des substances chimiques du sol comme les acides humiques et fulviques. Bien que les extraits d'ADN soient purifiés, ces substances chimiques peuvent échapper à la purification. La difficulté principale est de concilier la purification de molécules d'ADN de très petite taille avec l'élimination du maximum de molécules inhibitrices. Le compromis parfait entre ces deux contraintes est difficilement atteignable.
L'obtention des séquences génomiques d'ADNa nécessitent des réactions enzymatiques diverses. Les enzymes utilisées, comme la Taq ADN polymérase et l'ADN ligase, indispensable pour la construction des banques génomiques et l'amplification des molécules d'ADN par PCR, sont sensibles à la présence de molécules environnementales qui interfèrent avec leur activité, inhibant ainsi les réactions qu'elles catalysent.
Les impuretés contenues dans les extraits d'ADN fossile sont considérées comme des facteurs limitants de l'étude de l'ADNa.

Ces inhibiteurs sont mal connus. Ils peuvent provenir de l'échantillon lui-même ou de l'environnement où l'échantillon était excavé. Ces inhibiteurs peuvent cibler l'enzyme impliqué dans la réaction ou interagir avec les éléments nécessaires à l'équilibre physico-chimique du milieu réactionnel (Eckhart et al., 2000). Les acides humiques, fulviques et certains ions complexes comme le Fer (Fe2+) sont présent en grande quantité dans les extraits fossiles (Hagelberg, 1991; Hummel, 1992) mais leurs pouvoirs inhibiteurs n'a jamais été mis en évidence clairement bien qu'il soit hautement suspecté.

L'ADN une fois purifié peut être analysé soit par PCR ciblée, l'approche qui a été utilisée pendant les 20 premières années de la paléogénétique, soit après construction des banques génomiques par séquençage de l'ensemble de l'ADN, le plus souvent en utilisant la technologie Illumina. Dans ce cas, on peut soit choisir de tout séquencer, approche dite « shotgun » (séquençage aléatoire), soit choisir de ne séquencer que des régions spécifiques qu'on aura capturées en solution avec des appâts. J'ai utilisé ces différentes approches dans mon laboratoire d'accueil lors de ma thèse et j'ai produit essentiellement des résultats avec la technologie Illumina, aussi bien par capture que par shotgun (voir plus loin).

Chapitre II: Méthodes d’analyse phylogénétique

L’analyse des séquences d’ADNa est une étape clé et critique. Pour ne pas avoir des interprétations biaisées, il est nécessaire de paramétrer ces méthodes avec précautions et fidélité. Les paramétrages des méthodes d’analyse d’ADNa dépendent de la question posée et du type d’interprétations qu’on cherche à en tirer. Il y a, d’une part, les aspects liés aux traitements des millions de séquences courtes produites lors d’un séquençage NGS et d’autre part les aspects de traitement des séquences de mitogénomes ou de génomes une fois que les informations de séquences brutes sont converties soit en séquences de mitogénomes, soit en fichier compilant les variations génomiques identifiés. Je traiterai ici les aspects liés à la deuxième partie de ces analyses, avec un accent mis sur les méthodes d’analyse phylogénétique des mitogénomes. L’analyse phylogénétique permet d’estimer le taux de filiation des espèces et de déterminer leurs divergences en se basant sur différentes séquences d’ADN. Ces estimations doivent être vérifiées par des tests statistiques qui doivent être significatifs et reproductibles.

Dans la partie suivante je vais mentionner les différentes méthodes d’études phylogénétiques utilisées dans le domaine de la paléogénomique.

I) Alignement multiples des séquences :

Une fois que les données de séquençage sont produites et converties en séquences, par exemple des mitogénomes, un alignement multiple des séquences doit être effectué. L’alignement des séquences permet de faire ressortir les régions homologues et non homologues. L'objectif de l'alignement est d’identifier les zones de concordance. Ces alignements sont réalisés par des programmes informatiques dont l'objectif est de maximiser le nombre de coïncidences entre les nucléotides des différentes séquences.

Il y a deux types d’alignement : (1) L’alignement global est une méthode qui permet d’aligner les séquences sur la totalité de leur longueur. (2) l’alignement local se restreint à des régions limitées dans lesquelles la similarité est forte, à l'exclusion du reste des séquences. Avec des séquences très voisines, les résultats obtenus par les méthodes d'alignement local ou global sont très proches. Pour cette raison, les méthodes d'alignement local, plus flexibles, sont plus souvent utilisées aujourd'hui. Elles permettent à la fois d'aligner des séquences localement ou globalement similaires. Pour aligner le millier de mitogénomes que j’ai analysés, j’ai utilisé le programme d’alignement « Muscle » qui offre un bon compromis entre puissance et fiabilité (Edgar, 2004). Il est toutefois important de bien vérifier manuellement la fidélité des alignements, surtout lorsque l’on effectue des alignements de groupes de séquences de manière itérative, ce que j’ai fait régulièrement car j’ai petit à petit augmenté le nombre de séquences produites et analysées.

II) Analyse phylogénétique :

Il est nécessaire d’analyser les séquences d’ADNa avec précaution et de déterminer à priori les informations à tirer et donc les paramètres des méthodes utilisées pour éviter ainsi des résultats et des interprétations biaisées. Je vais expliquer par la suite les spécificités des méthodes d’analyse phylogénétiques, leurs paramètres et leurs limites pour l’analyse de la variabilité nucléotidique au sein ou entre les espèces.

L’analyse de phylogénie moléculaire permet d’estimer l’histoire évolutive des organismes vivants et donc estimer leurs liens de parenté. Ces liens de parenté ou phylogénie sont généralement représentés sous forme d’arbres ou réseaux phylogénétiques qui représentent des estimations de l’histoire évolutive ce qui nous oblige de tester par des tests statistiques la fiabilité des ramifications de l’arbre ou du réseau phylogénétique.

La construction d’arbres phylogénétiques est une étape complexe car l’analyse phylogénétique de différents taxons peut aboutir à l’obtention de plusieurs arbres phylogénétiques différents (Nei, 1996). En effet, un arbre phylogénétique est un arbre schématique qui montre les relations de parenté entre des groupes d'êtres vivants. Chacun des nœuds de l'arbre représente l'ancêtre commun de ses descendants. Ces arbres peuvent être enracinés ou pas. La longueur des branches représente la distance génétique évolutive entre les différents individus analysés. Cette distance est mesurée en nombre de substitutions nucléotidiques accumulées au cours du temps. On peut ainsi estimer l’âge du dernier ancêtre commun à plusieurs séquences. Les dates indiquées au niveau des nœuds représentent l’estimation du temps passé depuis la divergence entre deux ou plusieurs branches. Pour les arbres datés la longueur des branches reflète le temps écoulé.

Pour compléter la construction de l’arbre phylogénétique, une analyse de robustesse doit être effectuée, d’habitude par ‘boostrap’ (Efron et al., 1996; Felsenstein, 1985; Lemoine et al., 2018). Cette technique consiste à échantillonner les positions de l’alignement pour relancer la construction phylogénétique de manière itérative puis de comparer les résultats obtenus après plusieurs répétitions. Le résultat est présenté sous la forme d’un arbre consensus dans lequel figurent les regroupements. Une valeur de ‘boostrap’ (de 0 à 100%) est associée à chaque branche indiquant le nombre de fois que cette branche a été retrouvée au fil des répétitions et juger ainsi de leur crédibilité. La valeur de Boostrap indique alors une évaluation de la résistance d’un nœud à la perturbation des données. La branche est considérée robuste quand la valeur est supérieure ou égale à 95%.

III) Les méthodes de construction d’arbres phylogénétiques :

Il existe 3 méthodes de construction d’arbres phylogénétiques :

1) Les méthodes de distance :

Ces méthodes sont basées sur les similarités entre paires de séquences. Les substitutions vont déterminer la distance évolutive entre les différentes séquences et va former une matrice qui va servir pour la construction de l’arbre phylogénétique. Un outil de construction d’arbres phylogénétique basé sur cette méthode est UPGMA (Unweighted Pair Group Method with Arithmetic Mean) mais il a vite été délaissé au profit de la méthode NJ (Neighbour Joining) qui est plus adaptée aux études phylogéniques moléculaires et utilisée pour faire des arbres de plusieurs milliers de séquences.L’algorithme Neighbor Joining (NJ) n’assume pas l’horloge moléculaire bien qu’il fonctionne mieux quand l’horloge est respectée (N Saitou & M Nei, 1987). La méthode autorise un taux d’évolution différent entre les lignées étudiées et utilise une approche de regroupement combinée à une approximation efficace du principe d’évolution minimum.

Elle permet d’inférer des phylogénies sur des centaines de séquences et elle permet de trouver la vraie phylogénie si la matrice de distance est une réflexion exacte de la phylogénie.

2) Méthodes basées sur des caractères :

L'approche basée sur les caractères correspond à des méthodes plus robustes statistiquement que les méthodes de distances mais elles sont plus lentes. Cette approche regroupe les méthodes de parcimonie, de maximum de vraisemblance et les méthodes bayésiennes.

a) Les méthodes de parcimonie :

Ces méthodes estiment le nombre minimal de changements pour expliquer les différences observées entre les séquences. Ces méthodes sont très appréciées car elles sont rapides en temps de calcul mais moins précises. La méthode de parcimonie consiste à rechercher parmi tous les arbres possibles et toutes les séquences possibles de nœuds ancestraux, la combinaison qui minimise le nombre d'évènements mutationnels présents dans l'arbre reconstruit. Elle s'appuie sur deux hypothèses principales. D'abord, tous les sites évoluent indépendamment les uns des autres et la vitesse d'évolution est lente et constante à travers les lignées évolutives.

Pour rechercher l'arbre le plus parcimonieux, plusieurs approches peuvent être utilisées. Lorsque le nombre de taxons est inférieur à 10, il est possible d'effectuer une recherche exhaustive. Dans le cas contraire, il faut se contenter des constructions heuristiques d'un arbre parcimonieux ou utiliser une approche de type « branch-and-bound » (séparation et évaluation). La recherche exhaustive examine toutes les topologies possibles par une approche gloutonne. Au début, elle regroupe trois espèces choisies arbitrairement ou selon des critères heuristiques. Puis elle ajoute une espèce sur la branche conduisant à une longueur d'arbre minimale. Une autre espèce est ensuite ajoutée sur l'une des cinq branches possibles de l'arbre. Les itérations de l'algorithme se terminent lorsque toutes les espèces ont été placées. Les algorithmes de Fitch (1971) et Sankoff (1975) sont parmi les plus connus pour cela.

Dans les constructions heuristiques d'un arbre parcimonieux, des algorithmes itératifs s'ajoutent à la construction gloutonne de la recherche exhaustive. Ces algorithmes sont utilisés lorsque le nombre de taxons est supérieur à 20. Les constructions heuristiques partent d'un arbre initial et essaient de trouver un arbre plus parcimonieux en effectuant des réarrangements de branches. L'arbre initial est souvent inféré par des méthodes qui présentent une solution acceptable en un temps très court. Mais ces approches heuristiques ne couvrent généralement que les topologies similaires à la topologie initiale. Par ailleurs, il existe plusieurs stratégies de réarrangement ou de transformation de branches. Les stratégies les plus connues sont: i) le réarrangement local (Nearest-Neighbor Interchanges "NNI"), ii) le réarrangement global (Subtree Pruning and Regrafting "SPR"), iii) la bissection et reconnexion d'arbres (Tree Bisection and Reconnection "TBR"), iv) les fusions d'arbres (tree-fusing), v) les algorithmes génétiques (Genetic algorithms).

b) Les méthodes de maximum de vraisemblance (Maximum Likelihood- ML) :

Ces méthodes font partie des méthodes dites de caractère(s). Elle repose sur un ou plusieurs caractères à étudier. Il s'agit d'une méthode probabiliste qui nécessite un modèle d'évolution. Le choix de ce modèle est crucial pour la qualité de l'arbre obtenu.

On dit qu'il convient de l'utiliser à partir du moment où le nombre de caractères analysés est supérieur à la moitié du nombre de séquences analysées, sinon la reconstruction est considérée comme incorrecte. Elle est souvent décrite comme étant la meilleure méthode, c'est-à-dire la plus efficace pour trouver l'arbre le plus proche de la réalité. Le maximum de vraisemblance est considéré comme plus fiable que les méthodes de distances et de parcimonie. C'est la méthode qui conduit au résultat le plus proche de l'arbre évolutif réel en théorie. Il permet également d'appliquer les différents modèles d'évolution et d'estimer la longueur des branches en fonction du changement évolutif. Son désavantage se situe au niveau des temps de calculs qui peuvent être extrêmement longs bien que les améliorations de l'efficacité des algorithmes et de la puissance des ordinateurs a permis de rendre ces approches utilisables dans des temps raisonnables, même avec des centaines, voire des milliers de séquences (Guindon et al., 2010; Stamatakis, 2014).

Deux hypothèses principales sont émises lors de l'utilisation de cette méthode: i) selon l'arbre donné, l'évolution est indépendante sur les différents sites ii) l'évolution est indépendante selon les lignées.Le maximum de vraisemblance repose sur le calcul indépendant de la vraisemblance sur chaque site. Il recherche la vraisemblance des données sous différentes hypothèses évolutives d'un modèle d'évolution et en retient les hypothèses qui rendent cette vraisemblance maximale. Le maximum de vraisemblance cherche donc à trouver l'arbre dont la vraisemblance est maximale pour les séquences observées et le modèle d'évolution choisi. Pour trouver l'arbre le plus vraisemblable, les bases de toutes les séquences à chaque site sont considérées séparément et le logarithme de la vraisemblance est calculé pour une topologie donnée, en utilisant un modèle d'évolution particulier. Le logarithme de la vraisemblance est cumulé sur tous les sites et sa somme est maximisée pour estimer la longueur des branches de l'arbre. Cette procédure est répétée pour toutes les topologies possibles et la topologie ayant la plus grande vraisemblance est choisie.

c) Les méthodes bayésiennes :

Les méthodes bayésiennes sont similaires au maximum de vraisemblance. Elles diffèrent seulement par l'utilisation d'une distribution à priori des données qui sont en train d'être inférée. Elles sont plus rapides et permettent de traiter plus de taxons. Cette méthode utilise le théorème de Bayes. La probabilité postérieure d'un arbre pouvant être interprétée comme la probabilité que cet arbre soit vrai sachant les données, les inférences sont réalisées à partir de la distribution de probabilité postérieure des différents arbres évalués au cours de l'analyse. De manière analogue à la méthode du maximum de vraisemblance, l'arbre de probabilité postérieure maximale peut ainsi être déterminé facilement.

Mais l'approche bayésienne comporte certains problèmes comme la nécessité d'avoir une distribution à priori sur les hypothèses. Ceci pose un problème si nous n'en avons aucune ou si elle est controversée. La méthode bayésienne est donc souvent associée à un model MCMC (Monte Carlo Markov Chain). L'idée sous-jacente aux MCMC est qu'une chaîne de Markov, prenant la forme d'une marche guidée à travers l'espace multidimensionnel des paramètres, peut être utilisée pour estimer une distribution de probabilité en échantillonnant les valeurs de ces paramètres de façon périodique.

L’approximation de la distribution sera d’autant plus exacte que le nombre effectués par la chaîne de Markov sera élevé .
(http://www.info2.uqam.ca/~makarenkov_v/BIF7002/Rapport_Eric_Alexandre_2009/Rapport_HTML.html).

BEAST (Bayesian Evolutionary Analysis by Sampling Trees) (Drummond, 2005; Drummond & Rambaut, 2007) est l’algorithme le plus utilisé pour les analyses de la phylogénie. Il permet l’obtention d’arbres phylogénétiques en estimant les temps de divergence en tenant compte des dates de carbones 14 des échantillons fossiles et en fonction du taux de mutation entre les espèces. La méthode sur laquelle est basée le programme est une méthode probabiliste qui calcule les probabilités postérieures des arbres phylogénétiques en combinant la probabilité à priori avec la fonction de vraisemblance.

BEAST utilise la méthode MCMC pour faire la moyenne sur des arbres, de sorte que chaque arbre soit pondéré proportionnellement à sa probabilité postérieure. Nous utilisons un programme Beauti qui est un GUI (Graphical user-interface) pour la mise en place d'analyses standard et la préparation du fichier XML introduit par Beast et un ensemble de programmes pour analyser les résultats que je vais montrer par la suite. L’avantage de la méthode MCMC est de permettre l’implémentation de modèles de séquences complexes, tout en utilisant un nombre élevé de paramètres en un temps rapide de calcul.
Les régions génomiques ayant des caractéristiques similaires peuvent être regroupées et donc il est possible d’effectuer l’analyse phylogénétique en appliquant à chaque groupe un modèle d’évolution différent.

Le « Skyline plot » est une application intégrée dans le programme. Il s'est avéré très utile comme outil de sélection de modèle utilisé pour indiquer le modèle démographique le plus approprié pour un ensemble de données donné (Pybus & Rambaut, 2002). Le modèle démographique est basé sur la théorie de coalescence. Le modèle estime le nombre d’individus ancestraux contribuant à la diversité d’un échantillon de taille déterminée d’individus et l’évolution de ce nombre d’individus au cours du temps.
En effet, alors que le processus Wright-Fisher (voir plus loin) décrit naturellement l'évolution des séquences au sein des populations une génération après l'autre, les données génétiques des populations représentent généralement des individus échantillonnés à un moment donné. À des fins d'inférence, il convient donc de modéliser l'histoire du matériel génétique qui a donné naissance à l'échantillon.
La modélisation de l'ascendance d'un échantillon (également appelée généalogie) se fait généralement à rebours dans le temps, car chaque locus trouve un ancêtre commun dans le passé, jusqu'à l'ancêtre commun le plus récent (MRCA) de l'échantillon. La fusion de deux lignées dans le passé est appelée un événement de coalescence, et l'ensemble des outils mathématiques décrivant ce processus sous une variété de modèles démographiques est appelé la théorie de la coalescence. Kingman (Kingman JFC 1982 The coalescent Stoch process Appl) a d'abord décrit la coalescence standard, le modèle généalogique correspondant au modèle de Wright-Fisher. La coalescence standard est donc également appelée coalescence de Kingman.

3) L’Horloge moléculaire :

En génétique, l'hypothèse de l'horloge moléculaire est une hypothèse selon laquelle les mutations génétiques s'accumulent dans un génome à une vitesse constante. Elle permet ainsi théoriquement, en reliant le taux de mutation des gènes au rythme de diversification de leur espèce, d'établir une échelle chronologique, d'évolution et de lien des espèces entre elles.

La réconciliation de l'hypothèse de l'horloge moléculaire et de la théorie Darwinienne a été amorcée vers la fin des années 1960 par les travaux de Motoo Kimura, Allan Wilson et Vincent Sarich, et de l'élaboration de la théorie neutraliste de l'évolution. Celle-ci postule que la majorité des mutations génétiques accumulées et conservées sont neutres, c'est-à-dire qu'elles ne confèrent à l'individu subissant la mutation ni un avantage sélectif marqué ni un désavantage. L'horloge moléculaire a permis à de nombreux chercheurs de dater des événements de spéciations à l'aide de méthodes phylogénétiques de plus en plus développées.

Toutefois, alors que la quantité de données génétiques augmentait et que les méthodes statistiques se raffinaient, il devient de plus en plus clair que l'horloge moléculaire n'était pas valide, du moins dans certaines parties de la phylogénie des êtres vivants. Depuis, plusieurs modèles ont été proposés afin d'assouplir l'horloge moléculaire par des modèles statistiques plus sophistiqués (maximum de vraisemblance, méthodes bayésiennes), dits d'horloge moléculaire relaxée. Ces modèles ont pour avantage de donner des temps de divergence entre espèces plus précis car ils assument qu’il y a des variations de taux de substitutions entre différentes lignées. Cette relaxation du principe est d’un point de vue biologique beaucoup plus réaliste même s’il implique plus de puissance de calcul, et plus en accord avec les données paléontologiques (Soubrier et al., 2012).

Les taux de mutation diffèrent entre les espèces et même entre les différentes régions du génome d'une seule espèce. Ces différents taux de substitution nucléotidique sont mesurés en substitutions (mutations fixes) par paire de bases par génération. Par exemple, les mutations dans l'ADN intergénique ou non codant ont tendance à s'accumuler plus rapidement que les mutations dans l'ADN activement utilisé dans l'organisme (expression génique). Cela n'est pas nécessairement dû à un taux de mutation plus élevé, mais à des niveaux plus bas de sélection. Une région qui mute à une vitesse prévisible est candidate à l’utilisation comme horloge moléculaire. En fait, le taux de mutation d'un organisme peut changer en réponse au stress environnemental. Par exemple, les rayons UV endommagent l'ADN (Rastogi et al., 2010), ce qui peut entraîner des tentatives de réparation de l'ADN susceptibles d'entraîner des erreurs.

De plus, plusieurs événements affectant la démographie d’une population tels que les dérives génétiques, les goulots d’étranglement ou des erreurs de calibration de l’horloge peuvent biaisée l’estimation des dates de divergence. L’utilisation de séquences prélevées à des dates connues le long d’un processus évolutif permet d’estimer de manière beaucoup plus fiable la vitesse d’évolution des séquences et donc de mieux dater les événements de divergence dans l’arbre. Ainsi l’approche paléogénomique ouvre la possibilité de pouvoir dater avec plus de précision les radiations phylogénétiques. Elle permet aussi d’avoir une vision plus large des branches évolutives disparues.

4) Incertitude de la construction d’arbre phylogénétique :

La fiabilité de la topologie d’un arbre repose sur la robustesse des nœuds de l’arbre mesurée par des indices statistiques et sur les modèles d’évolution appliqués aux données.
Il existe un certain nombre d’artéfacts qui affectent les reconstructions phylogénétiques et ils sont particulièrement importants quand les séquences sont des séquences qui évoluent vite ou qui évoluent différemment de celles chez les autres organismes parce que cela va augmenter les chances d’obtenir des événements de convergence évolutive, avec pour résultats des espèces non apparentées présentant des traits morphologiques et/ou fonctionnels équivalents. Les événements de tri de lignées incomplètes (Incomplete Ligneage Sorting ILS), qui correspondent à des événements de dérive génétique qui n’ont pas eu assez de temps pour se fixer avant qu’une nouvelle divergence ne se produise, vont induire l’apparition de nœuds tardifs artéfactuels en bout de branche (Madisson, 2006). En plus les accumulations localisées de mutations dans une région spécifique, comme c’est le cas dans la région hypervariable mitochondriale, induisent un signal évolutif aléatoire qui peut biaiser la typologie de l’arbre en créant des points de divergence précoce entre espèce (Moreira & Philippe, 2000). Les phylogénies ne sont donc pas reconstruites avec certitude et il peut y avoir des erreurs ou des incertitudes dans les arbres inférés. C’est pourquoi il faut être prudent et vigilant lorsque l’on fait de la phylogénie.

IV) La dérive génétique :

Le processus le plus simple d'évolution des allèles au sein d'une seule population est appelé le modèle de Wright-Fisher. Il décrit l'évolution des allèles dans une population de taille fixe et constante, où tous les allèles ont la même aptitude, et donc la même chance d'être transmis à la génération suivante (ce qu’on appelle évolution neutre). On suppose que la population est panmictique, c'est-à-dire que les individus s'accouplent au hasard. Le temps est discrétisé en générations non chevauchantes de sorte que les allèles de la génération actuelle sont un échantillon aléatoire des allèles de la génération précédente, sans que de nouveaux allèles soient générés par mutation. Dans de telles conditions, les fréquences alléliques n'évoluent qu'à cause de la stochasticité dans l'échantillonnage des gamètes qui contribuera à la prochaine génération, un processus appelé dérive génétique.

Motoo Kimura a établi la théorie de la dérive génétique en 1968. Parce que les populations sont de taille finie, les allèles seront échantillonnés à leur fréquence réelle en moyenne seulement et le destin ultime de tout allèle en absence de sélection est soit d'atteindre la fréquence zéro dans la population et d'être perdu, alors que par hasard aucun individu porteur de cet allèle n'a de descendant dans la génération suivante ou de se fixer lorsque tous les autres allèles ont été perdus. Le temps jusqu'à la fixation dépend de la taille de la population: des populations plus petites montreront un effet d'échantillonnage plus fort et des temps de fixation plus courts.

Cette situation peut se produire au moment de l'apparition d'une espèce, ou après un goulot d'étranglement (quand une grande partie d'une espèce a disparu, à la suite de phénomènes épidémiques ou d'un changement climatique ou lors de la fondation d’une population à partir d’un nombre limité d’individus de la population d’origine).

La dérive génétique concerne tous les allèles même les allèles neutres. Elle peut faire disparaître un allèle favorable et fixer un allèle défavorable dans une population ce qui est fréquent pour des populations aux tailles très réduites. Chaque population isolée est en effet caractérisée par un génome spécifique, qui diffère de celui d'autres populations de la même espèce par la fréquence, voire l'existence de certains allèles et gènes. L'accumulation des modifications du génome par la dérive génétique peut aboutir à l'appauvrissement de la diversité génétique et provoquer la disparition d'une population, voire d'une espèce. Inversement, ces modifications peuvent engendrer une nouvelle espèce.

V) Migration et flux de gènes :

Le flux génétique est le transfert d'allèles par transfert de gènes d'une population à l'autre. En biologie de l'évolution, il se réfère à l'échange de matériel génétique entre deux populations d'une espèce, bien que l'essor des données génomiques a mis en évidence de plus en plus d'introgressions interspécifiques. Ce flux est responsable de changements significatifs dans le génome.
La migration est responsable en partie du flux de gènes car elle peut entraîner l'introduction de nouveau matériel génétique dans le patrimoine génétique établi d'une espèce ou d'une population particulière comme elle peut entraîner une perte de matériel génétique. Il y a un certain nombre de facteurs qui affectent le taux de flux génétique entre différentes populations. L'un des facteurs les plus importants est la mobilité.

Le brassage génétique (admixture), l'échange de matériel génétique en conséquence du flux de gènes entre les populations différenciées, a longtemps été reconnu comme un facteur important de l'évolution, par exemple chez les plantes (Gross & Rieseberg, 2005) ou les oiseaux (Rheindt & Edwards, 2011). La découverte de flux de gènes entre l'humain moderne et d'autres hominines comme les Néandertaliens a attiré l'attention sur ce sujet (Green et al., 2010; Kuhlwilm et al., 2016). L'admixture de populations à différentes échelles temporelle est un aspect bien établi de l'évolution humaine et de l'histoire de la population (Reich et al., 2010), qui utilise des méthodes reposant sur l'ADNa ou des déductions tirées de la variation génétique actuelle.

L'impact sur le génome dépendra de la divergence et de la diversité génétiques, des facteurs stochastiques et de l'environnement écologique des espèces concernées. En règle générale, l'introgression de matériel génétique de populations étroitement apparentées sera (principalement) neutre dans la population réceptrice, comme la plupart des autres modifications génétiques, bien que pour des populations vivant dans des environnements différents, on peut observer des introgressions adaptatives. Cependant, les changements génétiques qui différencient les populations les unes des autres peuvent avoir une gamme d'effets imprévus lorsqu'ils sont introduits soudainement par le flux de gènes.
En effet, un continuum de forces sélectives agit sur l'ADN introduit, allant des avantages adaptatifs causés par la diversité introduite à une exposition neutre à la sélection négative et à la dérive génétique en raison d'effets délétères sur la population réceptrice (Figure 2).
Il est connu par exemple que les allèles archaïques à basse fréquence contribuent à la variation phénotypique commune chez les humains actuels, en ce qui concerne la couleur de la peau et des cheveux, des phénotypes comportementaux tels que le chronotype (Dannemann & Kelso, 2017).

Dans certains cas, les différences génétiques apparues après la divergence des populations sont mal tolérées sur leurs fonds génétiques après l'introgression due aux incompatibilités hybrides de Dobzhansky-Muller.
Une sélection fortement négative peut se produire, par conséquence, dans des parties spécifiques du génome, permettant l'émergence de grandes régions génomiques appauvries en ascendance archaïque (Figure2). Ces régions sont appelées «déserts introgressifs». Il a été montré que le chromosome X, par exemple, est appauvri en signatures d'introgression chez l'Humain, les chimpanzés et les bonobos (cité dans (Fontsere et al., 2019)), à cause probablement de balayages sélectifs répétés sur ce chromosome qui est, pour sa majeure partie, sous une forme haploïde chez les mâles mais aussi fonctionnellement haploïde dans les cellules des femelles des organismes qui subissent l'inactivation du chromosome X.

Les conséquences fonctionnelles de la variation génomique nouvellement introduite peuvent être bénéfiques pour la population réceptive du fait d'une sélection positive (Figure2), comme le cas chez l'homme moderne (Dannemann & Kelso, 2017; Racimo et al., 2019). Souvent, des gènes ayant des fonctions dans le système immunitaire ont ainsi été impliqués dans des introgressions adaptatifs chez l'Humain et le chimpanzé, reflétant l'adaptation aux agents pathogènes de l'environnement (Dannemann et al., 2016; Nye et al., 2018). De multiples événements d'introgression complexe entre différentes espèces de *Bos* ont dû permettre certaines adaptations (Chen et al., 2018; Wu et al., 2018).

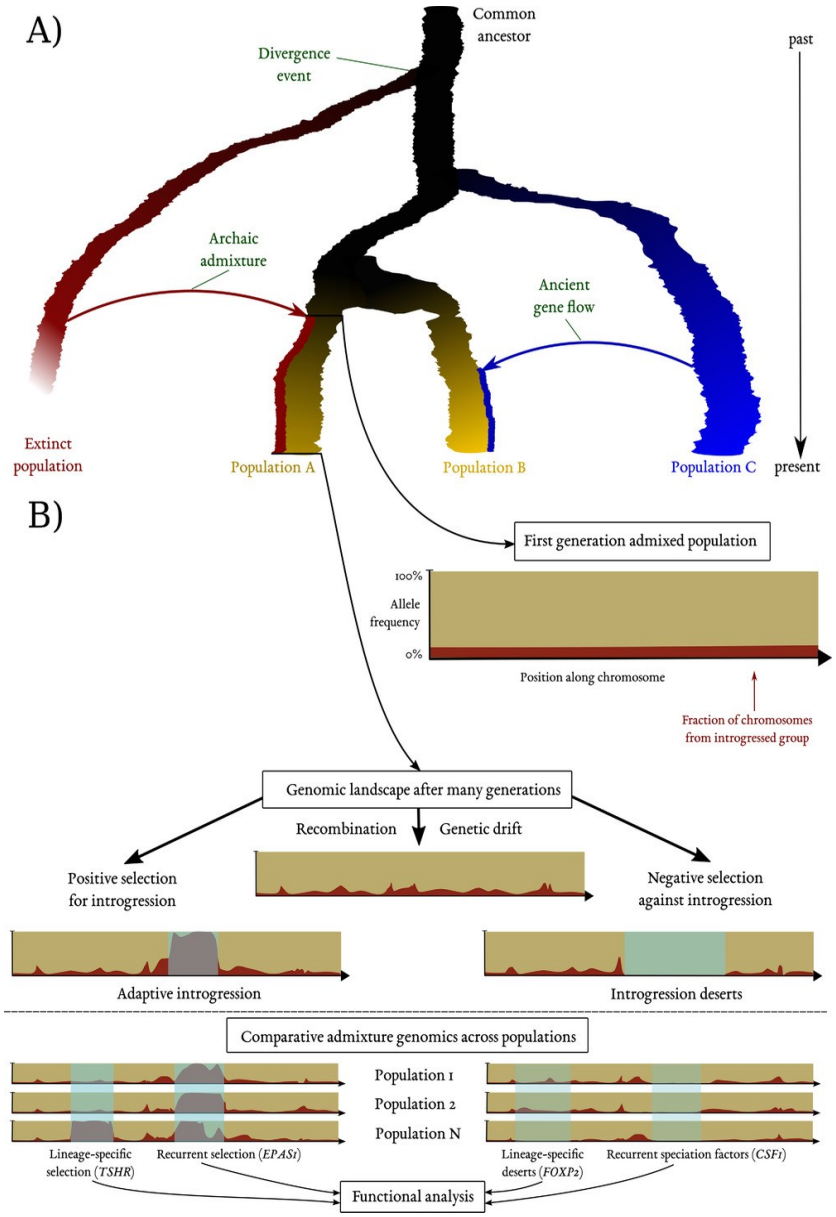


Figure 2: Schéma explicatif de flux génétique et ses conséquences sur les fréquences alléliques des populations (Fontsere et al., 2019).

En l'absence de panmixie, les échanges génétiques se produisent plus souvent entre certains individus, entraînant une structure de population avec plusieurs sous-populations. La structure de population est la présence d'une différence systématique dans la fréquence des allèles entre les sous-populations d'une population en raison de l'accouplement non aléatoire entre les individus. La structure de la population peut se produire pour différentes raisons telles que le chevauchement des générations, l'accouplement assorti ou l'isolement géographique (Wakeley, 2009).

L'accouplement assorti se produit lorsque les individus choisissent leurs partenaires en fonction d'une certaine similitude entre leurs phénotypes. Si le phénotype est déterminé génétiquement, l'accouplement assorti peut influencer le niveau d'hétérozygotie dans la population (Jiang et al., 2013). Le flux de gènes décrit la migration de variants génétiques entre sous-populations dans un scénario de structure de population. Il réduit la différenciation génétique entre les sous-populations (Sousa & Hey, 2013).

Les sous-populations peuvent diverger et devenir génétiquement isolées, ce qui peut donner lieu à une spéciation. Lorsque les événements de spéciation se produisent dans un temps court et que la taille des populations ancestrales est importante, le polymorphisme ancestral peut persister chez les espèces ancestrales, un phénomène appelé tri incomplet des lignées (ILS). La quantité attendue d'ILS dépend du nombre de générations entre deux événements d'isolement et de la taille effective de la population ancestrale.

VI) La coalescence et la recombinaison :

Chez les espèces se reproduisant sexuellement, la recombinaison fait référence à la fois au brassage des chromosomes non homologues et au réarrangement des chromosomes homologues au cours de la méiose. De tels événements de brassage amènent chaque chromosome à avoir deux chromosomes parents dans la génération précédente, qui sont eux-mêmes les produits d'événements de recombinaison dans les générations précédentes. Par conséquent, tout chromosome de la génération actuelle peut être considéré comme une mosaïque de chromosomes qui existait dans le passé. L'ensemble des événements de coalescence et de recombinaison qui décrit l'histoire des chromosomes échantillonnés jusqu'à ce que l'ancêtre commun le plus récent de chaque bloc non recombinant soit atteint est appelé le graphe de recombinaison ancestrale (ARG).

En remontant dans le temps, l'ancêtre commun le plus récent (MRCA) désigne le premier individu où l'ensemble de l'échantillon (population) se réunit pour un bloc non recombinant particulier. Les séquences d'ADN ne fournissent aucune information au-delà du MRCA dans un échantillon de génomes puisque tous les individus partageront toute mutation qui se produit plus loin dans le temps. En présence de recombinaison, différentes parties du génome auront des MRCA différents. Dans l'ARG, les segments nucléotidiques qui se trouvent à la fois dans les chromosomes passés et dans les échantillons contemporains sont appelés matériel génétique ancestral. À l'inverse, le matériel génétique non ancestral fait référence à des segments qui se trouvent dans les chromosomes passés, mais pas dans les échantillons contemporains. En outre, le matériel génétique non ancestral flanqué des deux côtés par du matériel génétique ancestral est appelé matériel génétique piégé. Dans ce contexte, les événements de recombinaison qui se produisent dans le matériel génétique piégé peuvent affecter le déséquilibre de liaison entre les nucléotides actuels. Le déséquilibre de liaison provient de la dérive génétique, du mélange de population et de la sélection, mais elle est réduite par recombinaison à chaque génération. Il est donc plus élevé entre les locus proches et se décompose avec l'augmentation de la distance physique. La sélection liée fait référence à la réduction de la diversité au sein d’une région génomique où un allèle sélectionné positivement entraîne un balayage sélectif des allèles du voisin qui se trouvaient dans le génome ancestral où l’allèle avantageux est apparu. Cependant, la recombinaison crée de nouvelles combinaisons alléliques et réduit cette corrélation à mesure que la distance physique du locus sélectionné augmente. La force de la sélection s’exerçant sur un allèle va influencer l’ampleur du balayage sélectif au sein d’une population.

Chapitre III: Les moteurs de l’évolution

I) Sélection naturelle vs sélection artificielle :

La sélection naturelle est l’un des moteurs de l’évolution. Il s’agit d’un processus qui aboutit à des changements génétiques évolutifs. Au sein d’une population, la sélection naturelle permet aux individus qui possèdent un trait particulier d’avoir plus de descendants fertiles, que les individus qui ont le variant alternatif du même caractère.
Une telle sélection naturelle peut se manifester de différentes manières. Par exemple, un variant qui confère une longévité plus importante, un potentiel attractif plus important pour le sexe opposé ou une progéniture plus importante par évènement de reproduction, peut être sélectionné positivement. Autrement dit, si un variant confère un pouvoir reproducteur plus important, il se répandra dans la population d’une génération à l’autre entraînant ainsi une évolution de cette population. La sélection positive et négative diminue la diversité génétique. Inversement, la sélection balancée agit en maintenant plusieurs allèles dans le pool génétique d'une population à des fréquences plus élevées que prévu par dérive seule. Trois mécanismes sont généralement reconnus : l'avantage hétérozygote, où les hétérozygotes ont une plus grande aptitude que les homozygotes favorisant le maintien d’un polymorphisme génétique, la sélection dépendante de la fréquence, où l'aptitude du génotype est soit proportionnelle soit inversement proportionnelle à sa fréquence dans la population, et finalement quand l'adaptation des génotypes varie en fonction des environnements occupés par l’espèce (également appelée adaptation locale) ou en fonction des stades de la vie.

La sélection naturelle n’aura un effet évolutif que lorsque le caractère est transmissible au cours des générations. Les scientifiques de l’évolution ont mis au point beaucoup de méthodes basées sur l’analyse phénotypique et génotypiques, afin d’étudier la sélection naturelle. D’autres phénomènes comme les migrations, les dérives génétiques, les goulots d’étranglement et les hybridations entre les individus de différents profils génétiques entraînent le remodelage des traits d’une population ce qui peut rendre délicat la détection de la sélection naturelle.

En conclusion, la sélection naturelle favorise les modifications génétiques naturelles adaptatives mais tous les changements ne sont pas adaptatifs. Depuis que l’être humain a commencé à dominer son environnement, des mécanismes d’évolution appelés « sélection artificielle » ou « sélection dirigée » apparaissent où les processus d’évolution sont influencés par l’être humain pour répondre à ses besoins.

1) Identification de la sélection sur le plan génomique :

L'identification des variants candidats est importante non seulement parce qu'ils documentent une évolution et éclairent l'histoire des espèces, mais aussi parce qu'ils peuvent représenter une variation biologiquement significative. Étant donné que la sélection opère au niveau du phénotype, les allèles montrant des preuves de sélection sont susceptibles d'être fonctionnellement pertinents. À l'échelle génomique, la sélection se manifeste par une propagation différentielle non aléatoire d'un allèle en raison de son effet phénotypique.

Les mutations aléatoires sont plus susceptibles d'être délétères que bénéfiques, de sorte que de nombreux nouveaux allèles sont immédiatement soumis à une sélection négative et sont retirés du pool de gènes avant de pouvoir atteindre une fréquence détectable au sein de la population.
Cette élimination continue des mutations délétères est une forme de sélection négative. Dans les régions génétiques soumises à une forte sélection négative, les mutations sont rapidement éliminées du pool de gènes, ce qui entraîne des étendues hautement conservées du génome, c'est-à-dire des régions où la variation est moins observée.

D’autres types de sélection, autre que la sélection positive et négative, donnent lieu à d’autres tendances évolutives communes, en particulier dans les organismes diploïdes et polyploïdes, où le phénotype dépend de l'interaction de plusieurs allèles au même locus. Un tel phénomène est la sélection balancée, dans laquelle plusieurs allèles sont maintenus à une fréquence appréciable dans le pool de gènes. Cela peut se produire en raison, par exemple, d'un avantage hétérozygote (c'est-à-dire d'une surdominance) ou d'une sélection dépendante de la fréquence (Charlesworth, 2006; Turelli, 1984). Si les allèles maintenus conduisent à des effets phénotypiques opposés, par exemple, si de grandes et petites tailles corporelles sont maintenues au sein de la population à l'exclusion des tailles intermédiaires, alors la tendance est souvent décrite comme une sélection diversifiée ou perturbatrice.
En revanche, lorsque les valeurs phénotypiques intermédiaires sont privilégiées, que ce soit en équilibrant la sélection d'allèles codominants ou en sélectionnant positivement les allèles sous-jacents aux phénotypes intermédiaires, la tendance est appelée sélection stabilisante. Malgré cette diversité de modes de sélection, de nombreuses recherches se sont concentrées, ces dernières années, sur le développement de méthodes génomiques pour identifier une sélection positive. L'une des raisons de cet accent sur la sélection positive est pratique : alors que la sélection négative est principalement observable dans les régions hautement conservées et que l'effet de la sélection d'équilibrage sur le génome est souvent subtil, la sélection positive laisse une empreinte plus visible sur le génome qui peut être détectée à l'aide d'un certain nombre de approches différentes. Une autre raison de l'intérêt pour la sélection positive est théorique : la sélection positive serait le principal mécanisme d'adaptation, c'est-à-dire aboutissant à des phénotypes adaptés à un environnement ou à une niche spécifique, ce qui à son tour présente un grand intérêt théorique pour les chercheurs (Akey, 2009).

2) Différentes approches pour l’identification de la sélection positive :

a) Identification de sélection au niveau macroévolutif :

Au niveau macroévolutif, les méthodes utilisées sont basées sur la comparaison des espèces différentes et de leurs taux relatifs de changement génétique. Ces méthodes sont le plus souvent utilisées pour identifier des événements sélectifs qui ont eu lieu dans un passé lointain et qui reflètent les tendances macroévolutives qui se produisent à la suite de la sélection entre, plutôt qu'à l'intérieur des espèces. Les méthodes de détection de la sélection au niveau macroévolutif reposent généralement sur des comparaisons de traits ou de séquences homologues parmi des taxons apparentés.

Ces méthodes identifient les séquences susceptibles d'être fonctionnelles, soit parce qu'elles codent pour des protéines, soit parce qu'elles sont conservées entre différentes espèces et puis rechercher des excès de substitutions par rapport aux taux de mutation de base qui peut être calculé soit à partir du taux de mutations synonymes ou à partir du taux global de substitutions entre espèces (Chamary et al., 2006).

Parmi ces méthodes, je cite la «Gene-based method» qui calcule le rapport Ka/Ks. Elle compare le taux de substitutions non synonymes, qui modifient l'acide aminé, pour chaque site avec le taux de substituions synonymes pour chaque site.
Parce que les changements synonymes sont supposés être fonctionnellement neutres ou silencieux, leur taux de substitution fournit une base de référence par rapport à laquelle le taux d'altérations des acides aminés peut être interprété. Un excès relatif de substitutions non synonymes indique une sélection positive en cours. Ceci est résumé par une valeur de Ka / Ks supérieure à 1, tandis que des valeurs plus petites indiquent une sélection négative continue contre des mutations délétères et la préservation conséquente de la structure protéique.
Ces méthodes peuvent également être appliquées à travers un cadre de lecture ouvert entier ou une subdivision de celui-ci (jusqu'à un codon individuel), car différentes régions peuvent être soumises à différentes pressions sélectives (Yang & Bielawski, 2000).

Une autre méthode basée sur le même principe est le test McDonald-Kreitman (MKT). Il permet d'étudier la divergence interspécifique et la diversité intraspécifique. MKT compare deux valeurs du ratio Ka / Ks, une entre les espèces et une au sein des espèces. Dans ce cas, une sélection positive est mise en évidence quand le taux de mutation non synonymes entre les espèces est supérieur au taux de mutation non synonymes au sein de l'espèce.
Selon la théorie de neutralité, ces taux devraient être égaux, compte tenu des taux constants de mutation et de substitution. Si le rapport inter-espèces dépasse significativement le rapport intra-espèce, l'hypothèse nulle peut être rejetée, suggérant une sélection positive s'exprimant différemment entre les espèces. En effet, la théorie de neutralité explique la diversité génétique par la dérive génétique sans donner une importance à la sélection naturelle (Kimura, 1989). Cependant, alors que les chercheurs commençaient à développer des méthodes pour distinguer les changements neutres des changements adaptatifs dans le génome, beaucoup en sont venus à rejeter les versions les plus fortes de la théorie neutre et ont tourné leur attention vers la quantification des contributions relatives de la dérive et de la sélection à l'évolution moléculaire (Nielsen, 2005).

Semblable au MKT, le test Hudson-Kreitman-Aguade (HKA) utilise à la fois des données de divergence et de diversité pour comparer les taux relatifs de changement. Plus précisément, le test HKA examine les rapports des différences interspécifiques fixes (D, c'est-à-dire des substitutions) aux polymorphismes intra-espèces (P) au sein des loci (Hudson et al., 1987).
Le test repose sur la supposition que, pour un site neutre, D et P sont des fonctions du taux de mutation du site, qui est supposé avoir été à peu près constant au moins depuis le point de divergence des espèces.

En utilisant un test de qualité d'ajustement (par exemple, χ2), il est possible d'identifier des sites individuels ayant un écart au rapport D / P neutre, ce qui permet le rejet de l'hypothèse nulle et peut donc être interprété comme une preuve de sélection. Des valeurs de divergences D / polymorphisme P relativement élevées indiquent que : 1) soit le changement contribuant à la spéciation a été accéléré, donc une sélection directionnelle différente entre les espèces ou bien 2) la diversité au sein de l'espèce est réduite. Des valeurs relativement petites de ce ratio suggèrent une sélection équilibrée entre les espèces.

L'un des avantages de l'approche HKA est qu'elle peut être appliquée à n'importe quelle région génétique, pas seulement à celles qui codent pour des protéines.

En pratique, le taux d'évolution neutre dans les régions codant pour les protéines est beaucoup plus facile à déduire (c'est-à-dire en examinant le taux de substitution synonyme). Ces dernières années, les chercheurs ont élargi cette approche dans un cadre du maximum de vraisemblance pour permettre des comparaisons multilocus plus efficaces (S. I. Wright & Charlesworth, 2004). En examinant plusieurs sites, on peut dériver le rapport D / P neutre attendu pour une lignée tout en tenant compte de la variation du taux de mutation. D'autres études ont utilisé des données génomiques comparatives pour identifier des éléments du génome qui sont hautement conservés entre des espèces mais qui montrent un taux de substitution significativement accéléré dans une espèce ou une lignée particulière (Prabhakar et al., 2006). Par exemple, le gène HAR1F, exprimé au cours du développement cérébral, est hautement conservé entre les chimpanzés et les autres vertébrés mais a 40 fois plus de substitutions chez l'humain que prévu selon la neutralité (Pollard et al., 2006). D'autre part, le principe de comparer des espèces apparentées et d'identifier des différences frappantes peut également être appliquée aux phénotypes. Les traits qui sont conservés dans de nombreuses espèces étroitement apparentées (et donc susceptibles d'être fonctionnels) mais qui montrent une différenciation extrême dans une ou quelques-unes de ces espèces sont de bons candidats pour la sélection naturelle (Romero et al., 2012).

b) Identification de sélection au niveau microévolutif :

Au niveau microévolutif, les méthodes de détection de sélection sont basées sur la fréquence allélique. Une sélection positive amène un allèle bénéfique à atteindre rapidement une prévalence élevée ou une fixation (prévalence de 100%) au sein d'une population.

Lorsqu'un allèle bénéfique et les variants environnementaux sur le même haplotype atteignent ensemble une prévalence élevée, cela produit une réduction à l'échelle de la population de la diversité génétique entourant l'allèle causal (Maynard & Haigh, 2007). Cette réduction, qui persiste jusqu'à ce que la recombinaison et la mutation rétablissent la diversité de la population au locus sélectionné, est la marque d'un balayage sélectif. Donc, les chercheurs cherchent et identifient ces balayages en se basant sur le taux hétérozygotie pour détecter les sélections.

c) Méthodes basées sur le spectre de fréquence allélique :

Le suivi du spectre de fréquences alléliques permet d'identifier les balayages sélectifs. Quand un allèle sélectionné et sa région génétique «auto-stop» voisine se dirigent vers la fixation, ils modifient la distribution des allèles dans la population.

Le balayage entraîne une réduction à l'échelle de la population de la diversité génétique autour du locus sélectionné.

De nouvelles mutations apparaissent sur ce fond homogène, mais ils sont initialement rares car ils ne sont apparus que récemment dans la population. Cela crée un surplus d'allèles rares c'est-à-dire que de nombreux sites proches du variant sélectionné ont des allèles qui ségrégent à de basses fréquences. Le D de Tajima quantifie ce phénomène en comparant le nombre de différences par paires entre les individus avec le nombre total de polymorphismes de ségrégation. Parce que les allèles de basse fréquence contribuent moins au nombre de différences par paires dans un échantillon que les allèles de fréquence modérée, un excédent d'allèles rares gonfle la dernière valeur de manière disproportionnée par rapport à la première. Ainsi, des valeurs négatives de D suggèrent un excès d'allèles rares, ce qui peut indiquer une sélection positive ou une expansion de la population.

d) Méthodes basées sur le déséquilibre de liaison ou LD :

Le déséquilibre de liaison est l'association non aléatoire des allèles à différents loci dans une population. Leurs fréquences d'association est plus élevées ou plus faibles que celle qu'ils auraient s'ils étaient indépendants. Le LD est influencé par la sélection, le taux de recombinaison, le taux de mutation, la dérive, la structuration de population et la liaison génétique. Par conséquent, le modèle de LD dans un génome est un signal puissant des processus génétiques de la population qui le structurent. En effet, un allèle sélectionné reste en fort déséquilibre de liaison avec les variants voisins qui sont entraînés par l'effet auto-stop. Cette forte association sera rompue par la recombinaison. Ce variant causal sélectionné et ses variants voisins liés définissent un haplotype. La recherche de ces régions étendues de fort LD permet d'étudier la sélection positive. Ces régions doivent atteindre rapidement une prévalence élevée car sinon la recombinaison provoquera à terme la décomposition du déséquilibre de liaison et le raccourcissement de l'haplotype.
Les mutations bénéfiques apparues plus récemment ou soumises à une pression sélective moins extrême sont encore plus susceptibles de rester polymorphes dans la population sélectionnée, et beaucoup n'atteindront jamais la fixation parce que les pressions sélectives peuvent changer considérablement sur des dizaines de milliers d'années.
Les approches basées sur LD peuvent également être utilisées pour identifier la sélection balancée à court terme, où le signal est comparable à celui d'un balayage incomplet, au cours duquel une mutation augmente en fréquence dans une population sans atteindre sa fixation. Par exemple, un certain nombre d'articles ont mis en évidence des signaux d'haplotype long associé à la mutation drépanocytaire en Afrique de l'Ouest (N. Hanchard et al., 2007; N. A. Hanchard et al., 2006).

Un test statistique basé sur le LD largement utilisé est «extended haplotype homozygosity statistics» ou EHH (Sabeti et al., 2002). Ce test calcule la probabilité que deux chromosomes choisis au hasard dans la population portant une région étendue ou un long haplotype soient identiques par descendance pour toute la région. EHH mesure la réduction de la diversité haplotypique en calculant la probabilité que deux haplotypes autour d'un locus donné soient identiques, étant donné qu'ils ont le même allèle au locus. Ainsi, à mesure que l'on s'éloigne de la région centrale haplotypique, l'EHH diminue, reflétant l'action de la recombinaison réduisant la longueur de l'haplotype au sein de la population.
Pour tester si l'haplotype a atteint rapidement une prévalence élevée et que la recombinaison n'a pas eu le temps de décomposer l'haplotype, un test appelé long-range haplotype ou LRH est utilisé. Il compare la fréquence d'un haplotype à son EHH.

Si des haplotypes étendus et communs sont trouvés, cela suggère que la recombinaison ne les a pas encore décomposés.

e) Méthodes basées sur la différenciation de population :

La valeur sélective d'un allèle dépend de l'environnement particulier dans lequel les individus existent. Différentes populations sont soumises à des pressions environnementales différentes et, par conséquent, les traits qui seraient adaptatifs dans chacune peuvent être différents.
Si la sélection agit sur un locus au sein d'une population mais pas au sein d'autres populations apparentées, alors les fréquences d'allèles à ce locus parmi les populations peuvent différer considérablement.
Ce principe est à la base d'un ensemble de tests qui reposent sur la différenciation de la population pour détecter les preuves de sélection. Le test commun de différenciation de population est Fst.
Ce test compare les fréquences alléliques entre et au sein des populations (Nagylaki, 1998; S. Wright, 1949). Des valeurs comparativement élevées de Fst à un locus indiquent une différenciation marquée entre les populations, ce qui suggère une sélection directionnelle. Des valeurs comparativement petites indiquent que les populations comparées sont homogènes.
Contrairement aux autres méthodes, les méthodes basées sur la différenciation des populations peuvent détecter différents types de sélection comme les balayages classiques, les balayages sur la variation préexistante et la sélection négative.

f) Méthodes composites :

Comme discuté ci-dessus, la sélection naturelle laisse un certain nombre d'empreintes sur le génome, et chaque test est conçu pour capter un signal légèrement différent. En conséquence, les chercheurs combinent parfois plusieurs métriques dans des tests composites dans le but de fournir une plus grande puissance et / ou résolution spatiale.

Ces tests se présentent sous deux formes distinctes, qui sont toutes deux généralement appelées composites. 1) Des méthodes qui forment un score composite pour une région génétique plutôt que pour un seul marqueur génétique en combinant des scores individuels à tous les marqueurs de la région. L'avantage de cette approche est que si des faux positifs peuvent se produire à n'importe quel site au hasard, une région de vrais positifs est beaucoup plus susceptible de présenter un vrai signal. En effet, comme les balayages sélectifs affectent des haplotypes entiers, on suppose que le signal de sélection s'étend à travers une région (Carlson, 2005). 2) Des méthodes composites qui intègrent le même test sur plusieurs sites améliorent la puissance et réduisent les faux positifs. Un exemple de cette approche est le test CLR de Kim & Stephan (Kim & Stephan, 2002), qui évalue la probabilité qu'un événement sélectif soit responsable d'un excès d'allèles dérivés sur plusieurs sites. Donc, pour conclure parmi les tests composites, il y a des tests qui combinent un ou plusieurs tests sur plusieurs variants, d'autres combinent plusieurs tests sur un seul variant.

g) Sélection sur la variation préexistante :

Comme les mutations se produisent au hasard et non en réponse à des pressions sélectives spécifiques, les allèles peuvent apparaître à un moment où ils ne sont pas immédiatement bénéfiques. Ces allèles neutres pourraient atteindre une fréquence modérée au sein de la population par dérive génétique. Si les pressions environnementales évoluent plus tard pour rendre un tel variant avantageux, le scénario est appelé «sélection sur variation préexistante».

La sélection sur variation préexistante crée un plus grand nombre de sites neutres liés qui ont des allèles à fréquence intermédiaire. Comme la distinction entre les signatures de sélection sur les variants *de novo* (balayage dur) ou la variation préexistante peut être subtile, Peter et al (Peter et al., 2012) proposent un cadre de calcul bayésien approximatif (ABC) pour les distinguer.
Un cas spécial de sélection sur variation préexistante se produit lorsque le variant préexistant apparaît sur plusieurs arrière-plans d'haplotypes distinctifs, par exemple, à la suite d'une mutation ou d'une migration récurrente. Ce phénomène est appelé balayage doux.
Bien que le terme de balayage doux soit parfois utilisé à tort pour indiquer la sélection la variation préexistante, les deux doivent être distingués, à cause de la signature sélective que ces tendances laissent, et, par conséquent, la ou les méthodes développées pour les détecter diffèrent (Pritchard et al., 2010).

II) Domestication et sélection artificielle :

Les agriculteurs et les éleveurs du Néolithique ont commencé à sélectionner les plantes et les animaux et en ont modifié certains caractères à travers le temps. Ce processus, connu sous le nom de domestication, est également qualifié de sélection artificielle ou sélection dirigée car il s'agit d'une sélection exercée par l'être humain et non pas par une adaptation à une condition environnementale donnée. De nos jours, la domestication de certaines espèces est orientée, vers la production massive de lait et de la viande.

Il y a moins d'unanimité dans les différentes définitions de la domestication. En effet, certains considèrent que pour qu'un animal soit domestiqué, les êtres humains doivent maîtriser tous les aspects de l'animal, sa reproduction, son mouvement, son alimentation et sa protection. D'autres considèrent que la domestication est une forme de mutualisme entre les deux partenaires dont chacun tire des bénéfices (O'Connor, 1997). D'autres encore soutiennent que les animaux domestiques ont manipulé les êtres humains d'une manière involontaires dans des relations qui leur ont donné un grand avantage évolutif au détriment de leur forme physique (Budiansky, 1997). Une bonne synthèse serait que la domestication résulte du mutualisme et que les deux partenaires tirent des avantages de leur interdépendance.
Les relations de co-évolution entre les êtres humains et les animaux domestiques cibles sont largement motivées par la capacité humaine à créer spontanément de nouveaux comportements adaptatifs qui répondent le mieux à leurs objectifs. Cependant, la capacité humaine d'apprentissage sociale place les êtres humains dans un rôle dominant dans un mutualisme de plus en plus asymétrique qui évolue à un rythme considérablement accéléré. Le processus de domestication implique l'expression phénotypique des changements génétiques qui transforme l'animal de son phénotype sauvage à son phénotype domestique. Ces changements génétiques au cours du processus de domestication sont causés par un certain nombre de processus sélectifs qui varient selon l'animal et la nature de sa relation avec l'être humain (Jensen, 2014; Price, 2002). La domestication a aidé les êtres humains de passer d'une dépendance totale vis-à-vis les populations libres à une économie agricole dans laquelle les animaux domestiques font jusqu'à 40 à 60 % de l'apport calorique humain (Smith et al., 2001).
Les voies de domestication varient en fonction des contraintes morphologiques, physiologiques et comportementales de l'animal domestique, de l'intensité de l'investissement humain et du contexte environnemental global dans lequel la relation se déroule.

Le défi alors est d'identifier les moyens de tracer les voies variables de la domestication et d'identifier les forces qui la dirigent en cours du temps. Les réponses à la domestication se sont probablement concentrées au début du processus de domestication principalement sur les caractéristiques comportementales chez les animaux (Barker, 1983; Price, 2002) . Par conséquent, l'identification des comportements sélectionnés et l'impact de cette sélection sur les animaux domestiques est essentielle pour comprendre la domestication des animaux.
Les êtres humains ont dû gérer les communautés animales en se basant sur des caractères pré-adaptatifs en déterminant par exemple les partenaires et le calendrier de la reproduction ou en s'occupant de jeunes animaux et de nouveaux nés. Ces caractères pré-adaptatifs se résument généralement dans la structure du groupe, le comportement sexuel, l'interaction et la relation entre les parents et les jeunes animaux et les réactions des animaux envers les êtres humains (Zeder, 2012). De plus, la flexibilité pour répondre aux exigences alimentaires et autres conditions environnementales de survie facilite la relation entre l'animal domestique et l'être humain. En général, le degré auquel une espèce est pré-adaptée à la domestication est alors positivement corrélé avec le degré auquel son comportement dans son environnement naturel ressemble à son comportement dans son environnement captif (Price, 2002).

1) Domestication , Commensalisme et Apprivoisement :

a) Domestication :

La domestication des végétaux, puis des animaux au début du Néolithique au Croissant Fertile, à partir d'il y a environ 12000 ans, a été une étape décisive de l'évolution des sociétés anciennes qui est à la base de nos sociétés actuelles.
La culture des plantes et la domestication des animaux ont permis aux groupes humains de s'installer durablement en donnant un surplus de nourriture stockable. Les plantes et animaux domestiqués ont permis aux populations de devenir plus grandes, d'augmenter l'accumulation et l'échange de biens matériels, fondant ainsi les bases de la société moderne.
Du côté animal ou végétal, la domestication conduit à une croissance du nombre d'individus des espèces domestiquées et donc à une réduction de leur risque de disparition, à l'élargissement de leurs aires de répartition et à l'augmentation de la taille de leurs populations. Du côté humain, les animaux et végétaux domestiqués satisfont des besoins alimentaires, économiques, sociaux et culturels.

Comme le cas pour les bovins domestiqués qui ont constitué pour les sociétés anciennes comme actuelles une force de travail sur les champs importante, mais aussi une source de lait, de peau et de viande. La domestication conduit à des modifications génétiques induites volontairement ou pas par les êtres humains contrôlant leur reproduction. Elle correspond alors à l'acquisition, la perte ou le développement de nouveaux caractères comportementaux, physiologiques ou morphologiques.

Les effets de cette sélection dépendent de la taille effective des populations. La domestication est considérée comme une sélection orientée d'une manière progressive et elle a plusieurs effets sur les populations d'intérêt. En effet, comme la domestication a commencé par l'isolement d'un petit nombre d'individus à partir d'une population sauvage plus grande, ceci peut conduire à la réduction de la diversité génétique initiale car certains allèles ne seront pas présents dans les populations d'animaux domestiques et certains gènes faiblement présentés dans les populations sauvages peuvent être perdus au cours du processus de la domestication. Ces individus isolés vont alors subir une évolution différente de celle du reste de la population initiale et qui est susceptible d'aboutir à des adaptations aux nouvelles conditions environnementales. Ces adaptations sont les conséquences de l'expression de nouveaux traits génotypiques sélectionnés au cours de la domestication.

Cependant, l'élevage intensif dont l'objectif majeur est d'augmenter le taux de reproduction tout en gardant les traits particuliers d'intérêt agronomiques, conduit à l'augmentation d'évènements de recombinaisons génétiques transmises de générations en générations entraînant ainsi l'apparition de combinaisons alléliques inédites qui vont développer des caractères inédits.
Il est connu que la première conséquence de la domestication est le changement de comportement, c'est-à-dire que les animaux deviennent plus dociles (Clutton-Brock, 1992; E. O. Price, 2002). Certaines conséquences morphologiques et physiologiques pourraient être liées aux changements de comportements tandis que d'autres apparaîtraient plus tard (Zeder, 2012).

La domestication est un processus évolutif progressif. Bien que les processus génétiques à l'œuvre au cours de la domestication des différents espèces animales ne sont pas tous bien caractérisés, il paraît déjà clairsemant que les différents changements génétiques ne se seraient pas produits de la mème manière. De plus, une fois que l'animal était domestiqué, s'il peut revenir à l'état sauvage, il est fort probable qu'il ne reviendra pas identique à l'animal sauvage initialement domestiqué.
Il a été montré que les voies de domestication des animaux étaient très variables et dépendaient de paramètres biologiques et culturels largement définis, ainsi que de plusieurs facteurs qui ont façonné les trajectoires de la domestication animale (Zeder, 2009). Ces voies variées peuvent, cependant, être regroupées en trois scénarios principaux de domestication qui semblent représenter le spectre complet des animaux domestiques : La voie commensale, l'apprivoisement et la sélection dirigée.

b) Le commensalisme :

La voie commensale est le plus souvent parcourue par les animaux qui entrent en contact initial avec les êtres humains pour se nourrir de déchets ou pour s'attaquer à d'autres animaux attirés par les établissements humains. Ces animaux développent des liens sociaux, économiques, plus étroits avec leurs hôtes humains. Le chien est un exemple classique d'un animal qui a probablement parcouru cette voie vers la domestication. Sa domestication aurait commencé lorsque des loups moins méfiants ont été attirés par des campements humains pour récupérer des déchets humains (Darcy Morey, 1994). Ces espèces commensales peuvent soit être considérées comme des nuisances soit apporter des bénéfices mutuels, soit être tolérées par les êtres humains sans qu'ils soient jugées intéressantes. Ce dernier type de relation entre les êtres humains et les espèces animales a existé beaucoup avant que les êtres humains décident d'élever des animaux.

Il a probablement fallu beaucoup de temps aux animaux empruntant la voie commensale pour passer d'une simple accoutumance aux humains et aux habitats humains au développement d'un partenariat actif avec eux. Le moment et la nature des forces qui ont propulsé l'élevage dirigé par l'être humain variaient probablement selon les différents animaux domestiques commensaux.

c) L'apprivoisement :

C'est le procédé par lequel l'être humain habitue un animal sauvage à son contact. L'apprivoisement et la domestication diffèrent essentiellement dans le fait que le premier n'est pas définitivement acquis. L'être humain ne contrôle pas forcément l'alimentation ni la reproduction de l'animal sauvage, même quand il parvient à l'habituer à être touché, caressé et même transporté sans frayeur. Il arrive aussi qu'un animal apprivoisé retourne de lui-même à la vie sauvage, de même que sa descendance. Par contre, le processus de domestication implique l'élevage de lignées animales sur plusieurs générations, ce qui n'est pas le cas pour l'apprivoisement ou le recrutement se fait en permanence par prélèvement d'animaux sauvages. Le processus de domestication a été accéléré lorsque les humains utilisaient les connaissances acquises grâce à la gestion d'animaux déjà domestiqués pour domestiquer une espèce sauvage qui possède une ressource ou un ensemble de ressources que les humains considèrent comme souhaitables.

d) Pression de sélection dirigée forte :

Après la publication de «L'origine des espèces» en 1859, les biologistes sont restés sur l'idée que l'évolution d'une population s'effectue sur une longue durée. Et ceci, en conséquence de la phrase que Darwin avail écrit dans son œuvre «We see nothing of these slow changes in progress, until the hand of time has marked the long lapse of ages» (Darwin 1859). Aujourd'hui, beaucoup affirment que Darwin avait tort et que l'évolution peut se faire avec une vitesse rapide. Plusieurs études ont montré que la sélection naturelle peut entraîner l'apparition de nouveaux caractères à une vitesse rapide. Citons l'exemple des guppies de Trinidad qui sont généralement plus colorés dans un environnement sans prédateurs (Templeton, 2004). En 1991, Endler a déplacé les poissons des ruisseaux qui contenaient des prédateurs vers des ruisseaux sans prédateurs. Il a remarqué l'apparition rapide d'une modification dans la couleur des poissons (Endler, 1991). Ce changement de couleur a été expliqué par la préférence des femelles pour les mâles les plus colorés et, par conséquent, les plus visibles. Ce phénomène s'est réalisé en sécurité, en absence de prédateurs, et a entraîné une évolution au bout de 14 générations.

2) Les caractères de la domestication et leur vitesse d'apparition :

Il est connu que Darwin a concentré beaucoup d'attention sur la domestication en tant que processus au cours duquel il y a une variation génétique. Il a admis que la variabilité générale est limitée. Il a soulevé à plusieurs reprises la question de savoir pourquoi les animaux domestiques sont si variables. Dans son analyse des causes de la variation sous la domestication, Darwin pensait que les changements doivent être exclusivement dus à des influences environnementales.

Depuis que l'être humain a commencé à contrôler la reproduction des animaux il y a 10000 ans, il a induit une pression de sélection forte entraînant ainsi des modifications rapides dans les génomes des animaux domestiques. Ces derniers ont évolué de manière spectaculaire par rapport à leurs ancêtres sauvages. Pour comprendre la vitesse d'apparition des caractères de domestication, plusieurs études ont été réalisées. Parmi ces études, je cite celle qui a porté sur la domestication des renards faite par Dmitri Konstantinovitch Belyaev (Belyaev, 1979). Elle a permis de montrer qu'au bout de 25 générations, les renards se comportent comme des chiens en cherchant le contact avec leurs soignants et en répondant à leurs noms (Trut et al., 2009). Même si le processus semble prendre plus de temps, car dans l'expérience de Belyaey, des renards présélectionnés ont été utilisés (Lord et al., 2020), il semble alors que la domestication ait commencé par la sélection des animaux moins sensible à un type de stress. L'étude de Hemmer en 1983 a montré que les animaux soumis à un stress prolongé présentent des dysfonctionnements hormonaux, une plus grande sensibilité aux infections et une moindre fertilité (Hemmer, 1983).

3) Signature génétique de la domestication :

La domestication impliquant le contrôle de la reproduction, de l'adaptation, l'isolement des individus, la sélection dirigée des mutations adaptatives, la dérive génétique et la migration a conduit à une grande variabilité d'animaux domestiques (Andersson, 2013; Groeneveld, 2010).
En effet, la sélection dirigée des mutations adaptatives rendent les animaux mieux adaptés aux besoins humains. Pendant longtemps, elle s'est basée sur la sélection phénotypique, où les humains ont gardé des animaux avec des phénotypes favorables entre autre à la reproduction. Après l'élaboration de la théorie de la génétique quantitative, des procédures statistiques de plus en plus sophistiquées ont été mises au point pour sélectionner les animaux présentant des valeurs de reproduction estimées exceptionnelles. Cela a conduit à une amélioration remarquable de la production animale au cours des 50 dernières années. De plus, on pense que les êtres humains ont eu des préférences pour la diversité des phénotypes chez les animaux domestiques. Ils ont sélectionné, par exemple, des mutants qui ont conféré une apparence phénotypique qui leur semblait intéressante, à condition que cette apparence n'interfère pas avec l'utilité des animaux. Ceci expliquerait pourquoi il y a des cochons noirs avec des ceintures blanches ou des chiens avec des crêtes dorsales (Andersson, 2013).

Le nombre important des races bien définies au sein de différentes espèces domestiques représente une source importante pour la recherche en génétique moléculaire. Cela a permis de reconstituer en partie l'histoire des animaux domestiques et celle de leurs ancêtres sauvages. Les différentes études sur les animaux domestiques ont conduit aux développements de programmes de sélection pour améliorer le potentiel génétique de plusieurs races (Daetwyler et al., 2014). La détection de la signature génétique de la domestication se fait par l'étude des gènes responsables des traits comportementaux et morphologiques acquis par les espèces à différentes étapes du processus de domestication. Mais, pour pouvoir analyser ces traits quantitatifs, il faut bien maîtriser les processus auxquels ils régissent. De plus, les ancêtres sauvages des espèces domestiques, dans plusieurs cas, sont éteints. Ceci représente un obstacle de l'analyse et les signatures ancestrales sont donc étudiées par une comparaison entre les différents groupes apparentés. En effet, connaître l'allèle ancestral augmente le pouvoir de détecter la sélection (Boitard, 2012).

Les approches paléogénétiques permettent l'obtention de données génétiques à partir des extraits fossiles et accélèrent les études de l'impact de la domestication sur l'évolution du génome de différentes espèces. Pour explorer l'évolution d'espèces domestiques, les généticiens se sont concentrés sur l'étude des marqueurs génétiques mitochondriaux qui présentent une variabilité importante et non soumis à la sélection (Ho & Gilbert, 2010; Paijmans et al., 2013).

III) ADN mitochondrial :

Les mitochondries sont des organites cellulaires dont le rôle principal est la respiration cellulaire. Elles produisent de l'énergie sous forme d'ATP, par phosphorylation oxydative, nécessaire à la survie et au fonctionnement des cellules des organismes eucaryotes. De plus, elles sont impliquées dans le métabolisme des lipides et des acides aminés et jouent un rôle important dans divers processus cellulaires tels que la prolifération cellulaire, l'apoptose et la différenciation cellulaire. L'origine de cet organite est une endosymbiose d'une cellule procaryote phagocytée par une cellule eucaryote primitive il y a environ 2 milliards d'années.

Le génome mitochondrial des mammifères est haploïde et de taille comprise entre 14.3 et 20 Kb. Il s'agit d'une molécule d'ADN double-brin circulaire contenant des régions non-codantes et des régions qui codent pour des protéines, des ARNs de transfert et des ARNs ribosomaux. La molécule d'ADNmt est formée d'un brin léger (L), riche en bases pyrimidiques, et d'un brin lourd (H), riche en bases puriques, et d'une région non-codante appelée la D-Loop (ou boucle de déplacement) qui contient des régions hypervariables (HVR).

La région hypervariable (HVR) ne subit qu'une pression de sélection faible ce qui explique pourquoi son taux de mutation est élevé. Une cellule eucaryote peut contenir des milliers de copies de génome mitochondrial. L'abondance de ces copies fait du génome mitochondrial un marqueur intéressant pour les études en paléogénétique. Il est plus facile d'obtenir des séquences mitochondriales présente en plusieurs milliers de copies par cellule, qu'une séquence génomique nucléaire présente seulement en deux copies. Le génome mitochondrial a aussi d'autres particularités qui font de lui un outil pour l'étude de l'évolution des populations. En effet, l'ADNmt est à hérédité maternelle. Les mitochondries sont transmises par l'intermédiaire des ovocytes. Au moment de la fécondation, seul le noyau du spermatozoïde pénètre l'ovocyte, le flagelle qui contient les mitochondries, afin de produire l'énergie nécessaire à la mobilité, ne pénètre pas. Cependant, dans quelques cas, l'ADNmt paternel peut pénétrer dans l'ovocyte par des mécanismes qui ne sont pas encore connus. Néanmoins, les mitochondries paternelles et leur ADN sont éliminés et ne sont pas transmis à la progéniture. Deux hypothèses ont été proposées pour expliquer le mécanisme sous-jacent à l'héritage maternel de l'ADNmt. Selon le «modèle de dilution simple», l'ADNmt paternel, présent à un nombre de copies beaucoup plus bas, est simplement dilué et éliminé par l'excès d'ADNmt des ovocytes et, par conséquent, il est difficilement détectable dans la progéniture (Ulf Gyllensten,1991). D'autre part, dans le «modèle de dégradation active», les chercheurs pensent que l'ADNmt paternel ou les mitochondries elles-mêmes sont dégradés de manière sélective, avant ou après la fécondation, afin d'empêcher activement la transmission de l'ADNmt paternel à la génération suivante.

La transmission uniparentale et plus précisément maternelle de l'ADNmt fait que ce dernier ne subit généralement pas des événements de recombinaison. Le génome est passé tel qu'il est d'une génération à une autre. La D-Loop contient des régions points chauds («Hot Spot») au niveau desquelles les mutations sont plus fréquentes (Meyer, 1999). La D-Loop est le marqueur mitochondrial le plus couramment utilisé pour les analyses phylogénétiques des populations. Depuis que les avancées technologiques ont pris leurs essors, le génome mitochondrial complet est plutôt utilisé dans les analyses phylogénétiques en tenant compte de la fréquence des différentes mutations, les positions des nucléotides mutés, le nombre de mutations à chaque position et le taux d'évolution de chaque région de l'ADNmt (pour revue, (Moritz et al., 1987).

L'ADNmt mitochondrial permet alors d'étudier l'évolution sur des périodes relativement courtes, comme c'est le cas pour le processus de domestication, en analysant les polymorphismes nucléotidiques ou « single nucleotide polymorphisms » (SNP). Il s'agit d'un outil puissant pour reconstruire le passé et retracer l'évolution des espèces. Le taux de mutations élevé dans la région hypervariable de l'ADNmt fait que c'est un excellent traceur de la phylogénie à courte distance évolutive.
Certains phénomènes comme les insertions nucléaires (*numts*) et, dans une moindre mesure, l'hétéroplasmie peuvent biaiser l'analyse sur l'ADNmt.

1) Les insertions mitochondriales nucléaires ou les NUMT :

Le génome nucléaire contient des insertions d'origine mitochondriales de différentes tailles (Mishmar et al., 2004; Nomiyama et al., 1985). Elles ont été détectées lors du séquençage du génome entier de différents organismes. Après la libération de l'ADNmt dans le cytoplasme, en raison de l'altération mitochondriale et de changements morphologiques, l'ADNmt est transféré dans le noyau. Les mécanismes de réparation cellulaire reconnaissent les extrémités des fragments de l'ADNmt et les ligaturent aux fragments d'ADN nucléaire. Par la suite, les fragments d'ADNmt vont être intégrés dans le génome nucléaire et forment des régions sans fonctions qui seront transmises au cours des générations de la même manière que le génome nucléaire. Ces intégrations mitochondriales ont été trouvés dans le génome nucléaire humain (Fukuda et al., 1985).
Les séquences mitochondriales intégrées évoluent à la même vitesse que dans le génome nucléaire. Il a été montré qu'une séquence mitochondriale de la région contrôle intégrée dans le génome nucléaire évolue 10 à 40 fois plus lentement dans le génome nucléaire que dans le génome mitochondrial.

Ces insertions représentent une source d'erreur pour déterminer les haplogroupes mitochondriaux des individus. Elles peuvent biaiser les résultats d'études phylogénétiques car elles peuvent être amplifiées au lieu de leurs homologues mitochondriaux. L'identification de l'origine de la séquence est alors primordiale pour qu'elle soit incluse dans les analyses phylogénétiques (un problème déjà rencontrés chez des données de chat anciens analysées par notre équipe).

Dans les analyses de l'ADNa, il est possible d'amplifier une insertion nucléaire d'ADNmt provenant de l'ADN moderne contaminant parce que l'ADNa dans les extraits est généralement dilué dans l'ADN environnemental. Un exemple frappant de ce phénomène a affecté une étude publiée en 1994 (Woodward et al., 1994), au tout début des études de l'ADNa. En effet, il a été montré que le fragment d'ADN de dinosaure est phylogénétiquement plus proche de l'humain moderne que d'autres espèces (Hedges & Schweitzer, 1995). L'ADN du dinosaure a été considéré comme impliquant une contamination avec de l'ADN moderne du manipulateur-expérimentateur. Pendant la même année, une équipe a montré que la séquence de dinosaure était identique à la séquence d'une insertion nucléaire d'ADNmt chez l'humain (Zischler et al., 1995).

2) L'hétéroplasmie :

La présence de plus d'un génotype mitochondrial chez un même individu ou même à l'intérieur d'une même cellule est un événement appelé héteroplasmie qui est relativement fréquent chez les animaux (Comas et al., 1995). En effet, des mutations peuvent affecter quelques-unes des copies alors que d'autres ne sont pas affectées. Il y a deux types d'hétéroplasmies: Le premier type se caractérise par un nombre limité de positions sur le génome mitochondrial. Dans ce cas, l'apparition aléatoire d'une mutation suivie de sa réplication sera transmise aux descendants. Ces derniers vont avoir des copies de génomes mitochondriaux différents sur une ou plusieurs positions. Les mécanismes de ségrégation de l'ADNmt sont responsables de la proportion des deux génomes mitochondriaux différents.

Alors que le deuxième type se caractérise par la présence simultanée de deux génomes mitochondriaux distincts qui diffèrent par un grand nombre de nucléotides. Dans ce cas, l'hétéroplasmie est causée par la présence chez l'individu de deux génomes mitochondriaux d'origines distinctes, probablement d'origine maternelle et paternelle. Comme on le sait, bien que le taux en soit très faible, un accident lors de la fécondation et la pénétration du spermatozoïdes dans l'ovocyte peut entraîner l'entrée des mitochondries contenues dans le flagelle (Kaneda et al., 1995). Par conséquence, l'hétéroplasmie représente un problème et une source d'erreur pour l'étude des phylogénies mitochondriales.

IV) Histoire de l'analyse des mitogénomes anciens :

Le domaine de la paléomitogénomique est un domaine récent car ce n'est qu'en 2001 que des génomes mitochondriaux complets de Moa, oiseaux aptères non volants ont été publiés (Cooper et al., 2001). Depuis, l'ADN mitochondrial a été abondamment utilisé pour l'étude des phylogénies (Ho & Gilbert, 2010). Au début du domaine, la PCR a été la méthode utilisée pour l'obtention des premiers mitogénomes. Les fragments d'ADNa ont été amplifiés par PCR puis séquencés par la méthode classique de Sanger (Cooper et al., 2001; Kalmar, 2000; Rogaev et al., 2006).

La nature dégradée de l'ADNa limite l'efficacité de la PCR car pour permettre l'amplification du mitogénome complet, il est nécessaire d'effectuer plusieurs PCR à partir des extraits d'ADNa pour augmenter les chances de couvrir tout le mitogénome. Cependant, ces extraits sont précieux et sont obtenus en faible volume et la multiplication des réactions entraîne la perte définitive de l'information génétique de l'échantillon fossile correspondant. De plus, l'utilisation d'amorces pour initier l'amplification des fragments d'ADN pose de nombreux problèmes. Premièrement, pour certaines espèces éteintes, la divergence des séquences anciennes par rapport aux séquences modernes qui servent à la conception des amorces peut rendre inefficace ces amorces pour amplifier l'ADNa. Deuxièmement, la dégradation importante de l'ADN impose l'amplification de produits de PCR de très petite taille. Plus le fragment à amplifier est de taille importante, plus faible est la probabilité qu'il existe des molécules d'ADN correspondantes dans l'extrait fossile. Troisièmement, pour amplifier des molécules cibles très rares, il faut que les amorces ne participent pas à l'amplification d'autres produits que celui ciblé. Même si la fréquence d'événements d'amplification indésirable est faible, si la cible de ces événements indésirables est présente en concentration beaucoup plus élevée que la cible visée, ces amplifications parasites peuvent être dominantes.

En particulier, les amorces elles-mêmes étant les molécules d'ADN les plus concentrées dans la réaction, les réactions parasites les plus fréquentes sont celles où les amorces sont aussi des matrices utilisées pour l'amplification, aboutissant à la production de produits appelés dimères d'amorces. Les amorces doivent être conçues avec la plus grande attention pour minimiser les risques de ce problème. Les amorces doivent aussi être testées systématiquement et celles ayant tendance à produire ne serait-ce que des traces de dimères doivent être proscrites. Finalement, lorsque les amorces sont optimisées, le risque d'artefact le plus important devient de la contamination avec des traces d'ADN moderne. Si on peut être en mesure d'amplifier une molécule initiale d'ADN, il est important de pouvoir éviter la contamination par même une seule molécule d'ADN moderne. Cela requiert des précautions très importantes et une vigilance de tous les instants lors des manipulations au laboratoire. Ces difficultés sont responsables d'un taux important de faux positifs dans les expériences, même dans celles qui ont été publiées. L'approche de la PCR ciblée limitée à une seule cible analysée à la fois est aussi expérimentalement extrêmement lourde et laborieuse et ne permet d'obtenir qu'une très petite quantité d'information génétique à la fois. Il était donc pertinent d'utiliser la PCR multiplex.

La première stratégie utilisée dans le domaine de la paléogénétique visait à économiser les extraits d'ADN en effectuant une PCR multiplex dans un premier temps suivi par une PCR simplex avec chaque couple d'amorces individuellement dans un second temps, c'est-à-dire, autant de PCR simplex que de couples d'amorces utilisés dans la PCR multiplex (Römpler et al., 2006). Une telle stratégie porte toutefois un risque élevé de contaminations qui sont très difficiles à contrôler et elle n'a pas été utilisée, ou quand elle l'a été, elle a produit des résultats douteux. De plus, elle ne résout pas le problème de l'inefficacité et du faible rendement du travail expérimental.
Le couplage de la PCR multiplex avec le séquençage de nouvelle génération (NGS) permet de résoudre le problème de l'efficacité car, dans ce cas, l'ensemble des produits de PCR sont séquencés simultanément.

En effet, plusieurs molécules du même produit de PCR sont séquencées, ce qui permet de minimiser les erreurs de séquences dues à des transformations diagénétiques ou des erreurs de réplication pendant la PCR et il est même possible de séquencer simultanément les produits de PCR provenant de multiples échantillons différents en ajoutant avant le séquençage des indexes différents aux produits provenant de chaque échantillon (Guimaraes et al., 2017).

L'optimisation de cette PCR multiplex est toutefois beaucoup plus délicate que pour une PCR unique. En effet, comme discuté plus haut, il faut éviter la production de dimères d'amorces pour que la PCR soit capable d'amplifier des cibles très rares, voire des molécules uniques. Or, plus il y a d'amorces différentes dans une réaction, plus la probabilité d'avoir une combinaison donnant lieu à des dimères est élevée. Il faut donc essayer de minimiser ces risques avec des programmes bioinformatiques et ensuite systématiquement tester les combinaisons d'amorces (voir les conditions d'optimisation de la technologie baptisée aMPlex dans (Guimaraes et al., 2017). Ce travail est laborieux et délicat, et il est souvent nécessaire de ne combiner ensemble que 5 à 10 couples d'amorces et d'utiliser plusieurs PCR multiplex différentes pour couvrir les différentes régions ciblées. De plus, tous les autres défis de la PCR appliquée à l'ADNa précédemment évoqués se posent aussi avec la PCR multiplex. Par contre, cette approche peut être économique et efficace quand on cherche à analyser un très grand nombre d'échantillons mal conservés. Elle a permis d'étudier différentes questions dans mon laboratoire d'accueil, de détecter de l'ADN de rongeurs du Pléistocène tardif en Afrique du Nord (Guimaraes et al., 2017), de génotyper des parasites dans de nombreux sédiments d'âge différents (Côté et al., 2016), et d'analyser la domestication des chats au cours des derniers 10 000 ans en analysant des centaines d'ossements de toute les régions circumméditerranéennes (Ottoni et al., 2017).

Le domaine de la paléogénétique a été transformé lorsque les technologies de séquençage NGS ont été utilisées pour séquencer l'ensemble de l'ADN contenu dans les fossiles (Noonan et al 2009). Cette approche est restée toutefois assez inefficace et coûteuse pendant plusieurs années car l'ADN endogène ne représentait qu'une toute petite partie de l'ADN séquencé, la majeure partie de l'ADN provenant de l'environnement dans lequel l'échantillon, la plupart du temps des os, était enfoui.
Au cours des années, les techniques de construction des banques génomiques se sont affinées pour les adapter aux propriétés particulières de l'ADNa (petite taille des fragments et dommages à l'ADN) de manière à ce qu'il soit mieux représenté dans la banque génomique (Gansauge et al., 2017; Gansauge & Meyer, 2013; Meyer & Kircher, 2010). De plus, les méthodes d'extraction de l'ADN se sont améliorées pour mieux purifier les fragments d'ADN courts (Dabney et al., 2013; Glocke & Meyer, 2017), des prétraitements de la poudre d'os pour réduire la quantité d'ADN environnemental ont été développés (Korlević et al., 2015).
Finalement, les résultats se sont sensiblement améliorés avec la mise en évidence de la meilleure conservation de l'ADN dans les os les plus denses et compacts, comme dans l'os pétreux (Gamba et al., 2014; Geigl & Grange, 2018; Pinhasi et al., 2015). Toutes ces évolutions, associées à la diminution régulière des coûts de séquençage NGS, ont rendu les approches de construction de banques génomiques beaucoup plus efficaces que les PCR ciblées pour l'analyse de l'ADNa.

Toutefois, tous les échantillons ne se prête pas à un séquençage direct par Shotgun quand ils contiennent encore un trop grand excès d'ADN environnemental, et même un échantillon contenant jusqu'à 10% d'ADN endogène, ce qui est souvent déjà considéré comme un bon échantillon bien conservé, sera encore très coûteux à séquencer et il reste encore de nombreux échantillons intéressants qui en contiennent encore moins. L'approche de capture de cibles d'intérêt par hybridation permet alors d'analyser de tels échantillons à des coûts encore raisonnables (Gnirke et al., 2009). Il existe de nombreuses déclinaisons de cette approche de capture. Elles utilisent le même principe mais les protocoles différent par la nature des sondes utilisées pour l'hybridation, ADN (Maricic et al., 2010) ou ARN (Carpenter et al., 2013), et par le support utilisé pour l'hybridation, solide (Enk et al., 2014) ou en solution (Bekaert et al., 2016; Horn, 2012). Mon laboratoire d'accueil a développé des approches économiques de capture en solution utilisant des sondes ARN biotinylées adaptées à l'ADN mitochondrial ancien (Massilani et al., 2016), mais aussi à des marqueurs nucléaires (Brunel et al., 2020).

Selon les questions posées et les échantillons analysés, la méthode d'analyse la plus adaptée peut être différente et il est utile de considérer ce qui est ciblé et ce que l'on obtient par les différentes approches. Le séquençage Shotgun permet la mesure d'une proportion relative de l'ADN présent dans l'extrait fossile mais il est difficile d'évaluer la quantité absolue avant d'avoir séquencé l'extrait de façon importante. Une approche PCR comme l'aMPlex, par contre, dépend surtout de la quantité d'ADN présente et est moins sensible à la présence d'ADN environnemental.

Chapitre IV: Les bovinés

I) Les bovinés :

En plus des *Bovinés*, la famille des Bovidés comprend d'autres sous-familles et notamment celle des caprinés avec les caprins et les ovins. Les bovinés représentent un groupe monophylétique avec des histoires évolutives diverses leurs permettant d'être un modele pour suivre les processus d'évolution sous pression de sélection naturelle ou artificielle par l'intervention de l'être humain. Les *Bovinés* (*Bovinae*) sont une sous-famille de mammifères ruminants de la famille des bovidés. Les bovinés sont des grands ruminants domestiques (bœuf, zébu, buffle d'eau) par opposition aux petits ruminants domestiques (les ovins et les caprins).
Cette sous-famille comprend plusieurs espèces importantes d'animaux sauvages ayant subi ou non la domestication et des animaux domestiques tels que *Bos taurus*. En effet, ce dernier est considéré comme une espèce agricole qui a joué un rôle important dans l'établissement des sociétés humaines depuis le Néolithique.

Ces herbivores ont un estomac formé de quatre compartiments (le rumen, le réseau, le feuillet et la caillette) permettant la rumination et donc digestion efficace des herbes mangées. Lorsque l'animal broute l'herbe, mastiquée imparfaitement, s'arrête d'abord dans la panse. Au moment où se fait la rumination, l'aliment remonte de la panse dans la bouche, où il est soumis à une nouvelle, mastication, puis il tombe dans le feuillet. De là il passe dans la caillette, ensuite dans les intestins, qui ont une très grande longueur. Leur mâchoire supérieure est dépourvue d'incisives. Sur le devant de la mâchoire inférieure, on compte généralement 6 incisives et 2 canines incisiformes. Pour couper l'herbe, l'animal la serre entre ses incisives inférieures et un bourrelet corné recouvrant le devant de la mâchoire supérieure, puis il la brise en donnant un coup de tête. Pour mastiquer, ils font effectuer à leur mâchoire inférieure des mouvements de droite à gauche et réciproquement. L'herbe étant divisée naturellement en lanières, les canines sont inutiles, et à leur place se trouve un grand espace vide. Cependant, il est curieux de constater que les Ruminants qui manquent de cornes, tels que les chameaux, les chevrotains, les llamas, etc., sont pourvus de deux canines supérieures, comme si ces dernières étaient chargées de remplacer les organes de défense qui manquent. La famille des ***Bovidae*** se divise en neuf sous-familles, dont l'une est la celle des Bovinés.

La sous famille de Bovinae qui est elle-même diphylétique comprenant d'une part, les genres ***Bison*** et ***Bos*** qui forment les bovinés au sens strict et d'autre part, les buffles d'Afrique (*Syncerus*) et les buffles d'Asie (*Bubalus*).
Le genre ***Bison*** comprend :

- *B.bonasus*: bison d'Europe
  B.bison: bison d'Amérique
  B.priscus: bison des steppes
  Ainsi que d'autres formes de bison encore plus anciennes

Le genre ***Bos*** comprend les sous-genres :

- *Bos*: *B.primigenius primigenius* (aurochs) et *B.primigenius primigenius cf. taurus* (bœuf taurin ou *Bos taurus*) et *B. primigenius primigenius cf. indicus* (zébu ou *Bos indicus*)

  Bibos: *B.javanicus* (banteng) et *B.frontalis* (gaur/gayal)

  Poephagus: *B.grunniens* (yack)

  Navibos: *B.sauveli* (kouprey)

La sous-famille des Bovinés est divisée en 3 tribus principales les Tragelaphini, les Boselaphini et les Bovini. La plupart des représentants des deux premiers groupes sont chassés pour leur viande et leur peaux. Le groupe des Bovini comprend les principaux bovins domestiques, ainsi que des espèces sauvages dont certaines sont éteintes ou en voie d'extinction (Hassanin & Ropiquet, 2004; Schaller & Liu, 1996). MacEachern et al en 2009 ont décrit dans un tableau les différents genres du groupes des bovinés leur classification et leur statut relatif à la domestication .

BOVINAE			
Tribus	Espèces représentatives	Domestication	Notes
Tragelaphini	*Taurotragus oryx* (Eland)	Non	
Boselaphini	*Boselaphus tragocamelus* (Antilope Nilgaut)	Non	
Bovini	*Syncerus caffer* (Buffle d'Afrique)	Non	
Bovini	*Bubalus bubalis* (Buffle Asiatique)	Oui	Domestiqué en Asie
Bovini	*Bubalus carabanesis* (Buffle des marais - Carabao)	Oui	Domestiqué en Asie de l'Est
Bovini	*Bison bison* (Bison américain)	Non	
Bovini	*Bison bonasus* (Bison européen)	Non	
Bovini	*Bos grunniens* (Yak)	Oui/Non	Domestiqué en Asie, Individus sauvages existants
Bovini	*Bos taurus* (Boeuf domestique et Zébu)	Oui	Domestiqué au Moyen Orient (Bœuf) et en Asie (Zébu)
Bovini	*Bos javanicus* (Banteng)	Oui/Non	Domestiqué en Indonésie, Individus sauvages existants
Bovini	*Bos gaurus* (Gaur)	Non	Domestication possible mais est considéré comme non domestiqué
Bovini	*Bos frontalis* (Gayal ou Mithan)	Oui	Possibilité d'être la version domestique de Bos gaurus

Tableau 1: Résumé des espèces représentatives de la sous-famille bovinés et leurs statuts relatif à la domestication (d'après MacEachern, 2009).

Au sein de la tribu des Bovini, la divergence des différentes espèces a été estimé à environ 2,8 millions d'années (Hartl et al., 1988). Cependant, des radiations plus récentes ont donnée d'autres espèces sans avoir contribué à une spéciation complète. En effet, les zébus et les taurins sont aujourd'hui capables de se reproduire et de donner une descendance hybride viable et interfertile alors que les croisements entre d'autres espèces de *Bovina* donnent généralement des mâles stériles et des femelles fertiles. Des séries de rétrocroisements répétés sont capables dans certains cas de restaurer la fertilité des mâles (Hassanin & Ropiquet, 2004).

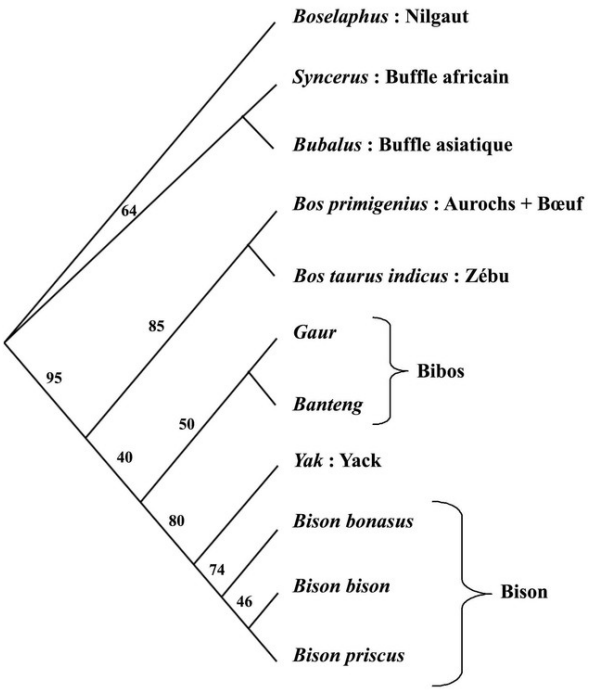


Figure 3: Arbre phylogénétique des bovidés (WJA Payne & J Hodges, 1997) .

Le genre *Bos* représente l'ensemble des bovins domestiques et leurs ancêtres sauvages. Plusieurs auteurs utilisent le nom grec Bos (signifiant « bœuf ») comme désignation principale du bœuf domestique proprement dit *Bos taurus* et son homologue le zébu *Bos indicus*. *Bos taurus* et *Bos indicus* représentent les espèces bovines les plus trouvées. Le genre contient aussi d'autres espèces retrouvées essentiellement en Asie du Sud-Est et de l'Est : Le bison indien, gaur ou gayal qui correspond à l'espèce *Bos frontalis* retrouvée en Inde centrale et orientale, le yak ou *Bos grunniens* vivant essentiellement au Tibet , le tembadou ou banteng (*Bos javanicus*) retrouvé en Indonésie et le kouprey (*Bos sauveli*) retrouvé au Cambodge. Les espèces domestiques sont interfertiles et donnent des hybrides fertiles ou partiellement fertiles.

Les espèces du genre *Bos* qui sont anciennement domestiquées, ont joué un rôle important dans l'établissement des sociétés humaines actuelles. Le bétail a suivi les mouvements humains à travers le monde et il s'est individualisé localement et s'est subdivisé en différents rameaux dans le monde. Chaque rameau a accumulé des nouvelles caractéristiques qui les différencient des autres groupes donnant naissance alors à différentes races de bœufs domestiques.

Le bétail joue aujourd'hui un rôle économique très important. C'est l'une des principales ressources de lait, de viande, de cuir et d'autres produits secondaires. Ces animaux sont aussi exploités pour le travail.

Pour obtenir des animaux de plus en plus productifs et afin d'améliorer le pouvoir génétique des troupeaux, les humains ont mis au point plusieurs systèmes de contrôle de la reproduction des bœufs domestiques, appelés aussi des systèmes artificiels de reproduction, qui entraînent l'augmentation de la pression de sélection subie par les animaux domestiques (Pieper, 2016).

✔ L’insémination artificielle : est pratiquée chez les bovins depuis le début du XXème siècle. Elle consiste à introduire dans l'utérus de la vache des paillettes du sperme sans qu'il y ait de rapport sexuel. Cette technique permet d’obtenir un nombre de descendants beaucoup plus important pour un taureau que par reproduction naturelle. Les taureaux utilisés en insémination artificielle doivent avoir des caractéristiques d’intérêt pour l’éleveur. Ce type de reproduction entraîne la réduction de la diversité génétique.

✔ Le transfert d’embryons : est une technique qui consiste à faire produire un nombre important d’embryons à une même vache par le biais de traitement hormonaux, puis de les transférer dans l’utérus d’autres vaches dites porteuses. Cela permet d’obtenir un nombre élevé de veaux de qualité.

✔ Le clonage : consiste à créer artificiellement des individus identiques. Il existe deux types de clonage : Le clonage somatique, qui consiste à recréer un animal à partir d’une cellule somatique d’un individu vivant. La vache Margueritte clonée par l’INRA en 1998 est le premier bovin issu de cette méthode. Le clonage somatique permet de recréer un animal avec des caractéristiques d’intérêt afin d’améliorer la sélection des animaux d’élevage ou aussi de créer des animaux transgéniques. En plus du clonage somatique, on trouve le clonage embryonnaire qui consiste en une scission de l’embryon de manière à obtenir des animaux identiques.

La quantité de lait produite par vache et la quantité de viande a augmenté considérablement dans les élevages intensifs et continue d’augmenter avec les méthodes efficaces de sélection.

II) Le genre *Bos* :

1) Les bovins et la culture humaine :

Depuis les périodes préhistoriques, les bovins ont fasciné les êtres humains comme le témoignent les diverses représentations des bovins sauvages des grottes préhistoriques. Ils occupent alors depuis une place importante dans de nombreuses catégories socio-culturelles. Pour des nombreuses croyances et religions, le taureau symbolise la virilité, la vigueur et l’énergie ainsi que taureau et vache la fertilité.Le bovin a été un animal de sacrifice prestigieux pour beaucoup de civilisations anciennes mésopotamienneses (Sumériens, Assyriens, Babyloniens et Hittites), égyptiennes, minoennes, gauloises, romaines, grecques et autres mais aussi parfois modernes, comme pour la Corrida en Espagne et pour les riches Musulmans en Turquie. Certaines civilisations anciennes ont aussi pratiqué le culte d’un dieu-taureau, comme en Mésopotamie, en Égypte et en Gaule.

De plus, dans la mythologie égyptienne, Hathor protectrice des nouveau-nés et déesse de l'amour, de la joie et de la danse est représentée sous la forme d'une vache ou sous la forme d'une femme avec des cornes de vaches. La vache qui symbolise la fécondité est associée aussi au Nil qui fécondait la terre. Au 4ème siècle n.e. (de notre ère), le bœuf apparaît dans les représentations de la nativité de Jésus Christ et symbolise la patience qui le réchauffe de son haleine. Les Hindous considèrent la vache comme l'incarnation de tous les dieux et ils refusent qu'elle soit tuée. La vache est en effet vue en Inde comme une « Mère universelle ». La vache, sous le terme *gaya*, signifie aussi « douceur », du fait qu'elle donne son lait à tous, même à ceux qui ne sont pas ses veaux. En Inde, la vache n'est pas seulement sacrée en tant que telle, bien qu'étant décrite dans la littérature hindoue comme la source et le fruit de tout sacrifice aux dieux, elle représente la sacralité de toutes les créatures. Aujourd'hui en Inde, les vaches sont libres de circuler dans les rues et ne sont mangées qu'après leur mort naturelle.Depuis la préhistoire, les manifestations artistiques de ces croyances sont nombreuses et jusqu'aujourd'hui, les bovins sont représentés dans les films de cinéma et dans des marques commerciales.

2) L'aurochs: *Bos primigenius :*

L'aurochs est l'ancêtre commun de tous les bovins domestiques eurasiatiques et du zébu. Selon les données archéologiques et archéozoologiques, cet herbivore a été domestiqué il y a environ 10 000 ans au Moyen Orient et en Inde. L'aurochs a co-existé avec les bovins domestiqués et a disparu au cours de l'Holocène à cause d'une chasse intensive, la fragmentation et réduction de son habitat et la compétition avec le bétail domestique. Le dernier animal a été tué au début du 17ème siècle dans une forêt de Pologne. Il existe donc des textes décrivant cet animal dans la période précédant sa disparition. Par exemple, dans « Il bello gallico », Jules César décrit cet animal comme suit : « Une troisième espèce porte le nom d'urus. La taille de ces animaux est un peu moindre que celle des éléphants, leur couleur et leur forme les font ressembler au taureau. Leur force et leur vélocité sont également remarquables, rien de ce qu'ils aperçoivent, hommes ou bêtes, ne leur échappe. On les tue, en les prenant dans des fosses disposées avec soin. Ce genre de chasse est pour les jeunes gens un exercice qui les endurcit à la fatigue ; ceux qui ont tué le plus de ces urus en apportent les cornes en public, comme trophée, et reçoivent de grands éloges. On ne peut les apprivoiser, même dans le jeune âge. La grandeur, la forme et l'espèce de leurs cornes diffèrent beaucoup de celles de nos bœufs. On les recherche avidement, on les garnit d'argent sur les bords, et elles servent de coupes dans les festins solennels ».

Dans les périodes historiques, du fait de l'abondance des bovins domestiqués, il n'est pas complètement clair si les derniers aurochs les mieux décrits ne correspondent pas en réalité à des hybrides. Par contre, les informations recueillies à partir des ossements et des peintures rupestres des périodes préhistoriques sont quant à elles peu susceptibles d'être faussées par cette hybridation. L'analyse des peintures rupestres et des ossements est le seul moyen pour étudier la morphologie de l'aurochs du Pléistocène supérieur et du début de l'Holocène.

Le dimorphisme sexuel, très prononcé chez cette espèce, apparaît sur les peintures des grottes de Lascaux et autres : les femelles sont beaucoup plus petites que les mâles avec des membres fins et longs alors que les mâles sont plus grands et ont des cornes plus épaisses. La couleur de la robe des mâles était noire alors que celle des femelles étaient plutôt brune. Les peintres paléolithiques ont coloré les aurochs mâles moins que les femelles. Mâles et femelles présentent un garrot grand plutôt saillant, un fanon bien développé, un thorax plus volumineux que l'abdomen et l'arrière train, un dos creusé et une longue queue (Dementiev, 1958)

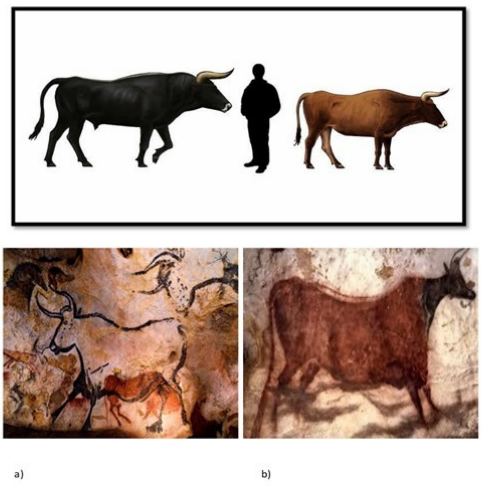


Figure 4 : Représentation anatomique de l'aurochs mâle (à droite de la personne) et femelle (à droite de la personne). a) Aurochs mâle peint sur les murs de la grotte de Lasceaux b) Aurochs femelle peinte sur les murs de la grotte de Lascaux.

(https://fr.wikipedia.org/wiki/Fichier:Aurochs_liferestoration.jpg)(https://en.wikipedia.org/wiki/Lascaux)

Avant la domestication, selon les pièces anatomiques, l'attribution des ossements à un aurochs peut soit être effectuée de manière non ambiguë, soit peut être confondue avec les bisons.
Une fois que les animaux domestiques coexistent avec les aurochs, il est difficile de distinguer l'aurochs des bœufs domestiques néolithiques car la conformation des pièces squelettiques est similaire. C'est alors la grande taille qui est le caractère marquant pour l'attribuer à *Bos primigenius*. Par contre, le dimorphisme sexuel peut fausser ces attributions. En effet, on peut attribuer de manière fiable un ossement à un aurochs seulement si ses dimensions sont supérieures à celles des plus grands mâles domestiques connus et les restes des femelles domestiquées seulement si elles sont plus petites que les petites femelles d'aurochs (Vigne, 2011).

De plus, il y a des variations de taille selon les zones géographiques et les périodes (Guintard, 1999). Les aurochs qui ont vécu en Europe du Sud étaient plus petits que ceux ayant vécu dans les contrées centrales et septentrionales (Davis, 1981). La taille des mâles adultes au Pléistocène a été estimée à 1.8 mètres au garrot en moyenne (Guintard, 1999; Vuure, 2005) et pouvait atteindre jusqu'à 2 mètres. La hauteur est moins importante pour les aurochs de l'Holocène qui mesuraient au garrot entre 1.5 et 1.7 mètres mais conservent des os de taille importante et une constitution imposante (Guintard, 1999). Le poids des aurochs a été estimé entre 800 et 1000 kg pour le mâle et 650 à 800 kg pour la femelle (Guintard, 1999). Leur longévité a été estimée à 15 à 20 ans. Les aurochs vivaient en petits troupeaux constitués essentiellement de femelles et de jeunes animaux et probablement avec un mâle dominant tandis que les jeunes mâles adultes vivaient à l'écart sauf à l'automne en période de rut. La femelle devait donner naissance à un veau par an (Guintard,C & Neron de Surgy, 2014). Les bovins néolithiques étaient plus petits que les aurochs et ont continué de diminuer en taille jusqu'au Moyen Age (Ajmone, 2010), à l'exception des bovins romains qui étaient de plus grande taille. Un autre effet de la domestication est la diminution de la taille des cornes.

En effet, les grandes cornes pour le combat sont devenues inutiles et même indésirables dans un environnement agricole, qui conduit à l'émergence de cornes courtes et même du bétail sans cornes (Ajmone, 2010). Ces deux critères ne sont pas le stade final de l'évolution des bovins car pendant les derniers siècles la différenciation a été accentuée par le développement de centaines de races spécialisées. Le développement le plus récent, impulsé par une approche plus industrielle de l'élevage et la mondialisation de la société humaine, a été une expansion des races les plus productives aux dépens des races rustiques.

3) Origine de l'aurochs :

Les premières étapes d'évolution qui ont permis l'émergence des aurochs sont entachées d'incertitudes et de controverses. C'est dû au fait que plus on remonte dans le temps, plus les ossements sont rares, incomplets et qu'il est difficile sans ambiguïté à une branche ancestrale ou à une branche latérale éteinte. La tendance naturelle des scientifiques est d'attribuer les ossements sur lesquels ils travaillent à des branches ancestrales mais l'analyse génétique des ossements anciens montre qu'il y a un grand nombre de branches latérales éteintes ce qui suggère que la probabilité est généralement plus importante lorsque les ossements sont rares de trouver des ossements correspondant à des branches latérales éteintes. Après ces précautions d'usage, je vais tacher de décrire des points qui semble faire le plus consensus au sein de la communauté des paléontologues.

L'ancêtre des aurochs serait originaire du sous-continent indien (Vuure, 2005) bien que certains auteurs proposent une origine africaine (Martínez-Navarro et al., 2010). La forme la plus ancienne serait *Leptobos* au Pliocène tardif ou Pléistocène ancien qui pourrait être un ancêtre commun aux aurochs et aux bisons (Pfeiffer, 1999). Le plus ancien spécimen a 2 millions d'années et a été trouvé dans le Siwalik en Inde du nord (Lydekker, 1898). Le précurseur des aurochs pourrait être *Bos acutifrons* (Vuure, 2005). *Bos acutifrons* aurait vécu en Inde jusqu'au milieu du Pléistocène et on pense généralement que *Bos primigenius* aurait évolué à partir de cette espèce entre 1,5 et 2 millions d'années. Certains auteurs propose un *Bos namadicus* comme intermédiaire entre *Bos acutifrons* et *Bos primigenius* (Groves, 2009; Pilgrim, 1947).

Au cours du Pléistocène, l'aurochs se serait distribué à partir d'Inde et se serait d'abord propagé vers le sud de l'Europe où il serait arrivé en Espagne il y a 700 000 ans (Jordi Estevez & Maria Sana, 1999). Il a aussi été identifié dans le delta du Tibre dans une période interglaciaire appelée Günz-Mindel (il y a 800 à 500 000 ans AP) (E.Cerilli & C.Petronio, 1992). Il se serait distribué en Europe centrale probablement en passant par la Russie (van Nuure, 2005). Le premier spécimen, un crâne, a été identifié en Allemagne à Steinheim an der Murr et date de 275 000 ans, une période très chaude entre deux âges glaciaires (Lehmann, Ulrich, 1949). La répartition des aurochs au cours du Pléistocène fluctuait en fonction des glaciations. L'aurochs se distribuait en Europe Centrale et du Nord uniquement pendant les périodes chaudes et se rétractaient vers le sud pendant les périodes glaciaires. Ces expansions et contractions ont certainement créé des goulots d'étranglement important pendant chaque période glaciaire et la dérive génétique, voire les liens d'affiliations des formes européennes trouvées pendant les différentes périodes interglaciaires ne sont pas très clairs.
On ne connaît où se situaient les refuges pendant les périodes glaciaires et quels refuges aurait permis de recoloniser les territoires pendant les périodes interglaciaires.

En Inde, l'aurochs a évolué pour donner lieu à *Bos primigenius namadicus* qui est à l'origine de la forme domestiqué du zébu ou *Bos indicus* (Figure 5). Une forme répartie sur la partie méditerranéenne de l'Afrique du nord a été proposée mais il n'est pas clair s'il s'agit véritablement d'une lignée spécifique ou d'une expansion territoriale de l'aurochs *Bos primigenius primigenius* (Vuure, 2005).

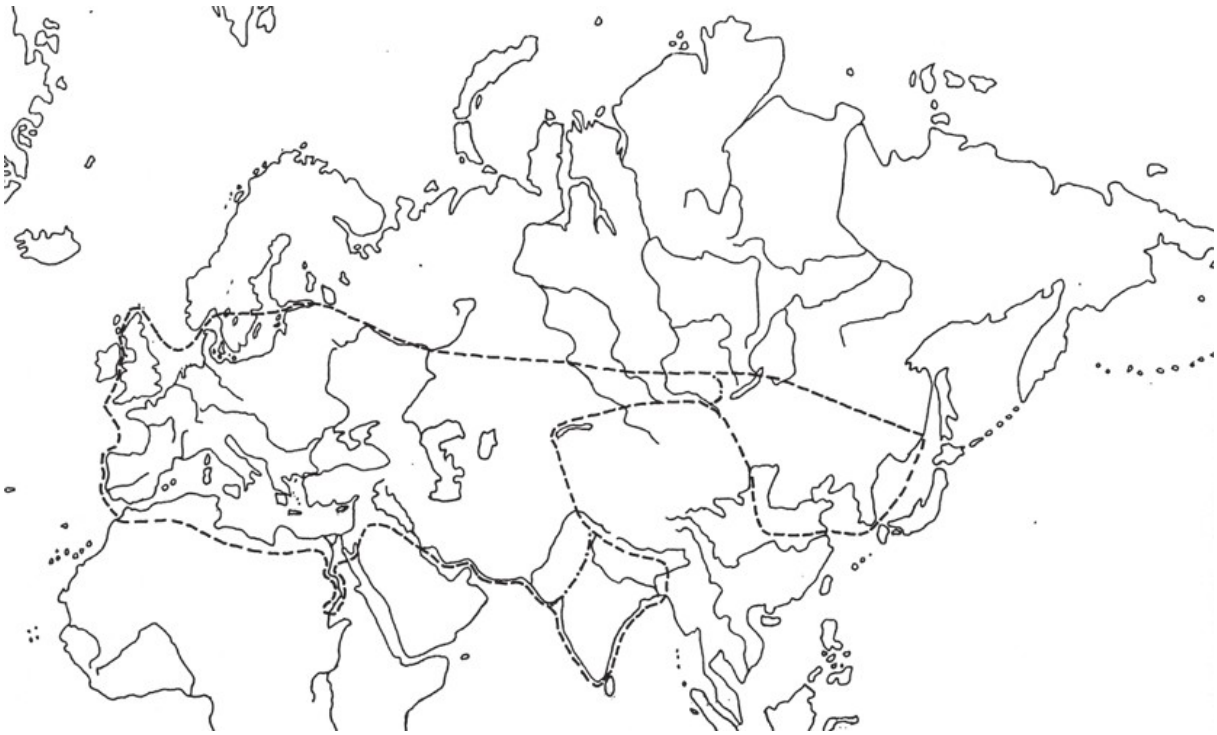


Figure 5 : Aire de distribution maximale cumulée des aurochs à la fin du Pléistocène et du début de l'Holocène (Vuure, 2005).

L'aurochs était peu abondant en Europe de l'Est et en Asie Centrale, car, contrairement au *Bison priscus*, l'aurochs ne pouvait pas passer les plaines continentales à cause de leur sécheresse. Il n'a pas pu non plus atteindre l'Amérique du Nord par le Détroit de Béring (Vuure, 2005).

Les peintures rupestres des grottes de Lascaux et de Chauvet témoignent de la présence de l'aurochs en Europe de Sud-Ouest à l'Aurignacien/Gravettien et au Magdalénien alors qu'il était absent en Europe centrale et de l'Est. Cependant, l'aurochs reste relativement rare parmi les bovidés jusqu'au Pléistocène supérieur et c'est à partir du dernier maximum glacier, il y a environ 20 000 qu'il s'est répandu en nombre (Guintard,C & Neron de Surgy, 2014). Le climat tempéré et humide de la période postglaciaire lors de la fin du Pléistocène et au début du l'Holocène a dû favoriser l'accroissement de ses effectifs et l'extension de son habitat. A partir de ce moment, l'aurochs a peuplé avec succès une large partie de l'Europe de l'Océan Atlantique à la mer Baltique ainsi que l'Afrique du Nord et les régions tempérés de l'Asie orientales (Guintard,C & Neron de Surgy, 2014).

4) Disparition de l'aurochs :

L'aurochs aurait disparu en Asie et en Afrique depuis 2000 ans alors qu'il a subsisté en Europe de l'Est jusqu'au 17ème siècle. Sa disparition est causée par la chasse intensive et la réduction des espaces sauvages boisés (Vuure, 2005; E. Wright, 2013). Les données archéologiques et archéozoologiques ont permis de caractériser le mode de vie des troupeaux d'aurochs. Ils suggèrent que les aurochs avaient besoin de vastes espaces pour survivre. Il a été proposé qu'ils broutaient le feuillage mort et les glands en hiver alors qu'en été ils se nourrissaient d'herbes (Guintard,C & Neron de Surgy, 2014). En plus des forêts qui leur servaient de refuge, l'aurochs était particulièrement bien adapté aux prairies humides, aux plaines et aux marais comme il s'en trouvait sur les bords des cours d'eau (Bokonyi, 1974).

L'Europe a connu une déforestation au Moyen Age à cause de l'augmentation de la taille des populations humaines. Cette déforestation a dû réduire les habitats pour les aurochs car ils ont dû se réfugier dans les forêts à cause de la compétition avec les bovins domestiqués et les agriculteurs. Le bois était utilisé pour le chauffage, pour la construction de maisons et pour l'alimentation des fours. En Europe, dès le 5ème s. n.e., l'aurochs a disparu successivement du sud vers le nord et de l'ouest vers l'est et au 13ème siècle, il n'était présent qu'en Europe de l'Est (Guintard, 1999). Quelques troupeaux survivaient en Pologne et dans la partie la plus occidentale de la Russie au siècle suivant. La Pologne est le dernier pays à abriter des aurochs à la fin du 16ème siècle dans la forêt de Jaktorow. Cette forêt était le dernier refuge de *Bos primigenius*. Les animaux qui ont vécu avec l'aurochs ont accéléré la réduction de leurs ressources alimentaires. Ainsi les pathologies du bétail passant à proximité des aurochs pourraient être la cause principale de l'effondrement de l'effectif d'aurochs. En 1601, on ne comptait plus que trois mâles et une femelle. Cette femelle représentait le dernier spécimen connu de l'espèce. Elle mourra en 1627 à l'âge de 30 ans (Guintard,C & Neron de Surgy, 2014).

III) La domestication de l’Aurochs aux bovins domestiques eurasiatiques :

1) La domestication de *Bos primigenius primigenius* :

Les données zoologiques et moléculaires suggèrent que les premiers centres de domestication des bovins eurasiatiques étaient au Proche-Orient et plus précisément dans le Croissant fertile (Edwards et al., 2007a, 2007b; Helmer, 1992; Loftus et al., 1994; Perkins, 1969; Troy et al., 2001). Les bovins domestiques ont été ensuite déplacés vers l’Ouest suivant deux courants de migration humaine au Néolithique : Le courant cardial et le courant rubané. Selon les données archéologiques et les études génétiques récentes, la domestication de l’aurochs *Bos primigenius primigenius* aurait débuté entre -10 000 et -8000 ans av. n. ère (Zeder, 2008).
La première domestication des bovins peut être imaginée comme un événement spontané et naturel dû à des conditions environnementales, biologiques et sociales qui ont été présentes simultanément pendant un temps relativement court dans une petite région de l’Asie du Sud-Ouest, dans la région Urfa au sud-est de la Turquie actuelle.

L’être humain a augmenté les aptitudes génétiques de la population animale induisant des changements morphologiques, physiologiques et comportementales.
Ces derniers dépendent de la capture et de la gestion de l’animal ainsi que du type d’agriculture utilisée pour l’exploitation de l’animal. La réduction de la taille, spécialement pour les mâles, est une des caractéristiques de la domestication. En effet, la réduction de la taille du corps peut être induite par les changements d’alimentation, d’ensoleillement et de sélection que subissent les animaux maintenus en captivité.

Comme le produit final de la domestication n’a pas pu être anticipé dans la phase initiale, on peut considérer que les motivations initiales pouvaient être autre que celles qui ont justifié la poursuite du processus une fois qu’il a été initié. La chasse de l’aurochs étant une activité dangereuse, elle devait donner du prestige et un statut important à celui qui en était capable. D’autre part, l’importante quantité de viande disponible lorsqu’un animal a été tué, devait se traduire par un partage de cette viande au sein de cette communauté assez large. Ceci aussi devait augmenter le prestige de celui qui procurait cette nourriture au groupe. Par exemple, à Çatal Höyük, 9000 BP, des crânes de taureaux sauvages sont encastrés dans le mur des habitations ou délimitent un espace (Figure 6). Cette recherche de prestige pourrait être à l’origine du maintien des aurochs en captivité qui aurait pu être la première étape nécessaire pour enclencher le processus de domestication.

Figure 6 : Plateforme funéraire décorée avec des cornes d'aurochs provenant du site de Çatal Höyük en Turquie (Hadad, 2018).

Bien avant son intégration dans les économies agro-pastorales, *Bos primigenius* a manifestement acquis un statut symbolique particulier au Néolithique au Proche-Orient.
Son importance dans les premières communautés néolithiques est illustrée par des bas-reliefs (Hadad, 2018), peintures murales (Çatal Höyük, Mellaart 1967), figurines en argile (Çayönü, Özdogan 1999), figurines en pierre (Nevali Çori, Schmidt 1998a).

Nous pouvons donc considérer la possibilité que le moteur de la domestication de l'espèce ait été son statut dans le monde spirituel des groupes humains provoquant ainsi les efforts visant à exercer un contrôle culturel sur l'aurochs.

2) Les bœufs domestiques :

Des aurochs de taille plus petite que les aurochs du Pléistocène du Nord de l'Europe ont peuplé au Néolithique l'Europe du Sud, l'Asie du Sud-Ouest et l'Afrique du nord (Davis, 1981). Il est difficile de distinguer les aurochs des premiers bœufs domestiques néolithique car les effets morphologiques d'une sélection ne sont visibles qu'après un certain nombre de générations. On ne peut pas dire que tous les bovins domestiques sont plus petits que l'aurochs, surtout que les vaches de la race italienne actuelle Chiannina font plus de 160 centimètres au garrot, les taureaux de plus de 180 centimètres et des extrêmes sont référencés à 190 cm.

Les peinture dans l'Égypte ancienne comportent des représentations de bovins exploités par les êtres humains, tels des taureaux qui font le labour des champs et des vaches encordées les unes aux autres, probablement pour faciliter la traite (Figure 7) (Guintard,C & Neron de Surgy, 2014).
Depuis les périodes historiques les représentations de taurins et les récits augmentent, ce qui prouve l'importance du rôle pratique et symbolique du bovin.

Figure 7: Illustration de peintures d'aurochs du Pléistocène (grotte de Lascaux), de l'Egypte ancienne et d'une vache actuelle. (https://en.wikipedia.org/wiki/Lascaux) (https://fr.wikipedia.org/wiki/Agriculture_dans_l%27%C3%89gypte_antique)

IV) Analyse génétique comparative des populations bovines anciennes et modernes:

La quantité des données de séquences obtenues à partir d'échantillons fossiles a augmenté considérablement au cours de la dernière décennie. Les progrès de l'archéologie, de la génomique, de la bioinformatique, et de la protéomique ont révolutionné notre compréhension de l'histoire et de la préhistoire. L'étude du passé brouille maintenant les frontières entre la science et les sciences humaines. Des méthodes statistiques rigoureuses et des techniques de calcul puissantes sont mises en œuvre pour comprendre les énormes quantités de données produites par toutes ces disciplines. Une connaissance approfondie de ces méthodologies, ainsi que des possibilités et des limites de chaque domaine, nécessitera de nouveaux niveaux de communication interdisciplinaire.

1) Les haplogroupes mitochondriaux bovins et leurs divergences :

Différentes analyses génétiques ont permis d'étudier la diversité génétique et l'évolution des bovins. Les premières études basées sur des séquences du génome mitochondrial ont suggéré que *B.taurus* et *B.indicus* aient divergé à une date estimé autour de 300.000 ans avant leurs domestications (Achilli et al., 2008, 2009). L'étude de la diversité mitochondriale des populations bovines confirme les données archéologiques et indique que les deux sous-espèces sont issues de deux événements de domestication distincts. La domestication de *Bos primigenius primigenius* dans le Croissant fertile serait à l'origine des populations de bovins domestiques eurasiatiques modernes et la domestication de *Bos primigenius namadicus* serait à l'origine des populations de zébus modernes (Achilli et al., 2008; J. A. Lenstra, 1999; Naik, 1978).

Une analyse des mitogénomes complets des bovins actuels a été effectuée en ne séquençant que les mitogénomes des individus dont une analyse préalable montrait qu'il devait porter une séquence mitochondriale distincte. Cette analyse a permis de construire l'arbre phylogénétique de ces mitogénomes et de dater les nœuds de cet arbre (Achilli et al., 2008). Cet arbre phylogénétique indique que les mitogénomes des zébus sont moins diversifiés que ceux des bovins taurins (Figure 8). En se basant sur l'ensemble des régions codantes du mitogénome, les auteurs ont appliqué un taux de mutations de 2 x 10^{-8} substitutions par nucléotide et par an pour estimer les âges des nœuds en utilisant une approche de maximum de vraisemblance (ML). Ceci correspond à 3172 ans par substitution dans les 15.4 kb considérées. Cette valeur avait été estimée en combinant un âge de divergence de 2 millions d'années entre les bisons (et les yaks) et les bovins taurins, basée sur les archives fossiles et en combinant quelques estimations plus anciennes de taux de mutations afin de garder une cohérence entre les différentes études (Achilli et al., 2008). Tous les zébus appartiennent à un seul haplogroupe I qui se serait séparé des taurins il y a 335 ka. Cet haplogroupe comporte deux sous-haplogroupes I1 et I2 qui auraient divergés il y a 31 ka.

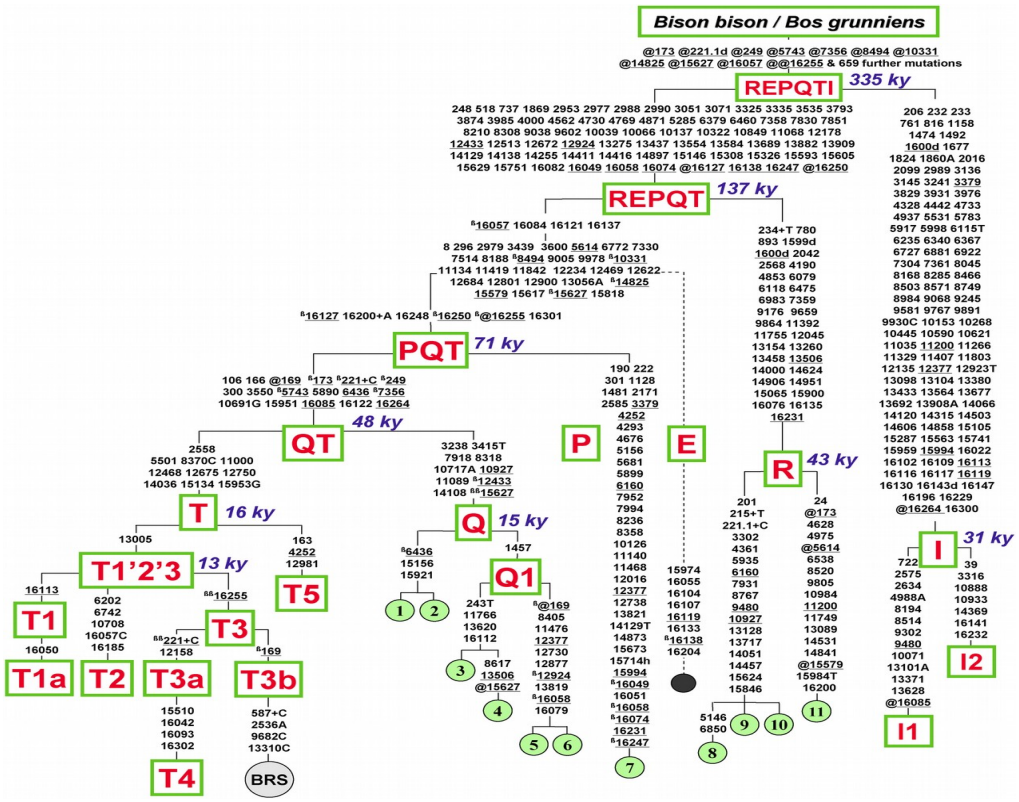


Figure 8: Arbre phylogénétique des différents haplogroupes mitochondriaux bovins (Achilli et al., 2009).

Les taurins présentent différents haplogroupes mitochondriaux distincts (P, R, Q et T). L'haplogroupe R, le plus divergent des haplogroupes mitochondriaux taurins, se serait séparé des PQT il y a à peu près 140 ka. Deux sous-haplogroupes R1 et R2 se seraient séparés il y a à peu près 43 ka. La branche P se serait séparée des QT il y a à peu près 70 ka. Cette divergence a conduit à la séparation de l'haplogroupe européen P des autres haplogroupes Q et T. 20 ka plus tard, les haplogroupes mitochondriaux T et Q ont divergé et chacun a développé des mutations qui ont permis l'apparition de nouveaux sous-haplogoupes (T2, T3 pour l'haplogroupe T et Q1, Q2 pour l'haplogroupes Q) (Figure 8).

Une étude plus ancienne portant uniquement sur la région hypervariable de l'ADNmt, un nombre élevé de bovins actuels de l'Europe, du Proche Orient et d'Afrique et typant systématiquement tous les individus analysés a été faite par Troy et al. En 2001 (Troy et al., 2001). Elle a permis de caractériser la fréquence des différents haplogroupes autour du bassin méditerranéen (Figure 9).

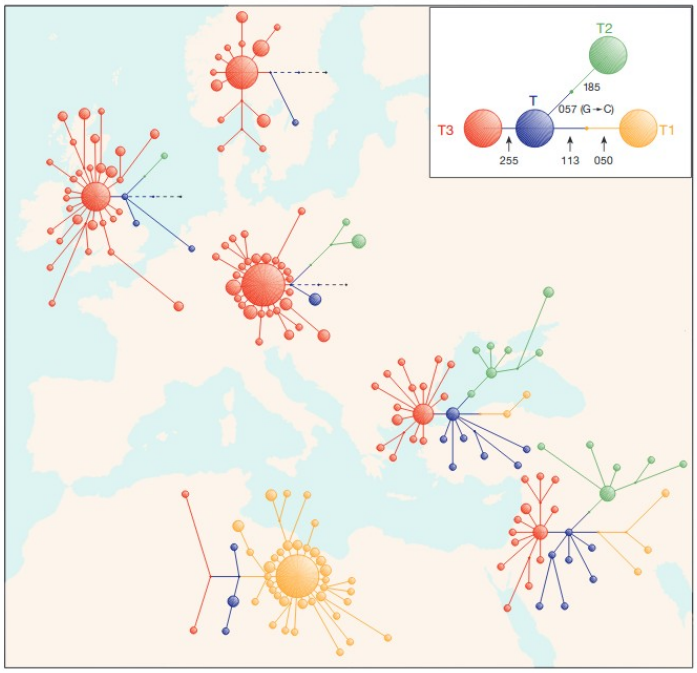


Figure 9: Distribution phylogéographique des haplogroupes mitochondriaux (Troy et al., 2001).

Les auteurs ont observé que la plus grande diversité des haplogroupes se trouvait en Asie du Sud-Ouest (ASO) tandis que l'haplogroupe T3 était dominant en Europe et l'haplogroupe T1 en Afrique du Nord. Ils ont interprété que la zone de plus grande diversité devait correspondre au centre de domestication ce qui était en accord avec les données archéologiques. La réduction de diversité génétique autour de l'haplogroupe T3 en Europe pourrait avoir été la résultante d'un effet fondateur concernant la population anatolienne initiale qui aurait été dispersée en Europe lors des migrations néolithiques. En ce qui concerne l'haplogroupe mitochondrial T1 en Afrique, cette origine est moins claire car les événements qui auraient pu documenter cet effet fondateur ne sont pas clairement caractérisés. A cette époque, la région hypervariable utilisée ne permettait pas d'identifier sans ambiguïté l'haplogroupe Q.

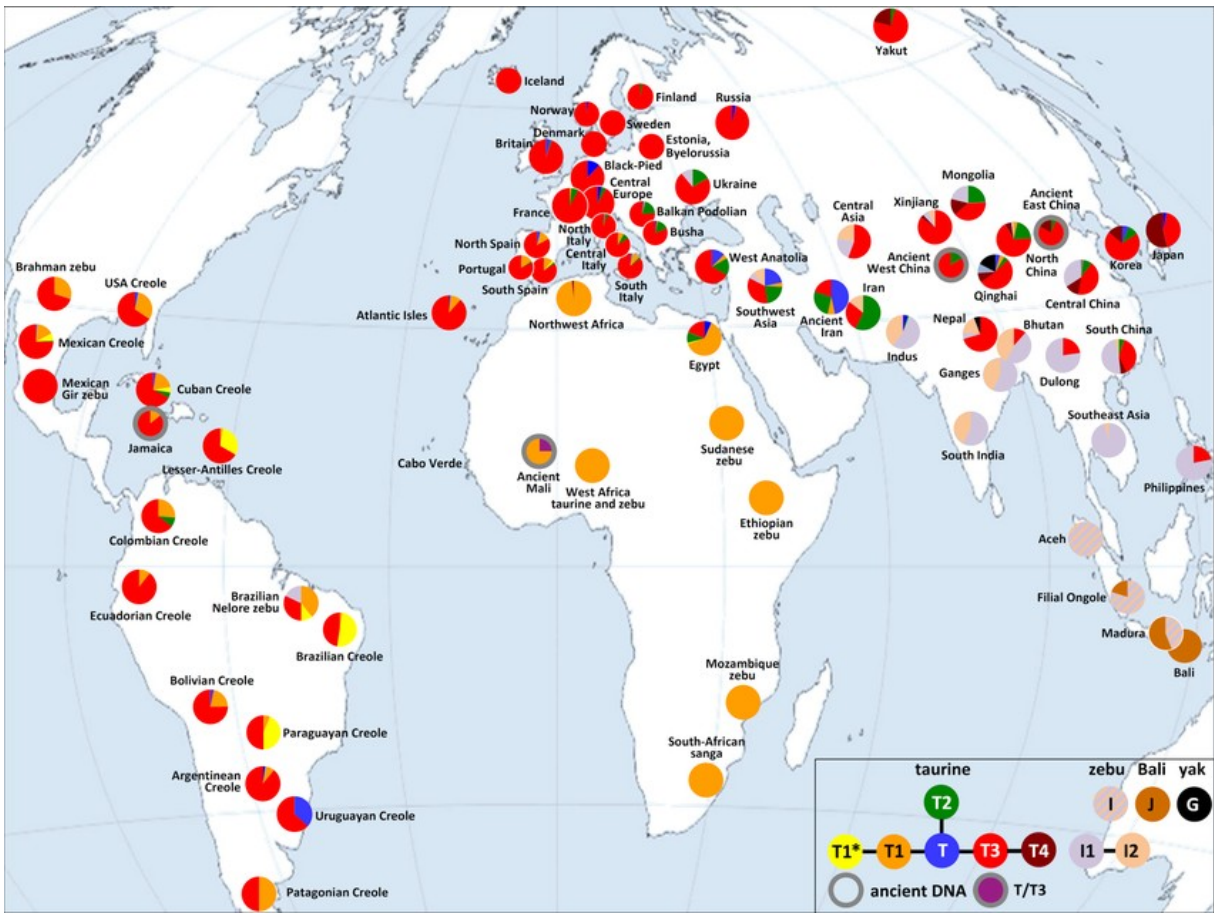


Figure 10: Distribution géographique des haplogroupes modernes (J. Lenstra et al., 2014a.

Par la suite, le typage des haplogroupes mitochondriaux a été étendu au cheptel bovin réparti sur différentes régions du globe (voir synthèse (J. Lenstra et al., 2014a)). L'haplogroupe T1 est présent en faible proportion dans les populations de la péninsule ibérique, en Sicile, au Sud et dans le centre de l'Italie (Figure 10). Cette faible proportion suggère qu'il y aurait eu une introgression du patrimoine génétique des bœufs africains dans les populations européennes autours de la Méditerranée, peut-être à l'époque historique (J. Lenstra et al., 2014a). La distribution de l'haplogroupe des bovins américains montre que l'haplogroupe mitochondrial T3 est majoritaire avec une présence minoritaire de l'haplogroupe mitochondrial T1. Ceci pourrait être dû à une origine ibérique du cheptel américain, en accord avec l'histoire de la colonisation de l'Amérique par les Espagnols à la fin du 15ème siècle. Par contre, cette image décrit la situation actuelle. Depuis la domestication, environ 10.000 ans se sont passés et les données génétiques actuelles ne reflètent pas forcément la situation du passé car des goulots d'étranglements et des effets de dérive génétique peuvent l'obscurcir et la diversité génétique perdue ne peut pas être retrouvée dans les données actuelles. Pour avoir un accès direct à la situation du passé, il faut obtenir de l'information génétique à partir des échantillons anciens.

L’haplotype P est la signature de l’aurochs européen et il est extrêmement rare chez les bovins actuels : jusqu’à l’année dernière, seules trois séquences mitochondriales de cet haplotype ont été déposées dans la banque des données et elles sont originaires de vaches Holstein en Chine et en Corée (Ajmone, 2010) ce qui suggère que les femelles aurochs européennes ont très peu contribué au pool génétique des bœufs modernes. Depuis, il a été montré que 46% des bovins de la race japonaise « shorthorn » portent l’haplogroupe P (Mannen et al., 2020). Il y a donc eu au moins un événement d’introgression en provenance d’une femelle aurochs qui a laissé des traces jusqu’à présent. Il est surprenant que cet événement ne concerne que les races asiatiques alors qu’il s’agit d’un aurochs européen. Soit il existe des populations européennes rustiques qui ont échappé jusque-là à la détection et qui portent aussi l’haplotype P, soit il y a eu un événement historique particulier impliquant des relations européo-asiatiques.

D’autres haplogroupes mitochondriaux non-T ont été trouvés en Europe et qui sont Q et R (Figure 10). La faible fréquence de ces haplogroupes a été à l’origine d’une hypothèse suggérant des événements de domestication en Europe, mais il a été montré que l’haplogroupe Q est proche génétiquement de l’haplogroupe T (Figure 10) ce qui suggère qu’ils appartiennent au même événement de domestication. Il a été suggéré que l’haplogroupe R ait une origine italienne, car les bœufs porteurs de l’haplogroupe R ont été trouvés exclusivement en Italie, le pays d’origine des auteurs qui ont conduit cette analyse (Bonfiglio et al., 2010). Cette observation a été interprétée par le fait que l’Italie était l’un des refuges européens pendant le dernier maximum glaciaire (LGM) et que les Alpes ont limité les expansions des animaux après le refuge. Ceci aurait contribué à une introgression des aurochs sauvages dans les bœufs domestiques, c’est-à-dire qu’il y aurait eu une femelle aurochs d’haplogroupe mitochondrial R qui a contribué au pool génétique des populations modernes (Achilli et al., 2008). Cependant, une étude faite par l’équipe de Bradley (Verdugo et al., 2019) a montré qu’un aurochs épipaléolithique marocain est d’haplogroupe mitochondrial R2 ce qui suggère probablement que l’haplogroupe mitochondrial R2 est originaire du Nord d’Afrique plutôt que d’Italie.

2) Les marqueurs nucléaires :

Le génotypage à haut débit a permis aux généticiens des populations d’utiliser des ensembles de marqueurs à l'échelle du génome pour analyser les histoires évolutives de nombreuses espèces (pour revue voir (Kwok, 2000). Les études sur les marqueurs nucléaires fournissent des informations plus complètes sur le processus de domestication, la migration et l’élevage des bovins. En effet, les études sur l’ADNmt sont limitées car elles nous permettent de suivre l’évolution des populations à travers leurs lignées maternelles et ne traduisent donc pas la diversité génétique des bovins actuels. Les séquences génomiques de différentes races bovines déposées sur les bases de données vont permettre d’étudier la diversité génétique au sein des populations (Boussaha, 2015), de comparer cette diversité avec les données génétiques d’échantillons fossiles, d’identifier alors les allèles ancestraux afin de suivre leurs évolutions et de déterminer les processus génétiques qui ont conduit à leurs diversités.

Les marqueurs moléculaires, comme les données génétiques mitochondriales, montrent une divergence profonde entre *Bos taurus* et *Bos indicus* (Decker et al., 2009). Decker et al en 2014 (Decker et al., 2014) ont montré que les bovins africains sont plus diversifiés que les bovins eurasiatiques. Comme les données archéologiques (Ajmone, 2010) ont montré que les bovins domestiques de l'Europe et de l'Afrique ont été importés par les migrants à partir du centre de domestication, on s'attendait que la diversité génétique soit comparable.
Cela peut être expliqué par deux scenarios : soit il y avait une sélection moins intense et moins dirigée en Afrique soit il y avait d'introgression de bovins d'origines différentes. Le cheptel bovin africain comporte différentes races à bosse. Ces dernières ne sont pas des races pures de zébu car ils présentent un ADNmt taurin (Bradley et al., 1996). De plus, des études basées sur les marqueurs du chromosome Y ont montré que ces derniers sont très diversifiés, en grande partie parce que beaucoup de races africaines sont des hybrides taurus-indicus (Decker et al., 2014).

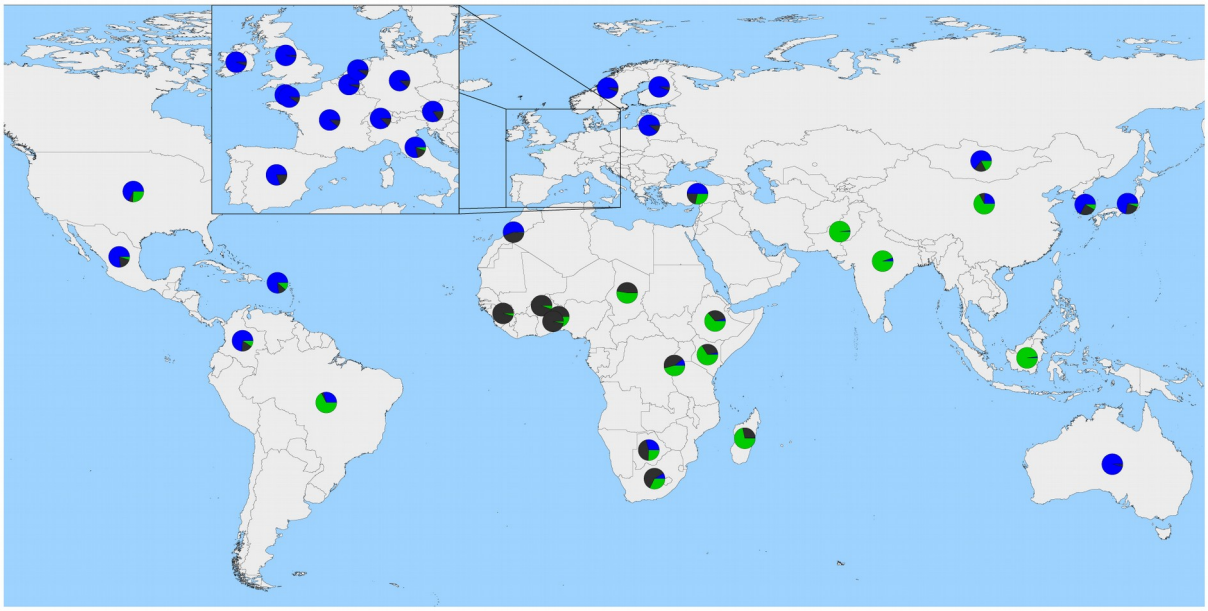


Figure 11 : Distribution des proportions des variants ancestraux dans le cheptel bovin (Decker et al., 2014) : La couleur bleue, verte, et noire représente respectivement les variants ancestraux eurasiatiques de Bos taurus, les variants ancestraux sud-asiatique Bos indicus et les variants ancestraux de Bos taurus africain.

Decker et al. ont mis en évidence par génotypage une forte introgression du patrimoine génétique des zébus dans les populations de bovins domestiques africains résultant essentiellement de l'apport des zébus mâles (Decker et al., 2014) (Figure 11). Les croisements entre le zébu et le bovin taurin ont dû permettre une meilleure adaptation aux conditions environnementales du continent africain. Un tel croisement permet de modifier la composition génétique et obtenir des hybrides avec un pouvoir productif plus élevé qu'un bœuf et plus résistants aux conditions environnementales qu'un zébu.

En effet, une analyse effectué par (Kwondo Kim & al, 2020) a permis d'identifier un événement majeur d'hybridation, en Afrique, entre les bovins taurins et les indicus datant d'environ 750 à 1050 ans. 16 loci ont été identifiés chez les animaux hybrides, et ont été liés à des adaptations environnementales africaines incluant des gènes liés à la reproduction et des gènes immunitaires et de tolérance à la chaleur. De plus, un variant très divergent chez les bovins taurins africains a été retrouvé chez les hybrides *Bos taurus/Bos indicus*. Ce variant est lié à la tolérance au trypanosomes transmis par la mouche Tsé-Tsé (Kwondo Kim & al, 2020).

Comme discuté auparavant, les données mitochondriales des bovins américains suggèrent que ces derniers seraient issus d'un déplacement des bovins ibériques vers le continent américain lors de sa colonisation au 15ème siècle (A.T. Primo, 1992). Cependant les marqueurs du chromosome Y montrent un autre profil (Decker et al., 2014). En effet, ce profil montre une introgression de gènes de zébu dans le cheptel américain qu'on ne voit pas dans les races européennes sauf dans quelques races italiennes (Figure 13) ce qui suggère que cette introgression aurait été faites spécifiquement en Amérique. Mukasa-Mugerwa, a indiqué en 1989 qu'au 19ème siècle des zébus asiatiques ont été introduits et qu'il y a eu des croisements entre ces derniers et les bœufs domestiques américains dans le but de produire des races hybrides plus résistantes aux conditions climatiques (Mukasa-Mugerwa, 1989).

V) Étude de la domestication des bovins par des approches paléogénomiques :

L'analyse des données des séquences hypervariables de l'ADNmt et de mitogénomes complets obtenus à partir d'échantillons fossiles d'aurochs et de bovins domestiques européens et proche-orientaux ont permis d'analyser les haplogroupes mitochondriaux anciens. Ces données ont permis de caractériser l'haplogroupe mitochondrial (P) qui a été retrouvé dans plusieurs ossements anciens bovins d'Europe et qui est donc considéré comme la signature de l'aurochs européen (Edwards et al., 2007a, 2010). Par contre, les haplogroupes mitochondriaux des bovins domestiques du Néolithique en Europe et en ASO appartiennent pour la plupart aux haplogroupes mitochondriaux T avec une distribution similaires aux populations modernes (Bollongino et al., 2006, 2012; Edwards et al., 2007b; Scheu et al., 2015; Troy et al., 2001). Ceci suggère que l'haplogroupe T est la signature mitochondriale des aurochs proche-orientaux qui ont été domestiqués. Ces données sont en accord avec les données archéologiques qui suggèrent que la domestication des bovins a eu lieu à partir des aurochs de l'ASO. Les ossements bovins trouvés dans les sites archéologiques néolithiques en Europe pourraient donc témoigner d'une diffusion de ces bovins domestiqués par les agriculteurs néolithiques d'origine anatolienne qui ont peuplé l'Europe à partir d'il y a environ 8500 ans (e.g., (Lazaridis et al., 2016)) ce qui a été confirmé par l'analyse de leurs mitogénomes (Edwards et al., 2007a; Scheu et al., 2015; Verdugo et al., 2019). En Europe, il y aurait donc eu coexistence entre bovins domestiqués d'origine anatolienne et les aurochs européens. Cette coexistence aurait pu entraîner des hybridations entre les populations sauvages et les bovins domestiques. Mais certains auteurs suggèrent que pendant cette période, les aurochs européens et les bœufs domestiques ont été séparés ce qui a limité les croisements. Edward et al, 2007 suggèrent que ces croisements sont rares et il ne devrait impliquer que des femelles domestiques et des mâles sauvages (Edwards et al., 2007b).

Comme discuté auparavant, on trouve très rarement l'haplogroupe P dans les bovins taurins modernes sauf en Asie, en particulier au Japon (Mannen et al., 2020). Ceci montre que les aurochs européens femelles ont très rarement contribué à une descendance domestique. L'analyse du génome nucléaire d'un aurochs britannique d'environ 6 750 ans du sud de l'Angleterre a mis en évidence que les aurochs ont contribué faiblement aux génomes des races bovines actuelles, en particulier à certaines races anglaises ce qui indique que quelques événements d'hybridation ont eu lieu dans le passé entre aurochs et bovins domestiqués (Park et al., 2015).

D'autres haplogroupes mitochondriaux non-T ont été trouvés en Europe et qui sont Q et R (Achilli et al., 2009). L'haplogroupe Q a été identifié avec juste un ou deux SNPs dans la région hypervariable mitochondriale amplifiée par PCR dans des conditions de prévention de la contamination des réactifs ou de « carry-over » insuffisantes compte tenu de la difficulté de produire des résultats fiables avec cette approche (Scheu et al., 2015). Avec ces réserves, les auteurs observent une fréquence assez importante de l'haplotype Q au Néolithique en ASO et dans le sud de l'Europe (Scheu et al., 2015).

Il a été suggéré que l'haplogroupe R a une origine italienne, car les bovins modernes porteurs de l'haplogroupe R ont été trouvés exclusivement en Italie (Bonfiglio et al., 2010). Par contre, la recherche systématique d'haplogroupes mitochondriaux rares n'a été effectuée pratiquement que sur le cheptel italien. Pour interpréter cette observation, les auteurs ont proposé que l'origine de l'haplogroupe R résultait d'aurochs présents dans le refuge italien pendant le dernier maximum glaciaire (LGM) et que les Alpes ont limité les expansions des animaux après le refuge. D'après ces auteurs, les aurochs italiens auraient laissé une trace mitochondriale dans le pool domestique italien. Mais Pereira Verdugo et al, 2019 ont détecté l'haplogroupe R2 dans un aurochs marocain épipaléolithique. Ce résultat suggère que la lignée R pourrait être originaire du Nord de l'Afrique. La présence de traces génomiques de *Bos indicus* dans les génomes des races italiennes (Decker et al., 2014) serait compatible avec un apport des races taurines africaines en Italie, ces races africaines portant souvent une hérédité mixte taurus-indicus. Cependant, pour confirmer ou infirmer cette hypothèse, il faut analyser d'autres échantillons fossiles du Nord de l'Afrique de différentes périodes historiques.

Le génome mitochondrial d'une mandibule, prélevé d'un fossé au Nord-Est de la Chine, daté à environ 10 500 ans a été séquencé. L'haplogroupe mitochondrial de cet individu appartient à un haplogroupe du genre Bos nommé «C», distinct de ceux retrouvés chez les populations modernes (Zhang et al., 2013). Cet haplogroupe est donc considéré comme spécifique d'une population d'aurochs d'Asie du Nord-Est. Les altérations sur la paire de mandibules prouvent que l'animal a probablement été soumis à un régime alimentaire particulier et un traitement social spécifique (Zhang et al., 2013). Ces altérations peuvent témoigner d'une interaction entre l'animal et l'être humain et peut-être d'un essai de domestication au Nord-Est de la Chine. Si ces tentatives ont eu lieu, elles n'ont toutefois pas laissé des traces à long terme.

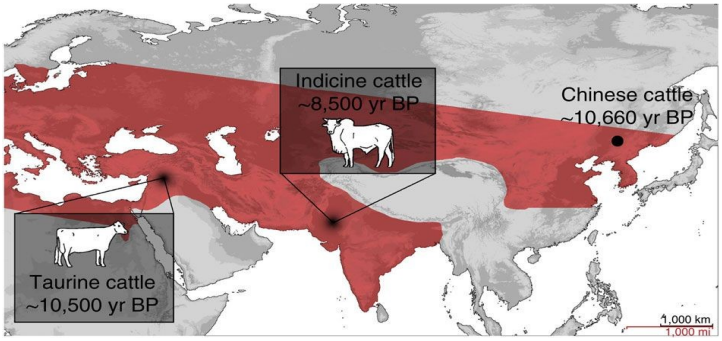


Figure 12 : Distribution géographique des centres de domestication de l'aurochs selon (Zhang et al., 2013).

Pour étudier l'histoire évolutive des populations bovines, il faut analyser des échantillons fossiles de périodes anciennes couvrant tout le processus de domestication. En effet, le génome mitochondrial ne reflète que la diversité maternelle et il ne permet d'observer les événements d'introgressions génétique d'aurochs dans les populations domestiques que si celles-ci sont introduites par des femelles sauvages ce qui fait de ce marqueur un marqueur limité. Ces introgressions impliquent un croisement entre un aurochs femelle et un taureau domestique, et les individus issus de ce croisement ont une plus grande probabilité de naître parmi les animaux sauvages, sauf si la femelle aurochs a été capturée et maintenue en captivité. Par contre, le croisement entre un aurochs mâle et une vache domestique a, quant à lui, plus de chances de donner une descendance qui naîtrait au sein de l'élevage et dont les gènes sont susceptibles d'être passés aux générations suivantes et pourraient encore être présents dans les populations de bœufs modernes. Pour déterminer l'éventualité d'une contribution génétique des aurochs dans les populations modernes, il est nécessaire d'analyser des génomes nucléaires.

Une analyse phylogénétique qui a impliqué plusieurs génomes entiers de bovins actuels et le génome d'un aurochs anglais a placé ce dernier sur une branche isolée du groupe des races européennes de bœufs domestiques modernes (Park et al., 2015). Par contre, étrangement l'aurochs anglais se trouvait sur une branche proche des zébus indiens et des bovins africains hybrides. La divergence avec les taurins européens modernes a été interprétée comme en faveur de l'hypothèse de l'origine proche-orientale des bovins domestiques européens. Toutefois, la proximité avec les zébus est troublante et paraît peu compatible avec les origines évolutives connues des aurochs européens. L'analyse effectuée reposant uniquement sur 10 000 SNPs alors que la séquence du génome complet avait été déterminée pourrait avoir biaisé les résultats.
L'analyse comparative du génome britannique et de plusieurs génomes de bœufs domestiques a suggéré qu'auraient été sélectionnés au cours de la domestication des gènes impliqués dans des fonctions immunologiques et neurobiologiques, des gènes du métabolisme et de croissance (Park et al., 2015).

Dernièrement, l'équipe dirigée par Dan Bradley (Verdugo et al., 2019) a analysé des données génomiques de différents échantillons anciens du Néolithique jusqu'au Moyen Âge. Ces échantillons sont originaires du Sud du Levant, d'Iran, d'Iraq, d'Asie centrale, d'Anatolie et des Balkans. Des échantillons de l'Europe de l'Ouest n'ont pas été inclus dans cette étude.

Le séquençage à faible couverture et l'appel des SNPs en se basant sur des données des puces couvrant la variabilité moderne a permis une analyse en composante principale (ACP) projetée (projection Procrustes) sur les données modernes des *Bos indicus*, *Bos taurus* européens, africains et hybrides africains taurus-indicus (Figure 13). La première composante principale (CP) a séparé *Bos taurus* des *Bos indicus* et la 2ème CP a séparé *Bos taurus* européen des *Bos taurus* africain. Dans cette analyse, l'équipe de Bradley a montré que l'hybridation entre les taurins et les indicus domestiqués dans deux régions différentes (Croissant fertile et Vallée de l'Indus, respectivement) se serait produites en ASO aux Âges des métaux, de manière concomitante avec un changement climatique présumé majeur appelé « l'événement 4.2k ».

Nous nous sommes particulièrement intéressés aux informations concernant les présumés aurochs dans ce jeu de données. La PCA suggère que l'aurochs britannique précédemment discuté, CPC98, est un peu plus proche des bovins néolithiques des Balkans et de certains bovins européens modernes que des bovins néolithiques anatoliens et de la plupart des bovins modernes européens. Cette donnée est clairement en contradiction avec l'arbre phylogénétique présenté dans l'étude de (Park et al., 2015) qui montrait que cet aurochs était plus proche des zébus et des hybrides taurins-zébus, ce qui montre qu'une partie des conclusions de l'étude de (Park et al., 2015) n'a pas été confirmée par les analyses ultérieures du même génome par la même équipe. Des présumés aurochs du Caucase, d'Anatolie et du Levant sont aussi intéressants à discuter. Il s'agit d'un échantillon néolithique anatolien provenant d'un site clef du Néolithique en Anatolie: Çatal Höyük (CH22) (environ 7500 ans) et de 2 échantillons d'Israël (Abu1, Abu2) d'environ 9000 ans ainsi que d'un échantillon arménien Gyu2 d'environ 7000 ans. Ils sont tous plus proches des bovins néolithiques domestiques d'Iran, d'Anatolie et d'Asie centrale que des bovins néolithiques des Balkans. Puisque les échantillons ont été considérés, dans cette étude, comme étant des aurochs, ce qui est jugé peu crédible par certains archéozoologues pour l'échantillon de Çatal Höyük, cela pourrait indiquer que les bovins domestiqués dans la région proviendraient bien d'une population sauvage présente dans la région. Par contre, cette étude repose uniquement sur des échantillons très mal conservés qui ont fourni peu de données génomiques. Il a donc été nécessaire de projeter ces données anciennes fragmentaires sur la variabilité moderne et compte tenu de la très faible couverture, le risque est important que les interprétations soient inexactes.

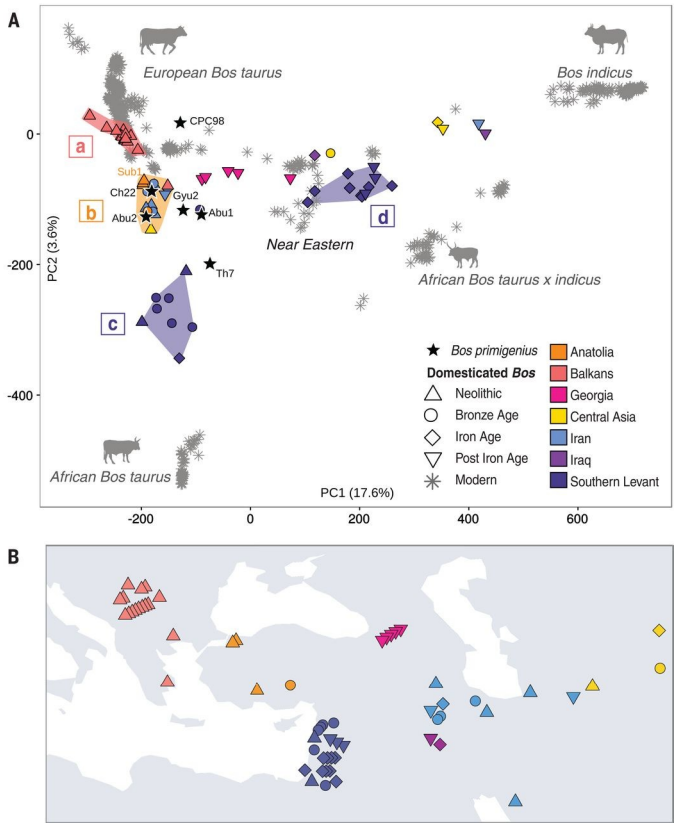


Figure 13: Analyse en composante principale projetée des échantillons anciens. A) les échantillons anciens sont projetées sur des données SNP bovins 770k. B) Distribution géographique des sites archéologiques des échantillons fossiles (Verdugo et al., 2019).

Le groupe, appelé **b** sur la figure, des bovins anatoliens et iraniens est proche de 4 aurochs de l'ASO. Cette proximité génétique suggère que ces aurochs appartenaient probablement à la population ancestrale qui a été domestiquée si tant est que les échantillons identifiés comme des aurochs le sont bien. Le groupe **a** des échantillons néolithiques des Balkans sont proches et chevauchent les bovins modernes européens. La proximité relative des génomes des échantillons des Balkans et des génomes anatoliens suggère que soit les échantillons des Balkans proviennent d'une population anatolienne qui aurait migré en Europe durant le Néolithique avec les populations humaines anatoliennes, soit pourrait refléter la proximité géographique des deux populations sauvages à l'origine du cheptel domestiqué. Une partie des bovins modernes ne chevauche toutefois pas les bovins anciens originaires d'Anatolie et des Balkans et se situe en position intermédiaire sur la CP2 avec l'aurochs britannique. Est-ce que l'introgression des aurochs européens dans le pool des bovins domestiques initiaux aurait contribué de façon très importante aux pool génétique des bovins actuels ? Les échantillons levantins sont divisés en deux groupes : un groupe **c** regroupant les échantillons levantins néolithiques et de l'Âge du Bronze. Ce groupe est proche de l'aurochs marocain épipaléolithique. Le deuxième groupe est constitué d'échantillons plus récents et proche des hybrides *Bos indicus-Bos taurus* du Proche-Orient. C'est ce résultat qui a donné lieu à l'interprétation mentionnée auparavant de l'hybridation des bovins levantins avec des zébus indiens après environ 2000 BCE, au moment du changement climatique.

Toutefois, ces interprétations doivent être relativisées car les données et les analyses ne permettent pas toujours de conclure de façon très robuste. Par exemple, il n'est pas clair sur cette figure que les échantillons du Levant ont contribué à la spécificité des bovins européens et pourquoi les échantillons néolithiques des Balkans ne sont pas similaires aux échantillons européens modernes. Il est à noter qu'une grosse partie de l'interprétation repose sur des différences visibles uniquement sur la CP2 qui ne représente que 3% de la variance des données. La CP1 qui représente 17% de la variance sépare surtout les taurins des zébus. Pour suivre correctement ce qui se passe, il faudrait certainement utiliser plus de variabilité génomique en utilisant des échantillons dont le génome a été séquencé plutôt que génotypé sur des puces à faible densité et visualiser cette variabilité en étudiant beaucoup plus que 3% de la variance. Si l'on prend ces analyses comme argent comptant les aurochs européens auraient beaucoup contribué à la variabilité des bovins domestiqués européens. Est-ce vraiment le cas ?

De plus, le pourcentage d'ADN endogène est très faible pour plusieurs échantillons qui jouent un rôle important dans les interprétations. Ainsi, 2 échantillons Abu1 et Abu2 contiennent si peu d'ADN endogène (0.2 et 0.1% respectivement) qu'ils procurent très peu de données et l'interprétation les concernant n'est pas robuste. En effet, les échantillons contenant peu d'ADN endogène sont plus susceptibles de produire des données faussées par les contaminations d'index lorsque les échantillons mal préservés sont séquencés simultanément avec des échantillons beaucoup mieux préservés. Le risque que le problème de contamination des indexes ait faussé les données publiées dans cette étude est élevé car en analysant les séquences brutes produites par l'équipe de Bradley, nous avons observé que certains échantillons contenaient un mélange d'indexes différents à l'extrémité 3' des fragments séquencés. Nous ne pouvons toutefois pas retracer si les contaminations d'indexes ont faussé les résultats des échantillons mal préservés car nous ne connaissons pas ni le plan de séquençage ni les indexes utilisés simultanément. Nous avons observé qu'il était indispensable d'utiliser des indexes double uniques pour séquencer des individus simultanément et de ne pas séquencer ensemble des échantillons avec des degrés de préservation très hétérogènes, ce que l'équipe de Bradley n'a pas fait pour le jeu de données produit ici.

D'autre part, n'étudier de la variabilité ancienne des aurochs que celle toujours présente dans les bovins domestiqués actuels en se focalisant sur le maximum de la variance génétique séparant les populations actuelles est une approche limitée et insatisfaisante et risque de gommer la variabilité ancienne qui n'aurait pas été transmise aux races actuelles issues du dernier siècle de sélection intensive. Pour bien caractériser le processus de domestication, nous avons visé à obtenir des données génomiques complètes pour éviter d'avoir recours à la projection sur la variabilité moderne et à bien sélectionner plusieurs échantillons fossiles couvrant une large distribution géographique et temporelle.

VI) Le genre *Bison* :

1) Les bisons actuels :

A l'heure actuelle, les bisons sont les plus gros animaux terrestres en Europe et en Amérique du Nord.

Le **bison d'Europe** ou **wisent** (***Bison bonasus***) vit actuellement dans la forêt où il se nourrit d'herbes, de feuilles, d'écorces et de branchages, mais des analyses isotopiques de bisons du Pléistocène suggèrent que la forêt n'est qu'un refuge parce que les bisons auraient été chassés des espaces ouverts (Bocherens et al., 2015a), En fait, ces animaux évitent les Humains et prennent la fuite à la vue des Humains à moins que ceux-ci ne se rapprochent trop (Haidt et al., 2018). Ils préfèrent les habitats mixtes et une alimentation mixte mais broutent quand c'est possible (Bocherens et al., 2015b; Markova et al., 2015). Ils vivent en petits groupes d'environ 10 à 20 femelles, veaux et jeunes de 2 à 3 ans d'âge. Après, les mâles adultes quittent le troupeau des femelles pour former des petits troupeaux de mâles. Les mâles ne défendent pas un territoire. Cette espèce atteint une hauteur d'épaule moyenne d'environ 180 cm (maximum ~ 210 cm) pour les mâles et d'environ 170 cm (maximum ~ 197 cm) pour les femelles (Semenov, 2014). Ses cornes se courbent vers le haut, avec une courbure plus prononcée chez les femelles (Krasińska & Krasiński, 2013).
Trois sous-espèces de *B. bonasus* ont été reconnues, à savoir *B. bonasus bonasus*, *B. bonasus* caucasicus et *B. bonasus hungarorum* (Pucek et al., 2004). Les populations contemporaines peuvent être divisées en deux lignées génétiques: la lignée des basses terres (descendant de *B. b. bonasus*) et la lignée des basses terres du Caucase (descendant de *B. b. bonasus* et *B. b. caucasicus*) (Pucek, 2004).

Le **bison américain** des plaines des Etats-Unis **(*Bison bison bison*)** habite les grandes prairies en grands troupeaux (parfois plusieurs centaines de têtes).
Les quantités importantes d'herbes dont ils se nourrissent les obligent de se déplacer régulièrement et par conséquent le phénomène de migration, petits déplacements journaliers et longues migrations saisonnières, est indispensable pour leur survie.
A la fin de l'été, après la saison de reproduction, les animaux s'amassent en grands troupeaux qui migrent. Les troupeaux consistent de femelles et leurs veaux, ainsi qu'adolescents, et, parfois, quelques mâles. Pendant la saison de la reproduction, les vieux mâles forment des harems et fécondent les femelles au début de la période alors que les plus jeunes mâles n'y arrivent qu'à la fin de la saison s'il y a des femelles pas encore fécondées.
Le bison américain possède un manteau d'hiver aux longs poils bruns foncé et un pelage d'été plus léger, d'un brun plus clair. L'animal peut atteindre 2 mètres au garrot, 3,6 mètres en longueur. Il pèse en moyenne entre 450 et 900kg. Les plus grands spécimens peuvent dépasser une tonne. La tête de l'animal est énorme. Les femelles comme les mâles sont dotés de deux cornes courtes et incurvées, qu'ils utilisent dans leur lutte pour obtenir un meilleur rang à l'intérieur du troupeau et pour la défense. Les bisons américains sont adultes à l'âge de trois ans et ont une espérance de vie de 18 à 22 ans, ou de 35 à 40 ans en captivité.

Il existe aussi une sous-espèce de **bisons au Canada (*Bison bison athabascae*)**, appelés « bisons de forêt », qui est adapté aux forêts boréales et de tremble avec leurs espaces ouverts et qui est spécialisé dans les herbes, cypéracées (carex), lichens, écorces etc.

La hiérarchie chez les mâles est fréquemment remise en cause, en particulier pendant la période de rut, où la compétition est grande pour les bisonnes reproductrices, les mâles de rang élevé ayant priorité pour la reproduction. Cette structuration explique que les femelles puissent se reproduire à l'âge de deux ou trois ans, tandis que les mâles ne se reproduisent pas avant d'avoir atteint une position élevée, généralement à l'âge de six ans.

Les bisons consacrent à leur hygiène une bonne partie de leur journée : ils se roulent dans des trous bourbeux ou dans le sable ou la poussière, toujours aux mêmes endroits, puis se frottent la tête et les flancs contre des branches, des rochers ou des troncs d'arbre, afin de se débarrasser de leurs parasites extérieurs. Cette activité de soin et de confort se transforme, en période de rut, en expression de leur surexcitation. Pendant la période du rut, les bisons mâles engagent des combats intenses. Ils déploient toute la force pour faire reculer l'autre. Mais la lutte a ses règles : les attaques latérales sont évitées parce que dangereuses et chacun attend que l'adversaire soit de face pour reprendre le combat, qui s'accompagne de mugissements, de piétinements et de forts renâclements.

Le bison américain possède 15 cotes et 4 vertèbres lombaires alors que le bison européen possède 14 cotes et 5 vertèbres. Les bisons européens adultes sont plus grands et plus minces dans leur constitution ayant des pattes plus longues que les bisons américains qui sont plus massifs. Les cornes pointues du bison européen font une courbe, ce qui le rend plus habile et experts dans les luttes à la différence du bison américain dont les cornes lui donnent un avantage pour l'attaque. La tête du bison américain est plus lourde et tombante et sa bosse plus développée que chez le bison européen.

Figure 14: Morphologie bison américain à droite et bison européen à gauche. (https://www.bisoneurope.com/le-bison-deurope-damerique/)

Les Bisons européens et américains sont classés en deux espèces différentes et malgré cela, ils sont complètement interfertiles et donnent une descendance « hybride » fertile et viable. Il existe d'ailleurs des troupeaux d'hybrides vivant en liberté dans le Caucase russe depuis les années 1950.

2) Histoire récente des bisons :

Les deux espèces de *Bison* sont passée par un goulot d'étranglement dans le passé récent.

Dès le 18è siècle les colonies de Virginie et de Pennsylvanie abattaient des milliers de bisons parce qu'ils détruisaient leurs clôtures et dévastaient leurs champs. Au 19ème siècle le massacre s'organisait. Le but de ce carnage était d'occuper des milliers d'êtres humains désœuvrés par la fin de la guerre de sécession. Mais l'objectif principal de l'extermination du bison est réduire les ressources des Amérindiens. De 1870 à 1875, 12,5 millions de bisons ont été tués officiellement recensés par la vente de leurs peaux. En 30 ans, au début du 20ème siècle, le nombre d'individus de *Bison bison* a diminué drastiquement à cause de la chasse extensive, essentiellement pour sa fourrure. Le cheptel initial de 70 millions de têtes était descendu, en 1889, à 1091 bisons américains vivant dans le monde.

Le bison européen était, au début de l'Holocène, reparti sur toute l'Europe, depuis le territoire de la France actuelle jusqu'à la rivière Volga en Russie. La population a diminué drastiquement avec la fragmentation de leur habitat, l'urbanisation pour se retrouver limités au Moyen Âge à quelques réserves naturelles. En particulier, il y avait une population, réservée et protégée pour la chasse royale, dans les forêts polonaises. Néanmoins, la taille et nombre des troupeaux ont continué à diminuer. Pendant la première guerre mondiale, les derniers bisons en liberté ont été tués par des soldats prussiens et par la population polonaise affamée. Le dernier animal a été tué en 1921. A partir de 1923, une association a été créée à Berlin qui a œuvré pour la restauration des bisons en réunissant des animaux des différents parcs zoologiques. Les aléas du Troisième Reich et de la guerre ont fait que ce projet n'a vu le jour qu'en 1952 dans la forêt de Białowieża, Pologne, à partir d'une douzaine d'individus de zoos.

3) Origine des Bisons :

Un élément caractéristique de l'ère du Pléistocène était la grande diversité de sa mégafaune. Beaucoup de ces animaux étaient déjà éteints avant la fin du Pléistocène (Barnosky, 2004), tandis que certains ont survécu jusqu'à l'Holocène. Parmi eux se trouvent également les bisons. La première espèce est peut-être apparue au Pliocène supérieur (durée: 3,6-2,58 Ma) ou au début du Pléistocène précoce (durée: 2,58-0,773 Ma) en Asie du Sud et en Chine (Akbar Khan et al., 2010). Le genre *Bison* a été largement étudié par les paléontologues en raison de la richesse des restes fossiles disponibles (Kirillova et al., 2013, 2015).Traditionnellement, les fossiles sont classés sur la base de la morphologie, ce qui rend souvent difficile la distinction entre les espèces (revue dans (Grange et al., 2018)). Cette difficulté résulte surtout de la fragmentation des restes, de la variabilité des caractéristiques morphologiques (Grange et al., 2018) et du dimorphisme sexuel (Drees, 2005).

Idéalement, l'identification taxonomique des restes de bisons basée sur la morphologie serait réalisée en considérant un grand nombre d'éléments squelettiques différents et bien conservés du même individu. Comme ceux-ci ne sont généralement pas présents, des crânes complets permettent la meilleure identification (Wilson et al., 2008). Les os longs des membres sont un autre bon identifiant. Cependant, comme les crânes, les humérus, les radius, les fémurs et les tibias sont souvent trouvés fragmentés ou sont complètement absents des archives paléontologiques, les os du pied les plus robustes et donc généralement mieux préservés sont fréquemment utilisés à la place. Bien que ceux-ci permettent une distinction entre les genres *Bos* et *Bison*, ils ne peuvent pas être utilisés pour distinguer de manière fiable les espèces individuelles de bisons européens du Pléistocène tardif (cité dans Grange et al., 2018, (Pacini, 1987) (Martin et al., 2018)).

Il est donc important de compléter l'analyse morphologique par une analyse moléculaire, afin de mieux comprendre la divergence des espèces.

Les restes fossiles ont permis d'identifier trois bisons archaïques: *Bison sivalensis*, *Bison palaeosinensis* et *Bison hanaizumiensis*. Au début du Pléistocène, le genre *Bison* s'étendait de l'Europe en Asie. Cette répartition a inclus deux espèces différentes bien distinguées. *Bison priscus* ou bison des steppes, éteint depuis l'Holocène et le *Bison schoetensacki* ou bison des bois, éteint depuis le Pléistocène moyen (Palacio et al., 2017). *B. priscus* est aussi connu par les peintures rupestres des grottes préhistoriques comme celles de Lascaux qui témoignent de sa large distribution allant de l'Europe de l'Ouest à l'Amérique du Nord en passant par l'Asie centrale.

4) Les Bisons et leur évolution en Amérique de Nord :

En Amérique du Nord, il y avait *Bison antiquus*, *Bison latifrons* et *Bison occidentalis* et le bison des steppes ou *Bison priscus*. Le bison des steppes (*Bison priscus*) est une espèce disparue de bison à longues cornes ayant vécu au Pléistocène. Le bison des steppes était plus haut et plus massif que les bisons actuels. Ses cornes étaient longues et dirigées vers le haut, ses épaules étaient puissantes. On pense que sa robe était plus semblable à celle du bison d'Europe qu'au bison d'Amérique du Nord. Les plus gros mâles devaient atteindre environ 2 mètres au garrot pour plus d'une tonne.
Leurs cornes atteignaient parfois plus de 1,50 m d'envergure. Il devait certainement constituer de grands troupeaux lorsque les prairies où ils étaient abondantes et devaient être chassé par de nombreux prédateurs. Des restes fossiles de grands bisons de steppes sont abondamment retrouvés dans les sites archéologiques du Pléistocène en Asie, en Europe, dans le Détroit de Béring et en Amérique du Nord. Après sa répartition et son existence pendant plus qu'un million d'années, *Bison priscus* s'est éteint en Europe pendant la période du Dryas vers la fin du Pléistocène (Kirillova et al., 2013), mais a évolué vers *B. bison* sur le continent américain pendant l'Holocène.

Les changements climatiques qui ont eu lieu au Pléistocène et principalement au niveau de Détroit de Béring qui se manifestaient par la diminution du niveau d'eau ont permis l'échange entre la faune d'un continent à un autre. L'expansion des steppes de toundra a constitué un refuge terrestre allant de l'Est de la Sibérie au Nord-Ouest du Canada. A la fin du Pléistocène, il y a eu une grande extinction de la faune du Détroit de Béring qui pourrait être expliqué par les changements climatiques lors du dernier maximum glaciaire (Froese et al., 2017).

Les données archéologiques suggèrent que les bisons des steppes auraient colonisés le Détroit de Béring au Pléistocène moyen (il y a entre 300 000 ans et 130 000 ans). Les restes fossiles indiquent que sur le continent américain, plusieurs lignées de *Bison priscus* ont évolué différemment en plusieurs morphologies classifiées parfois comme des espèces propres, comme *Bison latifrons* et *Bison antiquus*. Le premier fossile de *Bison latifrons* a été trouvé dans le Kentucky. Cette espèce fossile a été décrite pour la première fois par Harlan en 1825, puis par Leidy en 1852. Des fossiles de *B. latifrons* ont été retrouvés dans de nombreux États américains dont la Géorgie, le Californie, la Floride, le Dakota du Nord et jusqu'au Canada. Cette espèce occupait des régions boisées et formait des petites hordes lors de la dernière glaciation. Ses grandes cornes lui auraient servi à se défendre contre les prédateurs tels que le lion américain, ou encore certaines espèces du genre *Smilodon*.

Le *Bison latifrons* se serait éteint à la fin du Pléistocène entre 30 000 et 21 000 avant notre ère. Cette espèce de bison aurait été remplacée par la suite par *Bison antiquus* vers environ 25 000 ans avant notre ère. Ce dernier aurait évolué en une autre espèce *Bison continentalis* qui est l'ancêtre du *Bison bison* américain moderne qui était apparu il y a environ 10 000 à 5000 ans (Lott, 2002). Il est à noter que cette multiplication des espèces fossiles n'a probablement pas une base génétique solide (Froese et al., 2017).

5) Les Bisons en Eurasie :

Les bisons sont apparus en Europe au Pléistocène inférieur en se repartissant de l'Est à l'Ouest du continent. Ces bisons étaient, tout d'abord, décrits sous le nom de *Bison schoetensacki*, un spécimen de petite taille ayant de petites cornes. Ses restes fossiles ne sont pas abondants, ils sont même rares. L'espèce se serait éteinte à la fin du Pléistocène moyen. *Bison schoetensacki* était généralement semblable au bison d'Europe actuel. Par rapport à *Bison priscus*, *Bison schoetensacki* était de taille plus petite ou similaire, mais avec des os des membres plus fins, et des cornes plus courtes et de forme différente. Les restes fossiles indiquent que les bisons européens étaient plus grands avec des cornes plus massives pendant l'ère glaciaire de Mindel, qui est la deuxième glaciation du Quaternaire (-650 000 à -350 000 ans environ).

Le *Bison priscus* est apparu en Europe pendant la glaciation de Riss (-300 000 à -130 000 ans environ) et de Würm (-115 000 à -12 000 ans environ). Degerbol et Ivergen en 1945 indiquent qu'à la fin du Pléistocène, une lignée de *B. priscus* serait devenue plus petite (cité dans (Kowalski, K, 1967)). Cette lignée observée au Danemark et nommée *B.bonasus arbustotundrorum* est morphologiquement plus proche du bison européen actuel (cité dans (Kowalski, K, 1967)).
En Europe de l'Ouest, *B. priscus* a été remplacé à l'Holocène par le bison d'Europe ou *Bison bonasus* dont les vestiges ostéologiques morphologiquement distincts apparaissent pour la première fois dans les archives fossiles au début de l'Holocène, environ 12000 à 11700 ans avant présent (Verkaar et al., 2004). Par contre, ni la date ni le processus n'avaient été clairement élucidés par les analyses paléontologiques.

Dans une étude approfondie, un ensemble de données morphométriques pour les os métacarpiens III-IV a été comparé aux données génétiques (Grange et al., 2018). Le métacarpe est l'un des os de bison les mieux conservés et les plus communs provenant d'assemblages fossiles, avec une diversité morphologique et une corrélation suffisante avec la morphologie générale du squelette pour permettre une étude de l'évolution du squelette. Pour éviter les effets confondants du dimorphisme sexuel, seuls les os attribués à des individus de sexe féminin ont été pris en compte. Les résultats n'ont pas pu confirmer l'existence de différences morphologiques significatives dans les populations européennes du Pléistocène tardif. En général, les os métacarpiens de *B. bonasus* sont très différents de ceux de *B. priscus*, qui sont globalement plus grands avec des dimensions diaphysaires plus grandes et des extrémités distales plus épaisses. En outre, la comparaison des mesures de *B. priscus* provenant de sites sibériens, yakoutiens et ukrainiens entre eux et avec le *B. bonasus* existant a confirmé de manière convaincante une homogénéité morphologique claire de la série fossile de *B. priscus* ainsi que sa grande taille (Grange et al., 2018).

Depuis le début du 20ème siècle, plusieurs hypothèses ont été discutées pour comprendre les origines et les liens entre les bisons européens du Pléistocène et le bison européen moderne.
Il a été supposé qu'il y avait deux lignées au Pléistocène, une des deux étant celle qui mènerait du *Bison schoetensacki* directement au bison européen moderne incluant ainsi les bisons de petite taille. Le bison *priscus* de grande taille aurait évolué de façon différente sans laisser de descendance en Europe et en Asie à la fin de la dernière période glaciaire 9 500 ans AP (Kirillova et al., 2015). Selon Gromova (1965) (cité dans (Kowalski, K, 1967)), il n'y avait qu'une seule lignée en Europe au Pléistocène qui a amené du *B. schoetensacki* au *B. priscus* jusqu'au *B. bonasus* européen moderne. Cependant, l'absence des restes fossiles d'espèces intermédiaires entre *B. priscus* et *B. bonasus* a conduit les auteurs de supposer que le bison européen moderne descend directement de *B. priscus* (Kowalski, K, 1967).

Pour comprendre l'origine de bisons européens modernes et résoudre les questions relatives à leurs évolutions, il a fallu attendre jusqu'à l'année 2016 puis 2018, où deux articles d'études basées sur les données d'ADNmt des spécimens anciens de bison ont été publiées (Grange et al., 2018; Massilani et al., 2016).Ces études ont inclus des spécimens provenant de différents sites paléontologiques et archéologiques allant de l'Ouest à l'Est de l'Europe et du Caucase à la Sibérie du nord est. Les spécimens les plus récents proviennent de Pologne et du Nord du Caucase et datent du début du 20ème siècle, juste avant l'extinction de *B. bonasus* dans le milieu naturel. Les plus anciens spécimens datent d'environ 45 000 ans AP et proviennent d'un site paléontologique au nord du Caucase. L'analyse de l'ADNmt conservé dans des restes fossiles de bisons d'Eurasie couvrant les 50 000 dernières années a permis de retracer les dynamiques de population qui ont eu lieu pendant le Pléistocène supérieur et l'Holocène, y compris les migrations, les extinctions et les remplacements de populations.

Il a été montré qu'il y avait, en Europe de l'Ouest, une alternance entre des lignées mitochondriales de type *B. bonasus* et des lignées mitochondriales de type *B. priscus* qui étaient liées aux fluctuations climatiques. Le *Bison bonasus* était répandu en Europe de l'Ouest lorsque le climat était plus tempéré menant à une végétation plus boisée, semblable à celle de l'Europe de l'Ouest actuelle. Par contre, la population de bisons des steppes originaires du nord de l'Eurasie était prédominante en Europe occidentale pendant les périodes plus froides du Pléistocène supérieur avec leurs environnements ouverts. Les auteurs ont avancé l'hypothèse que ces fluctuations peuvent avoir été enregistrées dans les peintures rupestres paléolithiques, en particulier dans la grotte de Chauvet (Figure 15) qui avait été occupée par les humains pendant une longue période et où deux types de bisons distincts sont représentés (Massilani et al., 2016).

Figure 15: Peinture préhistorique Grotte de Chauvet, Ardèche France. (Imprimé avec l'autorisation du Centre National de Préhistoire, France. (Copyright: Ministère de la Culture et de la Communication, archeologie.culture.fr/chauvet; Arnaud Frich, Centre National de Préhistoire / MCC).

Les deux peintures ont été datées au radiocarbone. La date a été estimée entre 38500 et 34100 ans pour la bison supérieur, et à 36300–34600 ans pour le bison inférieur (Quiles et al., 2016). Massilani et al. ont proposé que le bison de grande taille dans la partie supérieure représenterait une ancienne lignée de *B. bonasus* avec une tête très positionnée, des cornes courbées, une bosse moyennement grande et une crinière faible, et des proportions corporelles plutôt équilibrées entre l'avant et l'arrière. La partie inférieure représenterait *B. priscus* avec sa grosse bosse, sa tête basse, sa crinière abondante et ses cornes en forme de croissant, bien que l'image soit quelque peu estompée, la forte inclinaison de la ligne arrière et l'arrière-train plus fort peuvent être distingués.

Dans cette étude (Massilani et al., 2016), l'alignement des séquences de l'ADNmt a montré que les groupes de bisons européens du Pléistocène supérieur présentent la plus grande diversité génétique. Cette observation confirme les résultats des données paléontologiques qui indiquent qu'au Pléistocène supérieur, jusqu'au dernier maximum glaciaire il y a environ 20 000 ans, les troupeaux des bisons eurasiatiques étaient nombreux et occupaient un territoire géographique important (Markova et al., 2015).

Un échantillon ancien d'environ 13 000 ans AP (14300-13800 BP) provenant du site archéologique de Kesslerloch en Suisse à la frontière avec l'Allemagne du Sud-Ouest possède une signature génétique mitochondriale identique à celle des spécimens modernes. Cet échantillon semble donc appartenir à la lignée de bisons européens ayant survécu en Europe jusqu'au début du 20ème siècle. Ceci a été vérifié par l'analyse phylogénétique: En effet, les échantillons modernes du Caucase (jaune sur la figure 16) et de Pologne (vert sur la figure 16) datant de la fin du 19ème jusqu'au début du 20ème siècle groupent en une branche très peu diversifiée à l'intérieur d'un clade nommé Bb2 (*Bison bonasus 2*) (Massilani et al., 2016). L'échantillon du site Kesslerloch (KSL) de la transition Pléistocène-Holocène groupe avec les wisents modernes sur cette branche ce qui confirme donc que ce spécimen appartient à la population ancestrale directe des *Bison bonasus* modernes.

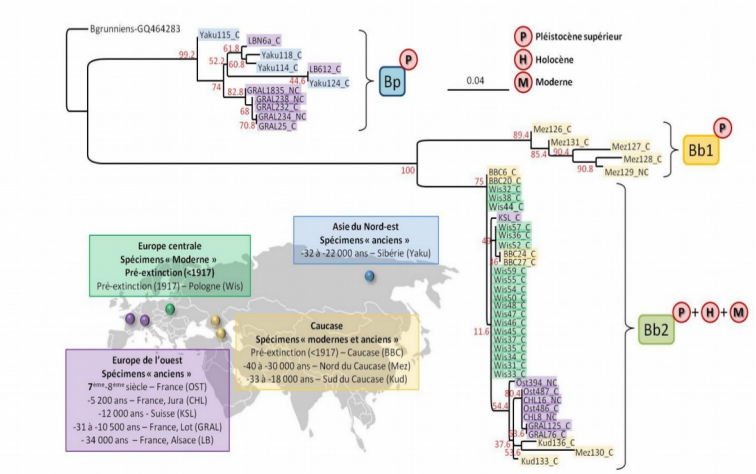


Figure 16: Arbre phylogénétique de maximum de vraisemblance des séquences de la région hypervariable des spécimens de bisons européens anciens (Massilani et al., 2016).

Les échantillons anciens d'Europe de l'Ouest (violet sur la figure 16) qui datent de l'Holocène sont groupés ensembles au sein du groupe Bb2 et divergent de la lignée originaire des bisons modernes. Les échantillons anciens de la France datant de l'Holocène et du Moyen Age sont assez divergents les uns par rapport aux autres et groupent avec les échantillons du Pléistocène supérieur dont deux originaires du sud du Caucase (Kud136 et Kud133) datant de 33 et 18 000 AP et un du nord du Caucase (Mez130) datant d'environ 40 000 ans. Les échantillons du Pléistocène analysés dans cette étude appartiennent aux 3 différents groupes de bisons présentés sur l'arbre phylogénétique (Figure 16). En effet, un autre groupe, distinct du groupe Bb2 contenant les échantillons de bisons européens les plus récents, a été nommé Bb1. Celui-ci, très diversifié, englobe les échantillons anciens du Pléistocène supérieure originaires du nord de Caucase provenant du site Mezmayskaya (Mez, jaune sur la figure 16).

Les échantillons provenant du nord-est de la Sibérie et de l'Europe de l'Ouest datant du Pléistocène supérieur sont groupés au sein du groupe des *Bisons priscus* nommé Bp. Ce dernier est plus distant des deux autres groupes (Bb1 et Bb2). Le groupe Bp est le groupe des bisons de grande taille dont les peintures historiques de la grotte de Lascaux, en France, témoignent de sa présence au Pléistocène en Europe de l'Ouest. Les *Bison priscus* ou bison des steppes ont peuplé les steppes froides de l'Ouest jusqu'à l'Alaska en passant par la Sibérie, comme cela peut être déduit de la proximité génétique des bisons de Yakoutie et de France (Massilani et al., 2016).

La majorité des échantillons anciens français provenant du site « l'Igue du Gral » datant de 31 à 15 000 ans appartiennent au groupe des *Bison priscus*. Cela confirme la présence de *Bison priscus* en France jusqu'à la fin du Pléistocène. Par contre, deux échantillons du même site datant de l'Holocène d'environ 10 000 ans appartiennent au groupe Bb2 des *Bisons bonasus* modernes ce qui permet de conclure que sur le site de l'Igue du Gral, il y a eu un replacement des populations de bisons durant le réchauffement climatique pendant la transition Pléistocène-Holocène.

De plus, l'arbre phylogénétique suggère qu'il y avait une réduction de la diversité génétique des bisons d'Europe qui coïncide avec le changement climatique et la transition Pléistocène-Holocène. Ce changement climatique est corrélé avec la disparition des *Bison priscus* qui sont plus adaptés au climat froid du Pléistocène et le repeuplement de l'Europe de l'Ouest et plus précisément de la France par de nouvelles populations plus adaptées au climat doux de l'Holocène (Massilani et al., 2016). Par contre, Soubrier et al ont analysé des échantillons provenant essentiellement de l'Oural et ont proposé une alternance inverse ou les Bb1 étaient plus abondants dans les périodes froides et les *Bison priscus* étaient plus abondants dans les périodes plus chaudes (Soubrier et al., 2016). Les alternances de populations varient en fonction de la géographie et donc vraisemblablement en fonction de l'impact des changements climatiques sur les écosystèmes locaux. Soubrier et al ayant extrapolé les données de l'Oural à la situation en France, en absence de données, ont interprété de manière erronée les variations affectant les morphotypes représentés dans les peintures rupestres en France.

La divergence de la région hypervariable mitochondriale des *B. bonasus* Bb1 et Bb2 a été estimée dans cette étude à environ 428 000 ans, alors que la date de divergence des *B. pricus*, qui montre une plus grande diversité, n'est que d'environ 149 000 ans (Figure 17).
Cela suggère que bien qu'étant présent depuis plus longtemps en Europe, l'ancêtre du *B. bonasus* est passé par un goulot d'étranglement sévère avant la fin du Pléistocène.

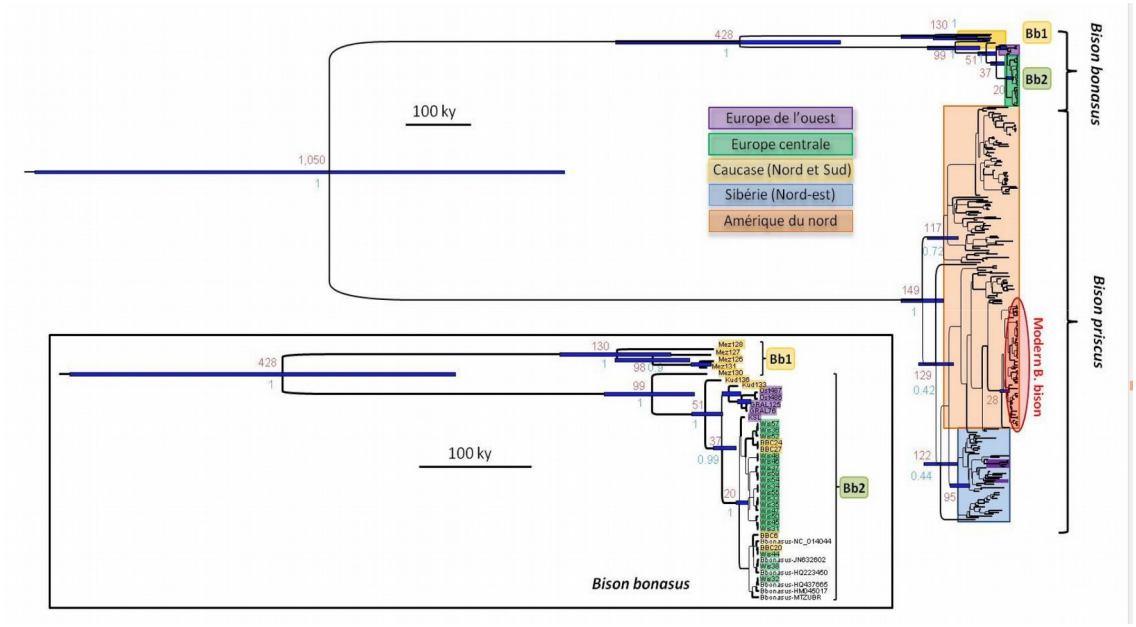


Figure 17: Arbre de phylogénie bayésienne de la région hypervariable des bisons anciens et modernes (Massilani et al., 2016).

Il a été montré que les bisons européens et américains sont divergents au niveau des lignées mitochondriales mais le sont beaucoup moins au niveau génomique (Hassanin et al., 2013; Verkaar et al., 2004).

Pour comprendre l'histoire évolutive du bison européen au Pléistocène supérieur à l'Holocène, un autre article a été publié en 2018 (Grange et al., 2018) reprenant et approfondissant les études basées sur des mitogénomes complets d'un nombre plus élevé d'échantillons anciens de bisons en utilisant aussi ceux publiés par une autre équipe (Soubrier et al., 2016).

Dans cette étude, les dates de divergence des différentes branches de *Bovina* ont été réestimées via une analyse bayésienne de tous les mitogènomes publiés à ce jour-là et de 86 nouveaux mitogénomes complets de bisons anciens provenant de zones complémentaires et chevauchantes d'Europe et d'Asie, ainsi qu'en Amérique. De plus, les données de séquences génomiques qui suggéraient une hybridation interspécifique entre *Bos* et *Bison* (Soubrier et al., 2016) ont été revisitées à la lumière des génomes complets récents publiés entre temps (Gautier et al., 2016; Węcek et al., 2016). Le flux génétique entre le genre *Bison* et *Bos* a été étudié en comparant des génomes modernes et anciens de Bisons.

La figure 18 représente l'arbre phylogénétique bayésien des bisons anciens et modernes. Cette figue résume les différentes estimations d'âge de divergence des lignées mitochondriales ainsi que les deux hypothèses expliquant le schéma évolutif des deux lignées de bisons. La couleur représente l'origine et la période de chaque spécimen analysé. Il s'agit de 5 régions géographiques (Europe occidentale et centrale, région Caucase-Oural, Sibérie et continent nord-américain), les 3 dernières phases climatiques correspondant aux 3 étapes isotopiques marines MIS. Le MIS1 correspond approximativement à l'Holocène mais commence il y a environ 14 500 ans avec un climat tempéré stable en Europe, en Asie du Nord et en Amérique du Nord à l'exception de la chute de température du « Dryas récent» (il y a environ 12 500-11 500 ans). Le MIS2, à partir d'il y a environ 29 000 ans, correspond à la période finale très froide et stable du Pléistocène connue sous le nom de Dernier Maximum Glaciaire (LGM). Le MIS3, commençant il y a environ 57000 ans correspond à une période généralement plus douce avec des baisses régulières de température.

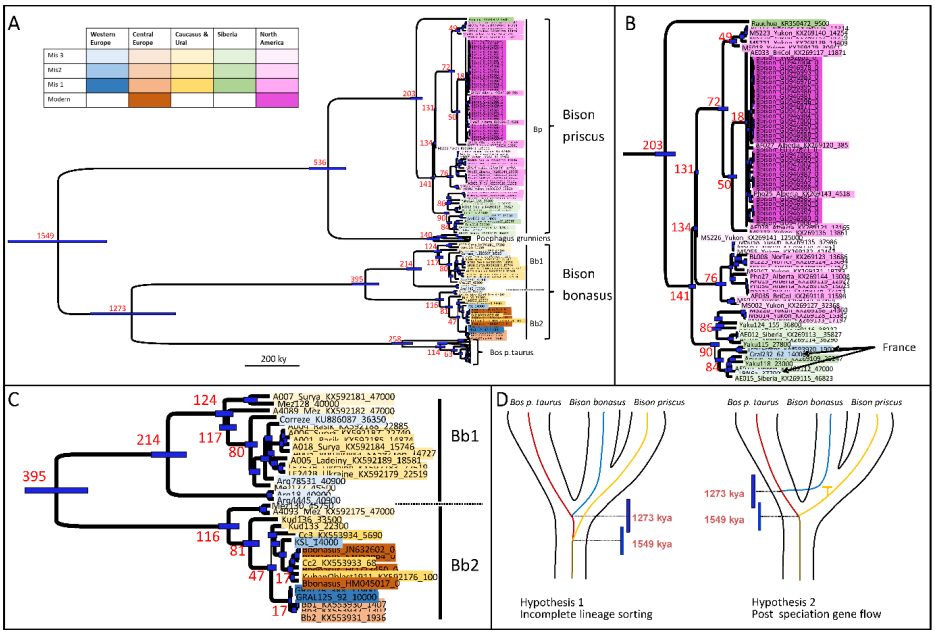


Figure 18: Phylogénie bayésienne de mitogénomes complets de genre Bison et Bos d'après (Grange et al., 2018).

La partie A de la figure 18 montre une affinité plus étroite entre les mitogénomes du groupe *Bison priscus/Bison bison* avec ceux du Yak et entre les mitogènomes de *Bison bonasus* avec ceux de l'aurochs. Dans cette étude et comparant avec les dates publiées par Massilani et al. en 2016, les estimations d'âges des ancêtres communs les plus proches ont légèrement changé à cause de l'augmentation de l'échantillonnage et de l'inclusion de mitogénomes de bisons anciens couvrant une large échelle spatio-temporelle. En effet, l'inclusion d'une plus grande diversité de séquences anciennes tend à améliorer la précision de l'estimation des taux évolutifs des séquences et ainsi à augmenter l'âge des nœuds par rapport aux estimations précédentes (Figure 17, 18).

Dans cette étude, l'âge de la racine de tous les mitogénomes de *Bovina* est estimé à environ 1,55 million d'années (Ma), avec une HPD de 95% entre 1,74 et 1,37 Ma alors que l'âge du nœud entre les lignées de mitogénomes taurines et des bisons européens est estimé à environ 1,27 Ma. Toutes ces dates sont incluses dans les dates estimées de la période de divergence des lignées *Bison* et *Bos* qui semble s'être produite entre 1,7 et 0,85 Ma à travers un processus de spéciation impliquant une période prolongée de flux génétique limité, comme déduit des analyses comparatives de génomes modernes taurin et de bison européen (Gautier et al., 2016). Ainsi, il a été montré que le schéma évolutif des deux lignées de mitogénomes de bison résulte très probablement d'un tri incomplet des lignées au sein d'une métapopulation ancestrale de *Bovina* au cours d'une période prolongée de spéciation plutôt que d'un flux génique ultérieur après la fin du processus de spéciation. En effet, cette dernière hypothèse exige qu'un événement de spéciation très rapide ait eu lieu précisément dans la courte période entre les deux événements de radiation, ce qui semble peu probable (voir la partie D de la figure 18 et la figure 19).

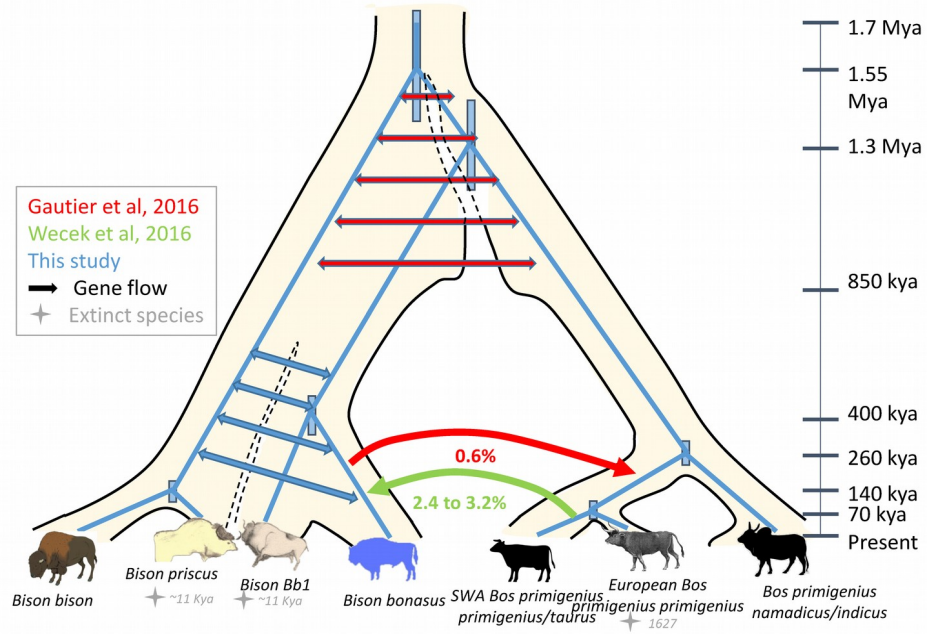


Figure 19: Représentation schématique de l'arbre évolutif des génomes de Bovina (Grange et al., 2018).

L'arbre récapitule les divers événements évolutifs déduits des analyses du mitogénome et des génomes nucléaires. Les lignes bleues représentent les lignées mitochondriales en utilisant les estimations de date et leurs densités postérieures les plus élevées estimées à partir des analyses bayésiennes. Les cadres beiges autour de ces lignes représentent les populations correspondantes déduites des analyses génomiques. Les lignes pointillées indiquent l'incertitude concernant l'étendue de la séparation des populations et/ou des espèces subissant encore un flux de gènes bidirectionnel. Les périodes de divergence estimées et proposées avec un flux génétique limité sont représentées par des séries de flèches à double face, tandis que les divers flux de gènes directionnels identifiés sont représentés par des flèches à simple face à côté de la fraction des génomes qui résulte de ce ou de ces événements de flux de gènes. Le code couleur indique l'étude d'origine des données recueillies comme indiqué dans l'encadré. Les dates estimées pour les différents événements évolutifs sont indiquées sous la forme d'une échelle de temps linéaire sur le côté droit. Les lignées/espèces éteintes sont indiquées par une étoile à côté de leur période d'extinction estimée.

Les analyses morphométriques et génomiques révèlent que les bisons eurasiatiques appartenant à différentes lignées de *Bison priscus* et *Bison bonasus* ont maintenu des trajectoires évolutives parallèles avec un flux génétique pendant une longue période de spéciation incomplète qui a cessé seulement après la migration de *B. priscus* vers le continent américain établissant la lignée du bison américain (Grange et al., 2018). L'analyse du génome nucléaire de *B. bonasus* a permis de rejeter l'hypothèse précédente selon laquelle il s'agit d'un hybride *B. priscus* et *Bos primigenius* (Soubrier et al., 2016).

Selon les études actuelles sur le comportement des bisons d'Europe et d'Amérique, Grange et al. proposent que des contradictions puissent être conciliées en considérant que les bisons femelles favorisaient la spécialisation des populations de bisons dans différentes niches écologiques alors que les bisons mâles permettaient des échanges génétiques homogénéisant les génomes des différentes populations.Finalement, les analyses morphométriques prenant en compte le dimorphisme sexuel n'ont pas validé une affirmation antérieure (Palacio et al., 2017) selon laquelle *Bison schoetensacki* était présent en France à la fin du Pléistocène.
Pour conclure, afin de comprendre l'histoire de ce clade et pour répondre aux questions relatives aux ancêtres des *Bisons bonasus*, il faut combiner les résultats morphométriques et moléculaires réalisés sur les animaux modernes.

La paléogénétique est une approche essentielle pour étudier le processus d'évolution du clade de *Bison bonasus* compte tenu du goulot d'étranglement qu'a subi l'espèce récemment qui ne reflète pas, par conséquent, la diversité génétique ancienne. L'un des objectifs de ma thèse est de continuer l'exploration de l'évolution des bisons à travers les séquences d'ADN anciennes de plusieurs échantillons eurasiatiques qui couvrent une large échelle spatio-temporelle.

Aspects méthodologiques de l’analyse de l’ADN ancien

Chapitre I: Traitement pré-séquençage : Extraction et purification de l’ADN

I) Extraction d’ADN ancien :

L’ADN ancien peut être obtenu à partir d’os, de dents, de tissus momifiés, de poils et cheveux, de cornes, ongles et sabots et de sédiment. Chez les vertébrés, les dents et les os sont les tissus les plus durs. La nature compacte de ces tissus leur permet, une fois enfouis dans le sédiment, de résister parfois pendant des centaines de milliers d’années. Les molécules d’ADN contenus dans ces tissus peuvent être préservées pendant des milliers d’années quand les conditions environnementales empêchent les processus de décomposition. Une préservation à long terme de l’ADN des restes animaux a été observée, par exemple, pour des restes fossiles des régions de pergélisol comme les restes de mammouths datés d’environ 1,2 millions d’année qui ont été retrouvé en Sibérie (van der Valk et al., 2021). Dans les régions à climat modéré à chaud, les restes fossiles sont moins bien préservés et plus la température moyenne est élevée, moins bonne est la préservation (e.g., (Pruvost et al., 2008; Smith et al., 2001). Dans les régions chaudes, la préservation de l’ADN endogène est mauvaise et peut même être absente. Les régions tempérées et chaudes jouent un rôle important dans l’évolution de beaucoup d’espèces. Il est donc important de récupérer de l’ADN à partir d’échantillons de régions chaudes. Par contre, ceci nécessite des optimisations méthodologiques spécifiques pour en tirer le maximum d’ADN préservé.

La préservation de l’ADN dépend de plusieurs facteurs taphonomiques qui interviennent depuis la mort jusqu’à la fossilisation de l’organisme (cité dans (Geigl, 2002), (Pruvost et al., 2007)). Il a été montré que les spécimens de pergélisol sont une exception car ils contiennent souvent des quantités relativement élevées d'ADN qui sont moins dégradées (Poinar et al., 2006). En effet, la basse température peut augmenter considérablement la stabilité de l’ADN (Smith et al., 2001).
De plus, une modélisation du taux de dégradation post-fouille de l'ADN dans les fossiles préservés dans les régions climatiques tempérées, comme celles du nord de la France, a montré qu’elle pourrait être environ 70 fois plus rapide que pendant l’enfouissement dans les sédiments (Pruvost et al., 2007). La dégradation est probablement due à la dépurination de l'ADN (Lindahl & Nyberg, 1972). Cette dégradation post-fouille de 90 % des molécules d'ADN endogènes peut être causée par une température moyenne de stockage supérieure à la température de préservation, combinée à une diminution du pH ou à un effet dessalant dû au lavage de l'os (Pruvost et al., 2007). La préservation d’un tissu comme l’os ou les dents dépend des caractéristiques du tissu lui-même, des conditions environnementales et des conditions de décomposition post-mortem.

L'hydroxyapatite de formule (Ca10(PO4)6(OH)2), est la principale composante minérale de l'émail dentaire, la dentine et l'os. Les molécules d’ADN ont une affinité à l’hydroxyapatite (Hagelberg et al., 1989; Romanowski et al., 1991), discuté et confirmé dans (Gotherstrom et al., 2002). Cette adsorption est probablement basée sur des groupes phosphates chargés négativement dans la molécule d'ADN et des sites d'adsorption hydroxyle sur l'hydroxyapatite (Kawasaki et al., 1985).

Le collagène quant à lui, joue un rôle protecteur des molécules d'ADN fixées à l'hydroxyapatite (Lee & Glimcher, 1989). En effet, le collagène est une protéine secrétée par les cellules des tissus conjonctifs. Contrairement à l'élastine présente aussi dans les tissus conjonctifs, le collagène est inextensible et résiste bien à la force mécanique (Fratzl, 2008). Les molécules d'ADN ont une forte affinité à l'hydroxyapatite. Cette affinité est exploitée pour isoler les acides nucléiques (Tiselius et al., 1956). Des tests ont été effectué pour étudier le rôle de l'hydroxyapatite sur la dégradation de l'ADN. En effet, il a été montré que la liaison des molécules d'ADN à l'hydroxyapatite réduit d'un facteur de deux le taux de dépurination de l'ADN (Lindahl, 1993). Les molécules d'ADN pourraient donc échapper aux attaques chimiques et enzymatiques dans ce que Eva-Maria Geigl a appelé «niches moléculaires» (Geigl, 2002).

La fossilisation, au cours de laquelle la matière organique des tissus durs est progressivement remplacée par des minéraux comme les acides phosphoriques et carboniques, peut être plus ou moins complète en fonction de l'environnement (Paabo, 1989). Par exemple, dans les grottes, les températures sont relativement stables, ce qui a été montré comme conditions environnementales favorisant la préservation de l'ADN dans les échantillons fossiles (Smith et al., 2001; van der Valk et al., 2021). Par contre, les régions chaudes ne permettent pas la bonne préservation de l'ADN a, alors que les climats tempérés, continentaux modérés ou méditerranéens s'avèrent souvent moins propices à la préservation de l'ADN dans les échantillons anciens que l'on pourrait le penser (Pruvost et al., 2008). En conclusion, la préservation de l'ADN dépend d'une multitude de facteurs et leur synergie est toujours méconnue et mal comprise.

Les premières analyses utilisant l'ADN de restes anciens se sont portées sur les tissus mous des momies égyptiennes. Une étape très importante était de réussir à étudier l'ADN à partir de tissus durs comme les os et les dents (Hagelberg et al., 1989; Horai et al., 1989). On s'est alors rendu compte que dans les mêmes conditions de conservation, l'ADN serait mieux préservé dans les os que dans les tissus mous, probablement grâce à l'adsorption des acides nucléiques sur l'hydroxyapatite (Cooper, 1994).

L'environnement dans lequel se trouve l'os définit la vitesse avec laquelle la fossilisation se réalise et conditionne sa colonisation par les microorganismes et par les contaminants organiques et inorganiques (D. E. G. Briggs, 2003). Ces derniers sont en générale des acides humiques et fulviques et encore des sels minéraux contenus dans les sols (Pruvost et al., 2007). Ces composants peuvent être extraits avec l'ADN endogène des fossiles ce qui pose problème car ce sont des inhibiteurs d'enzymes (Hagelberg, 1991; Hummel et al., 1992).

La partie de l’ADN extrait qui ne correspond pas à l’ADN de l’individu provient en grande partie des microorganismes du sol ayant colonisé l’os ainsi que d’autres microorganismes encore inconnus à l’heure actuelle. Cette partie de l’ADN contenu dans les extraits fossiles est appelée « ADN environnemental ». Ces ADN environnementaux posent un problème de rendement et de coût de séquençage parce qu’au lieu de séquencer seulement l’ADN d’intérêt, nous séquences aussi l’ADN contaminant ce qui réduit la couverture des génomes d’intérêt analysés et nous oblige de séquencer en profondeur ce qui devient plus cher. C’est pour cela que les méthodes d’extraction d’ADNa doivent être optimisées pour permettre de récupérer le maximum possible d’ADN endogène de l’échantillon fossile et minimiser la récupération de l’ADN environnemental. Lors de l’extraction d’ADN, la matrice solide des tissus doit être désintégrée pour libérer les molécules d’ADN dans une solution aqueuse qui sera par la suite purifiée. La dégradation de la matrice se fait par un chélateur, EDTA (acide éthylènediaminetétraacétique) et la protéinase K.

1) Partie pétreuse de l’os temporal :

Récemment, des analyses ont montré que l’ADN est mieux préservé dans les os pétreux que dans les os longs ou les dents. En effet, la partie pétreuse est un os disposé sous l'écaille temporal. Il provient de la capsule otique et contient l'oreille interne et il peut produire jusqu’à 100 fois plus d’ADN endogène que d’autres éléments squelettiques (Pinhasi et al., 2015).

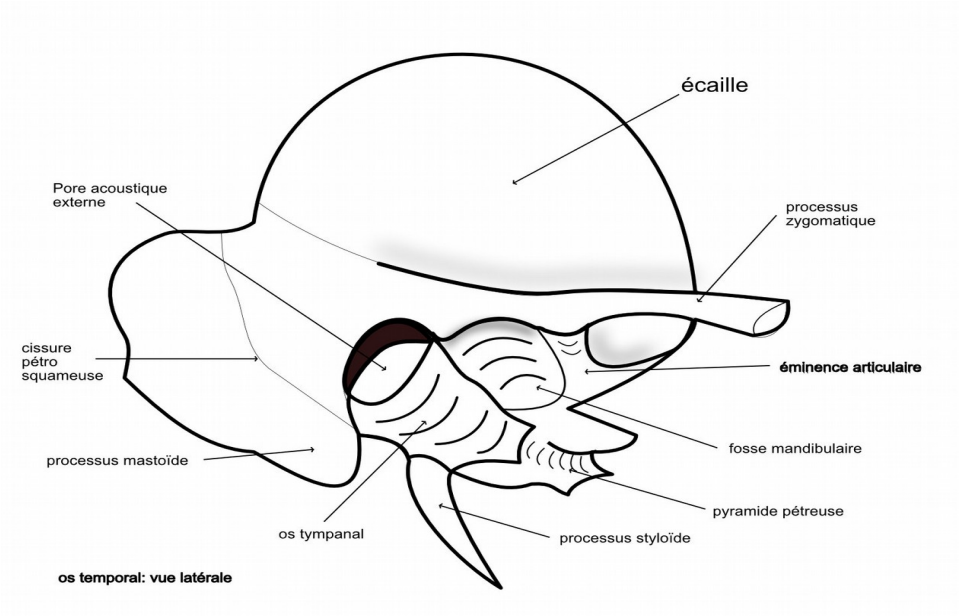


Figure 20 : Description des différentes parties de l’os temporal.

Il est généralement admis que plus l'os est vascularisé et poreux, plus il est facile pour les micro-organismes d'y pénétrer. Le fait que l’os pétreux soit peu vascularisé est sans doute une des causes de la bonne préservation de l’ADN endogène et de la faible teneur en ADN d’origine microbienne puisque cela limite les possibilités de pénétration des bactéries.
De plus, un mécanisme discuté pour rendre compte de la préservation à long terme de l’ADN dans les os est la quasi-momification des ostéocytes ayant échappé à la phagocytose par les ostéoclastes (Sørensen & Bretlau, 1997).

Le fait que l'os pétreux n'est plus remodelé après la petite enfance (Sørensen & Bretlau, 1997) pourrait préserver les noyaux des ostéocytes en état de condensation et permettrait à cet ADN d'échapper à la digestion microbienne. En 2018, une analyse comparative entre les différents compartiments de l'os temporal et les os long vis-à-vis du contenu en ADN endogène a montré que la capsule otique est la partie dans laquelle l'ADN est le mieux préservé (Geigl & Grange, 2018) (Figure 21). En effet, dans certains os pétreux analysés récemment par l'équipe Epigénome & Paléogénome, la partie de l'ADN endogène s'est élevée à >90%.

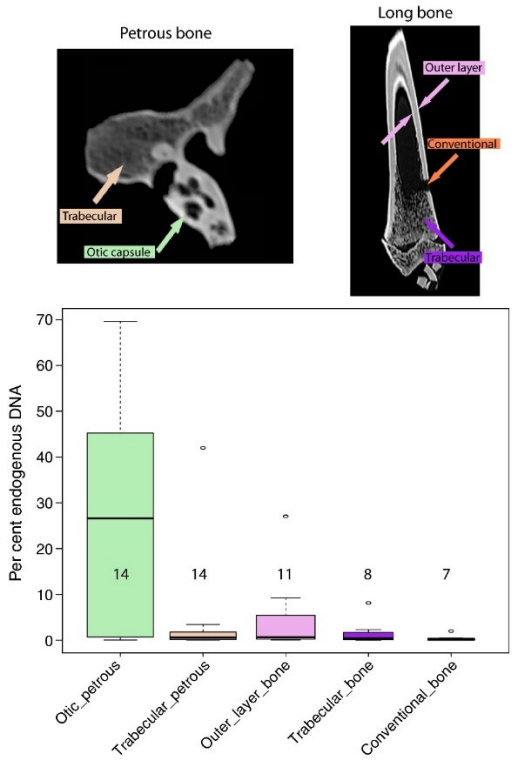


Figure 21: Pourcentage d'ADN endogène en fonction de la nature de l'os, d'après (Geigl & Grange, 2018).

Dans la présente étude, la partie la plus dense des os pétreux autour de la cochlée a été prélevée et a servi pour l'extraction de l'ADN.

2) Os longs et dents :

L'os est formé à partir (1) d'une matrice organique (25%) appelée ostéoïde, composée essentiellement de collagène de type I (90%), mais aussi d'autres protéines comme l'ostéonectine ou l'ostéocalcine, une protéine synthétisée par les ostéoblastes qui sont les cellules responsables de la mise en place et du renouvellement de l'os ; (2) d'une matrice minérale (70%) très riche en phosphate de calcium sous forme d'hydroxyapatite de calcium $Ca_{10}(PO_4)_6(OH)_2$) et (3) à partir de cellules osseuses spécifiques (5%) (Ostéoblastes et ostéoclastes) et de l'eau. Le processus d'ossification des os longs chez l'humain débute au cours du 3è mois de la vie embryonnaire et s'achève pratiquement vers l'âge de 25 ans (Beniash, 2011; Boskey, 2007; Boskey & Mendelsohn, 2005), cité dans (Balasse et al., 2015).

Les dents, quant à elles, ont des points en commun avec les os. Elles sont aussi des biocomposites dont la partie minérale est un phosphate de calcium, alors que leurs fonctions, structures et compositions diffèrent de l'os. Les dents sont composées de deux tissus principaux : l'émail et la dentine. L'émail est la partie extérieure de la dent, alors que la dentine constitue la plus grande partie de la dent et elle est séparée de l'os alvéolaire au niveau de la racine par un autre tissu minéralisé : le cément (Davit-Béal et al., 2007; Fincham et al., 1999). La structure des dents est très différente chez les Artiodactyles par rapport aux Hominidés (Tanaka et al., 2008). Chez les Artiodactyles, l'ADN n'est pas mieux préservé dans les dents que dans les os longs (Pruvost et al., 2008) alors que l'ADN dans la partie du cément chez les carnivores et les homininés peut être très bien préservé (e.g., (Svensson et al., 2021). C'est pour cette raison que j'ai analysé plus d'os que de dents des *Bovina*.

3) Passage de l'os à l'extrait d'ADN :

Les échantillons anciens analysés pendant ma thèse et mon stage de M2 ont des âges et des origines géographiques différentes et représentent le genre *Bos* et *Bison* de la famille des Bovidés. La majorité d'entre eux correspond à des os temporaux et plus précisément leurs parties pétreuses.

Toutes les préparations pour l'obtention d'extraits fossiles ont été réalisées dans un laboratoire de haut confinement dédié à l'analyse de l'ADNa (https://www.ijm.fr/440/pole-paleogenomique-et-taphonomie-moleculaire.htm), au 6ème étage de l'Institut Jacques Monod. A l'intérieur de ce laboratoire, les différentes étapes expérimentales sont effectuées dans des pièces séparées et dans des enceintes dédiées. En revanche, les étapes de purification des banques d'ADN et de la capture par hybridation ont été faites dans le laboratoire de génétique moléculaire des échantillons actuels au 5éme étage. Cette séparation physique est nécessaire pour l'analyse de l'ADNa qui peut être facilement contaminé par l'ADN moderne présent partout dans un laboratoire non confiné, surtout l'ADN issu des étapes d'amplification antérieurs. Pour minimiser ce risque, les échantillons archéologiques sont traités avec beaucoup de précautions dans un laboratoire de haut confinement. Avant de monter au laboratoire fossile, nous changeons les vêtements, portons des vêtements et une blouse lavés à l'eau de javel ainsi que des chaussures en plastique, et portons sur des mains lavées des gants qui sont eux-mêmes nettoyés à l'eau de javel.Les sacs en plastique contenant les échantillons sont aussi javellisés et, d'une manière générale, chaque objet ne monte au laboratoire fossile qu'après être nettoyé à l'eau de Javel qui permet d'éliminer l'ADN (Champlot et al., 2010). En entrant dans le sas du laboratoire fossile, on enfile une combinaison, un masque, une charlotte et des chaussures en plastique nettoyés à l'eau de Javel.

La première étape consiste à découper les os pétreux pour récupérer la partie la plus dense et la plus compacte : la capsule otique ou les os longs dont la surface ou cortex est récupérée. D'abord, la partie spongieuse a été enlevée et la surface des morceaux d'os ont été nettoyés avec un coton-tige imbibé d'eau, puis un autre imbibé d'hypochlorite, et finalement un troisième imbibé d'eau.
Cette étape consiste à enlever les sédiments et un maximum de molécules d'ADN exogène ainsi que des molécules, comme les substances humiques, qui inhibent les réactions enzymatiques. Puis, on laisse sécher le fragment d'os pendant une nuit avant de les transférer dans des porte-échantillons contenant un barreau métallique, destinés au cryo-broyage par un cryobroyeur (6770 Freezer/Mill®Spex SamplePrep) dans un champs magnétique.

Le broyage dans l'azote liquide permet d'éviter l'augmentation de température qu'occasionne le broyage, source de dégradation de l'ADN. Suivant le broyage, les porte-échantillons, après s'être réchauffés à température ambiante, sont ouverts avec le SPEX « vial opener » dans une enceinte en plexiglass. Les tubes sont ouverts doucement et le contenu est transféré dans un tube Eppendorf de 1.7 ml. Après chaque expérience, le matériel du cryobroyage utilisé est décontaminé à l'eau de javel, ou, pour les parties métalliques, RNaseAway puis rincé à l'eau, séché et exposé aux UV (Champlot et al., 2010).

L'extraction de l'ADNa se fait dans la deuxième pièce du laboratoire confiné. Avant d'entrer dans cette pièce une deuxième combinaison est mise.En général, 100 mg de poudre d'os sont lavés 3 fois pendant 15 minutes avec un tampon phosphate (0.5M, pH8) pour enlever les molécules d'acides nucléiques qui sont en interaction faible avec l'hydroxyapatite, à priori surtout l'ADN environnemental, ainsi que d'autres composants chimiques environnementaux qui proviennent du sédiment et ne sont pas associé à la matrice de l'os, ainsi que d'autres composants chimiques environnementaux qui proviennent du sédiment et ne sont pas associés à la matrice de l'os. Le surnageant est enlevé après chaque lavage avec le tampon phosphate. Finalement, la poudre est rincée avec 1 ml du tampon TT (Tris 10mM, Tween20 0.1%). Le tampon utilisé pour l'extraction est constitué de 0.5 M EDTA pH8, 0.14 M β-mercaptoéthanol, pH 8.0, 250 µg/ml protéinase K (PK) et de Tween20 10 %. Les poudres sont incubées dans le tampon d'extraction pendant 30h à 37°C. Par la suite, le surnageant est, récupéré. 1 ml du même tampon est ajouté aux poudres non digérées et incubées pendant 48h à 37°C.

Dans le cas des échantillons de peau, nous les avons réduit en poudre et l'ADN a été extrait dans un tampon d'extraction contenant une concentration plus élevée de PK par rapport à celle utilisée pour les ossements. En effet, nous avons utilisé 1 mg de PK par ml de tampon d'extraction car la peau contient plus de protéines que l'os et la quantité de PK utilisée pour l'os peut ne pas être suffisante pour la digestion des protéines de la peau. Ce tampon contient en plus de la PK, de l'EDTA et de beta-mercaptoéthanol, puis un détergent anionique, le NLS 10 % (sodium N-Lauroylsarcosyl), pour permettre la lyse cellulaire. Les poudres de peaux ont été incubées dans le tampon d'extraction pendant 48h à 37°C. Le surnageant a été récupéré et la poudre non digérée était remise en suspension dans un autre ml du même tampon d'extraction.

II) Purification de l'ADN :

Après l'incubation de la poudre d'os dans le tampon d'extraction, l'ADN extrait doit être séparé du reste des composants afin d'obtenir un extrait d'ADN purifié qui servira à la construction de banques génomiques. L'étape de purification se fait sur une colonne de silice, une méthode qui a été adoptée depuis 2007, quand Rohland & Hofreiter (Rohland & Hofreiter, 2007) l'avaient optimisée. Par contre, de nombreuses modifications ont été apportées depuis, surtout par (Dabney et al., 2013; Glocke & Meyer, 2017). A notre tour, nous avons testé, modifié et optimisé ce protocole. Deux tampons de purification d'ADN ont été utilisés pour récupérer le plus d'ADN endogène, en éliminant le maximum d'ADN contaminant et environnemental qui peuvent coexister avec l'ADNa. La logique de cette démarche est que les molécules d'ADN ciblées par les différents tampons ne sont pas identiques mais le contenu en ADNa dans un extrait ne peut être anticipé à priori. Cette purification s'effectue en trois étapes : Adsorption sur une matrice de silice, lavage et élution.

L’étape d’adsorption consiste à déposer sur une phase solide une matrice de silice, le surnageant d’extraction mélangé avec un tampon de liaison chaotropique qui va augmenter l’affinité entre les molécules d’ADN et la silice alors que les sels et les autres composants chimiques présents dans le surnageant passent à travers la matrice. Le tampon de liaison contient des sels de guanidine, en particulier guanidine-hydrochloride. Dans un second temps, l’ADN fixé sur la colonne de silice est nettoyé avec le tampon PE (10 mM Tris-HCl pH 7.5, 80% éthanol). Finalement, l’ADN est élué dans un tampon d’élution TET (10 mM Tris HCl pH 8.0, 1 mM EDTA, 0.1% Tween-20), qui a été irradié (254 nm pendant 20 min) pour éviter la contamination. Les sels de guanidine présents dans le tampon de liaison permettent à l’ADN, en présence d’isopropanol, de s’accrocher à la silice. Le tampon de rinçage PE contient de l’éthanol 80% qui permet à l’ADN de rester accroché sur la colonne de silice lors du lavage en absence d’agent chaotropique. Les colonnes de silice nécessitent un système d’aspiration (une pompe produisant un vide) ou une centrifugation pour faire passer les différents tampons à travers les membranes. Nous utilisons les colonnes de purification du kit QiAquick (Qiagen) et un système d’aspiration sous vide (vacuum Manifold, Qiagen) (Gorgé et al., 2016).

Les extraits ont été purifiés avec deux tampons de purification qui diffèrent par leur concentration en guanidine hydrochloride. Le tampon de purification appelé 2M70 est composé de 2M guanidine hydrochloride et de 70% isopropanol (Glocke & Meyer, 2017), alors que le tampon de purification 5M40 est composé de 5M guanidine hydrochloride et de 40% isopropanol auquel on ajoute l’acétate 3M (Dabney et al., 2013). Le system manifold de Qiagen par aspiration du vide est installé sous la hotte et les colonnes de silice insérées. Pour chaque extrait, on a transféré 1 ml dans des tubes Falcon de 15 ml auquel on a ajouté 10 ml du tampon de liaison et on les a transvasés directement dans les réservoirs(«extenders») au-dessus des petites colonnes de silice. Le liquide s’écoule plus ou moins doucement la vitesse étant variable selon les différents échantillons. Une fois tout le liquide est passé, les colonnes de silice sont rincées deux fois avec le tampon de rinçage (Qiagen PE) en ajoutant 1 ml à chaque fois. L’ADN est élué deux fois dans 30 microlitres de tampon TET (10 mM Tris HCl pH 8.0, 1 mM EDTA, 0.1% Tween-20).

Chapitre II: Traitement pré-séquençage: Stratégies de construction de banques d’ADN génomiques

I) Différentes stratégies de construction de banques d’ADN ancien :

Le séquençage de nouvelle génération a permis d’élargir le nombre d’études basées sur les données génomiques. Les appareils de séquençage, utilisés aujourd’hui, permettent l’obtention de milliards de fragments d’ADN. Ces quantités de données génomiques obtenues permettent la reconstruction de génomes complets en un temps court. Ces appareils utilisent le même principe de préparation des échantillons à séquencer. En effet, pour permettre le séquençage de nouvelle génération, il faut que les fragments d’ADN contenus dans les extraits fossiles soient convertis en banques génomiques. Cette conversion consiste à ajouter au niveau des extrémités des séquences adaptatrices spécifique à chaque plateforme de séquençage. Ces séquences adaptatrices permettent l’amplification et le séquençage de l’ensemble des fragments d’ADN contenus dans l’extrait fossile. La plupart des approches de préparation de banques disponibles dans le commerce pour le séquençage Illumina fonctionnent mal avec de l'ADN endommagé et dégradé (Stiller et al., 2016).

Les défis liés à l’étude de l’ADNa ont conduit au développement de plusieurs approches spécifiques pour la préparation de banques génomiques. L’approche de construction de banques est bien adaptée aux études de l’ADNa car la préparation d’une banque de séquençage nécessite une étape de fragmentation préalable des fragments d’ADN, qui est déjà une caractéristique de l’ADNa. Les adaptateurs liés aux extrémités des fragments d’ADNa permettent l’amplification de molécules courtes ce qui représente un avantage car la majorité des fragments d’ADNa sont de petite taille. Ces banques sont aussi utilisées pour l’enrichissement en ADN cible. Les adaptateurs des plateformes de séquençage contiennent des séquences appelés «index» qui permettent de distinguer les fragments d’ADN provenant de banques différentes. L’avantage de l’utilisation de deux index différents pour chaque banque permet d’éviter le problème de contamination croisée et de séquencer plusieurs banques à la fois (A. W. Briggs & Heyn, 2012).

Les caractéristiques particulières de l’ADNa nécessite l’utilisation d’un protocole adapté à l’ADNa car l’utilisation d’un protocole standard entraîne la perte préférentielle des fragments d’ADN correspondant à l’information génétique ancienne, précieuse et importante. En effet, les molécules d’ADNa présentent la plupart du temps des bases manquantes sur l’un ou l’autre des brins, conséquence de l’hydrolyse d’une liaison N-glycosidique qui génère un site abasique. Ce dernier sera par la suite éliminé par hydrolyse du désoxyribose entraînant ainsi la rupture de la molécule d’ADN double brin ce qui provoque sa fragmentation en des molécules de taille plus petite, double brin avec des extrémités simple brin. De plus, les fragments d’ADNa présentent des résidus uraciles dus à la désamination des cytosines qui s’accumulent principalement sur les extrémités simple brins des fragments d’ADN.

Ces uraciles vont entraîner des erreurs lors de l'amplification par PCR de type transition de cytosine vers thymine, ou guanine vers adénine sur le brin complémentaire. Ces caractéristiques particulières de l'ADNa font que l'optimisation de protocole de construction de banque soit nécessaire.

Trois stratégies peuvent être adoptées et discutées pour construire des banques d'ADNa : Les banques d'ADN double brin, les banques d'ADN simple brins et les banques en utilisant des adaptateurs Y (Figure 22) (Bennett et al., 2014).

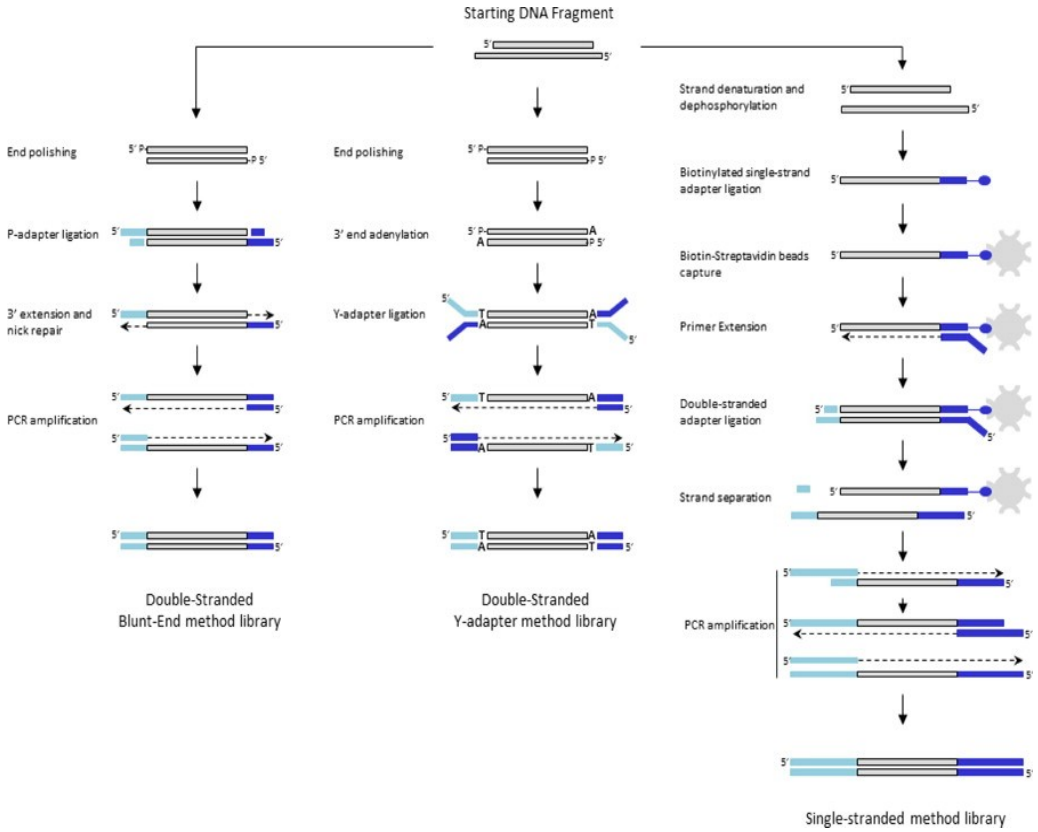


Figure 22: Différentes stratégies de construction de banques d'ADNa d'après (Bennett et al., 2014).

1) Stratégies de construction de banques double brin :

L'approche de construction de banque double brin est basée sur la ligation de deux adaptateurs distincts aux niveaux des extrémités franches d'un fragment d'ADN double brin. Cette méthode a été utilisée par la compagnie 454 Roche Life Sciences pour construire les premières banques d'ADN génomiques en 2005. Les adaptateurs sont des molécules partiellement double brin avec une une des deux extrémités 5' sous forme simple brin . Pour éviter la ligation des adaptateurs entre eux, leurs extrémités 5' et 3' ont un groupe hydroxyle et sont dépourvus de groupement phosphate. Chaque fragment d'ADNa contient deux adaptateurs différents liés de part et d'autres. Cependant, comme la ligation des adaptateurs se fait de façon aléatoire, environ 50% des molécules d'ADN auront le même adaptateur et seront donc perdus car il faut deux adaptateurs différents pour que la molécules soient lues par le séquenceur. Le principe de la stratégie double brin consiste à réparer les molécules d'ADNa pour obtenir des extrémités franches. Cette étape est suivie par une purification pour éliminer les enzymes de réparation.

Par la suite, les adaptateurs partiellement double brin sont ajoutés à des quantités équimolaires. La T4 ADN ligase assure la formation de liaison phosphodiester entre les groupements 5'P des fragments d'ADN et les groupement hydroxyles 3'OH libres des adaptateurs. Les brins courts d'adaptateurs qui ne sont pas liés de façon covalente au fragment d'ADN seront éliminés par dénaturation pour que l'ADN polymérase Bst puisse polymériser à partir de l'extrémité 3' OH du fragment d'ADN à séquencer. A l'issu de cette étape, nous obtenons un fragment d'ADNa prêt à être amplifié puis séquencé contenant de part et d'autre un adaptateur distinct. Cette stratégie est la plus utilisée pour la construction de banques d'ADNa (par exemple (Grange et al., 2018; Massilani et al., 2016; Orlando et al., 2013; Park et al., 2015).

Cette méthode a été optimisée par mon équipe. Cette optimisation a concerné l'ADN polymérase Bst qui intervient après la ligation des adaptateurs.En effet, dans le protocole standard de la construction de banque double brin, il faut se débarrasser du fragment court de l'adaptateur qui n'est pas lié de façon covalente à la molécule d'ADN. Ceci se fait avec une enzyme qui déplace le brin,le grand fragment de l'ADN polymérase Bst qui va ensuite synthétiser un nouveau brin d'adaptateur à partir de l'extrémité 3' du fragment ligaturé, pendant 20 minutes à 37°C (Meyer & Kircher, 2010).

Pour simplifier la procédure et minimiser les étapes d'ouverture de plaques ou de tubes,mon équipe a testé une autre polymérase qui a, en plus de l'activité polymérase 5'→3', une activité 5'→3' exonucléase lui permettant de faire des translations de césure ou «Nick-Translation» (Figure 23). Avec cette polymérase, le fragment court d'adaptateur est hydrolysé au fur et à mesure de la synthèse du nouveau brin d'adaptateur par l'activité exonucléase de la polymérase choisie.

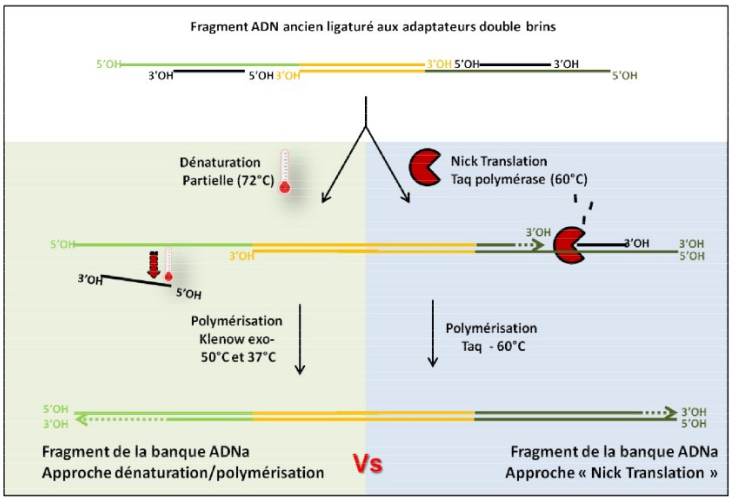


Figure 23: Comparaison du protocole standart (Dénaturation/Polymérisation) et optimisé (translation de cessure) de construction de banque double brin pour l'étape post-ligation des adaptateurs (Thèse de Massilani Diyendo).

Le mix enzymatique utilisé était le mix One Taq de NEB, qui contient 2 ADN polymérases (la Taq et la DeepVent) et des aptamères qui bloquent leur site actif à une température inférieure à 45°C. La comparaison des deux stratégies a permis de conclure qu'il n'y pas de différence entre les deux protocoles. Par contre, avec le mix One Taq, il n'était pas nécessaire de passer par une étape de purification, après l'étape de ligation des adaptateurs. De plus, puisque le mix OneTaq étant un mix d'amplification par PCR, il suffisait d'ajouter les amorces d'amplification de banque Illumina pour enchaîner directement par les cycles d'amplification par PCR, ce qui a rendu cette étape économique et rapide.

2) Stratégie de construction de banques avec les adaptateurs Y :

La stratégie de construction de banque double brin en utilisant des adaptateurs en forme Y a été développée par la plateforme de séquençage Illumina. Les adaptateurs Y sont utilisés dans le protocole de construction de banques double brin. Ces adaptateurs en forme de Y présentent une région complémentaire de quelques bases formant une extrémité double brin et une autre région non complémentaire formant alors deux régions simple brins.

L'avantage de cette méthode de construction de banque génomique est que la ligation des adaptateurs sur le fragment d'ADN, contenu dans l'extrait fossile, est orienté ce qui diminue la perte des fragments à séquencer qu'on trouve en utilisant le protocole de construction de banque double brin avec des adaptateurs double brin.L'étape de ligation d'adaptateurs permet d'obtenir des fragments d'ADNa contenant deux adaptateurs de chaque coté qui sont distincts au niveau des séquences non complémentaires. Les adaptateurs sont phosphorylés à l'extrémité 5' pour permettre la ligation. Pour éviter la ligation d'adaptateurs entre eux, leurs extrémités simple brin possèdent un dT flottant en 3'. Cela permet de diriger la ligation de l'adaptateur sur les fragments d'ADN à séquencer qui ont à leurs extrémités 3' un dA flottant ajouté au préalable grâce à l'action de l'ADN polymérase de type Klenow exo-, qui est dépourvue d'activité exonucléase 3'→5'. Cette polymérase permet d'ajouter un dA à l'extrémité 3' de l'ADN.

Le principe de cette méthode consiste d'abord à réparer les extrémités des fragments d'ADN contenus dans les extraits fossiles pour obtenir des fragments à bouts francs. Une étape d'adénylation de l'extrémité 3' est effectuée. Les fragments réparés sont par la suite ligaturés avec les adaptateurs. La ligation est facilité par l'hybridation du dA en 3' du fragment et du dT en 3' de l'adaptateur. Cependant, des dimères d'adaptateurs peuvent être formés à cette étape ce qui réduit l'efficacité de séquençage de banques d'ADN par la suite. La proportion de dimères obtenue est fonction de la quantité relative de fragments d'ADN dans l'extrait et d'adaptateurs ajoutés. Trop peu d'adaptateurs produit une proportion importante de fragments n'ayant pas d'adaptateurs aux deux extrémités. Trop d'adaptateurs produit beaucoup de dimères. Les quantités d'ADN dans les extraits anciens étant très variables et difficilement estimables avec la fluorescence du fait de la présence de molécules environnementales faussant les mesures, cette approche est délicate à mettre en œuvre quand on souhaite traiter beaucoup d'échantillons à la fois.Les fragments liés aux adaptateurs sont par la suite amplifié par PCR. Les indexes des adaptateurs Y sont par la suite ajoutés aux banques.

3) Stratégie de construction de banques simple brin :

La stratégie de construction de banque simple brin a été développée en 2013 par Gansauge & Meyer pour permettre l'incorporation plus efficace de fragments d'ADNa. Le principe de cette stratégie consiste à convertir en banques tous les fragments d'ADN simple brin sans passer alors par l'étape de réparation des extrémités de molécules d'ADN, qui est une étape primordiale et indispensable pour les deux stratégies discutées en dessus et qui entraîne la perte de bases nucléotidiques. Les extraits passent par une étape de dénaturation et de déphosphorylation par une phosphatase limitant ainsi le risque de ligation de deux fragments d'ADNa ou la formation des hétéroduplex. Les fragments dénaturés seront incubés avec une ligase et un adaptateur simple brin dont son extrémité 5' est phosphorylé et son extrémité 3' est reliée à une molécule de biotine. Les molécules ligaturées aux adaptateurs sont fixées sur des billes de streptavidine. Une amorce complémentaire à la séquence d'adaptateur s'hybride à celui ci pour amorcer l'élongation en présence d'une ADN polymérase. La polymérase utilisée dans cette approche (Bst 2.0) est dépourvue d'activité exonucléase 3'→5'. Elle ajoute indépendamment du brin matrice un dA à l'extrémité 3' du nouveau brin synthétisé. Après un traitement à la T4 ADN polymérase dont l'activité 3'→5' exonucléolytique simple brin va enlever les dA flottants et générer une extrémité double brin franche, un adaptateur partiellement double brin va être ligaturé à cette extrémité de manière covalente sur le brin antisens.

Bien que cette stratégie soit laborieuse, elle présente plusieurs avantages. En effet, les différentes réactions sont réalisées sur des molécules d'ADN fixées sur un support solide grâce à l'utilisation de la biotine minimisant ainsi la perte de molécules durant les étapes de purification. Cette approche permet de convertir en banques tous les les fragments endommagés présentant une cassure sur un des deux brins, contrairement à la stratégie double brin,parce qu'au moment de la dénaturation ces fragments se séparent en fragments simple brins et seront incorporés individuellement dans la banque. En effet, il a été observé que l'efficacité de la méthode simple brin est supérieure à celle utilisant le double brin et permet de récupérer plus de fragments qu'avec la méthode double brin (Gansauge & Meyer, 2013; Meyer et al., 2012). De plus, il a été montré qu'avec la méthode simple brin, les fragments sont de plus petite taille qu'avec la méthode double brin (Bennett et al., 2014). Contrairement aux banques simple brin, les banques double brin favorisent l'incorporation de fragments de taille plus importante mais sont moins efficaces pour la récupération des fragments les plus dégradés, donc pour l'analyse des échantillons moins bien conservés.

Le choix de la stratégie de la construction de banque reste un critère décisif des chercheurs tout en tenant compte de la caractérisation des échantillons fossiles (site d'excavation, âge, contexte archéologique...). Mais aussi, il est possible de combiner les deux méthodes double et simple brin pour augmenter la complexité de l'information génétique contenues dans l'extrait fossile.
Dernièrement, une nouvelle méthode de construction de banques génomiques simple brin dédiée à l'ADNa a été publiée par (Kapp et al., 2021). Bien que cette méthode permette de convertir les fragments d'ADNa en banques génomiques dans un temps plus court que les autres méthodes et qu'elle ne nécessite pas une étape de purification diminuant ainsi la perte des petits fragments, elle reste peu efficace à cause des dimères d'adaptateurs qu'elle produit en abondance à cause de la stratégie choisie pour simplifier la procédure.

II) Construction des banques d'ADN ancien :

Nous avons adopté la stratégie double brin de construction de banques d'ADN pour tous nos extraits fossiles (Figure 24). Les banques d'ADN sont construites à partir d'une population de molécules d'ADN uniques présentes dans les extraits auxquelles sont ajoutés des adaptateurs contenant des code-barres permettant leurs amplifications et l'obtention d'une population de molécules filles chacune identique à la molécule mère initiale. Cette étape est importante pour obtenir un nombre élevé de molécules d'ADNa à partir de celles qui sont présentes en faible quantité dans les extraits fossiles.

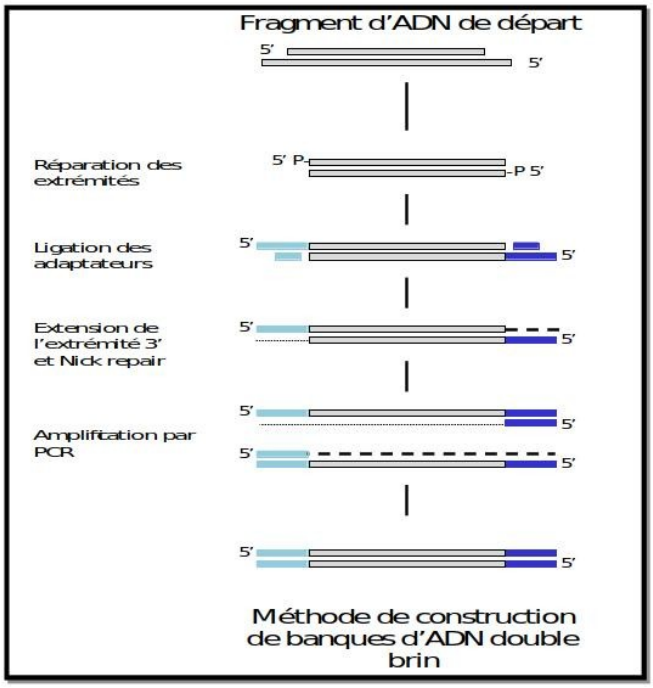


Figure 24: Schéma représentatif de la stratégie double brin de la plateforme Illumina (Bennett et al., 2014).

Aucun des deux brins de l'adaptateur n'est phosphorylé en 5' empêchant ainsi une ligation entre deux adaptateurs. L'adaptateur est composé de 3 régions: une région qui sert à la fixation de l'adaptateur sur la lame en verre et à la bridge PCR permettant la production des clusters lors du séquençage Illumina, une deuxième région qui contient une séquence de 12 nucléotides reconnue par l'amorce (P5 ou P7) permettant ainsi l'initiation de l'élongation de séquençage, et un Index qui diffère d'un adaptateur à un autre ce qui permettra de distinguer les différents échantillons fossiles.
Par exemple, pour un adaptateur de la série 700, le 701, a la séquence suivante :

P701 5' OH
CAAGCAGAAGACGGCATACGAGATTCGCCTTAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC 3'OH

Un adaptateur de la série P700 est constitué de 66 nucléotides alors qu'un adaptateur de la série P500 comprend 70 nucléotides. Nous utilisons des indexes sur les deux adaptateurs avec une séquence unique par échantillon aussi bien pour l'index i7 que i5 (Indexes duals uniques), ce qui

permet de minimiser, lors du démultiplexage, les risques de contamination croisée entre échantillons différents séquencés simultanément.

1) Traitement des extraits d'ADN ancien avec l'enzyme USER :

Un traitement avec l'enzyme USER (New England Biolabs) a été utilisé pour enlever les cytosines désaminées transformées en Uraciles. En effet, l'enzyme USER est un mélange d'Uracil DNA glycosylase (UDG) et d'une ADN glycosylase-lyase l'endonucléase VIII. L'UDG catalyse l'excision d'uraciles présentes dans l'ADN, ce qui forme un site abasique (apyrimidique) tout en laissant le squelette phosphodiester intact. L'activité lyase de l'endonucléase VIII casse le sucre de part et d'autre du site abasique, de sorte que la liaison phosphodiester est rompue (Figure 25). 3 µl du tampon (1.5µl du CutSmart Buffer, 1µl de H2O et 0.5µl USER) ont été mélangés et incubés avec 12 microlitres d'extraits. Pour les échantillons non traité avec l'enzyme USER, nous avons ajusté le volume avec de l'eau de façon à avoir un volume final de 15µL. Les extraits fossiles ont été incubés avec l'enzyme USER pendant 30 minutes à 37°C, le temps et la température nécessaires pour l'activité de l'enzyme USER. (voir liste des échantillons traités ou non par USER).

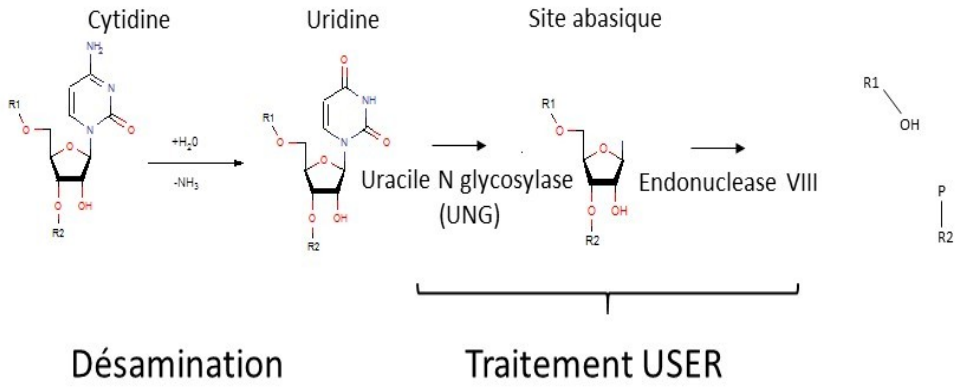


Figure 25: Effet du traitement USER sur les fragments d'ADN.

2) Réparation des extrémités d'ADN :

L'ADNa, dégradé, est susceptible d'avoir des bouts cohésifs et déphosphorylés. Pendant cette étape ses extrémités sont réparées de façon à avoir des bouts francs et phosphorylés en 5' permettant ainsi la ligation des adaptateurs. La réparation des extrémités d'ADN se fait par un mélange de deux enzymes (la T4 polynucleotide kinase et la T4 DNA polymerase) contenues dans le mélange EndRepair du Kit NEBNext EndRepair Module de NewEngland Biolabs. 15µl du mix d'enzyme End Repair (1.5µl NEBNext Ultra II End Prep Enzyme mix, 3.45µl NEBNext Ultra II End Prep Reaction Buffer, 10.05 H2O) ont été ajoutés au mélange réactionnel et incubé 30 min à 20°C puis 30 min à 60°C.

3) Purification des fragments d'ADN réparés :

La purification a été réalisée en utilisant le Kit NucleoSpin Gel and PCR clean-up de MachereyNagel pour récupérer l'ADNa réparé tout en se débarrassant de tous les composants des tampons utilisés et des enzymes.

Cette purification est nécessaire pour l'efficacité de l'étape suivante de ligation bien qu'elle puisse causer des pertes d'ADN. La purification a été réalisée en ajoutant 500 µL du tampon de liaison 2M70 pré-distribué dans une plaque 96 à puits profonds. Après 5 min d'incubation, le mélange est chargé sur une plaque 96 puits de matrice de silice installée sur un support pouvant être mis sous vide dans lequel est installée en dessous des matrices une plaque percée dédiée au lavage (wash plate Macherey Nagel). Un vide léger est appliqué à l'appareil et lorsque la solution est passée au travers de la matrice, les matrices sont rincées avec 1ml de tampon PE, toujours sous vide puis lorsque la plaque est sèche, nous l'avons centrifugé 6min à 4600rpm pour éliminer toute trace d'Ethanol. Pour éluer l'ADN, on a ajouté entre 46 et 71µL de H_2O avec un temps d'incubation de 5 min et on a centrifugé 7min à 4600rpm. Le volume ajouté a été ajusté afin de récupérer un volume d'au moins 37 µL pour l'étape suivante. Il faut noter que plus le volume d'élution est faible plus la concentration d'ADN extrait est élevée mais plus la perte est importante.

4) Ligation des adaptateurs P5/P7 d'Illumina :

La ligation des adaptateurs est une étape critique car elle permet d'identifier les échantillons anciens qui différent par deux indexes. Une contamination des adaptateurs ou une erreur lors de la manipulation est susceptible de fausser l'attribution aux bons échantillons des séquences d'ADN générées lors du séquençage. La ligation entre l'extrémité 3'OH de l'adaptateur et l'extrémité 5'du fragment d'ADN réparé est effectuée par la T4 ADN ligase. Une combinaison unique des adaptateurs P5 et P7 (20µM) a été ajoutée dans chaque puit de la plaque 96. Pour chaque échantillon on a transféré 36.5µL de l'ADN extrait réparé et purifié, et ajouté 10µL du tampon NEB Quick ligase réaction 5X (concentration finale 10mM $MgCl_2$, 1mM ATP, 1mM DTT, 66 mM Tris HCl pH8.0, 6% Polyethylene glycol) et 1.5µL de Quick T4 DNA ligase (NEB) et laissé incuber 20min à 16°C. Le petit brin d'adaptateur qui n'est pas lié de façon covalente à l'ADN va être enlevé à l'étape suivante pour permettre la synthèse du nouveau brin d'adaptateur.

5) Elongation des adaptateurs partiellement double brin et amplification par PCR :

Cette étape appelée aussi Nick Repair, fait intervenir la One Taq (NEB) qui est un mélange de deux ADN polymérase thermorésistantes associées à des aptamères bloquant pour leur conférer une activité qu'au-delà de 60°C. Les deux ADN polymérases sont la Taq et une enzyme avec activité de correction (exonucléase 3'-5') et le mélange possède en plus de l'activité 5'-3' polymérase une activité 5'-3' exonucléase qui lui permet de faire des translations de césures ou Nick-Translation en hydrolysant le petit brin non ligaturé au fur et à mesure de la synthèse du nouveau brin d'adaptateur. La réaction de ligation est mélangée avec un volume du mélange NEB One Taq 2X (concentration finale 20 mM Tris HCl, 0.05% Tween20, 1.8 mM $MgCl_2$, 22 mM KCl, 22 mM NH_4Cl, 5% Glycerol, 0.2 mM dNTPs et 25 unités/ml One Taq DNA polymerases), à laquelle nous avons ajouté les amorces d'amplification de banque Illumina P5m/P7m (10µM) pour enchaîner le end-repair avec une réaction de PCR. La réaction de « Nick translation » est réalisée à 60°C pendant 20min.

L'amplification par PCR qui est l'étape finale de construction de banques permet d'avoir une quantité importante et suffisante de fragments d'ADN pour réaliser le séquençage à haut débit. 10 cycles de PCR ont été effectués sur un Mastercycler Eppendorf selon le programme suivant : une dénaturation pendant 1min et 30s à 95°C, une deuxième dénaturation pendant 20s à 95°C, une hybridation de 35s à 60°C et une élongation pendant 70s à 68°C.6) Quantification des banques d'ADNa par qPCR avant amplification par PCR :La PCR quantitative en temps réel (Higuchi et al., 1993) permet de quantifier le nombre de molécules initialement présents dans les extrait fossiles afin d'évaluer l'authenticité des séquences obtenues (Pruvost & Geigl, 2004).

Suite à la construction de banques d'ADN génomiques, un aliquote a été prélevé pour être caractérisé par qPCR avec les amorces d'amplification universelles de la plateforme de séquençage Illumina. En effet, cette quantification permettra de quantifier la présence des fragments d'ADN qui contiennent des adaptateurs à leurs extrémités. Un robot pipeteur epMotion5070 Eppendorf est utilisé pour remplir une plaque de 384 puits. Il pipette 2µL des banques d'ADNa diluées au $50^{ème}$ et 5µL d'un mélange réactionnel 1X de QPCR contenant 0.02µL de chaque amorce P5 et P7 de concentration initiale 100µM. Le mélange réactionnel s'appelle LightCycler 480 SYBRGreen I de Roche et contient la FastStart Taq DNA Polymerase du SYBR Green I qui s'intercale entre les brins d'ADN permettant sa quantification. Le Mix 1X est mis dans un porte-tube chauffé à 37°C pour améliorer la précision du pipetage rendu imprécis par la viscosité. Dans ces conditions, la concentration réactionnelle finale du mélange est 0.7X mais cette situation assure une plus grande précision des mesures. Suite aux résultats de la qPCR, nous avons choisis le nombre de cycles d'amplification pour chaque banque. Un cycle PCR de réassociation a été fait pour dénaturer les grandes molécules aberrantes qui peuvent se former entre deux adaptateurs complémentaires et deux inserts différents lorsque la PCR atteint le plateau ce qui rend difficile la purification par la suite.

Le mélange préparé pour chaque échantillon contient 4µL du tampon Roche 10X ,1µL de la Taq polymérase HotStart, 1µL de dNTPs (2mM), 1µL du mix des amorces P5m/P7m (10µM),10µL d'extrait d'ADNa et on a ajusté le volume avec de l'eau jusqu'à 30µL afin de les ajouter à 10 µL des premières réactions de PCR. La PCR a été programmée pour une première étape de dénaturation à 95°C pendant 5min et un cycle comportant une dénaturation de 40s à 95°C, une hybridation de 2 min à 60°C et une étape d'extension de 10min à 68°C.

7) Purification des banques d'ADNa :

Cette étape est indispensable pour éliminer le maximum possible des dimères d'adaptateurs qui peuvent être présents dans les banques d'ADNa suite à la ligation. Elle se fait par l'utilisation des billes magnétiques NucleoMag® NGS Clean-up and Size Select, qui ont la particularité de permettre la sélection de taille des fragments d'ADN que l'on souhaite conserver à l'issue de la purification. La sélection de taille est dépendante de la concentration en Polyéthylène glycol (PEG) et en NaCl. En effet, plus on augmente la concentration en PEG, plus on sélectionne les fragments courts (Lis & Schleif, 1975). Les billes magnétiques fixent l'ADN quand il est aggloméré avec le PEG. La purification de l'ADN avec les billes SPRI (Solide phase reversible immobilization) se fait en plusieurs étapes: Pour un volume de 45 µL des banques d'ADN génomique, on a ajouté un volume des billes qui représente 1.3 fois le volume des banques (58.5µL).

On a mélangé 10 fois et on a incubé 5min à température ambiante pour que les fragments d'ADN s'accrochent aux billes magnétiques. La plaque 96 contenant les banques d'ADN est placée sur une plaque magnétique, et laissée environ 2-5 min jusqu'à ce que le surnageant soit devenu clair car les billes magnétiques fixant l'ADN sont attirées sur la paroi des tubes. On a pipeté et éliminé le surnageant et on a rincé les billes avec 200 µL d'éthanol 80% fraîchement préparé. On a déplacé la plaque autour des aimants pour dissocier les billes de la paroi et on a laissé 2 min jusqu'à ce que le surnageant soit devenu clair. L'Ethanol a été éliminé et les billes sont légèrement séchées. 52 µL du tampon d'élution TET (10 mM Tris HCl pH 8.0, 1 mM EDTA, 0.1% Tween-20) ont été ajoutés dans chaque puits de la plaque 96 pour resuspendre l'ADN. On a transféré par la suite 50 µL du surnageant dans une 2éme plaque pour réaliser une seconde purification dans les mêmes conditions.

Pour réaliser le séquençage de nouvelle génération Illumina, il est nécessaire de mélanger toutes les banques d'ADN génomiques ensemble. Le mélange des banques a ensuite été purifié 3 fois avec le même ratio du volume de billes SPRI tout en gardant 40 % du mélange de banques purifié 2 fois avant que le reste soit purifié 3 fois. Autrement dit, après avoir purifié le mélange de banques, nous arrivons a obtenir deux mélanges: le premier est purifié 2 fois et le deuxième 3 fois. En effet, nous avons mélangé 10 µL de chaque banque et nous avons ajouté 1.3 fois le volume obtenu de billes SPRI. Nous avons resuspendu les billes en premier temps dans 300 µL du tampon TET d'élution au quel nous avons ajouté 390 µL de billes SPRI. La resuspension suivante était réalisée dans 130 µL aux quels nous avons prélevé 50 µL qui correspondaient au mélange 2. Les 80µL qui sont restés ont été mélangés avec 104 µL de billes de SPRI et puis l'élution a été réalisée avec 50 µL correspondant au mélange appelé 3. Les banques ont donc été purifiées 4 ou 5 fois de suite avant le séquençage, ceci pour minimiser les dimères d'adaptateurs.

8) Quantification des mélanges purifiés de banques d'ADN par QPCR, dosage Qubit et Bioalyser (2100 AgilentBioanalyser) :

Une analyse BioAnalyser a été réalisée pour quantifier le mélange de banques d'ADN génomiques et vérifier l'absence ou la présence des dimères d'adaptateurs et déterminer la molarité de l'ADN pour estimer le bon facteur de dilution du mélange des banques génomiques pour le séquençage. Le BioAnalyser 2100 de la société Agilent Biotechnologies est un système d'électrophorèse en capillaires automatisé qui permet l'analyse qualitative et quantitative d'acides nucléiques à partir d'un volume de 1µL déposé sur la puce du BioAnalyser en utilisant un marqueur fluorescent qui permet la visualisation de l'ADN. Cet appareil est utilisé pour le contrôle qualité des ARNs, des protéines et des cellules entières ainsi que pour le contrôle qualité des banques d'ADN qui vont être par la suite séquencées par les séquenceurs de nouvelle génération NGS. Quelques banques ont été analysées par BioAnalyser pour avoir une idée du pourcentage de dimères d'adaptateurs, de la quantité d'ADN ainsi que de la distribution de taille des fragments d'ADN présents dans la banque. De plus, la concentration d'ADN dans les différents mélanges de banques purifiées a été mesurée en suivant le protocole Quant-It dsDNA HS kit (Invitrogen). Nous faisons aussi des quantifications par qPCR des mélanges de banques qui sont largement dilués (1/10000 et 1/100000) car les banques sont amplifiées et peuvent être plus concentrées quand on mélange une centaine de banques d'ADN à la fois.

Chapitre III : La PCR multiplexe couplée au séquençage de nouvelle génération (NGS)

La PCR multiplex est une technique de biologie moléculaire qui permet l'amplification de plusieurs séquences d'ADN cibles dans une seule expérience de PCR en utilisant plusieurs paires d'amorces qui sont désignées de façon à minimiser les dimères d'amorces pour ne pas diminuer l'efficacité de la PCR. En comparant avec la PCR classique, cette technique a le potentiel de générer des économies considérables de temps et d'efforts au sein du laboratoire et grâce au développement de technologies de séquençage, on parle aujourd'hui d'aMPlex qui est une combinaison entre la PCR multiplex et le séquençage de nouvelle génération (NGS) (Guimaraes et al., 2017). Pour analyser des séquences d'ADN de petite taille contenues dans des vestiges fossiles, il faut utiliser des méthodes de haut débit pour générer des données suffisantes pour une analyse phylogénétique. Pour analyser une série d'un grand nombre d'échantillons d'âge et de géographies différents, une aMPlex a été réalisée. L'objectif de cette méthode d'analyse est de séquencer les produits PCR amplifiés grâce à la présence des paires d'amorces qui ciblent des régions connues de l'ADN mitochondrial bovin afin de déterminer les différents haplogoupes mitochondriaux.

Les deux méthodes aMPlex et séquençage Shotgun, bien qu'elles présentent des principes différents, exigent la présence de séquences d'ADN en quantités suffisantes pour permettre une détermination fiable et précise de la séquence. La méthode Shotgun permet de séquencer toutes les molécules uniques qui existent dans l'extrait fossile mais à l'issue du séquençage, le pourcentage d'ADN endogène peut être faible et dilué avec l'ADN contaminant et environnemental d'où la nécessité de capturer et enrichir les banques d'ADN en ADNmt. Par contre, le principe est différent pour l'aMPlex où on amplifie en premier lieu des régions cibles de taille connue par des paires d'amorces spécifiques et on séquence en deuxième lieu les produits de PCR obtenus. La différence principale est que le Shotgun permet la mesure d'une proportion relative de l'ADN présent dans l'extrait fossile alors que l'aMPlex ne dépend que de la quantité absolue d'ADN présente presque indépendamment de la proportion d'ADN environnemental.

Contrairement au séquençage shotgun qui permet de détecter les désaminations des cytosines aux extrémités des fragments, avec une PCR ciblée, les séquences des extrémités des fragments proviennent des amorces et pas de la molécule ancienne, et donc cette approche ne permet pas d'évaluer correctement l'étendue de ces dommages, et donc d'évaluer la probabilité d'avoir affaire à une molécule ancienne authentique. L'approche aMPlex peut toutefois rester utile pour étudier à moindre coût de nombreux échantillons mal conservés car malgré la puissance de la méthode de capture par hybridation pour la génération d'une information génétique fiable, elle est limitée par l'état de conservation et peut s'avérer coûteuse et assez lourde à mettre en œuvre lorsque la plupart des échantillons sont mal conservés (Gansauge & Meyer, 2014). Si la technologie aMPlex a été précédemment optimisée, elle peut être encore utile pour répondre à certaines questions. Je l'ai ainsi utilisé pour génotyper des mitogénomes bovins sur une série d'échantillons mal préservés pour pouvoir évaluer la distribution des haplotypes mitochondriaux au sein d'une population bovine méditerranéenne au Néolithique.

I) Conceptualisation des amorces pour la PCR multiplex :

Pour identifier les fragments d'ADN contenant les polymorphismes nucléotidiques qui distinguent les haplogroupes mitochondriaux T et Q du genre Bos un alignement multiple a été créé pour toutes les séquences mitochondriales présentes dans les bases des données. Des amorces qui ciblent la région contenant le SNP déterminant l'haplogroupe ont été déterminées pour les utiliser lors de la PCR multiplex. La sélection des amorces est l'étape la plus sensible de l'aMPlex. Elle se fait in silico à l'aide de programmes informatiques Primer3 et Oligo7. Cette étape permet de tester la possibilité des amorces de former des dimères en estimant l'énergie libre de formation des hybrides appariés en 3' (delta G en kcal/mol). La stabilité du dimère d'amorce formé est fonction de la valeur de ce delta G. En effet un delta G élevé en valeur absolue traduit une stabilité importante des dimères d'amorce. Plusieurs combinaisons ont été faites pour déterminer les paires d'amorces qui n'ont pas tendance à former des dimères pour les mettre ensemble lors de la PCR. Il faut éviter la formation des dimères du coté 3' car ceci permet d'amplifier les séquences d'amorces et fausser les résultats alors que les dimères d'amorces à l'extrémité 5' sont tolérés car ils ne permettent pas l'amplification des dimères d'amorces. Un exemple explicatif est représenté en dessous.
Dimère d'amorces à l'extrémité 3' :

5'ATGTCCTGTGACC ATTGACTGT 3'---------->
< ---------- 3'AATAACTGACATGTATCATGTAATACAGTT 5'

Dimère d'amorces à l'extrémité 5' :
5' CTTG CTTAACTGCATCTTGAGCACCATTCGACCGGT3'
3'GAACGAATTGACGTAGAACTCGTGGT 5'

Différents paires d'amorces ont été choisies et le tableau 2 montre les haplogroupes mitochondriaux et les polymorphismes nucléotidiques ciblés par chaque paire d'amorce ainsi que la région hypervariable ciblée par 3 paires d'amorces.

Haplogroupe mitochondrial ciblé	Nom du couple d'amorces	Séquence de l'amorce 5'-3'	Taille d'amorce (pb)	Taille de produit PCR	Taille de produit PCR+ adaptateurs Illumina
HVR	BB34s	Sens(BB3S)	20	84	220
		Antisens(BB4S)	30		
	BB34m	Sens(BB3m)	30	94	230
		Antisens(BB4m)	20		
	BB3rev4mod	Sens(BB3rev)	21	158	294
		Antisens(BB4mod)	21		
T	SNP10729	Sens	24	75	211
		Antisens	22		
	SNP16207	Sens	21	85	221
		Antisens	22		
	SNP163	Sens	22	66	202
		Antisens	22		
	SNP13002	Sens	22	83	219
		Antisens	20		
	SNP12751/771	Sens	25	83	219
		Antisens	20		
Q	Q2	Sens	18	71	207
		Antisens	20		
	Q3	Sens	23	75	211
		Antisens	20		
	Q5	Sens	23	76	212
		Antisens	23		
	Q7	Sens	22	69	205
		Antisens	23		
	Q8	Sens	22	56	192
		Antisens	20		
	Q10	Sens	23	91	227
		Antisens	21		
	Q11	Sens	19	68	204
		Antisens	20		

Tableau 2: Les haplogroupes mitochondriaux, la région hypervariable et la taille des fragments d'ADN ciblés par les différentes paires d'amorces (Equipe Epigénome & Paléogénome, IJM).

II) Choix de mélange d'amorces :

Après la vérification in silico de la tendance des amorces à former des dimères, une qPCR a été réalisée sur des contrôles négatifs pour s'assurer de la compatibilité des paires d'amorces. Trois mélanges de paires d'amorces ont été choisis et utilisés pour la PCR multiplex pour cibler des séquences connues d'ADN. Le premier mélange, contient les paires d'amorces qui ciblent les SNPsT10729, T16207, T13002, T12751 qui déterminent l'haplogroupe mitochondrial T et la paire d'amorce BB34s qui cible une séquence de la région hypervariable. Le deuxième mix d'amorces contient les paires d'amorces Q2, Q3, Q5, Q7, Q10 qui ciblent les SNPs déterminant l'haplogroupe mitochondrial Q, et la paire d'amorce BB34m qui cible une séquence de la région hypervariable. Le dernier mélange contient la paire d'amorce BB3r4m qui cible un long fragment de la région hypervariable qui chevauche avec les séquences ciblées par les paires d'amorces BB34s et BB34m, Q11 qui ciblent un SNP déterminant l'haplogroupe mitochondrial Q et T163 qui cible un SNP déterminant l'haplogroupe mitochondrial T. La taille des fragments ciblés par les différentes paires d'amorces est représentée dans le tableau 2

III) Amplification par PCR des échantillons anciens par les différents mélanges de paires d'amorces choisis :

Trois traitements de décontamination ont été faits, les deux premiers servent à minimiser le problème de contamination des réactifs et le deuxième permettra d'enlever les cytosines désaminées converties en uraciles et les éventuels produits de PCR provenant des réactions précédentes.

Les traitements de décontamination des réactifs avec le monoazide d'éthidium (EMA) sont réalisés sur tout le mélange réactionnel alors que le traitement hlDnase a concerné seulement la Taq ADN polymérase. Le troisième traitement impliquant l'UNG est réalisé sur tout le mélange réactionnel lorsque les extraits ont été ajoutés.

a) Traitement de l'enzyme Taq NEB HotStart avec la hl-dsDnase (heat-labile dsDNase) : La hl-dsDNase est conçue pour éliminer les contaminants d'ADN génomique. C'est une endonucléase qui clive les liaisons phosphodiester dans l'ADN pour donner des oligonucléotides avec des extrémités 5'-phosphate et 3'-hydroxyle. Elle a une activité hautement spécifique pour l'ADN double brin. La dsDNase est facilement inactivée par traitement thermique à 55°C. Ces caractéristiques constituent un choix excellent pour la décontamination des ADN polymérases thermo-résistantes. La hl-dsDNase est supplémentée d'un mélange de $MgCl_2$ (10mM finale)/$CaCl_2$ (1mM finale) et 300mM de DTT qui sont requis pour l'activité ou l'inactivation de la dsDNase. Un volume de 270µl de Taq NEB a été mis en présence de 12µL de hl-dsDNase de concentration finale 1.5U/µl et de 14.85µl du mélange $MgCl_2$ (10mM finale)/$CaCl_2$ (concentration finale 1mM). La hl-dsDNase est incubée à 25°C pendant 30min. Après incubation, 1µl de DTT (300µM) est ajouté pour permettre l'inactivation de la hl-dsDNAse à une température de 55°C pendant 30min. Le temps et la température de l'inactivation de hl-dsDNase ont été optimisés de façon à minimiser les dommages qui peuvent affecter le mélange de PCR (Champlot et al., 2010).

b) Traitement de décontamination à l'EMA : En plus de l'irradiation aux UV du tampon10x de la PCR (200mM Tris HCl, 500 mM KCl), de la HSA (sérum albumine de cheval) et du $MgCl_2$, le mélange réactionnel a été décontaminé en l'incubant avec l'EMA à la concentration finale de 6µM, pendant 20 min à 4°C suivie de 15min dans la lumière Phastblue aussi à 4°C.

Les PCR sont faites dans un volume de 50µl contenant 4µl d'extrait d'ADN et 46µl du mélange réactionnel comportant pour chaque échantillon 5µl de HSA de concentration finale 100µg/ml, 5µl du tampon 10X roche Fast Start avec $MgCl_2$, 1µl du $MgCl_2$(25mM),0.4µl de l'enzyme Taq Hot Start NEB traité avec la hl-DNase,0.25µl de dNTPs (50mM) et 0.1µl de l'enzyme UNG (Uracile N-Glycosylase) à 5U/µl. un volume de 1.5µl du mélange d'amorces à 5µM est ajouté. Le volume final est ajusté avec de l'eau. La PCR a été programmée de la manière suivante : une étape d'activation de l'enzyme UNG de 15min à 37°C, une deuxième étape d'inactivation d'UNG à 95°C. Cette étape qui dure 10min consiste à inactiver l'UNG qui est une enzyme thermosensible et dégrader l'ADN ciblé par cette enzyme et puisque la Taq polymérase est de type Hot Start (à démarrage chaud) la température de 95°C est adéquate pour son activation. 35 cycles de PCR ont été faits par la suite comportant pour chaque cycle une dénaturation de l'ADN pendant 10s à 95°C, une hybridation des amorces pendant une min à 60°C et une élongation des amorces à 72°C pendant 4 min. Chaque échantillon a été amplifié en présence de chaque mélange d'amorces séparément. Les produitsde PCR obtenus avec chaque mélange d'amorces vont être mélangés ensemble pour l'étape suivante qui est la construction des banques d'ADNa.

IV) Construction et purification des banques d’ADNa à partir des produits PCR :

La construction des banques d’ADNa a été réalisée de la même manière que la construction des banques génomiques allant de l’étape de réparation des extrémités pour obtenir des extrémités à bouts francs jusqu’à la purification des banques. La différence entre les banques d’ADN génomiques et les banques construites à partir des produits PCR est l’étape de purification après la ligation des adaptateurs qui se fait par le E-Gel Size Selection. Cette étape consiste à sélectionner les fragments de taille attendue, obtenus par l’élongation des paires d’amorces lors de la PCR multiplex. La migration sur l’E-gel permet de séparer les dimères d’adaptateurs des produits PCR et ceci en utilisant un marqueur de taille qui sert de repère pour récupérer les bandes d’ADN d’intérêt. Une fois que les banques d’ADN sont construites et amplifiées, un BioAnalyser a été fait sur toutes les récupérations issues de la purification par le E-gel et un mélange a été préparé et quantifié contenant tous les fragments de taille attendue. Ce mélange a servi pour le séquençage Illumina.

Chapitre IV: Capture des mitogénomes

Pour étudier la phylogénie des populations bovines et leurs distributions phylogéographiques qui est basée, dans le cadre de ma thèse, sur l'analyse du génome mitochondrial, il est nécessaire de reconstruire le mitogénome complet et d'enrichir donc les banques génomiques en ADN mitochondrial. Le séquençage Shotgun de la plateforme Illumina permet de séquencer de façon aléatoire les fragments d'ADN inclus dans les banques d'ADNa. Donc les fragments d'ADN mitochondriaux représentent un pourcentage faible de l'ADN total séquencé.

La capture du génome mitochondrial ou de certaines portions du génome nucléaire permet l'obtention de résultats à partir d'extraits contenant une faible proportion d'ADN endogène. La capture permet de reconstruire le mitogénome complet de nombreux spécimens anciens à des coûts raisonnables.Par conséquent, cette méthode de capture des génomes consiste à augmenter considérablement la proportion d'ADN d'intérêt ou cible. En enrichissant l'ADN cible, on se débarrasse de l'ADN contaminant ou environnemental qui est considéré comme le problème majeur de l'ADNa. En effet, l'enrichissement permettra d'augmenter la présence des régions ciblées même en présence de quantités importantes d'ADN contaminant. Les stratégies de capture, augmentent le rendement et réduisent le coût prévu et réel des projets de séquençage à haut débit de l'ADNa (Enk et al., 2014).

I) Principe général de la capture :

Le principe de la capture du génome mitochondrial consiste à hybrider les fragments d'ADN mitochondriaux contenus dans les banques d'ADNa à des sondes d'ARN complémentaires biotinylées (Figure 26). Les hybrides issus de cette étape sont séparés du reste des fragments d'ADN non ciblés par l'utilisation des billes magnétiques couplées à la streptavidine. Certaines équipes ont réalisé la capture avec des appâts d'ADN doubles brins en utilisant soit des molécules simple brins biotinylés immobilisées sur des billes de streptavidine pour empêcher leurs ré-appariement (Maricic et al., 2010), soit une synthèse directe sur une plaque de verre (Paijmans et al., 2016).

Les molécules accrochées de façon non spécifique vont être éliminées par lavage et les fragments capturés vont être élués et purifiés. Des fragments d'environ 1,6kb du génome mitochondrial bovin moderne portant un promoteur d'ARN polymérase T3 ont été synthétisés par PCR. Ces fragments d'ADN sont chevauchant et couvrent tout le mitogénome. Ces fragments ont été par la suite transcrits in vitro en présence d'un nucléotide biotinylé pour obtenir les sondes ARNs qui sont utilisées comme appâts lors de l'hybridation. Ces ARN ont ensuite été hydrolysés pour avoir une taille moyenne de quelques centaines de nucléotides et utilisés pour la capture d'ADN de bovins anciens. En plus des sondes biotinylées, des ARN dits « bloquants » complémentaires à la séquence des adaptateurs P5/P7 de la plateforme Illumina sont utilisés en large excès pour s'hybrider aux adaptateurs et neutraliser toute interaction possible entre eux. En effet, les banques d'ADNa dénaturées sont incubées en présence des appâts d'ARN dans une solution saline qui permet l'hybridation.

Si les séquences qui sont très concentrées n'étaient pas neutralisées, elles seraient susceptibles de s'hybrider entre elles et d'entraîner la formation de chaînes de fragments simple brin, induisant ainsi la capture de séquences non spécifiques ce qui diminuerait l'efficacité d'enrichissement de la banque. C'est pourquoi il est nécessaire de bloquer les adaptateurs. Les séquences répétées vont être bloquées avec de l'ADN bovin Cot1.

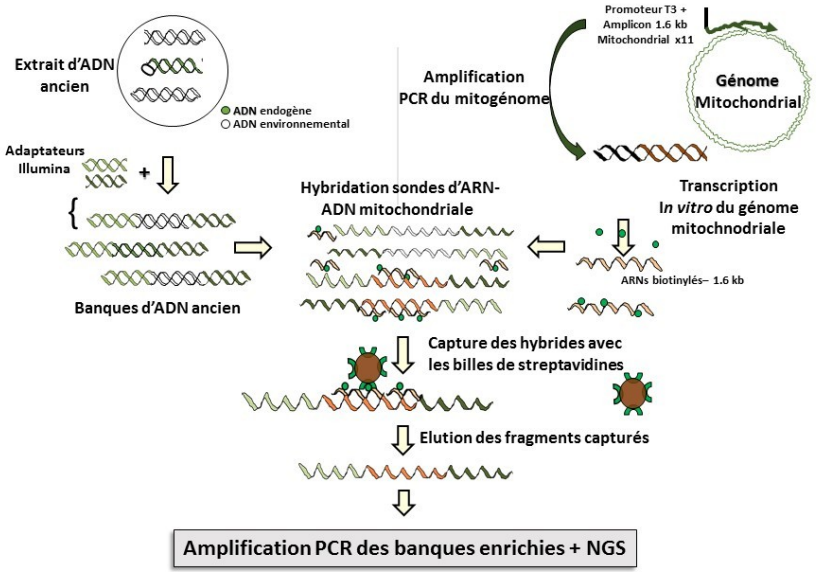


Figure 26 : Principe de la capture par hybridation du génome mitochondrial, d'après la thèse de Diyendo Massilani, 2016.

Pour bien enrichir les banques, on a réalisé deux hybridations qui ont duré chacune 56h. L'étape d'hybridation a été réalisée dans des conditions stringentes pour que seulement les séquences d'intérêt soient hybridées. Cette étape est critique car il faut permettre l'hybridation des séquences qui contiennent des mutations, et ce d'autant plus que les fragments d'ADNa et que certaines régions du mitogénome sont très riches en AT. Il faut alors trouver la bonne stringence. Les hybrides issus de cette étape sont séparés du reste des fragments d'ADN non ciblés par l'utilisation de billes magnétiques couplées à la streptavidine. Les molécules accrochées de façon non spécifique vont être éliminées par lavage et les fragments capturés vont être élués et purifiés. Les lavages ont été faits avec des solutions salines ou la stringence a été bien maîtrisée afin de récupérer seulement les séquences hybridées aux sondes ARN et pour minimiser la perte des petits fragments riches en AT (voir plus loin).

L'efficacité de la capture est influencée par les adaptateurs et les séquences répétées qui peuvent causer la capture simultanée de séquences non complémentaires à l'ADN mitochondrial qui vont diluer les séquences désirées.Les adaptateurs sont capables de s'apparier entre eux par complémentarité ce qui entraîne la formation de chaîne de séquences permettant la capture des séquences non spécifiques ce qui diminue l'efficacité de l'enrichissement de la banque.

Dans des banques d'ADNa les fragments d'ADN contiennent 140 bp d'adaptateurs alors que la plupart des inserts ont une taille inférieure à 50 bp. On a donc synthétisé des ARN appelés ARN bloquants complémentaires à la séquence des adaptateurs P5/P7 de la plateforme Illumina. Ces ARN ont été ajoutés en large excès pour s'hybrider aux adaptateurs et minimiser les interactions entre eux.

Dans le génome bovin comme tous les autres génomes, il y a une fraction répétée importante: des séquences hautement répétées comme les séquences répétées en tandem, des séquences modérément répétées, des séquences faiblement répétées. Lors de l'hybridation, ces séquences vont former des chaînes, en s'hybridant par complémentarité, car la cinétique de réassociation de la fraction répétée est plus rapide que celle des séquences uniques lors de l'hybridation. C'est pour cela qu'on les bloque avec des séquences répétées présentes en excès en utilisant l'ADN Cot1 qui correspond aux séquences hautement répétées présentes dans le génome bovin.

II) 1ère étape de la capture par hybridation: Synthèse des appâts :

La capture des séquences nucléotidiques par hybridation dépend des propriétés spécifiques des appâts et de la qualité des fragments d'ADN obtenus dans les banques et de la couverture des séquences à capturer. En effet, plus les fragments à capturer sont abondants dans la banque, plus la capture sera efficace. Autrement dit, la proportion des fragments d'ADN d'une région génomique spécifique dans une banque dépend de la proportion initiale de cette région au sein du génome de l'organisme ciblé mais aussi de la fraction d'ADN endogène dans l'extrait. La représentativité de certaines régions dans l'extrait fossile dépend de l'ampleur de la dégradation de l'ADN qui peut être variable selon les régions, et aussi de la capacité à purifier cet ADN qui est aussi variable selon les régions. Ainsi les régions riches en AT sont généralement moins bien représentées dans les banques (voir plus loin). La diversité inter-échantillons anciens est prise en compte lors de l'enrichissement des banques par l'utilisation de différentes températures d'hybridation utilisé lors de la capture. Les protocoles sont optimisés pour permettre de capturer le maximum de fragments possibles. La réussite de l'étape de capture dépend de la nature des appâts, de leur taille et de leur concentration. Diyendo Massilani a mis au point pendant sa thèse, dans notre laboratoire, les différents paramètres nécessaires au bon déroulement de la capture (Massilani et al., 2016). J'ai encore raffiné les paramètres pour améliorer la couverture des mitogénomes.

1) Les conditions de synthèse des appâts :

Pour la synthèse des appâts, différentes questions ont été posées pour pouvoir augmenter les chances de rencontres entre les appâts et les fragments d'intérêt et la probabilité d'hybridation de tous les fragments ciblés contenus dans les banques sans avoir des hybridations partielles. Mon équipe se demandaient si (1) les régions adaptatrices trouvées des deux cotés des fragments d'ADN peuvent ou non gêner l'hybridation et déstabiliser l'hybride si l'oligonucléotide appât est plus long qu'il doit être, (2) la taille des appâts doit être comparable ou non à la région à capturer.

Le choix de mon équipe s'est tourné vers une stratégie d'hybridation en solution en utilisant des appâts d'ARN simple brin au lieu d'appâts d'ADN, souvent utilisés pour la capture (Massilani et al., 2016). L'intérêt majeur de l'utilisation des appâts ARN est la possibilité de distinguer, après capture, les appâts d'ARN. En effet, après capture, les banques seront amplifiés par PCR. Les molécules d'ARN qui pourraient encore rester avec l'ADN capturé ne seront pas amplifié par les ADN polymérases et donc ne pourront pas entraîner la production des molécules chimériques entre les appâts et les molécules cibles qui aboutiraient à des données de séquences artéfactuelles. De plus, il est facile et simple par une transcription in vitro de produire une quantité importante d'ARN simple brin biotinylés à partir de produits de PCR amplifiés à partir d'ADN de vache européenne moderne. Une fois produits, les appâts peuvent être conservés et utilisés plusieurs fois. Cette méthode est donc efficace et économe.

2) Synthèse d'appâts d'ARN et transcription :

a) Amplification des fragments d'ADN moderne :

Puisque les appâts utilisés pour la capture sont des appâts ARN qui couvrent l'ensemble du génome mitochondrial, il fallait que nous amplifions, au préalable, les fragments d'ADN recouvrant tout le génome mitochondrial servant comme matrice pour la synthèse des appâts ARN. Nous avons conçus des amorces permettant de cibler et d'amplifier des fragments (voir tableau 3 et 4). Chaque fragment fait environ 1,6kb. La séquence du promoteur T3 de bactériophages a été ajoutée en 5' des amorces sur le brin H de chaque fragment, afin de permettre la transcription d'un seul des brins. A partir d'une séquences consensus du génome mitochondrial bovin, 11 couples d'amorces (Tableau 3) ont été conçus. En effet, la conception d'amorce se fait avec soin et précaution afin de minimiser le risque de formation de dimères entre les amorces du coté 3' en utilisant le logiciel OLIGO 7 (Rychlik, 2007). Les amorces ont une température d'hybridation (Tm) d'environ 58°C. Ces fragments sont partiellement chevauchants pour pouvoir amplifier l'intégralité du génome mitochondrial et ne pas perdre d'information génétique.

Les fragments d'ADN mitochondrial bovin moderne ne présentent pas la même composition en nucléotides ce qui fait que leur hybridation lors de l'amplification par PCR ne se fait pas à la même température. Pour cela, il fallait que nous testions différentes températures pour arriver à amplifier les 11 fragments qui couvrent le mitogénome d'une vache moderne et qui sont chevauchants (Tableau 4).

Nom de l'amorce	Séquence nucéoltidique	Taille de l'amorce	Taille du fragment ciblé (pb)
T3_MBos100-1500F	GCAAATTAACCCTCACTAAAGGGTGACCCGGAGCATCTATTGT	Sens 20pb	1464
MBos100-1500R	CATCGTTCCCTTGCGGTACT	anti-sens 20pb	
T3_MBos1500-3000F	GCAAATTAACCCTCACTAAAGGGCCAAAGATACCCTCTCGACTAAA	Sens 23pb	1556
MBos1500-3000R(Bos_Mito_R2)	CTCTGCCACCTTAACT	anti-sens 16pb	
T3_MBos3000_4500F	GCAAATTAACCCTCACTAAAGGGCACAAAACCCTGCCCTAGAAC	Sens 21pb	1745
MBos3000-4500R	GCCTCCAATTAGGATTGATAAAAC	anti-sens 24pb	
T3_MBos4500-6000F	GCAAATTAACCCTCACTAAAGGGCCAAATCTTCCCATCAATTAAC	Sens 22pb	1343
MBos4500-6000R	CCTGCCCCAGCTTCAACT	anti-sens 18pb	
T3_MBos6000-7500F	GCAAATTAACCCTCACTAAAGGGTCCCTCCCTCATTCCTACTACTC	Sens 23pb	1565
MBos6000-7500R	GGCGGGCAGAATGGTTCAG	anti-sens 19	
T3_MBos7500-9000F	GCAAATTAACCCTCACTAAAGGGATGACCACACGCTAATAATTGTCT	Sens 24pb	1555
MBos7500-9000R	GGTCAAGGGCTTGGGTTTACT	anti-sens 21pb	
T3_MBos9000-10500F	GCAAATTAACCCTCACTAAAGGGACAACACATAATGACACACCAAAC	Sens 24pb	1670
MBos9000-10500R	ATGAGGAGGAGGCTTGTAAAGCTA	anti-sens 24pb	
T3_MBos10500-12000F	GCAAATTAACCCTCACTAAAGGGCCTACTAGTCTTCGCAGCCTGT	Sens 22pb	1758
MBos10500-12000R	GGGTAGTTGGAAGGTTTGTAGGTGT	anti-sens 25pb	
T3_MBos12000-13500F	GCAAATTAACCCTCACTAAAGGGACCCCTACTACTCCTAACCCTAAA	sens 24pb	1553
MBos12000-13500R	GGGTAGGGAATCGGGGTTGT	anti-sens 20pb	
T3_MBos13500-15000F	GCAAATTAACCCTCACTAAAGGGACGCCTGAGCCCTTCTAATAAC	Sens 22pb	1540
MBos13500-15000R	GGCTATTACTGTGAGCAGAAGGATTAC	anti-sens 27pb	
T3_MBos15000-100F	GCAAATTAACCCTCACTAAAGGGACGGAGCTTCAATGTTTTTTATCT	Sens 24pb	14589
MBos15000-100R	TTTATGTCCTGTGACCATTGACTGT	anti-sens 25pb	

Tableau 3: Couples d'amorces utilisés pour l'amplification des différents fragments mitochondriaux bovins modernes. Toutes les amorces sens contiennent le promoteur T3 alors que que les amorces anti-sens sont vides. S : brin sens, AS : brin antisens. La séquence promotrice T3 est indiquée en rouge.

Fragment d'ADN ciblé	Couples d'amorces	Température d'hybridation
fragment 1	T3 15000F_100 R (F11_R11)	58°C
fragment 2	T3 100F_1500 R (F1_R1)	58°C
fragment 3	T3 1500F_3000 R (F2_R2)	54°C
fragment 4	T3 3000F_4500 R (F3_R3)	58°C
fragment 5	T3 4500F_6000 R (F4_R4)	58°C
fragment 6	T3 6000F_7500 R (F5_R5)	58°C
fragment 7	T3 7500F_9000 R (F6_R6)	58°C
fragment8	T3 9000F_10500 R (F7_R7)	58°C
fragment 9	T3 10500F_12000 R (F8_R8)	56°C
fragment 10	T3 12000F_13500 R (F9_R9)	58°C
fragment 11	T3 13500F_15000 R (F10_R10)	58°C

Tableau 4 : La température d'hybridation utilisée pour l'amplification de chaque fragment d'ADN mitochondrial bovin moderne.

Nous avons réussi à amplifier 9 (F1, F2, F4, F5, F6, F7,F8, F10 et F11) fragments, comme le montre la figure suivante, à une température d'hybridation de 58°C en utilisant la Taq NEB Hot Start et les amorces sens et anti sens qui permettent l'amplification du fragment désiré (Figure 27).

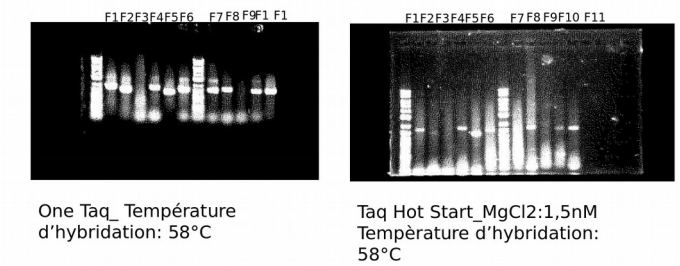


Figure 27: Gel d'agarose 1% de 11 fragments d'ADN bovin moderne à une température d'hybridation de 58°C en utilisant soit One Taq,soit TaqHot Start de NEB.

Pour permettre l'amplification des fragments qui n'étaient pas amplifiés dans les conditions des autres fragments amplifiés. Nous avons testé l'effet du changement de température d'hybridation sur la possibilité de leurs amplifications. Nous avons alors baissé la température d'hybridation à 56°C et 54°C. En plus de la température d'hybridation, nous avons essayé d'amplifier les fragments manquants en présence soit de l'enzyme One Taq de NEB (Option1) soit de l'enzyme Taq HotStart avec 3nM MgCl2 (Option2), soit de l'enzyme Taq HotStart avec 1.5nM MgCl2 (Option3). Le MgCl2 a été utilisé pour faciliter la réaction de l'amplification car les ions Mg2+ fonctionnent comme co-facteur pour l'activité de l'ADN polymérase en aidant incorporation des dNTPs pendant la polymérisation. Nous avons réussi à amplifier le fragments 9 à 56°C en utilisant l'enzyme One Taq, alors que nous n'avons pu amplifier le fragment 3 qu'avec One Taq à 54°C (Figure 28).

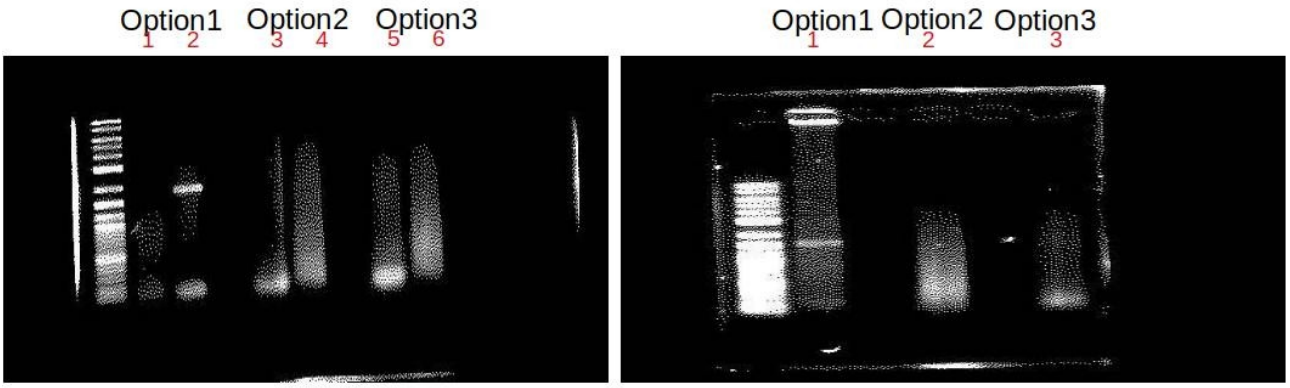


Figure 28: A gauche : Gel d'agarose 1%: Amplification de fragment 3 et 9 dans 3 conditions différentes à une température d'hybridation de 56°C. Les chiffres impaires et paires correspondent, respectivement, au fragment 3 et 9 amplifiés en suivant les différentes options déjà discutées ci-dessus. A gauche: Gel d'agarose 1%: Amplification du fragment 3 dans 3 conditions différentes à une température d'hybridation de 54°C. Les chiffres correspondent, respectivement, aux options d'amplification déjà discutées ci-dessus.

b) Transcription in vitro :

Nous avons utilisé le kit MegaScript de la compagnie Thermo Fisher qui nous a permis d'obtenir 100 µL de transcrits très concentrés. Notre mélange réactionnel contenait, 1µg de produit de PCR, 10 µL de chaque dNTP 75mM (ATP UTP et GTP) à l'exception du CTP dont le ratio CTP/CTPbiotine a été ajusté à 1.5 avec CTP à 75mM et CTPbiotinylé à 10mM,le tampon de transcription 10x, Ribolock RNase inhibitor (40U/µl), de mix de l'enzyme T3 et de l'eau Nuclease free. Le mélange transcriptionnel a été incubé pendant 4 h à 37°C.Par la suite, 2 µl de Dnase ont été ajoutés et incubés pendant 15 min à 37°C pour dégrader les fragments d'ADN. L'ARN produit est purifiée à l'aide Ambion NucAway Spin Column de la compagnie Thermo Fisher. C'est un outil rapide, efficace et facile à utiliser pour enlever les nucléotides non incorporés. Les ARNs transcrits ont été fragmentés suite à une incubation pendant 3 min à 94°C dans un tampon de fragmentation d'ARN 10X, suivie par l'ajout d'une solution stop de la fragmentation.

Les fragments d'ARNs ont été précipites par l'acétate de sodium 3M, pH 5.2 et l'éthanol 100% conservé à -20°C. Le mélange a été incubé pendant 30 min à -20°C puis centrifugé 25 min à 14000 rpm à 4°C. Le surnageant à été enlevé et nous avons nettoyé le culot d'ARN avec 300 µL d'éthanol froid à 70% conservé à -20°C. Après avoir centrifugé 3 min à 14000 rpm à 4°C et enlevé l'éthanol nous avons mis en suspension les appâts d'ARN dans 100 µL d'eau dépourvue de Rnase. Les sondes d'ARN ont été quantifié par qbit et nous avons obtenu alors 2.7 µg/ µL qui est une quantité importante car nous n'utilisons que 0.06 µg/ µL par réaction de capture. Après chaque étape, nous avons prélevé un microlitre du mélange réactionnel pour analyse sur gel d'agarose afin de s'assurer du bon déroulement de la manipulation (Figure 29). Nous observons que nous avons une bande sur le gel d'agarose après chaque étape, ce qui montre le bon déroulement de l'expérience.

Figure 29: Gel d'agarose 1%: Les puits 1, 2, 3, 4 et 5 correspondent respectivement aux fragments d'ADN mitochondrial bovin moderne, aux transcrits de ces fragments, aux transcrits après traitement avec la Dnase, et en fin aux sondes d'ARN purifiés.

III) 2ème étape: Synthèse des ARNs bloquants :

Des ARN dits « bloquants » complémentaires à la séquence des adaptateurs P5/P7 de la plateforme Illumina sont utilisés et présents en large excès pour s'hybrider aux adaptateurs et neutraliser toute interaction possible entre eux (Figure 30). En effet, les banques d'ADNa dénaturées sont incubées en présence des appâts d'ARN dans une solution saline qui permet l'hybridation. Si les séquences qui sont très concentrées n'étaient pas neutralisées, elles seraient susceptibles de s'hybrider entre elles et d'entraîner la formation de chaînes de fragments simple brin, entraînant la capture des séquences non spécifiques ce qui diminue l'efficacité d'enrichissement de la banque.

Nous avons synthétisé les ARN bloquants en utilisant le kit Thermo Fisher Trxn3. En effet, toutes les banques d'ADN construites contenaient des adaptateurs P5-P7 de la plateforme Illumina. A partir d'une matrice d'oligonucléotides simple brin d'adaptateurs P5 et P7, des molécules d'ADN ont été synthétisées par PCR permettant l'addition d'un promoteur T3 sur le brin H. Ces dernières ont servi pour la transcription in vitro en présence d' ARN polymérase T3 dont l'incubation à duré 30 min à 37°C. Par la suite, nous avons effectué un traitement à la Dnase I pour hydrolyser les molécules d'ADN.

Les transcrits d'ARN bloquants ont été purifiés par extraction au phénol-chloroforme et précipitation à l'EtOH, contrairement aux sondes d'ARNs car les colonnes de silice sont adaptées à purifier des fragments de taille supérieure à100pb. Après purification, les transcrits sont précipités à l'éthanol et l'acétate de sodium. Les ARNs précipités ont été mis en suspension dans 100 µl d'eau dépourvue de RNase et DNase. Nous avons obtenu alors une concentration de 5 µg/µL.

Figure 30: Principe générale de la stratégie d'ARN bloquants des séquences adaptatrices des fragments de banque.

IV) Les séquences répétées :

Les analyse Cot permettent de déterminer le temps nécessaire pour l'hybridation de deux oligonucléotides. Il s'agit de techniques basées sur les principes de cinétique de réassociation de molécules d'ADN en solution (Britten & Davidson, 1960). En effet, lors de l'hybridation, les séquences d'ADN simple brins complémentaires vont se réassocier pour reformer les séquences double brins (Figure 31). La vitesse de réassociation de ces séquences dépend de leurs concentration dans l'extrait d'ADN. Plus les séquences sont abondantes dans l'extrait plus la réassociation est rapide. Dans le génome, la cinétique de réassociation des séquences répétées est donc plus rapide que celle des séquences uniques. Les analyses Cot correspondent alors au produit de la concentration intiale d'ADN (Co) par le temps (t) de réassociation en seconde, d'où Cot (Britten & Davidson 1960).

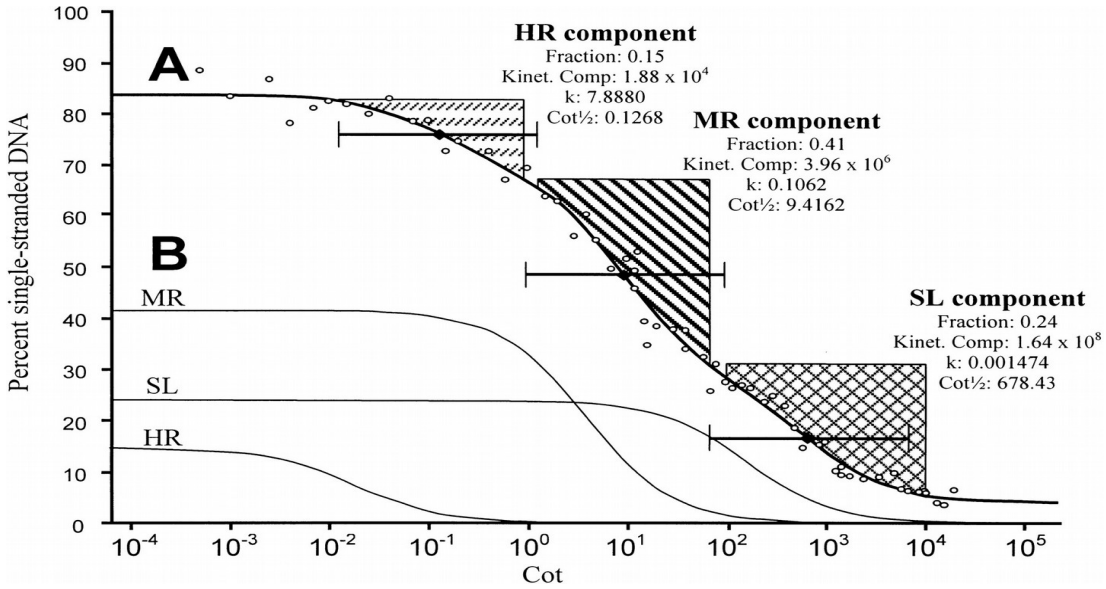


Figure 31: Cinétique de réassociation des séquences d'ADN en se basant sur l'analyse Cot d'après (Peterson et al., 2002).

Comme montré sur la figure 31, dans un extrait génomique d'ADN dénaturé, les régions de séquences répétées se réassocient à des valeurs de Cot faibles, en comparaison avec les autres régions rares et uniques. Lors de l'étape d'hybridation, les séquences répétées peuvent former des chaînes de séquences en s'hybridant par complémentarité. Ces chaînes de séquences seront donc capturées, ce qui entraîne la diminution de l'efficacité de la capture du mitogénome. Pour minimiser la formation de ces chaînes, des séquences répétées complémentaires aux séquences répétées contenues dans l'extrait d'ADNa ont été ajoutées en excès. Ces séquences correspondent aux séquences Cot qui correspondent à la fraction hautement répétée de l'ADN bovin moderne.

V) Hybridation :

Le choix des échantillons qui ont été capturés par hybridation afin d'obtenir les génomes mitochondriaux complets est basé sur leurs pourcentages d'ADN endogène, leurs géographies et leurs âges qui permettent de parcourir une large période temporelle.

Pour réaliser la capture, nous sommes partis de banques d'ADN déjà construites pour le shotgun. Le mélange des banques pour chaque réaction de capture est basé sur le pourcentage d'ADN endogène ainsi que le nombre total des séquences d'ADN totales générées par le séquenceur MiSeq. L'objectif est de rassembler dans un même mélange d'hybridation d'une part, toutes les banques correspondants à un même échantillon, d'autre part, 4 à 5 échantillons différents ayant des propriétés similaires, en particulier en ce qui concerne le pourcentage d'ADN endogène.

L'objectif de mélanger plusieurs extraits est de minimiser le nombre total de captures à traiter afin de pouvoir traiter ensemble un grand nombre d'échantillons. En effet, la capture implique de nombreuses étapes de lavages qui doivent être réalisées de façon reproductible en ce qui concerne la température et les temps d'incubations, ce qui limite le nombre total de captures qui peuvent être réalisées simultanément.

Pour mélanger ensemble plusieurs échantillons, les cinétiques d'hybridation étant fonction de la concentration des molécules cibles, il faut éviter qu'un échantillon occupe tous les appâts plus rapidement que les autres, et il faut donc éviter de mélanger ensemble des échantillons bien conservés, avec une concentration plus élevée d'ADN mitochondrial que les échantillons mal conservés. Finalement, pour bien distinguer les échantillons, il faut que les deux indexes de chaque banque soient différents de ceux des autres banques présentes dans le mélange. Les mélanges des banques d'ADN génomiques ont été quantifiés par le Dosage fluorimétrique Qubit DNA High Sensitivity. Ces banques déjà amplifiées contenant des fragments d'ADN doubles brins sont dénaturées. Des conditions réactionnelles ont été créées pour favoriser la formation d'hybrides entre les fragments d'ADN ciblés et les appâts d'ARN. L'hybridation est favorisée par un large excès des oligonucléotides appâts.
Les conditions de l'étape d'hybridation sont maîtrisées pour permettre la récupération de toutes les molécules d'intérêt, en particulier les molécules les plus divergentes.

1) Hybridation des ARNs aux fragments d'ADN mitochondrial :

Le mélange d'hybridation a été préchauffé à 62°C et 25µl ont été distribué dans chaque réaction de capture. Il contenait 1µl de sonde d'ARN (300ng/µl), 1µl d'ARN bloquants (3ng/µl) et 23µl du tampon d'hybridation 2x. Ce dernier est un mélange d'une solution saline SSPE 10x , EDTA 0.01M pH8, Tween20_0.20% et de l'eau. A ce mélange, nous avons ajouté 23 µl de nos mélanges de banques d'ADN à capturer dont la quantité d'ADN a varié entre 80 et 300 ng/µl. La 1ère réaction a duré 5 jours pendant lesquels la température est abaissée d'un degré chaque 12h, de 62°C à 56°C. Après 3,5 jours, l'hybridation reste à 56°C pendant le temps restant. La deuxième hybridation a aussi duré 5 jours .

2) Capture des ARNs biotinylés par la streptavidine :

Cette étape permet de capturer les fragments d'ADN hybridés aux sondes d'ARNs grâce à une interaction entre la biotine et la streptavidine (DynabeadsMyOne). Par réaction de capture, 30µl de billes ont été lavées 3 fois avec 1.5ml d'une solution saline (1M NaCl, 10mM Tris-HCl pH7.5, 1mM EDTA). Aprés lavage des billes, celles ci ont été resuspendues dans 120µl de la même solution saline, pour chaque réaction de capture, auxquels nous avons ajouté du Denhardt 50x (1% de Ficoll, 1% de Polyvinylpyrollidone et 1% d'albumine sérique bovine) avec une concentration finale d'1x pour chaque réaction de capture. Denhardt 50x est un mélange d'agents de blocage qui permet de réduire les interactions non-spécifiques de l'ADN avec les billes de steptavidine. Le mix d'hybridation a été mis en présence de 120µL de billes de streptavidine et incubé 40 min à 56°C tout en agitant à des intervalles de temps définis. La plaque contenant les réactions de capture a été mise sur une plaque magnétique pour permettre d'enlever le surnageant par pipetage et laisser les billes accrochées sur la paroi des tubes.

3) Lavages :

Les lavages ont été faits pour éliminer toutes les molécules non spécifiques qui peuvent se lier aux sondes d'ARNs. Ils ont été réalisés avec des tampons stringents dont la stringence augmente quand la concentration en sels diminue (Massilani et al., 2016).

Le premier lavage a été fait avec 150µl d'un tampon concentré en sels contenant 1X SSPE et 0.1% Tween pendant 15min à température ambiante. Le deuxième lavage a été fait avec aussi 150µl d'un tampon moins concentré en sels: 0.2X SSPE et 0.1% Tween pendant 15 min à température ambiante puis un deuxième lavage avec le même tampon et le même volume pendant 10 min à 56°C.

4) Élution des banques enrichies :

L'élution de l'ADN a été faite dans 30µl d'un tampon d'élution TET (10 mM Tris HCl pH 8.0, 1 mM EDTA, 0.1% Tween-20) à température élevée (95°C) pendant 5min. En effet, cette incubation permet de séparer les ARNs hybridés sur les molécules d'ADN capturées. Deux élutions ont été réalisées et nous avons récupéré, pour chaque réaction, 60µL de banques capturées par les sondes d'ARNs.

5) Quantification des banques enrichies par qPCR et amplification par PCR :

Après avoir quantifié les banques enrichies par Q-PCR, une PCR a été réalisée pour avoir des quantités suffisantes d'ADNmt capturé.En effet, comme nous avons fait deux hybridations pour chaque réaction de capture, nous avons effectué une amplification par PCR après chaque hybridation. La première PCR vise à produire une grande quantité d'ADN pour que la deuxième hybridation soit efficace. Nous avons alors fait 25 cycles d'amplification dont chaque cycle contient une étape de dénaturation qui dure 20 s à 95°C, une étape d'hybridation qui dure 35s à 60°C et une étape d'élongation qui dure une minute et demi à 68°C. Alors que l'amplification par PCR après la deuxième hybridation vise à obtenir une quantité d'ADN suffisante pour effectuer le séquençage de nouvelle génération Miseq. Pour cela, nous n'avons fait que 12 cycles d'amplification par PCR avec les même temps et température pour chaque étape.

6) Purification des banques enrichies après amplification :

Après chaque hybridation, les banques enrichies ont été purifiées avec des billes magnétiques NucleoMag® NGS Clean-up and Size Select en utilisant un volume qui représente 1.3x le volume de notre mélange de banques enrichies. A la suite de deux tours d'hybridation et purification des banques enrichies, nous avons mélangé 10 µl de chaque mélange de banques enrichies et purifiées. Ce mélange a été, lui même, purifié avec les billes magnétiques en utilisant le même ratio 1.3x. Deux tours de purification ont été effectués et nous avons obtenu deux mélanges dont un a subi deux tours de purification (un tour sur les banques enrichies individuellement et un autre tour sur le mélange des banques) et un deuxième subissant 3 tours de purifications (un tour sur les banques enrichies individuellement et 2 tours sur le mélange des banques). Une fois purifié, les 2 mélanges des banques enrichies et purifiées sont quantifiés par un dosage fluorimétrique Qubit et par BioAnalyser. On a eu des quantités d'ADN comparables entre le dosage Qubit et le BioAnalyser. Le mélange de banques enrichies en ADN mt qui n'a pas présenté un profil de dimères d'adaptateurs sur le BioAnalyser a été séquencé par le séquenceur MiSeq de la plateforme Illumina.

VI) Les facteurs influençant l'efficacité de la capture :

L'efficacité de la capture par hybridation dépend de plusieurs paramètres qui ont été bien maîtrisés dans notre étude pour permettre la récupération et la capture de tous les fragments d'intérêt ciblés.

1) Effet des appâts d’ARN sur l’efficacité de la capture :

Pour que la capture avec les appâts soit efficace, il faut que les séquences de ceux-ci soient complémentaires à celles des régions d’intérêt ciblées, et qu’il y ait une forte interaction entre les appâts et les billes de streptavidine pour que les appâts puissent isoler efficacement les fragments capturés. De plus, les séquences des appâts ne doivent pas être trop divergentes des séquences cibles. Un tel problème peut entraîner la formation d' hybrides trop peu stables pour qu'ils résistent à la stringence des lavages. Plus les séquences cibles sont proches de la séquence des appâts, moins le problème de divergence est important. Dans le cas des aurochs, la divergence peut être plus importante, en particulier en ce qui concerne la région hypervariable.

Dans le cas des bisons, le problème est plus prégnant car la divergence est encore plus élevée, voire très élevée dans certaines parties de la région hypervariable. Ce problème de divergence est particulièrement problématique dans les régions riches en AT car les hybrides sont déjà moins stables et donc encore plus sensibles aux mésappariements. Les conditions de stringence que nous avons utilisée visaient à assurer la meilleure couverture possible du mitogénome même pour les séquences divergentes. Nous avons donc fait le compromis de garder un certain nombre d'hybrides non-spécifiques pour garder les hybrides divergents d'intérêt.

2) Effet de la quantité de biotine sur l’efficacité de la capture :

L’accrochage des biotines aux billes de streptavidine dépend de la quantité de biotine que doit porter chaque appât. Plus la quantité est importante, plus l’accrochage est efficace car lors de la transcription la biotine sera incorporée aux transcrits par l’intermédiaire des cytosines biotinylées et par conséquent, plus il y aura de biotines incorporés plus l’accrochage aux billes sera efficace. Lors de la transcription, il faut s’assurer que les biotines soient en quantité suffisante pour permettre une distribution homogène sur les grandes molécules car, après la transcription, les ARNs d’environ 1.6 kb seront fragmentés. En effet, si la distribution des biotines n’est pas homogène, en fragmentant les grandes molécules, on risque d’avoir des petits fragments ne contenant pas de biotine empêchant ainsi la capture de celui en utilisant les billes de streptavidine. Plus les fragments sont petits, plus le risque est élevé. Si on prévoit d’avoir des fragments de petite taille, il faut que ce ratio soit important. Par contre, le ratio cytosines/cytosines biotinylées affecte la stabilité des hybrides. Le compromis que nous utilisons, qui est un ratio de 1.5 avec une concentration initiale de 75mM pour CTP et 10mM pour CTP cytosines biotinylées, donne des résultats satisfaisants sachant que l'ARN polymérase incorpore mieux le CTP non biotinylé, c'est à dire que le ratio initial n'est pas exactement celui qui existe dans les ARN appâts.

Résultats et Discussion

Une fois que l'on a produit les données de séquençage à partir des banques, nous avons réalisé plusieurs analyses avec des logiciels informatiques pour vérifier la qualité de ces séquences. Les fragments d'ADN sont séquencés à partir de chaque extrémité sur 75 nucléotides en mode paired-end. Les fragments étant pour la plupart plus petits que 150 bp, une partie de la région centrale, voir l'intégralité de la séquence sera lue sur les deux brins. Il est donc plus efficace d'essayer de fusionner ces deux brins avant de les aligner sur le génome d'intérêt. Les séquences étant produites sous forme d'un fichier « FastQ » qui associe la séquence avec sa probabilité d'erreur, les séquences des deux brins d'un même fragment sont fusionnées avec le logiciel leeHom (Renaud et al., 2014).

Les proportions des séquences de mauvaise qualité, des dimères d'adaptateurs sont quantifiées pour chaque banque. De plus, les séquences de taille inférieure à 28pb sont éliminées avant la cartographie sur le génome car ces séquences sont trop courtes pour être cartographiées avec fiabilité. La proportion de ces séquences trop courtes est aussi quantifiée. Après ces contrôles qualités des banques, les séquences restantes sont alignées au génome moderne de la vache européenne avec le logiciel d'alignement des séquences d'ADN BWA (Li & Durbin, 2009). Le nombre des séquences alignées sur le génome de référence pour tous les échantillons anciens est mesuré ainsi que la proportion dont la qualité d'alignement est supérieure à mapQ10. Toutes ces quantifications sont récapitulées dans un tableau Excel qui ne peut pas être présenté dans le rapport à cause de sa grande taille. Certaines des conclusions tirées de ces quantifications seront présentées.

Chapitre I: Optimisation méthodologique et son effet sur la récupération de l'ADN endogène extrait à partir d'échantillons fossiles diversifiés

Nous avons essayé différentes techniques d'extraction d'ADN dans le but d'augmenter les chances de récupération de toutes les molécules uniques contenues dans l'extrait fossile pour augmenter par conséquent la complexité des banques génomiques. En effet, dans ce chapitre, je discute les résultats techniques qui influencent les résultats phylogénétiques basés sur l'ADN mitochondrial et sur les analyses, basées sur l'ADN nucléaire, qui étudient les flux génétiques, la structure de la population et la sélection naturelle. Des études robustes dépendent de la qualité de données de départ. Plus la couverture du mitogénome ou du génome nucléaire est bonne, plus nous pouvons tirer des interprétations et des conclusions avec confiance. Nous avons donc travaillé sur la méthodologie pour obtenir des génomes et mitogénomes bien couverts pour aller plus loin dans nos perspectives et nos objectifs.

I) Relation entre ADN total et ADN endogène :

La quantité d'ADN extraite reflète l'ADN total extrait c'est-à-dire l'ADN endogène et environnemental. Dans des analyses précédentes (lors de mon stage de M2), nous avons montré qu'il n'y a pas une corrélation claire entre la quantité d'ADN totale obtenue et le pourcentage d'ADN endogène. Il paraît clairement que l'on récupère peu d'ADN endogène quand la quantité d'ADN de départ est élevée, ce qui indique effectivement que l'on a souvent beaucoup d'ADN environnemental quand on a beaucoup d'ADN dans un ossement ancien. Par contre, la corrélation réciproque n'est pas observée car on peut avoir des échantillons qui contiennent à la fois peu d'ADN total et peu d'ADN endogène. Une faible quantité d'ADN récupérée ne permet pas de prédire que l'ADN récupéré sera d'intérêt.

On voit dans la figure 32 que les pourcentages d'ADN endogène n'arrivent jamais à 100%, ce qui montre que les échantillons présentent des taux de préservation différents et des niveaux de contamination différents. Ici nous présentons le pourcentage d'ADN endogène vs la quantité totale d'ADN extrait pour les échantillons dont les extraits ont été purifiés avec le tampon de purification 2M70 et QG de Qiagen.

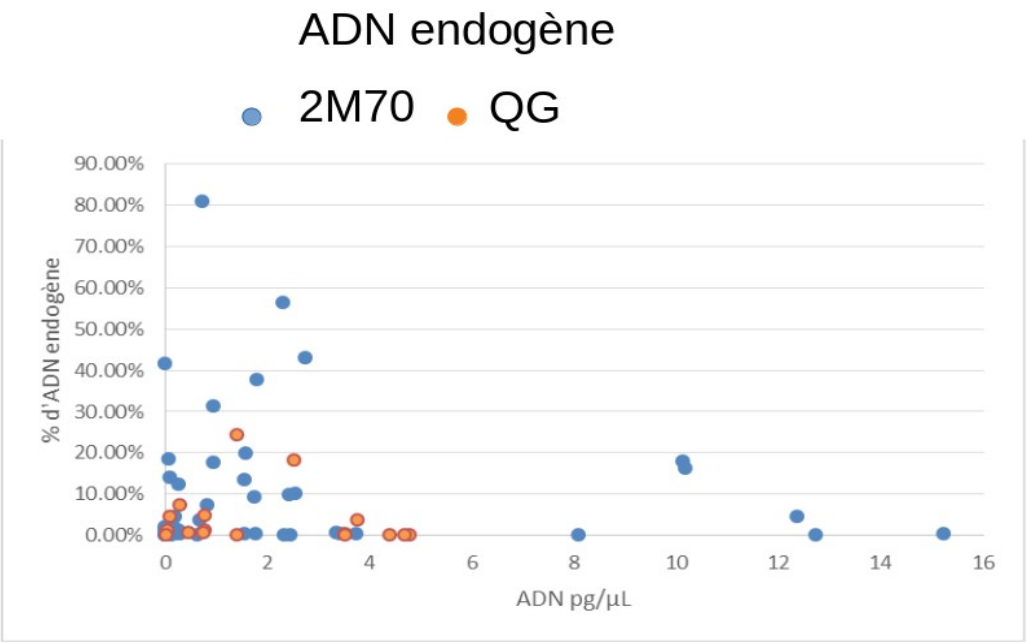


Figure 32: Relation entre la quantité totale d'ADN quantifiée et le pourcentage d'ADN endogène pour tous les échantillons anciens. L'axe des abscisses représente la quantité d'ADN en pg/µL et l'axe des ordonnées représente le pourcentage d'ADN endogène des différents échantillons. Les couleurs bleue et orange représentent respectivement les tampons 2M70 et QG de purification utilisés pour la purification des échantillons.

La figure 32 indique que le tampon 2M70 de purification permet souvent de récupérer une quantité d'ADN endogène plus importante que le tampon QG.

II) Comparaison de l'effet des tampons de purification de l'extrait d'ADN sur la quantité des séquences obtenues par séquençage de nouvelle génération :

1) Comparaison entre les tampons de purification 2M70 et QG :

Les extraits des échantillons montrés sur la figure 33 sont purifiés à la fois par le tampon 2M70 et QG. Ils correspondent à des échantillons vieux de genre *Bos* et *Bison* dont la préservation en ADN endogène n'est pas aussi importante.

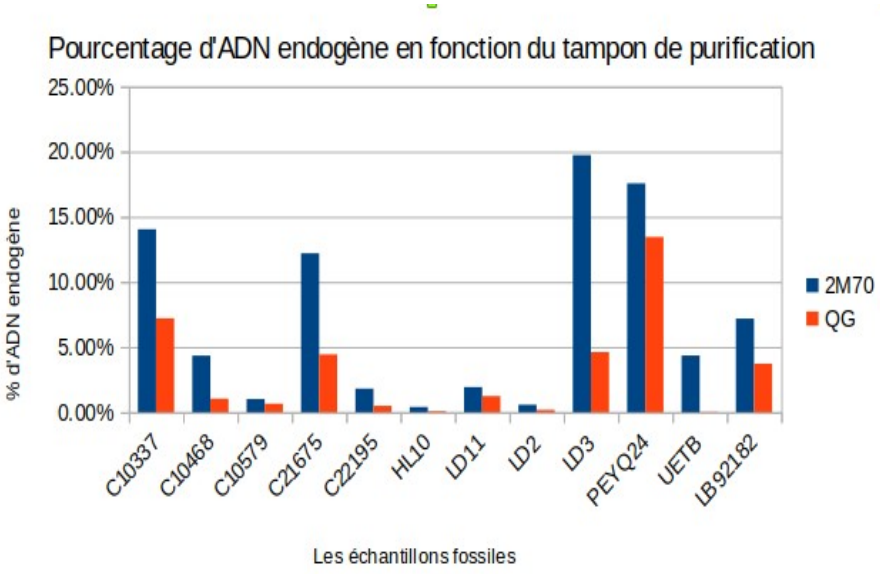


Figure 33: Comparaison du pourcentage d'ADN endogène pour chaque échantillon ancien purifié avec deux tampons (2M70 et QG). La couleur bleue correspond au tampon 2M70 de purification et la couleur orange correspond à la méthode QG.

Malgré la grande variabilité entre les 2 tampons utilisés vis-à-vis la récupération de l'ADN total, on constate que le pourcentage d'ADN endogène est plus important quand les extraits sont purifiés avec les tampons 2M70 plutôt que QG. Nos résultats indiquent qu'une plus grande proportion d'ADNa relativement à l'ADN environnemental peut être obtenue avec le tampon 2M70 de purification. Cette proportion peut atteindre entre 20 et 80% avec des os pétreux, même très anciens, ce qui élargit le champ des possibles des analyses paléogénomiques. Le niveau de la préservation de l'ADN endogène reste toutefois très variable et il dépend essentiellement de la composition de l'os ainsi que des conditions environnementales comme la température, l'hydratation et le pH.

Puisque nous avons observé que la différence de récupération de l'ADN endogène diffère en fonction de tampon de purification et que le pourcentage le moins élevé pour un extrait est retrouvé avec le tampon de purification QG, nous avons décidé de changer le tampon QG par le tampon 5M40. Cette décision était basée sur les résultats de la distribution de taille des séquences et leurs contenus en GC obtenues en fonction des tampons 2M70 et QG (voir plus loin).

2) Comparaison entre les tampons de purification 2M70 et 5M40:

Nos données de séquençage ont montré que le pourcentage d'ADN endogène obtenu pour le même échantillon fossile diffère quand l'extrait de ce dernier est purifié avec le tampon 5M40 (Dabney et al 2013) et le tampon 2M70 (Glocke et al 2017). Le pourcentage d'ADN endogène obtenu quand l'extrait d'ADN est purifié avec le tampon 2M70 est légèrement supérieure à celui obtenu avec le tampon 5M40 (figure 34). Cependant, la différence n'est pas aussi frappante que celle obtenu entre le tampon 2M70 et QG (Figure 33).

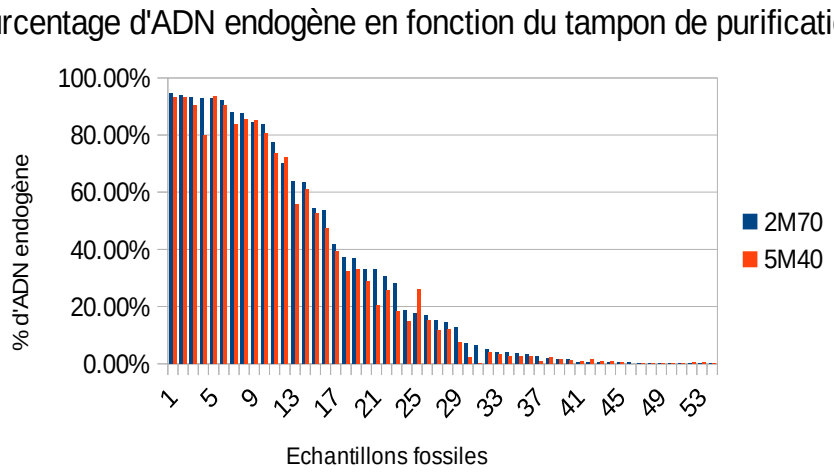


Figure 34: Comparaison du pourcentage d'ADN endogène pour chaque échantillon ancien purifié avec les tampons 2M70 (bleu) et 5M40 (orange).

Nous avons noté aussi que dans certains cas le pourcentage d'ADN endogène obtenu avec le tampon 5M40 est plus élevé que celui obtenu avec 2M70 tel le cas pour l'échantillon 5, 9, 12 et 25. Ce-là pourrait s'expliquer par le fait que le tampon 2M70 laissait trop d'inhibiteurs qui interféraient avec la construction des banques génomiques. Dans certains cas, on ne peut donc pas obtenir de bon résultats avec ce tampon. J'ai exploré alors la différence de quantités d'ADN incorporé dans les banques selon les méthodes de purification et ceci par la comparaison des Deltas Ct des PCR quantitatives avant amplification des fragments d'ADN converties en banques d'ADN génomique (paragraphe suivant).

3) Effet du tampon de purification sur l'élimination des inhibiteurs:

Même dans les cas d'une très bonne préservation de l'ADN, la part de l'ADN environnementale dans les extraits fossiles reste importante. Nous pouvons anticiper la part de l'ADN contaminant dès l'amplification en PCR quantitative en temps réel avant même le séquençage. Généralement, dans le cas d'ADNa une amplification précoce de fragments d'ADN contenus dans les banques génomiques indique que l'ADN de l'échantillon correspondant est mal préservé et donc faiblement présent dans les banques génomiques ou bien l'extrait d'ADN contient des inhibiteurs de l'ADN polymérase qui n'étaient pas efficacement éliminés lors de l'étape de purification.

En effet, le principe de l'amplification en temps réel ou qPCR repose sur la possibilité de suivre la quantité d'ADN présente dans la réaction à tout instant et non pas seulement à la fin de la PCR. Des sondes fluorescentes se fixent sur l'ADN double brin si nous utilisons la technologie SYBRGreen. Ces sondes ne fluorescent qu'une fois fixées à l'ADN. Un seuil de fluorescence est établi par le programme de l'appareil de PCR en temps réel (Higuchi et al., 1993). Une fois que la quantité d'ADN permet aux sondes fluorescentes de dépasser ce seuil alors on obtient un numéro de cycle PCR appelé «Ct» pour «Threshold Cycle» soit «cycle seuil». C'est cette valeur qui est à la base des calculs pour quantifier l'ADN de façon absolue ou relative (Mackay, 2002; S.A. Deepak et al., 2007).

Pour voir l'effet des inhibiteurs sur l'efficacité de la conversion des fragments d'ADN en banques d'ADNa, j'ai construit deux boîtes à moustaches des valeurs Ct obtenues quand les banques sont construites à partir d'extrait purifié avec le tampon 2M70 ou 5M40 et avant leurs amplifications (Figure 35). J'ai choisi 248 banques d'ADNa construites à partir d'extraits traités avec l'enzyme USER et purifiés soit avec le tampon 2M70 soit 5M40. Les valeurs Ct obtenues avec le tampon 2M70 sont plus élevées que celles obtenues avec le tampon 5M40, ce qui montre que l'amplification est retardée quand le tampon 2M70 est utilisé et ceci est du probablement aux inhibiteurs d'enzymes qui ne sont pas bien éliminés lors de la purification et qui sont efficacement enlevés quand la purification de l'extrait d'ADN est réalisée avec le tampon 5M40.

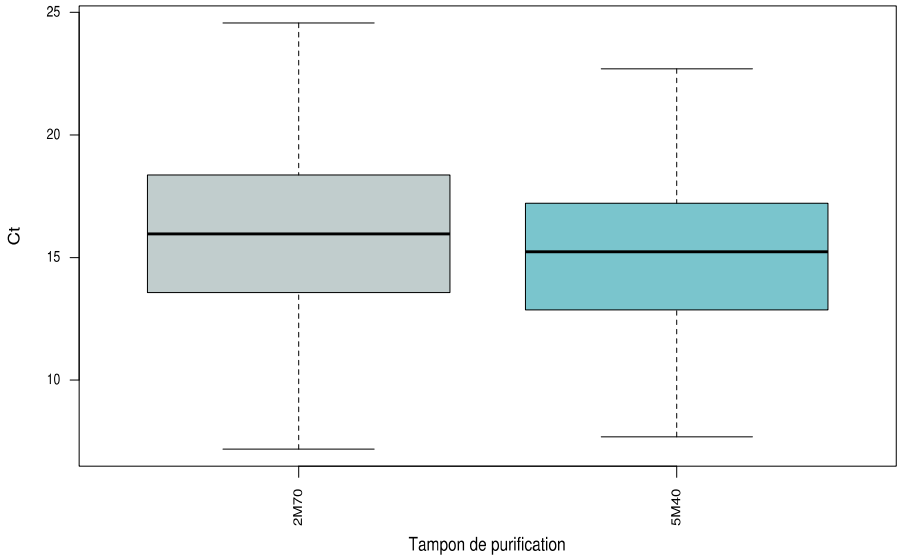


Figure 35: Boîtes à moustaches des valeurs Ct obtenues pour les banques d'ADNa non amplifiées construites à partir d'extraits purifiés soit avec le tampon 2M70 (gris) soit le tampon 5M40 (bleu).

Cet effet est également clair dans les comptages de lecture obtenus après séquençage des banques amplifiées où certains échantillons ont montré une forte inhibition lorsque le tampon de purification 2M70 a été utilisé. Cette inhibition est traduite par une différence remarquable du nombre de séquences totales obtenues avec 2M70 par rapport au nombre de séquences obtenues avec 5M40 (données non présentées).

III) Comparaison de l'effet des tampons de purification de l'extrait d'ADN sur la distribution de la taille des séquences obtenues par séquençage de nouvelle génération :

Pour couvrir le mitogénome complet, il est nécessaire de récupérer tous les fragments d'ADN endogène qui existent dans l'extrait fossile et qui sont de tailles différentes. Une comparaison de la distribution de la taille des séquences d'ADN obtenues avec les 3 tampons de purification a été réalisée pour les échantillons les mieux conservés présentant un nombre élevé de séquences d'ADN.

1) Différence entre le tampon 2M70 et QG sur la distribution de la taille de fragments d'ADN:

Pour étudier l'effet des tampons de purification sur la distribution de taille de fragments d'ADN incorporés dans les banques d'ADN génomiques, nous avons commencé par comparer la taille médiane des fragments obtenues par 2M7O et QG. Nous avons observé une différence de taille des fragments récupérés avec les deux tampons d'extractions. La figure 36 montre que la taille médiane des fragments obtenus pour 30 banques dont les extraits ont été purifiés avec le tampon QG est significativement plus grande que celle obtenue quand les même extraits ont été purifiés avec le tampon 2M70.

Cela suggère que le tampon 2M70 de purification permet de récupérer les fragments de petite taille qui sont présents à un pourcentage élevé, alors que le tampon QG de purification favorise l'obtention des fragments de grande taille. Nos résultats valident les constatations de Matthias Meyer et son équipe qui ont montré que le tampon 2M70 permet de récupérer des fragments de petite taille (Glocke & Meyer, 2017).

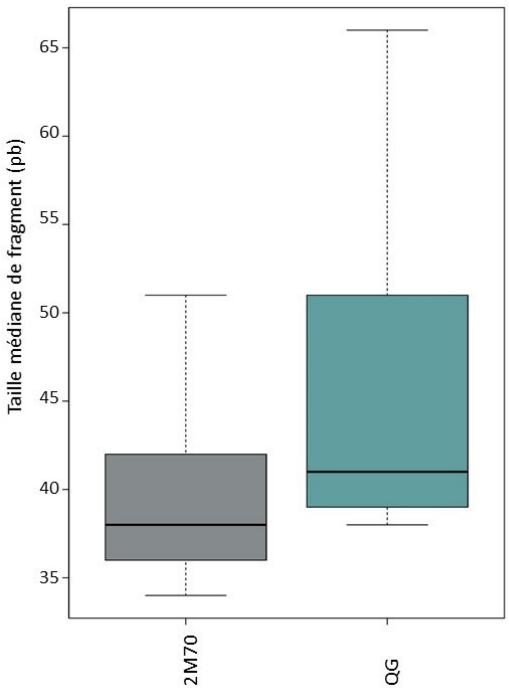


Figure 36: Boîtes à moustaches de la taille médiane des fragments d'ADN obtenus à partir de banques construites à partir d'extrait purifié avec le tampon 2M70 (en gris) et avec le tampon QG (en bleu).

Pour éviter la perte des petits fragments d'ADN observés quand l'extrait est purifié avec le tampon QG, nous avons décidé de changer le tampon QG par le tampon 5M40 en changeant le type de guanidine ainsi que sa concentration. Nous avons alors basculé de de 5,5 M Guanidine isothiocyanate pour le tampon QG à 5M Guanidine hydrochloride.

2) Différence entre le tampon 2M70 et 5M40 sur la distribution de la taille de fragments d'ADN:

La figure 37 représente la taille médiane de chaque banque construite à partir d'extrait d'ADN purifié soit avec le tampon 2M70 soit avec le tampon 5M40. Cette figure a été faite à partir des tailles médianes de 301 banques dont les extraits ont été purifiés avec 2M70 et 271 banques dont les extraits ont été purifiés avec le tampon 5M40.

Comme montré dans le paragraphe précédent, on observe que la taille médiane des séquences d'ADN varie avec le tampon de purification. En effet, le tampon 2M70 permet de récupérer les fragments de plus petite taille qui sont présents à un pourcentage élevé, alors que le tampon 5M40 de purification favorise l'obtention des fragments de plus grande taille. La différence entre les deux méthodes de purification est toutefois moindre que celle observée lorsque nous utilisions le tampon QG (Figure 36). En comparant avec le tampon QG de purification, le tampon 5M40 améliore la récupération de petits fragments d'ADN et il est alors moins dénaturant. Cependant, nous filtrons les séquences de taille inférieure à 28pb lors du traitement bio-informatique car celles ci peuvent s'aligner de façon non spécifique à la séquence de référence. Ce traitement permet donc la perte de petits fragments. La perte est plus remarquable quand l'extrait est purifié avec 2M70 montrant pour une autre fois que 2M70 favorise la récupération de petits fragments dont une proportion est éliminée par nos soins (Figure 38). Nous commençons alors de remettre en question l'utilité de ce tampon (voir plus loin).

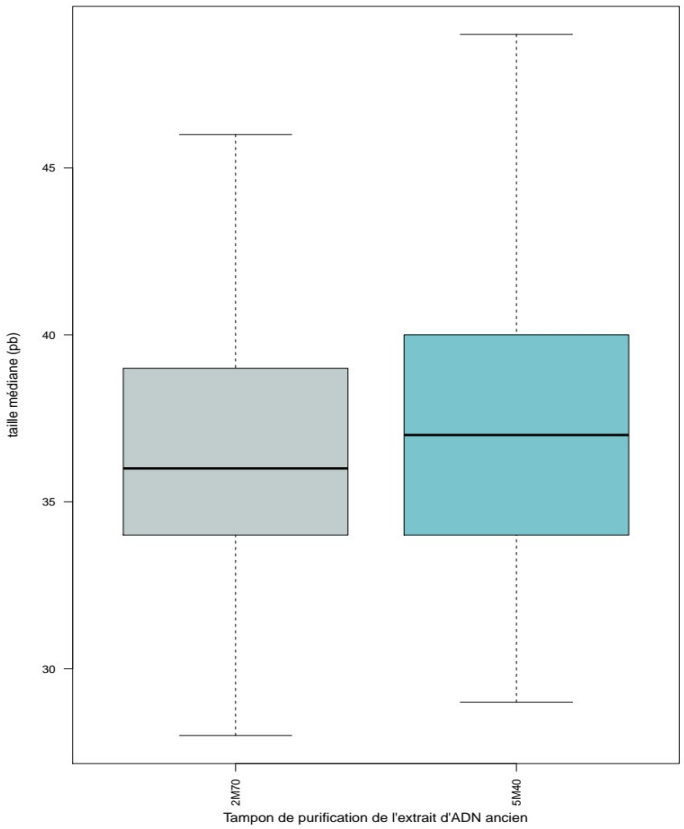


Figure 37: Boîtes à moustaches de la taille médiane des séquences d'ADNa obtenues à partir de banques génomiques construites à partir de fragments d'ADN purifiés soit avec le tampon 2M70 (gris) soit le tampon 5M40 (bleu).

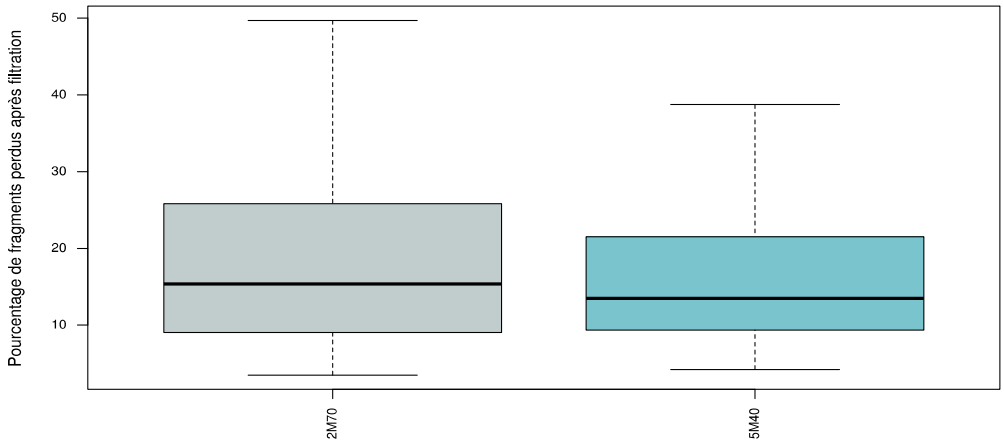


Figure 38 : Boîtes à moustaches de pourcentage d'ADN endogène perdu après la filtration des séquences ayant une taille inférieure à 28pb selon les deux tampons de purification.

Cette constatation nous a poussé à étudier le contenu en GC des séquences obtenues pour chaque échantillon en fonction du tampon de purification avec lequel l'extrait a été purifié pour avoir une meilleure vision sur la qualité des séquences purifiées.

IV) Comparaison de l'effet des tampons de purification de l'extrait d'ADN sur la teneur en GC des séquences obtenues par séquençage de nouvelle génération :

1) Impact des tampons de purification 2M70 et QG sur la teneur en GC:

Après avoir étudié la taille des séquences obtenues, il est important d'étudier le taux de couverture des régions riches en G/C et A/T en analysant le pourcentage en GC des séquences obtenues pour chaque échantillon. En effet, un biais compositionnel de récupération des séquences d'ADN va entraîner un biais de représentation des séquences lues. La fraction d'ADN total constituée principalement d'ADN environnemental est plus riche en GC que l'ADN endogène, ce qui est probablement dû aux microorganismes du sol qui ont colonisé l'os. Nous avons observé pour tous les échantillons fossiles que le pourcentage en GC de la fraction environnementale est plus élevé que celui obtenu avec la fraction alignée sur le génome bovin de référence (données non présentées).

Nous avons observé que le contenu en GC varie en fonction du tampon de purification utilisé . Comme présenté dans la figure 39, correspondant à un seul échantillon qui représente la teneur en GC des fragments en fonction de leur taille, nous avons trouvé que le contenu en GC est plus élevé que le contenu attendu pour le génome bovin moderne (0.42) quand les extraits sont purifiés avec le tampon QG, en particulier pour les fragments de petite taille. Ceci indique que le tampon QG contenant 5.5M de Guanidine isothiocyanate est plus dénaturant et entraîne une perte plus importante des petits fragments riches en AT alors que le tampon 2M70 contenant 2M de Guanidine

hydrochloride permet une meilleure récupération de petits fragments ce qui diminue le risque de perte de l'information génétique.

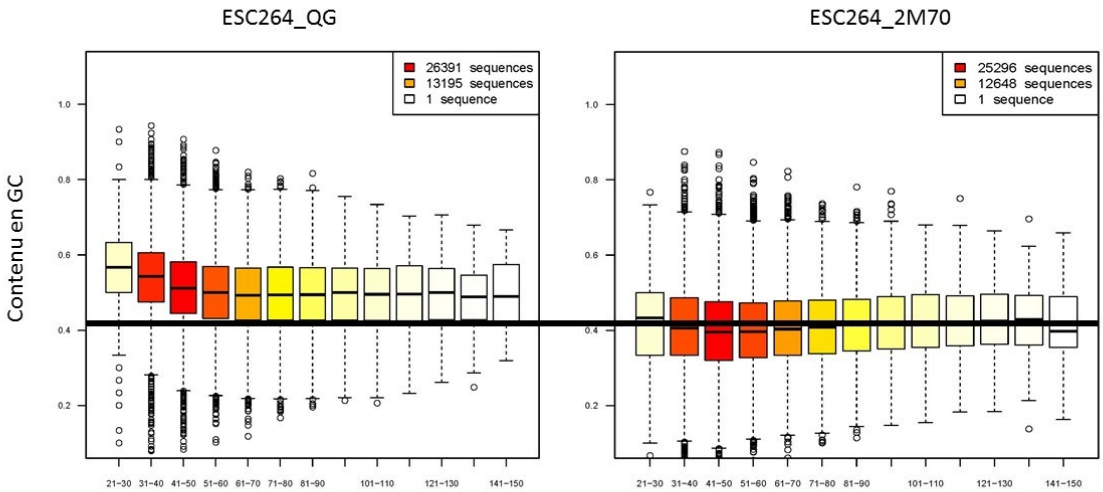


Figure 39: Boîtes à moustache du taux de bases G/C des fragments des banques ESC264_QG, et ESC2642M70. La couleur des boîtes est fonction de leur richesse en séquence. Les boîtes rouges sont celles avec le plus de fragments tandis que les jaunes sont intermédiaires et les blanches sont les plus pauvres en fragments. L'axe des abscisses représente la taille des séquences d'ADNa groupées par incrément de 10 bp et l'axe des ordonnées représente le pourcentage en GC.

2) Impact des tampons de purification 2M70 et 5M40 sur la teneur en GC:

Dans le but de tester l'effet du tampon 5M40 sur l'amélioration de la récupération de petits fragments d'ADN riches en GC, nous avons étudié l'effet du tampon 5M40 sur le contenu en GC des séquences d'ADN produites lors de la phase initiale. Nous avons exploré la différence du contenu en GC et la taille des fragments de l'ADN endogène obtenu avec les deux tampons de purification. En effet, quand nous comparons le taux en GC obtenu avec le tampon 2M70 et 5M40 pour l 'échantillon fossile Coupe Gorge par rapport au contenu en GC attendu qui est de l'ordre de 0.42 pour le génome bovin moderne, nous remarquons qu'avec le tampon 5M40 le contenu en GC est légèrement plus élevé que l'attendu, surtout pour les petits fragments (Figure 40).

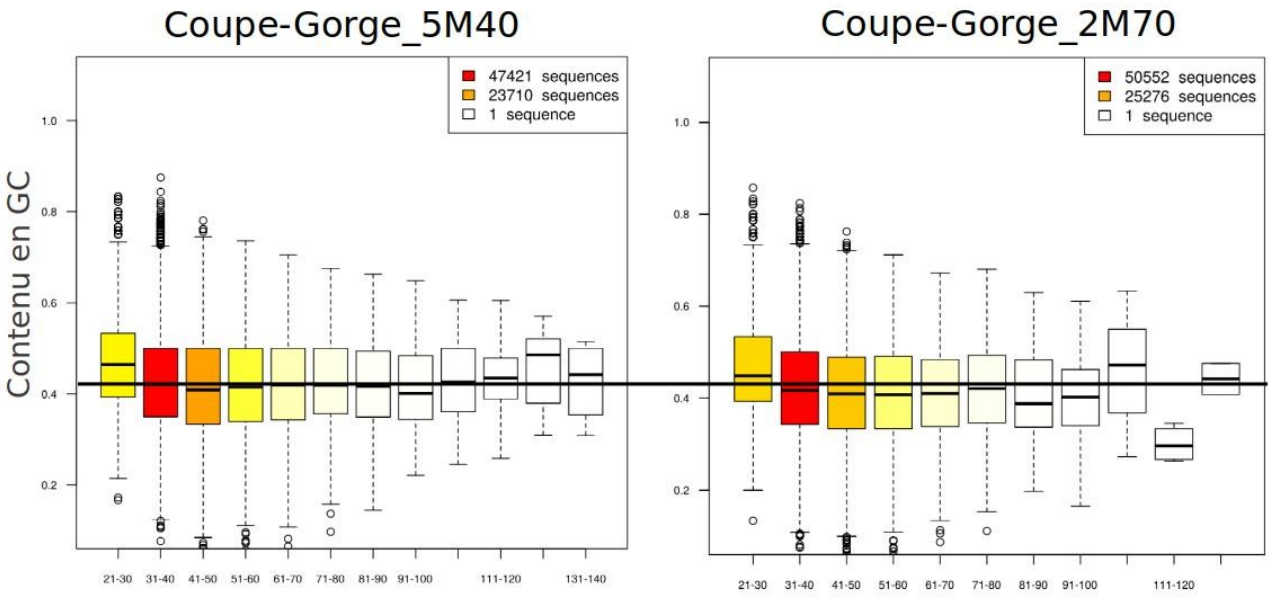


Figure 40: Boîtes à moustache du taux de G/C des fragments des banques Coupe-Gorge_5M40, et Coupe-Gorge_2M70. La couleur des boîtes est fonction de leur richesse en séquence. Les boîtes rouges sont celles avec le plus de fragments tandis que les jaunes sont intermédiaires et les blanches sont les plus pauvres en fragments. L'axe des abscisses représente la taille des séquences d'ADNa groupées par incréments de 10 bp et l'axe des ordonnées représente le pourcentage en GC.

En prenant cet échantillon comme exemple, nous avons observé que le contenu en GC ne diffère que légèrement en fonction du tampon de purification. Cependant, compte tenu de la diversité des échantillons fossiles, plusieurs facteurs peuvent influencer l'efficacité de la purification des extraits et le choix de l'utilisation systématique du d'un tampon est difficile.

Notre étude comparative des caractéristiques des fragments purifiés avec les différents tampons testés, nous a permis de conclure que le type de la guanidine utilisée influence sur la taille des fragments récupérés. En effet, la guanidine isothiocyanate est plus dénaturante que la guanidine hydrochloride et entraine par conséquent la perte de petits fragments qui sont généralement riches en AT. Une molarité de 5 pour la guanidine hydrochloride permet de mieux récupérer les petits fragments perdus et mieux éliminer les inhibiteurs.

V) Effet de traitement enzymatique et bio-informatique :

Les génomes anciens ont des biais particuliers, tels que les petites tailles des fragments d'ADN, la difficulté de couverture des régions riches en AT, et la transformation diagénétique des séquences. Ces caractéristiques provoquent un biais de séquences. Particulièrement, la désamination des cytosines en uraciles induit des transitions C vers T et G vers A sur le brin complémentaire.

Donc, pour minimiser les biais dus à des transformations diagénétiques, surtout la désamination des cytosines, les banques sont construites après traitement à l'uracile DNA glycosylase: l'UNG catalyse l'excision d'uracile laissant un site abasique avec un squelette phosphodiester intact. L'activité lyase de l'endonucléase VIII hydrolyse le desoxyribose de part et d'autres du site abasique de sorte que le brin d'ADN est rompue.

1) Effet de l’enzyme USER sur la taille des fragments d’ADN séquencés :

Certains des extraits d’ADNa ont été traités en présence et en absence de l’enzyme USER (Uracil DNA glycosylase (UDG) et l'ADN glycosylase-lyase (endonucléase VIII)) afin d’étudier l’effet de l’enzyme sur la taille des fragments obtenus et la relation entre l’âge des échantillons anciens et le nombre des cytosines désaminés contenues dans les séquences d’ADN. Différents échantillons ont été choisis en raison de leurs différences d’âge, entre 130 000 ans et 5000 ans, et du nombre des séquences alignées sur le génome mitochondrial de référence. Les résultats sont représentés dans la figure 41.

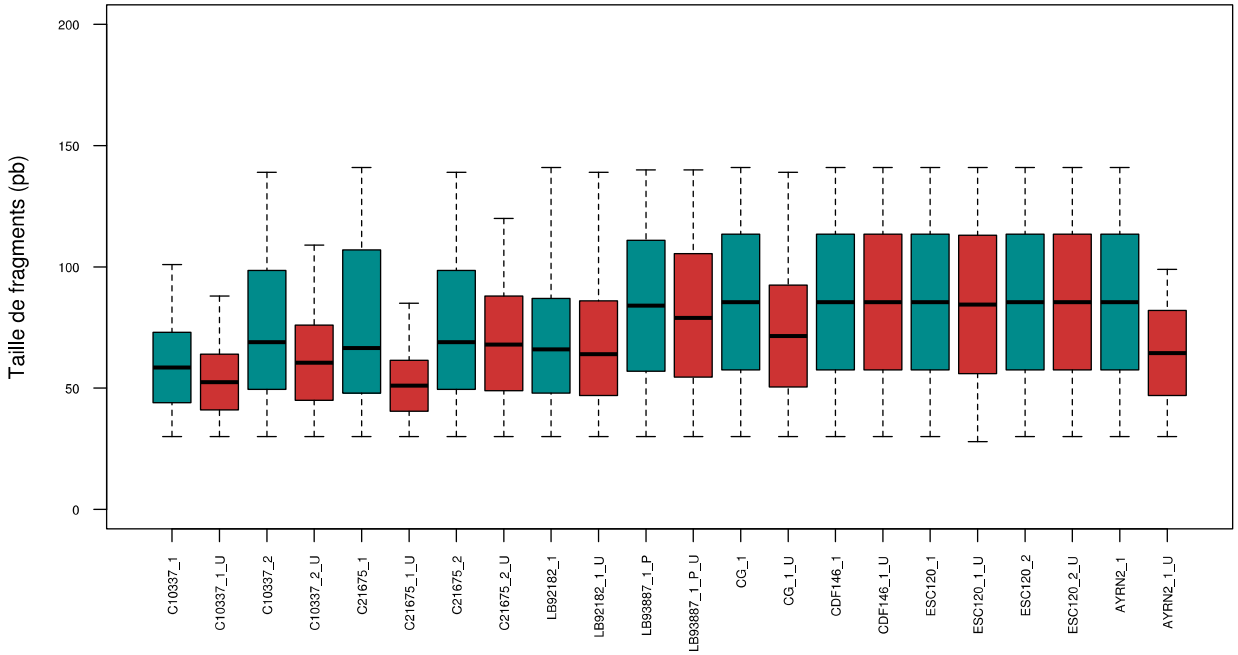


Figure 41: Effet de l’enzyme USER sur la distribution de taille des fragments d’ADN. Les couleurs vert et rouge représentent respectivement la taille médiane des séquences obtenues sans et avec traitement à l’enzyme USER.

Les fragments d’ADN de taille inférieur à 28pb ne sont pas présentés dans cette figure puisqu’on les a enlevés pour pouvoir réaliser l’alignement. Une réduction de taille en présence de l’enzyme USER a été observée pour tous les échantillons anciens de différents âges ce qui montre que l’enzyme USER permet d’enlever les cytosines désaminées transformées en uraciles et couper le fragment en plusieurs morceaux. Par contre, cette réduction varie selon l’échantillon. L’échantillon C21675_1 présente un profil de distribution de taille intéressant car la réduction de la taille des séquences d’ADN en présence d’USER est très significative. Pour les échantillons les plus vieux d’âge estimé de 130 000 ans, la réduction de taille est plus significative que pour les échantillons de 5 000 ans.

Pour l'échantillon AYRN2 qui est plus récent (5000 ans) et qui présent des fragments d'ADN de taille importante, on remarque que l'enzyme USER a un grand effet car la taille moyenne des fragments d'ADN a diminué d'une manière remarquable. Elle reste toutefois plus importante que pour les échantillons plus anciens. L'effet de l'enzyme USER dépend de plusieurs paramètres : La taille des fragments d'ADN, le nombre des cytosines désaminées, la position des cytosines et l'âge de l'échantillon.

En effet, quand le fragment d'ADN est de grande taille, la probabilité de la présence de cytosine désaminée est plus élevée que pour un fragment de petite taille donc la différence de distribution de taille sera plus remarquable. De plus, si le résidu de cytosine désaminée se trouve proche de l'extrémité de fragments d'ADN, ce bout de séquence de taille plus petite va être perdu lors de l'alignement qui ne prend pas en considération les fragments de taille inférieur à 28pb. Ainsi, une corrélation entre la diminution de la distribution de taille et l'âge de l'échantillon a été observée comme déjà montré par (Sawyer et al., 2012).

2) Effet de l'enzyme USER sur la qualité des séquences obtenues :

En plus de la réduction de la taille des fragments d'ADN obtenues en traitant les extraits d'ADNa avec USER, ce traitement permet de corriger les dommages et les incertitudes obtenus quand les extraits ne sont pas traités avec cette enzyme. Ces dommages correspondent aux transversions C vers T et autres comme l'oxydation. Ceci est visible après capture lorsque l'on analyse les séquences alignées sur le mitogénome. Pour mettre en valeur ce phénomène, je présenterai de manière anticipée une partie de ces résultats (Figure 42).

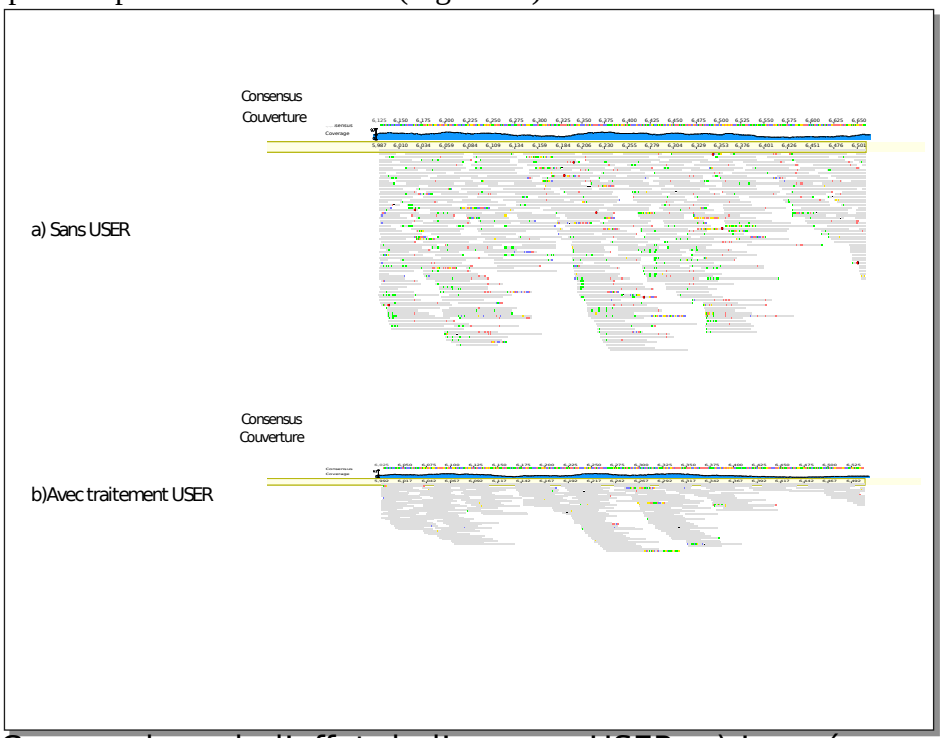


Figure 42: Comparaison de l'effet de l'enzyme USER. a) Les séquences d'ADNa obtenues sans traitement USER. b) Les séquences d'ADNa obtenues en présence de l'enzyme USER. Les erreurs de séquence sont visibles sous la forme de traits de couleur.

L’alignement des séquences ayant été réalisé en mode local (Li & Durbin, 2009), des molécules chimériques issues de deux fragments dont un seul correspond au génome mitochondrial, peuvent être identifiées. Ici, elles apparaîtront comme des séquences à une des deux extrémités des fragments contenant une forte proportion de mésappariements à la séquence de référence.

Cet effet de l’enzyme USER joue un rôle important sur l’efficacité de l’alignement de séquences car USER diminue le taux de mésappariement et augmente alors l’efficacité et la certitude de l’alignement des séquences d’ADN sur le génome de référence. Bien que l’enzyme USER entraîne la diminution de la taille des fragments et la perte des fragments qui deviennent plus courts que 28pb et qui sont donc éliminés par notre script d’alignement de séquences, la cartographie est devenue meilleures surtout quand les séquences sont peu divergentes.

3) Profil de dommages de l’ADN ancien obtenu avec le logiciel MapDamage :

Pour analyser la distribution des mutations induites par la désmination (C->T, ou G->A sur l’autre brin), on utilise un programme informatique appelé mapDamage. Il permet de caractériser, par correspondance avec la séquence de référence, le patron de dommage des séquences obtenues et la localisation de ces dommages sur les fragments séquencés. Le programme va principalement mettre en évidence et quantifier les mutations artéfactuelles induites par la désamination des cytosines qui s’accumule surtout aux extrémités des séquences car la désamination se produit beaucoup plus sur l’ADN simple brin que l’on trouve aux extrémités. Ceci permet d’attester du caractère ancien des séquences. La désamination se produit au cours du temps et en fonction de la température de préservation de l’ossement. Le programme permet de mettre en évidence aussi l’effet du traitement User (Figure 43).

Le traitement avec l’enzyme USER permet alors de réduire le taux des transitions C vers T sur un brin et G vers A sur l’autre brin. Cependant, il peut rester, pour quelques échantillons, une bonne partie des mutations C->T et G->A, principalement sur la dernière base car l’enzyme a du mal à agir sur la base à l’extrémité du fragment d’ADNa, son interaction avec la matrice étant moins stable à l'extrémité. L’efficacité de l’enzyme User peut être réduite à cause des inhibiteurs d’enzyme contenus dans les extraits fossiles. De plus, les fragments d’ADN accrochés sur les parois du tube peuvent échapper à l'action de l’enzyme mais être libérés pour participer aux réactions suivantes.
Pour enlever les cytosines désaminées ayant échappé au traitement User et pour minimiser le taux d’erreur et le taux de faux SNPs appelés, on ajoute une étape de traitement bioinformatique. Cette étape consiste à une réévaluation de la qualité des bases en se basant sur la distribution particulière, enrichis aux extrémités des transitions C->T (en 5’) et G->A (en 3’). Le nombre de base enlevée dans l’étape de réévaluation dépend de l’échantillon, de ses conditions de préservation et du taux d'inhibiteurs dans l’extrait. Les traitements enzymatique et bio-informatique en combinaison permettent de réduire le taux d’erreurs et de minimiser alors les biais dans les séquences d’ADNa (Figure 43).

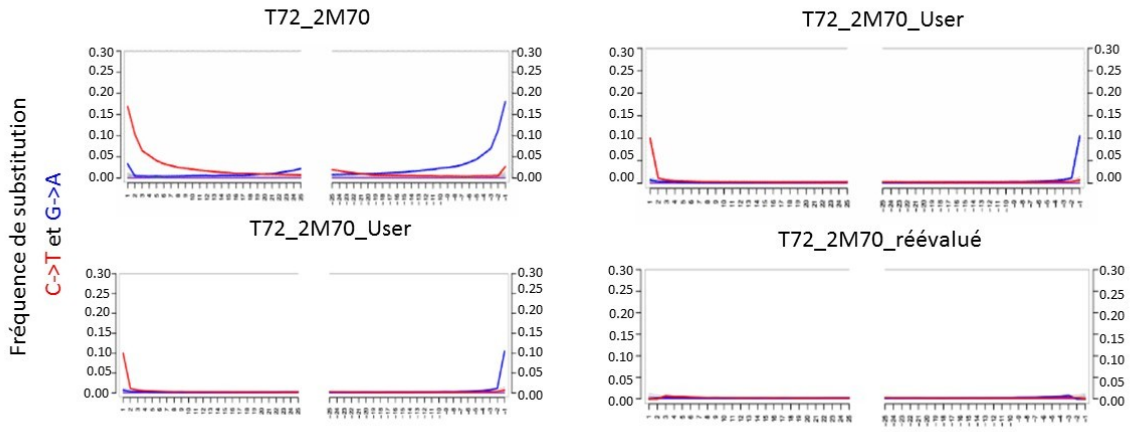


Figure 43 : Profil MapDamage des séquences alignées au génome bovin de banques d'ADN construites à partir d'ADN extrait du reste fossile T72 non traité (à gauche en haut) et traité (à droite en haut) avec l'enzyme USER. Les deux graphes du bas montrent l'effet de la réévaluation de cytosines désaminés.

VI) Conclusion :

Ces explorations primordiales des données obtenues lors de la phase initiale nous ont permis de choisir les tampons de purification à utiliser. En effet, nous avons décidé de ne plus utiliser le tampon QG parce que nous perdons les petits fragments riches en AT ce qui influence la couverture du génome. Nous avons alors remplacer ce tampon par le tampon 5M40 qui est moins dénaturant que le tampon QG et qui permet de récupérer les fragments de petite taille riches en AT. Nous avons observé dans certains cas que le tampon 2M70 n'élimine pas efficacement les inhibiteurs. Pour ces différentes raisons, chaque extrait d'ADN fossile était purifié avec les deux tampons de purification 2M70 et 5M40 pour permettre la récupération de fragments de tailles différentes et de maximiser la récupération de toutes les molécules uniques qui existent dans l'extrait fossile.

Les étapes de purifications des extraits fossiles visent à enlever les protéines et toutes substances chimiques contenues dans l'extrait, mais à l'issue de la purification, l'ADN endogène récupéré est dilué dans l'ADN environnemental ce qui nécessite de faire varier les tampons de purification de l'ADN pour pouvoir obtenir des quantités d'ADN suffisantes pour réaliser l'analyse paléogénomique. Ces résultats révèlent la difficulté de l'étude de l'ADNa qui, en plus de sa fragmentation en petits morceaux la plupart du temps inférieur à 50pb, est le plus souvent contaminé avec de l'ADN environnemental.

Pour maximiser la récupération de toutes les molécules uniques qui existent dans l'extrait fossile, notre choix s'est orienté vers la purification de chaque extrait d'ADN fossile par deux tampons de purification contenant 2M ou 5M de Gunidine Hydrochloride. Ce choix a été fructueux parce que nous avons réussi à couvrir au mieux les génomes des échantillons dont le niveau de préservation l'ADN a été variable et à avoir une couverture suffisamment bonne pour inclure des échantillons originaires des régions chaudes dont l'ADN n'était pas bien préservé dans l'analyse.

Nous avons trouvé que les deux tampons (2M70 et 5M40) sont complémentaires pour les échnatillons les plus vieux dont l'ADN est moins bien préservé, pour récupérer aussi bien les petits et les grands fragments et de réduire le biais de représentation des régions riches en A/T, un problème soulevé lors de la purification avec le tampon QG (Protocole standard de purification des banques d'ADN de Qiagen, 5.5 M guanidine thiocyanate, 20 mM Tris HCl pH 6.6 et 25% d'isopropanol).

En effet, l'agent chaotropique utilisé dans le protocole de purification sur colonne de silice dénature l'ADN. Plus le fragment est petit et riche en AT plus il est dénaturé. La guanidine isothiocyanate est un agent cahotropique plus fort que la guanidine hydrochloride. C'est pour cela que nous perdons plus de petits fragments avec le tampon QG, que 5M40 et 2M70. De plus, une concentration plus élevée du même agent cahotropique entraîne une dénaturation plus importante des petits fragments riches en AT et donc leur moins bonne rétention sur la matrice en silice des colonnes.

Chapitre II: Optimisation des conditions de capture par hybridation du génome mitochondrial

I) Comparaison de l’effet de la diminution de la stringence :

Au fur et à mesure de la production des résultats, nous avons optimisé la capture des mitogénomes par hybridation avec des sondes d’ARN. Dans cette partie, je vais vous présenter une analyse comparative des résultats obtenus lors de ces optimisations.

1) Différence de distribution de taille des séquences d’ADN et du contenu en GC après Shotgun et Capture:

Pour étudier l’efficacité de la capture, nous avons déterminé la distribution de taille de fragments d’ADN capturés et les comparer avec ceux obtenus avec le séquençage shotgun selon la stringence utilisée pour les lavages. Après capture et dans le cas de l’utilisation de stringence élevée de lavage (1x SSPE puis 0.1x SSPE), les séquences obtenues alignées sur le génome mitochondrial ont été comparées aux séquences obtenues avant capture qui s’alignaient sur le génome total. On a d’abord comparé la distribution de taille des fragments d’ADN générés par le séquençage Shotgun pour les échantillons bien conservés de la France, et la taille des fragments d’ADN générés après la capture par hybridation des séquences mitochondriales (Figure 44). On remarque qu’après capture des séquences mitochondriales, la taille moyenne des séquences d’ADN augmente pour tous les échantillons anciens purifiés avec les deux tampons de purification.

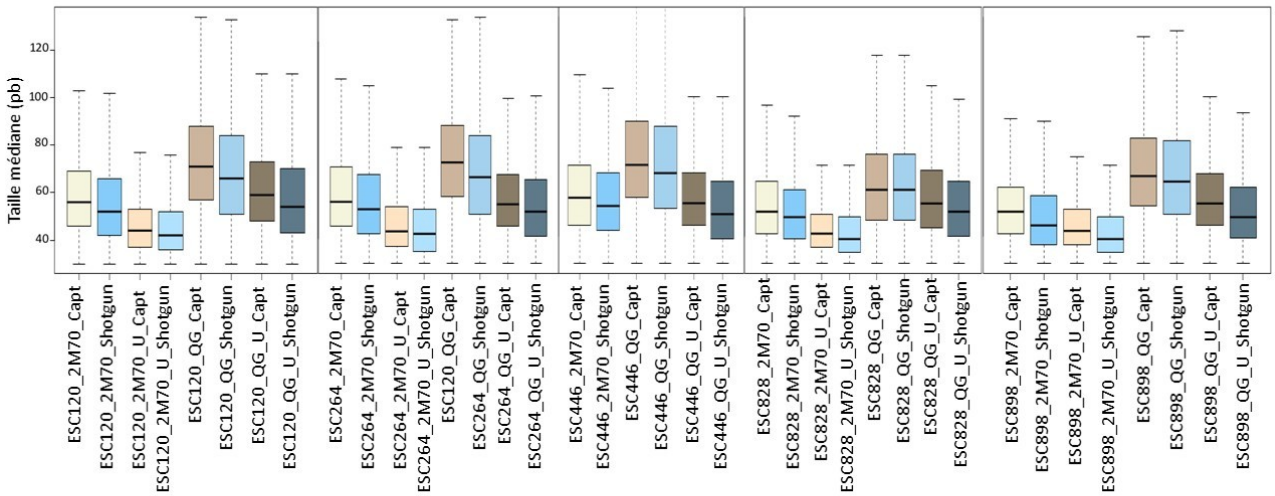


Figure 44 : Distribution de la taille des fragments d’ADN obtenus après séquençage Shotgun et capture. Les deux couleurs déterminent la taille des séquences obtenues avant et après capture. Les nuances de couleur beige correspondent aux échantillons capturés avec les différents paramètres expérimentaux et les nuances de couleur bleue correspondent aux données de séquençage shotgun.

Cette observation peut être expliquée par la perte des fragments d'ADN de petite taille lors de la capture à cause des paramètres stringents de lavage. Les résultats obtenus présentés dans la figure suivante (Figure 45) montrent que les petits fragments riches en AT purifiés avec le tampon 2M70 sont perdus lors de la capture et donc le pourcentage en GC des petits fragments augmente. Les conditions de capture utilisées au paravent nous ont fait donc perdre une partie de l'avantage que conférait l'utilisation du tampon 2M70.

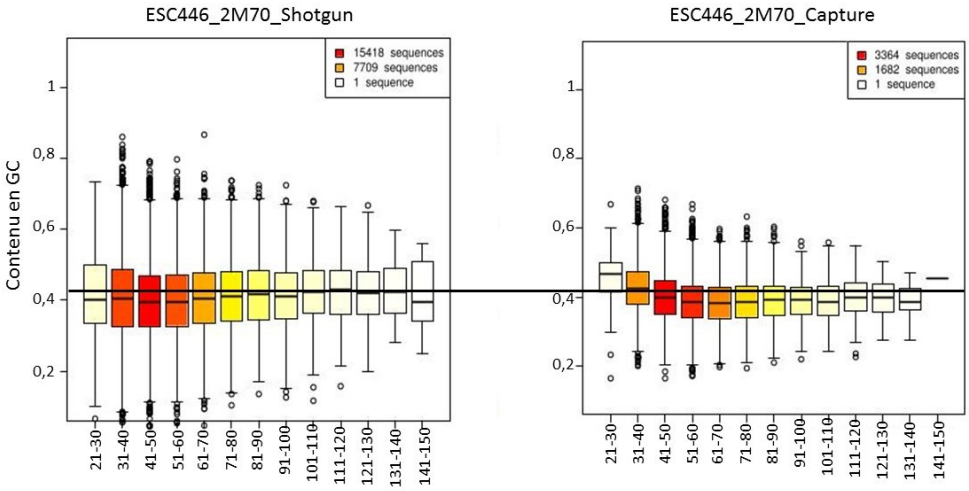


Figure 45: Différence de distribution de taille des séquences d'ADN et du pourcentage en GC entre le séquençage Shotgun et la capture pour l'échantillon ESC446 purifié avec le tampon 2M70.

2) Couverture des régions riches en GC après capture :

Suite à cette observation, on a analysé la couverture du mitogénome et étudié la corrélation avec la richesse en GC. La figure 46 montre la comparaison des couvertures du mitogénome obtenues après capture en utilisant les banques construites avec et sans traitement USER pour deux méthodes de purification (2M70 et QG) pour un échantillon représentatif. On remarque que la couverture du mitogénome est toujours hétérogène et semble suivre les variations du taux de GC du mitogénome. Les régions riches en AT sont les moins bien couvertes mais elles sont toutefois mieux couvertes avec le tampon de purification 2M70 qu'avec le tampon QG. On remarque aussi que le traitement de l'enzyme USER accentue le biais de couverture ce qui résulte du fait que ce biais est plus important lorsque les fragments d'ADN sont de petite taille, comme on le voit aussi lorsque l'on compare différents échantillons entre eux (données non montrées).

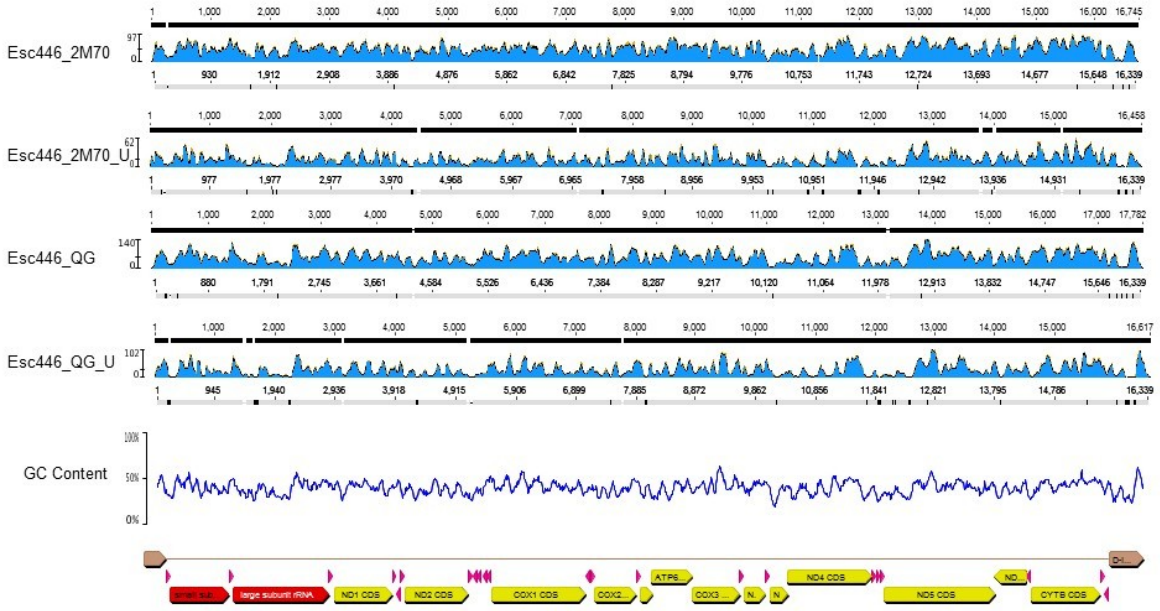


Figure 46: Comparaison du degré de couverture du génome mitochondrial après capture pour l'échantillon ancien ESC446 purifiés avec les deux tampons 2M70 et QG et traités ou non avec l'enzyme USER lors de la construction des banques.

Afin de s'assurer que l'hétérogénéité de couverture était bien due à des biais compositionnels, nous avons découpé le génome mitochondrial en fenêtres glissantes de 100 paires de base toutes les 50 paires de bases et mesuré pour chaque fenêtre le taux de couverture et le pourcentage en AT. Le graphe de densité en résultant (Figure 47) montre clairement que plus la séquence est riche en AT, moins bien elle est couverte, et ceci quelles que soient les conditions expérimentales lors de la purification de l'ADN ou la construction des banques comme le montre la pente négative de la droite de régression.

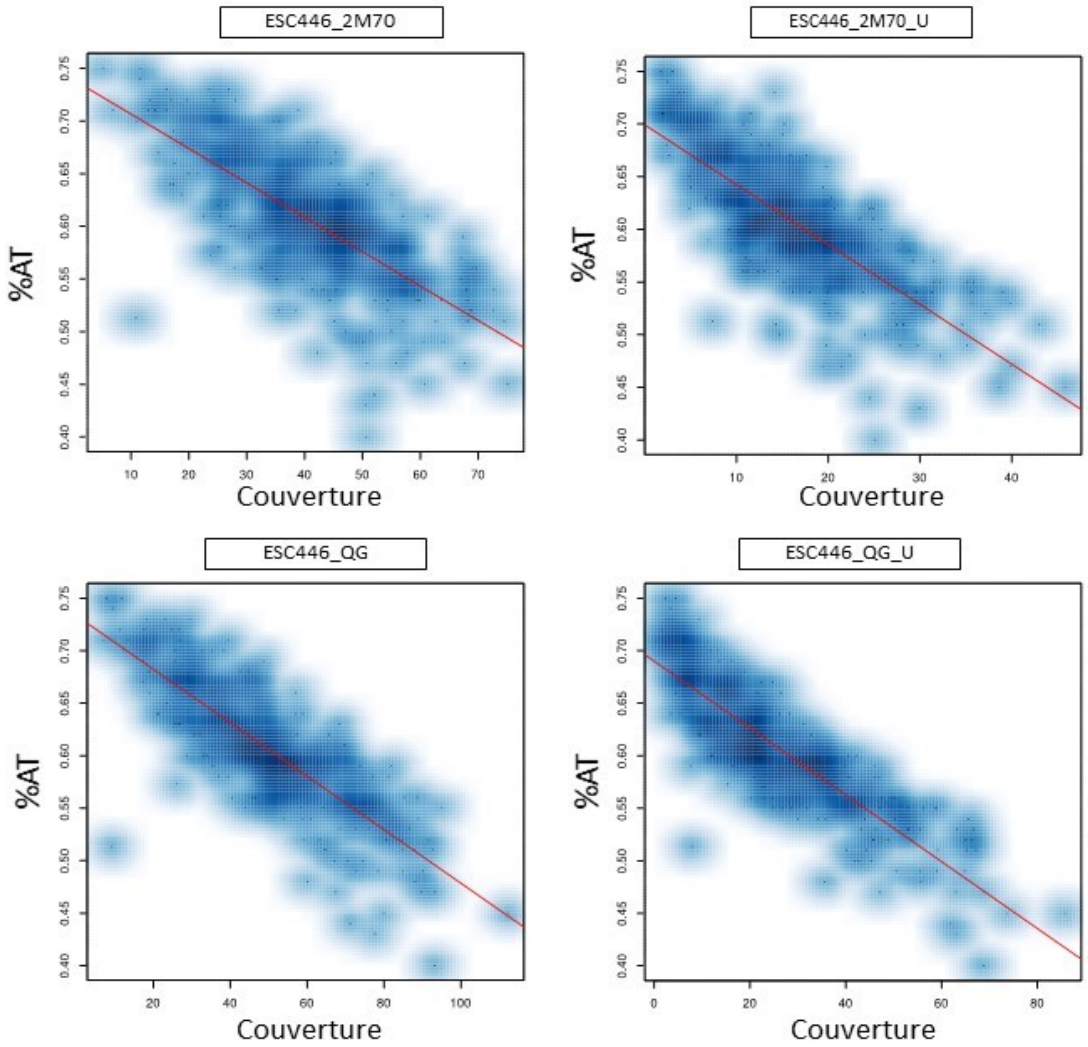


Figure 47 : Comparaison de la corrélation entre le degré de couverture et la richesse en AT après capture pour l'échantillon ancien ESC446 dont l'extrait a éré purifié avec les deux tampons 2M70 et QG et traité en présence ou en absence de l'enzyme USER. L'axe des abscisses représente la taille des fragments d'ADN et l'axe des ordonnées représentent la richesse en AT.

Pour les échantillons purifiés avec le tampon 2M70 et traités avec USER, on remarque que la couverture des régions riche en AT diminue et que le nuage des points se déplace vers la gauche par rapport à l'absence de traitement USER ce qui montre bien que la réduction de taille des fragments se traduit par une plus mauvaise représentation des séquences riches en AT. On observe le même phénomène pour les échantillons purifiés avec le tampon QG. Lorsque l'on compare les échantillons correspondants (avec ou sans USER) selon que la purification a été effectuée avec 2M70 ou QG, on voit aussi un déplacement vers la gauche des points pour la purification avec QG ce qui montre bien que la perte initiale des fragments courts riches en AT accentue le problème de couverture des régions génomiques correspondantes.

En conclusion, la capture de séquence dans les conditions de stringence utilisées cause une sous-représentation des séquences riches en AT et ce d'autant plus que les fragments d'ADN sont courts. Alors que le biais de cette représentation avait été pratiquement éliminé en utilisant le tampon 2M70 plutôt que le QG lors de la purification, cet avantage a été en partie perdu (mais pas totalement) lors de la capture. C'est pour cela était donc nécessaire de modifier la stringence des conditions de capture pour explorer la possibilité de minimiser l'importance de ce biais afin d'obtenir une couverture plus homogène des génomes anciens.

3) Effet de la diminution de la stringence des lavages sur l'efficacité de la capture des mitogénomes :

Pour mettre en évidence l'effet de la diminution de la stringence lors de l'étape de lavage de la capture par hybridation, j'ai déterminé pour les mêmes banques ayant été capturées par les deux conditions de capture le contenu en GC (Figure 48). La première condition de lavage correspond à l'utilisation de 2 tampons de lavage de concentration 1X SSPE et 0.1X SSPE. Pour la deuxième condition de lavage, nous avons utilisé un tampon de lavage de 1X SSPE puis un deuxième lavage de 0.2X SSPE. Nous remarquons qu'après capture avec la deuxième condition, le contenu en GC diminue ce qui montre que nous récupérons les fragments riches en AT qui ont été perdus avec la première condition.

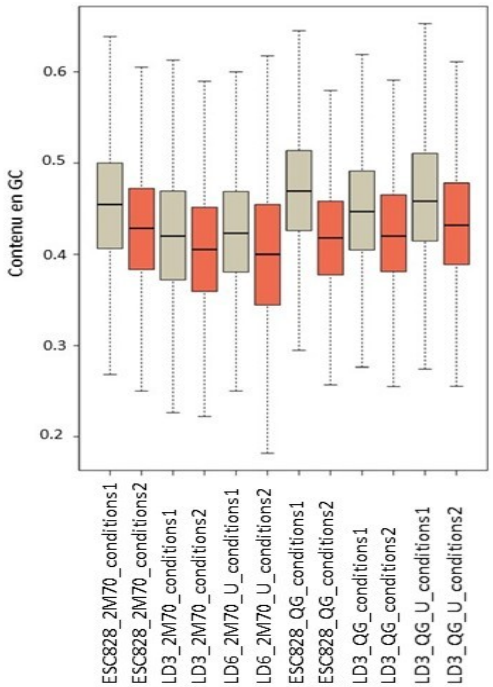


Figure 48 : Boîtes à moustaches du contenu en GC pour chaque échantillon dont les fragments d'ADN des banques génomiques ont été capturés selon les deux conditions de capture.

Bien que la deuxième condition de capture semble meilleure que la première, son efficacité semble dépendre de la divergence entre les séquences ciblées et les séquences des sondes d'ARNs et de la taille des fragments d'ADN de départ contenus dans les extraits fossiles (Figure 49).

En effet, en capturant les séquences mitochondriales de bisons et des aurochs, les plus anciens, la taille médiane des fragments capturés augmente ce qui pourrait s'expliquer par la perte des petits fragments qui sont probablement les fragments les plus divergents (Coud10337, Coud21675, Meng1, Meng21, Meng49, NK71, LP4, LP1874) (Figure 49). De plus, quand les fragments de départ sont de petite taille (exemple NK71 et KerT70), la capture semblerait moins efficace et elle ne récupère que les grands fragments.

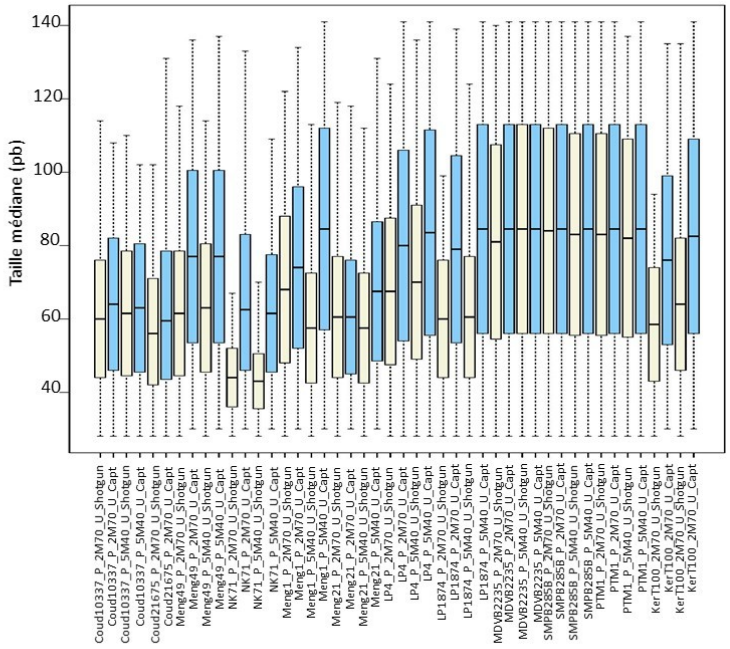


Figure 49 : Distribution de taille de fragments obtenus après séquençage aléatoire et après capture des échantillons bisons, aurochs et de bœufs domestiques.

II) Conclusion :

Nos optimisations du protocole de capture de mitogénomes nous ont permis d'améliorer la couverture des mitogénomes et ceci en augmentant la représentation des fragments riches en AT qui sont généralement de petite taille. Ces optimisations méthodologiques ont réduit le coût de séquençage en profondeur. Le problème de la capture des séquences mitochondriales de bisons et d'aurochs anciens dû à leurs divergences est résolu avec le séquençage en profondeur des banques génomiques qui a permis d'améliorer leurs couvertures et de boucher les trous sur le génome mitochondrial et plus précisément sur la région hypervariable.

Chapitre III: Stratégie d’appel du polymorphisme nucléotidique ou SNPs

I) Introduction : Appel du polymorphisme :

L’hérédité uniparentale de l’ADN mitochondrial limite son utilisation dans la phylogénie seulement et ne permet pas d’effectuer une étude comparative entre les aurochs et les bovins domestiques car il ne représente pas l’ensemble de la diversité génétique bovine. Pour cela, il est nécessaire d’avoir recours aux différents flux génétique qui ont eu lieu entre les bœufs domestiques originaires d’Asie du sud-ouest et les populations d’aurochs et il est nécessaire, alors, de compléter les études phylogénétiques basées sur les génomes mitochondriaux par des études basées sur les autosomes ou le chromosome Y.

L’obtention de séquences génomiques nucléaires et la détection des SNPs ancestraux d’aurochs permettent d’explorer les questions concernant les flux génétiques, la sélection naturelle et artificielle ainsi que les introgressions et les hybridations entre les populations locales d’aurochs et les bœufs domestiques. L’extinction de l’aurochs au 17ème siècle fait que la réponse à ces questions nécessitent une approche paléogénomique. Grâce au séquençage de nouvelle génération (NGS), la première séquence génomique d’un aurochs britannique CPC98 âgé d’environ 6740 a été publiée par (Park et al., 2015). Leurs résultats étaient en faveur d’une origine Proche-Orientale des bœufs européens car les analyses effectués sur le génome de CPC98, placent cet aurochs sur un groupe phylogénétique des aurochs européens qui est distinct de celui des bœufs domestiques européens. Avec ce génome, les chercheurs de cette étude ont pu mettre l’accent sur la sélection de gènes impliqués dans le système immunitaires et dans le métabolisme chez les bœufs domestiques.
Ce génome nous a permis d’avoir une idée sur les relations entre aurochs et bœufs domestiques et d’entamer l’étude sur les chemins évolutifs des populations bovines depuis la domestication il y a 10000 ans mais aussi de comprendre la diversité génétique des populations d’aurochs au Pléistocène en Europe. Pour répondre à ces questions, un axe de ma thèse porte sur l’optimisation des méthodes d’options de SNPs pour avoir un jeu de données robustes et fiables permettant l’obtention de résultats robustes.

L'appel de SNPs (Single Nucleotide polymorphism) vise à déterminer dans quelles positions il y a des polymorphismes ou dans quelles positions au moins une des bases diffère d'une séquence de référence. La manière d’appeler les SNPs est un critère important pour obtenir des SNPs avec une bonne certitude. L’appel de SNPs est toujours accompagné d'une estimation de la fréquence des variants et d'une certaine mesure de confiance. L’appel de SNP avec le Joint Genotyping permet d’appeler les variants d’ADN sur un ensemble d’échantillons à l’aide de l’outil haplotypeCaller du l’algorithme d’appel de SNPs GATK ce qui permet de créer un fichier gvcf (Genomic Variant Call Format) ou toutes les positions sont appelées et gardées.

Le génotypage joint s’effectue en deux temps: Tout d’abord, les variants sont appelés individuellement sur chaque échantillon, générant un fichier gVCF par échantillon qui répertorie les probabilités de génotype et leurs annotations. Toutes les positions sont appelées de façon à distinguer les positions identiques au génome de référence des positions manquantes. Par la suite, les variants des fichiers gVCF sont combinés entre eux et appelés simultanément de façon à génotyper pour chaque individu les SNPs pour lesquels au moins un des individus est différent du génome de référence et réévaluer la probabilité associée à chaque SNP identifié en fonction des SNPs lus à la même position sur tous les autres échantillons.

Pour effectuer le génotypage joint sur l’ensemble des génomes modernes et anciens publiés que nous avons téléchargés et nos échantillons anciens, nous avons réalisé une étude sur un génome produit dans le laboratoire, celui d’un aurochs anatolien âgé d’environ 9700 calBP (7838-7599 calBCE , 9675 calBP) et couvert 7 fois pour identifier et mettre en place toutes les conditions optimales pour l’obtention d’un jeu de données de SNPs robustes, réellement positifs.

II) Aurochs anatolien et appel de SNPs :

1) Présentation de l’aurochs anatolien: AS5

Un échantillon intéressant est AS5 car il s’agit d’un aurochs âgé de 9700 calBP. Il provient d’Anatolie, et plus précisément du site archéologique Asikli Höyük, proche du centre de domestication des aurochs (Figure 50). Les ossements bovins de cette période à cet endroit sont considérés correspondre aux formes sauvages de la population qui a été domestiquée. Les données d’AS5 nous aideront à analyser l'origine des bœufs domestiques en Europe et à étudier l'impact de la domestication sur l'évolution du génome des bovins.

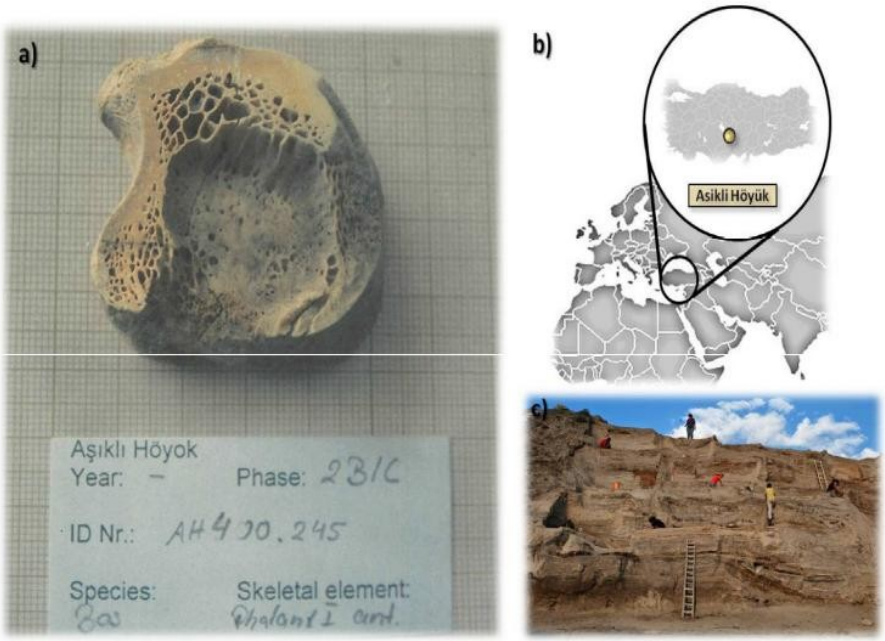


Figure 50: L'échantillon AS5. a) Photo du reste fossile AS5 (phalange d'aurochs). b) Localisation du site Asikli Höyük. c) Photo de la coupe stratigraphique du site Asikli Höyük (Thèse de Diyendo Massilani).

2) Détermination du sexe de l’aurochs anatolien AS5 :

Le caryotype des bovins est constitué de 29 paires d’autosomes et d’une paire de chromosomes sexuels qui sont ’XX’ pour les femelles et ‘XY’ pour les mâles. Le génome de référence qu’on a utilisé pour l’alignement des séquences contient les séquences qui correspondent au chromosomes X et Y. Le chromosome Y contient beaucoup plus de séquences répétées que les autres chromosomes. Lorsque l'on filtre les lectures pour ne garder que les lectures cartographiées de manière unique et que l'on les normalise sur la taille totale des chromosomes, on obtient des valeurs de lectures normalisées très proches pour chaque autosome. Si on séquence le génome d'une femelle on a aussi un nombre de lectures normalisé par la taille similaire, c'est à dire que le rapport du nombre de lectures normalisé X/A (A= valeur moyenne de tous les autosomes) est de 1. La valeur est très faible pour le chromosome Y dans ce cas bien qu'elle ne soit pas nulle car le génome de référence ne rend pas compte parfaitement de la réalité complexe du chromosome Y. Si l'on séquence le génome d'un mâle, le rapport X/A tombe à 0.5 et l'on obtient un nombre significatif de lectures sur le chromosome Y. Le rapport Y/A est de l'ordre de 0.2 lorsque l'on ne considère que les lectures cartographiées à une seule position du fait du fort taux de séquences répétées sur le chromosome Y. Pour AS5, nous avons observé qu’il y a moitié moins de lectures qui s’alignent sur le chromosome X que sur les autosomes et que le ratio Y/A est de 0.51 ce qui montrent qu’AS5 est un aurochs mâle.

III) Appel de SNPs :

1) Mutation génétique :

La variation génétique est la différence dans les séquences d'ADN entre les individus d'une population. Les mutations sont la source originale de cette variation. Une mutation est une altération permanente d'une séquence d'ADN.

En plus des mutations spontanées, des mutations dites « de novo » peuvent se produire lorsqu'il y a une erreur lors de la réplication de l'ADN qui n'est pas corrigée par les enzymes de réparation de l'ADN. Ce n’est que lorsque l’erreur est copiée par réplication de l’ADN et fixée dans l’ADN qu’elle est considérée comme une mutation. Les mutations spontanées peuvent être bénéfiques pour l'organisme, nocif ou neutre c’est-à-dire qu’elle n'a aucun effet sur l'aptitude de l'organisme. Ces mutations permettent d’étudier la variation au sein de populations d’individus dans le temps et dans l’espace. Des logiciels informatiques sont capables aujourd’hui de détecter ces variations génétiques avec une précision qui diffère d’un algorithme d’appel de variant à un autre.

2) Définition d’appel de variant :

L’appel de variant ou appel de polymorphisme est une approche par laquelle on identifie les variants contenus dans les données génomiques. Cette approche est divisée en deux classes: La première classe est «Indel calling» ou «Appel des insertions-deletions» qui est la classe des algorithmes qui permettent d’identifier les insertions et les délétions contenues dans les séquences d’ADN. La deuxième classe est la classe des «SNP calling» ou «Appel de SNPs» qui appellent les variants nucléotidiques contenus dans les séquences d’ADN.

Plusieurs raisons peuvent causer une variation nucléotidique: (1) un vrai Single Nucleotide Polymorphism (SNP) (2) Une erreur produite durant la préparation des banques d'ADN (3) une modification diagenétique (4) une erreur de séquençage (5) une erreur de «base calling » ou d'appel de base. Ce dernier type d'erreur peut être réduit par l'utilisation de méthodes plus performantes mais le risque d'erreur n'est jamais éliminé (discuté par la suite) (6) une erreur lors de l'alignement de séquences (7) une erreur dans la séquence de génome de référence. Cependant, dans le cas des études basées sur l'analyse de l'ADNa, un type d'erreur diagénétique peut se produire si on ne traite pas les extraits d'ADNa avec l'UNG (Uracile N-Glycosylase) qui enlève les cytosines désaminées.

3) Appel du polymorphisme du génome de l'aurochs anatolien :

Pour étudier l'origine génétique des bovins domestiques en Europe, une analyse de détection des SNPs ancestraux de l'échantillon aurochs AS5 a été réalisée. Le génome d'AS5 est couvert 7 fois ce qui limite et complique son analyse (bien que cette couverture reste bonne en comparant avec d'autres échantillons fossiles). A cause des lectures courtes et de la mauvaise couverture de l'ADNa il est nécessaire de tester différents algorithmes d'appel de SNPs pour éviter d'appeler les faux positifs et pour détecter le plus possible de SNPs réellement positifs. L'identification précise des variants génomiques est donc un facteur déterminant du succès des études évolutives.

Trois algorithmes ont été utilisés pour détecter le plus possible de SNPs ancestraux qui ne sont pas nécessairement encore présents chez les bovins domestiques modernes. L'intérêt de l'utilisation de trois algorithmes d'appel de SNPs est de déterminer celui qui permettra de détecter le plus possible de SNPs informatifs ou de trouver un compromis entre les trois algorithmes qui permettra d'obtenir le plus possible de SNPs partagés entre eux. Deux d'entre eux filtrent les SNPs selon le contrôle qualité de base et le troisième filtre selon la couverture de la région contenant le SNP. Les deux algorithmes qui filtrent selon le contrôle qualité de base sont GATK et Varscan. Ils diffèrent en plusieurs points : GATK supprime les lectures avec une qualité d'alignement faible et suppose que les erreurs de séquençage sont indépendantes alors que Varscan utilise toutes les lectures par défaut et suppose que la deuxième erreur a plus de chance. Mais, ces deux algorithmes peuvent être facilement dupés par manque de couverture et de lectures courtes parce que plus la couverture est faible, plus la détection des SNPs rares est limitée et plus la différence entre les algorithmes d'appel de SNPs est large.

FreeBayes est un détecteur de variants génétiques conçu pour détecter les petits polymorphismes, en particulier les SNP (polymorphismes mononucléotidiques), les indels (insertions et délétions), les MNP (polymorphismes multi-nucléotidiques) et les événements complexes (événements composites d'insertion et de substitution) plus petits que la longueur d'un alignement de séquence à lecture courte. Freebayes, qui filtre selon le nombre de lectures qui couvre le SNP, et GATK différent aussi en un point important qui est le type de SNP appelé. GATK provoque des erreurs qui ignorent l'allèle de référence ce qui fait que GATK appelle certains SNPs hétérozygotes comme des homozygotes alternatifs. Par contre, Freebayes provoque des erreurs en ajoutant l'allèle de référence même s'il n'est pas présent, ce qui fait qu'il peut appeler des homozygotes alternatifs comme des SNPs hétérozygotes (Hwang et al., 2015).

4) Recalibration de contrôle qualité de base (BQSR) :

La recalibration du score de qualité des bases est un processus dans lequel on applique un apprentissage automatique pour modéliser les erreurs produites par le séquenceur de manière empirique et ajuster les scores de qualité. Il s'agit d'une étape de prétraitement des données qui détecte les erreurs systématiques commises par le séquenceur lorsqu'il estime le score de qualité de chaque appel de base. La recalibration est utilisée donc pour minimiser le taux d'erreur et pour augmenter par conséquence la précision de l'algorithme d'appel de SNPs. Son principe repose sur un modèle de covariation basé sur les données et un ensemble de variants connus pour ajuster les scores de qualité des bases en se basant sur le modèle. L'algorithme BQSR suppose que tout SNP du fichier de données qui n'est pas dans le fichier de l'échantillon des SNPs connus a plus de chances de correspondre à une erreur de lecture par la machine. L'algorithme va chercher à identifier si les différentes erreurs potentielles se produisent dans des contextes de séquence similaires qui indiquerait des sources d'erreur systématique. Une étape fortement recommandée est de créer un deuxième modèle après recalibration et de générer des tracés avant et après pour visualiser le processus de recalibration. Cette étape est réalisée par l'outil GATK baserecalibrator. Mais, ce dernier ne peut pas corriger les appels de bases eux-mêmes c'est-à-dire qu'il ne peut pas déterminer si un C de faible qualité aurait dû être un G. Mais, il peut au moins indiquer plus précisément à l'algorithme de SNPs à quel point il peut être confiant dans l'identité de la base. (https://software.broadinstitute.org/gatk/documentation/tooldocs/3.8/org_broadinstitute_gatk_tools_walkers_bqsr_BaseRecalibrator.php).

5) Filtration des SNPs obtenus avec les SNP callers :

La filtration des SNPs consiste à choisir des seuils spécifiques et à exclure tout variant ayant des valeurs d'annotation supérieures ou inférieures aux seuils définis. Elle est utilisée pour améliorer la qualité de SNP. Plusieurs filtres peuvent être appliqués et dépendent du type des données et de l'objectif de l'analyse.

Dans le cas de l'ADNa et pour AS5, nous avons choisis d'utiliser les filtres suivants: (1) QualityByDepth qui est la confiance du variant divisée par la profondeur allélique (QD<2). (2) Mapping Quality ou MQ qui contrôle la qualité de cartographie de toutes les lectures sur un site (MQ<40). (3) StrandOddsRatio ou SOR qui est utilisé pour récupérer les régions riches en AT parce que FS (Fisher Strand) a tendance à pénaliser les variants localisés aux extrémités des séquences d'ADN (SOR>3). Fisher Strand est la probabilité en utilisant une échelle de Phread qu'il existe un biais de brin sur le site. Il nous indique si l'allèle alternatif a été vu plus ou moins souvent sur le brin direct ou inverse que l'allèle de référence. (4) MQRankSum qui compare la qualité de l'alignement des lectures qui soutiennent l'allèle de référence et l'allèle alternatif (MQRankSum < -12,5). (5) ReadPosRankSum, quant à lui, compare si les positions des allèles de référence et des allèles alternatifs sont différentes au sein des lectures, car le fait de ne voir un allèle que près des extrémités des lectures est révélateur d'une erreur, car c'est là que les séquenceurs ont tendance à commettre le plus d'erreurs (ReadPosRankSum < -8).

Les filtres recommandés par GATK ne sont appliqués que sur les SNPs appelés par GATK car cet algorithme calcule ces paramètres pour chaque SNP. Les autres algorithmes calcule d'autres paramètres avec d'autres modèles d'erreur, et donc, malheureusement, il n'est pas aisé, voire possible, de comparer les différents algorithmes après filtration.

IV) Analyse des résultats de l'appel de SNPs :

Tout d'abord, il est important de savoir que les algorithmes d'appel de SNPs doivent être appliqués sur les fichiers «bam» après avoir enlevé les duplications parce que l'appel de SNPs est difficile pour les lectures courtes avec duplications.

1) Les algorithmes de SNPs et les paramètres utilisés :

GATK HaplotypeCaller et varscan mpileup2cns ont été appliqués sur le fichier «bam» du notre échantillon ancien après avoir enlevé les duplications. Le génome de référence utilisé pour tous les SNP callers est celui produit par le projet «1000 bull genomes» et qui contient les séquences qui correspondent au chromosome Y (Séquence de référence: ARS-UCD1.2_Btau5.0.1Y.fa). On a utilisé quatre contrôles qualité de base pour chaque SNP caller: (1) q10 qui signifie qu'on tolère une erreur sur 10 donc un taux d'erreur de 0.1. (2) q15 c'est-à-dire qu'on tolère un taux d'erreur de 0.2 (3) q20 tolérant un taux d'erreur de 0.01 et (4) q30, le contrôle qualité de base le plus stringeant qu'on a utilisé qui tolère une erreur sur 1000. Freebayes a été appliqué sur le fichier bam sans duplication. On a testé cinq profondeurs de lecture égale à 4, 5, 6, 8 et 10 sachant que le génome d'AS5 est couvert 7 fois.

a) Optimum entre le contrôle de la qualité des bases et la couverture :

Après avoir réalisé l'appel de SNPs, nous avons cherché à trouver un optimum ente les trois algorithmes d'appel de SNPs c'est-à-dire trouver les conditions qui permettent d'obtenir le nombre de SNPs le plus important, partagés entre les trois algorithmes. Pour cela, les fichiers «vcf» (Variant Calling Format) obtenus à l'issue de l'appel de SNPs ont été utilisés pour faire des intersections entre soit deux fichiers «vcf» générés par deux algorithmes différents soit trois fichiers «vcf» générés par les 3 algorithmes d'appel ce SNPs. Pour ce faire, j'ai calculé le nombre de SNPs obtenus et appelés avec un seul algorithme d'appel de SNP, avec deux algorithmes à la fois et avec les 3 algorithmes et, à l'aide du package «eulerr» de R, j'ai présenté les comptages sous formes de diagramme de Venn pour mieux les visualiser. Les cercles sont proportionnelles au nombre de SNPs obtenus par les différentes conditions d'appel de SNPs (Figure 51).

Nous avons testé différentes intersections et recombinaisons entre GATK et Varscan en premier lieu et Freebayes dans un deuxième temps. Nous avons trouvé que Freebayes utilisé avec une profondeur de lecture égale à 5, a le nombre de SNPs partagés le plus importants avec la stringence q30 du score de qualité des bases de GATK et de Varscan. Nous avons utilisé l'intersection de GATK et Varscan avec un contrôle qualité de base égale à 30 parce que nous avons aussi observé qu'avec cette stringence on a obtenu le nombre le plus élevé de SNP partagés entre les deux algorithmes (données non montrées).

A l'issue de cette analyse, nous avons décidé de travailler avec ces conditions pour les 3 algorithmes d'appel de SNPs qui sont le contrôle de la qualité des bases qui autorise une erreur sur 1000 pour Varscan et GATK et une couverture de 5 lectures pour Freebayes (Figure 51).

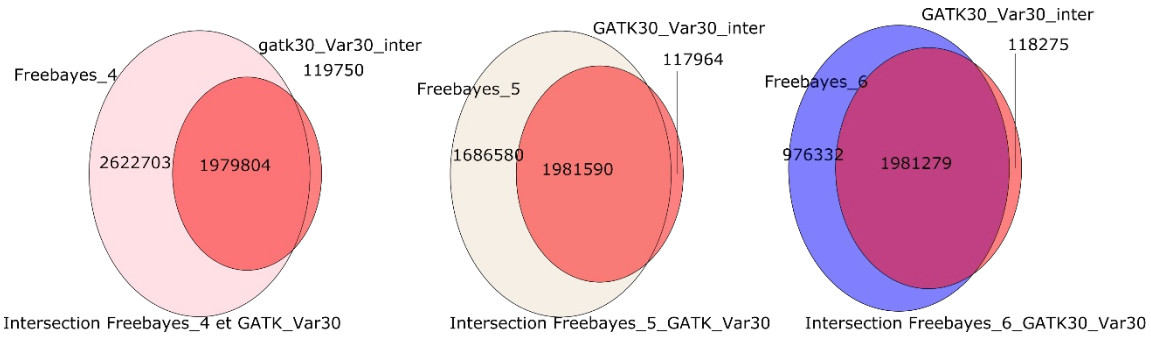


Figure 51: Représentation de nombre de SNPs obtenus avec les différents algorithmes d'appel de SNPs selon les conditions d'appel utilisées. a) A gauche : Intersection entre Freebayes_4 et l'intersection GATK_q30_Varscan_q30. b) Au centre de la figure: Intersection entre Freebayes_5 et l'intersection GATK_q30_Varscan_q30). c) A droite: Intersection entre Freebayes_6 et l'intersection GATK_q30_Varscan_q30.

b) Effet de la recalibration sur le nombre de SNPs obtenus par les différents algorithmes d'appel de SNPs :

La recalibration a été réalisée sur le fichier «bam» de l'alignement des séquences de l'aurochs AS5 sur le génome de la référence. Comme indiqué dans la figure 52, après recalibration, qui élimine les erreurs de séquençage, le nombre des SNPs obtenus a diminué d'une manière brutale. Le nombre de SNPs partagés par les 3 algorithmes a diminué 39 fois.

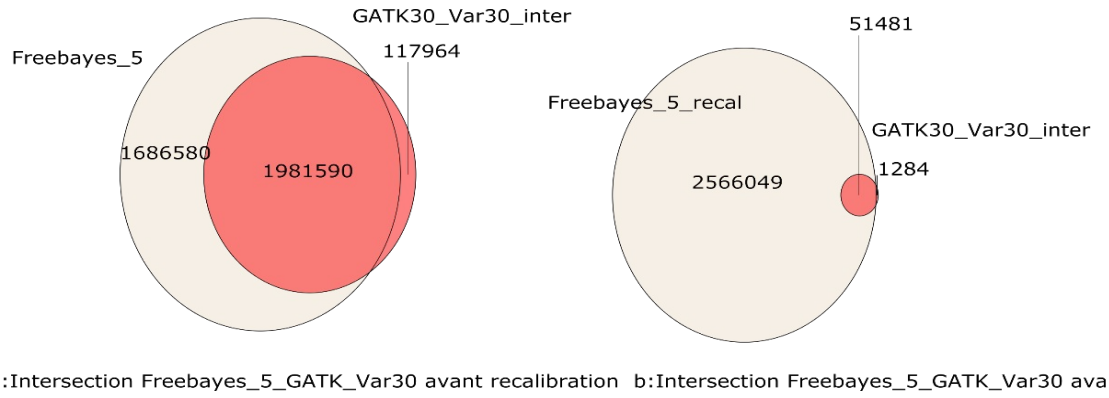


Figure 52 : Effet de la recalibration de score qualité de base sur l'appel de SNPs.

Nous remarquons ici que la diminution du nombre de SNPs a affecté beaucoup plus GATK et Varscan que Freebayes. Ceci est dû à une diminution très forte du score de qualité des bases à partir des données anciennes et il est donc important de détailler maintenant ce qui se passe lors de la recalibration pour expliquer pourquoi, pour la suite, nous avons choisi de diminuer le seuil minimal de la qualité des bases à des valeurs de 15 et 20.

En effet, comme nous séquençons des fragments d'ADNa courts ayant une taille médiane généralement inférieure à 50 bases en paired-end de 2*100bp, la plupart des lectures sont faites sur les deux brins. Le merging avec leeHom a induit un doublement des scores de qualité de bases pour toutes les positions lues sur les deux brins. Dans le cas des lectures sur les deux brins, les erreurs de lecture de séquence deviennent donc très peu probables. GATK considère ces scores de qualité de base comme trop élevé et va donc les ramener à des valeurs plus raisonnables selon lui, avec un score maximum de 30 (Figure 53). De fait, un score de 60 n'est pas réaliste car il y a d'autres sources d'erreurs de séquences que celles produites lors du séquençage: il y a les erreurs dues au transformations diagénétiques et les erreurs d'incorporation de bases par les polymérases lors de l'amplification des bases. Ces erreurs seront lues de façon fiable au niveau du séquenceur mais la position sera fausse. Bien que la recalibration vise à corriger les erreurs faites lors du séquençage, nous avons exploré si elle pouvait aussi corriger les erreurs dues aux transformations diagénétiques ou produites lors de la construction des banques.

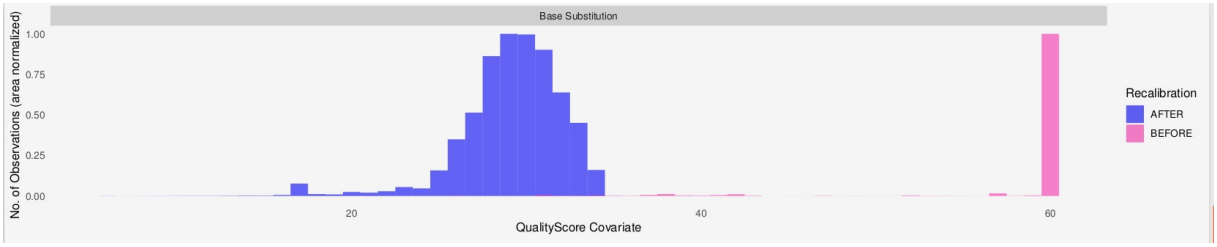


Figure 53: Distribution du score de qualité des bases avant et après recalibration avec le fichier BQSR.

c) Effet de la diminution du contrôle qualité de base après recalibration de score qualité de base :

Un appel de SNPs a été réalisé avec GATK et Varscan en utilisant pour chacun soit un contrôle qualité de base égale à 15, soit égale à 20. Nous avons comparé alors le nombre de SNPs obtenus après recalibration en utilisant les différents seuils de qualité de base testés. Nous avons trouvé que GATK_q15, Varscan_q15 et Freebayes_5 partagent le nombre de SNPs le plus élevé par rapport aux autres intersections: GATK_q20, Varscan_q20 et Freebayes_5 d'une part et GATK_q15, Varscan_q20 et Freebayes_5 d'autres part (Figure 54).

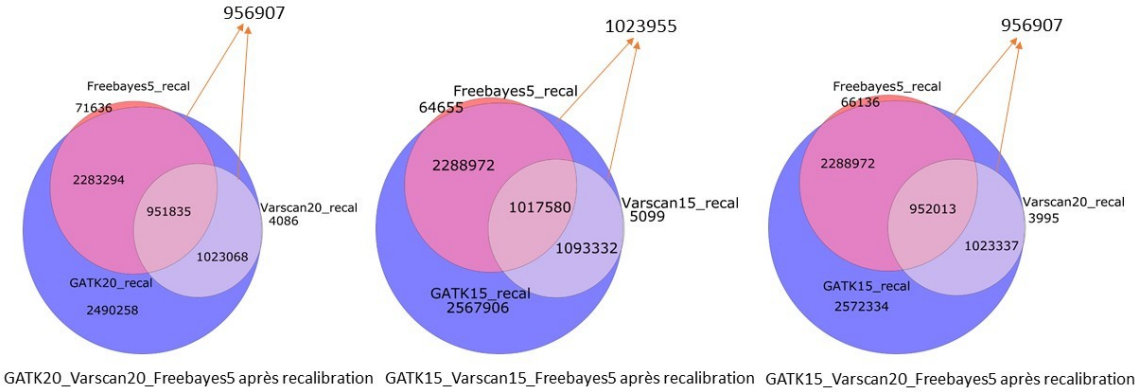


Figure 54: Représentation des différentes intersections entre les 3 algorithmes d'appel de SNPs après recalibration avec différents contrôles qualité de bases.

Nous avons trouvé aussi qu'avec ces paramètres, le nombre de SNPs obtenus après recalibration est proche de celui obtenu avant recalibration avec la stringence la plus élevée q30 (Une erreur sur 1000 est tolérée). Ce qui veut dire que nous n'avons pas perdu énormément de SNPs à la recalibration sauf que nous avons diminué de deux fois la qualité des bases comme observé auparavant.

2) Effet de la filtration et modification des seuils de filtres GATK:

Il a été recommandé par la communauté mettant en place l'outil GATK de modifier les seuils selon les besoins et les objectifs des chercheurs. Les seuils des filtres utilisés par GATK sont alors nécessaires à étudier car ils dépendent de la nature des données et de l'objectif de l'analyse. Mais, le but est toujours d'enlever, le plus possible, les SNPs faux positifs et de garder les vrais SNPs pour minimiser le taux d'erreur et obtenir des résultats robustes. Nous avons étudié la distribution des filtres appliqués sur les SNPs obtenus avec GATK. Tout d'abord, nous avons utilisé les seuils des filtres recommandés par GATK et sur cette base, nous avons modifié les seuils pour qu'ils soient adaptés à nos données.

a) Etude du filtre DP ou nombre de lectures couvrant le SNP :

Nous nous sommes intéressés à l'étude du filtre DP. Ce dernier filtre porte sur la profondeur de la couverture. Les SNPs appelés par GATK sont filtrés selon les seuils de filtres recommandés par GATK (QD<2, SOR>2, MQ<2, MQRankSum< -12.5 et ReadPosRankSum < -8). Les SNPs appelés par Varscan et Freebayes ne sont pas filtrés car les filtres GATK ne s'appliquent que sur les SNPs appelés par GATK.

Nous avons trouvé que par défaut, les 3 algorithmes d'appel de SNPs n'ont pas la même distribution de la DP (Figure 55). La valeur par défaut utilisée par GATK est une seule lecture qui supporte le SNP. Varscan exige, par défaut, qu'il faut 8 lectures pour que le SNP soit appelé alors que Freebayes, appelle les SNPs couverts par 5 lectures. Ce seuil de DP pour Freebayes était fixé en étudiant la cohérence entre la couverture et la stringence (Figure 51). Ces différences dans les valeurs de la DP, prises par défaut par les différents algorithmes d'appel de SNPs utilisés, influencent le nombre des SNPs obtenus (voir plus loin).

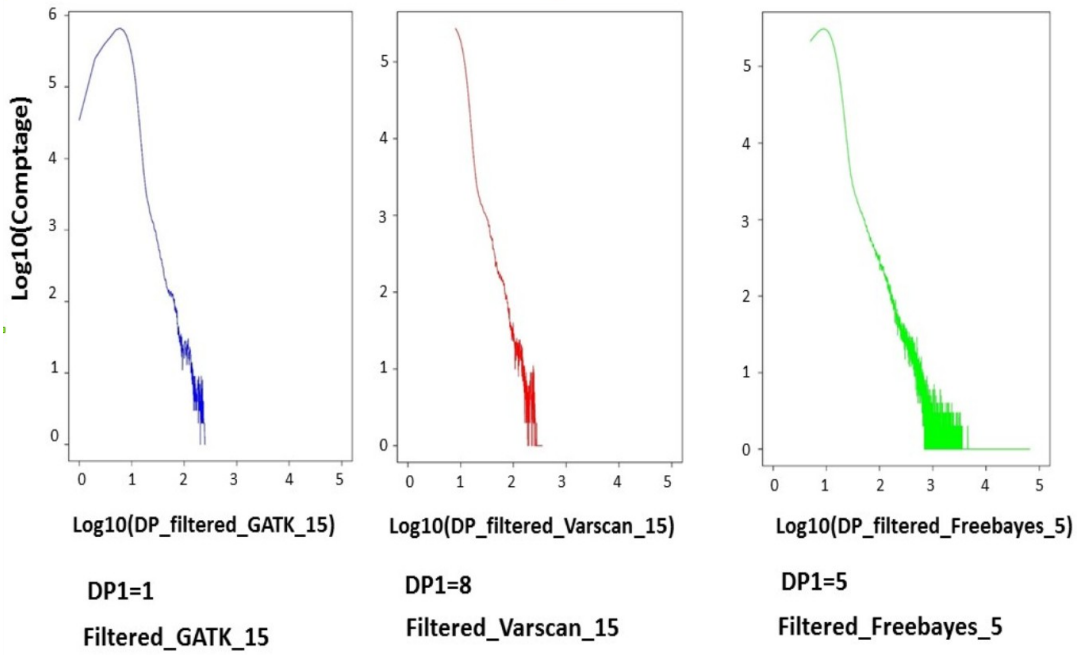


Figure 55 : Distribution de la DP selon l'algorithme d'appel de SNPs. La couleur bleue et rouge correspondent respectivement à la distribution de la DP pour les SNPs appelés avec GATK et Varscan avec un score de qualité des bases d'au moins 15. La couleur verte correspond à la distribution de la DP pour les SNPs appelés avec Freebayes avec une couverture d'au moins 5 lectures supportant le SNPs.

Puisque les différents algorithmes d'appel de SNPs exigent différentes couvertures de SNP pour qu'il soit appelé, nous avons cherché à comprendre si les SNPs appelés avec GATK mais qui ne sont pas appelés avec Freebayes sont des SNPs qui ont une couverture entre 1 et 5 (figure 56 à gauche). Pour mieux comprendre cette observation, j'ai généré une courbe du log du comptage de SNPs appelés par GATK et Freebayes en fonction de la DP. Nous avons trouvé que les SNPs non chevauchants entre ceux appelés et filtrés par GATK et ceux appelés par Freebayes et non filtrés (car les filtres GATK s'appliquent seulement sur les SNPs appelés par GATK) étaient dus au nombre de lectures couvrant le SNP inférieure à 5 car nous avons observé que les SNPs appelés seulement par GATK ayant une DP entre 1 et 5 sont nombreux et les SNPs de couverture plus élevé sont moins abondants.

Nous avons trouvé (voir figure 56 à gauche) qu'avec Freebayes qui appelle les SNPs couverts par au minimum 5 lectures, nous récupérons un nombre important de SNPs à couverture élevée (le maximum de couverture représenté dans cette figure est 30) ce qui ne reflète pas la réalité parce que notre génome est couvert 7 fois en moyenne. Freebayes a donc tendance à privilégier les régions du génome ou il y a trop de lectures cartographiées par rapport à la couverture moyenne. Ces régions pourraient être enrichies en lectures cartographiées de manière erronées, en particulier parce que le génome séquencé et le génome de référence diffère, soit à cause de duplications et de CNV dans le génome séquencé qui sont absentes dans le génome de référence, soit parce que le génome de référence est imparfait et représente mal certaines séquences dupliquées dans la plupart des génomes bovins. Le risque paraît donc élevé que Freebayes introduise beaucoup de faux positifs avec les paléogénomes généralement moins bien couverts que les génomes modernes.

On a tendance à dire pour GATK qu'il prend seulement les SNPs à couverture faible autour de la moyenne de la couverture de notre génome. GATK limite l'appel des SNPs qui ont une couverture élevée car il les considère probablement comme des couvertures aberrantes qui ne reflètent pas la réalité.

Pour affirmer ou infirmer cette interprétation, j'ai ajouté à la courbe précédente les comptages des SNPs appelées seulement avec GATK (nommés ainsi «complement») pour avoir une idée sur la couverture de ces SNPs spécifiques de GATK (Figure 56 à droite). Nous avons trouvé qu'on appelle peu de SNPs avec une couverture élevée et la majorité des SNPs appelés ont une couverture autour de la couverture moyenne du génome de l'aurochs anatolien ce qui soutient notre interprétation.
Cela suggère que quand nous sélectionnons les SNPs appelés à la fois par GATK et Freebayes, nous éliminons tous les SNPs ayant une couverture élevée non réelle.

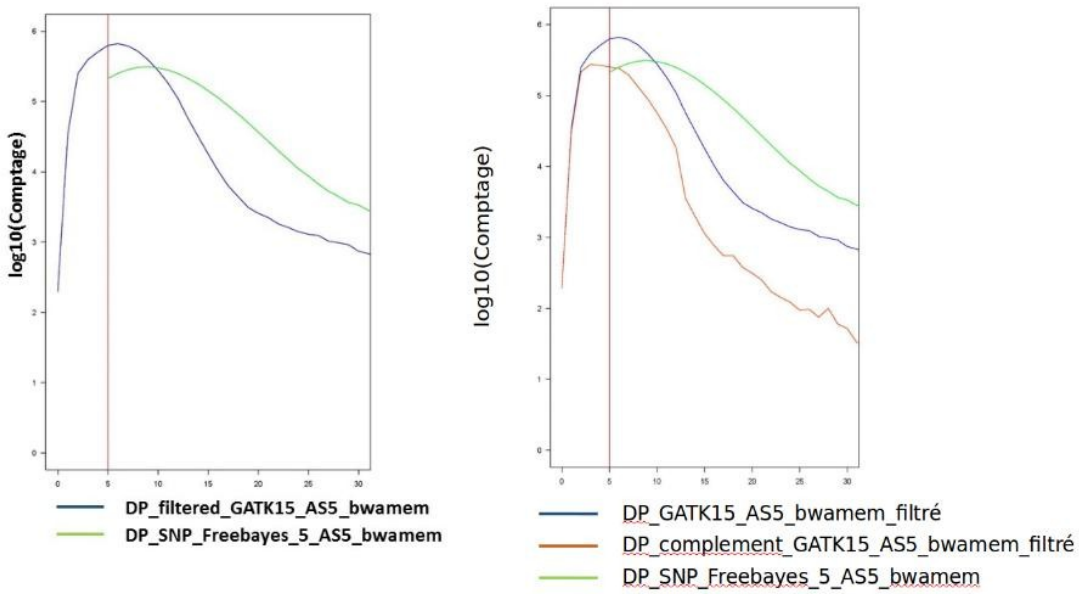


Figure 56: Couverture des SNPS appelés par les algorithmes d'appel de SNPs, GATK et Freebayes. Courbe à droite: Log de comptage de SNPs appelés en fonction de la DP(couverture). La couleur bleue et verte correspondent, respectivement, aux SNPs obtenus avec GATK et Freebayes. La couleur rouge correspond aux SNPs appelés seulement avec GATK.

J'ai étudié par la suite la distribution du taux d'hétérozygotie en fonction de la DP pour ajuster le seuil de ce filtre. J'ai déterminé pour chaque valeur de DP le ratio d'hétérozygotie et j'ai généré la courbe suivante.

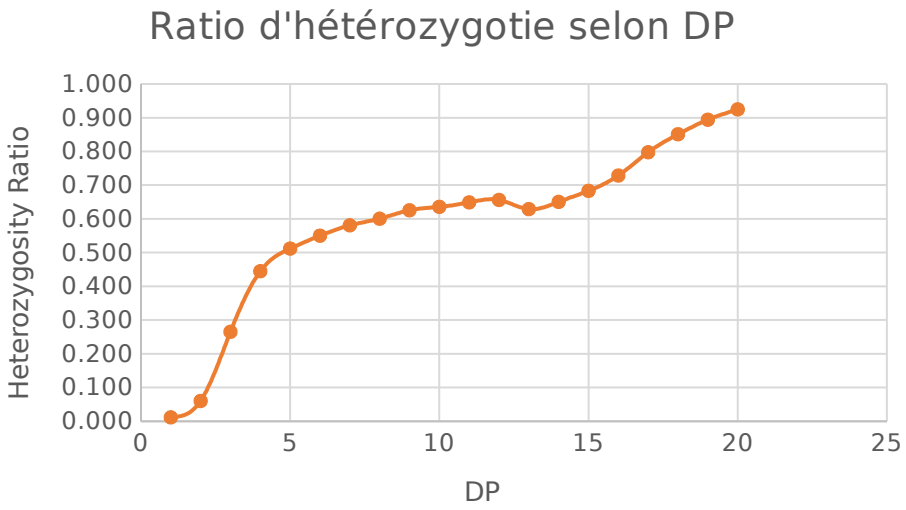


Figure 57: Courbe représentative du taux d'hétérozygotie en fonction de la DP. Les SNPs étaient appelés avec GATK en utilisant un filtre de qualité de base de 15 et ne sont pas filtrés.

Le ratio d'hétérozygotie varie de 0 à 1. Il atteint la valeur de 1 à une couverture égale à 25. Pour une DP inférieur à 5, l'algorithme d'appel de SNPs (GATK) appelle mal les SNPs hétérozygotes. Les taux d'hétérozygotie qu'on a obtenu dans cet intervalle (une DP entre 1 et 5) peuvent être dus à ce qu'on appelle «Allelic dropout» ou «Décrochage allélique». Le décrochage allélique résulte de données manquantes dans lesquelles une ou les deux copies alléliques au niveau d'un locus ne sont pas amplifiées par la polymérase au cours de la PCR soit par hasard, soit car les deux allèles ne s'amplifient pas avec la même efficacité. Ce problème, en particulier pour les échantillons dont l'ADN est de mauvaise qualité, comme le cas de l'ADNa, entraîne une estimation biaisée de l'hétérozygotie. Cet inconvénient diminue la précision du génotypage d'un échantillon unique (Wang et al., 2012). A partir d'une DP égale à 5 jusqu'à une DP égale à 15, il y a un plateau. Mais au-delà de cette DP, il y a une augmentation artificielle du taux d'hétérozygotie à cause de la contribution des alignements parasites, c'est-à-dire l'alignement des séquences qui sont répétées mais qui ne sont pas identifiées comme répétées dans le génome de référence ce qui fait qu'il y a beaucoup de lectures d'origine différente qui s'alignent sur le même endroit et qui ne sont pas filtrés par le filtre «Mapping Quality»

Nous avons étudié par la suite la distribution des SNPs appelés avec GATK en fonction de la DP. Nous avons trouvé que pour une DP inférieure ou égale à 3 et une DP supérieure ou égale à 14, le nombre de SNPs récupéré est faible (Figure 58).

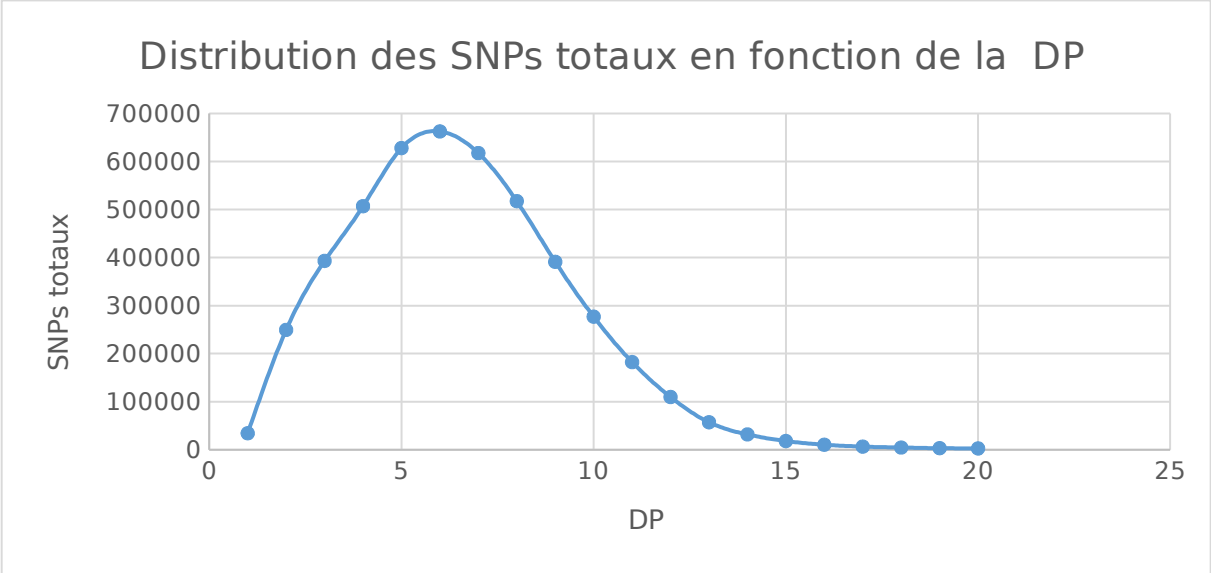


Figure 58: Courbe représentative de la distribution des SNPs totaux obtenus avec GATK en fonction de la DP.

Sur la base de ce résultat, nous avons décidé d'éliminer les SNPs ayant une DP inférieure ou égale à 3 et ceux ayant une DP supérieure ou égale à 14 pour ne garder que les SNPs ayant un taux d'hétérozygotie qui se rapproche de la réalité.

b) Étude du filtre QUAL :

Le filtre QUAL détermine à quel point nous sommes convaincus qu'il existe une sorte de variation sur un site donné. La variation peut être présente dans un ou plusieurs échantillons. Le filtre QUAL ou sa forme normalisée QD (ou «Quality by depth» dont je parlerai prochainement) est recommandé dans un contexte multi-échantillons. Malgré cette recommandation de GATK, nous avons étudié sa distribution pour voir l'effet de la recalibration du score de qualité des bases ainsi que l'effet de la filtration sur la DP des SNPs sur la distribution de QUAL (Figure 59).

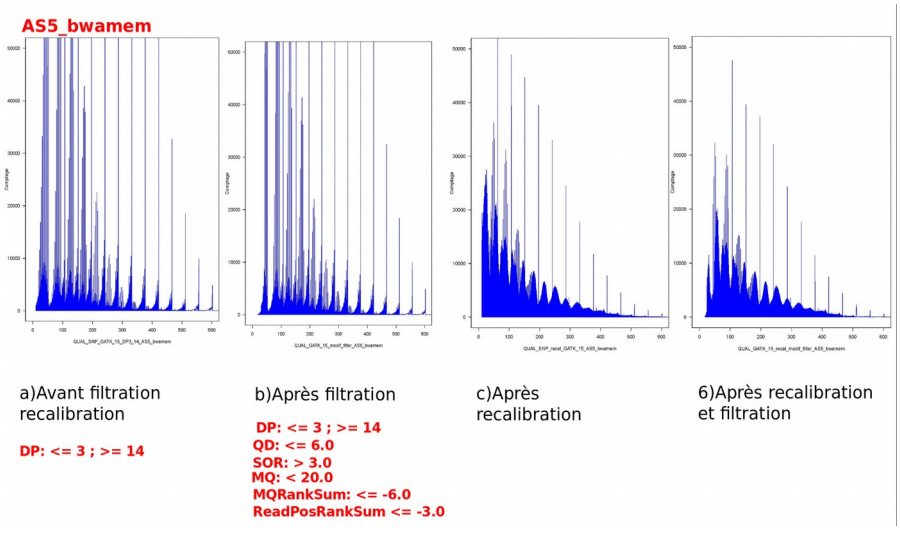


Figure 59: Distribution du filtre QUAL (a) après filtration seulement sur la DP et après recalibration, (b) après filtration sur la DP et sur la QD,SOR,MQ, MQRankSum et ReadPosRankSum avec les seuils recommandés par GATK, (c) après recalibration et (d) après filtration et recalibration.

Nous avons gardé seulement les SNPs ayant une valeur de QUAL allant jusqu'à 800 car les SNPS de valeurs plus élevées que 1000 n'étaient pas abondants. La distribution du filtre QUAL après filtration sur la DP montre que tous les SNPs qui avaient des valeurs QUAL élevées ont été éliminés parce qu'ils étaient probablement causés par des couvertures élevées et aberrantes.

Après recalibration, la distribution est changée d'une manière remarquable. En effet, après recalibration du score de qualité des bases, les pics bas trouvés avant recalibration se déplacent vers des valeurs de QUAL plus élevée et n'ont pas la même périodicité. Après recalibration, les pics plus hauts et plus étroits obtenus avant recalibration et filtration sont redistribués sur les pics les plus bas et ont augmenté leurs valeurs et changé leurs formes. Les pics les plus haut obtenus avant recalibration sont probablement dus aux lectures sur les deux brins et ils diminuent et se mélangent avec les autres pics lors de la recalibration (Comme déjà expliqué dans la partie de la recalibration).

c) Étude du filtre GQ ou Qualité de génotype :

Pour mieux comprendre la périodicité des pics obtenues avec le filtre QUAL, nous avons étudié la distribution du filtre GQ (Figure 60). En effet, alors que le filtre QUAL détermine à quel point nous sommes convaincus qu'il existe une sorte de variation sur un site donné, le filtre GQ nous indique à quel point nous sommes convaincus que le génotype attribué à un échantillon particulier est correct.

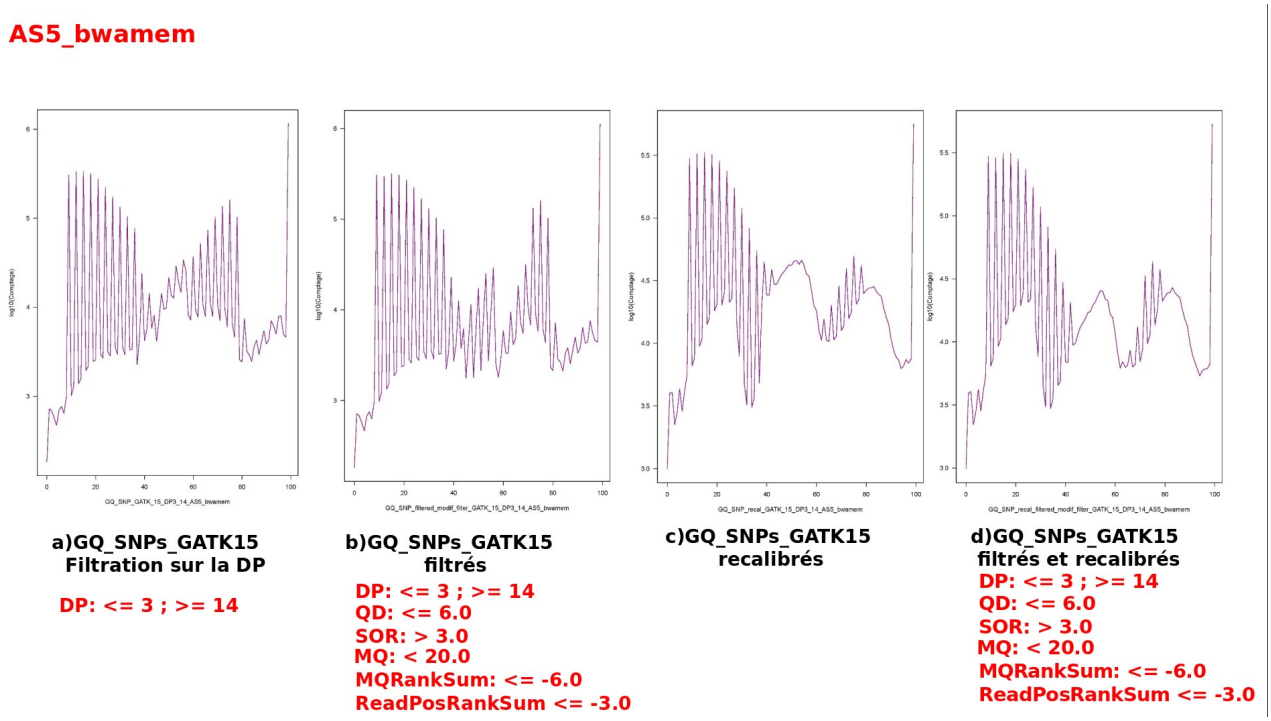


Figure 60: Distribution du filtre GQ (a) Après filtration sur la DP seulement, (b) après filtration avec les seuils de filtres modifiés notés en rouge, © après filtration sur la DP et recalibration du score de qualité des bases et (d) après filtration et recalibration.

La distribution du filtre GQ nous a montré que la filtration diminue significativement les SNPs ayant une GQ entre 40 et 70 alors qu'elle ne touche pratiquement pas ceux ayant une valeur entre 0 et 40. La recalibration enlève les SNPs ayant une valeur de GQ inférieure à 10 et redistribue complètement les valeurs entre 30 et 100. Pour comprendre la différence de distribution des valeurs de GQ entre 30 et 100, j'ai étudié l'effet du filtre QD («Quality by depth») sur la distribution des valeurs de GQ.

d) Effet du filtre de qualité par profondeur QD (Quality by depth) sur la distribution de la qualité de génotypes :

Comme chaque lecture contribue au score de QUAL, les variants correspondant aux régions à couverture élevée peuvent avoir des scores QUAL gonflés artificiellement ce qui donne l'impression que l'appel de SNPs est étayé par plus de preuves que ce qu'il n'est réellement. Dans ce cas, normaliser la confiance du variant par la profondeur donne une image plus objective de l'appel de SNPs.

Nous avons modifié le seuil de filtre recommandé par GATK qui est QD <2 pour le monter à QD ≤ 6. Cette modification nous a permis de conclure que la qualité de profondeur influence sur la qualité de génotype. En effet, la distribution de GQ est différente en fonction du seuil de filtration du filtre QD (Figure 61) ce qui suggère que les SNPs ayant une QD entre 2 et 6 (ceux éliminés à la suite de l'élévation du seuil de QD) avaient probablement des valeurs de GQ élevées. La zone de GQ entre 40 et 60 varie en fonction du seuil de filtration de QD et en fonction de l'utilisation ou non du filtre QD.

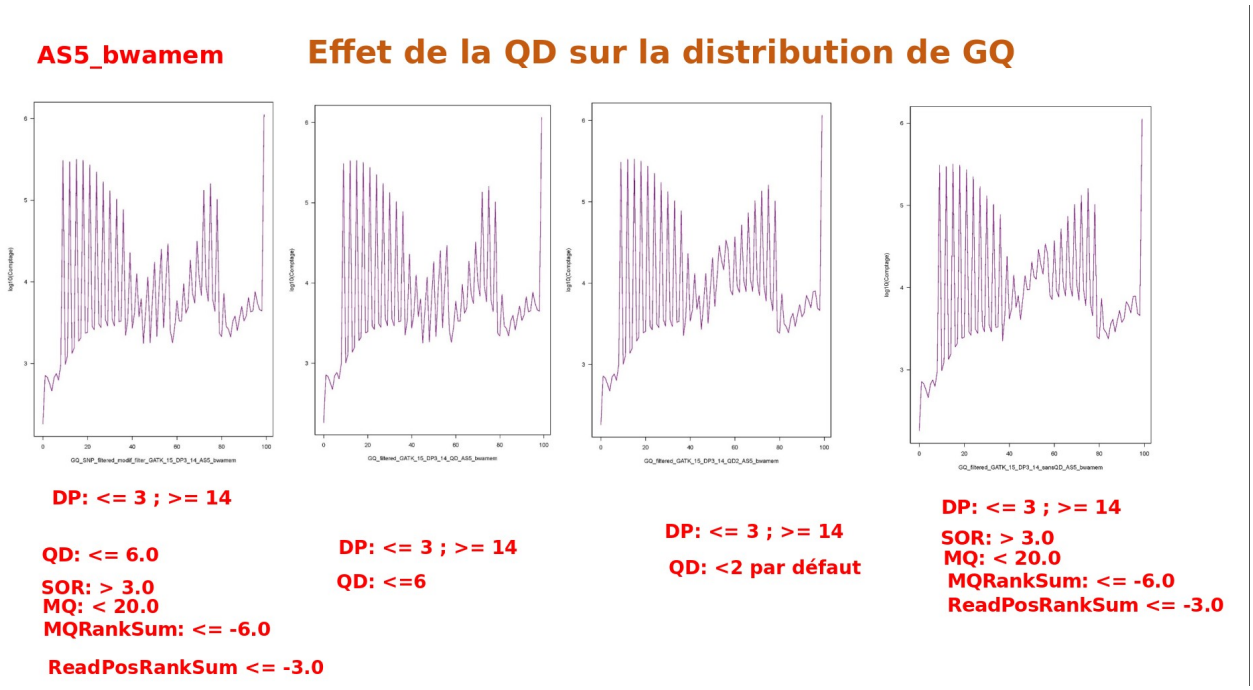


Figure 61: Distribution du filtre GQ (a) Après modification de tous les seuils de filtres, (b) Après modification des seuils de DP et de QD, © Après filtration sur la DP seulement en gardant le seuil QD recommandés par GATK et (d) Après filtration sur les différents filtres sauf QD.

Pour comprendre pourquoi le nombre de SNPs ayant une GQ entre 40 et 70 diminue quand nous appliquons le filtre QD≤ 6, j'ai déterminé la distribution de GQ pour les SNPs ayant une DP entre 3 et 14 et ayant une QD entre 2 et 6, c'est-à-dire les SNPs qu'on élimine avec le seuil de filtre QD modifié. D'après la figure 62, un grand nombre de SNPs enlevés par le filtre QD sont des SNPs ayant une GQ entre 40 et 70 expliquant ainsi la baisse de nombre de SNPs dans cet intervalle de valeur de GQ.

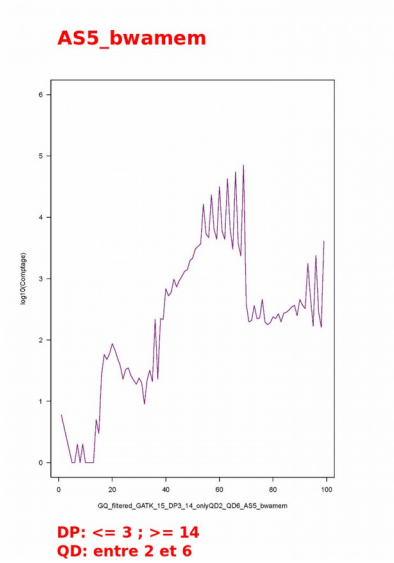


Figure 62: Distribution de valeur GQ des SNPs ayant une QD entre 2 et 6.

Nous avons étudié par la suite le chevauchement des différents algorithmes d'appel de SNPs avant et après recalibration et filtration pour étudier l'effet de chacun de ces traitements sur le nombre de SNP obtenus avec chacun des trois algorithmes.

3) Chevauchement entre la diversité génétique moderne connue et AS5 :

Nous avons essayé de déterminer par la suite la proportion des SNPs ancestraux obtenus avec les différents algorithmes d'appel de SNPs et qui ne sont pas présents dans le pool génétique moderne. Pour cela, nous avons exploré le chevauchement entre les variants d'AS5 obtenus avec les différents algorithmes d'appel de SNPs et la diversité moderne connue représentée dans le fichier BQSR du projet 1000 bull genome (Daetwyler et al., 2014) et avons déterminé le pourcentage de SNPs partagés avec la liste des variants connus. En effet, nous avons étudié le chevauchement entre notre échantillon ancien et la liste des SNPs modernes après recalibration de contrôle qualité de base, après filtration et après recalibration et filtration.

a) Chevauchement entre les SNPs appelés avec les 3 algorithmes de SNPs après recalibration :

Après avoir appliqué la recalibration sur les séquences alignées sur le génome de référence bovin moderne avec l'outil GATK Baserecalibrator, j'ai compté le nombre de SNPs obtenus avec chaque algorithme d'appel de SNPs, le nombre de SNPs appelés à la fois par deux algorithmes (GATK-Varscan, GATK-Freebayes et Varscan-GATK) et par les 3 algorithmes ainsi que le nombre de SNPs partagé avec les SNPs modernes connus pour chaque algorithme d'appel de SNPs (Figure 63).

Nous avons remarqué que le nombre de SNPs obtenus par les différents algorithmes d'appel de SNPs ainsi que le nombre de SNPs partagés par les algorithmes d'appel de SNPs deux à deux a diminué après recalibration. Alors que le nombre de SNPs partagés par les 3 algorithmes d'appel de SNPs a augmenté après recalibration et qu'il est passé de 1015736 à 1017580 (gain de 1844 SNPS) ce qui montre que la recalibration enlève les faux positifs et augmente l'aire des SNPs partagés par les différents algorithmes d'appel de SNPs. Nous avons remarqué aussi que le pourcentage des SNPs partagés avec les SNPs modernes a augmenté pour GATK et Freebayes alors qu'il a diminué de 2 % avec Varscan ce qui suggère que l'algorithme d'appel de SNPs Varscan n'est pas aussi efficace pour appeler des SNPs plus probablement authentiques.

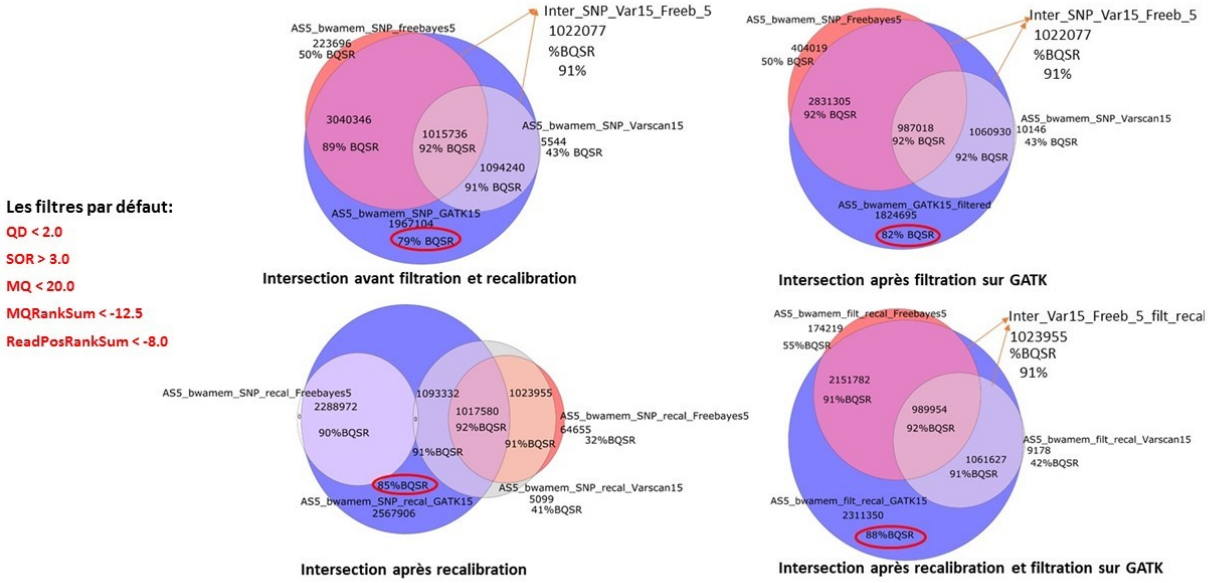


Figure 63: Nombre de SNPs appelés par les 3 algorithmes d'appel de SNPs avant filtration et recalibration, après filtration avec les filtres recommandés par GATK, après recalibration et après filtration et recalibration. La couleur bleue correspond aux SNPs appelés par GATK, la couleur rose correspond aux SNPs appelés par Freebayes et la couleur grise correspond aux SNPs appelés par Varscan. Les valeurs en pourcentage représentent le pourcentage des SNPs trouvés dans le fichier de SNPs modernes et appelés par chacun des algorithmes d'appel de SNPs.

b) Chevauchement entre les SNPs appelés par les 3 algorithmes de SNPs après filtration :

L'application des filtres recommandés par GATK a diminué le nombre de SNPs. En effet, en plus de la diminution du nombre de SNPs appelés par les différents algorithmes, le nombre de SNPs appelés par les 3 algorithmes à la fois a diminué aussi contrairement à la recalibration (Figure 63). La perte de SNPs partagés entre les 3 algorithmes d'appel de SNPs a été estimée à 2.83 fois. Le pourcentage de SNPs partagés avec les SNPs connus a augmenté après filtration pour les SNPs appelés avec GATK et Freebayes mais pas avec Varscan. Cela montre que la filtration enlève les faux positifs et augmente la précision de la détection des SNPs partagés avec les modernes.

c) Effet de la recalibration de score qualité de base et la filtration sur l'appel de SNPs :

Quand nous avons étudié l'effet de la recalibration sur les SNPs filtrés, nous avons remarqué que le nombre de SNPs a diminué sauf pour la zone des SNPs partagés par les différents algorithmes d'appel de SNPs ce qui montre une deuxième fois que la recalibration élimine les SNPs faux positifs et augmente le nombre de SNPs partagés par les différents algorithmes d'appel de SNPs.
De plus, quand nous avons filtrés les SNPs après avoir effectué la recalibration, nous avons remarqué que le pourcentage de SNPs appelés seulement par GATK et partagés avec les SNPs modernes connus a augmenté de 3% ce qui montre que la filtration nous a fait gagné des SNPs réellement positifs. Ce résultat nous permet de conclure que la filtration et la recalibration du score de qualité des bases sont deux étapes augmentant la robustesse d'un jeu de données justifiant notre choix de les utiliser .

d) Effet de modification des filtres GATK :

Après avoir étudié la distribution des filtres de GATK, nous avons défini nos propres seuils pour chaque filtre (QD, GQ, DP, SOR, MQRankSum et ReadPosRankSum) afin de répondre a nos objectifs, une stratégie recommandés par GATK (Tableau 5).

Filtre GATK	Seuil recommandé de filtre GATK	Seuil modifié de filtre GATK
QD (Quality by depth)	<2	<= 6
SOR (StrandOddsRatio)	>3	>3
MQ (Mapping Quality)	<20	<20
MQRankSum (Mapping Quality RankSum Test)	< -12.5	<= -6
ReadPosRankSum (Read PosRankSum Test)	< -8	<= -3 et >= 3
DP	x	<= 3 et >= 14

Tableau 5: Seuil de filtres utilisés lors de la filtration des SNPs appelés par les 3 algorithmes d'appel de SNPs.

A l'issue de la modification des seuils de certains filtres, nous avons constaté que le nombre de SNPs a diminué, ce qui est plutôt attendu (Figure 64). Après filtration seulement sur la DP, le chevauchement entre les 3 algorithmes d'appel de SNPs est amélioré, le nombre de SNPs partagés a augmenté. La filtration sur la DP a permis d'enlever les SNPs aberrants qui viennent de couvertures aberrantes. En modifiant les seuils de filtres, la plupart des SNPs appelés avec Freebayes et Varscan sont aussi obtenus avec GATK. Très peu de SNPs sont appelés seulement avec Freebayes ou Varscan, GATK semble donc être l'algorithme d'appel de SNPs le plus pertinent car il est capable d'appeler les SNPs appelés par l'un ou l'autre des autres algorithmes ainsi que des SNPs spécifiques qui sont présents dans le jeu de SNPs connus avec une fréquence à peine inférieure à celles des SNPs appelés par plusieurs algorithmes.

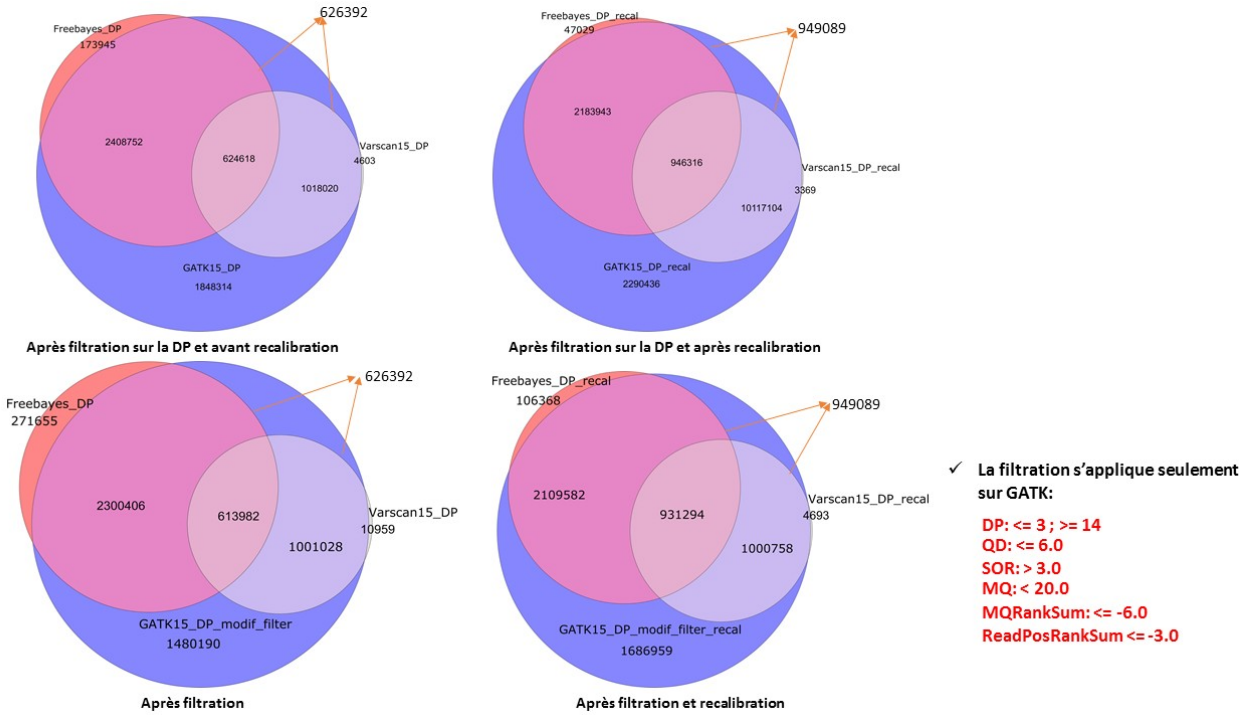


Figure 64: Nombre de SNPs appelés par les 3 algorithmes d'appel de SNPs après filtration sur la DP, après filtration avec les filtres dont les seuils ont été modifiés et après filtration et recalibration. La couleur bleue correspond aux SNPs appelés par GATK, la couleur rose correspond aux SNPs appelés par Freebayes et la couleur grise correspond aux SNPs appelés par Varscan.

Le tableau 6 montre le pourcentage de SNPs enlevés après recalibration et sans filtration sur la DP et après recalibration et filtration sur la DP obtenus avec chaque algorithme d'appel de SNPs ou partagés par deux ou 3 algorithmes d'appel de SNPs. Quand on appelle les SNPs avec GATK, nous perdons 0,4% des SNPs après recalibration et filtration sur la DP. Nous n'avons pas constaté une différence avec Varscan. Avec Freebayes, nous éliminons deux fois moins de SNPs quand nous filtrons sur la DP, ce qui explique l'amélioration du chevauchement entre GATK et Freebayes et indique que les SNPs qui étaient appelés seulement avec Freebayes étaient dus aux couvertures élevées aberrantes.

Effet de la recalibration	sans filtre DP	avec filtre DP
AS5_bwamem_all_SNP_GATK_15		
AS5_bwamem_all_SNP_recal_GATK_15	1.5%	1.1%
AS5_bwamem_all_SNP_Varscan_15		
AS5_bwamem_all_SNP_recal_Varscan_15	0.1%	0.1%
AS5_bwamem_all_SNP_freebayes_5		
AS5_bwamem_all_SNP_recal_freebayes_5	16.2%	7.3%

Tableau 6: Pourcentage de SNPs perdus après recalibration et filtration sur la DP (garder seulement les SNPs ayant une DP entre 3 et 14).

V) Effet du nouveau fichier de recalibration de score qualité de base :

Le fichier BQSR de référence a été mis à jour en 2020. Le nouveau fichier combine celui d'avant avec des variant appelés indépendamment dans diverses autres espèces, plus particulièrement les zébus, et il est donc approprié pour notre étude. Les deux fichiers qui ont précédé celui-là n'étaient pas appropriés à la recalibration des out-groupes. Ce fichier a permis la recalibration des espèces plus éloignées comme bison européen, Bison bison, Yak, Gaur et Banteg mais il n'a pas permis la recalibration des génomes de Buffle.

Comme nous séquençons des fragments d'ADNa courts ayant une taille médiane généralement inférieure à 50 bases en paired-end de 75 ou 100bp, la plupart des lectures sont faites sur les deux brins, ce qui limite considérablement le risque d'erreur. Les scores qualités de base s'additionnent alors et on obtient une valeur de 60 avant recalibration. L'algorithme de score qualité de base considère cette valeur comme une valeur qui ne représente pas la réalité. Il va alors attribuer aux SNPs nouvelles valeurs de scores qualité de base en tenant compte d'un seul brin. On perd donc a priori le gain d'une séquence sur les deux brins. Par contre, bien que la séquence puisse ne pas avoir d'erreur machine, le fragment peut porter une modification de bases artéfactuelle introduite soit à cause d'une modification diagénétique, soit une simple erreur de copie lors de la PCR. Ce type d'erreur pourrait aussi être détectée par l'algorithme de recalibration. Nous nous sommes convaincus que tel était le cas au vu du type de résultats présentés ici, que nous avons réanalysés pour apprécier le gain du nouveau jeu de données de référence pour le BQSR.

J'ai testé alors l'effet du nouveau fichier BQSR sur le nombre et la qualité des SNPs obtenus. La recalibration a été faite sur le génome de l'aurochs anatolien de 9700 calBC.

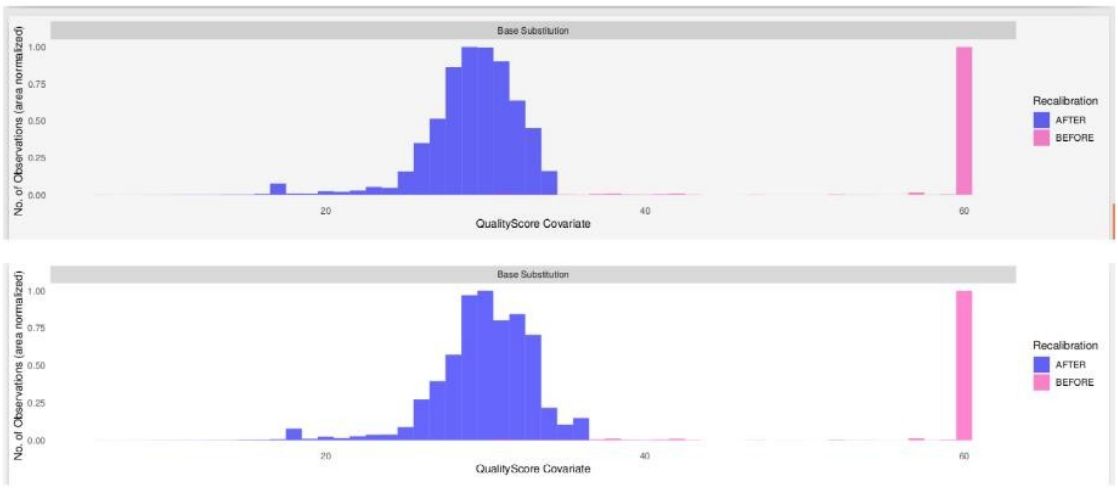


Figure 65: Distribution du score de qualité des bases avant et après recalibration avec l'ancien fichier BQSR (en haut) et le nouveau fichier BQSR (en bas).

Nous observons que les distributions de score qualité sont différentes en fonction du fichier de recalibration. La recalibration avec le nouveau fichier BQSR a permis d'obtenir une distribution plus large surtout pour les valeurs de haute valeur de score de qualité et un déplacement vers les meilleurs scores de qualité. On en conclue que le nouveau fichier BQSR a permis l'amélioration des scores de qualité des bases (Figure 65).

Je voulais par la suite évaluer le nombre de SNPs récupéré après recalibration avec le nouveau fichier BQSR et le comparer avec le nombre et la qualité des SNPs obtenu quand on recalibre avec l'ancien fichier BQSR. En effet, après avoir fait la recalibration avec la commande «gatk BaseRecalibrator», j'ai utilisé la commande «gatk HaplotypeCaller» pour appeler les SNPS produisant ainsi un fichier vcf (Variant Call Format). J'ai fait une intersection entre le fichier vcf obtenu après recalibration avec l'ancien fichier BQSR de référence et appel de SNPs et le fichier vcf obtenu après recalibration avec le fichier BQSR de référence mis à jour et appel de SNPs pour obtenir un autre fichier vcf qui contient seulement les SNPs partagés entre les deux fichiers vcf de départ ceci avec la commande «vcf-isec» de bcftools. Le nombre de SNPs dans chaque fichier a été compté et les comptages ont permis de faire un diagramme de Venn (Figure 66).

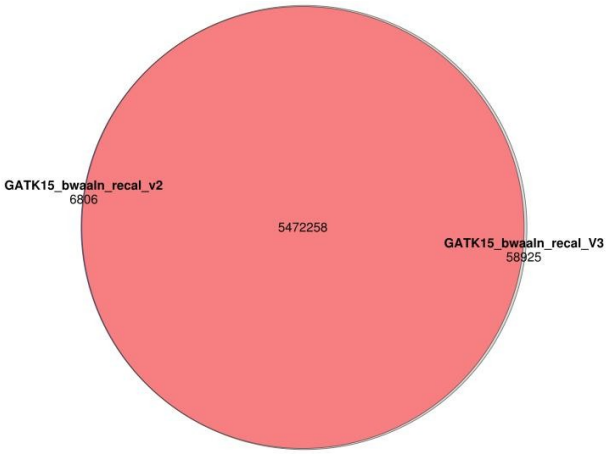


Figure 66: Nombre de SNPs obtenus après recalibration avec l'ancien fichier BQSR (v2) et le nouveau fichier BQSR (v3).

Le nombre total de SNPs obtenu après recalibration du score de qualité des bases avec le nouveau fichier BQSR est plus élevé que celui obtenu avec l'ancien fichier BQSR. Presque 5.5 millions de SNPs sont retrouvés à la fois avec le nouveau et l'ancien fichier BQSR. 0.1 % des SNPs sont appelés seulement quand le score qualité de base est recalibré avec l'ancien fichier BQSR et 1 % sont retrouvés après recalibration du score de qualité des bases avec le nouveau fichier BQSR. La fraction de SNPs qu'on ne récupère pas avec le nouveau fichier BQSR pourrait soit correspondreaux faux SNPs qui ont été éliminés par l'utilisation d'un panel de référence plus diversifié, aux vrais SNPs qui sont devenus des faux SNPs par exemple parce qu'ils correspondaient à des SNPs rares spécifiques des bœufs modernes et qui sont passés sous le seuil de fiabilité du fait de l'ajout de génomes d'autres espèces ne présentant pas ces SNPs. Dans ce cas, c'est difficile de choisir entre les deux hypothèses.

Les SNP retrouvés seulement après recalibration avec le nouveau fichier BQSR représente l'information génétique gagné avec ce nouveau fichier BQSR qui doit correspondre à de la variabilité présente chez les aurochs et les bovins ayant divergé auparavant mais qui est absente chez les bœufs modernes et doit être enrichie en vrais positifs du type qui nous intéressait à priori.

1) Effet de l'amélioration du score de qualité des bases sur la qualité des génotypes :

Pour évaluer l'effet de l'amélioration du score de qualité des bases due à l'utilisation du nouveau fichier BQSR sur la qualité des génotypes, j'ai étudié la distribution de qualité des génotypes obtenus après utilisation des deux fichiers BQSR (Figure 67). La figure 67 (à droite) montre sur l'axe des abscisses la qualité des génotypes et sur l'axe des ordonnées le nombre de SNPs en log 10. On observe que la différence concerne les SNPs ayant une qualité de génotypes entre 40 et 60 ce qui suggère que la recalibration avec le nouveau fichier BQSR (V3) permet de récupérer des SNPs avec une meilleure qualité de génotypes de SNPs appelés.

J'ai déterminé par la suite la qualité des SNPs appelés seulement quand le fichier de séquence de l'échantillon AS5 est recalibré avec l'ancien fichier BQSR de référence et ceux appelés seulement quand le fichier de séquence de l'échantillon AS5 est recalibré avec le nouveau fichier BQSR de référence (Figure 67 à gauche). La couleur rouge et bleue correspondent aux SNPs obtenus après la recalibration avec V2 et après la recalibration avec V3, respectivement. Cette figure montre qu'on a une fraction de SNPs trouvés seulement avec V3 qui a de meilleure qualité de génotypes. La recalibration avec V3 augmente peu la valeur GQ de certains SNPs, alors que tous les SNPs qui ont une valeur GQ plus élevée trouvés avec V3 sont trouvés aussi avec V2. On conclue alors qu'avec V3, les GQ des SNPs s'améliorent.

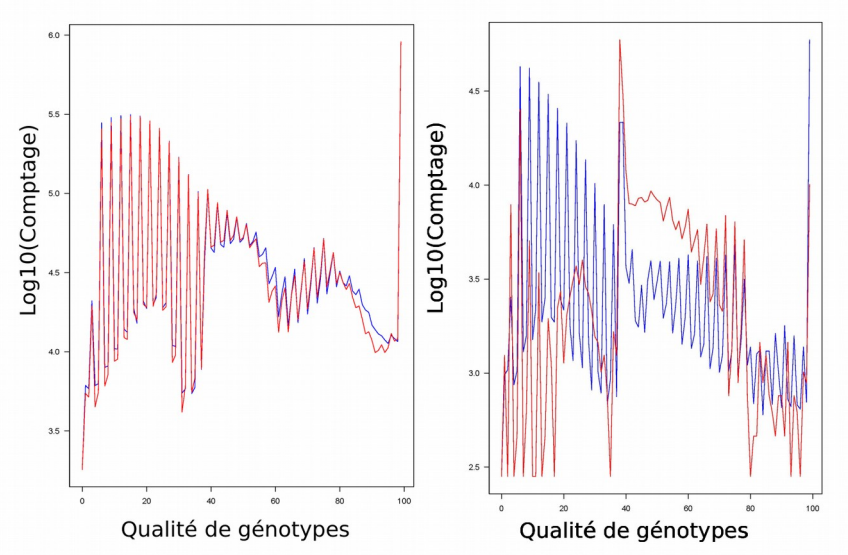


Figure 67: A droite: Qualité de génotypes obtenus après recalibration du score de qualité des bases en utilisant l'ancien (en rouge) et le nouveau (en bleu) fichier BQSR. A gauche: Qualité de génotypes des SNPs obtenus exclusivement après recalibration du score de qualité des bases en utilisant l'ancien (en rouge) ou le nouveau (en bleu) fichier BQSR.

2) Effet du nouveau fichier BQSR sur l'appel des SNPs par les 3 algorithmes d'appel de SNPs utilisés dans notre étude :

Comme précédemment, nous avons comparé les différents algorithmes d'appel de SNPs Après recalibration avec le nouveau fichier BQSR. Cette figure montre à gauche le diagramme de venn du nombre de SNP obtenu après recalibration avec le nouveau fichier BQSR mis à jour et à droite celui obtenu avec l'ancien fichier BQSR. On a trouvé qu'avec les différents algorithmes d'appel de SNP, le nombre de SNP obtenus après recalibration avec le nouveau fichier BQSR est plus élevé que celui obtenu avec l'ancien. Si on regarde le nombre de SNPs appelés par les différents callers après recalibration avec le nouveau fichier BQSR, on trouve qu'il est plus important que celui obtenu après recalibration avec l'ancien. La différence est de 2148 SNPs (Figure 68).

L'effet positif est moins important que celui que l'on avait observé en comparant la non-recalibration avec le calibration, mais le nouveau fichier BQSR a encore permis de gagner des SNPs qui ont des propriétés assez convaincantes. Le nouveau fichier BQSR nous a permis d'augmenter l'aire des SNPs partagés par les 3 SNPs callers et de minimiser en quelque sorte, l'appel des SNPs faux positifs, bien que le risque existe toujours que des SNPs faux positifs soit toujours présents. Ainsi, le pourcentage de SNPs appelés seulement avec GATK et partagés avec les SNPs modernes a diminué quand on recalibre le génome de l'aurochs anatolien de 9000 ans avec le nouveau fichier BQSR. Est-ce que cela signifie que le nouveau fichier BQSR nous a permis de récupérer des SNPs authentiques qui ne sont pas présents dans le jeu de données modernes ?

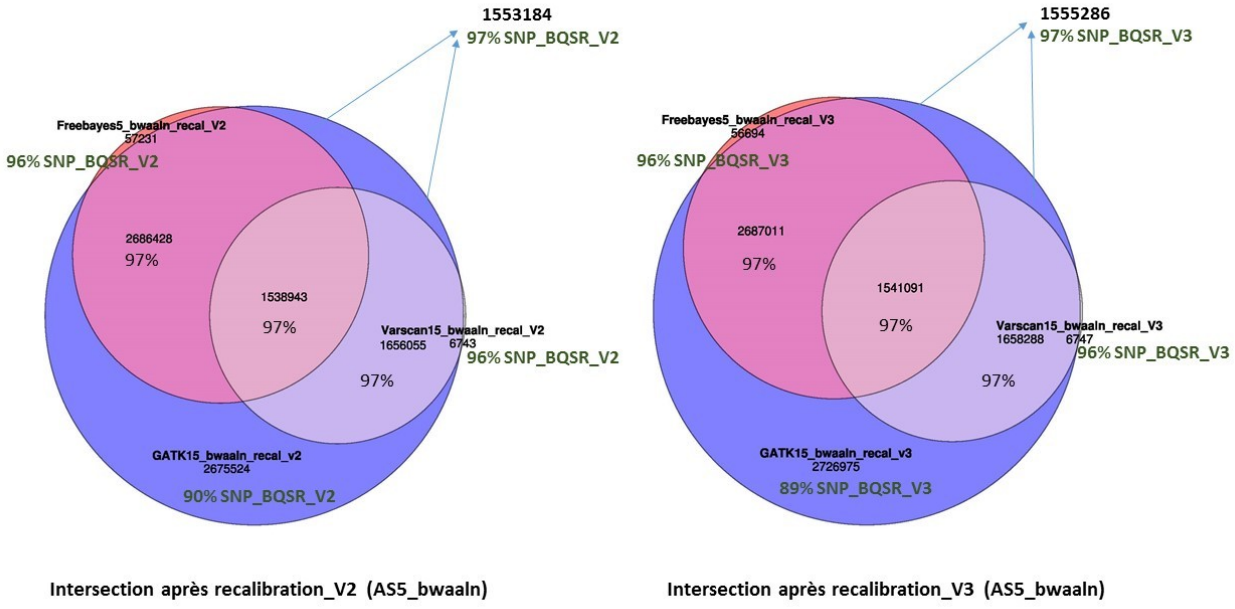


Figure 68: Comparaison du comptages des SNPs après recalibration du score de qualité des bases en utilisant soit l'ancien fichier BQSR de référence (V2-à gauche) soit le nouveau fichier BQSR de référence (V3-à droite).

VI) Effet du logiciel d'alignement des séquences d'ADN BWA sur l'appel de SNPs :

Nous avons exploré la différence des algorithmes 'mem' et 'aln' du logiciel BWA (Li & Durbin, 2009a) sur le nombre de SNPs appelés. Bien que les deux partagent des fonctionnalités similaires telles que la prise en charge de la lecture longue et l'alignement fractionné, BWA-mem, est un algorithme d'alignement local, c'est à dire qu'il tolère la présence de bases non alignées aux extrémités des fragments séquencés. Bwa aln, par contre, effectue des alignements globaux, c'est à dire que les deux extrémités du fragment doivent être alignées sur le génome de référence. BWA-mem a de meilleures performances pour les lectures Illumina de 70-100bp. BWA-aln induit de faibles taux d'erreur pour les séquences courtes alors qu'elles sont plus élevées pour bwa mem (Oliva et al., 2021). Après avoir aligné les séquences du génome de l'aurochs anatolien en utilisant BWA-mem et BWA-aln, j'ai appelé les SNPs avec GATK ayant un score qualité de base égale à 15 et une DP entre 3 et 14.

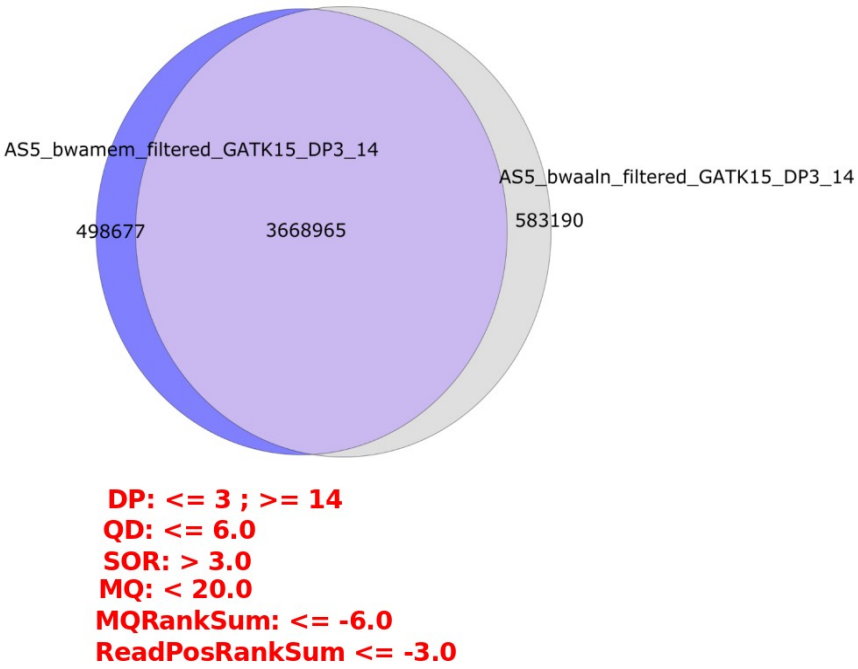


Figure 69: Diagramme de Venn du nombre de SNPs appelés par GATK et filtrés selon les seuils de filtres indiqués selon l'outil d'alignement des séquences d'ADN (bwa-mem en bleu et bwa-aln en gris).

Le comportement de l'algorithme d'appel de SNPs varie en fonction de l'algorithme d'alignement (Figure 69). Nous appelons plus de SNPs avec BWA-aln qu'avec BWA-mem. Le chevauchement du nombre de SNPs est significativement important mais il existe quand même un nombre non négligeable de SNPs qui ne sont appelés que par bwa-aln ou bwa-mem. Pour comprendre ce comportement, nous avons étudié le taux de transitions qui était presque le même indépendamment de l'outil d'alignement ainsi que le taux hétérozygotie en fonction de la couverture des SNPs appelés quand les séquences sont mappées avec bwa-aln et de ceux appelés quand les séquences sont alignées avec bwa-mem ainsi que les SNPs chevauchants.

1) Distribution de la couverture des SNPs obtenus en fonction de l'outil d'alignement des séquences d'ADN :

A partir des fichiers «vcf» obtenus en appelant les SNPs avec GATK en autorisant un score qualité de base égale à 15, j'ai déterminé le nombre de SNPs obtenus pour chaque couverture qui sont soient spécifiques à l'outil d'alignement de séquences bwa-mem, soient spécifiques à l'outil d'alignement de séquences bwa-aln ou communs aux deux.

Comme le diagramme de Venn l'avait montré, les SNPs communs sont les plus abondants et bien couverts avec 3 à 14 lectures couvrant le SNP et une chute du nombre de SNPs vers les couvertures les plus élevées. Par contre, nous avons observé que la couverture des SNPs spécifiques à bwa-aln et bwa-mem est différente. En effet, avec bwa-mem nous récupérons moins de SNPs à couverture faible (entre 3 et 6) et plus de SNPs à couverture plus élevé (à partir de 10), contrairement à bwa-aln, avec lequel nous récupérons plus de SNPs à couverture faible (entre 3 et 6) et moins de SNPs à couverture élevée (à partir de 10). Les SNPs à couverture élevée étant susceptibles d'être enrichis en lectures mésattribuées à la position génomique considérée, ceci est un indice que bwa mem pourrait induire plus d'erreurs d'appel de SNPs (Figure 70).

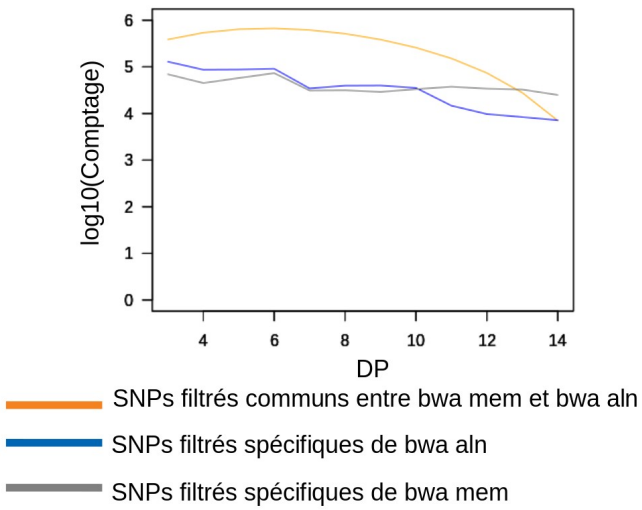


Figure 70: Distribution de la DP des SNPs obtenus soit par bwa-mem soit bwa-aln ou communs.

2) Taux d'hétérozygotie en fonction de l'outil d'alignement des séquences d'ADN :

Avec l'outil vcf-stats de vcftools, nous avons déterminé la quantité et la qualité des SNPs appelés. Nous obtenons plus de SNPs homozygotes alternatifs AA avec bwa-mem et plus de SNPs hétérozygotes référence-alternatif RA avec bwa-aln (Tableau 7). Sachant que bwa mem augmentait le nombre de SNPs avec une DP élevée, le fait qu'il y ait plus de SNPs homozygote alternatifs suggère que les lectures supplémentaires possiblement alignées de manière illégitime n'amènent pas particulièrement une hétérozygotie supplémentaires.

Par contre, les lectures courtes positionnées exclusivement par aln introduisent plus de variabilité. S'agit-il d'une variabilité authentique manquée par bwa mem qui tendrait à défavoriser les allèles alternatifs lorsque les séquences sont courtes causant un biais de référence qui est une préoccupation importante dans le domaine de la paléogénomique (Günther & Nettelblad, 2019)? Ou bien bwa aln tend-il à favoriser la prise en compte de lectures avec des erreurs?

	AS5_SNPs_spécifiques_bwaaln_filtrés_GATK15_DP3_14_complement	AS5_SNPs_spécifiques_bwamem_filter_GATK15_DP3_14_complement	AS5_SNPs_communs_bwamem_aln_filtrés_GATK15_DP3_14_complement
Comptages des SNPs hétérozygotes RA	427739	324156	1758612
Comptage des SNPs homozygotes AA	1394	897	2303
Comptages des SNPs hétérozygotes AA	19	21	26
Nombre de SNPs hétérozygotes	427758	324177	1758638
Nombre total de SNPs	429152	324198	1758664

Tableau 7: Comptages des SNPs homozygotes et hétérozygotes obtenues en fonction de l'algorithme d'alignement de séquences d'ADNa sur le génome de référence.

Pour explorer ces aspects, nous avons étudié le taux d'hétérozygotie des SNPs spécifiques à bwa-mem, spécifiques à bwa-aln ou les SNPs communs en fonction de la DP (Figure 71).

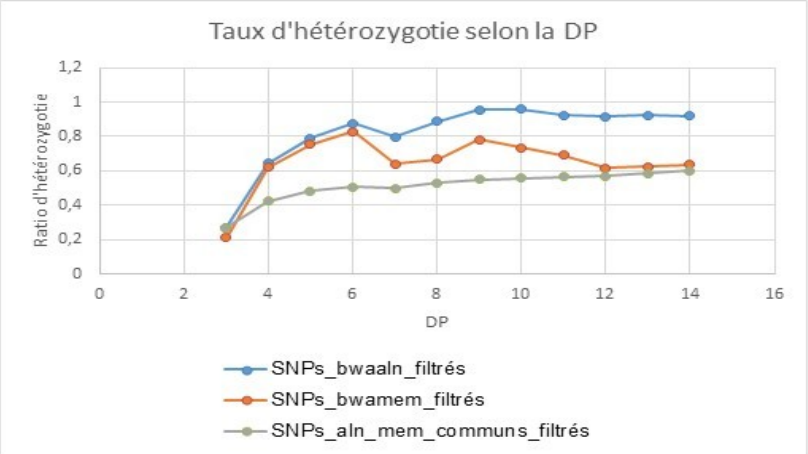


Figure 71: Taux d'hétérozygotie selon la couverture (DP). La couleur bleue correspond aux SNPs obtenues spécifiquement quand les séquences génomiques ont été alignées avec bwa-aln, la couleur orange correspond aux SNPs obtenues spécifiquement quand les séquences génomiques ont été alignées avec bwa-mem, la couleur grise correspond aux SNPs obtenus avec les deux outils d'alignement.

Nous avons trouvé que les SNPs communs ont un taux d'hétérozygotie entre 0,2 et 0,6, approchant du plateau de 0.6 à partir d'une DP de 6. Ces SNPs représentent la plupart des SNPs appelés (77% des SNPs totaux). Les SNPs spécifiques de bwa-aln (12% des SNPs totaux) ont des taux d'hétérozygotie plus élevés qui s'approche de 1 à partir d'une DP de 6. Les SNPs spécifiques de bwa-mem (10 % des SNPs totaux) ont des ratios hétérozygotie similaires à ceux spécifiques d'aln quand les SNPs sont couverts avec 4 à 6 lectures mais à partir d'une DP de 7, le ratio d'hétérozygotie diminue et se rapproche de ceux obtenus avec les deux algorithmes à partir d'une DP de 12.

Cela montre qu'avec l'outil bwa-aln, nous récupérons plus de SNPs hétérozygotes quelle que soit la DP. Cette différence pourrait être due à l'absence de soft clipping aux extrémités avec cet outil contrairement à ce que fait bwa mem qui clippe les bases mésalignées avec la référence aux extrémités des lectures.

3) Emplacement des allèles alternatifs et de référence sur les séquences d'ADN obtenues en utilisant les deux outils d'alignement de séquences :

L'obtention de taux d'hétérozygotie différents pour les SNPs obtenus seulement quand les séquences sont alignées au génome de référence soit avec bwa aln ou bwa mem, nous a poussé à déterminer les positions des allèles de référence et alternatifs pour explorer si la position sur la lecture est bien responsable de la différence entre les deux aligneurs. Une accumulation aux extrémités des séquences pourrait être un indice d'erreur parce que c'est là où les séquenceurs ont tendances à commettre le plus d'erreurs. En effet, les erreurs de séquençage au niveau des extrémités des fragments d'ADN pourraient être compensées pour toutes les lectures qui sont lues complètement sur les deux brins, dans notre cas, tous les fragments ayant une taille plus petite que 100pb.

De plus, les erreurs dues modifications diagénétiques des bases d'ADN sont plus problématiques car elles ont tendance de s'accumuler au niveau des extrémités des fragments d'ADN (A. W. Briggs et al., 2007). Pour cela, nous avons étudié la distribution du filtre GATK ReadPos RankSum (Figure 72). En effet, pour ce filtre, une valeur idéale est une valeur proche de 0 ce qui indique qu'il y a une faible différence quant à l'emplacement des allèles par rapport aux extrémités des lectures. Une valeur négative indique que l'allèle alternatif est trouvé plus sur les extrémités des séquences que l'allèle de référence et inversement une valeur positive indique que l'allèle de référence est trouvé plus sur les extrémités des séquences que l'allèle alternatif.

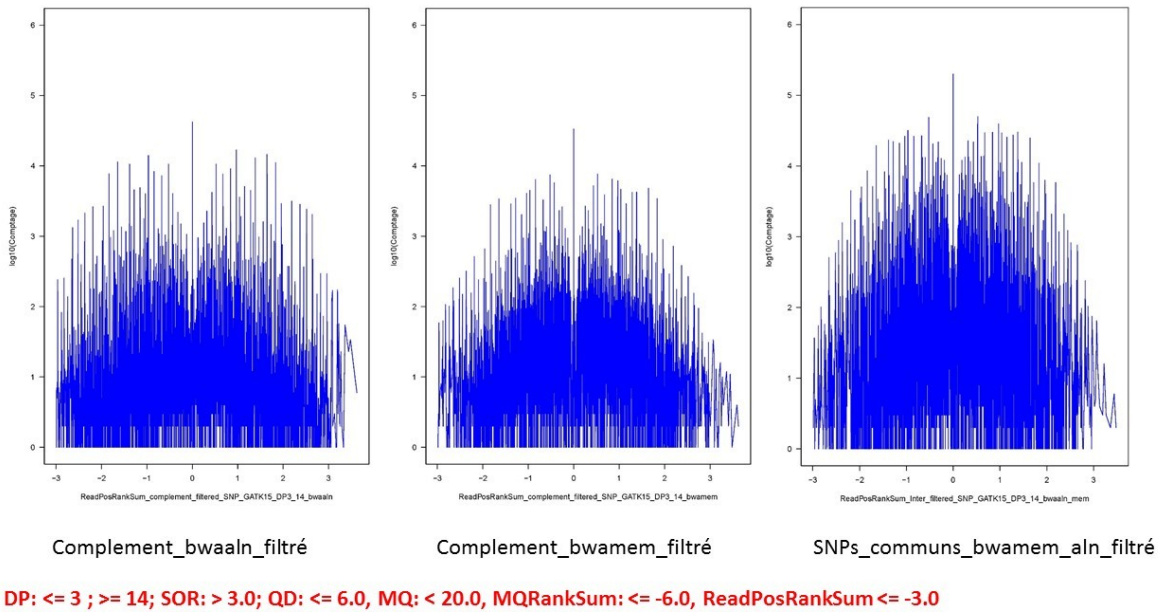


Figure 72: Distribution du filtre ReadPosRankSum de GATK pour les SNPs trouvés spécifiquement avec l'outil d'alignement bwa-aln, bwa-mem ou les SNPs communs.

La distribution des valeurs du filtre ReadPosRankSum diffère selon l'algorithme de cartographie. La distribution des SNPs appelés aussi bien après cartographie avec l'un ou l'autre des algorithmes est la plus enrichie en SNPs avec une valeur proche de 0, se traduisant par la forme la plus pointue de distribution (à droite) indiquant le moins de biais de position des SNPs appelés. Les SNPs appelés exclusivement après cartographie avec l'un ou l'autre algorithme sont plus distribués avec des valeurs s'éloignant du zéro, donnant des profils plus aplatis. L'effet est toutefois moindre pour les SNPs spécifiques de bwa mem que de bwa aln. Ceci montre que les SNPs appelés uniquement lorsqu'un des algorithmes de cartographie sont enrichis en SNPs plus douteux, l'enrichissement étant plus important pour les lectures cartographiées avec bwa aln. Il est probable que la différence entre mem et aln soit liée au fait que les SNPs aux extrémités des fragments sont clippés avec bwa mem et pris en compte avec une pénalité avec bwa aln. De tels SNPs sont susceptibles d'être enrichis en faux positifs correspondant à des transformation diagénétique et des erreurs de séquençage.

VII) Appel de SNP simultané sur les données modernes et les données anciennes :

Nous avons travaillé sur le téléchargement de près de 500 génomes bovins modernes de toutes les races bovines domestiques. Un tel panel représente la diversité génétique existante et détermine les SNPs caractéristiques des bœufs domestiques, des zébus et de quelques bovins plus éloignés : bisons européens et américains, gaur, gayal et yack. On a travaillé aussi à élargir le jeu de données anciennes pour couvrir une large période historique et une large distribution géographique, Les deux ensembles vont nous permettre de suivre le processus de domestication à haute résolution. L’appel des SNPs simultané pour les échantillons anciens et modernes est un exercice délicat car la qualité des génomes obtenus diffère, les génomes anciens ayant des biais particuliers: petites taille des fragments d’ADN, difficulté de couverture de région riches en AT, ou dupliquées, et transformation diagénétique des séquences.

Nous avons choisi d’effectuer un génotypage joint parce que, comme discuté au paravent, nous appelons tous les SNPs c’est à dire nous obtenons plus de SNPs RR (Référence-Réference) car contrairement à l’appel de SNPs, sur un seul échantillon avec lequel on appelle seulement les SNPs alternatifs, différents de la référence, donc nous obtenons que des SNPs RA ou AA, nous appelons tous les SNPs qui ont un allèle A dans au moins un des individus génotypés simultanément. Le taux hétérozygotie dépend alors de l’ensemble d’individus génotypés car si nous incluons des individus d’une autre espèce, le taux des SNPs RR va augmenter entraînât ainsi la diminution du taux hétérozygotie.

Pour permettre l’obtention d’un panel de SNP qui représente la diversité génétique moderne et ancienne, on a travaillé sur les conditions de traitement bioinformatique pour appeler les SNPs à partir d’échantillons anciens. Il y a deux types de stratégie pour l’appel de SNPs.

1) Appel ciblé de SNPs spécifiques :

On peut dans un premier cas appeler uniquement les SNPs sur les positions connues et fiables de variabilité présente dans les séquences modernes. Ce type d'appel génère un fichier vcf (Variant call format) où seulement les positions des variants sont gardées. Cette approche est simple mais si on se focalise seulement sur l'appel des SNPs chevauchants les SNP modernes caractérisés sur des puces, il ne permet pas d'identifier les SNPs présents dans les séquences anciennes qui auraient été perdus lors de la domestication et lors des sélections au cours du temps. L'approche simple nous prive donc d'une variabilité génétique particulièrement intéressante pour suivre le processus de domestication au cours du temps.

2) Appel de SNPs de Novo :

Pour permettre la récupération des SNPs spécifiques des échantillons anciens qui ne sont pas présents dans les échantillons modernes, je me suis donc intéressée à évaluer les approches qui permettrait d'appeler ces SNPs, c'est-à-dire de faire de l'appel de novo des séquences anciennes. L'approche utilisée pour faire l'appel de novo consiste à faire un génotypage joint de toutes les séquences ensemble. La difficulté est que les séquences anciennes et les séquences modernes ont des propriétés très différentes, les génomes anciens ont généralement des couvertures plus faibles, des lésions diagénétiques et des régions moins bien couvertes, du fait des fragments courts, les régions riches en AT sont sous-représentés car elles se dénaturent en cours de préparation et de constructions des banques. Les risques sont donc élevés d'avoir un taux de faux positifs plus élevé dans les séquences anciennes que modernes. Finalement, il y a beaucoup plus de génomes modernes disponibles que de génomes anciens, donc lors du génotypage joint, les SNPs provenant des génomes anciens peuvent être mal pris en compte. J'ai donc essayé d'évaluer les méthodes de traitement bioinformatique des séquences liées à l'appel de SNP afin de minimiser l'appel de SNP faux positifs. Je me suis focalisée pour cela sur le premier génome ancien bovin que nous avons produit pour avancer en attendant que les autres génomes que j'ai produit au cours de ma thèse soit analysés.

L'avantage de travailler sur un seul génome et de pouvoir faire une analyse fine de l'importance des biais pour chaque paramètre et de bien comprendre comment sont produits ces biais en fonction des conditions de traitement des données. Il faudra toutefois que nous utilisions ces acquis avec discernement lors du génotypage joint car nous devons anticiper que les biais peuvent varier entre les deux approches de génotypage. Les biais du génotypage joint sont en plus susceptibles de varier selon la distribution des séquences analysées simultanément. Par exemple, pour favoriser l'identification de SNPs absents des bœufs modernes mais présent dans les aurochs anciens, en particulier les aurochs les plus anciens, précédant le dernier maximum de glaciation, que l'on anticipe être les plus différents, nous avons choisi d'inclure une bonne représentation de bovins modernes plus divergents, gaur, gayal, yack et une bonne représentation des zébus. Nous avons ainsi inclus des bisons européens et américains car nous avons aussi étudié les génomes de bisons anciens.

Il a fallu réaliser une étude minutieuse pour identifier les conditions qui nous ont a paru optimales, sachant qu'il est délicat d'identifier quand on élimine plus de faux positifs que l'on ne perd des SNPs anciens authentiques absent dans les génomes modernes (Faux négatifs) lorsque l'on ne sait pas a priori si les SNPs détectés spécifiquement sur les génomes anciens sont des SNPs authentiques ou artéfactuels. Ce travail était nécessaire pour disposer des meilleurs outils pour détecter les régions génomiques sélectionnées au cours du processus de domestication.

VIII) Discussion :

Nous avons réalisé plusieurs tests pour réussir à trouver un compromis entre les conditions d'appel de SNPs par les 3 algorithmes d'appel de SNPs. Nous avons trouvé qu'un contrôle qualité de base égale à 15 pour GATK et Varscan et une couverture égale à 5 pour Freebayes permettent d'avoir le maximum de SNPs partagés entre les différents algorithmes d'appel de SNPs et une bonne cohérence. Nos tests ont montré que la recalibration réduit les erreurs produites lors de la construction des banques et leur séquençage. Cette réduction entraîne une diminution du nombre de SNPs appelés et élargie le % de SNPs partagés avec les SNPs de référence. La filtration réduit plus le nombre de SNPs que la recalibration. La filtration a un grand effet sur l'augmentation du % des SNPs partagés avec les SNPs connus. L'exploration de la distribution des filtres nous a permis de modifier les seuils des filtres et d'augmenter le chevauchement des 3 algorithmes d'appel de SNPs.
Le test sur les algorithmes d'appel de SNPs nous a amené à choisir GATK qui semble être le meilleur outil d'appel de SNPs. Il a le plus faible taux de faux négatifs des trois algorithmes et il permet d'appeler la grande majorité des SNPs appelés par au moins un ou les deux autres outils, Varscan et Freebayes. Suite à cette étude comparative de la performance de 3 algorithmes d'appel de SNPs afin d'identifier celui qui permet d'appeler le maximum de SNPs ancestraux qui n'existent plus dans les données modernes, nous utilisons maintenant l'outil GATK. Le fichier BQSR de référence mis à jour nous a permis de récupérer plus de SNPs probablement authentiques avec une meilleure qualité de génotypes.

L'étude de distribution de filtres GATK nous a permis d'améliorer la qualité des SNPs appelés par les différents algorithmes d'appel de SNPs et de choisir entre les 3 algorithmes testés. En effet, nous avons réussi à augmenter le chevauchement entre les 3 algorithmes de façon que la majorité des SNPs appelés par Varscan et Freebayes sont appelés avec GATK. A la suite de ce résultat, nous avons décidé d'utiliser GATK pour appeler les SNPs car c'est l'algorithme le plus pertinent. De plus, une comparaison entre les outils d'alignement de séquences n'avaient pas le même effet sur le nombre et la qualité des SNPs. Nous sommes orientés vers l'utilisation de bwa aln car bwa mem pose plus de problème du fait que les SNPs aux extrémités des fragmentes sont clippées et sous représente alors les fragments courts qui représentent la majorité des fragments d'ADNa. Il a été montré que le paramétrage de bwa mem proposé par (Xu et al., 2021) diminue la précision de l'alignement par rapport à BWA-mem en utilisant les paramètres par défaut pour le séquençage des lectures inférieures à 70 bases, qui sont particulièrement abondantes dans les échantillons d'ADN anciens. De plus, il a été montré que bwa aln réduit le biais de génome de référence qui est un biais liés à l'alignement qui peut gonfler les faux positifs et est particulièrement problématique pour les études d'ADN anciennes (Oliva et al., 2021).

Il était nécessaire de bien optimiser les paramètres d'appel de SNPs pour récupérer le maximum de SNPs ancestraux réellement positifs. En effet, les panels de SNPs modernes et anciens vont nous permettre d'effectuer une analyse comparative. Nous les utilisons pour réaliser une analyse en composante principale ACP en comparant directement la variabilité des spécimens anciens et modernes. Quand on ne fait que l'appel de SNPs sur la variabilité moderne connue, on peut analyser les génomes anciens en les projetant sur la variabilité moderne, mais dans ce cas, la géométrie de l'ACP sera déterminée uniquement par les génomes modernes et on ne fera que s'intéresser à comment se positionnent les anciens par rapport à ces modernes, c'est à dire visualiser les échantillons fossiles dans le contexte de la variation moderne.

Pouvoir intégrer la variabilité ancienne dans ce type d'analyse est déjà un gain. Toutefois les données génétiques riches ne sont qu'imparfaitement décrites avec les deux premières composantes principales, la variance des données étant classiquement distribuées sur un très grand nombre de composantes, les deux premières ne représentant que quelques % de la variance globale. Pour prendre en compte les autres composantes, on peut aussi utiliser une méthode, appelée UMAP, qui peut intégrer toutes les composantes pour à nouveau représenter l'ensemble de la variance en deux dimensions. UMAP devrait être une manière très puissante d'analyser les différences entre génomes anciens et modernes avec un appel des SNPs de novo. Nous allons aussi étudier les flux génétiques entre les populations anciennes et actuelles en utilisant différents tests statiques (Fst, F3, F4, Statistique D). Le test Fst compare les fréquences allélique entre et au sein de la population. Des valeurs élevées de Fst indiquent une différenciation marquée entre les populations. Des valeurs plus petites indiquent que les populations comparées sont homogènes. Les tests f3, f4 et D sont des tests qui fournissent des informations sur les liens entre les génomes et des indications de mélange et sur la la directionnalité du flux génétique même si les événements de flux génétiques se sont produits il y a des centaines de générations. Une des questions posées concernera les introgressions des populations sauvages vers les premières populations domestiquées et comment ce processus a contribué à l'évolution des populations domestiquées. Ce panel de SNPs va aussi nous permettre de mettre l'accent sur les schémas de sélection positive.

Chapitre IV : Analyse phylogéographique des *Bovina* :

I) Analyse bayésienne des mitogénomes des genres *Bos* et *Bison* :

Après avoir capturé les fragments d'ADN hybridant avec le mitogénome bovin, les séquences correspondantes ont été alignées sur ce mitogénome afin de reconstituer des mitogénomes complets. Dans le cas où les séquences consensus obtenues étaient sensiblement différentes du mitogénome utilisé pour la capture, nous avons réaligné les séquences capturées sur un ou plusieurs mitogénomes les plus proches des séquences consensus obtenues. Cette démarche itérative était nécessaire car les fragments d'ADN qui pourraient avoir été capturés malgré la divergence avec la séquence de référence pourraient ne pas être cartographiés sur la séquence lorsqu'elle contenait un nombre trop important de divergences à cause des seuils utilisés.

Pour les séquences les plus divergentes pour lesquelles nous ne disposions pas d'une séquence de référence connue, nous avons utilisé deux algorithmes de cartographie des lectures, un global (bwa aln) et un local (bwa mem) qui peut faire un « soft clipping » des lectures aux extrémités, ce qui peut permettre de récupérer quelques nucléotides en particulier dans les zones d'extrême divergence dans la région hypervariable. Dans certains cas, nos avons dû recourir de manière itérative à cette stratégie pour établir une nouvelle séquence consensus qui a servi à recartographier les lectures. Certaines des séquences ont aussi pu être complétées en utilisant les lectures à haute profondeur qui visaient à obtenir les génomes nucléaires des individus correspondants. Ceci a permis de s'affranchir de la contrainte de la capture qui ne fonctionne plus lorsque les séquences sont trop divergentes, surtout si elles sont riches en AT. Une fois que les séquences consensus ont été obtenues, elles ont ensuite été alignées en utilisant un programme d'alignement multiple « Muscle » (Edgar, 2004) qui offre un compromis satisfaisant entre qualité de l'alignement et vitesse d'exécution en utilisant près d'un millier de mitogénomes complets. Toutes les phases de réalignement des lectures, d'établissement d'une séquence consensus d'alignements multiples ont été réalisés en utilisant l'environnement intégré du logiciel Geneious (Geneious prime 2021) (Kearse et al., 2012). La démarche utilisée pour se faire une première idée de la phylogénie au fur et à mesure de l'acquisition des séquences était de réaliser un arbre phylogénétique avec des logiciels de maximum de vraisemblance, en utilisant des algorithmes qui tournent rapidement. Le premier choix consistait à utiliser le logiciel «Fastree » (M. N. Price et al., 2010). Dans un deuxième temps, pour obtenir un peu plus de précision dans la phylogénie reconstituée, nous avons utilisé le programme « RaxML » (Stamatakis, 2014). Ce programme est plus lent, mais il permet de tester la robustesse des nœuds avec une approche de bootstrap (encore plus lente) ce qui est utile lorsque l'on souhaite avoir une confirmation des phylogénies avec deux approches différentes.

Je ne présenterai qu'une seule analyse avec cette approche, car nous avons surtout utilisé l'approche bayésienne pour obtenir des estimations des âges des nœuds et des évolutions de taille des populations. Toutefois, toutes les topologies qui sont robustes avec l'approche bayésienne sont aussi observées avec l'approche de maximum de vraisemblance.

L’analyse bayésienne a été réalisée en utilisant le programme BEAST (Bayesian Evolutionary Analysis by Sampling Trees) (Drummond & Rambaut, 2007). Les mitogénomes ont été partitionnés en quatre régions : 1) la partition de la région hypervariable ; 2) la partition de la région codant pour les ARNs ribosomaux et de transfert ; 3) la partition de la première et la deuxième position des codons ; et 4) la partition de la troisième position des codons. Ceci permet de mieux contrôler les différences de taux d’évolution des séquences qui ne sont pas les mêmes dans ces différentes régions soumises à des contraintes évolutives différentes.

La première étape a consisté à identifier les modèles d’évolution des séquences qui conviennent le mieux à chacune des partitions. Pour cela, nous avons utilisé le logiciel « ModelTest » (Darriba et al., 2020) qui construit des arbres de maximum de vraisemblance avec un grand nombre de modèles d’évolution des séquences, pour trouver le modèle qui décrit la topologie qui a le plus de vraisemblance. Plusieurs approches d’évaluation du meilleur modèle peuvent être utilisées : BIC (Bayesian Information Criteria), AIC (Akaike Information Criteria), et AICc (Corrected Akaike Information Criteria). Nous avons recherché les modèles d’évolution qui avaient le poids le plus élevé pour l’ensemble des trois critères, ce n’était pas toujours le même qui était le premier pour chaque critère. Le modèle retenu a donc été le modèle TPM1uf +I+G4 (I=invariant, G4=distribution gamma avec quatre catégories) pour la région hypervariable et les deux premières positions des codons. Le modèle TrN+I+G4 a été retenu pour la partition des ARNs et celle pour la troisième position des codons.

Nous avons appliqué une horloge moléculaire stricte, c’est-à-dire que nous avons fait l’approximation que le taux d’évolution des séquences de chaque partition ne changeait pas au cours du temps sur l’échelle évolutive considérée. En effet, le mode de vie des bovinés et le temps de génération peut être considéré suffisamment homogène pour que cette approximation soit valide.
Nous avons utilisé un modèle d’évolution de populations reposant sur le « Bayesian skyline ». Nous avons soit exploré différents nombres de populations allant de 5 à 10 pour déterminer celles qui permettaient d’obtenir la meilleure exploration des paramètres (effective sample size, ESS, maximal pour la plupart des paramètres estimés). Dans ce cas, la prise en compte de cinq populations donnait les meilleures valeurs. Nous avons aussi utilisé le modèle de population du « Bayesian skyline » étendu (EBSP) qui évalue le nombre de populations à partir des données et qui tend à produire une trentaine de populations. Les résultats finaux étaient très similaires. Le modèle EBSP est celui qui permet de reconstituer l’évolution des tailles effectives des populations au cours du temps. C’est aussi celui qui donne un petit peu plus d’ESS pour un nombre équivalent d’exploration. Pour définir les « priors », comme nous souhaitions mettre le maximum de poids sur l’âge des séquences anciennes incluses dans la phylogénie, nous avons choisi une distribution assez souple des autres paramètres.
Pour cela, nous avons utilisé pour tous les paramètres une distribution log normale et nous avons défini un intervalle de confiance de 95% très large, aussi bien pour estimer les tailles effectives des populations que les taux d’évolution des séquences.

Lors des analyses réalisées précédemment sur les mitogénomes des bisons (Massilani et al., 2016) il avait utilisé comme « priors » les taux d'évolution des différentes partitions en se basant sur ceux connus pour les mitogénomes humains. Les valeurs postérieures qui avaient été obtenues après analyse étaient à peu près cinq fois plus élevées, ce qui correspond bien à la différence de temps de génération entre l'humain et les bovins. Cette première analyse a montré que l'approche d'utiliser des séquences anciennes datées permet de sortir de la valeur médiane indiquée en « prior » avec une distribution log normale large et suffisamment de séquences anciennes. Ici, nous sommes partis des valeurs estimées lors de cette première analyse pour définir les « priors ». Pour les tailles des populations, nous avons utilisé une distribution log normale avec un intervalle de confiance très, très large variant de 150 à 1 milliard d'individus pour ne pas contraindre ce paramètre. Nous n'avons pas imposé un âge à la racine et avons laissé le programme l'estimer librement. Pour les autres paramètres de l'évolution des séquences, nous avons laissé les options proposées par défaut.

L'algorithme de Markov Chain Monte Carlo implémenté dans BEAST, a été utilisé pour estimer la distribution postérieure de chaque paramètre d'intérêt. Nous avons effectué une vingtaine d'analyses en explorant 20 millions d'itérations pour chaque analyse. Logcombiner a été utilisé pour assembler et comparer les différentes combinaisons d'arbre proposées et l'arbre le plus crédible a été estimé ainsi que la taille médiane des nœuds avec TreeAnnotator . Les résultats ont été transférés pour analyse de l'exploration des paramètres sur le logiciel Tracer (http://tree.bio.ed.ac.uk/software/tracer/). Les arbres ont été visualisés avec FigTree (v1.4.4) (http://tree.bio.ed.ac.uk/software/software/figtree/).

Pour réaliser l'analyse phylogénétique, nous avons utilisé 950 mitogénomes complets provenant d'une part de séquences mitochondriales téléchargées à partir du site du NCBI et d'autre part les séquences produites au laboratoire ainsi que celles issues de l'alignement des lectures téléchargées pour constituer une base de données locale de génomes bovins actuels. Nous avons inclus des mitogénomes des espèces suivantes : *Bison bison*, *Bison bonasus*, *Bison priscus*, *Bos indicus*, *Bos grunniens*, *Bos gaurus*, *Bos javanicus* et *Bos frontalis* correspondant à 280 séquences dont 89 anciennes. 45 de ces séquences anciennes correspondent à *Bison priscus* avec 36 provenant d'études antérieures et 9 de notre laboratoire dont j'en ai contribué quatre. Il y aussi 44 séquences anciennes correspondant aux deux clades de *Bison bonasus*, 16 provenant de (Soubrier et al., 2016), 11 provenant de (Massilani et al., 2016), et 17 produites par moi-même. 73 séquences appartiennent à *Bos indicus*, et aucune de ces séquences n'est ancienne. 594 séquences appartiennent aux différents haplogroupes de *Bos primigenius* et de ses descendants et parmi celles-ci 191 sont anciennes et sont datées de 100 ans à 130 000 ans. 54 de ces séquences anciennes proviennent d'études antérieures et notre laboratoire en a contribué 136 nouvelles dont 98 ont été produites par moi-même.

L'inclusion de 280 séquences mitogénomiques anciennes sur un total de 950 permet d'avoir une bonne estimation des paramètres d'évolution des séquences sur l'ensemble de la phylogénie en se basant sur l'âge des échantillons anciens. La topologie globale de l'arbre est représentée sur la figure 73.

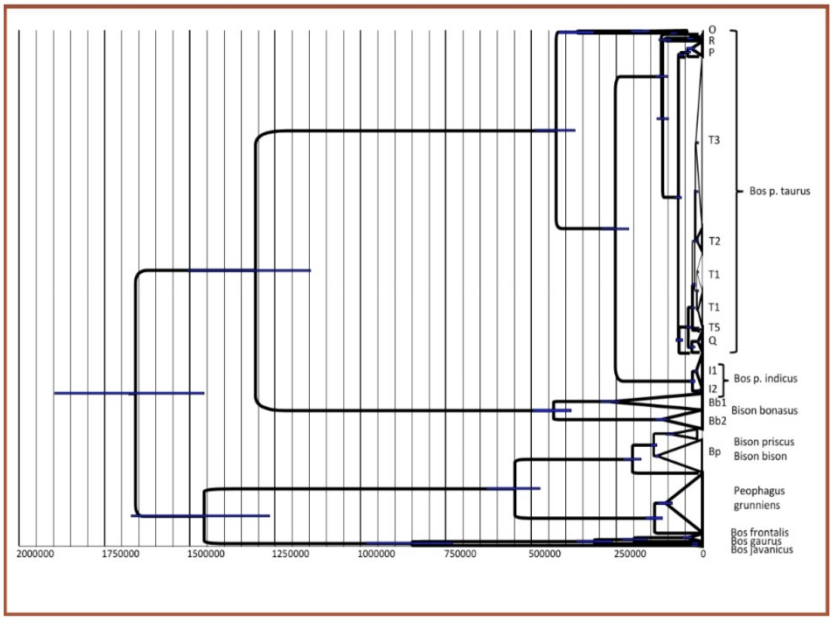


Figure 73 : Arbre phylogénétique bayésien des mitogénomes complets des spécimens d'aurochs, de bovins et de bisons européens anciens et modernes. L'échelle temporelle en années AP est représentée avec les lignes verticales, tandis que les traits bleus indiquent l'intervalle de confiance de 95% de l'âge estimé des nœuds.

Cette analyse bayésienne nous a permis d'estimer l'âge des nœuds de divergences des différentes branches de l'arbre présentées dans l'arbre phylogénétique (tableau 8).

Noeud du clade	Date de divergence	95%HPD-low	95%HPD_hi	Achilli et al 2009	Wu et al 2018	CI Wu et al
All	1515000	1400000	1650000		1181000	807000-1794000
Bp-Gru-Gay-Gau-Ban	1332000	1210000	1455000		1125000	794000-1718000
Gay-Gau-Ban	775000	698000	856000			
Bp-Gru	504000	450000	557000		880000	558000-1543000
Gay-Gau	287000	252000	325000			
All_Bp	193190	173470	213860			
Bt-Bb	1200000	1090000	1307000		1181000	807000-1794000
Bb1-Bb2	405000	366000	445000			
Bb1	251000	234000	277000			
Bb2	111000	97000	127000			
All_Bt	393000	357000	433000			
O	337000	302000	371000			
O1	170000	148000	192000			
O2	80000	70700	90000			
Bt-Bi	232000	205000	262000	335000	215700	42000-844000
Bi	25000	18000	33000			
I1	15700	11700	20500			
I2	8500	6600	10800			
C-RPQT	108000	95000	123000			
RPQT	10400	92000	118000	137000		
R3-R1R2	98000	85000	112000			
R1-R2	31000	22000	40000	43000		
R1	3000	1000	6000			
R2-Morocco	9000	8900	10600			
R2-Italy	400	0	2300			
P-QT	62000	55000	71000	71000		
LP4-P	54500	47000	62000			
Uz3-P	38700	31000	47000			
P	29000	22000	41000			
LF1284-QT	60000	52000	67000			
Q-T	37000	30000	45000	48000		
Gyu2-Q	28000	20000	35000			
Q	12900	10300	16500			
T	26000	20000	32000	18000		
HL9-T123	22500	18000	27000			
T1-T2-T3	17000	13000	21000	13000		
T1	13200	10500	16500			
T2	13200	10300	16900			
T3	12600	10600	15400			
Sv2-T5	8700	7600	11000			
T5	8200	6000	10000			

Tableau 8 : Âges estimés par l'analyse bayésienne des nœuds des différentes branches des lignées mitochondriales des Bovina anciens et modernes avec les intervalles de confiance à 95% (95% HPD). Les données que nous avons obtenues sont comparées avec celles obtenues par (Wu et al., 2018) à partir des génomes (avec l'intervalle de confiance CI) et par (Achilli et al., 2009) avec une analyse de maximum de vraisemblance à partir des mitogénomes .

Je vais très brièvement discuter la phylogénie globale et je discuterai ensuite plus en détail des points en traitant chaque clade individuellement. Le tableau 8 présente aussi les datations estimées des nœuds et les compare avec l'estimation des âges des nœuds effectuée par l'approche de maximum de vraisemblance à partir des mitogénomes actuels (Achilli et al., 2009; Bonfiglio et al., 2010) ainsi qu'à partir de génomes nucléaires (Wu et al., 2018). On peut voir que nos estimations de dates établies à partir des mitogénomes anciens chevauchent les estimations des dates établies à partir des génomes actuels, particulièrement en ce qui concerne les nœuds anciens séparant les différentes espèces.

Par contre, les intervalles de confiance établis à partir des génomes sont très larges. Ces âges de divergence ont été estimés en utilisant une vitesse d'horloge obtenue il y a longtemps d'une manière qui n'est pas forcément très robuste. L'âge de la racine de tous les bovinés est estimé à partir des mitogénomes à 1,5 millions d'années (95% HPD 1,4 – 1,65) alors qu'à partir des génomes, (Wu et al., 2018) ont estimé 1,2 millions d'années avec 0,8 – 1,8 millions d'années d'intervalle de confiance. Bien que nos valeurs soient inclues dans leur intervalle de confiance, nous sommes plutôt dans la partie la plus âgée de leur intervalle de confiance. Ceci est peut-être dû aux différentes façons d'estimer l'horloge moléculaire qui est forcément plus arbitraire lorsqu'on ne dispose que de génomes modernes.

Toutefois, il est attendu que le mitogénome ancestral le plus récent soit en fait antérieur à la date de séparation définitive des espèces qui est reflétée dans les génomes. Ainsi, pour l'ancêtre commun des lignées de *Bison priscus* / *bison*, des yaks, des gayals, des gaurs et des bantengs nous estimons une date d'1,33 millions d'années (95% HPD 1,21 – 1,45) alors que les génomes indiquent 1,12 millions d'années (95% HPD 0,8 – 1,7). Par contre, la date de l'ancêtre de *Bison priscus* / *bison* et des yaks au niveau du mitogénome est plus récente : 0,5 millions d'années (95% HPD 0,45 – 0,56), alors qu'au niveau des génomes, la radiation serait plus ancienne : 0,88 millions d'années (95% HPD 0,56 – 1,54). Ceci suggère que la lignée mitochondriale des yaks pourrait provenir d'une introgression des *Bison priscus* postérieure à la date de séparation des populations. Les dates de séparation des taurins et des zébus sont beaucoup plus proches avec les deux estimations : 232 000 ans pour les mitogénomes (95% HPD 205 – 262 000) alors qu'elle est de 216 000 ans pour les génomes (95% HPD 42 – 844 000). A partir de cette radiation, on ne peut plus comparer avec les données génomiques, mais on peut comparer avec les estimations d'Achilli et al., 2009 qui appliquaient un taux constant estimé en n'utilisant que des séquences actuelles. Ce que l'on voit sur le tableau, c'est que les estimations des nœuds anciens par Achilli et al. sont plus anciennes que celles que nous estimons ici. Par exemple, Achilli date la divergence *B. taurus – indicus* à 335 000 ans ce qui paraît précéder un peu trop la date estimée à partir des génomes. De la même manière, toutes les dates des nœuds estimées par (Achilli et al., 2009) antérieures à 40 000 ans, sont plus anciennes que celles que nous estimons. Par contre, les estimations d'âges des ancêtres communs pour les radiations les plus récentes au sein des haplotypes T sont postérieures de quelques milliers d'années pour Achilli et al. que pour nous, ce qui aura des conséquences non-négligeables sur les interprétations en terme de domestication (voir plus loin).

La robustesse de l'arbre phylogénétique proposé est élevée en ce qui concerne la séparation des différents clades entre eux mais à l'intérieur des haplogroupes les supports de branches sont moins robustes, les supports bayésiens sont plus faibles (l'épaisseur du trait est proportionnel à la probabilité postérieure) et les différentes itérations bayésiennes ainsi que les analyses par maximum de vraisemblance peuvent donner des topologies variables. C'est particulièrement le cas pour les radiations récentes issues d'expansions rapides de populations après la domestication qui obscurcissent la phylogénie (génomes peu différents et multiples événements d'évolution parallèle).

II) Phylogénie bayésienne du genre *Bison* :

Dans le cadre de mon travail de thèse, j'ai complété l'analyse génomique des aurochs et de leur domestication par l'analyse comparative de l'évolution d'une espèce cousine non-domestiquée, le bison (Geigl & Grange, 2019 dans (Peters et al., 2019)). Ce travail se situe dans la continuité d'un travail déjà initié dans le laboratoire et ayant donné lieu à deux publications déjà décrites dans l'introduction (Massilani et al., 2016 ; Grange et al., 2018). En parallèle de mes efforts de produire des données génomiques sur les spécimens anciens, j'ai complété le corpus disponible en contribuant 17 séquences mitochondriales des lignées *B. bonasus* Bb1 et Bb2 et 4 mitogénomes du type *B. priscus*. Ceci m'a permis d'élargir des analyses précédentes sur la dynamique des populations au Pléistocène supérieur et à l'Holocène en Europe de l'Ouest.

1) Échantillons analysés :

Les caractéristiques des échantillons sont représentés dans le tableau suivant incluant les âges et les origines géographiques ainsi que les haplotypes mitochondriaux.

Échantillon fossile	Haplotype mitochondriale	Âge approximatif	Site archéologique	Origine	Cadre chronologique/contexte archéologique	% ADN endogène de banques génomiques
Coud10337	Bb1	130000	CoudoulousII	France	Pleistocène moyen	29.98%
Coud21675		130000	CoudoulousII		Pleistocène moyen	9.46%
Coud22195		130000	CoudoulousII		Pleistocène moyen	1.15%
LB93887		33250	La Berbie		Pléistocène supérieur	33.10%
LP1874		48000	Les Plumettes		Pléistocène supérieur	42.52%
NK64		42670	Bobenheim-Roxheim	Haute Vallée du Rhin	Pléistocène supérieur	38.51%
NK65		43460	Bobenheim-Roxheim			30.70%
NK67		44210	Bobenheim-Roxheim			12.36%
NK68		42820	Bobenheim-Roxheim			0.28%
NK69		42820	Bobenheim-Roxheim			42.39%
NK83		42060	Bobenheim-Roxheim			42.46%
NK84		42820	Bobenheim-Roxheim			36.72%
NK85		45000	Bobenheim-Roxheim			26.11%
NKP1400		45000				3.30%
Cha1	Bb2	4300	La Cha	France/Ain	~4300	93.13%
ZIN8937		200	Ancienne Prusse de l'Est?	Russie	163 ± 31 BP	56.89%
ZIN8943		300	Ancienne Prusse de l'Est?		243 ± 32 BP	77.49%
MENG22	Bp	42150	Bensheim	Haute Vallée du Rhin	Pléistocène supérieur	34.95%
NKP1401		32000				94.76%
SD1		43000	Siegsdorf	Bavière du Sud		4.46%
LB92182		23250	La Berbie	France/Dordogne		10.06%

Tableau 9: Échantillons anciens de bisons européens inclus dans notre étude et ajoutés aux échantillons analysés dans les études publiés antérieurement (Grange et al., 2018; Massilani et al., 2016).

2) Capture du génome mitochondrial de bisons anciens :

En se basant sur le résultat de séquençage aléatoire, nous avons criblé les meilleurs banques génomiques présentant un pourcentage d'ADN endogène supérieur à au moins 1%. Nous avons capturé le génome mitochondrial à partir d'au moins deux banques génomiques pour chaque échantillon dont chacune est construite à partir d'extrait d'ADN purifié soit avec le tampon 2M70 ou 5M40 mais nous avons aussi inclus les banques construites à partir d'extraits purifiés avec le tampon QG pour les échantillons de Coudoulous II. Après deux cycles d'enrichissement de chaque banque génomique par hybridation en solution avec les sondes d'ARNs qui couvrent tout le mitogénome d'un bovin européen, nous avons séquencé les fragments double brins enrichies en ADN mitochondrial sur la plateforme Illumina.

La capture ayant été faite avec des appâts synthétisés à partir du mitogénome bovin qui diffère par endroit du mitogénome de bison, il fallait que la stringence des conditions d'hybridation et de lavage ne soit pas trop élevée, une condition que nous avons cherché à satisfaire de toute façon même pour les aurochs dont certains pouvaient porter des mitogénomes eux aussi assez différents de celui qui a servi à produire les appâts. La couverture de certaines régions était toutefois moins bonne chez les bisons mais les séquençages des génomes complets ont permis de combler la plupart des trous. Dans les études antérieures, les trous de séquence, en particulier dans la région hypervariable la plus riche en A-T et la plus divergente avaient été comblés par amplification par PCR de cette région.

3) Relation phylogénétique mitochondriale des bisons anciens et modernes :

Notre étude phylogénétique basée sur le génome mitochondrial des *Bovina* a inclus 21 nouveaux mitogénomes complets de deux clades de *Bison bonasus* (Bb1 et Bb2) et de *Bison priscus*. Nous avons ajouté aux mitogénomes inclus dans l'étude publié par notre équipe en 2018 (Grange et al., 2018), des mitogénomes d'échantillons originaires de la France du Sud datant du Pléistocène moyen et tardif, des mitogénomes d'échantillons de la Haute Vallée du Rhin datant du Pléistocène supérieur, un spécimen français datant d'environ 4300 ans et deux échantillons du Musée de St. Pétersbourg datés de 200 à 300 ans et provenant probablement de la Prusse orientale de l'époque. Ces mitogénomes sont assignées aux haplogroupes mitochondriaux de *Bison bonasus* et de *Bison priscus* (comme présenté sur le tableau 9). Les relations phylogénétiques à l'échelle intra et interspécifique entre les échantillons modernes et anciens de tous les sous-groupes de la famille des *Bovina* sont représentées dans la figure suivante.

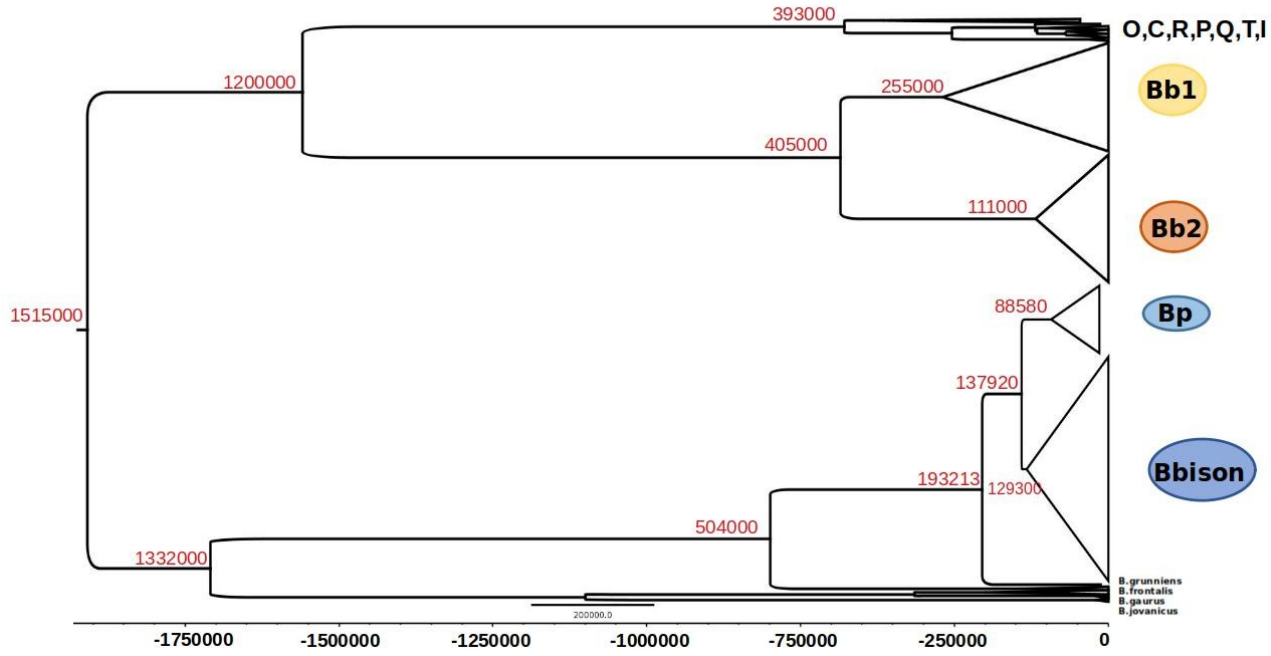


Figure 74: Arbre phylogénétique bayésien de la sous-tribu des Bovina montrant l'affinité mitochondriale entre les bisons européens Bb et les bovins taurins et l'affinité entre les bisons américains et les yaks (B. grunniens). Bb1 et Bb2 sont les 2 clades de B. bonasus, Bp est le clade de B. priscus, Bbison est le clade des bisons américains modernes.

Comme il a été observé précédemment (Verkaar et al., 2004; Zeyland et al., 2012), on voit ici que les mitogénomes des *B. bonasus* sont plus proches des mitogénomes taurins que ceux des *B. priscus* et *B. bison* tandis que les yaks sont plus proches des *B. priscus* et *B. bison* américains que ceux-ci ne sont proches des *B. bonasus* européens. Comme discuté (Grange et al., 2018), cette topologie s'explique mieux par un tri incomplet de lignées que par une introgression tardive chez l'ancêtre du *B. bonasus* provenant d'un ancêtre des aurochs, compte tenu des dates de divergence entre les bisons et les aurochs estimées à partir des génomes actuels (voir Introduction).

L'inclusion d'une plus grande diversité de séquences anciennes (279 mitogénomes anciens sur 950 totaux) et leur plus grande profondeur temporelle (comme celles pour les échantillons de Coudoulous II, les plus anciens inclus dans notre étude et remontant à environ 130 000 ans) permet d'affiner l'estimation bayésienne des taux d'évolution et de l'âge des nœuds. La tendance est à un léger ralentissement des taux d'évolution des nucléotides et à une légère augmentation de l'âge des nœuds par rapport à nos estimations précédentes (Grange et al., 2018). Je présente dans le tableau 10 les modifications des dates de radiation de différentes lignées de *Bison* entre notre étude actuelle et celle publiée en 2018 par notre équipe (Grange et al., 2018). La différence observée est toutefois minime car l'analyse de Grange et al., 2018 incluait déjà un *B. priscus* de plus de 100 000 ans et les 38 mitogénomes bovins déjà produits antérieurement par notre équipe (Tableau 10).

	Estimation d'age des noeuds en mille ans	
	Age (95% HPD)	
Noeuds	**Grange et al,2018**	**Etude actuelle**
Bovinae	1549 (1366-1736)	1551 (1400-1650)
Bos p.taurus/Bison bonasus	1273 (1117-1436)	1200 (1090-1307)
Bos grunniens/Bison priscus	536 (468-608)	504 (450-557)
B.priscus	203 (177-224)	193,19 (173,47-213,86)
B.bonasus	395 (343-445)	405 (366-445)
Bb1	214 (183-240)	255 (234-277)
Bb2	116 (98-130)	111 (97-127)

Tableau 10: Estimation des âges des nœuds à travers les analyses bayésiennes.

L'arbre des mitogénomes complets et cohérent avec les résultats obtenus par (Grange et al., 2018; Massilani et al., 2016) et la topologie globale est conservée. Nos résultats montrent que les lignées maternelles des ancêtres de *Bos primigenius* et de *Bison bonasus* se sont séparés il y a environ 1200000 ans (95% HPD 1090000-1307000 ans). Je discuterai les âges des autres nœuds au fur et à mesure en analysant l'ensemble des résultats dans chaque groupe.

a) *Bison bonasus* ou Bison européen :

Les spécimens de bison ayant un mitotype appartenant aux clades Bb sont distribués sur deux clades Bb1 et Bb2 qui ont divergé il y a environ 405 000 ans (95% HPD 366 000-445 000 ans).
Bb1, qui a également été nommé Bison X par (Soubrier et al., 2016), correspond à une lignée qui s'est éteinte à la fin du Pléistocène et qui a disparu de l'Europe de l'Ouest avant le dernier maximum de glaciation.

Le clade Bb2 correspond à la lignée qui inclut les *B. bonasus* modernes (Figure 75 à droite). La radiation du clade Bb1 daterait d'il y a environ 255 000 ans (95% HPD 234 000-277 000). Il s'agit donc du clade de bisons trouvés au Pléistocène supérieur dont la radiation est la plus ancienne. Il contient des échantillons du Pléistocène moyen et supérieur provenant de France, d'Allemagne et du Caucase (Figure 75 à gauche). Ce groupe présente une grande diversité génétique ce qui est en accord avec les données paléontologiques qui indiquent que la démographie des populations de bisons en Eurasie était maximale au Pléistocène et jusqu'au dernier maximum glaciaire (il y environ 20 000 ans). Les troupeaux étaient alors nombreux occupant une large aire géographique (Markova et al., 2015). Est-ce que les bisons du clade Bb1 correspondaient au clade majoritaire dans les populations de bisons antérieures à 50 000 ans ? La grande diversité génétique trouvée à l'intérieur de ce clade et l'âge particulièrement ancien de sa racine est compatible ave cette interprétation.

Les échantillons français de la fin du Pléistocène moyen/du début du Pléistocène supérieur de la grotte de Coudoulous II (Coud) datant d'environ 130 000 ans sont regroupés ensemble et forment un groupe monophylétique avec les échantillons français du Pléistocène supérieur de l'Aven de l'Arquet (Arq) datés entre 46 700 et 34 300 BP ce qui suggère la continuité de cette lignée dans cette région. Il est à noter que dans l'aven de l'Arquet, le troisième spécimen ayant permis d'obtenir un mitogénome se trouve en dehors de ce sous-clade avec les autres bisons du clade Bb1. Ce groupe aurait divergé des autres Bb1 il y a environ 255 000 ans (95% HPD 234 000-277 000), tandis que la radiation des autres Bb1 postérieurs à 50 000 ans dateraient de 120 000 ans (95% HPD 108 570-132 303) indiquant que le sous-clade Coud-Arq reflète une diversité ancestrale qui avait déjà largement été réduite il y a 50 000 ans.
Les membres du deuxième sous-clade se retrouvent aussi bien en France du Sud entre 48 000 et 23 000ans, l'Aven de l'Arquet dans le Gard (Arq78531), La Berbie en Dordogne (LB3887), Les Plumettes en Vienne (LP1874) et Siréjol en Corrèze (Correze)] que dans la Haute Vallée du Rhin (échantillons NK de la figure 81 à gauche) entre 45 000 et 42 000 ans, qu'en Ukraine, Caucase et dans l'Oural entre 47 000 et 14 700 ans. L'occupation des territoires par les membres de ce clade aux différentes époques corrèle avec les oscillations climatiques, mais d'une façon paradoxale. On les trouve dans le Caucase et en Europe de l'Ouest, aussi bien dans le sud de la France que dans la Haute Vallée du Rhin, pendant le stage marine isotopique MIS3 qui était relativement plus chaud tandis qu'on les retrouve dans l'Oural pendant la période MIS2 qui correspond au dernier maximum glaciaire.

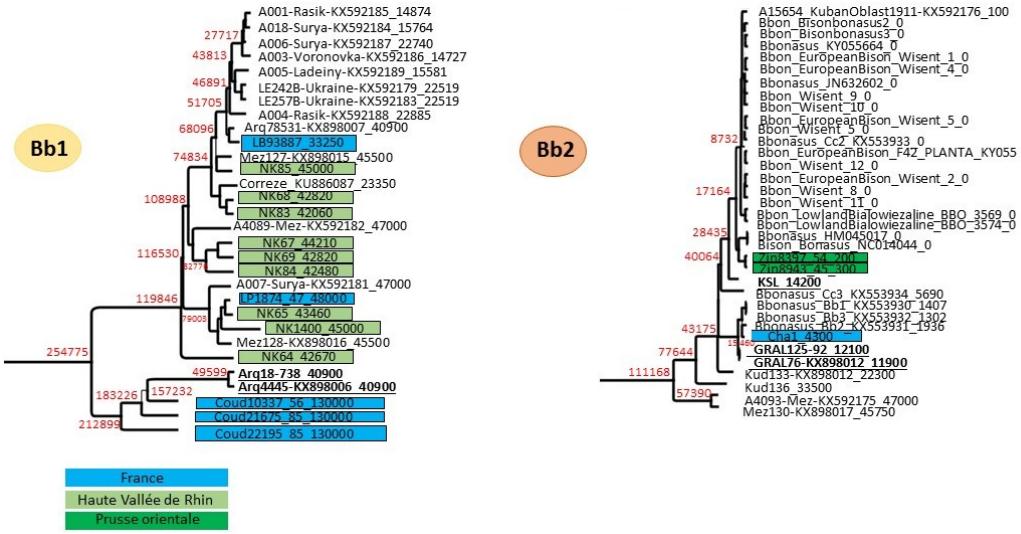


Figure 75: Phylogénie bayésienne des deux clades de B. bonasus (Bb1 à gauche et Bb2 à droite). Les couleurs correspondent à l'origine géographique des échantillons (Vert clair pour la Haute Vallée de Rhin, vert foncé pour la Prusse orientale et Bleu pour le sud de la France). Les âges des échantillons avant le présent (AP) sont ajoutés après le symbole «_». Les échantillons soulignés sont les échantillons discutés dans le texte.

La lignée Bb2 à laquelle appartient le de bison européen actuel a été trouvée plus rarement dans des spécimens anciens, particulièrement précédant le dernier maximum glaciaire (Massilani et al., 2016; Soubrier et al., 2016, étude présente). Parmi les échantillons anciens du clade Bb2 (Figure 75 à droite), on distingue 2 sous-groupes qui auraient divergé il y a environ 111 000 ans (95% HPD 97 000-127 000). Les plus anciens spécimens ont été trouvés dans le site de Mezmaiskaya dans le nord du Caucase autour de 47 000 à 40 000 ans (Massilani et al., 2016; Soubrier et al., 2016). On a mis en évidence deux spécimens dans le Sud du Caucase dans le site de Kudaro entre 33 000 et 22 000 ans (Massilani et al., 2016) et on a continué à en trouver dans la région, près du lac de Sevan en Arménie, jusqu'à 5 700 ans (Cc3 ; (Węcek et al., 2016)).

On a observé dans le site de l'Igue du Gral (Lot) la succession de bisons des clades Bp et Bb2, Bb2 apparaissant vers 12 000 av.n.e (Massilani et al., 2016). On le trouve aussi il y a 14 000 ans dans le gouffre de Kesslerloch (Canton Schaffhausen, Suisse, (Massilani et al., 2016)) et dans le gouffre « La Cha » (Ain) il y a environ 4 300 ans (présente étude) et en Autriche il y a entre 1 900 et 1 300 ans (Bb1-3, (Węcek et al., 2016)). La succession chronologique suggère que le clade Bb2 s'est réparti au début de l'Holocène du Caucase vers l'Europe de l'Ouest (Massilani et al., 2016).
Les *B. bonasus* actuels sont tous très proches phylogénétiquement et l'échantillon le plus proche et le plus ancien est celui de Kesslerloch. Les deux échantillons de l'ancienne Prusse orientale (ZIN), daté de 200 à 300 ans, font partie de cette faible diversité moderne tandis que les autres sous-clades incluant les échantillons de France et de l'Autriche de 12 000 à 1 300 ans n'ont pas laissé des traces détectables au présent. L'âge de l'ancêtre commun à toutes les séquences modernes Bb2 est d'environ 17 160 ans (95% HPD 11 000-23 000).

Toutes les radiations antérieures ont donné lieu à des branches éteintes. Les séquences que nous avons obtenues des échantillons de St. Pétersbourg, datés de 200 à 300 ans, montrent que la réduction de la diversité, visible chez les bisons actuels, s'est produite avant le goulot d'étranglement majeur de la première guerre mondiale. Ces échantillons, comme l'échantillon Cha1 datés d'environ 4 300 ans, sont de suffisamment bonne qualité pour donner lieu à des génomes complets ce qui permettra de bien caractériser la phase finale de l'évolution des bisons européens.

Les changements climatiques à la fin de l'Holocène pourraient avoir contribué à la disparition du clade Bb1, mais ceci n'est pas absolument clair compte tenu du fait qu'il a survécu dans l'Oural pendant le dernier maximum glaciaire. A-t-il été victime d'une synergie entre les changements climatiques et l'activité humaine ? Le clade Bb2, bien que moins diversifié que Bb1 dès le Pléistocène supérieur, a lui aussi subi une forte érosion de sa diversité après son expansion au début de l'Holocène. La chasse est probablement responsable de cette réduction et le bison a disparu de la France autour du Moyen Âge (Massilani et al., 2016), tandis qu'il a survécu en Europe centrale jusqu'au présent mais avec une diversité plus restreinte.

b) Bison priscus et bison américain moderne :

Le 3ème clade, Bp, est plus divergeant des autres clades et contient les B. priscus et les *B. bison* modernes ainsi que les échantillons de la fin du Pléistocène moyen et du Pléistocène supérieur. Ces échantillons sont distribués sur les 3 différents haplogroupes mitochondriaux. En plus de ceux décrits précédemment appartenant aux clades Bb1 et Bb2, le clade Bp regroupe essentiellement les spécimens du Pléistocène supérieur du sud de la France, de la Haute Vallée du Rhin et du nord de la Sibérie. On observe une proximité génétique entre les échantillons de France, de la Haute Vallée du Rhin et de Sibérie qui révèle la continuité de cette population à l'échelle du continent eurasiatique, continuité temporelle étalée sur près de 25 000 ans en Sibérie, et continuité géographique entre individus séparés de moins de 4 000 ans entre l'est et l'ouest de l'Eurasie provenant de sites distants de plus de 6 000 km (Figure 76).

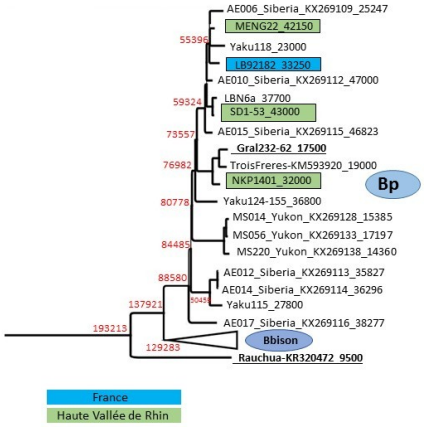


Figure 76: Phylogénie bayésienne des Bison priscus. Les couleurs correspondent à l'origine géographique des échantillons fossiles (vert pour la Haute Vallée du Rhin et bleu pour le sud de la France). Les âges des échantillons sont ajoutés après le symbole «_». Les échantillons soulignés sont les échantillons discutés dans le texte.

Nos données témoignent de la présence de *Bison priscu*s en Europe de l'Ouest au début du Pléistocène supérieur jusqu'à sa fin vers environ 17 500 ans, date correspondant à l'âge du plus jeune spécimen *Bison priscus* inclut dans nos études, originaire du site de l'Igue Gral. Notre équipe a analysé deux autres échantillons de l'Holocène du même site français de l'Igue du Gral et qui n'étaient pas assignés à l'haplogroupe mitochondrial de *Bison priscus* mais à celui de *Bison bonasus*, Bb2. Ces résultats ont permis de mettre en évidence, sur un même site, un remplacement de populations de bisons pendant la transition Pléistocène-Holocène (Massilani et al., 2016). En effet, les populations de *Bison priscus* ont été remplacées par les populations de bisons européens de manière concomitante avec le réchauffement climatique du début de l'Holocène.

Ce remplacement étaient précédé par une période de chevauchement en France dans le site «La Berbie» où nous avons trouvé deux échantillons de bisons contemporains (environ 34 000 ans) assignés à Bb1 (*Bison bonasus1*) et Bp (*Bison priscus*). Ce résultat corrobore encore une fois les résultats obtenus par (Massilani et al., 2016) qui ont montré qu'il y avait un chevauchement de ces deux populations en France au Pléistocène il y a environ 35 000 ans. Ce que nous observons en plus ici, c'est un chevauchement il y a 42 000 ans dans la Haute Vallée du Rhin entre les bisons des clades Bb1 (les neuf échantillons de bisons NK) et Bp (MENG22). Nous identifions aussi un autre individu du clade Bp en Bavière à la même époque. Ceci met en évidence un nouveau chevauchement temporel et géographique non-observé jusque-là. Au sein du clade Bp, une première bifurcation datée d'environ 193 190 ans (95% HPD 173 470-213 860) sépare des autres spécimens anciens et modernes le plus récent *Bison priscus*, daté d'environ 9 500 ans et caractérisé par (Kirillova et al., 2015). Cet échantillon, qui a été retrouvé dans la rivière Rauchua dans l'extrême nord-est de la Sibérie, suggère que la dernière population n'était pas un descendant direct de la population principale qui a dominé l'Europe de l'Ouest. Cette population cryptique n'ayant pas été identifiée jusque-là ailleurs et plus tôt, il est difficile de déterminer son origine. La radiation séparant les bisons ayant franchi le détroit de Béring et peuplé le continent américain date d'environ 138 000 ans (95% HPD 131 200-146 000). L'ancêtre commun de tous les *B. priscus* trouvés exclusivement sur le continent eurasiatique datent d'environ 88580 ans (95% HPD 78 600-100 087) ce qui montre que les populations de bison qui se sont distribuées de l'Est à l'Ouest du continent eurasiatique depuis 50 000 ans jusqu'à l'Holocène étaient relativement peu diverses génétiquement au niveau mitochondrial. Cette faible diversité mitochondriale est à contraster avec celle du clade Bb1 qui a co-existé sur le même territoire et qui, quant à elle, 2,5 fois plus élevée avec un ancêtre commun estimé à 255 000 ans (Grange et al., 2018; Massilani et al., 2016).

Cette population a également fait une incursion sur le continent américain car il contient trois échantillons trouvés dans le Yukon (Canada) et datant entre 17 000 et 14 000 AP. Cette incursion tardive ne semble pas avoir été fructueuse car cette lignée maternelle n'a pas laissé des descendants. L'âge de l'ancêtre commun de tous les mitogénomes actuels du bison américain *(Bison bison)* est estimé à environ 15 000 ans (95% HPD 13000-18000) ce qui indique qu'il a eu un goulot d'étranglement à la fin du Pléistocène, presque en même temps que *Bison priscus* s'est éteint. Cette période correspond à une période généralisée d'extinction de la mégafaune en Eurasie du Nord et en Amérique causée par une combinaison de changements climatiques et de pressions anthropiques (Cooper et al., 2015; Lorenzen et al., 2011).

Il est donc probable que l'extinction de *Bison priscus et l*es goulots d'étranglement que *Bison priscus* et son descendant américain *Bison bison* ont subi, étaient liés aux mêmes causes. Nos estimations d'âges des nœuds ont montré que l'ancêtre commun le plus proche entre *Bison priscus* et *Bison grunniens* daterait d'il y a environ 504 000 ans (95% HPD 450 000-557 000). Cette date est deux fois inférieur à celle estimée pour le clade de *Bison bonasus*. Presque le même facteur de différence entre les deux divergences a été estimé dans les deux publications précédentes de notre équipe (Grange et al., 2018; Massilani et al., 2016).

4) Discussion et Conclusion :

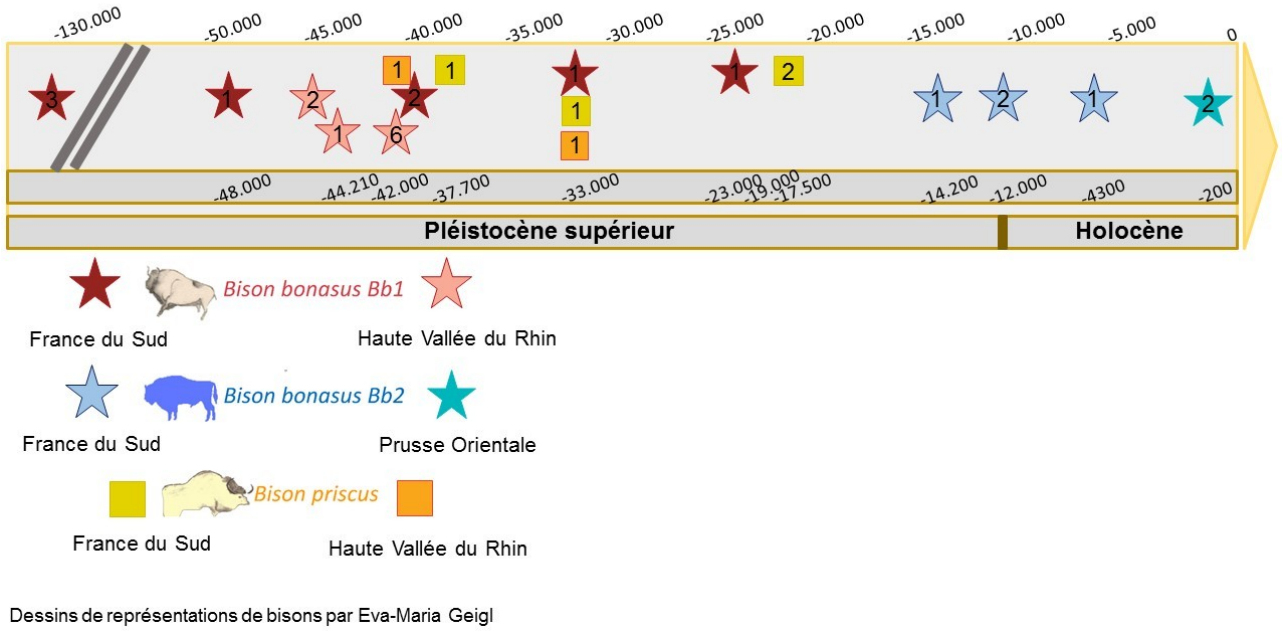
L'analyse des archives fossiles indiquait auparavant que l'Eurasie était peuplée par le *B. priscus* tout au long du Pléistocène moyen et supérieur et que le *B. bonasus* n'apparaissaient qu'à l'Holocène (voir Introduction).

L'approche paléogénomique des lignées maternelles a remis en cause cette grille d'analyse. Deux lignées maternelles bien distinctes cohabitaient dans le Pléistocène supérieur, la lignée caractéristique des *B. bonasus* actuels a été détectée dès le Pléistocène supérieur dans le Caucase.
L'interprétation morphologique ne rend-elle pas bien compte de la phylogénie des populations ? On pourrait considérer l'hypothèse que les différences de lignée maternelle ne reflètent pas des populations distinctes mais cette interprétation ne rendrait alors pas compte du fait que, dans une période donnée à un moment donné, l'un des clades est dominant. Ceci suggère qu'il s'agit de populations qui avaient des propriétés distinctes, possiblement adaptées à des environnements qui n'étaient pas identiques. Il a été proposé que les changements des populations que l'on observe dans un endroit sont souvent corrélées avec des associations climatiques (Grange et al., 2018; Massilani et al., 2016).

Lorsque l'on peut analyser suffisamment d'échantillons au cours du temps sur la même région, on observe toutefois que les transitions entre populations impliquent de longues périodes de chevauchement de populations (Figure 77). Ainsi, entre 130 000 et 40 000 ans, on peut identifier une dominance des Bb1 dans le sud de la France (6 individus) et dans la Haute Vallée du Rhin (9 individus), mais il y a déjà 2 Bp identifiés dans la Haute Vallée du Rhin il y a 43 000 ans. Entre 40 000 et 15 000 ans, on observe plus de Bp (4 dans le sud de la France et 1 dans la Haute Vallée du Rhin), mais on continue à observer des Bb1 (2 individus) dans le sud de la France à 33 000 et à 23 000 ans. A partir de 15 000 ans, on ne voit plus en France, Allemagne, Suisse et en Autriche que des Bb2, les Bb2 étant détectés plus précocement dans le Caucase entre 47 000 et 22 000 ans. Dans l'Oural et en Ukraine, les Bb1 ont dominé entre 23 000 et 15 000 ans (Soubrier et al., 2016) (Figure 83). Il n'est donc pas clairement évident que le climat favorise plus particulièrement les bisons des lignées Bb1 ou Bp mais plutôt qu'il y avait des expansions et contractions de populations des lignées Bb1 qui dominaient dans la partie Ouest du continent eurasiatique, et de Bp qui étaient dominant à l'Est du continent eurasiatique et qui faisaient des incursions à l'Ouest. Si l'on considère l'ensemble des résultats obtenus dans notre équipe et publiés par Soubrier et al., 2016, les deux populations Bb1 et Bp auraient coexisté pendant le Pléistocène supérieur avec juste des périodes où l'une des deux populations domine à un endroit particulier.

L’analyse comparative des génomes des bisons américains et européens actuels indique que ces deux « espèces » ont commencé à se séparer autour de 215 000 ans et que le flux génique entre elles aurait cessé il y a 100 000 ans (Grange et al., 2018). Ceci correspond à la période où les deux espèces occupent des continents différents. Ceci indique que les populations de bison Bb1, Bb2 et Bp devaient s’hybrider entre elles pendant qu’elles cohabitaient sur le continent eurasiatique. La dominance des mitogénomes observée localement selon les endroits et les périodes pourrait être due à une certaine homogénéité des troupeaux de femelles tandis que les mâles devaient contribuer à équilibrer le pool génétique en circulant d’un troupeau à l’autre (Grange et al., 2018). L’analyse des génomes nucléaires que j’ai produits à partir des échantillons anciens devrait permettre de mieux caractériser les éventuelles spécificités des deux populations identifiées par les mitogénomes, ainsi que l’ampleur des flux génétiques entre elles.

Finalement, les génomes des individus de la Haute Vallée du Rhin provenant de crânes avec leurs cornes, il sera possible de lier leur analyse morphologique avec une analyse génomique et donc d’explorer s’il existait vraiment des signatures morphologiques de ces deux lignées qui auraient pu être capturées sur les représentations sur les peintures rupestres (Massilani et al., 2016; Soubrier et al., 2016).



Dessins de représentations de bisons par Eva-Maria Geigl

Figure 77 : Transitions de populations de bisons au cous du temps impliquant de longues périodes de chevauchement.

III) Relation phylogénétique mitochondriale des aurochs et des bovins domestiques anciens et modernes :

Nous allons d'abord explorer comment l'analyse des mitogénomes nous a permis de suivre l'évolution des populations d'aurochs depuis 50 000 ans ainsi que le processus de la domestication de ces aurochs et de leur diffusion, particulièrement en Europe. Pour présenter cette masse importante de données, il m'a paru préférable de les regrouper par haplogroupe plutôt que par période. Ces haplogroupes sont ordonnés en fonction de l'ancienneté de leur radiation ce qui me permettra de discuter d'abord des événements les plus anciens pour finir par la domestication et la diffusion en Europe.

1) Découverte d'un nouvel haplogroupe mitochondrial chez les aurochs et son apport à la compréhension du peuplement de l'Europe de l'Ouest par les aurochs au Pléistocène :

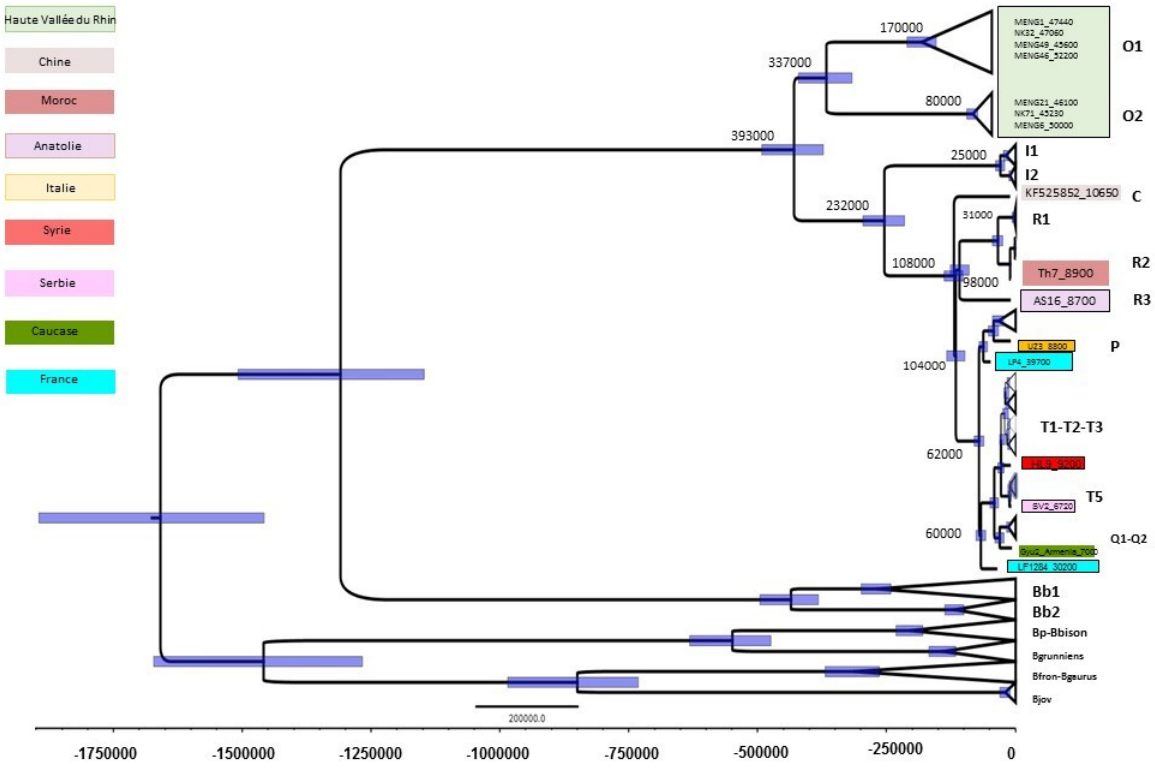


Figure 78: Arbre phylogénétique bayésien des mitogénomes anciens et modernes de Bos primigenius, Bos taurus et Bos indicus. Quelques échantillons anciens qui divergent à des positions ancestrales des clades principaux sont représentés sur cette figure dans des cadres colorés en fonction de leur origine géographique, ceux analysés par notre équipe sont entourés d'un trait noir.

Nous avons analysé un corpus important d'aurochs du Pléistocène tardif originaire de la Haute Vallée du Rhin provenant de deux anciennes collections privées et ayant été collecté au fil des années par des particuliers lors de découvertes fortuites. Tous ces échantillons ont été datés au carbone-14 car le contexte paléontologique n'est pas connu. Ces échantillons datent d'une période comprise entre 52 000 et 45 000 cal BP. Les caractéristiques, incluant les âges et l'endroit de découverte ainsi que le pourcentage d'ADN endogène et la couverture des mitogénomes appartenant à cet haplogroupe nouvellement identifié sont présentés dans le tableau suivant.

Échantillon fossile	Haplotype mitochondriale	Âge approximatif	Site	Origine	Cadre chronologique/contexte archéologique	% ADN endogène de banques génomiques	Couverture moyenne du mitogénome
MENG1	O1	47440	Eich	Haute Vallée du Rhin	Pleistocène supérieur	47.92%	140x
NK32	O1	47060	Bobenheim-Roxheim			33.60%	276x
MENG49	O1	45600	Crumstadt			59.78%	174x
MENG46	O1	52200	Biblis			13.24%	69x
MENG21	O2	46100	Eich			34.73%	44x
NK71	O2	46100	Bobenheim-Roxheim			4.40%	20x
MENG6	O2	46100	Groß-Rohrheim			15.87%	48x

Tableau 11: Échantillons fossiles d'aurochs anciens de la Haute Vallée du Rhin appartenant au nouvel haplogroupe mitochondrial O.

Bien que le pourcentage d'ADN endogène obtenu à partir des banques d'ADN génomique varie selon les échantillons, la capture des mitogénomes a été suffisamment efficace pour couvrir les mitogénomes de 20 à 275 X. Pour quelques mitogénomes, il reste un trou dans la région hypervariable à cause de la très grande divergence de cette région avec le mitogénome utilisé pour la capture mais aussi avec ceux utilisé pour cartographier les lectures. S'il a été possible de boucher ce trou pour l'haplogroupe O1, pour l'haplogroupe O2 qui ne contient que les spécimens moins préservés, il n'a pas été possible de boucher le trou même en utilisant pour la cartographie la séquence hypervariable de l'haplogroupe O1. Ceci suggère que les deux régions hypervariables sont différentes à l'endroit du trou car même avec les mitogénomes non-capturés, cette région s'est révélée réfractaire à la cartographie. Cette divergence entre les régions hypervariables des haplogroupes O1 et O2 est attendue compte tenu de la date importante de la divergence entre les deux mitogénomes : 337 000 ans (95% HPD 302 000 – 375 000).

Le point saillant de cet haplogroupe O, c'est que l'ancêtre commun le plus récent avec les autres haplogroupes (I, C, R, P, Q, T) est très ancien, 393 000 ans (95% HPD 357 000 – 433 000). Cet haplogroupe a donc divergé bien avant la séparation entre les zébus (I) et les taurins (C, R, P, Q, T). Pour interpréter l'évolution de ces divergences, il faut considérer les fluctuations de température au cours des derniers 900 000 ans (voir figure 79) et l'impact qu'elles ont eu sur les environnements. Les périodes les plus froides sont incompatibles avec la présence d'aurochs dans l'Europe centrale et du nord. Comme indiqué dans l'introduction, l'hypothèse la plus consensuelle est celle d'une origine du genre *Bos* se situerait dans le sous-continent indien.

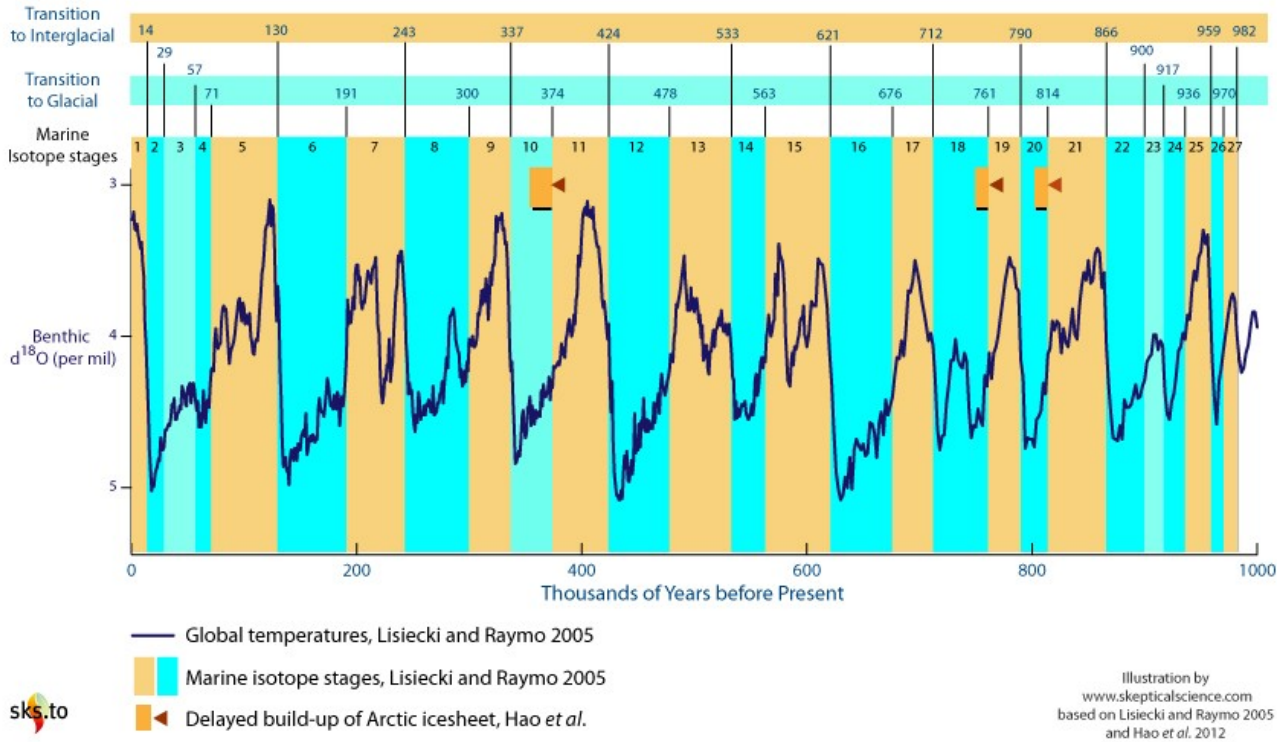


Figure 79 : Fluctuations de la température au cours d'un million d'années basées sur les changements de ^{18}O dans les sédiments océaniques avec les stades isotopiques marins déduits de cet enregistrement (Hao et al., 2012; Lisiecki & Raymo, 2005).

Les premières populations d'aurochs auraient atteint l'Europe au début du Pléistocène moyen, vers environ 700 000 ans (Vuure, 2005). Ceci correspond à une période où la température était plus clémente. La date de séparation des groupes O avec les autres clades taurins dont les zébus suggère qu'ils seraient issus d'une vague de migration des aurochs vers l'Europe postérieure à cette première date car cette migration devrait être légèrement postérieure à la date de divergence estimée ici à 393 000 ans. La séparation des groupes O1 et O2 estimée à il y a 337 000 ans, s'est probablement produite en Europe puisque l'on a trouvé des individus appartenant à ces deux haplogroupes au même endroit dans la Haute Vallée du Rhin il y a ~50 000 ans. Ceci suggère donc que l'haplogroupe O est issu d'une migration en Europe en provenance d'Inde comprise entre ces deux dates, c'est-à-dire, pendant la période chaude du MIS 11, le stade isotopique marin 11 (voir figure 79). Des descendants, par la lignée maternelle, de cette population auraient persisté en Europe jusqu'à il y a 50 000 ans (MIS 3). Cette population n'aurait pas survécu au dernier maximum glaciaire (MIS 2).

L’ancêtre commun pour les membres de l’haplogroupe O1 date d’environ 170 000 ans (95% HPD 148 000-192 000 ans) et celui de l’haplogroupe O2 date d’environ 80 000 ans (95% HPD 70 000-90 000 ans). Ces dates correspondent à MIS 6 et MIS 5a, respectivement, donc, à une période glaciaire et interglaciaire. L’âge de l’ancêtre de l’haplogroupe O1 suggère un goulot d’étranglement pendant la période glaciaire. Pour l’haplogroupe O2, le nombre d’échantillons étant petit et une petite partie de la région hypervariable étant manquante, il n’est pas certain que cette date persisterait si plus de données étaient disponibles.

Bien que nous disposions de peu de séquences provenant de cette période originaire du sud de l’Europe, les quelques séquences obtenues en provenance du sud de la France et du nord de l’Espagne n’appartiennent pas à cet haplogroupe O (voir paragraphe suivant), alors qu’aucun des aurochs de la Vallée du Rhin n’appartient aux haplogroupes que l’on retrouve plus au sud. Il est à noter qu’il y a un décalage temporel d’environ 5 000 ans entre l’échantillon le plus récent de la Haute Vallée du Rhin et les plus anciens que nous avons analysés du sud de la France et du nord de l’Espagne. Bien que le nombre d’échantillons soit petit, la probabilité que ces deux populations soient différentes par hasard est de 0,003 (test exact de Fisher). Est-ce que ces deux populations n’étaient donc pas connectées, du moins en ce qui concerne les troupeaux de femelles?

2) L’haplogroupe mitochondrial P: Signature des aurochs Européens :

Avant notre étude, 17 mitogénomes complets de l’haplogroupe P étaient disponibles : 11 d’entre eux provenaient de bovins actuels originaires d’Asie de l’Est, un en provenance de Corée (FC3, (Zhang et al., 2013)), et 10 provenant du Japon (Mannen et al., 2020). Les six aurochs anciens correspondaient à un aurochs anglais daté de 6 740 ans (CPC98) (Edwards et al., 2010), un polonais de 1 500 ans (JQ437479) (Zeyland et al., 2013), trois provenant de cornes à boire du Moyen Âge datés à environ 500 à 600 ans (Bro-Jørgensen et al., 2018, p.), et un espagnol de 9 100 ans (CL2) (Gurke et al., 2021).

Nous avons ajouté à cette liste 21 nouveaux mitogénomes complets d’échantillons aurochs appartenant à l’haplogroupe P dont 7 avaient été obtenus auparavant par Diyendo Massilani. L’échantillon le plus vieux est un spécimen français datant du Pléistocène supérieur est originaire du site « Les Plumettes » (LP4 ; Lussac-les-Châteaux, Vienne, Poitou-Charentes) datés à 40 186 – 39 171 ans AP (Figure 80). Cet échantillon date donc de MIS 3 et précède le dernier maximum glaciaire. Onze autres échantillons français datant entre environ 16 050 et 4 400 ans: deux aurochs du sud de la France, un de la grotte de Coupe Gorge (CG, Montmaurin, Comminges Pyrénées, Haute-Garonne, Occitanie), datant de 18 300 – 17 700 ans AP, donc du début du réchauffement climatique précédant la chute de température du Dryas récent (12 880 – 11 650 ans AP).

Tous les autres échantillons datent de l'Holocène et donc d'une période chaude et stable. Il y a un échantillon du site de « Les Pradelles-Marsat » (Auvergne-Rhône-Alpes) datant de 10 695 – 10 414 ans AP. Deux échantillons anciens du début du Néolithique originaire de la Vallée de l'Aisne (Picardie), «BFT951» (Bucy-le-long) vieux d'environ 7 100 ans, et «BGM310-311» (Bucy-le-long le Grand Marais) daté à 6 572 – 6 403 ans AP, ainsi qu'un échantillon du Néolithique moyen (Michelsberg Kultur) «MGA9» vieux de 6 100-5 800 ans AP (Maizy). Il s'y rajoutent 5 échantillons datant du Néolithique récent du site de Chalain, (CHL12 et CHL21) (Franche-Comté), datant de 3 200-3 000 av.n.e. (~5 000 ans AP) les deux échantillons du muséum de Nantes, (NANT1 et NANT2) (de la région Loire-Atlantique) datant de 4 520 – 4 240 ans AP et de 4 620 – 4 420 ans AP, respectivement. Les échantillons de l'aurochs de « Pontvallain » (Sarthe) (PVL), datant de 3 985 – 3 718 ans AP et un échantillon de l'Âge de Bronze du site d'Osly Courtil (Vallée de l'Aisne) (OTM2) datant d'environ 2 850 ans AP. L'échantillon le plus récent provient d'une corne des collections du Musée Vert (Le Mans, Sarthe), dont l'origine initiale est inconnue mais qui est suspectée comme étant celle d'un aurochs, qui date de la fin du 12ème, du début du 13ème siècle n.e.

Nous avons aussi caractérisé des aurochs du début de l'Holocène provenant tous du sud de l'Italie, trois de la Grotta del Uzzo (Trapani, Sicile) (UZ1, UZ2, UZ3) datant à environ 8 830 ans AP, un échantillon de la Grotta del Santuario della Madonna (Cosenza, Calabre) (DMAD2) datant d'environ 10 000 ans AP, et trois échantillons de Santa Maria di Agnano (Ostuni, Brindisi, Pouilles) (SMA161B, SMA28 et SMA149) datant, pour SMA149, à 9266 - 9013 ans AP. Des aurochs grecs sont aussi retrouvés dans la branche de l'haplogroupe P. Il s'agit d'un échantillon du site de Fyllotsairi Mavropigi (Grèce du Nord) (MAV2) et d'un échantillon de la grotte Kouveleiki B (Kouv1) datant d'environ 8 600-7 900 et 6 269-5 949 ans, respectivement.

Nous avons pu passer le corpus de mitogénomes complets d'aurochs antérieur 2 800 ans AP de 2 à 23 ce qui permet d'étudier la phylogénie de cet haplogroupe depuis son origine et dater son évolution au Pléistocène supérieur et à l'Holocène. L'échantillon LP4 est le plus distant des autres aurochs. L'ancêtre commun de tous les aurochs incluant LP4 daterait d'environ 54 500 ans AP (95% HPD 47 000 – 62 000ans) (Figure 80). Avec l'approche bayésienne, l'ancêtre commun d'UZ3 et de tous les autres aurochs daterait de 39 000 ans AP (95% HPD 31 000 – 47 000). Par contre, la position d'UZ3 est beaucoup moins certaine. En effet, l'analyse de maximum de vraisemblance montre qu'il groupe avec les mitogénomes des autres aurochs de la Grotta del Uzzo, ainsi qu'avec deux échantillons italiens, DMAD2 et SMA161B (Figure 81).

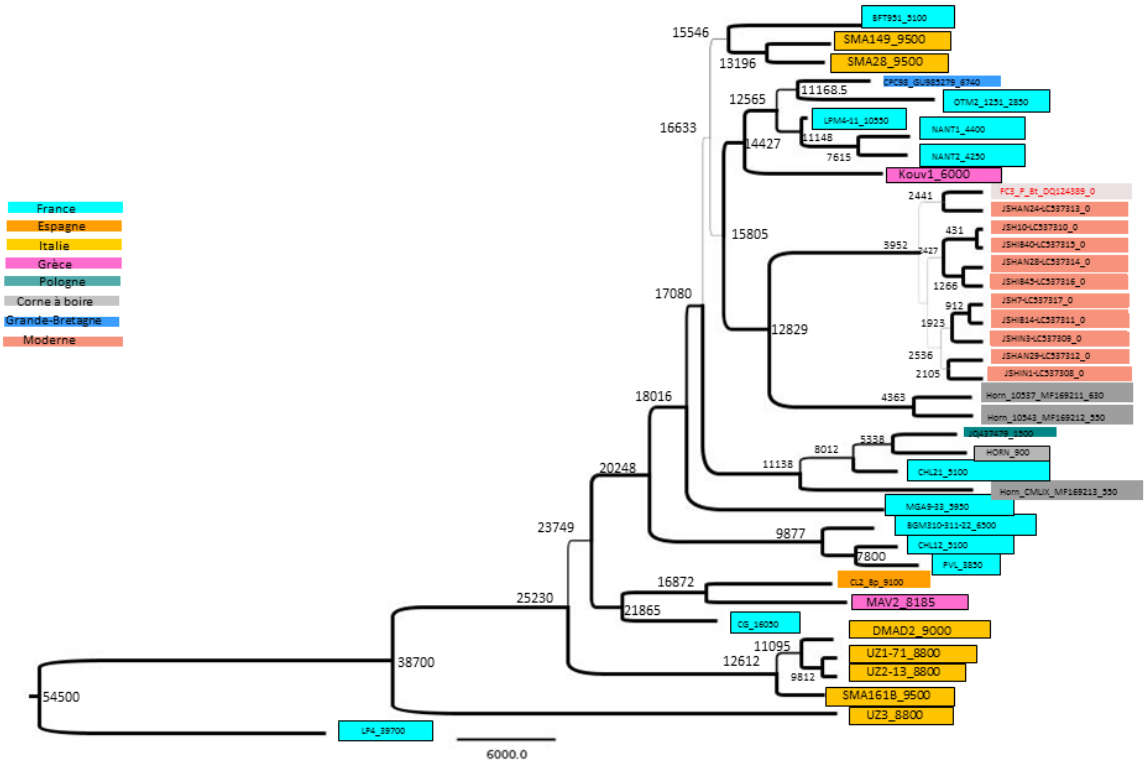


Figure 80: Phylogénie mitochondriale bayésienne de l'haplogroupe P et origines géographiques des spécimens d'aurochs européens. L'âge médian des nœuds est indiqué. L'épaisseur des branches est proportionnelle à la probabilité postérieure à laquelle elles sont associées.

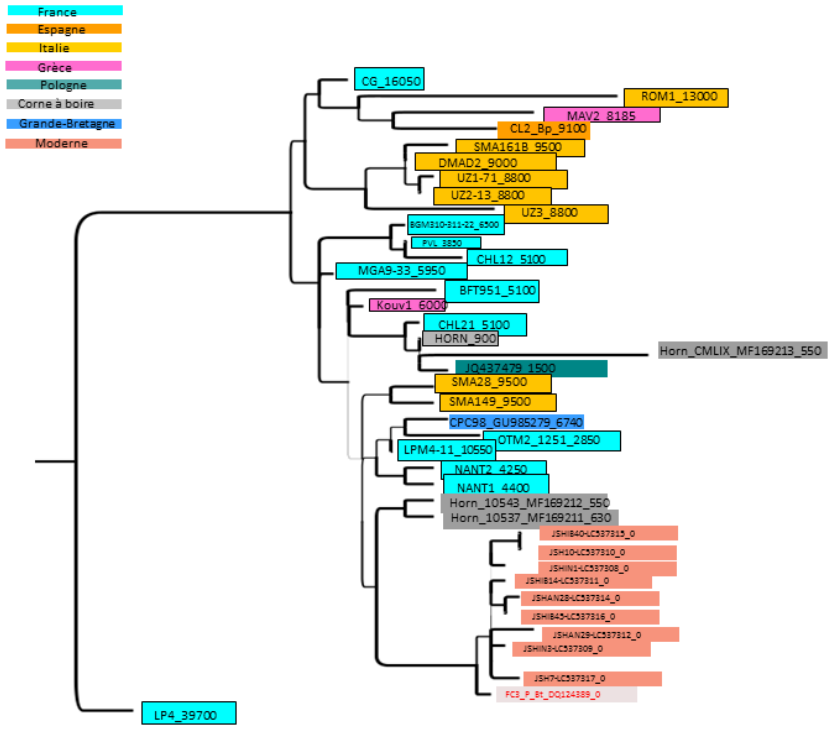


Figure 81 : Arbre de maximum de vraisemblance de l'haplogroupe P.

Tous les autres aurochs P sont plus semblables et l'ancêtre commun daterait de 29 000 ans AP (95% HPD 22 000 – 41 000). L'aurochs LP4 correspond à une diversité antérieure au dernier maximum glaciaire qui n'a pas survécu à cette période. Bien que la position dans l'arbre bayésien eût une bonne probabilité postérieure, la phylogénie avec le maximum de vraisemblance paraît peut-être plus cohérente. Les ancêtres des aurochs de la Grotta del Uzzo ont dû gagner la Sicile pendant le dernier LGM lorsqu'il y existait un pont terrestre entre cette île et le continent (Antonioli & al, 2012).

L'échantillon de Coupe Gorge, CG, du plus ancien aurochs de cet haplogroupe après LP4 correspond à une lignée que l'on retrouve 7 000 ou 8 000 ans plus tard aussi bien en Espagne (CL2) qu'en Grèce (MAV2). Un autre aurochs dont le mitogénome était trop incomplet pour figurer dans l'analyse bayésienne, l'échantillon ROM1 de la Grotta de Romanelli en Italie dans les Pouilles et datant à 13 750 – 9 200 ans AP, se retrouve aussi dans ce groupe lorsqu'on fait une analyse de maximum de vraisemblance en éliminant les parties manquantes (Figure 81).

L'ancêtre commun de la majeure partie des aurochs de l'Holocène (sans UZ3) datant du dernier maximum glaciaire (LGM), on peut en déduire que les aurochs de l'haplogroupe P ont subi un goulot d'étranglement sévère pendant cette période et qu'ils ont ensuite connu une expansion de population en même temps qu'une radiation dans toute l'Europe. Ils se trouvaient pendant la période glaciaire probablement dans un ou plusieurs refuges au sud, possiblement couvrant l'Espagne, le sud de la France et l'Italie. CG est le plus ancien de ce groupe et localisé dans le sud de la France avant le Dryas récent indiquant que cette région était soit un des refuges pendant le LGM, soit était accessible facilement après le réchauffement.

Entre il y a 25 000 et 27 000 ans AP, il y a une incertitude dans l'arbre bayésien. Si CG est toujours associé avec CL2 et MAV2, et que les quatre échantillons italiens SMA61B, DMAD2, UZ1 et UZ2 groupent ensemble dans tous les cas, l'incertitude concerne l'ordre dans lequel les deux sous-groupes divergent des autres aurochs qui varie selon les répétitions des analyses bayésiennes. Ces deux sous-groupes sont liés entre eux dans l'analyse de maximum de vraisemblance et séparés de tous les autres aurochs de l'Holocène. Par contre, les deux autres échantillons des Pouilles de la même époque (SMA149 et SMA28) forment un sous-groupe avec un aurochs français de 5 500 ans AP (BFT951) et sont inclus dans les différents clades apparentés dont la racine a ~20 000 ans. Au sein de ce clade on trouve tous les aurochs qui se sont repartis sur tout le continent européen et dont une petite sous-branche est encore présente en Asie de l'Est. Les aurochs français, italiens et espagnols du début de l'Holocène juste après la fin du LGM, sont donc apparentés et on ne peut pas clairement identifier s'ils étaient localisés dans un seul refuge sur ces trois territoires ou si ces derniers étaient suffisamment connectés pour ne correspondre qu'à un seul refuge. L'âge des racines indique de toute façon qu'un ancêtre commun date du plein milieu du dernier LGM ce qui indique un goulot d'étranglement sévère à cette période (Figure 82).

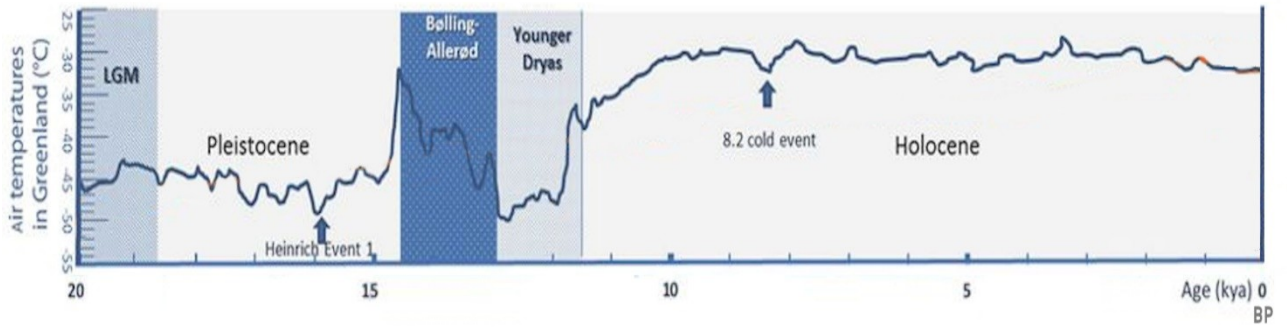


Figure 82 : Températures de l'air lors des derniers 20 000 ans reconstruites à partir des carottes glaciaires du Groenland GISP2 (Platt et al., 2017).

Nous avons donc utilisé l'approche bayésienne pour reconstituer l'évolution de la taille de population effective pouvant donner lieu à la diversité mitochondriale observée (« Bayesian skyline plot ») (Figure 83). Comme nous disposons d'un seul aurochs P précédant le LGM (LP4), nous ne voyons pas d'évolution de la taille de la population entre il y a 57 000 et 20 000 ans (entre deux points, on ne peut faire passer qu'une droite). On voit l'expansion de la population commençant à partir de la racine autour de 20 000 ans AP, mais il est en fait probable que l'expansion réelle date de la fin du LGM entre il y a 15 000 et 14 000 ans qui est effectivement la période à partir de laquelle on voit l'expansion la plus nette de la taille de la population.

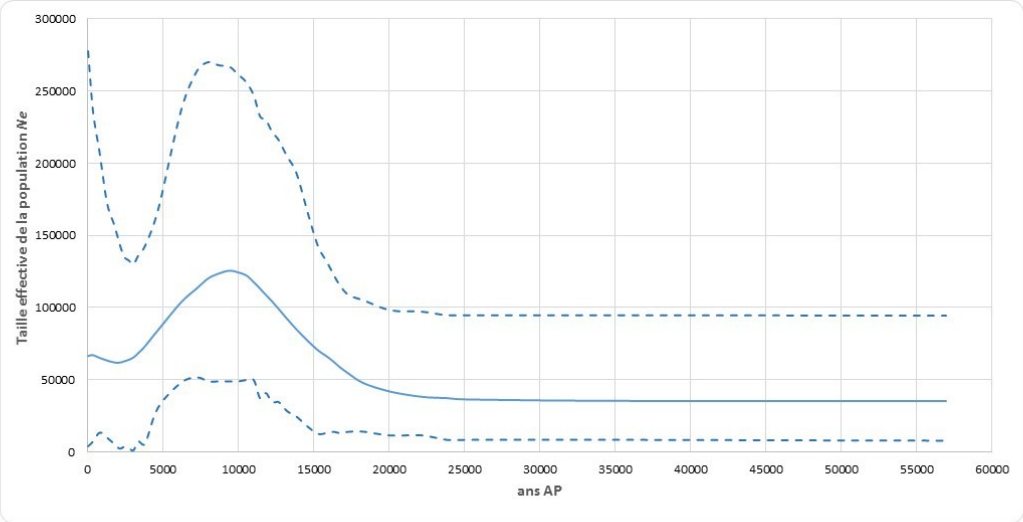


Figure 83 : « Bayesian skyline plot » correspondant à la modélisation de la taille de la population effective des aurochs de l'haplogroupe P à partir des mitogénomes obtenus (en trait plein la médiane, en pointillé les limites de l'intervalle de confiance à 95%).

Le gros de la diversification des séquences a effectivement eu lieu entre il y a 18 000 et 8 000 ans, comme on peut le voir avec l'âge des différents nœuds, et autour d'il y a 8 - 6 000 ans, la population commence à décliner. Les mitotypes P les plus récents du Moyen Âge, provenant des cornes à boire, et de l'époque actuelle, préservés à la suite d'une introgression dans les bovins domestiques d'Asie de l'Est correspondent à une toute petite partie de la diversité des aurochs de l'Holocène au pic de leur expansion. Deux des quatre cornes à boire, sont proches de la branche qu'a donné lieu aux séquences modernes (avec un ancêtre commun autour de 12 700 ans AP) tandis que deux autres cornes à boire ont un ancêtre commun avec ces derniers survivants remontant à 17 000 ans AP. Les séquences modernes d'Asie de l'Est sont toutes très proches avec un ancêtre remontant autour de 5 000 ans AP et une expansion rapide à partir de cet ancêtre qui donne une topologie de l'arbre très mal résolue. Cette dernière expansion est visible aussi avec la reconstitution démographique du « Bayesian skyline plot ».

Les aurochs P se sont répandus jusqu'en Grèce, mais la position des deux échantillons grecs (8 000 et 6 000 ans AP) au sein de la phylogénie montrent qu'ils sont apparentés aux échantillons espagnols, français, italiens et avec dans chaque cas des ancêtres communs entre il y a 17 000 et 145 00 ans ce qui indique clairement qu'ils sont arrivés en Grèce en provenance du sud-ouest de l'Europe. L'aurochs anglais CPC98 est apparenté aux aurochs français et l'ancêtre commun avec les français les plus proches (OTM2) remonte à environ 10 000 ans AP montrant que la colonisation de la Grande-Bretagne par les aurochs remonte à la période d'expansion après la sortie du refuge du sud. La connexion entre la Grande-Bretagne et le continent existaient jusqu'en 8 500 à 8 200 AP, quand les eaux ont monté et ont submergé la partie entre les deux formant la manche actuelle (Weninger et al., 2008) terminant le flux génique entre les populations d'aurochs britanniques et continentaux.

3) Autres aurochs européens :

Parmi les aurochs de l'Europe du sud qui précèdent le LGM, tous n'appartiennent pas à l'haplogroupe P. L'échantillon le mieux couvert que nous avons analysé est l'échantillon LF1284 (Le Flageolet I, Bezenac, Dordogne, Aquitaine) daté à 30 192 cal BP qui se trouve à la racine des haplogroupes Q et T et dont l'ancêtre commun avec les membres de ces haplogroupes date de 60 000 ans AP (95% HPD 67 000 - 52 000 ans AP) (Figure 84). Le dernier ancêtre commun à tous les haplogroupes P, Q et T a une date très peu différente de 62 000 ans AP (95% HPD 71 000 - 55 000 ans AP). Pour rappel, l'ancêtre commun entre l'aurochs LP4 et les autres membres de l'haplogroupe P était, quant à lui, daté à 54 500 ans AP (95% HPD 62 000 – 47 000 ans AP). On voit donc ainsi que les aurochs LF1284 et LP4 sont en fait très proches l'un de l'autre et eux-mêmes très proches de la racine commune, aussi bien aux aurochs européens de l'Holocène de l'haplogroupe P que des aurochs d'Asie du sud-ouest qui ont été domestiqués quelques milliers d'années après le début de l'Holocène. Il y avait donc en France avant le LGM une population d'aurochs ancestrale à tous les aurochs qui ont survécu au LGM.

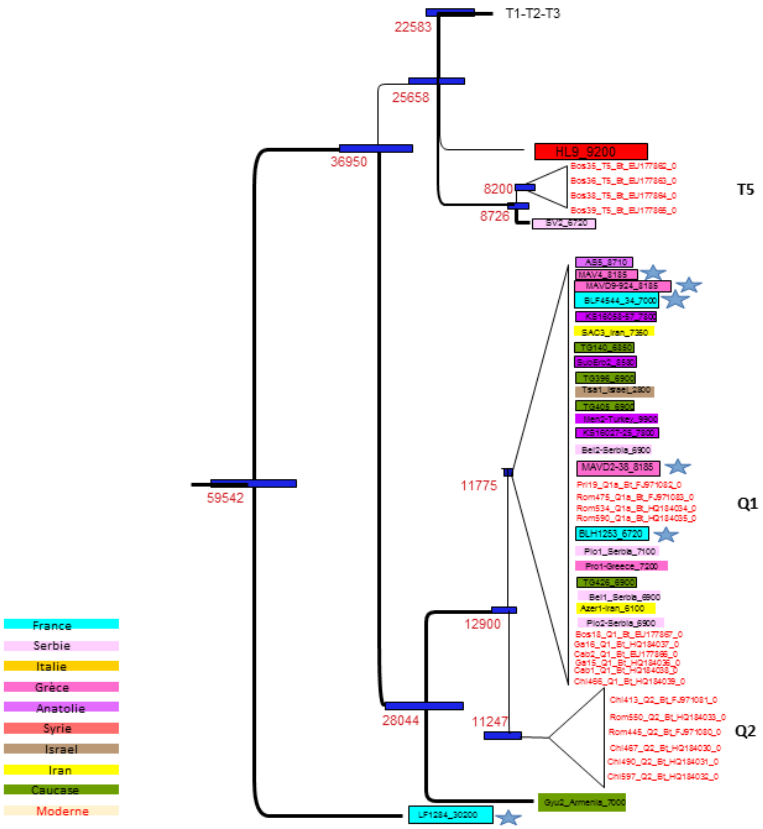


Figure 84: Arbre phylogénétique bayésien des lignées mitochondriales T et Q. Les aurochs européens non P sont marqués par une étoile.

Nous avons aussi obtenu un certain nombre de données de séquences mitochondriales à partir de deux autres aurochs du sud de l'Europe précédant le LGM, mais pas de quoi couvrir le mitogénome complet. Pour un de ces échantillons, CANO24, originaire de Riera dels Canyars (Gavà, Barcelona, Espagne), daté géologiquement à ~40 - 38 000 AP, nous avons obtenu environ 14 554 bases du mitogénome. Pour l'autre échantillon, GO2423, de la Grotte de l'observatoire (Monaco), daté à ~70 -30 000 ans AP, nous n'avons pu obtenir que 8 768 bases. Comme ces couvertures étaient insuffisantes, nous ne les avons pas inclus dans l'analyse bayésienne mais nous avons fait un arbre de maximum de vraisemblance en éliminant toutes les positions manquantes à l'échantillon CANO24 pour s'assurer de sa position phylogénétique (Figure 85). On voit que ces deux mitogénomes sont aussi à la racine des haplogroupes Q et T avec LF1284. Le GO2423 est très proche du LF1284, mais comme il lui manque 35% de la longueur totale (5 786 positions), la branche est anormalement longue. LF1284 et CANO24 sont extrêmement proches l'un de l'autre et tous deux à la racine de ces haplogroupes.

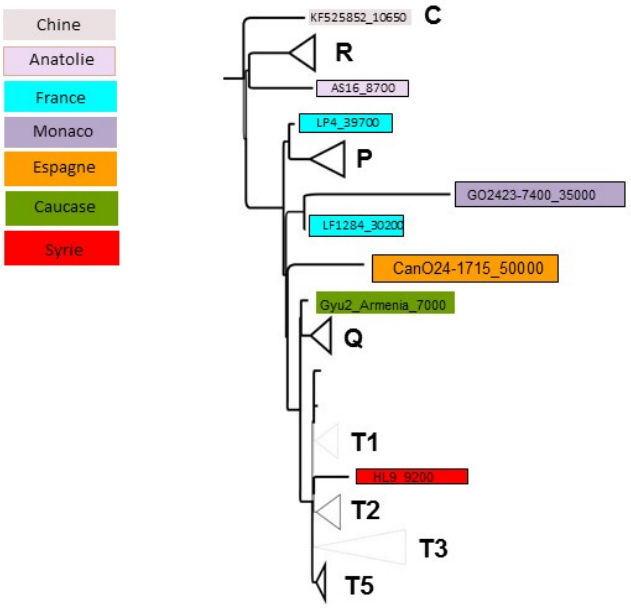


Figure 85 : Arbre de maximum de vraisemblance (RaXML) montrant que les aurochs LF1284, CANO24 et GO2423 sont à la racine des Q et des T. L'alignement utilisé a été purgé pour toutes les positions manquantes à CAN024 alors qu'il reste 5 786 positions manquantes pour GO2423, ce qui explique la longueur anormale de la branche.

Les quatre aurochs du sud de l'Europe précédant le LGM pour lesquels nous avons pu obtenir des mitogénomes sont tous à la racine des P ou des QT et tous assez proches les uns des autres. On voit donc que ces aurochs correspondent à la population qui a survécu au LGM. Les aurochs de l'haplogroupe P sont ceux qui se sont répandus en Europe à l'Holocène et on ne trouve pas les aurochs des groupes Q et T dans les individus qui procèdent l'arrivée des bovins domestiqués lors de l'expansion du Néolithique. Ceci suggère que les aurochs de la population pré-LGM que l'on a détecté en France et en Espagne, ont dû se repartir dans d'autres zones refuge plus à l'est et que les contingences du goulot d'étranglement ont permis la survie des précurseurs des haplogroupes Q et T en Asie du sud-ouest.

Notre phylogénie bayésienne et l'identification des séquences de ces spécimens permet de conclure que la population d'aurochs d'Asie du sud-ouest qui a été domestiquée n'aurait divergé que très tardivement de la population d'aurochs qui aurait survécu en Europe. Il reste une incertitude quant au moment où la population d'aurochs que nous avons identifiée dans ces spécimens d'Europe de l'ouest aurait pénétré en Asie du sud-ouest. Comme nous n'avons pas identifié d'aurochs du tout début de l'Holocène appartenant aux haplogroupes T et Q en Europe et tous appartenaient à l'haplogroupe P, nous favorisons l'hypothèse admise auparavant que T et Q n'ont pas survécu au LGM dans les refuges de la France du sud, de la péninsule ibérique et italique.

Il est à noter qu'il y a eu une publication qui a décrit un aurochs mésolithique (non daté mais indiqué comme ayant 11 450 ans) en Italie appartenant à l'haplogroupe T3 (Lari et al., 2011), mais cette séquence a été obtenue entièrement par PCR en amplifiant des fragments de grande taille, sans aucune prévention de la contamination des réactifs par l'ADN bovin provenant de la sérum albumine et la séquence obtenue correspond totalement à la séquence majoritaire dans les réactifs contaminés et que c'est une séquence dérivée ne correspondant pas aux séquences d'un tel âge. Par contre, ceux que nous avons échantillonné au début de l'Holocène à l'Est de l'Europe, particulièrement en Grèce mais aussi en Anatolie sont un peu moins anciens et datent d'environ 8 200 ans en Grèce, et 9 800 ans en Anatolie.

En Grèce, les échantillons du site de Mavropigi correspondent à la fois à des ossements identifiés ostéologiquement comme aurochs (cinq) et comme domestiques (six). Parmi les présumés aurochs, nous avons identifié deux de l'haplogroupe P, deux de l'haplogroupe Q et un de l'haplogroupe T (deux en capture et trois en aMPlex :voir chapitre I en annexes). Parmi les six échantillons identifiés comme domestiques, nous avons typé trois Q et trois T (quatre en capture, deux en aMPlex). La difficulté est qu'à ces dates-là, il n'est pas complètement clair qu'on puisse distinguer ostéologiquement sans faillir les sauvages et les domestiqués. Il est donc possible que les Q et les T associés aux aurochs ne soient pas des aurochs natifs mais des hybrides ou des domestiqués n'ayant pas encore une taille réduite bien que la distribution de taille des ossements dans ce site indique qu'il n'y avait que quelques individus de grande taille (S. Michalopoulou, A. Curci, communication personnelle). En Anatolie, les individus identifiés comme aurochs appartiennent aux haplogroupes Q, T et R3. L'haplogroupe que nous avons identifié comme R3 correspond à une région hypervariable qui a été identifié dans un spécimen allemand originaire d'Eilsleben et daté à 6 600 cal AP. Il y a donc clairement une incertitude quant à la localisation du refuge où les groupes Q-T auraient pu survivre. Etait-ce en Grèce ou en Anatolie ? Dans quelle direction les aurochs se seraient dispersés au début de l'Holocène ? On voit que le P de Mavropigi pour lequel on a obtenu le mitogénome complet est affilié aux aurochs P de France, d'Espagne et d'Italie et ses ancêtres se seraient vraisemblablement déplacés en provenance d'un refuge plus occidental. Est-ce que des Q et T de Mavropigi étaient endémiques ou se seraient déplacés, eux aussi, en provenance de refuges plus orientaux ? Il est difficile de trancher car la Grèce est vraisemblablement une zone d'interface des deux populations d'aurochs post-LGM et que la période considérée au tout début de la domestication rend délicate les interprétations ostéologiques fondées sur les différences de taille des ossements.

Depuis le Néolithique en Europe jusqu'au dernier aurochs recensé en Pologne datant du 17ème s. n.e., un certain nombre d'échantillons attribué aux aurochs, sur les critères de taille des ossements, porte en fait des mitotypes T3, et Q pour l'un d'entre eux. Si, pour certains os, la robustesse du critère d'attribution à des aurochs peut être discutée, pour d'autres on dispose de crânes (ou bucrânes) qui ont clairement des dimensions typiques d'aurochs.
L'haplotype T3 étant l'haplotype majoritaire au début du Néolithique en Europe de l'Ouest (voir plus loin) et l'haplotype encore majoritaire aujourd'hui, ces spécimens sont vraisemblablement des produits d'hybridations entre aurochs mâles et vaches domestiquées, bien que pas nécessairement de la génération F1.

On trouve de tels spécimens au Néolithique en Slovaquie sur le site de Svodin datés à 6 800 ans AP (SV2, SV9), en France (l'aurochs Néolithique NANT3 datant d'environ 4 900 ans AP), en Espagne (CF403 de la grotte « Cova Fosca » en Espagne de l'Est daté à environ 7 100 ans AP),puis au Chalcolithique dans le centre de l'Espagne datés de la deuxième moitié du 3ème millénaire av.n.e. (Camino de las Yeseras CY8, CY5, CY21, CY22), de Portugal daté entre le 4ème et 3ème millénaire av.n.e. (BOM). D'autres spécimens porteurs de mitogénomes T ou Q n'ont pas été clairement identifiés ostéologiquement comme appartenant aux aurochs et ne sont donc pas mentionnés ici. Finalement, les spécimens d'aurochs médiévaux originaires d'Alsace portaient tous des haplotypes T (notamment Ost446 du 10ème s. n.e.). Le dernier aurochs mâle connu date de 1620 et est mort dans la forêt de Sochaczewski en Pologne. La dernière aurochs femelle est morte, quant à elle, sept ans après dans la forêt de Jaktorów en Pologne. L'ADN d'une corne provenant du dernier mâle a été génotypé comme T3 (Bro-Jørgensen et al., 2018). Ceci montre que les derniers aurochs étaient déjà largement hybridés avec les bovins domestiqués.

En conclusion, avant le LGM, on trouve dans le sud de l'Europe une population ancestrale à tous les haplogroupes taurins de l'Holocène. Les aurochs européens de l'haplogroupe P correspondent aux individus sauvages qui se sont repartis au début de l'Holocène sur toute l'Europe de l'ouest en Grande-Bretagne, en Europe du nord jusqu'à la Grèce. Les aurochs des haplogroupes Q et T correspondent à la population qui a été domestiqués au début de l'Holocène en Asie du sud-ouest. Il n'est pas encore certain où cette population était réfugiée pendant le LGM, en Grèce, en Anatolie et/ou dans le Caucase. Elle était clairement connectée avec la population de l'Europe de l'ouest avant le LGM et, tout comme les aurochs P, elle s'est certainement redéployée rapidement à partir du refuge au début du réchauffement climatique, soit à la fin du Pléistocène supérieur juste avant le Dryas récent, soit après ce dernier au début de l'Holocène.

Notre échantillonnage dans cette région ne commençant que deux mille ans après cette période critique, les aurochs identifiés en Anatolie et en Grèce ont largement eu le temps de venir d'ailleurs. Nous avons détecté des aurochs P en Grèce au début du Néolithique (~8000 ans AP, Mavropigi), et même un peu plus tard (~6000 ans AP, Kouveleiki), mais ni nous ni d'autres n'ont jamais identifié d'aurochs P en Anatolie suggérant que la limite de l'extension des aurochs P pourrait se situer autour du Bosphore.

Les aurochs Q et T identifiés il y a 8000 ans à Mavropigi, pourraient soit correspondre à des aurochs locaux ou ayant naturellement passé le Bosphore avant qu'il ne s'ouvre, soit de manière catastrophique il y a 7600 ans (Ryan et al., 1997), soit de manière progressive à partir de 8000 ans AP (Ferguson et al., 2018). Comme la domestication avait déjà commencé à cette époque, il est aussi possible que les Q et T de Mavropigi aient été importés par les humains et que ce soit en fait des hybrides entre domestiqués et aurochs. La formation d'hybrides entre les premiers domestiqués et les aurochs est apparente lors de l'expansion néolithique en Europe. Ainsi on trouve des aurochs T3 dès le Néolithique en Europe centrale et de l'ouest, en France et en Espagne. On continue à en trouver au Chalcolithique en Espagne jusqu'au Moyen Âge en France et jusqu'au dernier aurochs en Pologne au 17ème s. n.e.

4) Première divergence au sein des taurins après la séparation avec les zébus : les haplogroupes mitochondriaux R et C

Après la séparation entre les zébus et les taurins, autour de 232 000 ans AP, une première série de radiations au sein du clade taurin, détectables parmi les membres ayant persisté jusqu'à l'Holocène, se produit autour de 100 000 ans AP (Figure 86). A 108 000 ans AP (95% HPD 123 000 – 95 000) on observe la radiation entre l'haplogroupe C et les autres haplogroupes. L'haplogroupe C a été trouvé dans des aurochs chinois (Zhang et al., 2013), celui ayant permis d'obtenir un mitogénome presque complet datant de 10 660 ans AP. Cet haplogroupe a persisté en Chine jusqu'à au moins entre 6 300 – 5 000 ans AP (Cai et al., 2018) mais n'est plus présent de nos jours. La radiation suivante est presque concomitante. Elle sépare les haplogroupes R et les P, Q, T et elle est estimé à 104 000 ans AP (95% HPD 118 000 – 92 000).

L'haplogroupe R a été détecté chez des bovins modernes italiens et deux haplogroupes, R1 et R2, ont été définis (Bonfiglio et al., 2010). Nous avons détecté en Anatolie, à Aşıklı Höyük, un membre très divergent de cet haplogroupe que nous avons appelé R3 dans le spécimen AS16, daté d'environ 8 700 ans AP (Figure 86). L'ancêtre commun entre R3 et la branche qui donnera lieu à R1 et R2 est lui daté à 98 000 ans AP (95% HPD 120 000 – 85 000). Cette date est aussi très proche des deux précédentes mais le support statistique pour l'ordre relatif de ces trois radiations est fort et elles sont observées aussi dans le même ordre sur les arbres de maximum de vraisemblance. L'haplogroupe R3 correspond à une séquence hypervariable qui a été observée dans un unique spécimen d'environ 6 600 ans AP à Eilsleben (Saxe-Anhalt, Allemagne) (Edwards et al., 2007a). Les auteurs avaient initialement appelé cet haplogroupe E sur la base du nom du site, mais il nous paraît plus pertinent de prendre en compte son affiliation avec l'haplogroupe R qui est clairement détectable avec le mitogénome complet mais qu'il est difficile de faire avec le petit bout de séquence hypervariable obtenue à l'époque.

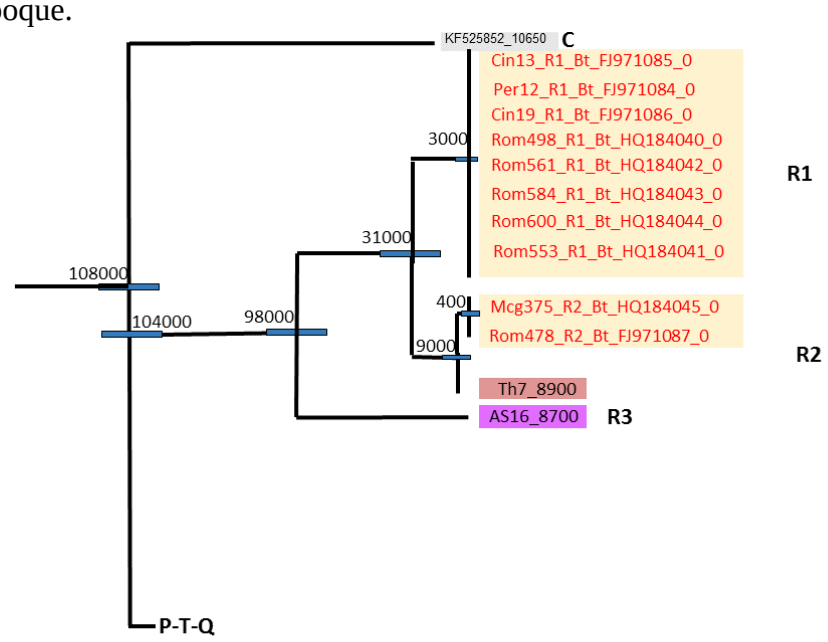


Figure 86: Arbre phylogénétique bayésien des lignées mitochondriales bovines d'haplogroupe R et C.

Lorsque les haplogroupes R1 et R2 ont été détectés dans le cheptel italien actuel, les auteurs ont proposé qu'ils provenaient d'une population nouvelle d'aurochs qui aurait pu se trouver dans les refuges de l'Europe du sud au LGM (Bonfiglio et al., 2010). Toutefois, le groupe de D. Bradley a détecté une séquence de type R2 dans un spécimen épi-paléolithique, Th7, daté d'environ 8 900 ans AP (Verdugo et al., 2019). Ceci suggère que la population d'aurochs qui portait les haplogroupes R1 et R2 était en Afrique du nord au début de l'Holocène. Le cheptel italien a une histoire connue d'introgression avec le cheptel africain qui a introduit une ancestralité provenant des zébus (Decker et al., 2014). Il est donc possible que les haplogroupes R1 et R2 ont été introduits beaucoup plus tardivement en provenance d'Afrique. L'ancêtre commun à tous les R1 date d'environ 3 000 ans AP (95% HPD 6 000 - 1 000) et celui des R2 italiens est seulement de 400 ans AP (95% HPD 2 300–0).

L'ancêtre commun au spécimen marocain de 8 900 ans et les R2 est de 9 000 ans AP (95% HPD 10 6000 – 8 900), c'est-à-dire, date de l'âge du spécimen ce qui indique qu'il est extrêmement proche de l'ancêtre de l'haplogroupe italien actuel. Ceci soutient fortement l'hypothèse d'une origine nord-africaine de cet haplogroupe et non d'une population d'aurochs italienne. Nos observations que les aurochs italiens du début de l'Holocène appartenaient à l'haplogroupe P corroborent aussi cette interprétation.

La période de radiation autour de 100 000 ans correspond à la fin de la période chaude du MIS 5e. La population ancestrale aux différents haplogroupes taurins se serait séparée à cette époque probablement en se répartissant dans différents refuges au moment du refroidissement qui a suivi, les aurochs donnant lieu à l'haplogroupe C allant vers l'Asie, les aurochs donnant lieu aux haplogroupes P, Q, T allant vers l'Europe du sud, et les aurochs donnant lieu aux haplogroupes R se répartissant entre l'Asie du sud-ouest et l'Afrique du nord, possiblement en se séparant en deux branches, R3 à l'est et l'ancêtre des R1 et R2 plus à l'ouest.

Le territoire dans lequel était la population ancestrale à cette époque, n'est pas clairement identifiable. Il était probablement différent de celui occupé à la même période par les aurochs des haplogroupes O car, comme discuté auparavant, on a observé une ségrégation territoriale entre ces différents haplogroupes 50 000 ans plus tard en Europe de l'ouest. L'élucidation de la dynamique des populations à cette époque-là requerrait un échantillonnage beaucoup plus important dans le pourtour de la Méditerranée du sud et de l'est sur les périodes entre 100 000 et 10 000 ans AP, ce qui est un défi majeur du fait de la très mauvaise préservation de l'ADN dans ces climats.

5) Origine de l’haplogroupe mitochondrial Q :

L’haplogroupe Q est proche de l’haplogroupe T mais possède une trentaine de SNPs caractéristiques qui le distingue de l’haplogroupe T. La répartition de ces SNPs est distribuée sur tout le mitogénome et il n’y a pas du tout de concentration au niveau de la région hypervariable, alors que ceci est généralement observé sur les différents mitogénomes. Ceci suggère que l’haplogroupe Q pourrait avoir une histoire évolutive particulière. On pourrait faire l’hypothèse qu’un haplogroupe ancestral des Q issu d’une divergence ayant accumulée des mutations ponctuelles sur le mitogénome incluant la séquence hypervariable aurait subi un événement très rare de recombinaison suite à une hétéroplasmie s’étant produite chez un individu portant à la fois cette séquence Q ancestrale et une séquence de la lignée T.

De tels événements sont considérés en première approximation improbable chez les animaux multicellulaires car il existe des mécanismes d’élimination des mitochondries paternelles par dégradation autophagique après la fertilisation (Rojansky et al., 2016). Par contre, il peut y avoir des rares cas où cette élimination est incomplète, entraînant une hétéroplasmie qui peut permettre la recombinaison (Ma & O’Farrell, 2015). Cela pourrait expliquer la distribution particulière de la variation génétique le long du mitogénome Q bien que l’on ne puisse pas complètement exclure que la région hypervariable était particulièrement contrainte d’un point de vue évolutif sélectivement au sein des lignées Q et T. Bien que la modélisation bayésienne servant à estimer les dates de divergence repose sur une évolution des séquences uniquement par mutations ponctuelles, la date de divergence entre Q et T que l’on peut estimer sur l’ensemble du mitogénome n’est probablement que légèrement sous-estimée car la région hypervariable concernée par cette évolution particulière ne fait que quelques centaines de paires de bases. Par contre, si l’on n’utilisait que la région hypervariable pour estimer cette divergence, on aurait des valeurs très différentes.

L’âge de l’ancêtre commun aux Q et T est estimé à 37 000 ans AP (95% HPD 45 000 – 30 000). Cette divergence se serait donc produite environ 20 000 ans après l’ancêtre commun aux séquences trouvées en Europe de l’ouest et discutées précédemment : LF1284, GO2423 qui en est proche, et probablement aussi CanO24, bien que ces deux dernières étaient insuffisamment couvertes pour être analysées dans la modélisation bayésienne (Figure 87).

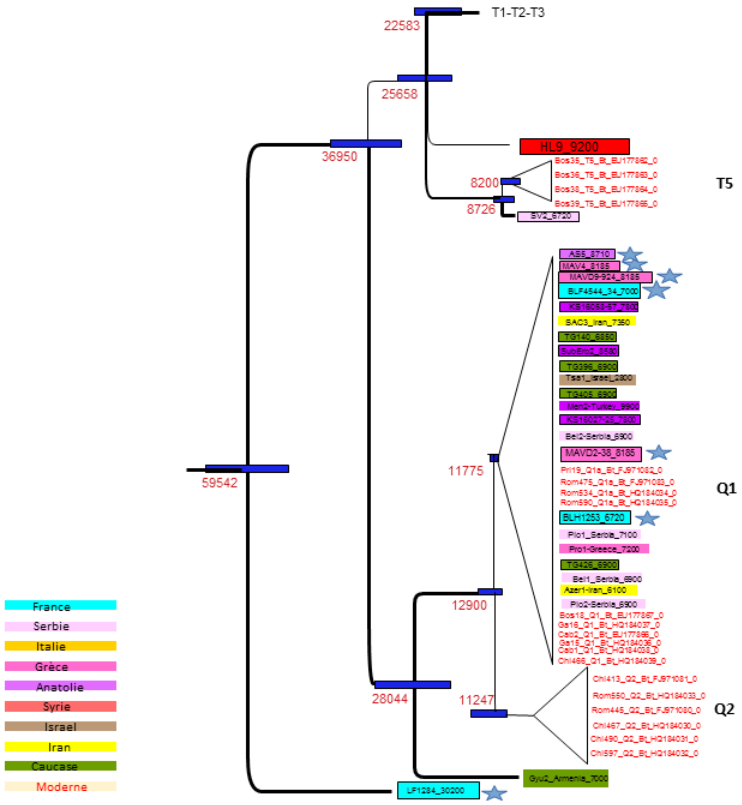


Figure 87: Arbre phylogénétique bayésien des lignées mitochondriales bovines d'haplogroupe Q. Les échantillons analysés dans notre étude sont encadrés.

L'haplogroupe Q est présent chez les bovins actuels et anciens et dans notre phylogénie nous avons 16 séquences modernes et 23 séquences anciennes. On détecte sa présence la première fois dans les aurochs anatoliens et les premiers domestiqués, à Aşıklı Höyük (AS5, ~9 700 ans AP), à Suberde ou Erbaba (SubErb2, ~8 580 ans AP), à Menteşe (Men2, daté de ~7 900 ans AP) (Verdugo, 2019), à Köşk Höyük (KS16058-57 et KS16027-25, ~7 800 ans AP). On les trouve aussi à Mavropigi, Grèce, approximativement à la même époque (MAV4, un présumé aurochs, et les domestiqués MAVD2 et MAVD9, tous ~8 250 ans AP, ainsi que deux autres spécimens identifiés par l'approche aMPlex dont un présumé aurochs).

De 7 000 à 5 000 AP, dans le Caucase, on trouve un aurochs en Arménie daté à environ 7 000 ans et originaire de Gyumri (Gyu2) et trois bovins domestiqués et un aurochs originaire de Géorgie (Tsiteli Gorebi) et datés, pour l'aurochs à ~6 850 ans (TG140) et pour les trois domestiqués (TG396, TG405, TG 426) autour de 6 000 ans. L'aurochs de Gyu2 (Verdugo et al., 2019) est très divergent des autres Q d'Anatolie ainsi que de tous les anciens analysés et tous les modernes. L'âge de l'ancêtre commun à Gyu2 et tous les autres Q est environ 28 000 ans AP (95% HPD 35 000 – 20 000) alors que la racine commune à tous les Q est de 12 900 ans AP (95% HPD 16 500 – 10 300), c'est-à-dire correspond à une radiation récente datant du début de l'Holocène. L'aurochs de Gyu2 doit donc correspondre à une lignée relique qui se serait séparée juste avant le LGM, qui a laissé peu de descendants après le LGM et qui n'aurait pas contribué au pool domestique maternel. Un Q a aussi été identifié chez deux bovins domestiqués en Iran à Tappeh-Sang-e-Chakhmaq (Sac3, ~7 350 ans AP) et à Kul Tepe (Azer1, ~6 100 ans AP) (Verdugo et al., 2019).

On retrouve les Q lors de l'expansion néolithique vers l'Europe au Néolithique récent en Grèce à Promachon (Pro 1, ~7 200 ans AP) et en Serbie à Pločnik et à Belovode-Veliko Laole (Plo1, Plo2, Bel1 et Bel2 à 6 900 ans AP) (Verdugo et al., 2019) et en France dans la Vallée de l'Aisne (BLF4544 à ~7 000 ans AP et BLH1253 à 6 720 ans AP). L'équipe a aussi identifié un autre Q dans la vallée de l'Aisne en utilisant l'approche aMPlex. Les Q sont toutefois rares et nous n'avons trouvé aucun en Espagne et dans le sud de la France alors que nous avons génotypé 80 spécimens en Espagne et 18 dans le sud de la France allant du Néolithique jusqu'au Moyen Âge. Ces différences suggèrent que les bovins domestiqués avec l'haplotype Q n'auraient pas participé à la voie de diffusion néolithique via la Méditerranée. L'haplotype Q est surtout représenté en Europe de l'est et en Asie de l'ouest. Aucune des séquences anciennes ne peut être attribuée à l'haplotype Q2 défini donc uniquement par six séquences modernes et la racine de ce groupe est estimé à 2 000 ans AP (95% HPD 4 200 – 200). Les 22 pour lesquelles nous avons des mitogénomes complets (neuf provenant de (Verdugo et al., 2019), les 13 autres ayant été produites dans notre étude) appartiennent toutes à l'haplotype Q1 ainsi que neuf séquences modernes. L'haplogroupe Q2 correspond donc à une radiation récente au sein du pool domestiqué et pas à un haplogroupe ancestral lié à l'origine du processus de domestication.

6) L'haplogroupe mitochondrial T :

L'haplogroupe T correspond à l'haplogroupe le plus représenté aussi bien au sein des populations modernes qu'anciennes. Cet haplogroupe peut se sous-diviser en quatre sous-groupes : T1, T2, T3, T5, défini d'abord sur la base de la région hypervariable (Troy et al., 2001), et ensuite sur la base des mitogénomes complets (Achilli et al., 2009). Un sous-groupe T4 a été proposé sur la base des séquences hypervariables de taurins asiatiques (Mannen et al., 2004), mais ce sous-groupe n'a pas de support statistique robuste avec les mitogénomes complets et peut être assimilé au sous-groupe T3. Aucune séquence ancienne complète actuellement obtenue correspond à cet haplogroupe, donc nous l'ignorerons ici. L'âge de la racine de l'haplogroupe T est de 26 000 ans (95% HPD 32 000 – 20 000). Toutes les lignées maternelles des bovins de l'Holocène de cet haplogroupe descendent donc d'une population ayant subi un fort goulot d'étranglement pendant le LGM.

La première radiation, au niveau de cette racine, sépare l'haplogroupe T5 des trois autres haplogroupes. Cet haplogroupe porte très peu de représentants (quatre séquences modernes) et une séquence ancienne correspondant à un spécimen (SV2) que nous avons analysé provenant de Svodin en Slovaquie et daté de ~6 700 ans AP. SV2 diverge des T5 actuels autour de 8 700 ans AP (95% HPD 11 000 – 7 600) et la racine des T5 actuels est à peu près chevauchante, autour de 8 200 ans AP (95% HPD 10 000 – 6 000), c'est-à-dire que le spécimen SV2 est très proche des ancêtres des séquences modernes. La séparation suivante concerne le spécimen HL9 originaire de Tell Halula en Syrie et daté à ~9 500 ans AP. Cette séparation date de 22 500 ans AP (95% HPD 27 000 – 18 000). Ce spécimen correspond donc à une séquence ancestrale qui n'a pas contribué au pool mitochondrial des bovins modernes. La radiation suivante correspond à la séparation de l'haplogroupe T2 des T1 + T3 qui s'est produite autour de 17 000 ans AP (95% HPD 21 000 – 13 000).

La séparation des T1 et T3 est datée à 15 200 ans AP (95% HPD 19 000 – 12 500) ce qui est très proche de la séparation précédente et le support statistique pour l'ordre dans lequel ces séparations se seraient faites n'est pas très bon. Il est donc raisonnable de les considérer approximativement simultanées. Les âges des nœuds incluant toutes les séquences connues T1, T2 et T3 sont pratiquement identiques : T1, 13 200 ans (95% HPD 16 500 – 10 500), T2, 13 200 ans AP (95% HPD 16 900 – 10 300) et T3, 12 600 ans 1P (95% HPD 15 400 – 10 600). Pour rappel, c'est aussi à cette période que le nœud de l'haplogroupe Q peut être daté (12 900 ans AP). On voit donc que tous ces haplogroupes se sont diversifiés au début de l'Holocène après être passés à travers un goulot d'étranglement sévère pendant le LGM. Effectivement, la reconstitution bayésienne des tailles effectives des populations avec l'approche « Bayesian skyline plot » montre que la taille effective de la population ancestrale pendant le Pléistocène final était faible, quelques dizaines de milliers d'individus, et qu'une expansion très forte se serait produite à l'Holocène de façon concomitante avec la domestication (Figure 88).

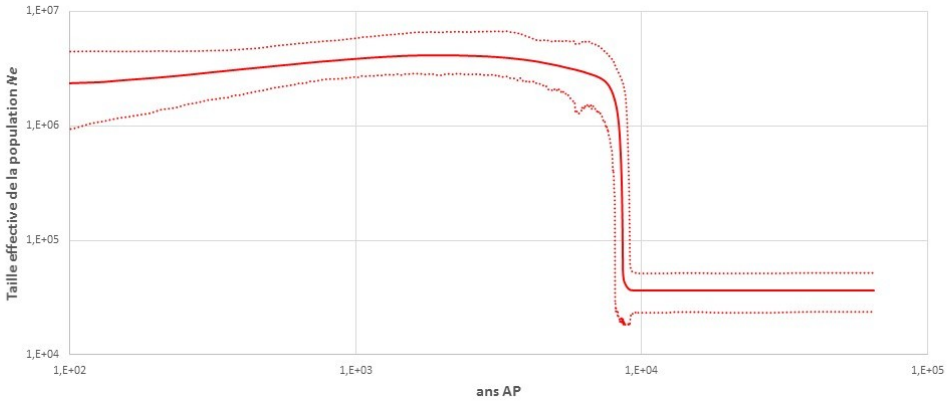


Figure 88 : « Bayesian skyline plot » correspondant à la modélisation de la taille de la population effective des haplogroupes T et Q à partir des mitogénomes obtenus (en trait plein la médiane, en pointillé les limites de l'intervalle de confiance à 95%).

Pour effectuer cette modélisation, nous avons considéré 539 séquences des haplogroupes Q et T comprenant 139 séquences anciennes en commençant à la racine de LF1284, le spécimen français discuté auparavant d'environ 30000 ans. Ne disposant d'aucun autre échantillon de cet haplogroupe entre 30 000 et 10 000 ans, nous avons très peu de résolution pour évaluer l'évolution des tailles des populations ce qui explique que la courbe est plane entre 10 000 ans et l'âge de la racine. Afin d'avoir une meilleure résolution pour évaluer l'évolution des tailles effectives des populations au Pléistocène récent, nous avons utilisé notre travail de génotypage des génomes modernes à partir des données du projet «1 000 Bull genomes».

Nous avons utilisé 220 génomes des races européennes en ne gardant que celles qui n'avaient clairement pas d'introgression provenant des zébus. Il s'agissait des races actuelles Holstein, Gelbvieh, Angus, Hereford, Jersey, Charolais, Limousin, Maine Anjou, Piémontaise, et quelques autres individus de France, du Portugal et de Grèce. Nous avons utilisé l'approche de modélisation smc++ (Terhorst et al., 2017). Cette modélisation donne un profil similaire avec celui obtenu par le mitogénome, mais les données génomiques permettent d'avoir une meilleure résolution au Pléistocène moyen/supérieur (Figure 89).
On observe un premier goulot d'étranglement entre 200 000 et 180 000 ans AP, correspondant à la période glaciaire MIS6, suivi d'un rebond de la taille de population pendant la période MIS5 commençant un nouveau déclin autour de 70 000 ans AP, correspondant à la période MIS4.
On ne voit pas de rebond pendant la période MIS3, mais plutôt un déclin régulier atteignant le niveau le plus bas, ici environ 10 000 individus, à la fin du LGM. Puis, la population augmente régulièrement jusqu'au présent (Figure 89).

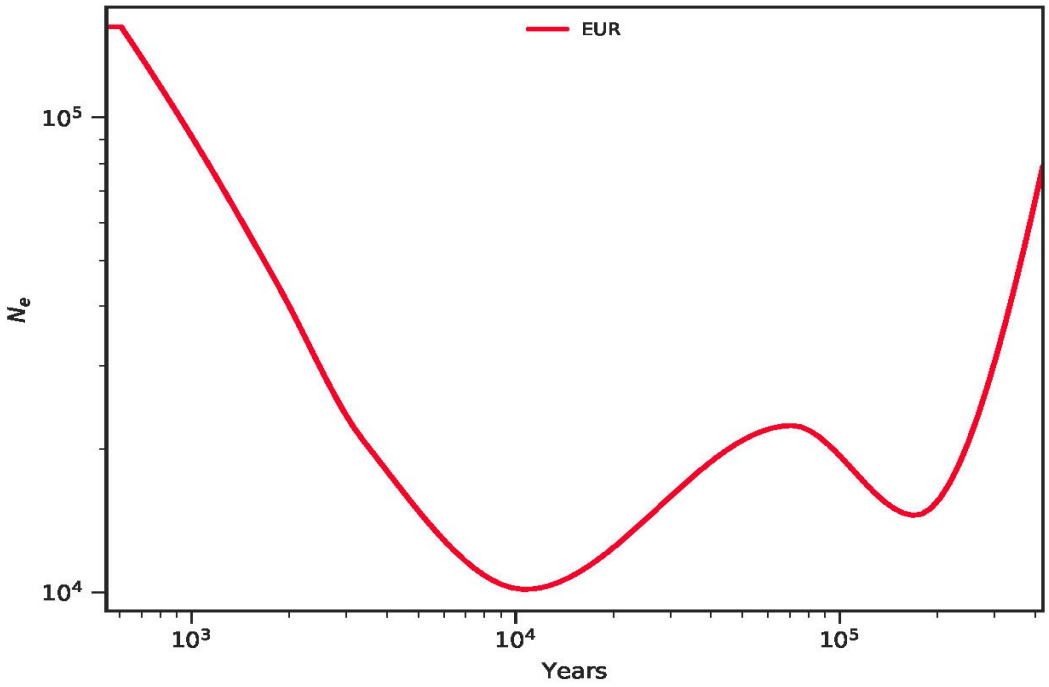


Figure 89: Modélisation des fluctuations de la taille effective des populations ancestrales à partir de 220 génomes de bovins actuels en utilisant l'approche smc++.

La différence notable entre les reconstitutions des tailles effectives des populations en utilisant l'analyse génomique et mitochondriale concerne la vitesse d'expansion des populations qui paraît plus progressive avec les génomes qu'avec les mitogénomes. Cette différence peut être due au fait qu'avec les mitogénomes nous avons interrogé des spécimens anciens distribués depuis le Caucase jusqu'à l'Espagne au cours des derniers 10 000 ans tandis que les génomes n'ont été obtenus essentiellement qu'à partir des races bovines d'Europe de l'ouest actuelles sélectionnés récemment, ce qui pourrait donner une image restreinte de la dynamique populationnelle sur l'ensemble du continent européen. Compte tenu des importantes différences entre les échantillons et le type de données utilisé (génomes nucléaires diploïdes versus mitogénomes), la congruence des estimations de l'expansion liée à la domestication et de l'impact du LGM sur le goulot d'étranglement est assez remarquable.

Sur la base d'une modélisation bayésienne du type ABC (approximate bayesian computation) et en utilisant une quinzaine de séquences hypervariables anciennes et 26 séquences hypervariables modernes couvrant à peu près 400 bases, Bollongino et al. ont proposé en 2012 que la domestication aurait impliqué 80 femelles (Bollongino et al., 2012). Cette modélisation reposait sur les valeurs de Fst, de Tajima D, et autres statistiques résumés comparant la diversité entre ces tout petits effectifs anciens et modernes. De plus, ils ont estimé que la taille effective de la population des aurochs d'Asie du sud-ouest était de 45 000 individus. Ici, en utilisant 400 séquences modernes de mitogénomes complets et 139 mitogénomes complets anciens, avec l'approche bayésienne, nous voyons un goulot d'étranglement lié au LGM qui laisse une population de 36 000 individus mais pas une réduction majeure associée au début de la domestication.

L'estimation de l'âge des racines des haplogroupes T1, T2, T3 et Q convergent vers une date autour de 13 000 ans correspondant à la fin du Pléistocène mais qui ne laisserait que 4 ancêtres communs uniques à tous les haplogroupes présents dans les domestiqués actuels à l'exception de ceux, beaucoup plus rares, appartenant aux haplogroupes R et P. Si la diversité ancestrale des populations d'aurochs d'Asie du sud-ouest était très élevée, on se serait attendu à trouver de nombreux cas de séquences différentes à ces quatre groupes (T1, T2, T3 et Q) parmi les aurochs au début de la domestication, alors que nous n'en trouvons que très rarement : R3 dans AS16, un T ancestral dans HL9 et un Q ancestral dans Gyu2 (trouvé par (Verdugo et al., 2019)). De plus, les séquences Q, P et T sont d'évolution très récente à partir d'une population d'aurochs précédant le LGM dont on trouvait quelques représentants en France et en Espagne (LP4, LF1284, CanO24, GO2423). Il n'y avait donc pas une large population d'aurochs très diverse en Asie du sud-ouest avant la domestication.

De plus, il paraît aussi improbable que les populations sauvages et domestiquées se sont séparées dès le début de la domestication sans qu'il y ait des flux génétiques entre ces deux groupes. Les données suggèrent plutôt un rôle primordial de la réduction de la taille des populations liée aux fluctuations climatiques du Pléistocène qui obscurcit considérablement la mise en évidence d'une réduction supplémentaire de la taille de la population liée à la domestication.

De plus, en utilisant les données génomiques de 220 individus actuels, on obtient aussi un goulot d'étranglement à la fin du Pléistocène autour de 10 000 individus ce qui soutient nos estimations bayésiennes basées sur l'ADN mitochondrial. La transition Pléistocène/Holocène et la domestication des aurochs étant à peu près concomitante, il paraît vain de vouloir chiffrer le nombre d'individus domestiqués en ignorant la dynamique des populations sauvages.

Revenons maintenant à l'analyse de l'évolution des populations des haplogroupes T1, T2 et T3. Pour l'haplogroupe T1, nous avons considéré 131 séquences actuelles et 11 séquences anciennes, 8 ayant été contribuées par notre étude. Pour l'haplogroupe T2, nous avons considéré 25 séquences actuelles et 17 séquences anciennes, 7 ayant été contribuées par notre étude. Pour l'haplogroupe T3, nous avons considéré 218 séquences actuelles et 90 séquences anciennes dont 69 ayant été contribuées par notre étude. T3 qui est l'haplogroupe majoritaire en Europe (Troy et al., 2001) est le plus représenté parmi les échantillons anciens (Figure 90).

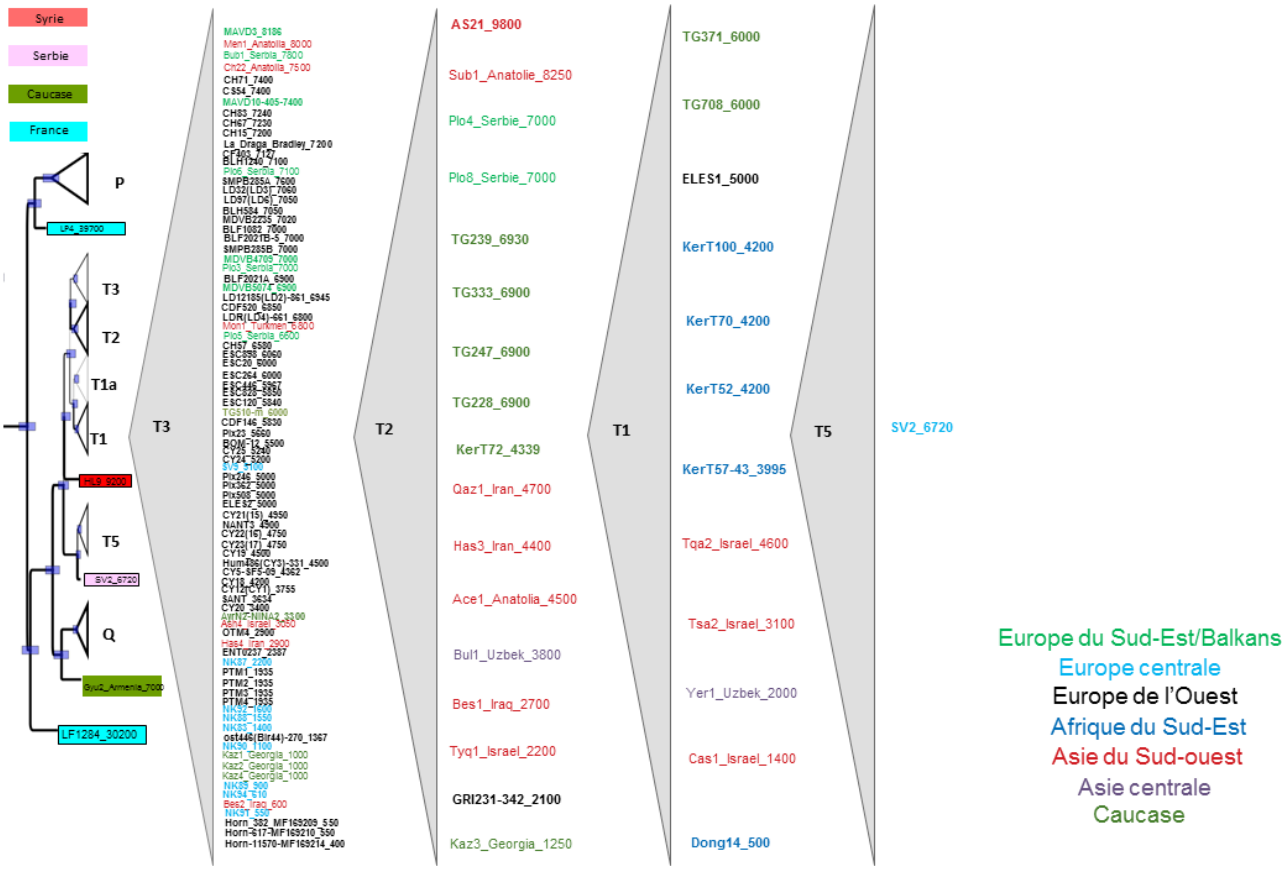


Figure 90: Arbre phylogénétique bayésien des haplogroupes mitochondriaux T. Les échantillons sont ordonnés selon leurs ages approximatifs et les couleurs correspondent à leurs origines géographiques.Les noms des échantillons en gras correspondent aux échantillons analysés dans la présente étude.

L'haplogroupe T1 étant majoritaire en Afrique de nos jours et est très peu représenté en Asie du sud-ouest, il a été proposé à l'époque qu'il pourrait correspondre à une population d'aurochs locale en Afrique du Nord-est qui aurait été domestiquée (Bradley et al., 1996; Hanotte, 2002; Troy et al., 2001), en accord avec une hypothèse émise en 2006 et basée sur des données archéologiques (Blench & MacDonald, 2006). Nos données ne corroborent pas cette hypothèse. L'équipe a génotypé la région hypervariable de deux spécimens de Mezraa Teilelat en Turquie du sud-est datant d'environ 8 800 ans AP comme appartenant à l'haplogroupe T1. Deux bovins domestiqués du site chalcolithique Tsiteli Gorebi en Géorgie et datant d'environ ~6 800 ans AP appartiennent aussi à cet haplogroupe, TG371 et TG708. On trouve les T1 plus tardivement en Israël (3 spécimens datés à environ 4 600, 3 100 et 1 400 ans AP) (Verdugo et al., 2019).

En Afrique elle-même, l'équipe a pu génotyper par approche aMPlex un aurochs T1 de la vallée du Nil (Hierakonpolis daté à environ 5 600 ans AP), et j'ai obtenu les mitogénomes de quatre échantillons de Kerma au Soudan datés à environ 4 200 ans AP (KerT52, KerT57, KerT70 et KerT100). Le cinquième échantillon de Kerma (KerT72) est un T2 et le sixième (KerT97) est un T3, génotypé par aMPlex. Beaucoup plus tard, au 14ème siècle de n.e., un échantillon de Old Dongola/KomA au Soudan (Dong14), est aussi un haplotype T1 bien que c'est beaucoup trop tardif pour être informatif.

Donc, l'haplotype T1 est assez tôt dominant en Afrique, mais nous l'avons mis en évidence tout de même ~3 000 ans plus tard que dans les échantillons originaires de la zone présumée de la domestication (J.-D. Vigne et al., 2005). Il a diffusé aussi bien vers le Caucase que vers l'Afrique sans que l'on puisse dire s'il a diffusé comme animal sauvage ou domestiqué, mais à ces dates, cela pourrait être une diffusion des domestiqués. On l'a trouvé aussi dans le centre de l'Espagne à El Espinillo, daté à environ 5-4 000 ans AP ce qui est un signe clair d'une translocation d'un animal domestique.

En conclusion, si T1 apparaît effectivement d'être un haplotype prédominant en Afrique au plus tard vers 4 200 ans AP, on ne peut pas conclure qu'il y a eu un événement local de domestication ou si c'est juste les aléas d'une diffusion d'un petit nombre d'animaux domestiques qui a induit les biais de représentation qui perdurent jusqu'à maintenant. Cette conclusion rejoint celle de Verdugo et al., 2019 qui déduit de leurs données que les bovins domestiqués en Afrique du nord-est auraient été domestiqués dans le Croissant fertile (Verdugo et al., 2019).

On trouve le T2 le plus ancien en Anatolie à Aşıklı Höyük (AS21, ~9 800 ans AP), et l'équipe de D. Bradley l'a trouvé un peu plus tard dans la même région à Suberde/Erbaba (Sub1) daté à environ 8 250 ans AP et elle continue à le trouver autour de 4 500 ans AP en Anatolie (Ace1) et en Iran (Has3 et Qaz1) (Verdugo et al., 2019). T2 est aussi très abondant à Tsiteli Gorebi en Géorgie autour de 6 800 ans AP où nous l'avons trouvé dans quatre sur six échantillons (TG228, 239, 247 et 333). L'équipe l'a aussi génotypé dans quatre spécimens datant à la période 5 500 - 3 200 ans AP à Didi Gora en Géorgie, et dans deux échantillons de la même période, un de Tsamakaberd en Arménie et un de Tell Mozan dans le nord de la Syrie.

Par ailleurs, l'haplotype T2 est resté dans la région du Caucase de l'Iraq et du Levant, en Ouzbékistan entre 3 800 et 2 200 ans AP (Bes1, Bul1, Tyq1) (Verdugo et al., 2019). Il était donc abondant dans le Caucase, mais a pu entrer dans le pool domestique plus tôt, en Anatolie éventuellement et a aussi été diffusé pendant la diffusion néolithique en Europe provenant de l'Anatolie puisqu'il a été trouvé dans deux échantillons de Pločnik en Serbie (Plo4 et Plo8) d'environ 7 000 ans AP (Verdugo et al., 2019), et nous l'avons identifié en France du Nord à l'Âge du Fer il y a ~2 100 ans AP (Gri231). Néanmoins, il a relativement peu contribué à la diffusion en Europe ce qui est reflété aussi dans sa distribution actuelle (Troy et al., 2001).

L'haplogroupe T3 est l'haplogroupe majoritaire en Europe d'où il a diffusé aussi bien sur le continent américain qu'en Asie et en Océanie (J. Lenstra et al., 2014b; Troy et al., 2001). Nous n'avons pas identifié des mitogénomes complets T3 dans les échantillons anatoliens anciens analysés, par contre, le génotypage réalisé par l'équipe a permis de le mettre en évidence dès 8 800 jusqu'à 7 800 ans AP (8 spécimens originaires de Mezraa Teleilat et Aşıklı Höyük en Turquie, de Tell Halula, Tell el Kerkh et Salat Cami Yani en Syrie et de Ein Zippori en Israël). L'équipe de D. Bradley a identifié deux mitogénomes complets en Anatolie entre 8 000 et 7 500 ans AP (à Çatal Höyük, CH22, et à Menteşe, Men1) (Verdugo et al., 2019). En Asie du sud-ouest, on en a trouvé des nombreux échantillons dans les périodes plus tardives, mais ce n'est plus pertinent pour comprendre la diffusion après la domestication.

Nous avons retrouvé l'haplotype T3 à peu près à la même époque en Grèce à Mavropigi à ~8 250 ans AP (MAVD3 et MAVD10), et Verdugo et al. en ont identifié un à Sarakenos (Sar38) à ~7 650 ans AP (Verdugo et al., 2019). Puis, il a été diffusé rapidement sur les sites néolithiques en Europe. Nous avons obtenu des mitogénomes complets à partir de dix spécimens en France du nord dans le Bassin Parisien datés autour de 7 100 ans AP, mais aussi en Espagne à Cueva de Chaves (cinq mitogénomes complets et huit mitogénomes génotypés par aMPlex) datés à ~7 200 ans AP, un mitogénome complet d'un spécimen de la Cova Fosca, également daté d'environ 7 200 ans AP, cinq mitogénomes complets de La Draga datés à ~7 060 ans AP, un mitogénome complet de Coro Trasito daté à environ 7 000 ans AP, deux mitogénomes complets et un mitogénome génotypé de la Cova del Frare datés à ~5 830 ans AP. Ce qui est remarquable en comparant ces plusieurs dizaines de mitogénomes très bien couverts c'est qu'au début de l'expansion néolithique, il y a très peu de variations entre les séquences et on obtient des séquences identiques depuis l'Anatolie, la Grèce, l'Espagne et la France du Nord, ce qui est en accord avec les données archéologiques et génétiques précédentes proposant que la domestication des bovins s'est effectuée en Asie du sud-ouest (voir Introduction).

Ensuite, nous avons suivi la répartition de ces séquences depuis le Néolithique moyen, le Chalcolithique, l'Âge du Bronze, l'Âge du Fer, les Antiquités jusqu'au Moyen Âge aussi bien en Espagne qu'en France, et on a obtenu des mitogénomes complets mais aussi génotypé des spécimens pour lesquels nous n'avons pas pu obtenir des mitogénomes complets. La diversité des séquences T3 des mitogénomes complets augmente légèrement au cours du temps. Il est aussi remarquable que pendant ces périodes plus tardives, on ne trouve que rarement d'autres haplogroupes. La domination de l'haplogroupe T3 en Europe s'est donc établi très précocement au cours de la migration néolithique.

7) Synthèse :

Notre étude phylogénétique basée sur l'ADNmt des bovins domestiques et de leur ancêtre, *Bos primigenius*, nous a permis de déterminer les liens évolutifs entre les différents haplogroupes mitochondriaux bovins et d'identifier un nouvel haplogroupe du Pléistocène supérieure et éteint depuis le LGM. Cette étude nous a permis d'explorer la diversité génétique des populations d'aurochs avant et après le processus de domestication et d'observer d'une part les liens entre les fluctuations climatiques et la démographie de ces populations et d'autre part les relations entre les différentes populations sauvages et celles qui ont contribué au pool génétique de la population bovine actuelle.

Nous avons construit un arbre phylogénétique robuste à partir de 950 mitogénomes complets de *Bos,* de *Bison* et de bovins asiatiques comprenant 280 mitogénomes anciens. Cet arbre nous a permis de déterminer les dates de radiations des différentes lignées mitochondriales des genres *Bos* et *Bison* en utilisant une approche bayésienne reposant sur la datation des feuilles (tip-dating).

Nous avons mis en évidence pour la première fois une nouvelle lignée d'aurochs qui a peuplé la Haute Vallée du Rhin il y a à peu près 50 000 ans. Cette lignée correspond à une divergence très ancienne antérieure à la divergence des zébus et des taurins qui reflète un peuplement ancien de l'Europe probablement en provenance du sous-continent indien qui aurait eu lieu dans une période comprise entre 390 000 et 340 000 ans AP. Cette lignée ne semble pas avoir survécu au refroidissement qui a suivi cette période.

A la même période, dans le sud de l'Europe de l'ouest, nous avons identifié quelques aurochs dont le mitogénome est ancestral aussi bien aux aurochs de l'haplogroupe P que l'on retrouvera au début de l'Holocène dans l'Europe de l'ouest qu'aux aurochs des haplogroupes Q et T qui correspondront à la population domestiquée en Asie du sud-ouest. Ceci montre qu'avant le dernier maximum glaciaire, il y avait une population d'aurochs dans la Vallée du Rhin qui n'était apparemment pas connectée avec la population du sud de l'Europe alors que cette population du sud devait pouvoir se déplacer tout autour du Bassin Méditerranéen assurant une continuité génétique avant la séparation de ces populations au moment du LGM.

En plus de ces deux populations identifiées il y a 50 000 ans, l'analyse des mitogénomes du début de l'Holocène met en évidence deux autres lignées, R et C, qui auraient divergé des autres populations d'aurochs il y a 110 000 et 100 000 ans AP. N'ayant pas identifié de spécimens de ces lignées antérieurs à 10 000 ans, on ne peut que spéculer sur le territoire qu'ils ont occupé depuis leur divergence sur la base des lieux où ils ont été identifiés il y a 10 000 ans AP. Nous avons caractérisé le plus ancien membre de la lignée R à peu près 10 000 ans et l'autre membre de cette lignée a été identifié au Maroc presqu'à la même époque. Cette lignée pourrait donc correspondre à une population s'étendant sur tout le pourtour sud de la Méditerranée, au moins à la sortie du LGM. La lignée R a survécu jusqu'à nos jours mais ne sont observés que deux sous-groupes ayant chacun des racines très récentes (3 000 ou 1 000 ans AP environ).

La seconde lignée qui a divergé presqu'au même moment, autour de 100 000 ans, a été retrouvée en Chine il y a aussi à peu près 10 000 ans AP. Cette lignée n'a pas persisté dans les bovins actuels. Ces deux lignées sont les témoins survivants d'un goulot d'étranglement qui a affecté les populations de aurochs autour de l'avant-dernière période de glaciation.

Le LGM a sévèrement affecté les populations d'aurochs qui se sont refugiés dans des refuges déconnectés, qui ont subi des goulots d'étranglement sévères et qui se sont ensuite reparti et ont recolonisé les territoires au moment du réchauffement climatique de la fin du Pléistocène bien que cette expansion a dû être fortement perturbée par le refroidissement brutal du Dryas récent. A l'ouest de l'Europe, l'haplogroupe P qui s'était refugié autour de la France du sud, de la péninsule ibérique, de l'Italie a repeuplé toute l'Europe de l'ouest et du nord donnant lieu aux aurochs qu'on a retrouvé aussi bien en Angleterre, au Danemark, qu'en Pologne et jusqu'en Grèce. La chasse et la compétition avec les bovins domestiqués a affecté cette population jusqu'à sa disparition au 17ème s., les derniers membres identifiés étant déjà hybridés avec des bovins domestiqués. La lignée P a survécu jusqu'à nos jours en Asie du nord-est grâce à un événement d'introgression dans le cheptel domestique.

Les lignées Q et T qui ont dû se refugier au moment du LGM dans l'est de l'Europe et/ou dans l'Asie du sud-ouest correspondent à la population qui a été domestiquée et qui a subi une formidable expansion démographique grâce à cette domestication.
Ces populations ont subi un goulot d'étranglement sévère lors du LGM et il n'est pas complètement clair qu'il y ait eu un goulot d'étranglement supplémentaire lié au processus de domestication.
En effet, la domestication a commencé au moment de l'amélioration climatique, simultanément à ce qui a permis aux populations d'aurochs de repartir, comme on a pu l'observer pour les populations de l'Europe de l'ouest qui n'ont pas été domestiquées. L'expansion des populations a toutefois été beaucoup plus forte une fois que la domestication a été initiée.

Lors de l'expansion néolithique des premiers agriculteurs d'Anatolie vers l'Europe de l'ouest, les bovins domestiques ont été diffusés dans toute l'Europe, mais ces événements de diffusion ont dû faire intervenir un tout petit nombre de bovins au départ, ou il y a eu des goulots d'étranglement sévères liés à cette expansion, car on observe une faible diversité au début du Néolithique avec quelques séquences que l'on retrouve identiques de l'Anatolie à l'Espagne ou au Nord de la France. Cette histoire initiale laisse encore des traces importantes de nos jours car ce même haplotype (T3) est toujours l'haplotype majoritaire en Europe. On observe une dynamique d'expansion similaire lors des migrations néolithiques vers l'Afrique du nord et le long de la vallée du Nil avec un autre haplogroupe (T1) qui domine dès les phases précoces bien que notre échantillonnage n'a pas la même profondeur temporelle dans cette région du fait de l'impact négatif des hautes températures sur la préservation de l'ADN.

Perspectives

Mon travail a consisté à extraire l'ADN d'un grand nombre d'ossements de bison, d'aurochs et de bovins domestiqués. Ceci m'a permis d'identifier les ossements les mieux préservés pour les sélectionner pour un séquençage à grande profondeur. En parallèle, j'ai effectué la capture des mitogénomes d'un maximum d'échantillons, même de ceux qui étaient trop mal préservés pour être analysables par séquençage génomique aléatoire à haute couverture. Ce sont ces mitogénomes et leurs analyses qui représentent le cœur de ma thèse présentée dans le manuscrit.

En parallèle, afin de définir les bonnes conditions d'analyse des génomes anciens en cours de production, j'ai travaillé à la mise au point des méthodes optimales pour faire l'appel des SNPs à partir de ces génomes anciens en utilisant le premier génome bovin ancien produit dans le laboratoire, provenant d'un spécimen appartenant à la population d'aurochs initialement domestiquée. Particulièrement, j'ai mis en évidence l'intérêt d'utiliser le logiciel GATK pour faire l'appel des SNPs et d'effectuer une recalibration de la qualité des bases en utilisant le jeu de données de référence du projet « 1000 bull genomes ».

Finalement, j'ai aussi collecté plus de 500 données publiques de séquençage de taurins, de zébus, de gaurs, de gayal, de bisons et de yaks actuels pour les cartographier sur le génome bovin afin de constituer une base de données de référence des *Bovina* actuels à laquelle nous pourrons comparer les génomes anciens produits. La production de ces données, et surtout leur analyse, prenant beaucoup de temps, il ne m'a pas été possible de mener à bien l'analyse génomique complète dans le temps imparti pour cette thèse. En effet, l'étape de génotypage joint de près de 600 individus peut prendre plusieurs mois, sans compter le temps nécessaire pour produire les fichiers gvcf (génotypage de toutes les positions du génome) pour chaque individu qui sont nécessaires pour faire le génotypage joint. Les séquençages des banques génomiques que j'ai produites sont toujours en cours et donc l'analyse de toutes ces données ne pourra commencer que quelques mois après ma soutenance. Mon travail de thèse est donc une étape fondatrice pour un projet qui continuera à vivre et à se développer après mon activité au laboratoire.

De même, mon travail de thèse a été construit sur les fondations issues du travail de thèse de Diyendo Massilani. J'ai pu le prolonger dans des multiples directions. Ce qui a fait le succès de ma contribution, ce sont à la fois les améliorations méthodologiques qui ont été apportées aussi bien par mon travail que par le travail effectué en parallèle par d'autres membres du laboratoire travaillant sur d'autres systèmes modèles mais aussi par la contribution active à ces améliorations de mes encadrants. J'ai pu aussi bénéficier dans la deuxième moitié de ma thèse de l'arrivée de nouveaux échantillons particulièrement intéressants correspondant aux périodes précédant le dernier maximum glaciaire car les analyses effectuées à partir du travail de Diyendo Massilani que j'avais complété pendant la première moitié de ma thèse montraient qu'il était essentiel pour nous d'étudier cette période pour bien comprendre la dynamique des populations d'aurochs précédant leur domestication. Ceci nous permet donc d'avoir un socle solide décrivant l'évolution de la lignée maternelle des aurochs pendant les derniers 50 000 ans et a fourni la matière pour analyser leurs génomes nucléaires. La fin d'une étape est donc aussi le commencement de l'étape suivante, comme c'est toujours le cas dans l'activité de recherche.

L’analyse des génomes que j’ai produits ici permettra d’éclairer de façon originale le processus de transformation de la forme sauvage vers la forme domestique, en disposant d’un jeu de données de taille suffisante pour démêler les effets confondants des changements démographiques liés au changements climatiques et environnementaux majeurs qui ont précédé la domestication. En effet, la coïncidence temporelle du passage du statut de chasseur-cueilleur à agriculteur et donc de l’invention du Néolithique avec l’amélioration climatique du début de l’Holocène est frappante. Il est donc très probable que cette amélioration climatique ait influencé aussi profondément le mode de vie des populations humaines qu’elle ait impacté la dynamique des populations animales.

De ce fait, cette coïncidence temporelle complique considérablement l’interprétation des changements génétiques qui se sont produits à cette époque car certains de ces changements peuvent avoir été causés par les changements de taille de populations tandis que d’autres ont résulté du processus de sélection effectué, plus ou moins consciemment, par les êtres humains. L’étape suivant ma thèse s’avère donc à la fois passionnante et complexe. Je me réjouis d’avoir pu contribuer ce jeu de données dont l’analyse offre tant de perspectives.

Annexes

Annexe I: Analyse phylogéographique des données de la PCR multiplexe :

I) Typage génétique des échantillons anciens et répartition géographique :

Les séquences d'ADN obtenues par PCR multiplex avec la stratégie aMPlex (Guimaraes et al., 2017) sont nettoyées des séquences d'adaptateurs et celles de mauvaise qualité avec le logiciel cutadapt et sont ensuite alignées sur un génome artificiel correspondant au concaténat des fragments de PCR amplifiés avec les versions locales et globales du logiciel bwa (Li & Durbin, 2009). Un comptage des séquences alignées a été fait et les haplogroupes mitochondriaux ont été déterminés. Les résultats obtenus sont présentés sur la carte géographique suivante :

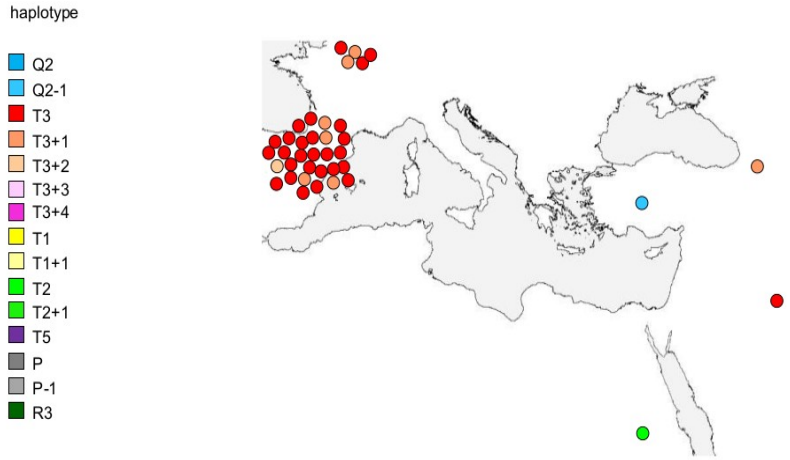


Figure 91: Distribution des haplogroupes mitochondriaux (selon le code couleur) trouvés dans les restes bovins du Néolithique et de l'Age des métaux.

Nos résultats montrent que les vestiges fossiles trouvés en France sont d'haplogroupe mitochondrial T3 mais qu'ils diffèrent entre eux d'un ou de deux nucléotides, alors que l'échantillon de Turquie de Soudan d'Iraq et d'Iran sont respectivement d'haplogroupes mitochondriaux Q2, T2, T3 et T3 avec un nucléotide diffèrent de l'haplogroupe T. Nos résultats sont en accord avec ce qui a été publié concernant les haplogroupes mitochondriaux et leurs origines (Achilli et al., 2008, 2009) mais on a observé que la diversité génétique de deux sites archéologiques en Espagne est différente. En effet, pour le premier site, la variabilité génétique était réduite puisque tous les échantillons ont le même haplogroupe mitochondrial alors que pour le deuxième site, la diversité génétique était plus élevée.

II) Conclusion :

Les résultats de l'aMPlex ont permis de génotyper un nombre élevé d'échantillons anciens, et d'élargir la gamme des échantillons analysés pour permettre une étude approfondie et élargie de l'origine des bœufs domestiques en Europe. Une différence de diversité génétique a été observée dans deux sites différents en Espagne. Ceci suggère une diversité des origines du cheptel bovin selon les sites néolithiques espagnols.

PCR multiplexe associée au séquençage de nouvelles générations				
Genre	Nom de l'échantillon	Origine	Site archéologique	Chronologie
Genre Bos	CDF0	Espagne	Cova del Frare	Néolithique ancien
	CDF256			
	CDF343			
	CDF520			
	CH15		Chaves	Néolithique
	CH16			
	CH19			
	CH48			
	CH56			
	CH57			
	CH67			
	CH71			
	CH83			
	CH98			
	CH266			
	CH386			
	CR277		Can Roig	Période romaine
	CR350			
	CS54		Can Sadurni	Néolithique moyen
	CS97			
	CT286		Cora Trasito	Néolithique
	Gava13		Mines de Gavà	Fin Néolithique_Chalcolithique
	MARS44		Marsilla	Age de bronze
	MARS45			
	MO32		Olvena	Néolithique
	OV546			
	PIX23		Cova de les Pixarelles	Néolithique moyen
	PIX104			
	PIX204			
	PIX246			
	PIX263			
	PIX362			
	PIX481			
	PIX483			
	PIX508			
	PIX556			
	RA683		Reina Amàlia	
	RA4537			
	RA6769			
	RA1196			
	RA10622			
	RA10902			
	ERB5A	Turquie centrale	Erbaba Höyük	Fin Néolithique_Chalcolithique
	ERB5B			
	ERB72			
	ERB227			
	ERB264			
	ERB295			
	ERB319			
	HL15	Vallé d'Euphrates	Tell Halula	Néolithique ancien
	HL16			
	HL17			
	HL18			
	HL19			
	HL21			
	HL22			
	HL23			
	HL25			
	HL26			
	HL27			
	HL28			
	HL76			

Tableau 12: Noms, origines et chronologies des échantillons anciens.

Annexe II : Description des échantillons anciens inclus dans l'analyse bayésienne des mitogénomes

Les échantillons inclus dans l'analyse bayésienne sont présentés dans le tableau suivant. Le nom, l'âge approximatif, l'origine géographique ainsi que le pourcentage d'ADN endogène sont représentés pour chaque échantillon et les échantillons sont groupés selon l'haplogroupe mitochondrial d'aurochs ou de bovins domestiques correspondant.

Échantillon fossile	Haplotype mitochondriale	Âge approximatif	Site archéologique	Origine	Cadre chronologique/contexte archéologique	% ADN endogène de banques génomiques
LP4	P	39700	Les Plumettes	France	Pleistocène supérieur	46.22%
CoupeGorge		16050	Coupe Gorge		Pléistocène supérieur	69.74%
LPM4		10560	Les Pradelles (Marsat) St4		Mésolithique	2.57%
BFT951		7100	Bucy-le-long		Néolithique	8.27%
BGM310-311		6500	Bucy-le-long le Grand marais		Néolithique	70.43%
NANT650 (NANT2)		4520	Saint-Nazaire/Loire Atlantique		2670 - 2470 calBCE (90.8%)	79.84%
NANT 657 (NANT1)		4400	Ancenis/Loire Atlantique		2570 - 2290BC calBCE (95.4%)	70.72%
CHL21		5100	Chalain		Néolithique récent	2.91%
SMA161B		9500	Santa Maria di Agnano	Italie	Mésolithique	5.78%
SMA28		9500	Santa Maria di Agnano		Néolithique	1.37%
SMA149		9140	Santa Maria di Agnano		Néolithique	31.88%
DMAD2		9000	Grotta della Madonna		Mésolithique	2.93%
UZ1		8800	Grotta dell'Uzzo		Mésolithique	0.52%
UZ2		8800	Grotta dell'Uzzo		Mésolithique	1.19%
UZ3		8800	Grotta dell'Uzzo		Mésolithique	0.52%
MAV2		8185	Mavropigi	Grèce	Néolithique	0.35%
Kouv1		6000	Kouveleiki B		Néolithique	0.33%
LF1284	ancêtre Q et T	30200	La Flageolet	France	Pléistocène supérieur	40.24%
AS5	Q1	8710	Aşıklı Höyük	Turquie	Néolithique	25.44%
Sub-Erb2		8580	Suberde-Erbaba		Néolithique	93.81%
BLF4544		7000	Bucy-le-long	France/Aisne	Néolithique	1.73%
TG396		6900	Tsiteli Gorebi 5	Géorgie	Chalcolithique	0.68%
TG405		6900	Tsiteli Gorebi 5		Chalcolithique	8.91%
TG140		6850	Tsiteli Gorebi 5		Chalcolithique	1.45%
TG426	Q2	6900	Tsiteli Gorebi 5		Chalcolithique	1.56%
BLH1253		6720	Bucy-le-long	France/Aisne	Néolithque	48.23%
HL9	T	9200	Tell Halula	Syrie/Vallée de l'Euphrate	Néolithique	0.09%
TG371	T1	6900	Tsiteli Gorebi 5	Géorgie	Chalcolithique	0.68%
KerT100		4200	Kerma	Soudan	Kerma	31.54%
KerT70		4200	Kerma		Kerma	3.95%
KerT52		4200	Kerma		Kerma	0.80%
KerT57		3995	Kerma		Kerma	0.22%
Dong14		500	Dongola		~500	0.25%
ELES1		5000	El Espinillo	Espagne	Chalcolithique	1.42%
TG339	T2	6900	Tsiteli Gorebi 5	Géorgie	Chalcolithique	95.59%
TG247		6900	Tsiteli Gorebi 5			22.45%
TG239		6930	Tsiteli Gorebi 5			95.59%
TG228		6900	Tsiteli Gorebi 5			2.27%
KerT72		4339	Kerma	Soudan	Kerma	89.52%
SV2	T5	6720	Svodin	Slovaquie	Néolithique	0.31%

Tableau 13 : Description des échantillons appartenant aux haplogroupes mitochondriaux P, Q, T1, T2 et T5.

Échantillon fossile	Haplotype mitochondriale	Âge approximatif	Site archéologique	Origine	Cadre chronologique/contexte archéologique	% ADN endogène de banques génomiques
CS54	T3	7400	Can Sadurní	Espagne, Begues, Barcelona	Néolithique moyen	0.73%
CH15		7240	Chaves	Espagne/Huesca	Néolithique	16.66%
CH83		7240	Chaves		Néolithique	26.57%
CH57		7200	Chaves		Néolithique	0.62%
CH71		7200	Chaves		Néolithique	13.49%
La-Draga-Bradley		7200	La Draga	Espagne, Banyoles, Girona	Néolithique ancien	27.36%
BLH1240		7100	Bucy-le-long	France/Aisne	Néolithque	81.10%
SMPB285A		7060	Saint Martin sur le Prés	France/Champagne	Néolithique	93.00%
SMPB285B		7060	Saint Martin sur le Prés		Néolithique	69.16%
BLH584		7050	Bucy-le-long	France/Aisne	Néolithque	2.74%
LD97(LD6)		7050	La Draga	Espagne, Banyoles, Girona	Néolithique	42.52%
MDVB2235		7020	Menneville derrière le village	France/Aisne	Néolithique ancien	49.70%
MDVB4709		7000	Menneville derrière le village		Néolithique	29.64%
BLF2021		6900	Bucy-le-long		Néolithique	72.72%
TG510-m		6900	Tsiteli Gorebi 5	Géorgie	Chalcolithique	0.71%
MDVB5074		6890	Menneville derrière le village	France/Aisne	Néolithique	83.94%
CDF520		6850	Cova del Frare	Espagne, Matadepera, Barcelona	Néolithique ancien	0.22%
CH67		6230	Chaves	Espagne/Huesca	Néolithique	4.61%
ESC898		6060	Escalles Mont d'Hubert	France/Nord Pas de Calais	Néolithique moyen	95.85%
ESC20		6000	Escalles Mont d'Hubert		Néolithique moyen	35.08%
ESC264		6000	Les Escalles		Néolithique moyen	92.46%
ESC446		5967	Les Escalles		Néolithique moyen	91.68%
ESC828		5850	Escalles Mont d'Hubert		Néolithique moyen	67.37%
ESC120		5840	Escalles Mont d'Hubert		Néolithique moyen	84.80%
Pix23		5660	Cova de les Pixarelles	Espagne, Banyoles, Girona	Néolithique moyen	0.63%
Pix246		5600	Cova de les Pixarelles		Néolithique moyen	3.32%
Pix362		5600	Cova de les Pixarelles		Néolithique moyen	0.57%
Pix508		5600	Cova de les Pixarelles		Néolithique moyen	1.98%
BOM		5500	Bombard	Portugal	Chalcolithique	1.47%
CY24		5200	Camino de las Yeseras	Espagne/Madrid	Chalcolithique	62.34%
SV9		5100	Bobenheim-Roxheim	Slovaquie	Néolithique	11.23%
ELES2		5000	El Espinillo	Espagne	Chalcolithique	0.29%
(NANT308 avant NANT3)		4900	Penhouët Basin Saint-Nazaire	France/Loire Atlantique	~ 4900	67.62%
CY25		4540	Camino de las Yeseras	Espagne/Madrid	Chalcolithique	78.51%
CY5-SF5-09		4362	Camino de las Yeseras		Chalcolithique	93.05%
SANT		3634	Santioste	Espagne	Âge du Bronze	93.51%
NK87		2200	Bobenheim-Roxheim	Haute Vallée du Rhin	Âge du Fer	34.33%
PTM1		1935	Les Trois Frères	France/Ile-de-France	Âge du Fer	71.61%
PTM2		1935	Les Trois Frères		Âge du Fer	61.40%
PTM3		1935	Les Trois Frères		Âge du Fer	48.12%
PTM4		1935	Les Trois Frères		Âge du Fer	68.45%
NK92		1600	Bobenheim-Roxheim	Haute Vallée du Rhin	Âge du Fer	50.41%
NK88		1550	Bobenheim-Roxheim		Âge du Fer	86.61%
NK93		1400	Bobenheim-Roxheim		~1400	91.20%
NK90		1100	Bobenheim-Roxheim		~1100	85.61%
NK89		900	Bobenheim-Roxheim		~900	53.51%
NK94		610	Bobenheim-Roxheim		~610	91.77%
NK91		550	Bobenheim-Roxheim		~550	50.41%

Tableau 14 : Description des échantillons appartenant à l'haplogroupe mitochondrial T3.

Annexe III : Lignes de commandes utilisées pour le l'appel des SNPs

Analyses bio-informatiques des données de séquençage :

1) Traitement des données génomiques :

a) Fusion des séquences des deux brins (merging) et contrôle qualité des séquences :

Les résultats de séquençage sont transférées du séquenceur à nos ordinateurs sous forme de fichier « .fastq ». Les séquences des deux brins d'un même fragment sont fusionnées, les proportions des séquences de mauvaise qualité, des dimères d'adaptateurs sont quantifiées pour chaque banque en utilisant le logiciel leeHom (Renaud et al., 2014) et en suivant les recommandations des programmeurs.

b) Filtration des séquences de taille inférieure à 28 bases :

Nous sélectionnons les séquences fusionnées obtenues ayant une taille supérieur à 28pb grâce à la commande « awk » en utilisant comme fichier de départ, le fichier « fq.gz ».

c) Alignement aux génomes de référence :

Une fois les séquences fusionnées et sélectionnées pour garder que celles supérieur à 28 bases, ces dernières sont alignées aux génomes de référence grâce à l'aligneur bwa ou aln.
La séquence du génome complet bovin (ARS-UCD1.2_Btau5.0.1Y.fa) est téléchargée sur le site du Consortium International des 1000 Génomes Bovins (http://www.1000bullgenomes.com/).
Les génomes mitochondriaux de référence des différents haplogroupes de bovins domestiques et d'aurochs européen utilisés pour aligner les séquences des banques enrichies ont été téléchargés de la banque de donnée génomique GenBank. Des traitements post-alignement comme l'élimination des dupliqués, la distribution de taille, le contenu en GC… sont effectués avec Samtools 1.9 (https://github.com/samtools).

d) Reconstruction des génomes mitochondriaux :

Après avoir aligné les séquences mitochondriales capturées aux différentes séquences de référence de génomes mitochondriaux, elles sont transférés dans le logiciel de traitement de données génétiques Geneious 8.1.9 (http://www.geneious.com, (Kearse et al., 2012). Les couvertures des génomes sont déterminées et les séquences de mauvaise qualité sont éliminées. Les séquences consensus des génomes de différents échantillons sont extraites.

e) Authentification des séquences d'ADN ancien :

L'authentification des séquences d'ADNa alignées à des génomes de référence est effectuée grâce au programme MapDamage (Ginolhac et al., 2011a; Jónsson et al., 2013) selon la commande suivante :

mapDamage -i nom_de_l’échantillon.bam -r ARS-UCD1.2_Btau5.0.1Y.fa -l 28 -Q 10 -d MAPDAMAGE

2) Appel du polymorphisme :

a) Appel et filtration de SNPs avec GATK :

Séléction des variants avec comme exemple une qualité d’alignement égale à 10 :

gatk HaplotypeCaller -I nom_de_l’échantillon.bam -O all_GATK_10.vcf -R ../ARS-UCD1.2_Btau5.0.1Y.fa --min-base-quality-score 10 --minimum-mapping-quality 10 -ERC "NONE"
gatk SelectVariants -R ARS-UCD1.2_Btau5.0.1Y.fa -V nom_de_l’échantillon_GATK_10.vcf -O nom_de_l’échantillon_GATK_10.SNP.vcf --select-type SNP

Filtrage des SNVs selon les filtres recommandés par GATK :

gatk VariantFiltration -R ../ARS-UCD1.2_Btau5.0.1Y.fa -V nom_de_l’échantillon_GATK_10.SNP.vcf -O nom_de_l’échantillon_GATK_10.SNP.prefilt.vcf --filter-expression "QD < 2.0 || FS > 60.0 || MQ < 40.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0" --filter-name "hard_filtering_snv"

Sélection des variants passant le filtre :

gatk SelectVariants -R ../ARS-UCD1.2_Btau5.0.1Y.fa -V nom_de_l’échantillon_GATK_10.SNP.prefilt.vcf -O nom_de_l’échantillon_GATK_10.SNP.filtered.vcf --exclude-filtered
bgzip -c a nom_de_l’échantillon_GATK_10.SNP.filtered.vcf > nom_de_l’échantillon_GATK_10.SNP.filtered.vcf.gz
tabix -p vcf nom_de_l’échantillon_GATK_10.SNP.filtered.vcf.gz

b) Appel de SNPs avec Varscan :

Conversion du fichier d’alignement “bam” en format “mpileup” car l'affichage de varscan se fait par mpileup :

Tester le controle qualité de base (q10, q30) :

samtools mpileup -q 10 -B -f ARS-UCD1.2_Btau5.0.1Y.fa nom_de_l’échantillon.bam.bam > nom_de_l’échantillon_q10.mpileup
samtools mpileup -q 30 -B -f ARS-UCD1.2_Btau5.0.1Y.fa nom_de_l’échantillon.bam.bam > nom_de_l’échantillon_q30.mpileup

Détection de variants avec Varscan en utilisant un score qualité de base égale à 10:

varscan mpileup2cns nom_de_l’échantillon_q10.mpileup --output-vcf --variants --min-avg-qual 10 > nom_de_l’échantillon_Varscan_10.vcf
bgzip -c nom_de_l’échantillon_Varscan_10.vcf > nom_de_l’échantillon_Varscan_10.vcf.gz
tabix -p vcf nom_de_l’échantillon_Varscan_10.vcf.gz

c) Appel de SNPs avec Freebayes :

Détection de variants avec Freebayes, -C correspond au nombre de lectures minimum supportant le variant pour qu'il soit appelé
Exemple C=5
freebayes -f ARS-UCD1.2_Btau5.0.1Y.fa -C 5 nom_de_l'échantillon.bam > nom_de_l'échantillon_freebayes_5.vcf
bgzip -c nom_de_l'échantillon_freebayes_5.vcf > nom_de_l'échantillon_freebayes_5.vcf.gz
tabix -p vcf all_freebayes_5.vcf.gz

d) Recalibration de score qualité de base BQSR :

Les deux fichiers de score qualité de base ont été téléchargés sur le site du Consortium International des 1000 Génomes Bovins (http://www.1000bullgenomes.com/). La recalibration avec la deuxième version du fichier BQSR est effectuée selon la commande suivante :

gatk BaseRecalibrator -R ARS-UCD1.2_Btau5.0.1Y.fa -I nom_de_l'échantillon.bam --known-sites ARS1.2PlusY_BQSR.vcf.gz -O nom_de_l'échantillon.recal_data.table

Application sur mes données :

gatk ApplyBQSR -R ARS-UCD1.2_Btau5.0.1Y.fa -I . nom_de_l'échantillon.bam --bqsr-recal-file nom_de_l'échantillon.recal_data.table -O nom_de_l'échantillon.recal.bam
Analyser la covariation après recalibration :
gatkBaseRecalibrator -R ARS-UCD1.2_Btau5.0.1Y.fa -I nom_de_l'échantillon.recal.bam --known-sites ARS1.2PlusY_BQSR.vcf.gz -O nom_de_l'échantillon.recal.post_recal_data.table
Génération de pdf de la covariation :
gatkAnalyzeCovariates -before nom_de_l'échantillon.recal_data.table -after nom_de_l'échantillon.post_recal_data.table -plots recal_plots.pdf

Apès recalibration, les SNPS ont été appelés en utilisant le fichier avec les 3 algorithmes d'appels de SNPs, en utilisant les même commandes montrés ci-dessus, pour étudier l'effet de la recalibration sur le nombre de SNPs obtenus.

e) Intersection entre les fichiers vcf ontenus avec les 3 algorithmes d'appel de SNPs :

Tester l'intersection entre deux fichiers « vcf » obtenus avec deux algorithmes d'appel de SNPs :
Exemple pour varscan30 et GATK30 après recalibration :
vcf-isec -f -n +2 nom_de_l'échantillon_Varscan_30.recal.vcf.gz nom_de_l'échantillon_GATK_30.recal.vcf.gz > GATK30_varscan30_recal_inter.vcf

Tester l'intersection entre les 3 fichiers « vcf » obtenus avec les 3 algorithmes d'appel de SNPs après recalibration :

vcf-isec -f -n +3 nom_de_l'échantillon_GATK_30.recal.vcf.gz nom_de_l'échantillon_Varscan_30.recal.vcf.gz nom_de_l'échantillon_freebayes_5.recal.vcf.gz > GATK30_var30_free5_recal_inter.vcf

f) Comptage du nombre de SNPs obtenus avec les 3 algorithmes d'appel de SNPs :

Exemple de comptage de SNPs filtrés et appelés avec GATK après recalibration :
grep -v "^#" nom_de_l'échantillon_GATK_30.recal.SNP.filtered.vcf | wc -l > GATK30_filtred_recal

g) Distribution des filtres GATK :

g1) Filtration sur la DP :
Exemple de filtration sur la DP des SNPs appelés avec GATK et recalibrés :

- bcftools filter -i 'TYPE="snp" && MIN(FORMAT/DP)>=3' nom_de_l'échantillon_SNP_recal_GATK_15.vcf > nom_de_l'échantillon_SNP_recal_GATK_15_minDP3.vcf
- bcftools filter -i 'TYPE="snp" && MAX(FORMAT/DP)<= 14' nom_de_l'échantillon_SNP_recal_GATK_15_minDP3.vcf > nom_de_l'échantillon_SNP_recal_GATK_15_DP3_14.vcf

g2) Distribution des filtres GATK :

- bcftools query -f '[%DP\n]' nom_de_l'échantillon_SNP_recal_GATK_15_DP3_14.vcf | sort -g | uniq -c > DP_count_nom_de_l'échantillon_SNP_recal_GATK_15_DP3_14_DP_count
- bcftools query -f '[%QUAL\n]' nom_de_l'échantillon_SNP_recal_GATK_15_DP3_14.vcf | sort -g | uniq -c > QUAL_count_nom_de_l'échantillon_SNP_recal_GATK_15_DP3_14
- bcftools query -f '[%GQ\n]' nom_de_l'échantillon_SNP_recal_GATK_15_DP3_14.vcf | sort -g | uniq -c > GQ_count_nom_de_l'échantillon_SNP_recal_GATK_15_DP3_14
- bcftools query -f '[%ReadPosRankSum\n]' nom_de_l'échantillon_SNP_recal_GATK_15_DP3_14.vcf | sort -g | uniq -c > ReadPOsRankSum_count_ nom_de_l'échantillon_SNP_recal_GATK_15_DP3_14
- bcftools query -f '[%MQ\n]' nom_de_l'échantillon_SNP_recal_GATK_15_DP3_14.vcf | sort -g | uniq -c > MQ_count_ nom_de_l'échantillon_SNP_recal_GATK_15_DP3_14
- bcftools query -f '[%QD\n]' nom_de_l'échantillon_SNP_recal_GATK_15_DP3_14.vcf | sort -g | uniq -c > QD_count_AS5_ nom_de_l'échantillon_SNP_recal_GATK_15_DP3_14

g3) Récupérer les SNPs hétérozygotes :

La commande pour récupérer un fichier « vcf » contenant seulement les hétérozygotes et récupérer le champ de DP après filtration :

for file in *_filtered.vcf; do bcftools view -f PASS -i 'GT~"0/1"' > ${*_filtered% %.vcf}_heterozygotes.vcf
Distribution de la DP pour les SNPs hétérozygotes :
for file in *_heterozygotes.vcf; do bcftools query -f '[%DP\n]' | sort -g | uniq -c > ${*_heterozygotes %%.vcf}_DP.count

3) Les diagrammes de venn :

R Script pour diagram de venn : Exemple de l'intersection de nombre de SNPs appelés après recalibration avec les deux fichiers de score qualité de base :

```
GATK15_bwamem_recal_v2_v3_Inter<-euler(c("GATK15_bwamem_recal_v2" = 525789,
"GATK15_bwamem_recal_V3" = 48149,
"GATK15_bwamem_recal_v2&GATK15_bwamem_recal_V3" = 4928321))
plot(GATK15_bwamem_recal_v2_v3_Inter, counts= TRUE, font=1, cex=1, alpha=0.5, quantities =
TRUE, fill=c("blue", "grey", "red"))
```

R Script pour diagram de venn : Exemple de l’intersection de nombre de SNPs appelés avec les 3 algorithmes d’appel de SNPs après recalibration :

```
gatk30_var30_freeb5_recal_inter<-euler(c("GATK30_recal" = 317810 , "Varscan30_recal" =
236016 , "Freebayes5_recal" = 2617530 , "GATK30_recal&Varscan30_recal" = 52765 ,
"GATK30_recal&Freebayes5_recal" = 241930 , "Varscan30_recal&Freebayes5_recal" = 225959 ,
"GATK30_recal&Varscan30_recal&Freebayes5_recal" = 51481 ))
plot(gatk30_var30_freeb5_recal_inter, counts= TRUE, font=1, cex=1, alpha=0.5, quantities =
TRUE, fill=c("pink", "antiquewhite2", "blue"))
```

Annexe IV: Contribution à l’obtention des données génomiques publiées dans l’article sous presse dans « Science Advances » et intitulé : The genetic identity of the earliest human-made hybrid animals, the kungas of Syro-Mesopotamia

Notre intérêt pour la domestication des bovins nous a amené à nous intéresser au site de Göbekli Tepe dans le sud-est de l'Anatolie, en Turquie actuelle. L'occupation humaine de la région de Göbekli Tepe a commencé vers environ 9500-8700 cal BCE et s’est achevée vers 8000 cal BCE (pour description du site voir (Schmidt, 2007)). Ce site témoigne donc du tout début du Néolithique et la faune présente représenterait la faune sauvage avant que les processus de domestication aient commencé à les changer au niveau de leurs génomes.

Nous avons analysé cinq échantillons originaires de ce site dont un os pétreux. Ils ont été identifiés, selon les critères morphologiques, comme des bovins.Les extraits d’ADN ont été purifiés et les banques construites comme décrites par ailleurs. Les séquences Fastq obtenues ont été alignées sur le génome de référence des différentes espèces étudiées au sein du laboratoire en utilisant le logiciel d’alignement de séquences bwa mem (Li & Durbin, 2009). Les échantillons de Göbekli Tepe correspondant à des os longs ont montré une mauvaise préservation d’ADN endogène dont le pourcentage n’a pas dépassé les 0.07%.

Cependant, d’une manière surprenante, le nombre de lectures de l’os pétreux, l’échantillon GT64, alignées sur le génome de référence de cheval (*E.caballus)* equCab2.fa était plus élevé que celui obtenu quand les séquences étaient alignées sur le génome de référence bovin (ARS-UCD1.2_Btau5.0.1Y.fa). En effet, le pourcentage moyen d’ADN endogène obtenu à partir de différentes banques construites à partir des extraits de cet échantillon était de 14.86% alors que si les séquences étaient alignées sur le génome de référence bovin, le pourcentage moyen d’ADN endogène était de 1.09%.
Nous en avons donc conclu qu’il s’agit d’un équidé et non pas d’un bovin. Néanmoins, puisqu’un projet en cours dans notre laboratoire portait sur une étude paléogénomique des équidés, la banque génomique que j’avais construite a été séquencée en profondeur. Nous avons produit des données à l'échelle du génome et atteint une couverture du génome nucléaire de 1,6 X.

Les séquences obtenues ont été inclues dans l’analyse des ânes sauvages en cours au laboratoire. L’analyse en composante principale incluant 4.1 millions de SNPs partagés entre les équidés éteints, deux ânes sauvages syriens (*Equus hemippus* ou hémippe) du 19ème siècle et l’échantillon de Göbekli Tepe a placé ce dernier avec les deux hémippes modernes. L’analyse des données mitochondriales et leur comparaison avec celles produites antérieurement par le laboratoire sur les ânes sauvages asiatiques (Bennett et al., 2017) regroupe cet échantillon dans le clade des hémippes avec une position ancestrale. La conclusion était donc que cet échantillon de Göbekli Tepe était un hémippe.

Les séquences génomiques, quant à elles, ont également été analysées et inclues dans l'étude en cours au laboratoire. Cette étude avait comme but d'analyser des ossements d'équidés enterrés dans des tombes prestigieuses du 3ème millénaire av.n.e. au nord de la Syrie dans un site appelé Umm el-Marra. Ces équidés ont été préalablement soumis à une analyse ostéologique qui a montré que ces individus avaient des caractères morphologiques particuliers qui n'ont pas permis une classification taxonomique claire (Jill A. Weber, 2008). Des traces d'usure sur les dents suggèrent que ces animaux portaient des mors en permanence et ont donc été nourris. Ces données laissaient penser que ces équidés pourraient être des « Kungas », équidés de grande valeur utilisés, avant l'introduction des chevaux domestiques en Asie du sud-ouest à la fin du troisième millénaire BCE (Guimaraes et al., 2020). Ces équidés de prestige servait la noblesse en temps de paix et de guerre ce qui a été documenté d'un côté sur le célèbre « étendard d'Ur » (British Museum, Londres) ainsi que sur des tablettes cunéiformes (voir Fig. 1 de l'article Bennett et al., 2021).

Au laboratoire, nous avons analysé des ossements de quelques équidés d'Umm el-Marra en utilisant différentes approches, une approche ciblée sur des SNPs pour analyser le chromosome Y et l'ADN mitochondrial car l'ADN dans ces ossements est extrêmement mal préservé. L'ADN mitochondrial s'est avéré d'être celui d'un âne domestiqué (*Equus africanus*) et les SNPs du chromosome Y étaient partagés avec des hémippes modernes du 19ème s. de n.e. (Fig. 3, Bennett et al., 2021). Ceci indiquait alors qu'il s'agit d'un hybride dont la mère était une ânesse et le père un âne sauvage syrien. Finalement, nous avons analysé le génome nucléaire d'un os qui avait préservé des traces d'ADN.

L'analyse en composante principale de ces données a placé cet échantillon à une position intermédiaire entre les ânes domestiques et les hémippes (anciens et modernes) suggérant une origine hybride (Fig. 4A, Bennett et al., 2021). Pour mieux explorer cette observation, une analyse ADMIXTURE a été effectuée et a montré que le génome de l'échantillon kunga était bi-composite, la moitié étant apparenté au génome de l'âne domestique, et l'autre moitié à celui de l'hémippe syrien (Fig. 4B, Bennett et al., 2021). Ce résultat a été confirmé par une autre analyse, TreeMix, dont l'arbre a également placé cet échantillon à mi-chemin entre les ânes et les hémippes, ce qui est attendu lorsqu'il s'agit d'hybrides F1 de ces espèces (Fig. 4D, Bennett et al., 2021).

La conclusion que notre équipe a pu en tirer était donc que les kungas étaient des hybrides F1 entre des ânes domestiques femelles et des hémippes mâles, documentant ainsi les premières preuves de la production par les êtres humains d'animaux hybrides.

Les analyses qui permettront d'exploiter l'information contenue dans les génomes bovins que j'ai obtenus au cours de ma thèse sont similaires à celles réalisées pour ce travail.

GENETICS

The genetic identity of the earliest human-made hybrid animals, the kungas of Syro-Mesopotamia

E. Andrew Bennett[1]*†‡, Jill Weber[2], Wejden Bendhafer[1], Sophie Champlot[1], Joris Peters[3,4], Glenn M. Schwartz[5], Thierry Grange[1]*†, Eva-Maria Geigl[1]*†

Before the introduction of domestic horses in Mesopotamia in the late third millennium BCE, contemporary cuneiform tablets and seals document intentional breeding of highly valued equids called kungas for use in diplomacy, ceremony, and warfare. Their precise zoological classification, however, has never been conclusively determined. Morphometric analysis of equids uncovered in rich Early Bronze Age burials at Umm el-Marra, Syria, placed them beyond the ranges reported for other known equid species. We sequenced the genomes of one of these ~4500-year-old equids, together with an ~11,000-year-old Syrian wild ass (hemippe) from Göbekli Tepe and two of the last surviving hemippes. We conclude that kungas were F1 hybrids between female domestic donkeys and male hemippes, thus documenting the earliest evidence of hybrid animal breeding.

INTRODUCTION

In the third millennium BCE, urbanized, socially stratified, and literate societies appeared for the first time in Syria and northern Mesopotamia (*1*, *2*). Part of this "second act" of the urban revolution was the breeding and employment of an equid of high status and prestige designated a "kunga." The precise taxonomical determination of the kunga and its identification in the archaeological record have been uncertain until now. Third millennium BCE cuneiform clay tablets from Syro-Mesopotamia describe several equids, using the generic term ANŠE associated with various logographs. Of these, the so-called kunga was represented by the cuneiform signs ANŠE.BARxAN (Fig. 1A) (*3*, *4*). Texts from the Diyala region in Mesopotamia and the kingdom of Ebla in the Levant state that the prices for these equids were considerable, costing up to six times the price of a donkey (*5*). References for these valuable equids are found in multiple clay tablets (*3*, *4*) (Fig. 1A) such as those detailing fodder expenses, e.g., barley for the equids of the god Shara and the deified king Shulgi from Umma (*6*), and dowries for royal marriages (*7*). Large-sized male kungas were used to pull the vehicles of "nobility and gods" (*6*), and their size and speed made them more desirable than asses for the towing of four-wheeled war wagons (*8*), which predate horse-pulled chariots. Smaller-sized male and female kungas were used in agriculture, where they were frequently reported pulling ploughs (*4*, *9*). Kunga foals were seldom born within the urban centers of Sumer and Syria, and Ebla purchased young kungas almost exclusively from what may have been the principal breeding center at Nagar (modern Tell Brak), in northern Mesopotamia, whose rulers also provided them as gifts to the elites of allied territories (*3*). Presumed kungas featured prominently on royal seals throughout the region (*10*), and images of these hybrids likely appear on both the "war" and "peace" panels of the standard of Ur, a Sumerian artifact excavated from the royal cemetery in the ancient city of Ur (in modern-day Iraq). In one of the first depictions (2600 BCE) of a military expedition in human history, warriors stand on four-wheeled war wagons, each drawn by a team of unspecified equids (Fig. 1B). An example of the rein ring featured in this image has been found in a royal grave at Ur (Fig. 1C), decorated with a small statue of a noncaballine equid, either a kunga or hemione. Kunga use and traditions decreased and eventually vanished following the introduction of domestic horses in the region (*9*, *11*). Early references to horses in cuneiform writing coincide with the Third Dynasty of Ur (late third millennium BCE), where they are referred to as anše-zi-zi and later anše-kur-ra (equids of the mountain) (*6*, *8*). An introduction of domestic horses in Mesopotamia by the end of the third millennium is also supported by paleogenetic data illustrating their late arrival in Anatolia around 2000 BCE, presumably through the Caucasus (*12*).

While the symbol for kunga (ANŠE.BARxAN) is used to describe a hybrid equid, the unambiguous assignment of this term to a species is difficult and controversially discussed. Some authors even argue that the kunga referred only to wild caught Persian onagers (also known as Iranian onagers; *Equus hemionus hemionus*, a subspecies of the Asiatic wild ass) rather than hybrid animals (*3*, *6*), although the difficulty in taming modern onagers, which are reportedly less tractable than zebras (*13*), does not support this interpretation. One of the likely parents of the kunga is the donkey (*Equus africanus asinus*), thought to be present in Sumer from at least the late fourth millennium BCE (*8*). The identity of the other parent, however, remains unclear. Another equid attested since the Early Dynastic period I (ca. 2800 BCE) is the anše-edin-na, literally translated as "equid of the desert." This animal was hunted for its meat and hide, but never used as a draught animal. The anše-edin-na is broadly considered to be a type of onager (*4*, *8*), although it is impossible to say whether it refers to the Persian onager (*Equus hemionus onager*) or to the Syrian wild ass, or hemippe (*Equus hemionus hemippus*) (*8*), sometimes also named "Syrian onager." Described as a light, swift animal (*14*), the hemippe was the smallest of all modern equids until the subspecies went extinct early in the 20th century (*11*). It

[1]Institut Jacques Monod, Université de Paris, CNRS, 75013 Paris, France. [2]Near East Section, The University Museum of Archaeology and Anthropology, Philadelphia, PA 19103, USA. [3]ArchaeoBioCenter, Institute of Palaeoanatomy, Domestication Research and the History of Veterinary Medicine, LMU Munich, 80539 Munich, Germany. [4]SNSB, Bavarian State Collection of Palaeoanatomy, 80333 Munich, Germany. [5]Department of Near Eastern Studies, Johns Hopkins University, Baltimore, MD 21218, USA.
*Corresponding author. Email: eva-maria.geigl@ijm.fr (E.-M.G.); thierry.grange@ijm.fr (T.G.); eabennett@gmail.com (E.A.B.)
†These authors contributed equally to this work.
‡Present address: Key Laboratory of Vertebrate Evolution and Human Origins, Institute of Vertebrate Paleontology and Paleoanthropology, Center for Excellence in Life and Paleoenvironment, Chinese Academy of Sciences, Beijing 100044, China.

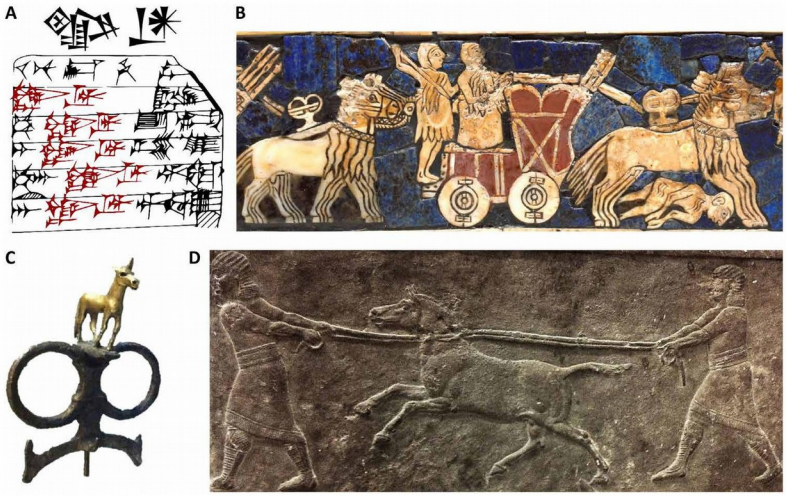


Fig. 1. Iconographic and textual depiction of the kunga. (**A**) Third millennium BCE cuneiform signs for the kunga (ANŠE.BARxAN) above a photo and drawing of a clay tablet from UrIII Girsu/Lagaš (British Museum BM23836) featuring multiple occurrences, highlighted in the juxtaposed drawing. The first two lines read "transmitted barley plots of 1 bur 6 iku (=8.64 ha) in area, (for the keeping of) ANŠE.BARxAN — equids of the king" (drawing and translation courtesy of K. Maekawa). (**B**) Detail from the Standard of Ur shows an equid team pulling a four-wheeled wagon in battle (photo credit: The British Museum Images). (**C**) Image of a rein ring with decorative equid from a royal grave at Ur, contemporary and similar to those visible in the Standard of Ur. (**D**) Nineveh panel: "hunting wild asses" (645 to 635 BCE) (British Museum, London). Figure S8 shows additional panels attesting that the equids depicted are noncaballine. (C and D) British Museum, London; photo credit: E. Andrew Bennett.

has been argued that in addition to its untamable, aggressive nature (*14*), its diminutive size made it an unlikely candidate for use in breeding kungas [(*3*) and references therein]. Some authors considered the nondonkey parent to be a horse [discussed in (*4*, *9*)].

In the elite burial complex of Tell Umm el-Marra (2600 to 2200 BCE), possibly belonging to the ancient city of Tuba, 55 km east of Aleppo in modern-day northern Syria (Fig. 2), men and women were interred with ceramics, bronze, and silver vessels; bronze weapons and tools; and personal ornaments made of bronze, silver, gold, and lapis lazuli (*15*). Within this royal burial complex, complete skeletons of 25 male equids and bones from six additional animals were buried separately from humans, either in a sequence of pits or in their own mud-brick structures (*15*), akin to the 3000 BCE donkey burials at Abydos, Egypt (*16*). While some animals were interred after natural deaths, more than half appear to have been deliberately killed for burial in the complex. Morphometric values obtained from these bones indicate that these animals constitute a population outside of the typical ranges of horses, asses, and onagers, and it has been proposed that these skeletons represent hybrids, presumably kungas [(*17*) and Supplementary Materials]. In absolute size, the skeletons are closer to hemiones, but are more robust; commonly used slenderness indices suggest greater affinities with asses than with hemiones. The leg characteristics of hemiones, responsible for a speed exceeding that of horses, is retained in these animals, suggesting that they were also fast (*18*). Discrepancies in wear between the incisors and cheek teeth of some of the equids indicate that the animals were foddered and not commonly grazed (*17*), features that would have been expected on the skeletons of the equids depicted on the standard of Ur, whose lip or nose rings would have made grazing difficult (Fig. 1B). These animals would have been stronger and faster than donkeys and must have been more tamable than hemiones (*19*).

Taxonomic classification of equids uncovered in tombs across Mesopotamia (Ur, Kish, and Lagash—now al-Hiba, Abu Salabikh, and Tell Madhhur) is often controversial [for discussion, see, e.g., (*5*, *8*, *20*, *21*)]. The degree of variation within ancient populations is not fully known, and the degree of variation between individuals within a population—especially of domesticated animals—is large, making it difficult to differentiate between *E. africanus* and *E. hemionus* using solely bone morphological and metrical characteristics [for discussion, see (*22*)].

To clarify whether the burials of Tell Umm el-Marra contained the remains of the politically and symbolically important hybrids referred to in numerous cuneiform tablets as kunga and to determine the taxonomic status of those animals, we investigated the genomes in samples from the skeletons of the equid installations at Umm el-Marra, an equid sample from the Early Neolithic site of Göbekli Tepe (Turkey), and the last survivors of the Syrian wild ass conserved in the Natural History Museum of Vienna.

RESULTS AND DISCUSSION

Analysis of the maternal and paternal lineages of the Umm el-Marra equids

An initial polymerase chain reaction (PCR) screening of equid samples from Umm el-Marra showed that DNA was extremely poorly preserved in these bones owing to the hot climate in Syria, detrimental to long-term DNA preservation, and the poor condition of the bones (phalanges and sternum) available for study (fig. S1). Therefore, we combined shotgun nuclear DNA sequencing with

Fig. 2. Map of third millennium BCE Syro-Mesopotamia showing the major historical and archaeological sites (modified from Wikipedia https://fr.m.wikipedia.org/wiki/Fichier:Syrie_3mil_aC.svg). The insert shows a representative equid burial in Umm el-Marra. Photo credit: G. Schwartz.

highly sensitive PCR, targeting taxonomically informative regions of both uniparental markers: mitochondrial DNA and the Y chromosome. To better pinpoint the genetic identity of the parental species, we increased the available Y-chromosome data by sequencing regions from additional populations of both modern and 19th and 20th century museum samples of hemiones and donkeys, for which the mitochondrial sequences were previously generated (*23*). Short, overlapping PCR products suited to the degraded DNA of the samples were designed to amplify a highly diagnostic mitochondrial control region fragment [324 base pairs (bp) long], including the site of a well-characterized 28-bp deletion exclusive to hemiones (*23*), and three separate regions of the Y-chromosomal DNA (in total 168 bp long) encompassing four single-nucleotide polymorphisms (SNPs), which we show to be diagnostic between *Equus ferus* (*caballus* and *przewalskii*), *E. africanus*, and *E. hemionus*.

The full targeted mitochondrial sequence was successfully amplified from two of the six individual equids tested from Umm el-Marra. At every position divergent between *E. ferus*, *E. africanus*, and *E. hemionus*, both of these sequences contained the *E. africanus*–specific bases and lacked the 28-bp deletion specific to *E. hemionus* (*23*). The maternal lineage of these equids thus unambiguously belongs to *E. africanus* as visible in a median-joining network (Fig. 3A) (*24*). All three Y-chromosome fragments were successfully amplified from these same two Umm el-Marra individuals. Within these three regions, four diagnostic positions differentiate *E. africanus* (T/G/T/A) from *E. hemionus* (C/A/G/G), two of which also differentiate *E. ferus* (C/G/G/A) from either *E. hemionus* or *E. africanus*. At each diagnostic position, the equids from Umm el-Marra were found to have the *E. hemionus*–specific base, and no diagnostic position of any product contained the *E. africanus*–specific base (table S1). The hemione-specific Y-SNPs were also confirmed previously in diverse hemiones from archaeological samples from the Caucasus, museum specimens from Tibet and Syria, and present-day specimens from the Gobi in Mongolia (Fig. 3B and table S1) (*23*).

In addition to the *E. hemionus* diagnostic positions, both Umm el-Marra sequences contained two additional Y-chromosome SNPs observed only in the two hemippes from the 19th and 20th centuries analyzed here (Fig. 3B), one of them being the last known member of the subspecies. This animal had been caught in the deserts north of Aleppo in 1911 and had been kept in the Schönbrunn Zoo in Vienna until its death in 1929 (see fig. S2 for images of two of the hemippes used in this study). Thus, the Umm el-Marra equids harbor the maternal lineage of the domestic donkey and the paternal lineage of the Syrian wild ass, suggesting that they could be F1 hybrids, since interspecific equid hybrids are generally sterile or poorly fertile.

Analysis of the nuclear genomes of Umm el-Marra and Göbekli Tepe equids and the last Syrian wild asses

To further establish the hybrid identity of these equids, we sequenced a subset of the nuclear genome of the best preserved Umm el-Marra equid bone (table S5). In addition, we established the genome sequence of the extinct Syrian hemippe by sequencing a ca. 11,000-year-old wild ass from the early Neolithic site of Göbekli Tepe, present-day Turkey, representing the first temple (*25*), and two 19th century specimens from the Schönbrunn Zoo (table S5). These four newly generated genomes were compared to six modern horse (*26*), six domestic donkey (*27*), three Mongolian khulans (an *E. hemionus* subspecies from the Gobi) (*27*, *28*), two kiang genomes (*E. h. kiang* or *E. kiang*) (*27*, *29*), and one Persian onager genome (an *E. hemionus* subspecies from Iran) (*29*) (table S4). A set of 15.5 million SNPs residing outside of repeated sequences and being variable in the modern genome equid panel was used for calling the ancient genomes (see Supplementary Materials and Methods and table S5). Although the best Umm el-Marra extract contained only 0.18% endogenous DNA, we could obtain 45.6K SNPs, 40.4K of which were shared with either hemippe. Of these, 15.2K SNPs (37%) were shared with both of the two best-covered hemippe

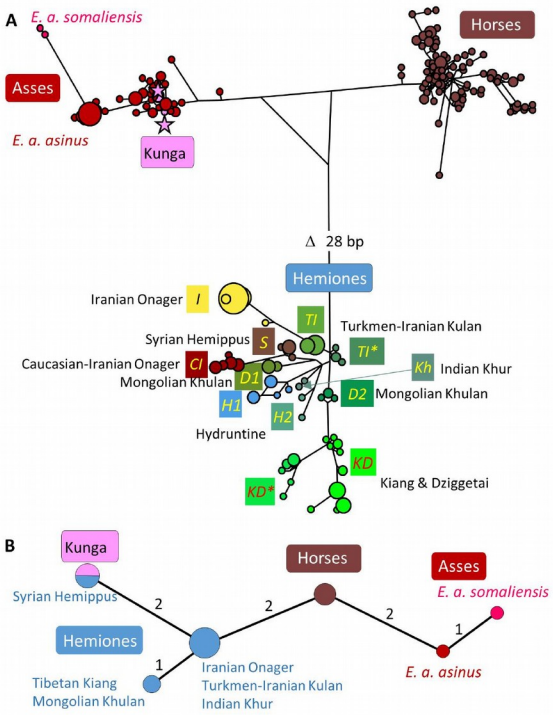


Fig. 3. Median-joining network of equid sequences. (**A**) Hypervariable region (324 bp) of mitochondrial DNA from 278 individuals belonging to Asiatic wild asses [*E. hemionus* subspecies (*23*)], to horses (*E. f. caballus* and *E. f. przewalskii*), and to African asses (*E. a. asinus* and *E. a. somaliensis*). The position of the sequences obtained from the Umm el-Marra samples is indicated with pink stars. The *E. hemionus* mitogenome clades (I, TI, TI*, CI, H1, H2, D1, D2, Kh, KD, and KD*) are as defined previously (*23*). (**B**) Three different fragments (168 bp) of the Y chromosome of equids (asses, horses, and hemiones). The position of the sequences obtained from the Umm el-Marra samples is indicated in pink.

genomes. First, we performed principal components analysis (PCA) to compare (i) the hemione and donkey genomes (Fig. 4A and fig. S4A) or (ii) hemione, donkey, and horse genomes (fig. S4, B to E). Identical results were obtained whether we used the 15.2K SNPs shared between the Umm el-Marra equid (UMM9), the Göbekli Tepe, and the 1864 hemippe, or whether we used the 4.1 million SNPs shared between the Göbekli Tepe and the 1864 hemippe genome and projected the UMM9 equid onto the PCA [compare Fig. 4A and fig. S4 (A to E)], showing that the 15.2K SNPs obtained allowed robust characterization of the status of the Umm el-Marra equid. When only donkey and hemione genomes are used, PC1 separates donkeys from hemiones and PC2 separates the hemiones (Fig. 4A and fig. S4A). The most differentiated in PC2 are the Persian onager and the kiang zoo specimens. The Mongolian khulan and the kiang from neighboring regions in China are very closely related and overlapping in the PCA, which is in accordance with the shared mitochondrial lineages we reported previously that led us to question the specific taxonomic status of the kiang as a separate species (*23*) (see also the phylogenetic trees of the mitogenomes and the genomes in figs. S6 and S7). The two modern hemippes and the Göbekli Tepe sample overlap as well (the lower coverage 1892 hemippe was projected) and are located at an intermediate position in PC2. The Göbekli Tepe sample, a mare, is an ancient hemippe as observed from the phylogenetic trees constructed from both mitochondrial and nuclear genomes (figs. S6 and S7). The UMM9 equid falls exactly halfway between the donkeys and the hemippes. When the PCA also includes the horses, PC1 differentiates the horse from the noncaballine equids, PC2 separates the donkeys from the hemiones, and PC3 separates the hemiones in a similar way as PC2 does when horses are not included (fig. S4, B to E). In all analyses, the results for the UMM9 equid illustrate an intermediate position between the donkeys and hemippes. The PCA analyses thus indicate that the UMM9 genome is a 50% mixture of donkey and hemippe.

We further explored this outcome through ADMIXTURE analysis (Fig. 4B) (*30*). A four-population model separates horses, donkeys, Mongolian khulans, and kiangs from onagers and the two best-covered hemippes. The UMM9 equid is modeled as an admixture of equal proportion between donkey and hemippe/onager (Fig. 4B). Likewise, when considering the 4738 UMM9 equid SNPs for which all six donkeys differ from all hemippes that have the position covered, the UMM9 equid harbors SNPs corresponding to roughly half of those specific to each putative parent (Fig. 4C).

Last, a bifurcating tree with gene flow analysis was performed using treemix (Fig. 4D) (*31*). For the reference equids, genomic tree topology is similar in both the full mitogenome and genome tree topology obtained with different methods (figs. S6 and S7). The Persian onager and the hemippes are closely related, and the Mongolian khulans and the Tibetan kiangs are even more closely related. In this respect, the genetic distances between the various hemiones correspond to the geographic distances between their native range (onager: Iran; hemippe: Syria; Mongolian khulan: Mongolia; kiang: Tibet). The UMM9 equid is represented on the tree as related to the donkey, but the residuals between the hemippe and the UMM9 equid are high, and a gene flow event from the hemippe to the UMM9 equid best describes the phylogeny (see also fig. S5). These results demonstrate the sufficiency of the Umm el-Marra SNPs to determine the phylogenetic relationships between the equids. The tree also placed the Umm el-Marra sample halfway between the asses and the hemippes (Fig. 4), which meets expectations when dealing with F1 hybrids of these species. Evidence from Y-chromosome analysis indicates that the Syrian hemippe rather than the Persian onager was used to father the Umm el-Marra equids, whereas the mitochondrial DNA reveals that a donkey contributed the maternal genome. The fact that the Umm el-Marra equids were F1 hybrids and not back-crossed hybrids is also supported by the relative hybrid sterility between donkeys and horses, as well as experiments in the 1940s crossing female donkeys with male hemiones, the Turkmenian kulans (*Equus hemionus kulan*), which produced sterile offspring (*32*).

Expectedly, the 19th to 20th century hemippes, representing some of the last survivors of the subspecies, are genetically similar, whereas the ~11,000-year-old Göbekli Tepe hemippe is more divergent (fig. S7). We also noted that the divergence between the three sequenced hemippes is much larger than that observed between the six domestic donkeys (fig. S7). The higher diversity between the sequenced hemippe genomes versus between the donkey genomes suggests that the donkey mother of the UMM9 equid is more closely related to present-day donkeys than the Syrian hemippe father of

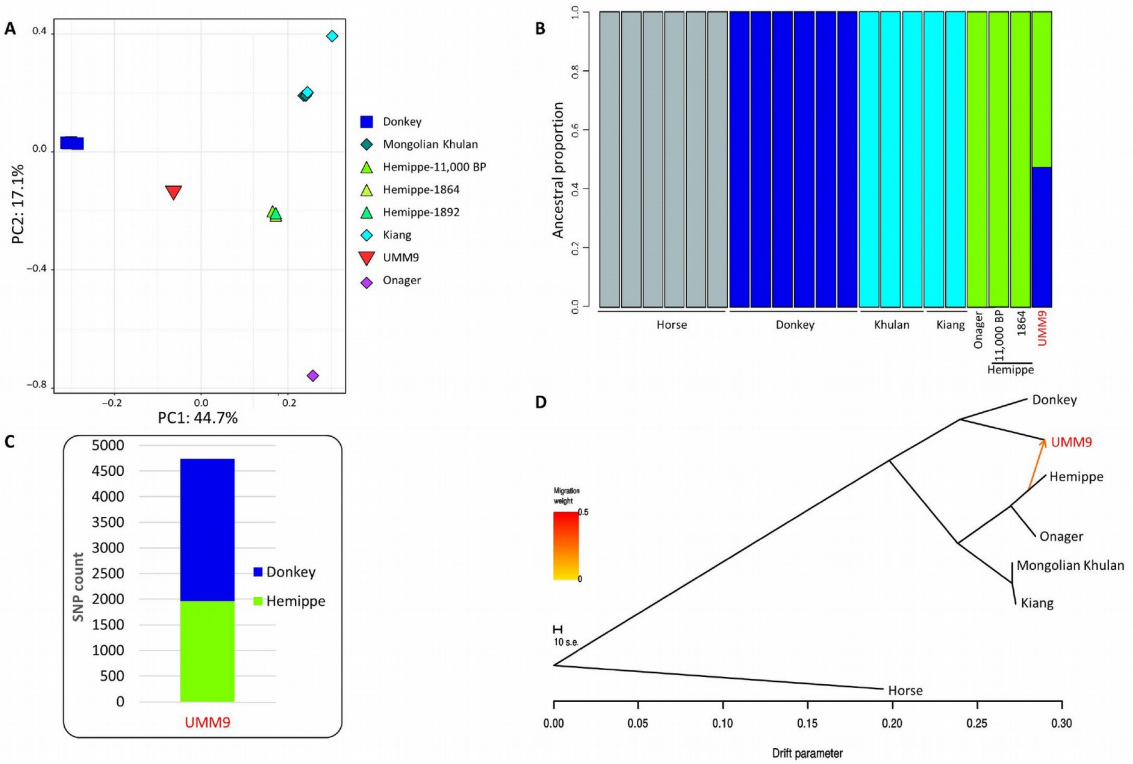


Fig. 4. Genome-scale analyses reveal that the UMM9 equid shares equal ancestry from donkey and hemippe. (**A**) PCA plots of noncaballine equids. The 15.2K SNPs shared between UMM9, the ca. 11,000-year-old Göbekli Tepe sample, and the 1864 hemippe were used, and only the 1892 hemippe was projected. (**B**) Admixture (*30*) analysis modeling four populations based on 15.2K SNPs shared between UMM9 and the higher-coverage Göbekli Tepe and 1864 hemippe (used in the analysis). (**C**) Counts of the UMM9 equid SNPs are identical to either hemippe or donkey using the 4738 SNPs, where all donkeys are identical and differ from the hemippes (of 40.4K total SNPs shared between the UMM9 equid and either of the three hemippes). (**D**) Bifurcating tree of equids with gene flow performed using treemix (*31*). The 40.4K SNPs shared between UMM9 and either of the three hemippes were used. Each equid group is represented by the following numbers of individuals: horse (six), donkey (six), hemippe (three), UMM9 equid (one), onager (one), Mongolian khulan (three), and kiang (two). Horses were used as the outgroup, and sample size correction was disabled. The tree obtained with one gene flow event is represented. The residuals with no or one gene flow event are plotted in fig. S5.

the UMM9 equid is to the other sequenced hemippes. This difference, albeit small, may account for the slightly higher affinity of the UMM9 equid to present-day domestic donkeys that is visible on fig. S5 (B and D). Both the genomic phylogenetic tree (fig. S7) and the PCA analyses (Fig. 4B and fig. S4) indicate that differentiation between the donkey genomes is low, far less than between the various present-day hemiones, presumably because donkeys went through a major bottleneck, possibly upon domestication and translocation to southwest Asia outside the range of the ancestors of donkeys. The observation that the ~4500-year-old UMM9 equid appears more closely related to present-day donkeys than to the last hemippe that disappeared a century ago suggests that the bottleneck of the donkey population had already taken place by the third millennium BCE.

It has been noted that the Syrian wild ass (hemippe), whose range once extended across the Levant, was the smallest form of modern equids (*18*). Both historical specimens analyzed in this study stood ca. 100 cm at the shoulder (*14*) (fig. S2). In contrast, the hybrids of Umm el-Marra were estimated to average 130 cm at the shoulder (*17*). Regarding this difference in size, previous work had recovered mitochondrial haplotypes from larger-sized Bronze Age equids recovered from Tell Munbaqa, situated in northern Syria east of Umm el-Marra, as well as from three historical hemippe samples dating from the mid-19th to early 20th century. Both the larger ancient and smaller more recent animals were shown to cluster together in a single separate mitochondrial clade (*23*). The Göbekli Tepe wild asses were, on average, even slightly larger than those of roughly contemporaneous Tell Mureybet (10th to 9th millennia BCE) and third-second millennia BCE Tell Munbaqa, two sites located in the direct vicinity of Umm el-Marra (Fig. 2) (*22*). It was concluded, therefore, that the small Syrian wild ass was likely to have been a dwarfed descendant of a genetically continuous population of larger, more robust animals populating Syria in the third millennium BCE and earlier (*23*). The genomic analyses of both ancient and historical hemippes in the present study support this earlier finding. No dwarf form has ever been reported from Late

Pleistocene and Holocene sites in Mesopotamia (*33*) or Anatolia (*34*). Nearly 2000 years after the equid burials of Umm el-Marra, sixth century BCE palace reliefs featuring hunted hemiones from Nineveh (in modern-day northern Iraq) show already relatively small animals (Fig. 1D and fig. S8).

To conclude, the genomic results from the rare equid burials at the elite mortuary complex of Umm el-Marra confirm earlier hypotheses based on morphological data that these animals are hybrids (*8*, *9*, *17*) and, given their interment in high-status tombs, are most likely identical with the valuable kungas frequently mentioned in cuneiform texts and depicted in images and royal seals throughout Mesopotamia. This study now offers a firm zoological classification of the historical kunga as an F1 cross between a female donkey and a male Syrian wild ass, or hemippe, putting to rest past speculations regarding the taxonomic identification of the BARxAN. We further show that the third millennium BCE ancestors of the hemippe were likely larger than those first described by European travelers visiting Syria in the 19th century. Our study also presents the earliest known case in human history of interspecies hybridization, which was practiced by Early Bronze Age breeders at sites such as Nagar (Tell Brak) (Fig. 2), to generate animals famous for their power, both physical and symbolic, in ancient warfare and diplomacy. This result also deepens our insight into the economic and political relationships between contemporary royal households of Greater Mesopotamia, and the dynamics by which these social elites fostered distant alliances. It also increases our understanding of the ways in which the earliest stratified urban societies of the Middle East developed and maintained their positions of authority. In this respect, genomic characterization of additional equids from comparable contexts, particularly from Nagar, may help clarify the scale of hybrid breeding in third millennium BCE Mesopotamian societies before the introduction of domestic horses.

MATERIALS AND METHODS

Sample description

Bone remains from equid skeletons dated between ca. 2550 and 2300 BCE and excavated in 2006 at Tell Umm el-Marra, a Bronze Age elite cemetery in northern Syria (*15*, *17*, *36*), were sampled for ancient DNA analysis. Further descriptions of the samples are given in the Supplementary Materials, and photos of the samples from the two individuals from which sufficient DNA was recovered appear in fig. S1. A petrous bone excavated from a layer dated between 9500 and 8300 BCE from the site of Göbekli Tepe in southeast Turkey [(*25*) and Supplementary Materials] was also analyzed in the present study. Furthermore, we analyzed two samples of the extinct *E. h. hemippus* originating from the desert of Aleppo in Syria and kept in the zoo of Schönbrunn, Vienna, Austria, a tooth from the NMW6048/ST345 specimen and a hair and skin sample from the NMW1308/B4690 specimen, corresponding to animals who died in the Schönbrunn zoo in 1864 and 1892, respectively (fig. S2) (*14*). Last, hair of a male Somalian ass (*E. africanus somaliensis*) from the "Réserve Africaine de Sigean" (Sigean, France) was provided for the analysis of the Y chromosome by E. Trunet (sample "As.Somalie").

Ancient DNA extraction, amplification, and sequencing

Hair samples were added to 1.5 ml of hair digestion buffer [100 mM tris-HCl (pH 8.0), 100 mM $NaCl_2$, 40 mM dithiothreitol, 3 mM $CaCl_2$, 2% *N*-lauryl sarcosyl, and proteinase K (250 μg/ml)] and incubated 4 to 24 hours at 50°C, shaken at 300 rpm. Solutions were then pelleted, and the supernatant was purified using a QIAquick Gel Extraction kit (Qiagen, Hilden, Germany) according to instructions.

DNA from archaeological bone samples was extracted, purified, and prepared for either quantitative PCR (qPCR) or sequencing in the ancient DNA laboratory described previously (*35*, *37*, *38*). Bone cleaning and treatment protocols were as described previously (*12*, *23*). Briefly, after removal of the surface with a razor blade or surface cleaning with bleach, the bones were either sawed using a flame-sterilized diamond disc of Dremel Fortiflex (Dremel Europe, The Netherlands) and grounded to fine powder in 6775 Freezer/MillSpex SamplePrep in liquid nitrogen or drilled at low speed with a flame-sterilized bit. The dense pyramidal part of the petrous bone GT64 was isolated using a flame-sterilized diamond disc of a Dremel and then grounded to fine powder in 6775 Freezer/MillSpex SamplePrep in liquid nitrogen. Half of the GT64 powder was treated with diluted hypochlorite (1:20), and both halves were washed with phosphate buffer according to Korlević *et al.* (*39*). DNA extraction was performed by incubating the bone powder at 37°C for 48 to 90 hours either in 1- to 10-ml extraction buffer A [0.5 M EDTA, 0.25 M PO_4^{3-} (pH 8.0), and 0.14 M ß-mercaptoethanol] or in twice 1-ml extraction buffer B [0.5 M EDTA, 0.05% Tween 20, proteinase K (250 μg/ml), and 0.14 M ß-mercaptoethanol] that was pooled before purification. Samples were purified using silica membrane spin columns (QIAquick Gel Extraction kit) with a vacuum manifold (Qiagen) and 25 ml of extenders (Qiagen) as described (*37*, *40*), as well as with either the 5 M guanidine HCl, 40% isopropanol (5M40) buffer as described by Dabney *et al.* (*41*) or the 2 M guanidine HCl, 70% isopropanol (2M70) buffer as described by Glocke and Meyer (*42*). The elution was performed twice in 25 μl of 10 mM tris-HCl (pH 8.0) and 0.05% Tween 20 (referred to as EBT) made from gamma-irradiated water (8 kGy).

Purified DNA was amplified by qPCR, the extract making up 5 to 20% total volume (10 to 20 μl per reaction). Inhibition characteristics were determined for failed samples indicating possible inhibition, and once optimal dilutions were determined, qPCR was attempted again. To protect against cross-contamination, the UQPCR [uracil *N*-glycosylase (UNG)–coupled quantitative PCR] method was used (*35*, *38*, *43*), in which uridine was substituted for thymidine in all PCRs, and incubation with UNG (extracted from *Gadus morhua*; Biotec Marine Biochemicals, Norway) was performed before each reaction. Mock extracts were included with each extraction and amplified to control for contamination. qPCRs varied slightly depending on the sample, but a typical reaction included 1.77 μl of LC FastStart DNA MasterPLUS mix1b; 0.23 μl of either FastStart DNA MasterPLUS mix1b, mix1a, or FastStart Taq (Roche Applied Science, Mannheim, Germany); a final concentration of 1 μM of each primer; and 1 U per reaction of UNG in 10-μl total volume. Primers were obtained from Sigma-Aldrich (St. Louis, USA). Mitochondrial primers were designed to amplify 357 bp of the hypervariable region (HVR) of *E. africanus* and *E. hemionus* mitochondria using short, overlapping fragments (table S3). Y-chromosome primers were designed to amplify three short sections of Y-chromosome DNA containing the target SNPs (tables S1 to S3). Several modifications of these primers were designed to increase sensitivity of qPCRs by minimizing the likelihood of primer dimers and artifacts and increasing primer efficiency. A list of primers used and product sizes is given in table S3. qPCR was performed using LightCycler 1.5 or LightCycler 2 (Roche Applied Science, Mannheim, Germany). qPCR programs varied depending on primer requirements and product length, but a typical program involved UNG incubation

at 37°C for 15 min, followed by polymerase activation at 95°C for 5 min, then two-step cycles of denaturation at 95°C for 10 s, then primer annealing and extension at 62°C for 40 s, and finally a temperature increase of 0.1°C/1 s from 62° to 95°C with continuous fluorescence measurement to generate melt curves of the products. Products were purified with a QIAquick PCR Purification kit (Qiagen, Hilden, Germany), and both strands were sequenced by capillary electrophoresis at Eurofins/MWG Operon (Ebersberg, Germany) using the ABI 3730xl DNA Analyzer (Life Technologies). Samples that yielded sequence results for the Y-chromosome are shown in table S1. An average of one nontemplate control (NTC) was run for every 6.6 samples (including mocks). No DNA was amplified in either NTCs or mocks, demonstrating that no detectable equid DNA was introduced during sample preparation or was present in reagents.

Samples from the two individuals from Umm el-Marra with the best preserved DNA identified via qPCR (UMM4 and UMM9) were selected for shotgun sequencing, and 24 double-stranded libraries using dual barcodes were prepared from DNA purified from three or four different areas of each bone using the protocol described by Massilani *et al.* (*44*). Seven of these libraries were treated with UNG to reduce the presence of cytosine deamination damage in the resulting sequences.

Double-stranded libraries of the two *E. h. hemippus* (hemippe) museum specimens were constructed using the NxSeq ampFREE Low DNA Library Kit (Lucigen, Middleton, WI, USA) following the protocol and the modifications described by Bennett *et al.* (*45*). Barcodes were added during an amplification reaction using dual-barcoded single-stranded library adapters (*46*) as primers, rather than those in the kit, where 20 μl of eluted library was added to 25 μl of OneTaq 2× Master Mix (Roche) and 0.6 μM of each adapter for 50-μl total volume, and amplified with the following protocol: 5 min at 95°C, 30 cycles of 30 s at 95°C, 30 s at 60°C, and 45 s at 68°C, followed by a 5-min cycle at 68°C. A library for a hair and skin sample belonging to specimen NMW5493/B 3625 was constructed using the NEBNext Ultra II DNA Library Prep Kit for Illumina (NEB, Ipswich, MA, USA).

Libraries from the Göbekli Tepe petrous bone GT64 extracts were constructed using the NEBNext Ultra II DNA Library Prep Kit for Illumina (NEB, Ipswich, MA, USA) after a pretreatment with USER enzyme mix (NEB, Ipswich, MA, USA).

Dual-barcoded libraries were then purified and size-selected using NucleoMag beads (Macherey-Nagel) for two rounds of purification following the supplied protocol at a ratio of 1.3× beads per reaction volume and eluted in 30 μl of EBT.

All libraries were quantified with the Qubit 2.0 Fluorometer (Thermo Fisher Scientific), with Bioanalyzer 2100 (Agilent, Santa Clara, CA, USA), and by qPCR. Screening by shotgun sequencing of Umm el-Marra samples and of the two hemippe samples was performed on an Illumina MiSeq system using a v3 reagent kit for 2 × 75 cycles. The libraries constructed from sample SP345, which came from a molar belonging to specimen NMW6048/ST345 (1864 hemippe) and a hair and skin sample belonging to specimen NMW1308/B4690 (1892 hemippe), and the two libraries from the UMM9 sample purified using the 5M40 and 2M70 buffers were selected for deep genomic sequencing.

Sequencing of the 1864 hemippe genome was performed on Illumina NextSeq using NextSeq 500/550 High Output Kit v2 (2 × 75 cycles). The custom sequencing primer CL72 (*46*) was substituted for the read 1 primer sequencing steps, which is compatible with the single-stranded adapters used for these samples. Sequencing of the 1892 hemippe and of two of the UMM9 libraries was performed first on Illumina MiSeq using a v3 reagent kit for 2 × 75 cycles and then on Illumina NovaSeq 6000 using an S2 flow cell for 2 × 50 cycles. Sequencing of the GT64 libraries was performed on NovaSeq 6000 using an S4 flow cell for 2 × 75 cycles.

Paleogenetic data analyses

Sequences from PCRs were manually curated, assembled, and aligned using the Geneious software suite (*47*). Median-joining network analysis (*24*) was performed on mitochondrial sequences covering 357 bp of the hypervariable region generated in this study by PCR combined with those previously reported (*23*) (accession numbers given in table S2) and Y-chromosome sequences generated in this study combined with those publicly available (samples and sources shown in table S1). Maximum likelihood (ML) analyses of the complete mitochondria of the two hemippes combined with donkey and hemione complete mitochondria after deletion of the 11-bp tandem repeat in the HVR were computed using RAxML (*48*) with a generalised time reversible (GTR) nucleotide substitution model, a gamma-distributed rate of variation among sites with four rate categories, and invariant sites (i.e., GTR-GAMMA-I) (fig. S6). We used 100 bootstraps to estimate node robustness.

Genomic analyses

Fastq reads from six modern horse genomes (*24*), six domestic donkey genomes (*25*), three Mongolian khulan genomes (an *E. hemionus* subspecies from the Gobi) (*25*, *26*), two kiang genomes (*E. h. kiang* or *E. kiang*) (*25*, *27*), and one Persian onager genome (an *E. hemionus* subspecies from Iran) (*27*) (table S4) were trimmed with cutadapt (v1.18) (*49*) and aligned to the *E. caballus* reference genome (eqCab2.0) using the BWA (v0.7.17) (*50*) mem program. PCR duplicates were removed using Picard MarkDuplicates (v2.20.0) (*51*), and reads aligning to the reference genome with mapping quality score below 30 were removed using samtools 1.9 (*52*).

We curated the 36 million biallelic variant list used to differentiate equids in the Zonkey workflow (*53*) to filter out variants found in repeated sequences using an EqCab2 genome repeat mask downloaded from the UCSC browser (https://genome.ucsc.edu/cgi-bin/hgTables). The rationale for this filtration was that these variants would be less reliably called using the short, damaged reads typical of ancient DNA libraries, in particular, when mapping reads from a noncaballine equid to the horse reference genome because these genomes are expected to differ markedly in repeat location and sequence variability. This filtration reduced the variant list to 22 million. We then called variants from this list on the modern equid genomes using bcftools (v1.9) (*54*) mpileup -B -q30 -Q30 and bcftools call -m. The vcf file was imported in plink (v1.9) (*55*) and filtered to include only SNPs, removing invariant and multiallelic positions. The final curated list contains 15.5 million SNPs.

Shotgun reads for the Umm el-Marra (UMM9) and Göbekli Tepe (GT64) samples were merged with leeHom (*56*) using the ancientdna option, while the historical hemippe reads were trimmed with cutadapt (v1.18) (*49*). Fragments smaller than 28 bp were discarded, and the remaining reads were aligned to the *E. caballus* reference genome (eqCab2.0) using the BWA (v0.7.17) (*50*) aln program with parameters "-n 0.01 -l 0" followed by samse (UMM9 and GT64) or sampe (hemippe). PCR duplicates were removed

using Picard MarkDuplicates (v2.20.0) (*51*), and reads aligning to the reference genome below a mapping quality score of 20 and a length of 28 bp were removed. To reduce the increase in spurious alignments from shorter reads described in (*57*), mapped reads less than 35 bp containing indels were also removed using an awk script. The ancient nature of the UMM9 and GT64 sequences was confirmed by analyzing the damage profile using mapDamage2 (*58*) of libraries generated from extracts not treated with USER-enzyme (fig. S3). To remove the C->T mutations at the end of the molecules that escaped the USER treatment (fig. S3), the base quality was rescaled at the last two bases using mapDamage2 (*58*). Since all Umm el-Marra samples had very low levels of endogenous equid DNA (0 to 0.18% of reads), they were additionally aligned to the human and bovine genome reference sequence (GRCh37 and ARS_UCD1.2, respectively). Only libraries that had at least fivefold more reads mapping to the horse than to the cow or human genome when a seed length of 18 was used during bwa aln mapping were kept. Summaries of the sequencing results are given in table S5. Hemippe reads were additionally aligned to the kiang mitochondrial genome (NC_016061.1) (*59*) using bwa aln and bwa mem. The resulting *E. h. hemippus* mitochondrial genomes had a mean coverage of 52× (1864 hemippe), 40× (1892 hemippe), and 51× (Göbekli Tepe GT64). Complete mitogenome sequences were generated by consensus calling of the bases using Geneious (*47*). To obtain a full-length mitogenome, gaps were filled using both targeted PCR data of the HVR (*23*) and by analyzing, at the boundary of the gaps, the soft clipped reads resulting from mapping with bwa mem rather than bwa aln.

Nuclear SNPs were called from the Umm el-Marra and hemippe bam files using the samtools (*54*) mpileup command with the following parameters: -B -A -Q20 and specifying only the 15.5 million SNP positions described above. Calling and selection of a single allele for all heterozygous sites were performed using pileupCaller (*60*). This resulted in 6.8 million shared positions between extant equids and the 1864 hemippe, 2.2 million shared with the 1892 hemippe, 10.9 million with the Göbekli Tepe GT64 sample, and 45,604 with the UMM9 sample (table S5).

PCA was performed using EIGENSOFT SmartPCA (v16000) (*61*, *62*) by projecting the samples with partial coverage onto eigenvectors calculated from all shared positions of well-covered equids (projectlsq: YES). For the PCA represented in Fig. 4A and fig. S4 (D and E), we used the 15.2K SNPs shared between UMM9 and both the GT64 and the 1864 hemippe, and only the 1892 hemippe was projected. For the PCA represented in fig. S4 (A to C), we used the 4.1M SNPs shared between the extant equids used and both the GT64 and 1864 hemippe, with both the UMM9 and 1892 hemippe being projected. Admixture (v1.3.0) (*30*) was used to estimate ancestry of the six horses, six donkeys, three Mongolian khulans, two kiangs, the onager, the 1864 and 11,000-year-old GT64 hemippe, and UMM9 using the 15.2K SNPs shared between UMM9 and both the GT64 and 1864 hemippe and a four-population model ($K = 4$). Figure 4B represents the admixture bar graph obtained in 70% of the 30 iterations (90% showed the UMM9 sample as a 1:1 admixture of onager/hemippe and donkey). The bifurcating tree with gene flow was performed using treemix (*31*) with the 40.4K SNPs shared between UMM9 and either of the three hemippes, and considering the following equid groups (number of individuals): horse (six), donkey (six), hemippe (three), UMM9 (one), onager (one), Mongolian khulan (three), and kiang (two). Horses were used as the outgroup, and sample size correction was disabled. The tree obtained with one migration/admixture event is represented in Fig. 4D, and the residuals are plotted in fig. S5. From these 40.4K SNPs, we identified those where all donkeys are identical and all hemippes that have the corresponding positions covered are identical and distinct from the donkeys (4738 SNPs) and counted the SNPs where the UMM9 SNPs are identical to either the donkey- or hemippe-specific SNPs. These counts are represented in Fig. 4C.

The genome phylogeny shown in fig. S7 was obtained using the 738.5K SNPs shared between all three hemippes and equids, after calculation of the pairwise distance matrix between all equids using plink (*55*) and construction of the phylogenetic tree using fastme with nni optimization (*63*).

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at https://science.org/doi/10.1126/sciadv.abm0218

View/request a protocol for this paper from *Bio-protocol*.

REFERENCES AND NOTES

1. P. M. M. G. Akkermans, G. M. Schwartz, *The Archaeology of Syria: From Complex Hunter-Gatherers to Early Urban Societies (c. 16,000–300 BC)* (Cambridge Univ. Press, 2003).
2. J. A. Ur, Cycles of civilization in northern Mesopotamia, 4400—2000 BC. *J. Archaeol. Res.* **18**, 387–431 (2010).
3. K. Maekawa, The donkey and the Persian onager in late third millennium B.C. Mesopotamia and Syria: A rethinking. *J. West Asian Archaeol.* **7**, 1–20 (2006).
4. K. Maekawa, The ass and the onager in Sumer in the late third millennium B.C. *Acta Sumerologica* **1**, 35–62 (1979).
5. J. Zarins, in *Equids in the Ancient World (Beihefte zum Tübinger Atlas des Vorderen Orients, Reihe A, Nr. 19/1)*, R. H. Meadow, H. P. Uerpmann, Eds. (Dr. Ludwig Reichert Verlag, 1986), pp. 165–191.
6. W. Heimpel, Towards an understanding of the term SiKKum. *Rev. Assyriol.* **88**, 5–31 (1994).
7. M. G. Biga, The marriage of Eblaite princess Tagris-Damu with a son of Nagar's king. *Subartu IV* **2**, 17–22 (1998).
8. J. N. Postgate, in *Equids in the Ancient World (Beihefte zum Tübinger Atlas des Vorderen Orients, Reihe A, Nr. 19/1)*, R. H. Meadow, H. P. Uerpmann, Eds. (Dr. Ludwig Reichert Verlag, 1986), pp. 194–205.
9. J. Zarins, *The Domestication of Equidae in Third-Millennium BCE Mesopotamia* (CDL Press, 2014).
10. R. Dolce, Equids as luxury gifts at the centre of interregional economic dynamics in the archaic urban cultures of the ancient near east. *Syria* **91**, 55–75 (2014).
11. J. Clutton-Brock, *Horse Power. A History of the Horse and the Donkey in Human Societies* (Harvard Univ. Press, 1992).
12. S. Guimaraes, B. S. Arbuckle, J. Peters, S. E. Adcock, H. Buitenhuis, H. Chazin, N. Manaseryan, H.-P. Uerpmann, T. Grange, E.-M. Geigl, Ancient DNA shows domestic horses were introduced in the southern Caucasus and Anatolia during the Bronze Age. *Sci. Adv.* **6**, eabb0030 (2020).
13. E. Mohr, Eine durch Hagenbeck importierte Herde des persischen Onagers. *Equus hemionus Onager* **1**, 164–189 (1961).
14. O. Antonius, Beobachtungen an Einhufern in Schönbrunn 1: Der Syrische Halbesel (Equus hemionus hemippus J. Geoffr.). *Zool. Gart. NF* **1**, 19–25 (1929).
15. G. M. Schwartz, H. H. Curvers, S. S. Dunham, B. Stuart, J. A. Weber, A third-millennium B.C. elite mortuary complex at Umm El-Marra, Syria: 2002 and 2004 excavations. *Am. J. Archaeol.* **110**, 603–641 (2006).
16. S. Rossel, F. Marshall, J. Peters, T. Pilgram, M. D. Adams, D. O'Connor, Domestication of the donkey: Timing, processes, and indicators. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 3715–3720 (2008).
17. J. A. Weber, in *Archaeozoology of the Near East VIII*, E. Vila, Ed. (Maison de l'Orient et de la Méditerranée, 2008), vol. TMO 49, pp. 499–519.
18. C. P. Groves, R. H. Meadow, H. P. Uerpmann, Eds. (Dr. Ludwig Reichert Verlag, 1986), vol. I of *Beihefte zum Tübinger Atlas des Vorderen Orients, Reihe A*, pp. 11–47.
19. J. Oates, in *Prehistoric Steppe Adaptation and the Horse*, M. Levine, C. Renfrew, K. Boyle, Eds. (McDonald Institute Monograph, 2003), pp. 115–125.
20. J. Clutton-Brock, R. Burleigh, The animal remains from Abu Salabikh: Preliminary report. *Iraq* **40**, 89–100 (1978).

21. J. Clutton-Brock, in *Equids in the Ancient World (Beihefte zum Tübinger Atlas des Vorderen Orients, Reihe A, Nr. 19/1)*, R. H. Meadow, H. P. Uerpmann, Eds. (Dr. Ludwig Reichert Verlag, 1986), pp. 207–225.
22. E. M. Geigl, T. Grange, Eurasian wild asses in time and space: Morphological versus genetic diversity. *Ann. Anat.* **194**, 88–102 (2012).
23. E. A. Bennett, S. Champlot, J. Peters, B. S. Arbuckle, S. Guimaraes, M. Pruvost, S. Bar-David, S. J. M. Davis, M. Gautier, P. Kaczensky, R. Kuehn, M. Mashkour, A. Morales-Muñiz, E. Pucher, J.-F. Tournepiche, H.-P. Uerpmann, A. Bălăşescu, M. Germonpré, C. Y. Gündem, M.-R. Hemami, P.-E. Moullé, A. Ötzan, M. Uerpmann, C. Walzer, T. Grange, E.-M. Geigl, Taming the late Quaternary phylogeography of the Eurasiatic wild ass through ancient and modern DNA. *PLOS ONE* **12**, e0174216 (2017).
24. H. J. Bandelt, P. Forster, A. Röhl, Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol.* **16**, 37–48 (1999).
25. K. Schmidt, *Sie Bauten die Ersten Tempel* (Beck Verlag, 2006).
26. L. Orlando, A. Ginolhac, G. Zhang, D. Froese, A. Albrechtsen, M. Stiller, M. Schubert, E. Cappellini, B. Petersen, I. Moltke, P. L. F. Johnson, M. Fumagalli, J. T. Vilstrup, M. Raghavan, T. Korneliussen, A.-S. Malaspinas, J. Vogt, D. Szklarczyk, C. D. Kelstrup, J. Vinther, A. Dolocan, J. Stenderup, A. M. V. Velazquez, J. Cahill, M. Rasmussen, X. Wang, J. Min, G. D. Zazula, A. Seguin-Orlando, C. Mortensen, K. Magnussen, J. F. Thompson, J. Weinstock, K. Gregersen, K. H. Røed, V. Eisenmann, C. J. Rubin, D. C. Miller, D. F. Antczak, M. F. Bertelsen, S. Brunak, K. A. S. Al-Rasheid, O. Ryder, L. Andersson, J. Mundy, A. Krogh, M. T. P. Gilbert, K. Kjær, T. Sicheritz-Ponten, L. J. Jensen, J. V. Olsen, M. Hofreiter, R. Nielsen, B. Shapiro, J. Wang, E. Willerslev, Recalibrating *Equus* evolution using the genome sequence of an early Middle Pleistocene horse. *Nature* **499**, 74–78 (2013).
27. C. Wang, H. Li, Y. Guo, J. Huang, Y. Sun, J. Min, J. Wang, X. Fang, Z. Zhao, S. Wang, Y. Zhang, Q. Liu, Q. Jiang, X. Wang, Y. Guo, C. Yang, Y. Wang, F. Tian, G. Zhuang, Y. Fan, Q. Gao, Y. Li, Z. Ju, J. Li, R. Li, M. Hou, G. Yang, G. Liu, W. Liu, J. Guo, S. Pan, G. Fan, W. Zhang, R. Zhang, J. Yu, X. Zhang, Q. Yin, C. Ji, Y. Jin, G. Yue, M. Liu, J. Xu, S. Liu, J. Jordana, A. Noce, M. Amills, D. D. Wu, S. Li, X. Zhou, J. Zhong, Donkey genomes provide new insights into domestication and selection for coat color. *Nat. Commun.* **11**, 6014 (2020).
28. J. Huang, Y. Zhao, D. Bai, W. Shiraigol, B. Li, L. Yang, J. Wu, W. Bao, X. Ren, B. Jin, Q. Zhao, A. Li, S. Bao, W. Bao, Z. Xing, A. An, Y. Gao, R. Wei, Y. Bao, T. Bao, H. Han, H. Bai, Y. Bao, Y. Zhang, D. Daidiikhuu, W. Zhao, S. Liu, J. Ding, W. Ye, F. Ding, Z. Sun, Y. Shi, Y. Zhang, H. Meng, M. Dugarjaviin, Donkey genome and insight into the imprinting of fast karyotype evolution. *Sci. Rep.* **5**, 14106 (2015).
29. H. Jonsson, M. Schubert, A. Seguin-Orlando, A. Ginolhac, L. Petersen, M. Fumagalli, A. Albrechtsen, B. Petersen, T. S. Korneliussen, J. T. Vilstrup, T. Lear, J. L. Myka, J. Lundquist, D. C. Miller, A. H. Alfarhan, S. A. Alquraishi, K. A. Al-Rasheid, J. Stagegaard, G. Strauss, M. F. Bertelsen, T. Sicheritz-Ponten, D. F. Antczak, E. Bailey, R. Nielsen, E. Willerslev, L. Orlando, Speciation with gene flow in equids despite extensive chromosomal plasticity. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 18655–18660 (2014).
30. D. H. Alexander, J. Novembre, K. Lange, Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
31. J. K. Pickrell, J. K. Pritchard, Inference of population splits and mixtures from genome-wide allele frequency data. *PLOS Genet.* **8**, e1002967 (2012).
32. V. A. Ščekin, A. V. Škurgin, Rezultaty gibridizacii lošadei i oslov s kulanami. *Trudy vsesojuznogo naučno-issledovatelskogo instituta konevodstva* **18**, 106–118 (1950).
33. S. Bökönyi, in *Equids in the Ancient World (Beihefte zum Tübinger Atlas des Vorderen Orients, Reihe A, Nr. 19/1)*, R. H. Meadow, H. P. Uerpmann, Eds. (Dr. Ludwig Reichert Verlag, 1986), pp. 302–317.
34. R. H. Meadow, in *Equids in the Ancient World (Beihefte zum Tübinger Atlas des Vorderen Orients, Reihe A, Nr. 19/1)*, R. H. Meadow, H. P. Uerpmann, Eds. (Dr. Ludwig Riechert Verlag, 1986), p. 284.
35. S. Champlot, C. Berthelot, M. Pruvost, E. A. Bennett, T. Grange, E.-M. Geigl, An efficient multistrategy DNA decontamination procedure of PCR reagents for hypersensitive PCR applications. *PLOS ONE* **5**, e13042 (2010).
36. G. M. Schwartz, H. H. Curvers, S. S. Dunham, J. A. Weber, From urban origins to imperial integration in Western Syria: Umm el-Marra 2006, 2008. *Am. J. Archaeol.* **116**, 157 (2012).
37. E. A. Bennett, D. Massilani, G. Lizzo, J. Daligault, E.-M. Geigl, T. Grange, Library construction for ancient genomics: Single strand or double strand? *Biotechniques* **56**, 289–298 (2014).
38. M. Pruvost, R. Schwarz, V. B. Correia, S. Champlot, S. Braguier, N. Morel, Y. Fernandez-Jalvo, T. Grange, E. M. Geigl, Freshly excavated fossil bones are best for amplification of ancient DNA. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 739–744 (2007).
39. P. Korlević, T. Gerber, M.-T. Gansauge, M. Hajdinjak, S. Nagel, A. Aximu-Petri, M. Meyer, Reducing microbial and human contamination in DNA extractions from ancient bones and teeth. *Biotechniques* **59**, 87–93 (2015).
40. O. Gorgé, E. A. Bennett, D. Massilani, J. Daligault, M. Pruvost, E.-M. Geigl, T. Grange, Analysis of Ancient DNA in microbial ecology. *Methods Mol. Biol.* **1399**, 289–315 (2016).
41. J. Dabney, M. Knapp, I. Glocke, M. T. Gansauge, A. Weihmann, B. Nickel, C. Valdiosera, N. Garcia, S. Paabo, J. L. Arsuaga, M. Meyer, Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 15758–15763 (2013).
42. I. Glocke, M. Meyer, Extending the spectrum of DNA sequences retrieved from ancient bones and teeth. *Genome Res.* **27**, 1230–1237 (2017).
43. M. Pruvost, T. Grange, E. M. Geigl, Minimizing DNA contamination by using UNG-coupled quantitative real-time PCR on degraded DNA samples: Application to ancient DNA studies. *Biotechniques* **38**, 569–575 (2005).
44. D. Massilani, S. Guimaraes, J.-P. Brugal, E. A. Bennett, M. Tokarska, R.-M. Arbogast, G. Baryshnikov, G. Boeskorov, J.-C. Castel, S. Davydov, S. Madelaine, O. Putelat, N. N. Spasskaya, H.-P. Uerpmann, T. Grange, E.-M. Geigl, Past climate changes, population dynamics and the origin of Bison in Europe. *BMC Biol.* **14**, 93 (2016).
45. E. A. Bennett, I. Crevecoeur, B. Viola, A. P. Derevianko, M. V. Shunkov, T. Grange, B. Maureille, E.-M. Geigl, Morphology of the Denisovan phalanx closer to modern humans than to Neanderthals. *Sci. Adv.* **5**, eaaw3950 (2019).
46. M. T. Gansauge, M. Meyer, Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. *Nat. Protoc.* **8**, 737–748 (2013).
47. M. Kearse, R. Moir, A. Wilson, S. Stones-Havas, M. Cheung, S. Sturrock, S. Buxton, A. Cooper, S. Markowitz, C. Duran, T. Thierer, B. Ashton, P. Meintjes, A. Drummond, Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647–1649 (2012).
48. A. Stamatakis, RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
49. M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17**, 10 (2011).
50. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
51. Broad Institute, "Picard Toolkit". GitHub Repository (2019); https://broadinstitute.github.io/picard/.
52. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
53. M. Schubert, M. Mashkour, C. Gaunitz, A. Fages, A. Seguin-Orlando, S. Sheikhi, A. H. Alfarhan, S. A. Alquraishi, K. A. S. Al-Rasheid, R. Chuang, L. Ermini, C. Gamba, J. Weinstock, O. Vedat, L. Orlando, Zonkey: A simple, accurate and sensitive pipeline to genetically identify equine F1-hybrids in archaeological assemblages. *J. Archaeol. Sci.* **78**, 147–157 (2017).
54. H. Li, A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
55. C. C. Chang, C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, J. J. Lee, Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience* **4**, 7 (2015).
56. G. Renaud, U. Stenzel, J. Kelso, leeHom: Adaptor trimming and merging for Illumina sequencing reads. *Nucleic Acids Res.* **42**, e141 (2014).
57. C. de Filippo, M. Meyer, K. Prüfer, Quantifying and reducing spurious alignments for the analysis of ultra-short ancient DNA sequences. *BMC Biol.* **16**, 121 (2018).
58. H. Jónsson, A. Ginolhac, M. Schubert, P. L. F. Johnson, L. Orlando, mapDamage2.0: Fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* **29**, 1682–1684 (2013).
59. Y. Luo, Y. Chen, F. Liu, C. Jiang, Y. Gao, Mitochondrial genome sequence of the Tibetan wild ass (Equus kiang). *Mitochondrial DNA* **22**, 6–8 (2011).
60. S. Schiffels, "pileupCaller". GitHub Repository (Max Planck Institute for the Science of Human History, 2019); https://github.com/stschiff/sequenceTools.
61. A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, D. Reich, Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
62. N. Patterson, A. L. Price, D. Reich, Population structure and eigenanalysis. *PLOS Genet.* **2**, e190 (2006).
63. V. Lefort, R. Desper, O. Gascuel, FastME 2.0: A comprehensive, accurate, and fast distance-based phylogeny inference program. *Mol. Biol. Evol.* **32**, 2798–2800 (2015).
64. A. von den Driesch, J. Peters, Vorläufiger Bericht über die archäozoologischen Untersuchungen am Göbekli Tepe und am Gürcütepe bei Urfa, Türkei. *Istanbuler Mitteilungen* **49**, 23–39 (1999).

Acknowledgments: We are very grateful to K. Maekawa for discussion of the cuneiform terms and for providing a drawing of the cuneiform tablet BM 23836, conserved in the British Museum, London. We are also grateful to F. Zachos and A. Bibl from the "Naturhistorisches Museum Wien" for providing the hemippe specimens. We thank G. Heindl from the Geschichtsforschung & Dokumentation department of the Schönbrunner Tiergarten GmbH, Vienna, Austria, for help with the search for photos of the last hemippe. We thank Elodie Trunet Réserve Africaine de Sigean, France, for providing the sample of the Somalian ass. We

thank T. Kovaleva for translations of Russian articles, C. Martin for critical reading of the manuscript, and O. Gorgé for assistance with some of the sequencing. **Funding:** The paleogenomic facility of the Institut Jacques Monod obtained support from the University Paris Diderot within the program "Actions de recherches structurantes." The sequencing facility of the Institut Jacques Monod, Paris, is supported by grants from the University Paris Diderot, the Fondation pour la Recherche Médicale (DGE20111123014), and the Région Ile-de-France (11015901). Moreover, we acknowledge support from the French national research center CNRS. We are grateful to the Directorate-General of Antiquities and Museums, Syria, for its support of the Umm el-Marra project. The excavations at Umm el-Marra were funded by the National Science Foundation (grants BCS-0137513 and BCS-0545610), the National Geographic Society, the Metropolitan Museum of Art, the Arthur and Isadora Dellheim Foundation, and the Johns Hopkins University. Faunal research at Göbekli Tepe was funded by the Deutsche Forschungsgemeinschaft (DFG) under grant PE 424/10-1-4 to J.P.

Author contributions: E.-M.G. and J.W. initiated the project. E.-M.G. and T.G. conceptualized and supervised the project. J.W., J.P., and G.M.S. provided material. E.A.B., W.B., S.C., E.-M.G., and T.G. performed the laboratory work. E.A.B., T.G., and E.-M.G. analyzed the data. E.A.B., T.G., and E.-M.G. wrote the paper with input from J.P. and G.M.S. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** Sequence data generated for this study are available from EBI European Nucleotide Archive PRJEB47929. Syrian wild ass (hemippe) mitochondrial sequences are available on GenBank MN990427, OK393913, and OK393914.

Submitted 20 August 2021
Accepted 22 November 2021
Published 14 January 2022
10.1126/sciadv.abm0218

ScienceAdvances

The genetic identity of the earliest human-made hybrid animals, the kungas of Syro-Mesopotamia

E. Andrew BennettJill WeberWejden BendhaferSophie ChamplotJoris PetersGlenn M. SchwartzThierry GrangeEva-Maria Geigl

Sci. Adv., 8 (2), eabm0218. • DOI: 10.1126/sciadv.abm0218

Bibliographie :

A

Achilli, A., Bonfiglio, S., Olivieri, A., Malusà, A., Pala, M., Kashani, B. H., Perego, U. A., Ajmone-Marsan, P., Liotta, L., Semino, O., Bandelt, H.-J., Ferretti, L., & Torroni, A. (2009). The Multifaceted Origin of Taurine Cattle Reflected by the Mitochondrial Genome. *PLoS ONE*, *4*(6), e5753. https://doi.org/10.1371/journal.pone.0005753

Achilli, A., Olivieri, A., Pellecchia, M., Uboldi, C., Colli, L., Al-Zahery, N., Accetturo, M., Pala, M., Kashani, B. H., Perego, U. A., Battaglia, V., Fornarino, S., Kalamati, J., Houshmand, M., Negrini, R., Semino, O., Richards, M., Macaulay, V., Ferretti, L., … Torroni, A. (2008). Mitochondrial genomes of extinct aurochs survive in domestic cattle. *Current Biology*, *18*(4), R157–R158. https://doi.org/10.1016/j.cub.2008.01.019

Ajmone, M. (2010). *On the origin of cattle: How aurochs became domestic and colonized the world*.

Akbar Khan, M., Kostopoulos, D. S., Akhtar, M., & Nazir, M. (2010). Bison remains from the Upper Siwaliks of Pakistan. *Neues Jahrbuch Für Geologie Und Paläontologie - Abhandlungen*, *258*(1), 121–128. https://doi.org/10.1127/0077-7749/2010/0090

Akey, J. M. (2009). Constructing genomic maps of positive selection in humans: Where do we go from here? *Genome Research*, *19*(5), 711–722. https://doi.org/10.1101/gr.086652.108

Allentoft, M. E., Collins, M., Harker, D., Haile, J., Oskam, C. L., Hale, M. L., Campos, P. F., Samaniego, J. A., Gilbert, M. T. P., Willerslev, E., Zhang, G., Scofield, R. P., Holdaway, R. N., & Bunce, M. (2012). The half-life of DNA in bone: Measuring decay kinetics in 158 dated fossils. *Proceedings. Biological Sciences*, *279*(1748), 4724–4733. https://doi.org/10.1098/rspb.2012.1745

Andersson, L. (2013). Molecular consequences of animal breeding. *Current Opinion in Genetics & Development*, *23*(3), 295–301. https://doi.org/10.1016/j.gde.2013.02.014
Antonioli & al. (2012). *The land bridge between Europe and Sicily over the past 40 kyrs: Timing of emersion and implications for the migration of Homo sapiens*.

A.T. Primo. (1992). *EL GANADO BOVINO IBERICO EN LAS AMERICAS: 500 AÑOS DESPUÉS*.

B

Balasse, M., Brugal, J.-P., Dauphin, Y., Geigl, E.-M., Oberlin, C., & Reiche, I. (Eds.). (2015). *Messages d'os: Archéométrie du squelette animal et humain*. Editions des archives contemporaines. https://doi.org/10.17184/eac.9782813001641

Barker, G. (1983). Domesticated Animals from Early Times. By J. Clutton-Brock. 25.5 × 20 cm. Pp. 208 + 100 figs. + 48 b/w, 12 col. pls. London: Heinemann and British Museum (Natural History), 1981. ISBN 434-13950-5. £9.95. *The Antiquaries Journal*, *63*(1), 138–139. https://doi.org/10.1017/S0003581500014414

Barnosky, A. D. (2004). Assessing the Causes of Late Pleistocene Extinctions on the Continents. *Science*, *306*(5693), 70–75. https://doi.org/10.1126/science.1101476

Bekaert, B., Ellerington, R., Van den Abbeele, L., & Decorte, R. (2016). In-Solution Hybridization for the Targeted Enrichment of the Whole Mitochondrial Genome. In W. Goodwin (Ed.), *Forensic DNA Typing Protocols* (Vol. 1420, pp. 173–183). Springer New York. https://doi.org/10.1007/978-1-4939-3597-0_14

Belyaev, D. K. (1979). Destabilizing selection as a factor in domestication. *Journal of Heredity*, *70*(5), 301–308. https://doi.org/10.1093/oxfordjournals.jhered.a109263

Beniash, E. (2011). Biominerals--hierarchical nanocomposites: The example of bone. *Wiley Interdisciplinary Reviews. Nanomedicine and Nanobiotechnology*, *3*(1), 47–69. https://doi.org/10.1002/wnan.105

Bennett, E. A., Champlot, S., Peters, J., Arbuckle, B. S., Guimaraes, S., Pruvost, M., Bar-David, S., Davis, S. J. M., Gautier, M., Kaczensky, P., Kuehn, R., Mashkour, M., Morales-Muñiz, A., Pucher, E., Tournepiche, J.-F., Uerpmann, H.-P., Bălăşescu, A., Germonpré, M., Gündem, C. Y., … Geigl, E.-M. (2017). Taming the late Quaternary phylogeography of the Eurasiatic wild ass through ancient and modern DNA. *PLOS ONE*, *12*(4), e0174216. https://doi.org/10.1371/journal.pone.0174216

Bennett, E. A., Massilani, D., Lizzo, G., Daligault, J., Geigl, E.-M., & Grange, T. (2014). Library construction for ancient genomics: Single strand or double strand? *BioTechniques*, *56*(6). https://doi.org/10.2144/000114176

Blench, R., & MacDonald, K. (Eds.). (2006). *The Origins and Development of African Livestock* (0 ed.). Routledge. https://doi.org/10.4324/9780203984239

Bocherens, H., Hofman-Kamińska, E., Drucker, D. G., Schmölcke, U., & Kowalczyk, R. (2015a). European Bison as a Refugee Species? Evidence from Isotopic Data on Early Holocene Bison and Other Large Herbivores in Northern Europe. *PLOS ONE*, *10*(2), e0115090. https://doi.org/10.1371/journal.pone.0115090

Bocherens, H., Hofman-Kamińska, E., Drucker, D. G., Schmölcke, U., & Kowalczyk, R. (2015b). European Bison as a Refugee Species? Evidence from Isotopic Data on Early Holocene Bison and Other Large Herbivores in Northern Europe. *PLOS ONE*, *10*(2), e0115090. https://doi.org/10.1371/journal.pone.0115090

Boitard, S. (2012). *Detecting Selective Sweeps from Pooled Next-Generation Sequencing Samples*.

Bokonyi, S. (1974). *History of Domestic Mammals in Central and Eastern Europe*.

Bollongino, R., Burger, J., Powell, A., Mashkour, M., Vigne, J.-D., & Thomas, M. G. (2012). Modern Taurine Cattle Descended from Small Number of Near-Eastern Founders. *Molecular Biology and Evolution*, *29*(9), 2101–2104. https://doi.org/10.1093/molbev/mss092

Bollongino, R., Edwards, C. J., Alt, K. W., Burger, J., & Bradley, D. G. (2006). Early history of European domestic cattle as revealed by ancient DNA. *Biology Letters*, *2*(1), 155–159. https://doi.org/10.1098/rsbl.2005.0404

Bonfiglio, S., Achilli, A., Olivieri, A., Negrini, R., Colli, L., Liotta, L., Ajmone-Marsan, P., Torroni, A., & Ferretti, L. (2010). The Enigmatic Origin of Bovine mtDNA Haplogroup R: Sporadic Interbreeding or an Independent Event of Bos primigenius Domestication in Italy? *PLoS ONE*, *5*(12), e15760. https://doi.org/10.1371/journal.pone.0015760

Boskey, A. L. (2007). Mineralization of Bones and Teeth. *Elements*, *3*(6), 385–391. https://doi.org/10.2113/GSELEMENTS.3.6.385

Boskey, A. L., & Mendelsohn, R. (2005). Infrared spectroscopic characterization of mineralized tissues. *Vibrational Spectroscopy*, *38*(1–2), 107–114. https://doi.org/10.1016/j.vibspec.2005.02.015

Boussaha, M. (2015). *Genome-Wide Study of Structural Variants in Bovine Holstein, Montbéliarde and Normande Dairy Breeds*.

Bradley, D. G., MacHugh, D. E., Cunningham, P., & Loftus, R. T. (1996). Mitochondrial diversity and the origins of African and European cattle. *Proceedings of the National Academy of Sciences*, *93*(10), 5131–5135. https://doi.org/10.1073/pnas.93.10.5131

Briggs, A. W., & Heyn, P. (2012). Preparation of Next-Generation Sequencing Libraries from Damaged DNA. In B. Shapiro & M. Hofreiter (Eds.), *Ancient DNA* (Vol. 840, pp. 143–154). Humana Press. https://doi.org/10.1007/978-1-61779-516-9_18

Briggs, A. W., Stenzel, U., Johnson, P. L. F., Green, R. E., Kelso, J., Prufer, K., Meyer, M., Krause, J., Ronan, M. T., Lachmann, M., & Paabo, S. (2007). Patterns of damage in genomic DNA sequences from a Neandertal. *Proceedings of the National Academy of Sciences*, *104*(37), 14616–14621. https://doi.org/10.1073/pnas.0704665104

Briggs, D. E. G. (2003). The Role of Decay and Mineralization in the Preservation of Soft-Bodied Fossils. *Annual Review of Earth and Planetary Sciences*, *31*(1), 275–301. https://doi.org/10.1146/annurev.earth.31.100901.144746

Bro-Jørgensen, M. H., Carøe, C., Vieira, F. G., Nestor, S., Hallström, A., Gregersen, K. M., Etting, V., Gilbert, M. T. P., & Sinding, M.-H. S. (2018). Ancient DNA analysis of Scandinavian medieval drinking horns and the horn of the last aurochs bull. *Journal of Archaeological Science*, *99*, 47–54. https://doi.org/10.1016/j.jas.2018.09.001

Brunel, S., Bennett, E. A., Cardin, L., Garraud, D., Barrand Emam, H., Beylier, A., Boulestin, B., Chenal, F., Ciesielski, E., Convertini, F., Dedet, B., Desbrosse-Degobertiere, S., Desenne, S., Dubouloz, J., Duday, H., Escalon, G., Fabre, V., Gailledrat, E., Gandelin, M., … Pruvost, M. (2020). Ancient genomes from present-day France unveil 7,000 years of its demographic history. *Proceedings of the National Academy of Sciences*, *117*(23), 12791–12798. https://doi.org/10.1073/pnas.1918034117

Budiansky, S. (1997). *The covenant of the wild: Why animals chose domestication*. Phoenix.

Cai, D., Zhang, N., Zhu, S., Chen, Q., Wang, L., Zhao, X., Ma, X., Royle, T. C. A., Zhou, H., & Yang, D. Y. (2018). Ancient DNA reveals evidence of abundant aurochs (Bos primigenius) in Neolithic Northeast China. *Journal of Archaeological Science*, *98*, 72–80. https://doi.org/10.1016/j.jas.2018.08.003

Carlson, C. S. (2005). Genomic regions exhibiting positive selection identified from dense genotype data. *Genome Research*, *15*(11), 1553–1565. https://doi.org/10.1101/gr.4326505

Carpenter, M. L., Buenrostro, J. D., Valdiosera, C., Schroeder, H., Allentoft, M. E., Sikora, M., Rasmussen, M., Gravel, S., Guillén, S., Nekhrizov, G., Leshtakov, K., Dimitrova, D., Theodossiev, N., Pettener, D., Luiselli, D., Sandoval, K., Moreno-Estrada, A., Li, Y., Wang, J., … Bustamante, C. D. (2013). Pulling out the 1%: Whole-Genome Capture for the Targeted Enrichment of Ancient DNA Sequencing Libraries. *The American Journal of Human Genetics*, *93*(5), 852–864. https://doi.org/10.1016/j.ajhg.2013.10.002

Chamary, J. V., Parmley, J. L., & Hurst, L. D. (2006). Hearing silence: Non-neutral evolution at synonymous sites in mammals. *Nature Reviews Genetics*, *7*(2), 98–108. https://doi.org/10.1038/nrg1770

Champlot, S., Berthelot, C., Pruvost, M., Bennett, E. A., Grange, T., & Geigl, E.-M. (2010). An Efficient Multistrategy DNA Decontamination Procedure of PCR Reagents for Hypersensitive PCR Applications. *PLoS ONE*, *5*(9), e13042. https://doi.org/10.1371/journal.pone.0013042

Charlesworth, D. (2006). Balancing Selection and Its Effects on Sequences in Nearby Genome Regions. *PLoS Genetics*, *2*(4), e64. https://doi.org/10.1371/journal.pgen.0020064

Chen, N., Cai, Y., Chen, Q., Li, R., Wang, K., Huang, Y., Hu, S., Huang, S., Zhang, H., Zheng, Z., Song, W., Ma, Z., Ma, Y., Dang, R., Zhang, Z., Xu, L., Jia, Y., Liu, S., Yue, X., … Lei, C. (2018). Whole-genome resequencing reveals world-wide ancestry and adaptive introgression events of domesticated cattle in East Asia. *Nature Communications*, *9*(1), 2337. https://doi.org/10.1038/s41467-018-04737-0

Clutton-Brock, J. (1992). The process of domestication. *Mammal Review*, *22*(2), 79–85. https://doi.org/10.1111/j.1365-2907.1992.tb00122.x

Comas, D., Paabo, S., & Bertranpetit, J. (1995). Heteroplasmy in the control region of human mitochondrial DNA. *Genome Research*, *5*(1), 89–90. https://doi.org/10.1101/gr.5.1.89

Cooper, A. (1994). DNA from Museum Specimens. In B. Herrmann & S. Hummel (Eds.), *Ancient DNA* (pp. 149–165). Springer New York. https://doi.org/10.1007/978-1-4612-4318-2_10

Cooper, A., Lalueza-Fox, C., Anderson, S., Rambaut, A., Austin, J., & Ward, R. (2001). Complete mitochondrial genome sequences of two extinct moas clarify ratite evolution. *Nature*, *409*(6821), 704–707. https://doi.org/10.1038/35055536

Cooper, A., Turney, C., Hughen, K. A., Brook, B. W., McDonald, H. G., & Bradshaw, C. J. A. (2015). Abrupt warming events drove Late Pleistocene Holarctic megafaunal turnover. *Science*, *349*(6248), 602–606. https://doi.org/10.1126/science.aac4315

Côté, N. M. L., Daligault, J., Pruvost, M., Bennett, E. A., Gorgé, O., Guimaraes, S., Capelli, N., Le Bailly, M., Geigl, E.-M., & Grange, T. (2016). A New High-Throughput Approach to Genotype Ancient Human Gastrointestinal Parasites. *PLOS ONE*, *11*(1), e0146230. https://doi.org/10.1371/journal.pone.0146230

Dabney, J., Knapp, M., Glocke, I., Gansauge, M.-T., Weihmann, A., Nickel, B., Valdiosera, C., Garcia, N., Paabo, S., Arsuaga, J.-L., & Meyer, M. (2013). Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proceedings of the National Academy of Sciences*, *110*(39), 15758–15763. https://doi.org/10.1073/pnas.1314445110

Daetwyler, H. D., Capitan, A., Pausch, H., Stothard, P., van Binsbergen, R., Brøndum, R. F., Liao, X., Djari, A., Rodriguez, S. C., Grohs, C., Esquerré, D., Bouchez, O., Rossignol, M.-N., Klopp, C., Rocha, D., Fritz, S., Eggen, A., Bowman, P. J., Coote, D., … Hayes, B. J. (2014). Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nature Genetics*, *46*(8), 858–865. https://doi.org/10.1038/ng.3034

Damgaard, P. B., Margaryan, A., Schroeder, H., Orlando, L., Willerslev, E., & Allentoft, M. E. (2015). Improving access to endogenous DNA in ancient bones and teeth. *Scientific Reports*, *5*(1), 11184. https://doi.org/10.1038/srep11184

Dannemann, M., Andrés, A. M., & Kelso, J. (2016). Introgression of Neandertal- and Denisovan-like Haplotypes Contributes to Adaptive Variation in Human Toll-like Receptors. *The American Journal of Human Genetics*, *98*(1), 22–33. https://doi.org/10.1016/j.ajhg.2015.11.015

Dannemann, M., & Kelso, J. (2017). The Contribution of Neanderthals to Phenotypic Variation in Modern Humans. *The American Journal of Human Genetics*, *101*(4), 578–589. https://doi.org/10.1016/j.ajhg.2017.09.010

Darcy Morey. (1994). *The Early Evolution of the Domestic Dog*. https://doi.org/10.2307/29775234
Darriba, D., Posada, D., Kozlov, A. M., Stamatakis, A., Morel, B., & Flouri, T. (2020). ModelTest-NG: A New and Scalable Tool for the Selection of DNA and Protein Evolutionary Models. *Molecular Biology and Evolution*, *37*(1), 291–294. https://doi.org/10.1093/molbev/msz189

Davis, S. J. M. (1981). The effects of temperature change and domestication on the body size of Late Pleistocene to Holocene mammals of Israel. *Paleobiology*, *7*(1), 101–114. https://doi.org/10.1017/S0094837300003821

Davit-Béal, T., Allizard, F., & Sire, J.-Y. (2007). Enameloid/enamel transition through successive tooth replacements in Pleurodeles waltl (Lissamphibia, Caudata). *Cell and Tissue Research*, *328*(1), 167–183. https://doi.org/10.1007/s00441-006-0306-1

Decker, J. E., McKay, S. D., Rolf, M. M., Kim, J., Molina Alcalá, A., Sonstegard, T. S., Hanotte, O., Götherström, A., Seabury, C. M., Praharani, L., Babar, M. E., Correia de Almeida Regitano, L., Yildiz, M. A., Heaton, M. P., Liu, W.-S., Lei, C.-Z., Reecy, J. M., Saif-Ur-Rehman, M., Schnabel, R.

D., & Taylor, J. F. (2014). Worldwide Patterns of Ancestry, Divergence, and Admixture in Domesticated Cattle. *PLoS Genetics*, *10*(3), e1004254. https://doi.org/10.1371/journal.pgen.1004254

Decker, J. E., Pires, J. C., Conant, G. C., McKay, S. D., Heaton, M. P., Chen, K., Cooper, A., Vilkki, J., Seabury, C. M., Caetano, A. R., Johnson, G. S., Brenneman, R. A., Hanotte, O., Eggert, L. S., Wiener, P., Kim, J.-J., Kim, K. S., Sonstegard, T. S., Van Tassell, C. P., … Taylor, J. F. (2009). Resolving the evolution of extant and extinct ruminants with high-throughput phylogenomics. *Proceedings of the National Academy of Sciences*, *106*(44), 18644–18649. https://doi.org/10.1073/pnas.0904691106

Dementiev, G.-P. (1958). QUELQUES NOTES SUR L'AUROCHS. *Mammalia*, *22*(1–4). https://doi.org/10.1515/mamm.1958.22.1-4.161

Der Sarkissian, C., Ermini, L., Jónsson, H., Alekseev, A. N., Crubezy, E., Shapiro, B., & Orlando, L. (2014). Shotgun microbial profiling of fossil remains. *Molecular Ecology*, *23*(7), 1780–1798. https://doi.org/10.1111/mec.12690

Drees, M. (2005). Sexual dimorphism in PleistoceneBison priscus (Mammalia, Bovidae) with a discussion on the position ofBison schoetensacki. *Senckenbergiana Lethaea*, *85*(1), 153–157. https://doi.org/10.1007/BF03043424

Drummond, A. J. (2005). Bayesian Coalescent Inference of Past Population Dynamics from Molecular Sequences. *Molecular Biology and Evolution*, *22*(5), 1185–1192. https://doi.org/10.1093/molbev/msi103

Drummond, A. J., & Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, *7*(1), 214. https://doi.org/10.1186/1471-2148-7-214

E

E.Cerilli & C.Petronio. (1992). *Biometrical_variations_of_Bos_primigenius_Bojanus_1827_from_middle_Pleistocene_to_Holocene*.

Eckhart, L., Bach, J., Ban, J., & Tschachler, E. (2000). Melanin Binds Reversibly to Thermostable DNA Polymerase and Inhibits Its Activity. *Biochemical and Biophysical Research Communications*, *271*(3), 726–730. https://doi.org/10.1006/bbrc.2000.2716

Edgar, R. C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, *5*(1), 113. https://doi.org/10.1186/1471-2105-5-113

Edwards, C. J., Bollongino, R., Scheu, A., Chamberlain, A., Tresset, A., Vigne, J.-D., Baird, J. F., Larson, G., Ho, S. Y. W., Heupink, T. H., Shapiro, B., Freeman, A. R., Thomas, M. G., Arbogast, R.-M., Arndt, B., Bartosiewicz, L., Benecke, N., Budja, M., Chaix, L., … Burger, J. (2007a). Mitochondrial DNA analysis shows a Near Eastern Neolithic origin for domestic cattle and no indication of domestication of European aurochs. *Proceedings of the Royal Society B: Biological Sciences*, *274*(1616), 1377–1385. https://doi.org/10.1098/rspb.2007.0020

Edwards, C. J., Bollongino, R., Scheu, A., Chamberlain, A., Tresset, A., Vigne, J.-D., Baird, J. F., Larson, G., Ho, S. Y. W., Heupink, T. H., Shapiro, B., Freeman, A. R., Thomas, M. G., Arbogast, R.-M., Arndt, B., Bartosiewicz, L., Benecke, N., Budja, M., Chaix, L., … Burger, J. (2007b). Mitochondrial DNA analysis shows a Near Eastern Neolithic origin for domestic cattle and no indication of domestication of European aurochs. *Proceedings of the Royal Society B: Biological Sciences*, *274*(1616), 1377–1385. https://doi.org/10.1098/rspb.2007.0020

Edwards, C. J., Magee, D. A., Park, S. D. E., McGettigan, P. A., Lohan, A. J., Murphy, A., Finlay, E. K., Shapiro, B., Chamberlain, A. T., Richards, M. B., Bradley, D. G., Loftus, B. J., & MacHugh, D. E. (2010). A Complete Mitochondrial Genome Sequence from a Mesolithic Wild Aurochs (Bos primigenius). *PLoS ONE*, *5*(2), e9255. https://doi.org/10.1371/journal.pone.0009255

Efron, B., Halloran, E., & Holmes, S. (1996). Bootstrap confidence levels for phylogenetic trees. *Proceedings of the National Academy of Sciences*, *93*(14), 7085–7090. https://doi.org/10.1073/pnas.93.14.7085

Ellegaard, M., Clokie, M. R. J., Czypionka, T., Frisch, D., Godhe, A., Kremp, A., Letarov, A., McGenity, T. J., Ribeiro, S., & John Anderson, N. (2020). Dead or alive: Sediment DNA archives as tools for tracking aquatic evolution and adaptation. *Communications Biology*, *3*(1), 169. https://doi.org/10.1038/s42003-020-0899-z

Endler, J. A. (1991). Variation in the appearance of guppy color patterns to guppies and their predators under different visual conditions. *Vision Research*, *31*(3), 587–608. https://doi.org/10.1016/0042-6989(91)90109-I

Enk, J. M., Devault, A. M., Kuch, M., Murgha, Y. E., Rouillard, J.-M., & Poinar, H. N. (2014). Ancient Whole Genome Enrichment Using Baits Built from Modern DNA. *Molecular Biology and Evolution*, *31*(5), 1292–1294. https://doi.org/10.1093/molbev/msu074

F

Felsenstein, J. (1985). Confidence Limits on Phylogenies: An Approach Using the Bootstrap. *Evolution*, *39*(4), 783. https://doi.org/10.2307/2408678

Ferguson, S., Warny, S., Escarguel, G., & Mudie, P. J. (2018). MIS 5–1 dinoflagellate cyst analyses and morphometric evaluation of Galeacysta etrusca and Spiniferites cruciformis in southwestern Black Sea. *Quaternary International*, *465*, 117–129. https://doi.org/10.1016/j.quaint.2016.07.035

Fincham, A. G., Moradian-Oldak, J., & Simmer, J. P. (1999). The Structural Biology of the Developing Dental Enamel Matrix. *Journal of Structural Biology*, *126*(3), 270–299. https://doi.org/10.1006/jsbi.1999.4130

Fontsere, C., Manuel, M., Marques Bonet, T., & Kuhlwilm, M. (2019). Admixture in Mammals and How to Understand Its Functional Implications: On the Abundance of Gene Flow in Mammalian Species, Its Impact on the Genome, and Roads into a Functional Understanding. *BioEssays*, *41*(12), 1900123. https://doi.org/10.1002/bies.201900123

Fratzl, P. (Ed.). (2008). *Collagen: Structure and mechanics*. Springer.

Froese, D., Stiller, M., Heintzman, P. D., Reyes, A. V., Zazula, G. D., Soares, A. E. R., Meyer, M., Hall, E., Jensen, B. J. L., Arnold, L. J., MacPhee, R. D. E., & Shapiro, B. (2017). Fossil and genomic evidence constrains the timing of bison arrival in North America. *Proceedings of the National Academy of Sciences*, *114*(13), 3457–3462. https://doi.org/10.1073/pnas.1620754114

Fukuda, M., Wakasugi, S., Tsuzuki, T., Nomiyama, H., Shimada, K., & Miyata, T. (1985). Mitochondrial DNA-like sequences in the human nuclear genome. *Journal of Molecular Biology*, *186*(2), 257–266. https://doi.org/10.1016/0022-2836(85)90102-0

G

Gamba, C., Jones, E. R., Teasdale, M. D., McLaughlin, R. L., Gonzalez-Fortes, G., Mattiangeli, V., Domboróczki, L., Kővári, I., Pap, I., Anders, A., Whittle, A., Dani, J., Raczky, P., Higham, T. F. G., Hofreiter, M., Bradley, D. G., & Pinhasi, R. (2014). Genome flux and stasis in a five millennium transect of European prehistory. *Nature Communications*, *5*(1), 5257. https://doi.org/10.1038/ncomms6257

Gansauge, M.-T., Gerber, T., Glocke, I., Korlević, P., Lippik, L., Nagel, S., Riehl, L. M., Schmidt, A., & Meyer, M. (2017). Single-stranded DNA library preparation from highly degraded DNA using *T4* DNA ligase. *Nucleic Acids Research*, gkx033. https://doi.org/10.1093/nar/gkx033

Gansauge, M.-T., & Meyer, M. (2013). Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. *Nature Protocols*, *8*(4), 737–748. https://doi.org/10.1038/nprot.2013.038

Gansauge, M.-T., & Meyer, M. (2014). Selective enrichment of damaged DNA molecules for ancient genome sequencing. *Genome Research*, *24*(9), 1543–1549. https://doi.org/10.1101/gr.174201.114

Gautier, M., Moazami-Goudarzi, K., Levéziel, H., Parinello, H., Grohs, C., Rialle, S., Kowalczyk, R., & Flori, L. (2016). Deciphering the Wisent Demographic and Adaptive Histories from Individual Whole-Genome Sequences. *Molecular Biology and Evolution*, *33*(11), 2801–2814. https://doi.org/10.1093/molbev/msw144

Geigl, E.-M. (2002). On the circumstances surrounding the preservation and analysis of very old DNA. *Archaeometry*, *44*(3), 337–342. https://doi.org/10.1111/1475-4754.t01-1-00066

Geigl, E.-M., & Grange, T. (2018). Ancient DNA: The quest for the best. *Molecular Ecology Resources*, *18*(6), 1185–1187. https://doi.org/10.1111/1755-0998.12931

Gelabert, P., Sawyer, S., Bergström, A., Margaryan, A., Collin, T. C., Meshveliani, T., Belfer-Cohen, A., Lordkipanidze, D., Jakeli, N., Matskevich, Z., Bar-Oz, G., Fernandes, D. M., Cheronet, O., Özdoğan, K. T., Oberreiter, V., Feeney, R. N. M., Stahlschmidt, M. C., Skoglund, P., & Pinhasi, R. (2021). Genome-scale sequencing and analysis of human, wolf, and bison DNA from 25,000-year-old sediment. *Current Biology*, *31*(16), 3564-3574.e9. https://doi.org/10.1016/j.cub.2021.06.023

Glocke, I., & Meyer, M. (2017). Extending the spectrum of DNA sequences retrieved from ancient bones and teeth. *Genome Research*, *27*(7), 1230–1237. https://doi.org/10.1101/gr.219675.116

Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E. M., Brockman, W., Fennell, T.,

Giannoukos, G., Fisher, S., Russ, C., Gabriel, S., Jaffe, D. B., Lander, E. S., & Nusbaum, C. (2009). Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature Biotechnology*, *27*(2), 182–189. https://doi.org/10.1038/nbt.1523

Gorgé, O., Bennett, E. A., Massilani, D., Daligault, J., Pruvost, M., Geigl, E.-M., & Grange, T. (2016). Analysis of Ancient DNA in Microbial Ecology. In F. Martin & S. Uroz (Eds.), *Microbial Environmental Genomics (MEG)* (Vol. 1399, pp. 289–315). Springer New York. https://doi.org/10.1007/978-1-4939-3369-3_17

Gotherstrom, A., Collins, M. J., Angerbjorn, A., & Liden, K. (2002). Bone preservation and DNA amplification. *Archaeometry*, *44*(3), 395–404. https://doi.org/10.1111/1475-4754.00072

Grange, T., Brugal, J.-P., Flori, L., Gautier, M., Uzunidis, A., & Geigl, E.-M. (2018). The Evolution and Population Diversity of Bison in Pleistocene and Holocene Eurasia: Sex Matters. *Diversity*, *10*(3), 65. https://doi.org/10.3390/d10030065

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H. Y., Hansen, N. F., Durand, E. Y., Malaspinas, A. S., Jensen, J. D., Marques-Bonet, T., Alkan, C., Prufer, K., Meyer, M., Burbano, H. A., … Paabo, S. (2010). A Draft Sequence of the Neandertal Genome. *Science*, *328*(5979), 710–722. https://doi.org/10.1126/science.1188021

Groeneveld. (2010). *Genetic diversity in farm animals*.

Gross, B. L., & Rieseberg, L. H. (2005). The Ecological Genetics of Homoploid Hybrid Speciation. *Journal of Heredity*, *96*(3), 241–252. https://doi.org/10.1093/jhered/esi026

Groves, C. P. (2009). Systematic relationships in the Bovini (Artiodactyla, Bovidae). *Journal of Zoological Systematics and Evolutionary Research*, *19*(4), 264–278. https://doi.org/10.1111/j.1439-0469.1981.tb00243.x

Guimaraes, S., Arbuckle, B. S., Peters, J., Adcock, S. E., Buitenhuis, H., Chazin, H., Manaseryan, N., Uerpmann, H.-P., Grange, T., & Geigl, E.-M. (2020). Ancient DNA shows domestic horses were introduced in the southern Caucasus and Anatolia during the Bronze Age. *Science Advances*, *6*(38), eabb0030. https://doi.org/10.1126/sciadv.abb0030

Guimaraes, S., Pruvost, M., Daligault, J., Stoetzel, E., Bennett, E. A., Côté, N. M.-L., Nicolas, V., Lalis, A., Denys, C., Geigl, E.-M., & Grange, T. (2017). A cost-effective high-throughput metabarcoding approach powerful enough to genotype ~44 000 year-old rodent remains from Northern Africa. *Molecular Ecology Resources*, *17*(3), 405–417. https://doi.org/10.1111/1755-0998.12565

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010). New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology*, *59*(3), 307–321. https://doi.org/10.1093/sysbio/syq010

Guintard, C. (1999). *On the size of the Ure-ox or Aurochs (Bos primigenius Bojanus)*.

Guintard,C & Neron de Surgy. (2014). *L'aurochs: De Lascaux au XXIe siècle*.

Günther, T., & Nettelblad, C. (2019). The presence and impact of reference bias on population genomic studies of prehistoric human populations. *PLOS Genetics*, *15*(7), e1008302. https://doi.org/10.1371/journal.pgen.1008302

Gurke, M., Vidal-Gorosquieta, A., Pajimans, J. L. A., Węcek, K., Barlow, A., González-Fortes, G., Hartmann, S., Grandal-d'Anglade, A., & Hofreiter, M. (2021). Insight into the introduction of domestic cattle and the process of Neolithization to the Spanish region Galicia by genetic evidence. *PLOS ONE*, *16*(4), e0249537. https://doi.org/10.1371/journal.pone.0249537

H

Hadad, R. (2018). Une illusion du vraisemblable. Mise en scène taphonomique et prospective néolithique à Çatalhöyük. *Gradhiva*, *28*, 112–141. https://doi.org/10.4000/gradhiva.3750

Hagelberg, E. (1991). *Isolation and characterization of DNA from archaeological bone*.

Hagelberg, E., Sykes, B., & Hedges, R. (1989). Ancient bone DNA amplified. *Nature*, *342*(6249), 485–485. https://doi.org/10.1038/342485a0

Haidt, A., Kamiński, T., Borowik, T., & Kowalczyk, R. (2018). Human and the beast—Flight and aggressive responses of European bison to human disturbance. *PLOS ONE*, *13*(8), e0200635. https://doi.org/10.1371/journal.pone.0200635

Hanchard, N. A., Rockett, K. A., Spencer, C., Coop, G., Pinder, M., Jallow, M., Kimber, M., McVean, G., Mott, R., & Kwiatkowski, D. P. (2006). Screening for Recently Selected Alleles by Analysis of Human Haplotype Similarity. *The American Journal of Human Genetics*, *78*(1), 153–159. https://doi.org/10.1086/499252

Hanchard, N., Elzein, A., Trafford, C., Rockett, K., Pinder, M., Jallow, M., Harding, R., Kwiatkowski, D., & McKenzie, C. (2007). Classical sickle beta-globin haplotypes exhibit a high degree of long-range haplotype similarity in African and Afro-Caribbean populations. *BMC Genetics*, *8*(1), 52. https://doi.org/10.1186/1471-2156-8-52

Hanotte, O. (2002). African Pastoralism: Genetic Imprints of Origins and Migrations. *Science*, *296*(5566), 336–339. https://doi.org/10.1126/science.1069878

Hao, Q., Wang, L., Oldfield, F., Peng, S., Qin, L., Song, Y., Xu, B., Qiao, Y., Bloemendal, J., & Guo, Z. (2012). Delayed build-up of Arctic ice sheets during 400,000-year minima in insolation variability. *Nature*, *490*(7420), 393–396. https://doi.org/10.1038/nature11493

Hartl, G. B., Göltenboth, R., Grilltsch, M., & Willing, R. (1988). On the biochemical systematics of the bovini. *Biochemical Systematics and Ecology*, *16*(6), 575–579. https://doi.org/10.1016/0305-1978(88)90065-8

Hassanin, A., An, J., Ropiquet, A., Nguyen, T. T., & Couloux, A. (2013). Combining multiple autosomal introns for studying shallow phylogeny and taxonomy of Laurasiatherian mammals: Application to the tribe Bovini (Cetartiodactyla, Bovidae). *Molecular Phylogenetics and Evolution*, *66*(3), 766–775. https://doi.org/10.1016/j.ympev.2012.11.003

Hassanin, A., & Ropiquet, A. (2004). Molecular phylogeny of the tribe Bovini (Bovidae, Bovinae) and the taxonomic status of the Kouprey, Bos sauveli Urbain 1937. *Molecular Phylogenetics and Evolution*, *33*(3), 896–907. https://doi.org/10.1016/j.ympev.2004.08.009

Hedges, S., & Schweitzer, M. (1995). Detecting dinosaur DNA. *Science*, *268*(5214), 1191–1192. https://doi.org/10.1126/science.7761839

Helmer, D. (1992). *La domestication des animaux par les hommes préhistoriques*. Masson.

Hemmer, H. (1983). *Domestikation: Verarmung der Merkwelt*. Vieweg.
Heyn, P., Stenzel, U., Briggs, A. W., Kircher, M., Hofreiter, M., & Meyer, M. (2010). Road blocks on paleogenomes—Polymerase extension profiling reveals the frequency of blocking lesions in ancient DNA. *Nucleic Acids Research*, *38*(16), e161–e161. https://doi.org/10.1093/nar/gkq572

Higuchi, R., Bowman, B., Freiberger, M., Ryder, O. A., & Wilson, A. C. (1984). DNA sequences from the quagga, an extinct member of the horse family. *Nature*, *312*(5991), 282–284. https://doi.org/10.1038/312282a0

Higuchi, R., Fockler, C., Dollinger, G., & Watson, R. (1993). Kinetic PCR Analysis: Real-time Monitoring of DNA Amplification Reactions. *Nature Biotechnology*, *11*(9), 1026–1030. https://doi.org/10.1038/nbt0993-1026

Ho, S. Y. W., & Gilbert, M. T. P. (2010). Ancient mitogenomics. *Mitochondrion*, *10*(1), 1–11. https://doi.org/10.1016/j.mito.2009.09.005

Hofreiter, M. (2001). *DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA.*

Horai, S., Hayasaka, K., Murayama, K., Wate, N., Koike, H., & Nakai, N. (1989). DNA amplification from ancient human skeletal remains and their sequence analysis. *Proceedings of the Japan Academy, Series B*, *65*(10), 229–233. https://doi.org/10.2183/pjab.65.229

Horn, S. (2012). Target enrichment via DNA hybridization capture. *Methods in Molecular Biology (Clifton, N.J.)*, *840*, 177–188. https://doi.org/10.1007/978-1-61779-516-9_21

Hudson, R. R., Kreitman, M., & Aguadé, M. (1987). A test of neutral molecular evolution based on nucleotide data. *Genetics*, *116*(1), 153–159.

Hummel, S. (1992). *Improved efficiency in amplification of ancient DNA and its sequence analysis*.

Hummel, S., Nordsiek, S., & Herrmann, B. (1992). Improved efficiency in amplification of ancient DNA and its sequence analysis. *Naturwissenschaften*, *79*(8), 359–360. https://doi.org/10.1007/BF01140179

Hwang, S., Kim, E., Lee, I., & Marcotte, E. M. (2015). Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Scientific Reports*, *5*(1), 17875. https://doi.org/10.1038/srep17875

J

Jensen, P. (2014). Behavior Genetics and the Domestication of Animals. *Annual Review of Animal Biosciences*, *2*(1), 85–104. https://doi.org/10.1146/annurev-animal-022513-114135

Jiang, Y., Bolnick, D. I., & Kirkpatrick, M. (2013). Assortative Mating in Animals. *The American Naturalist*, *181*(6), E125–E138. https://doi.org/10.1086/670160

Jill A. Weber. (2008). *Elite equids: Redefining equid burials of the mid- to late 3rd millennium BC from Umm el-Marra, Syria*.

Jordi Estevez & Maria Sana. (1999). *Auerochsenfunde auf der Iberischen Halbinsel*.

K

Kalmar, T. (2000). A simple and efficient method for PCR amplifiable DNA extraction from ancient bones. *Nucleic Acids Research*, *28*(12), 67e–667. https://doi.org/10.1093/nar/28.12.e67

Kaneda, H., Hayashi, J., Takahama, S., Taya, C., Lindahl, K. F., & Yonekawa, H. (1995). Elimination of paternal mitochondrial DNA in intraspecific crosses during early mouse embryogenesis. *Proceedings of the National Academy of Sciences*, *92*(10), 4542–4546. https://doi.org/10.1073/pnas.92.10.4542

Kapp, J. D., Green, R. E., & Shapiro, B. (2021). A fast and efficient single-stranded genomic library preparation method optimized for ancient DNA. *Journal of Heredity*, esab012. https://doi.org/10.1093/jhered/esab012

Kawasaki, T., Takahashi, S., & Ideda, K. (1985). Hydroxyapatite high-performance liquid chromatography: Column performance for proteins. *European Journal of Biochemistry*, *152*(2), 361–371. https://doi.org/10.1111/j.1432-1033.1985.tb09206.x

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P., & Drummond, A. (2012). Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, *28*(12), 1647–1649. https://doi.org/10.1093/bioinformatics/bts199

Keyser-Tracqui, C., Blandin-Frappin, P., Francfort, H.-P., Ricaut, F.-X., Lepetz, S., Crubézy, E., Samashev, Z., & Ludes, B. (2005). Mitochondrial DNA analysis of horses recovered from a frozen tomb (Berel site, Kazakhstan, 3rd Century BC). *Animal Genetics*, *36*(3), 203–209. https://doi.org/10.1111/j.1365-2052.2005.01316.x

Kim, Y., & Stephan, W. (2002). Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics*, *160*(2), 765–777.

Kimura, M. (1989). The neutral theory of molecular evolution and the world view of the neutralists. *Genome*, *31*(1), 24–31. https://doi.org/10.1139/g89-009

Kirillova, I. V., Zanina, O. G., Chernova, O. F., Lapteva, E. G., Trofimova, S. S., Lebedev, V. S., Tiunov, A. V., Soares, A. E. R., Shidlovskiy, F. K., & Shapiro, B. (2015). An ancient bison from the mouth of the Rauchua River (Chukotka, Russia). *Quaternary Research*, *84*(2), 232–245. https://doi.org/10.1016/j.yqres.2015.06.003

Kirillova, I. V., Zanina, O. G., Kosintsev, P. A., Kul'kova, M. A., Lapteva, E. G., Trofimova, S. S., Chernova, O. F., & Shidlovsky, F. K. (2013). The first finding of a frozen Holocene bison (Bison priscus Bojanus, 1827) carcass in Chukotka. *Doklady Biological Sciences*, *452*(1), 296–299. https://doi.org/10.1134/S0012496613050128

Kistler, L. (2012). Ancient DNA Extraction from Plants. In B. Shapiro & M. Hofreiter (Eds.), *Ancient DNA* (Vol. 840, pp. 71–79). Humana Press. https://doi.org/10.1007/978-1-61779-516-9_10

Korlević, P., Gerber, T., Gansauge, M.-T., Hajdinjak, M., Nagel, S., Aximu-Petri, A., & Meyer, M. (2015). Reducing microbial and human contamination in DNA extractions from ancient bones and teeth. *BioTechniques*, *59*(2), 87–93. https://doi.org/10.2144/000114320

Kowalski, K. (1967). *The evolution and fossil remains of the European bison*.

Krasińska, M., & Krasiński, Z. A. (2013). *European Bison*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-36555-3

Kuhlwilm, M., Gronau, I., Hubisz, M. J., de Filippo, C., Prado-Martinez, J., Kircher, M., Fu, Q., Burbano, H. A., Lalueza-Fox, C., de la Rasilla, M., Rosas, A., Rudan, P., Brajkovic, D., Kucan, Ž., Gušic, I., Marques-Bonet, T., Andrés, A. M., Viola, B., Pääbo, S., … Castellano, S. (2016). Ancient gene flow from early modern humans into Eastern Neanderthals. *Nature*, *530*(7591), 429–433. https://doi.org/10.1038/nature16544

Kwok, P. Y. (2000). High-throughput genotyping assay approaches. *Pharmacogenomics*, *1*(1), 95–100. https://doi.org/10.1517/14622416.1.1.95

Kwondo Kim & al. (2020). *The mosaic genome of indigenous African cattle as a unique genetic resource for African pastoralism*. https://doi.org/10.1038/s41588-020-0694

L

Lari, M., Rizzi, E., Mona, S., Corti, G., Catalano, G., Chen, K., Vernesi, C., Larson, G., Boscato, P., De Bellis, G., Cooper, A., Caramelli, D., & Bertorelle, G. (2011). The Complete Mitochondrial Genome of an 11,450-year-old Aurochsen (Bos primigenius) from Central Italy. *BMC Evolutionary Biology*, *11*(1), 32. https://doi.org/10.1186/1471-2148-11-32

Lazaridis, I., Nadel, D., Rollefson, G., Merrett, D. C., Rohland, N., Mallick, S., Fernandes, D., Novak, M., Gamarra, B., Sirak, K., Connell, S., Stewardson, K., Harney, E., Fu, Q., Gonzalez-Fortes, G., Jones, E. R., Roodenberg, S. A., Lengyel, G., Bocquentin, F., … Reich, D. (2016). Genomic insights into the origin of farming in the ancient Near East. *Nature*, *536*(7617), 419–424. https://doi.org/10.1038/nature19310

Lee, D. D., & Glimcher, M. J. (1989). The Three-Dimensional Spatial Relationship Between the Collagen Fibrils and the Inorganic Calcium-Phosphate Crystals of Pickerel and Herring Fish Bone. *Connective Tissue Research*, *21*(1–4), 247–257. https://doi.org/10.3109/03008208909050014

Lehmann, Ulrich. (1949). *Der Ur im Diluvium Deutschlands und seine Verbreitung. Auszug aus Band 90 aus der Reihe "Neues Jahrbuch für Mineralogie, Geologie und Paläontologie. Abhandlungen. Abteilung B"*.

Lemoine, F., Domelevo Entfellner, J.-B., Wilkinson, E., Correia, D., Dávila Felipe, M., De Oliveira, T., & Gascuel, O. (2018). Renewing Felsenstein's phylogenetic bootstrap in the era of big data. *Nature*, *556*(7702), 452–456. https://doi.org/10.1038/s41586-018-0043-0

Lenstra, J. A. (1999). *Systematics and phylogeny of cattle*.

Lenstra, J., Ajmone-Marsan, P., Beja-Pereira, A., Bollongino, R., Bradley, D., Colli, L., De Gaetano, A., Edwards, C., Felius, M., Ferretti, L., Ginja, C., Hristov, P., Kantanen, J., Lirón, J., Magee, D., Negrini, R., & Radoslavov, G. (2014a). Meta-Analysis of Mitochondrial DNA Reveals Several Population Bottlenecks during Worldwide Migrations of Cattle. *Diversity*, *6*(1), 178–187. https://doi.org/10.3390/d6010178

Lenstra, J., Ajmone-Marsan, P., Beja-Pereira, A., Bollongino, R., Bradley, D., Colli, L., De Gaetano, A., Edwards, C., Felius, M., Ferretti, L., Ginja, C., Hristov, P., Kantanen, J., Lirón, J., Magee, D., Negrini, R., & Radoslavov, G. (2014b). Meta-Analysis of Mitochondrial DNA Reveals Several Population Bottlenecks during Worldwide Migrations of Cattle. *Diversity*, *6*(1), 178–187. https://doi.org/10.3390/d6010178

Leonard, J. A., Shanks, O., Hofreiter, M., Kreuz, E., Hodges, L., Ream, W., Wayne, R. K., & Fleischer, R. C. (2007). Animal DNA in PCR reagents plagues ancient DNA research. *Journal of Archaeological Science*, *34*(9), 1361–1366. https://doi.org/10.1016/j.jas.2006.10.023

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Lindahl, T. (1993a). Instability and decay of the primary structure of DNA. *Nature*, *362*(6422), 709–715. https://doi.org/10.1038/362709a0

Lindahl, T. (1993b). Instability and decay of the primary structure of DNA. *Nature*, *362*(6422), 709–715. https://doi.org/10.1038/362709a0

Lindahl, T., & Nyberg, B. (1972). Rate of depurination of native deoxyribonucleic acid. *Biochemistry*, *11*(19), 3610–3618. https://doi.org/10.1021/bi00769a018

Lis, J. T., & Schleif, R. (1975). Size fractionation of double-stranded DNA by precipitation with polyethylene glycol. *Nucleic Acids Research*, *2*(3), 383–390. https://doi.org/10.1093/nar/2.3.383

Lisiecki, L. E., & Raymo, M. E. (2005). A Pliocene-Pleistocene stack of 57 globally distributed benthic δ^{18}O records: PLIOCENE-PLEISTOCENE BENTHIC STACK. *Paleoceanography*, *20*(1), n/a-n/a. https://doi.org/10.1029/2004PA001071

Loftus, R. T., MacHugh, D. E., Bradley, D. G., Sharp, P. M., & Cunningham, P. (1994). Evidence for two independent domestications of cattle. *Proceedings of the National Academy of Sciences*, *91*(7), 2757–2761. https://doi.org/10.1073/pnas.91.7.2757

Lord, K. A., Larson, G., Coppinger, R. P., & Karlsson, E. K. (2020). The History of Farm Foxes Undermines the Animal Domestication Syndrome. *Trends in Ecology & Evolution*, *35*(2), 125–136. https://doi.org/10.1016/j.tree.2019.10.011

Lorenzen, E. D., Nogués-Bravo, D., Orlando, L., Weinstock, J., Binladen, J., Marske, K. A., Ugan, A., Borregaard, M. K., Gilbert, M. T. P., Nielsen, R., Ho, S. Y. W., Goebel, T., Graf, K. E., Byers, D., Stenderup, J. T., Rasmussen, M., Campos, P. F., Leonard, J. A., Koepfli, K.-P., … Willerslev, E. (2011). Species-specific responses of Late Quaternary megafauna to climate and humans. *Nature*, *479*(7373), 359–364. https://doi.org/10.1038/nature10574

Lott, D. F. (2002). *American Bison: A Natural History*. University of California Press.

Lydekker, R. (1898). *Wild oxen, sheep & goats of all lands, living and extinct, by R. Lydekker.* R. Ward,. https://doi.org/10.5962/bhl.title.8851

M

Ma, H., & O'Farrell, P. H. (2015). Selections that isolate recombinant mitochondrial genomes in animals. *ELife*, *4*, e07247. https://doi.org/10.7554/eLife.07247

Madisson, W. (2006). *Madison WP, Knowles LL. Inferring phylogeny despite incomplete lineage sorting. Syst Biol 55: 21-30*.

Mannen, H., Kohno, M., Nagata, Y., Tsuji, S., Bradley, D. G., Yeo, J. S., Nyamsamba, D., Zagdsuren, Y., Yokohama, M., Nomura, K., & Amano, T. (2004). Independent mitochondrial origin and historical genetic differentiation in North Eastern Asian cattle. *Molecular Phylogenetics and Evolution*, *32*(2), 539–544. https://doi.org/10.1016/j.ympev.2004.01.010

Mannen, H., Yonezawa, T., Murata, K., Noda, A., Kawaguchi, F., Sasazaki, S., Olivieri, A., Achilli, A., & Torroni, A. (2020). Cattle mitogenome variation reveals a post-glacial expansion of haplogroup P and an early incorporation into northeast Asian domestic herds. *Scientific Reports*, *10*(1), 20842. https://doi.org/10.1038/s41598-020-78040-8

Maricic, T., Whitten, M., & Pääbo, S. (2010). Multiplexed DNA Sequence Capture of Mitochondrial Genomes Using PCR Products. *PLoS ONE*, *5*(11), e14004. https://doi.org/10.1371/journal.pone.0014004

Markova, A. K., Puzachenko, A. Yu., van Kolfschoten, T., Kosintsev, P. A., Kuznetsova, T. V., Tikhonov, A. N., Bachura, O. P., Ponomarev, D. V., van der Plicht, J., & Kuitems, M. (2015). Changes in the Eurasian distribution of the musk ox (Ovibos moschatus) and the extinct bison (Bison priscus) during the last 50 ka BP. *Quaternary International*, *378*, 99–110. https://doi.org/10.1016/j.quaint.2015.01.020

Martínez-Navarro, B., Rook, L., Papini, M., & Libsekal, Y. (2010). A new species of bull from the Early Pleistocene paleoanthropological site of Buia (Eritrea): Parallelism on the dispersal of the genus Bos and the Acheulian culture. *Quaternary International*, *212*(2), 169–175. https://doi.org/10.1016/j.quaint.2009.09.003

Massilani, D., Guimaraes, S., Brugal, J.-P., Bennett, E. A., Tokarska, M., Arbogast, R.-M., Baryshnikov, G., Boeskorov, G., Castel, J.-C., Davydov, S., Madelaine, S., Putelat, O., Spasskaya, N. N., Uerpmann, H.-P., Grange, T., & Geigl, E.-M. (2016). Past climate changes, population dynamics and the origin of Bison in Europe. *BMC Biology*, *14*(1), 93. https://doi.org/10.1186/s12915-016-0317-7

Maynard, J., & Haigh, J. (2007). The hitch-hiking effect of a favourable gene. *Genetics Research*, *89*(5–6), 391–403. https://doi.org/10.1017/S0016672308009579

Meyer, M. (1999). *Pattern of Nucleotide Substitution and Rate Heterogeneity in the Hypervariable Regions I and II of Human mtDNA*.

Meyer, M., & Kircher, M. (2010). Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing. *Cold Spring Harbor Protocols*, *2010*(6), pdb.prot5448-pdb.prot5448. https://doi.org/10.1101/pdb.prot5448

Meyer, M., Kircher, M., Gansauge, M.-T., Li, H., Racimo, F., Mallick, S., Schraiber, J. G., Jay, F., Prüfer, K., de Filippo, C., Sudmant, P. H., Alkan, C., Fu, Q., Do, R., Rohland, N., Tandon, A., Siebauer, M., Green, R. E., Bryc, K., … Pääbo, S. (2012). A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *Science*, *338*(6104), 222–226. https://doi.org/10.1126/science.1224344

Mikić, A. M. (2015). The First Attested Extraction of Ancient DNA in Legumes (Fabaceae). *Frontiers in Plant Science*, *6*. https://doi.org/10.3389/fpls.2015.01006

Mishmar, D., Ruiz-Pesini, E., Brandon, M., & Wallace, D. C. (2004). Mitochondrial DNA-like sequences in the nucleus (NUMTs): Insights into our African origins and the mechanism of foreign DNA integration. *Human Mutation*, *23*(2), 125–133. https://doi.org/10.1002/humu.10304

Moreira, D., & Philippe, H. (2000). Molecular phylogeny: Pitfalls and progress. *International Microbiology: The Official Journal of the Spanish Society for Microbiology*, *3*(1), 9–16.

Moritz, C., Dowling, T. E., & Brown, W. M. (1987). Evolution of Animal Mitochondrial DNA: Relevance for Population Biology and Systematics. *Annual Review of Ecology and Systematics*, *18*(1), 269–292. https://doi.org/10.1146/annurev.es.18.110187.001413

Mukasa-Mugerwa, E. (1989). *A review of reproductive performance of female Bos indicus (zebu) cattle*. International Livestock Centre for Africa.

N Saitou & M Nei. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*. https://doi.org/10.1093/oxfordjournals.molbev.a040454

N

Nagylaki, T. (1998). Fixation Indices in Subdivided Populations. *Genetics*, *148*(3), 1325–1332. https://doi.org/10.1093/genetics/148.3.1325

Naik, S. N. (1978). Origin and domestication of Zebu cattle (Bos indicus). *Journal of Human Evolution*, *7*(1), 23–30. https://doi.org/10.1016/S0047-2484(78)80032-3

Nei, M. (1996). PHYLOGENETIC ANALYSIS IN MOLECULAR EVOLUTIONARY GENETICS. *Annual Review of Genetics*, *30*(1), 371–403. https://doi.org/10.1146/annurev.genet.30.1.371

Nielsen, R. (2005). Molecular Signatures of Natural Selection. *Annual Review of Genetics*, *39*(1), 197–218. https://doi.org/10.1146/annurev.genet.39.073003.112420

Nomiyama, H., Fukuda, M., Wakasugi, S., Tsuzuki, T., & Shimada, K. (1985). Molecular structures of mitochondrial-DNA-like sequences in human nuclear DNA. *Nucleic Acids Research*, *13*(5), 1649–1658. https://doi.org/10.1093/nar/13.5.1649

Noonan, J. P. (2005). Genomic Sequencing of Pleistocene Cave Bears. *Science*, *309*(5734), 597–599. https://doi.org/10.1126/science.1113485

Nye, J., Laayouni, H., Kuhlwilm, M., Mondal, M., Marques-Bonet, T., & Bertranpetit, J. (2018). Selection in the Introgressed Regions of the Chimpanzee Genome. *Genome Biology and Evolution*, *10*(4), 1132–1138. https://doi.org/10.1093/gbe/evy077

O

O'Connor, T. P. (1997). Working at relationships: Another look at animal domestication. *Antiquity*, *71*(271), 149–156. https://doi.org/10.1017/S0003598X00084635

Oliva, A., Tobler, R., Cooper, A., Llamas, B., & Souilmi, Y. (2021). Systematic benchmark of ancient DNA read mapping. *Briefings in Bioinformatics*, bbab076. https://doi.org/10.1093/bib/bbab076

Orlando, L., Ginolhac, A., Zhang, G., Froese, D., Albrechtsen, A., Stiller, M., Schubert, M., Cappellini, E., Petersen, B., Moltke, I., Johnson, P. L. F., Fumagalli, M., Vilstrup, J. T., Raghavan, M., Korneliussen, T., Malaspinas, A.-S., Vogt, J., Szklarczyk, D., Kelstrup, C. D., … Willerslev, E. (2013). Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. *Nature*, *499*(7456), 74–78. https://doi.org/10.1038/nature12323

Ottoni, C., Van Neer, W., De Cupere, B., Daligault, J., Guimaraes, S., Peters, J., Spassov, N., Prendergast, M. E., Boivin, N., Morales-Muñiz, A., Bălăşescu, A., Becker, C., Benecke, N., Boroneant, A., Buitenhuis, H., Chahoud, J., Crowther, A., Llorente, L., Manaseryan, N., … Geigl, E.-M. (2017). The palaeogenetics of cat dispersal in the ancient world. *Nature Ecology & Evolution*, *1*(7), 0139. https://doi.org/10.1038/s41559-017-0139

P

Paabo, S. (1989). Ancient DNA: Extraction, characterization, molecular cloning, and enzymatic amplification. *Proceedings of the National Academy of Sciences*, *86*(6), 1939–1943. https://doi.org/10.1073/pnas.86.6.1939

Pacini. (1987). *Bison Schoetensacki Freud. From Isernia la Pineta (early Mid-Pleistocene—Italy) and Revision of the European Species of Bison*.

Paijmans, J. L. A., Fickel, J., Courtiol, A., Hofreiter, M., & Förster, D. W. (2016). Impact of enrichment conditions on cross-species capture of fresh and degraded DNA. *Molecular Ecology Resources*, *16*(1), 42–55. https://doi.org/10.1111/1755-0998.12420

Paijmans, J. L. A., Gilbert, M. T. P., & Hofreiter, M. (2013). Mitogenomic analyses from ancient DNA. *Molecular Phylogenetics and Evolution*, *69*(2), 404–416. https://doi.org/10.1016/j.ympev.2012.06.002

Palacio, P., Berthonaud, V., Guérin, C., Lambourdière, J., Maksud, F., Philippe, M., Plaire, D., Stafford, T., Marsolier-Kergoat, M.-C., & Elalouf, J.-M. (2017). Genome data on the extinct Bison schoetensacki establish it as a sister species of the extant European bison (Bison bonasus). *BMC Evolutionary Biology*, *17*(1), 48. https://doi.org/10.1186/s12862-017-0894-2

Park, S. D. E., Magee, D. A., McGettigan, P. A., Teasdale, M. D., Edwards, C. J., Lohan, A. J., Murphy, A., Braud, M., Donoghue, M. T., Liu, Y., Chamberlain, A. T., Rue-Albrecht, K., Schroeder, S., Spillane, C., Tai, S., Bradley, D. G., Sonstegard, T. S., Loftus, B. J., & MacHugh, D. E. (2015). Genome sequencing of the extinct Eurasian wild aurochs, Bos primigenius, illuminates the phylogeography and evolution of cattle. *Genome Biology*, *16*(1), 234. https://doi.org/10.1186/s13059-015-0790-2

Perkins, D. (1969). Fauna of Catal Huyuk: Evidence for Early Cattle Domestication in Anatolia. *Science*, *164*(3876), 177–179. https://doi.org/10.1126/science.164.3876.177

Perri, A. R., Feuerborn, T. R., Frantz, L. A. F., Larson, G., Malhi, R. S., Meltzer, D. J., & Witt, K. E. (2021). Dog domestication and the dual dispersal of people and dogs into the Americas. *Proceedings of the National Academy of Sciences*, *118*(6), e2010083118. https://doi.org/10.1073/pnas.2010083118

Peter, B. M., Huerta-Sanchez, E., & Nielsen, R. (2012). Distinguishing between Selective Sweeps from Standing Variation and from a De Novo Mutation. *PLoS Genetics*, *8*(10), e1003011. https://doi.org/10.1371/journal.pgen.1003011

Peters, J., McGlynn, G., & Goebel, V. (Eds.). (2019). *Animals: Cultural identifiers in ancient societies?* VML Verlag Marie Leidorf.

Peterson, D. G., Wessler, S. R., & Paterson, A. H. (2002). Efficient capture of unique sequences from eukaryotic genomes. *Trends in Genetics*, *18*(11), 547–550. https://doi.org/10.1016/S0168-9525(02)02764-6

Pieper, L. (2016). *Consumers' attitudes about milk quality and fertilization methods in dairy cows in Germany*.

Pilgrim, G. E. (1947). The Evolution of the Buffaloes, Oxen, Sheep and Goats. *Journal of the Linnean Society of London, Zoology*, *41*(279), 272–286. https://doi.org/10.1111/j.1096-3642.1940.tb02077.x

Pinhasi, R., Fernandes, D., Sirak, K., Novak, M., Connell, S., Alpaslan-Roodenberg, S., Gerritsen, F., Moiseyev, V., Gromov, A., Raczky, P., Anders, A., Pietrusewsky, M., Rollefson, G., Jovanovic, M., Trinhhoang, H., Bar-Oz, G., Oxenham, M., Matsumura, H., & Hofreiter, M. (2015). Optimal Ancient DNA Yields from the Inner Ear Part of the Human Petrous Bone. *PLOS ONE*, *10*(6), e0129102. https://doi.org/10.1371/journal.pone.0129102

Platt, D. E., Haber, M., Dagher-Kharrat, M. B., Douaihy, B., Khazen, G., Ashrafian Bonab, M., Salloum, A., Mouzaya, F., Luiselli, D., Tyler-Smith, C., Renfrew, C., Matisoo-Smith, E., & Zalloua, P. A. (2017). Mapping Post-Glacial expansions: The Peopling of Southwest Asia. *Scientific Reports*, *7*(1), 40338. https://doi.org/10.1038/srep40338

Poinar, H. N., Schwarz, C., Qi, J., Shapiro, B., MacPhee, R. D. E., Buigues, B., Tikhonov, A., Huson, D. H., Tomsho, L. P., Auch, A., Rampp, M., Miller, W., & Schuster, S. C. (2006). Metagenomics to Paleogenomics: Large-Scale Sequencing of Mammoth DNA. *Science*, *311*(5759), 392–394. https://doi.org/10.1126/science.1123360

Pollard, K. S., Salama, S. R., Lambert, N., Lambot, M.-A., Coppens, S., Pedersen, J. S., Katzman, S., King, B., Onodera, C., Siepel, A., Kern, A. D., Dehay, C., Igel, H., Ares, M., Vanderhaeghen, P., & Haussler, D. (2006). An RNA gene expressed during cortical development evolved rapidly in humans. *Nature*, *443*(7108), 167–172. https://doi.org/10.1038/nature05113

Prabhakar, S., Noonan, J. P., Paabo, S., & Rubin, E. M. (2006). Accelerated Evolution of Conserved Noncoding Sequences in Humans. *Science*, *314*(5800), 786–786. https://doi.org/10.1126/science.1130738

Price, E. O. (2002). *Animal domestication and behavior*. CABI Pub.

Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE*, *5*(3), e9490. https://doi.org/10.1371/journal.pone.0009490

Pritchard, J. K., Pickrell, J. K., & Coop, G. (2010). The Genetics of Human Adaptation: Hard Sweeps, Soft Sweeps, and Polygenic Adaptation. *Current Biology*, *20*(4), R208–R215. https://doi.org/10.1016/j.cub.2009.11.055

Pruvost, M. (2007). *Freshly excavated fossil bones are best for amplification of ancient DNA*.

Pruvost, M., & Geigl, E.-M. (2004). Real-time quantitative PCR to assess the authenticity of ancient DNA amplification. *Journal of Archaeological Science*, *31*(9), 1191–1197. https://doi.org/10.1016/j.jas.2002.05.002

Pruvost, M., Grange, T., & Geigl, E.-M. (2005). Minimizing DNA contamination by using UNG-coupled quantitative real-time PCR on degraded DNA samples: Application to ancient DNA studies. *BioTechniques*, *38*(4), 569–575. https://doi.org/10.2144/05384ST03

Pruvost, M., Schwarz, R., Bessa Correia, V., Champlot, S., Grange, T., & Geigl, E.-M. (2008). DNA diagenesis and palaeogenetic analysis: Critical assessment and methodological progress. *Palaeogeography, Palaeoclimatology, Palaeoecology*, *266*(3–4), 211–219. https://doi.org/10.1016/j.palaeo.2008.03.041

Pruvost, M., Schwarz, R., Correia, V. B., Champlot, S., Braguier, S., Morel, N., Fernandez-Jalvo, Y., Grange, T., & Geigl, E.-M. (2007). Freshly excavated fossil bones are best for amplification of ancient DNA. *Proceedings of the National Academy of Sciences*, *104*(3), 739–744. https://doi.org/10.1073/pnas.0610257104

Pucek, Z. (2004). *European bison: Status survey and conservation action plan*. IUCN - The world conservation union.

Pybus, O. G., & Rambaut, A. (2002). GENIE: Estimating demographic history from molecular phylogenies. *Bioinformatics*, *18*(10), 1404–1405. https://doi.org/10.1093/bioinformatics/18.10.1404

Q

Quiles, A., Valladas, H., Bocherens, H., Delqué-Količ, E., Kaltnecker, E., van der Plicht, J., Delannoy, J.-J., Feruglio, V., Fritz, C., Monney, J., Philippe, M., Tosello, G., Clottes, J., & Geneste, J.-M. (2016). A high-precision chronological model for the decorated Upper Paleolithic cave of Chauvet-Pont d'Arc, Ardèche, France. *Proceedings of the National Academy of Sciences*, *113*(17), 4670–4675. https://doi.org/10.1073/pnas.1523158113

R

Racimo, F., Woodbridge, J., Fyfe, R. M., Sikora, M., Sjögren, K.-G., Kristiansen, K., & Linden, M. V. (2019). *A geostatistical approach to modelling human Holocene migrations in Europe using ancient DNA* [Preprint]. Evolutionary Biology. https://doi.org/10.1101/826149

Rastogi, R. P., Richa, Kumar, A., Tyagi, M. B., & Sinha, R. P. (2010). Molecular Mechanisms of Ultraviolet Radiation-Induced DNA Damage and Repair. *Journal of Nucleic Acids*, *2010*, 1–32. https://doi.org/10.4061/2010/592980

Reich, D., Green, R. E., Kircher, M., Krause, J., Patterson, N., Durand, E. Y., Viola, B., Briggs, A. W., Stenzel, U., Johnson, P. L. F., Maricic, T., Good, J. M., Marques-Bonet, T., Alkan, C., Fu, Q., Mallick, S., Li, H., Meyer, M., Eichler, E. E., … Pääbo, S. (2010). Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*, *468*(7327), 1053–1060. https://doi.org/10.1038/nature09710

Renaud, G., Stenzel, U., & Kelso, J. (2014). leeHom: Adaptor trimming and merging for Illumina sequencing reads. *Nucleic Acids Research*, *42*(18), e141–e141. https://doi.org/10.1093/nar/gku699

Rheindt, F. E., & Edwards, S. V. (2011). Genetic Introgression: An Integral but neglected component of speciation in birds. *The Auk*, *128*(4), 620–632. https://doi.org/10.1525/auk.2011.128.4.620

Rogaev, E. I., Moliaka, Y. K., Malyarchuk, B. A., Kondrashov, F. A., Derenko, M. V., Chumakov, I., & Grigorenko, A. P. (2006). Complete Mitochondrial Genome and Phylogeny of Pleistocene MammothMammuthus primigenius. *PLoS Biology*, *4*(3), e73. https://doi.org/10.1371/journal.pbio.0040073

Rohland, N., & Hofreiter, M. (2007). Ancient DNA extraction from bones and teeth. *Nature Protocols*, *2*(7), 1756–1762. https://doi.org/10.1038/nprot.2007.247

Rojansky, R., Cha, M.-Y., & Chan, D. C. (2016). Elimination of paternal mitochondria in mouse embryos occurs through autophagic degradation dependent on PARKIN and MUL1. *ELife*, *5*, e17896. https://doi.org/10.7554/eLife.17896

Romanowski, G., Lorenz, M. G., & Wackernagel, W. (1991). Adsorption of plasmid DNA to mineral surfaces and protection against DNase I. *Applied and Environmental Microbiology*, *57*(4), 1057–1061. https://doi.org/10.1128/aem.57.4.1057-1061.1991

Romero, I. G., Ruvinsky, I., & Gilad, Y. (2012). Comparative studies of gene expression and the evolution of gene regulation. *Nature Reviews Genetics*, *13*(7), 505–516. https://doi.org/10.1038/nrg3229

Römpler, H., Dear, P. H., Krause, J., Meyer, M., Rohland, N., Schöneberg, T., Spriggs, H., Stiller, M., & Hofreiter, M. (2006). Multiplex amplification of ancient DNA. *Nature Protocols*, *1*(2), 720–728. https://doi.org/10.1038/nprot.2006.84

Ryan, W. B. F., Pitman, W. C., Major, C. O., Shimkus, K., Moskalenko, V., Jones, G. A., Dimitrov, P., Gorür, N., Sakinç, M., & Yüce, H. (1997). An abrupt drowning of the Black Sea shelf. *Marine Geology*, *138*(1–2), 119–126. https://doi.org/10.1016/S0025-3227(97)00007-8

Rychlik, W. (2007). OLIGO 7 Primer Analysis Software. In A. Yuryev (Ed.), *PCR Primer Design* (Vol. 402, pp. 35–59). Humana Press. https://doi.org/10.1007/978-1-59745-528-2_2

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z. P., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G. J., Ackerman, H. C., Campbell, S. J., Altshuler, D., Cooper, R., Kwiatkowski, D., Ward, R., & Lander, E. S. (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature*, *419*(6909), 832–837. https://doi.org/10.1038/nature01140

S

Salamon, M., Tuross, N., Arensburg, B., & Weiner, S. (2005). Relatively well preserved DNA is present in the crystal aggregates of fossil bones. *Proceedings of the National Academy of Sciences*, *102*(39), 13783–13788. https://doi.org/10.1073/pnas.0503718102

Sampietro, M. L., Gilbert, M. T. P., Lao, O., Caramelli, D., Lari, M., Bertranpetit, J., & Lalueza-Fox, C. (2006). Tracking down Human Contamination in Ancient Human Teeth. *Molecular Biology and Evolution*, *23*(9), 1801–1807. https://doi.org/10.1093/molbev/msl047

Sawyer, S., Krause, J., Guschanski, K., Savolainen, V., & Pääbo, S. (2012). Temporal Patterns of Nucleotide Misincorporations and DNA Fragmentation in Ancient DNA. *PLoS ONE*, *7*(3), e34131. https://doi.org/10.1371/journal.pone.0034131

Schaller, G. B., & Liu, W. (1996). Distribution, status, and conservation of wild yak Bos grunniens. *Biological Conservation*, *76*(1), 1–8. https://doi.org/10.1016/0006-3207(96)85972-6

Scheu, A., Powell, A., Bollongino, R., Vigne, J.-D., Tresset, A., Çakırlar, C., Benecke, N., & Burger, J. (2015). The genetic prehistory of domesticated cattle from their origin to the spread across Europe. *BMC Genetics*, *16*(1), 54. https://doi.org/10.1186/s12863-015-0203-2

Schmidt, K. (2007). *Sie bauten die ersten Tempel: Das rätselhafte Heiligtum der Steinzeitjäger die archäologische Entdeckung am Göbekli Tepe* (3. erw. und aktualisierte Aufl). C.H. Beck.

Smith, C. I., Chamberlain, A. T., Riley, M. S., Cooper, A., Stringer, C. B., & Collins, M. J. (2001). Not just old but old and cold? *Nature*, *410*(6830), 771–772. https://doi.org/10.1038/35071177
Sørensen & Bretlau. (1997). *Spatial organization of bone modelling and remodelling in the otic capsule*.

Soubrier, J., Gower, G., Chen, K., Richards, S. M., Llamas, B., Mitchell, K. J., Ho, S. Y. W., Kosintsev, P., Lee, M. S. Y., Baryshnikov, G., Bollongino, R., Bover, P., Burger, J., Chivall, D., Crégut-Bonnoure, E., Decker, J. E., Doronichev, V. B., Douka, K., Fordham, D. A., … Cooper, A. (2016). Early cave art and ancient DNA record the origin of European bison. *Nature Communications*, *7*(1), 13158. https://doi.org/10.1038/ncomms13158

Soubrier, J., Steel, M., Lee, M. S. Y., Der Sarkissian, C., Guindon, S., Ho, S. Y. W., & Cooper, A. (2012). The Influence of Rate Heterogeneity among Sites on the Time Dependence of Molecular Rates. *Molecular Biology and Evolution*, *29*(11), 3345–3358. https://doi.org/10.1093/molbev/mss140

Sousa, V., & Hey, J. (2013). Understanding the origin of species with genome-scale data: Modelling gene flow. *Nature Reviews Genetics*, *14*(6), 404–414. https://doi.org/10.1038/nrg3446
Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Stiller, M., Sucker, A., Griewank, K., Aust, D., Baretton, G. B., Schadendorf, D., & Horn, S. (2016). Single-strand DNA library preparation improves sequencing of formalin-fixed and paraffin-embedded (FFPE) cancer DNA. *Oncotarget*, *7*(37), 59115–59128. https://doi.org/10.18632/oncotarget.10827

Svensson, E., Günther, T., Hoischen, A., Hervella, M., Munters, A. R., Ioana, M., Ridiche, F., Edlund, H., van Deuren, R. C., Soficaru, A., de-la-Rua, C., Netea, M. G., & Jakobsson, M. (2021). Genome of Peştera Muierii skull shows high diversity and low mutational load in pre-glacial Europe. *Current Biology*, *31*(14), 2973-2983.e9. https://doi.org/10.1016/j.cub.2021.04.045

T

Tanaka, J. L. O., Medici Filho, E., Salgado, J. A. P., Salgado, M. A. C., Moraes, L. C. de, Moraes, M. E. L. de, & Castilho, J. C. de M. (2008). Comparative analysis of human and bovine teeth: Radiographic density. *Brazilian Oral Research*, *22*(4), 346–351. https://doi.org/10.1590/S1806-83242008000400011

Templeton, C. N. (2004). Multiple selection pressures influence Trinidadian guppy (Poecilia reticulata) antipredator behavior. *Behavioral Ecology*, *15*(4), 673–678. https://doi.org/10.1093/beheco/arh065

Terhorst, J., Kamm, J. A., & Song, Y. S. (2017). Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nature Genetics*, *49*(2), 303–309. https://doi.org/10.1038/ng.3748

Tiselius, A., Hjertén, S., & Levin, Ö. (1956). Protein chromatography on calcium phosphate columns. *Archives of Biochemistry and Biophysics*, *65*(1), 132–155. https://doi.org/10.1016/0003-9861(56)90183-7

Troy, C. S., MacHugh, D. E., Bailey, J. F., Magee, D. A., Loftus, R. T., Cunningham, P., Chamberlain, A. T., Sykes, B. C., & Bradley, D. G. (2001). Genetic evidence for Near-Eastern origins of European cattle. *Nature*, *410*(6832), 1088–1091. https://doi.org/10.1038/35074088

Trut, L., Oskina, I., & Kharlamova, A. (2009). Animal evolution during domestication: The domesticated fox as a model. *BioEssays*, *31*(3), 349–360. https://doi.org/10.1002/bies.200800070

Turelli, M. (1984). Heritable genetic variation via mutation-selection balance: Lerch's zeta meets the abdominal bristle. *Theoretical Population Biology*, *25*(2), 138–193. https://doi.org/10.1016/0040-5809(84)90017-0

V

van der Valk, T., Pečnerová, P., Díez-del-Molino, D., Bergström, A., Oppenheimer, J., Hartmann, S., Xenikoudakis, G., Thomas, J. A., Dehasque, M., Sağlıcan, E., Fidan, F. R., Barnes, I., Liu, S., Somel, M., Heintzman, P. D., Nikolskiy, P., Shapiro, B., Skoglund, P., Hofreiter, M., … Dalén, L. (2021). Million-year-old DNA sheds light on the genomic history of mammoths. *Nature*, *591*(7849), 265–269. https://doi.org/10.1038/s41586-021-03224-9

Verdugo, M. P., Mullin, V. E., Scheu, A., Mattiangeli, V., Daly, K. G., Maisano Delser, P., Hare, A. J., Burger, J., Collins, M. J., Kehati, R., Hesse, P., Fulton, D., Sauer, E. W., Mohaseb, F. A., Davoudi, H., Khazaeli, R., Lhuillier, J., Rapin, C., Ebrahimi, S., … Bradley, D. G. (2019). Ancient cattle genomics, origins, and rapid turnover in the Fertile Crescent. *Science (New York, N.Y.)*, *365*(6449), 173–176. https://doi.org/10.1126/science.aav1002

Verkaar, E. L. C., Nijman, I. J., Beeke, M., Hanekamp, E., & Lenstra, J. A. (2004). Maternal and Paternal Lineages in Cross-Breeding Bovine Species. Has Wisent a Hybrid Origin? *Molecular Biology and Evolution*, *21*(7), 1165–1170. https://doi.org/10.1093/molbev/msh064

Vigne, J. D. (2011). *Etat des connaissances archéozoologiques sur les débuts de l'élevage...*

Vigne, J.-D., Peters, J., Helmer, D., & International Council for Archaeozoology (Eds.). (2005). *The first steps of animal domestication: New archaeozoological approaches*. Oxbow.

Vuure, C. van. (2005). *Retracing the aurochs: History, morphology and ecology of an extinct wild ox*. Pensoft.

W

Wadman, M. (2008). James Watson's genome sequenced at high speed. *Nature*, *452*(7189), 788–788. https://doi.org/10.1038/452788b

Wakeley, J. (2009). *Coalescent theory: An introduction*. Roberts & Co. Publishers.

Wang, C., Schroeder, K. B., & Rosenberg, N. A. (2012). A Maximum-Likelihood Method to Correct for Allelic Dropout in Microsatellite Data with No Replicate Genotypes. *Genetics*, *192*(2), 651–669. https://doi.org/10.1534/genetics.112.139519

Węcek, K., Hartmann, S., Paijmans, J. L. A., Taron, U., Xenikoudakis, G., Cahill, J. A., Heintzman, P. D., Shapiro, B., Baryshnikov, G., Bunevich, A. N., Crees, J. J., Dobosz, R., Manaserian, N., Okarma, H., Tokarska, M., Turvey, S. T., Wójcik, J. M., Żyła, W., Szymura, J. M., … Barlow, A. (2016). Complex Admixture Preceded and Followed the Extinction of Wisent in the Wild. *Molecular Biology and Evolution*, msw254. https://doi.org/10.1093/molbev/msw254

Weninger, B., Schulting, R., Bradtmöller, M., Clare, L., Collard, M., Edinborough, K., Hilpert, J., Jöris, O., Niekus, M., Rohling, E. J., & Wagner, B. (2008). The catastrophic final flooding of Doggerland by the Storegga Slide tsunami. *Documenta Praehistorica*, *35*, 1–24. https://doi.org/10.4312/dp.35.1

Wilson, M. C., Hills, L. V., & Shapiro, B. (2008). Late Pleistocene northward-dispersing Bison antiquus from the Bighill Creek Formation, Gallelli Gravel Pit, Alberta, Canada, and the fate of Bison occidentalis. *Canadian Journal of Earth Sciences*, *45*(7), 827–859. https://doi.org/10.1139/E08-027

WJA Payne & J Hodges. (1997). *Tropical cattle: Origins, breeds and breeding policies*.
Woodward, Weyand, N., & Bunnell, M. (1994). DNA sequence from Cretaceous period bone fragments. *Science*, *266*(5188), 1229–1232. https://doi.org/10.1126/science.7973705

Wright, E. (2013). *The history of the European aurochs(Bos primigenius)from the Middle Pleistocene to its extinction: An archaeological investigation of its evolution, morphological variability and response to human exploitation*.

Wright, S. (1949). THE GENETICAL STRUCTURE OF POPULATIONS. *Annals of Eugenics*, *15*(1), 323–354. https://doi.org/10.1111/j.1469-1809.1949.tb02451.x

Wright, S. I., & Charlesworth, B. (2004). The HKA Test Revisited. *Genetics*, *168*(2), 1071–1076. https://doi.org/10.1534/genetics.104.026500

Wu, D.-D., Ding, X.-D., Wang, S., Wójcik, J. M., Zhang, Y., Tokarska, M., Li, Y., Wang, M.-S., Faruque, O., Nielsen, R., Zhang, Q., & Zhang, Y.-P. (2018). Pervasive introgression facilitated domestication and adaptation in the Bos species complex. *Nature Ecology & Evolution*, *2*(7), 1139–1145. https://doi.org/10.1038/s41559-018-0562-y

X

Xu, W., Lin, Y., Zhao, K., Li, H., Tian, Y., Ngatia, J. N., Ma, Y., Sahu, S. K., Guo, H., Guo, X., Xu, Y. C., Liu, H., Kristiansen, K., Lan, T., & Zhou, X. (2021). An efficient pipeline for ancient DNA mapping and recovery of endogenous ancient DNA from whole genome sequencing data. *Ecology and Evolution*, *11*(1), 390–401. https://doi.org/10.1002/ece3.7056

Y

Yang, Z., & Bielawski, J. P. (2000). Statistical methods for detecting molecular adaptation. *Trends in Ecology & Evolution*, *15*(12), 496–503. https://doi.org/10.1016/S0169-5347(00)01994-7

Zeder, M. A. (2008). Domestication and early agriculture in the Mediterranean Basin: Origins, diffusion, and impact. *Proceedings of the National Academy of Sciences*, *105*(33), 11597–11604. https://doi.org/10.1073/pnas.0801317105

Zeder, M. A. (2009). The Neolithic Macro-(R)evolution: Macroevolutionary Theory and the Study of Culture Change. *Journal of Archaeological Research*, *17*(1), 1–63. https://doi.org/10.1007/s10814-008-9025-3

Zeder, M. A. (2012). Pathways to Animal Domestication. In P. Gepts, T. R. Famula, R. L. Bettinger, S. B. Brush, A. B. Damania, P. E. McGuire, & C. O. Qualset (Eds.), *Biodiversity in Agriculture* (pp. 227–259). Cambridge University Press. https://doi.org/10.1017/CBO9781139019514.013

Zeyland, J., Wolko, Ł., Bocianowski, J., Szalata, M., Słomski, R., Dzieduszycki, A. M., Ryba, M., Przystałowska, H., & Lipiński, D. (2013). Complete mitochondrial genome of wild aurochs (Bos primigenius) reconstructed from ancient DNA. *Polish Journal of Veterinary Sciences*, *16*(2), 265–273. https://doi.org/10.2478/pjvs-2013-0037

Zeyland, J., Wolko, Ł., Lipiński, D., Woźniak, A., Nowak, A., Szalata, M., Bocianowski, J., & Słomski, R. (2012). Tracking of wisent–bison–yak mitochondrial evolution. *Journal of Applied Genetics*, *53*(3), 317–322. https://doi.org/10.1007/s13353-012-0090-4

Zhang, H., Paijmans, J. L. A., Chang, F., Wu, X., Chen, G., Lei, C., Yang, X., Wei, Z., Bradley, D. G., Orlando, L., O'Connor, T., & Hofreiter, M. (2013). Morphological and genetic evidence for early Holocene cattle management in northeastern China. *Nature Communications*, *4*(1), 2755. https://doi.org/10.1038/ncomms3755

Zischler, H., Geisert, H., von Haeseler, A., & Pääbo, S. (1995). A nuclear "fossil" of the mitochondrial D-loop and the origin of modern humans. *Nature*, *378*(6556), 489–492. https://doi.org/10.1038/378489a0

Abréviations

ADN : Acide Désoxyribonucléique
ADNa : ADN ancien
ADNmt : ADN mitochondrial
aMPlex : PCR multiplexe associé aux séquençage de nouvelle génération
ARN : Acide ribonucléique
AP : Avant le présent
A.n.ère : Avant notre ère
ASO : Asie du Sud-Ouest
dATP : Désoxyadénine triphosphate
dCTP : Désoxycytosine triphosphate
dCTP : Désoxyguanine triphosphate
dTTP : Désoxythymine triphosphate
dNTPs : Mélange des quatre nucléotides
EMA : Bromure d'éthidium monoazide
2M70 : Protocole de purification des banques d'ADN (2M70) contenant 2M Guanidine, 70% Isopropanol)
5M40 : Protocole de purification des banques d'ADN (5M40) contenant 5M Guanidine, 40% Isopropanol)
HVR : Région hypervariable
LGM : Last Glacial Maximum (dernier maximum glaciaire)
MRCA : Lost Recent Common Ancestor (Ancêtre commun le plus proche)
Mitogénome : Génome mitochondrial
min : minute
pb : paire de base
PCR : Polymerase Chain Reaction ou Réaction de Polymérisation en Chaine
PCRq : PCR quantitative
QG : Protocole standard de purification des banques d'ADN de Qiagen contenant 5.5 M guanidine thiocyanateet 25% d'isopropanol.
s : seconde
SNP : Single Nucleotide Polymorphism (polymorphisme nucléotidique d'une base)
UNG : Uracile-N-Glycosylase
USER : Mélange d'Uracile-N-Glycosylase et Endonucléase VIII