Génomique des populations intégrative : de la phylogénie à la génétique des populations

M Bastian

THÈSE DE DOCTORAT DE L'UNIVERSITÉ DE LYON

opérée au sein de
l'Université Claude Bernard Lyon 1

École doctorale n°341
Évolution, Écosystèmes, Microbiologie, Modélisation

Discipline : Évolution moléculaire

Soutenue publiquement le 16/12/2024 par

Mélodie Bastian

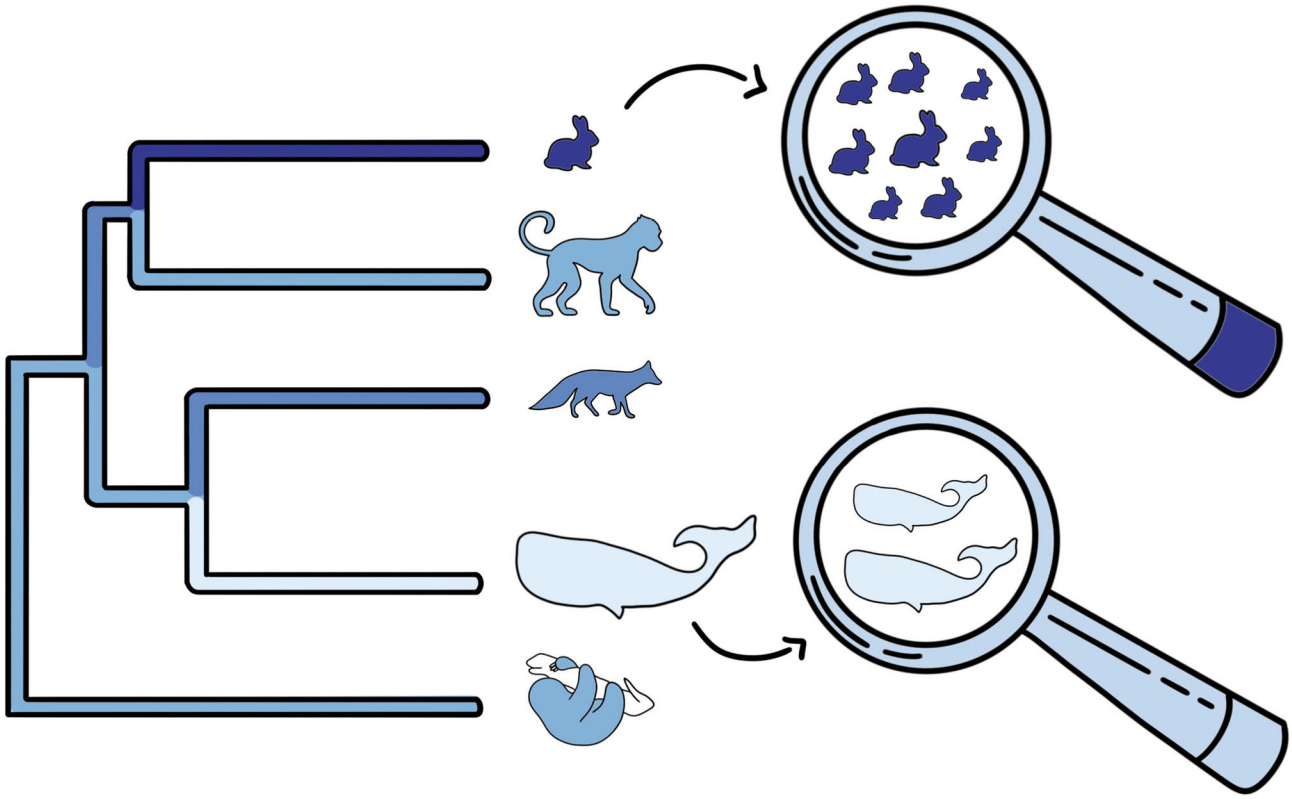
Génomique des populations intégrative : de la phylogénie à la génétique des populations

Devant le jury composé de :

Dr. Guillaume ACHAZ, Pr, CIRB, College de France, France	Rapporteur - **Président**
Dr. Maria ANISIMOVA, Pr, ICLS, Zurich University of Applied Science, Suisse	Rapportrice
Dr. Thomas BATAILLON, Pr, BIRC, Aarhus University, Danemark	Rapporteur
Dr. Marie FABLET, MCU, LBBE, Université Lyon 1, France	Examinatrice
Dr. Christelle FRAÏSSE, CR CNRS, Evo-Eco-Paleo, Université de Lille France	Examinatrice
Dr. Nicolas LARTILLOT, DR, LBBE, CNRS - Université Lyon 1, France	Directeur de thèse

Génomique des populations intégrative

De la phylogénie à la génétique des populations



Mélodie Bastian

Sous la direction de Nicolas Lartillot

Résumé

L'étude de l'évolution des génomes s'exerce à différentes échelles de temps évolutifs. Une première échelle, dite micro-évolutive (du domaine de la génétique des populations), se concentre sur les changements génétiques entre individus d'une même population. Si on étend cette échelle sur un temps plus long, on aboutit à une deuxième échelle temporelle dite macro-évolutive (du domaine de la phylogénie), basée sur l'étude des changements génétiques entre individus d'espèces différentes. Les génomes, notamment les séquences codantes, évoluent sous l'action combinée de plusieurs processus tels que la mutation, la sélection naturelle et la dérive génétique. Cette dernière, inversement proportionnelle à la taille efficace d'une population (N_e), est un processus évolutif stochastique particulièrement important, car il limite l'efficacité de la sélection naturelle. Dans un contexte où la majorité des mutations sont délétères, avec des effets allant de létal à faibles, la sélection naturelle purifiante va éliminer une mutation délétère d'autant plus vite qu'elle est impactante. Selon la théorie quasi-neutre formulée par Ohta, plus N_e est petit et plus la dérive génétique conduit rapidement à la fixation aléatoire de mutations délétères, empêchant la sélection de purifier les mutations à faible effet. Ainsi, l'efficacité de la sélection dans une population serait proportionnelle à N_e. Cependant, cette théorie manque encore de preuves empiriques décisives sous la forme d'une étude complète qui articule de manière cohérente les échelles micro et macro-évolutives sur un jeu de données suffisamment riche à la fois en gènes et en espèces.

Lors de mon premier travail de thèse, j'ai constitué un jeu de données contenant environ 150 génomes de mammifères. J'ai annoté chacun de ces génomes afin de construire un alignement de plus de 6000 gènes orthologues, globalement partagés à l'échelle du clade. Par ailleurs, puisque les individus séquencés sont diploïdes, j'ai estimé leur hétérozygotie sur ces mêmes gènes afin d'obtenir une estimation du polymorphisme de leurs populations. Dans mon deuxième travail de thèse, j'ai étudié la validité de l'approximation du polymorphisme par l'hétérozygotie d'un unique individu et j'ai notamment montré que cette approche est valable, bien qu'elle introduise une variance supplémentaire dans l'estimation du polymorphisme.

En associant les alignements de gènes inter-spécifiques à l'hétérozygotie intra-individu, j'ai pu estimer l'intensité de la sélection à la fois aux échelles macro-évolutive (d_N/d_S) et micro-évolutive (π_N/π_S). J'ai également intégré des estimateurs de la taille efficace (traits d'histoire de vie et π_S) pour chaque échelle. Le jeu de données a été analysé par une méthode intégrative bayésienne (FastCoevol) permettant d'étudier conjointement les relations entre chaque trait étudié, le long de la phylogénie. Une attention particulière a été portée sur la qualité des données et sur la reproductibilité de l'analyse.

Mon travail de thèse a permis de confirmer la corrélation positive entre d_N/d_S et traits d'histoire de vie, déjà connue par ailleurs, et a également permis de montrer une corrélation négative entre π_N/π_S et π_S, un résultat inédit dans le cas des génomes nucléaires de mammifères. Prises ensemble, ces deux observations valident la relation entre intensité de la sélection et N_e séparément pour les échelles de temps micro- et macro-évolutives, chez les mammifères. Aussi, le croisement des échelles montre, bien que de façon plus faible, une concordance entre micro- et macro-évolution sur les attendus quasi-neutres. L'ensemble de ces résultats fournissent une validation empirique de la théorie quasi-neutre à l'échelle des mammifères et permettent d'envisager une reconstruction quantitative des variations de N_e aux différentes échelles ainsi que de mesurer les contributions respectives des variations de court et long terme aux changements de N_e.

Abstract

The study of genome evolution is carried out on different evolutionary timescales. A first scale, called micro-evolution (in the field of population genetics), focuses on genetic changes between individuals in the same population. If we extend this scale over a longer period of time, we arrive at a second time scale known as macro-evolution (in the field of phylogeny), based on the study of genetic changes between individuals of different species.

Genomes, especially coding sequences, evolve under the combined action of several processes such as mutation, natural selection and genetic drift. This last, inversely proportional to the effective population size (N_e), is a particularly important stochastic evolutionary process, as it limits the efficacy of purifying selection against deleterious mutations. In a context where the majority of mutations are deleterious, with effects ranging from lethal to weak, purifying natural selection eliminates a deleterious mutation all the faster the greater its impact. According to Ohta's quasi-neutral theory, the smaller N_e is, the faster genetic drift leads to the random fixation of deleterious mutations, preventing selection from purifying low-effect mutations. Thus, the efficiency of selection in a population would be proportional to N_e. However, this theory still lacks decisive empirical proof in the form of a comprehensive study that coherently articulates micro and macro-evolutionary scales on a sufficiently rich dataset of both genes and species.

In my first thesis work, I built up a dataset containing around 150 mammalian genomes. I annotated each of these genomes to construct an alignment of over 6,000 orthologous genes, globally shared at clade level. In addition, since the individuals sequenced are diploid, I estimated their heterozygosity on these same genes to obtain an estimate of the polymorphism of their populations. In my second thesis project, I investigated the validity of approximating polymorphism by the heterozygosity of a single individual, and in particular showed that this approach is valid, although it introduces additional variance into the polymorphism estimate.

By associating inter-specific gene alignments with intra-individual heterozygosity, I estimate the intensity of selection at both macro-evolutionary (d_N/d_S) and micro-evolutionary (π_N/π_S) scales. I also integrated effective size estimators (life history traits and π_S) for each scale. The dataset was analysed using an integrative Bayesian method (FastCoevol) to jointly study the relationships between each trait studied, along the phylogeny. Particular attention was paid to data quality and the reproducibility of the analysis.

My thesis work confirmed the positive correlation between d_N/d_S and life-history traits, already known elsewhere, and also showed a negative correlation between π_N/π_S and π_S, a novel result in the case of mammalian nuclear genomes. Taken together, these two observations validate the relationship between selection intensity and N_e separately for micro- and macro-evolutionary time scales in mammals. Also, crossing the scales shows, albeit to a smaller extent, a concordance between micro- and macro-evolution on quasi-neutral expectations. Taken together, these results provide empirical validation of the quasi-neutral theory at the mammalian scale, and make it possible to envisage a quantitative reconstruction of N_e variations at different scales, as well as to measure the respective contributions of short- and long-term variations to changes in N_e.

Table des matières

Table des figures

Liste des boxes

Glossaire

Acides aminés Éléments composant les protéines. Il existe 20 acides aminés différents présents naturellement dans les organismes. (Voir Box 1 et figure 1.2). 20, 22, 39, X, XIV

ADN Acide DésoxyriboNucléique, assemblage de nucléotides disposés en deux brins antiparallèles enroulés l'un autour de l'autre pour former une double hélice. L'ADN porte l'information génétique nécessaire au développement, au fonctionnement et à la reproduction des organismes vivants. (Voir Box 1). 14, 19, 22, 29, 30, 35, 57, 61, IX–XI, XIII, XIV

Allèle Un variant possible d'un gène. 23, 25, 50, 71, 82, 83, 86, XI–XIII

ARN Acide RiboNucléique, très proche chimiquement de l'ADN. C'est une molécule à un brin issue de la transcription d'un gène porté par l'ADN. Parmi les différents types d'ARN on retouve l'ARN messager (ARNm), qui transporte l'information génétique de l'ADN vers les ribosomes, l'ARN de transfert (ARNt), qui amène les acides aminés pour former les protéines, et l'ARN ribosomal (ARNr), qui participe à la structure des ribosomes et catalyse la formation des liaisons peptidiques. (Voir Box 1). 14, 30, 39, 57, XIII, XIV

Biométrie Étude par la mesure des caractéristiques d'un individu ou d'une espèce. 16, 18

Branche (d'un arbre phylogénétique) Représentation schématique d'une relation entre deux unités évolutives dans une phylogénie. La longueur de la branche peut être informative et représenter la quantité d'évolution entre les deux unités reliées. (Voir Box.6). 74, 78, 79, 97, 179, X, XIII

Cadre de lecture Façon dont les nucléotides sont regroupés en triplets pour former des codons. Un décalage du cadre de lecture, dû par exemple à une insertion d'un nucléotide, peut entraîner une traduction incorrecte de la protéine. 67, 194, 195

Catastrophisme Théorie selon laquelle les observations géologiques sont compatibles avec les écrits bibliques. On fait souvent référence à Cuvier qui pensait que les différentes couches géologiques correspondent à différentes catastrophes. 10

Chromosome Séquence d'ADN repliée en 3D, contenant les gènes. Chez un individu diploïde, il y a 2 exemplaires de chaque chromosome, chacun issu d'un des deux parents. Chez l'humain, il y a 23 paires de chromosomes, dont une paire de chromosomes sexuels. 18, 19, 61, 71

Cladogramme Représentation phylogénétique dans laquelle les longueurs de branches sont de taille égale et ne sont pas informatives, contrairement à un phylogramme. (Voir Box.6). 74, XIII

Cluster de calcul Ensemble d'ordinateurs interconnectés qui travaillent ensemble pour exécuter des tâches complexes, offrant une puissance de calcul plus élevée que celle d'une machine individuelle. 68, 72

Code génétique Table de lecture qui permet de traduire une séquence nucléotidique en séquence protéique en associant des triplets de nucléotides à un acide aminé spécifique. (Voir figure 1.2). 20, 22, 30, 31, 70, 76

Codon Une suite de trois nucléotides dans une séquence génique. Il existe 64 codons possibles qui se traduisent en 20 acides aminés différents. Plusieurs codons peuvent donner le même acide aminé. On les appelle des codons synonymes. Certains codons signalisent la fin d'un gène, ce sont des codons stop. (Voir Box 1 et figure 1.2). 20, 36, 67, 76, IX

Coefficient de sélection Paramètre mesurant l'impact d'une mutation sur la fitness de l'individu qui la porte. Plus une mutation a un effet fort et plus son coefficient de sélection, en valeur absolue, est élevé. S'il est positif, la mutation est avantageuse, s'il est négatif, la mutation est délétère. En général, le coefficient de sélection est noté «s». 41, 45, 185, X

Couverture du génome ou d'un site La couverture est le nombre de fois qu'une position du génome est lue lors du séquençage. Plus une position est lue, plus la couverture est élevée et la position fiable. La couverture globale du génome est la moyenne des couvertures de toutes les positions. 62, 64, 71, 122, 188

Créationnisme Théorie selon laquelle l'univers, la Terre et toutes les espèces vivantes ont été créés par une entité divine. 10, 15

DFE (*Distribution of Fitness Effect*) Distribution des Effets sur la Fitness, représentation graphique qui montre, en ordonnée, le nombre de sites dans un génome ayant un certain coefficient de sélection, indiqué en abscisse. 44, 94, 110, 115, 184, 193

Diploïde Se dit d'un organisme qui possède deux jeux complets de chromosomes, soit deux copies de chaque chromosome, une héritée de chacun des parents. 18, 28, 34, 62, 71, 83, 88, 91, 117, 118, 121, 153, 154, X, XII

Divergence génétique Quantité de différence génétique accumulée entre deux individus d'espèces différentes. Le temps de divergence correspond au temps depuis la spéciation. 31, 32, 34, 36, 49, 91, 106, 121, 192

Droite de régression linéaire Droite ajustée pour représenter au mieux un nuage de points. Sa pente permet d'évaluer le lien de causalité entre deux variables. 18, 97

Dérive génétique Processus d'évolution aléatoire de la fréquence d'un allèle dans une population. Plus une population est petite et plus la dérive génétique est intense et peut engendrer la perte ou la fixation d'un allèle indépendamment de son effet. 24, 25, 28, 31, 33, 57, 76, 82, 87, 98, 106, 122, 176, 185, 204

Enzyme Protéine qui a pour fonction de faciliter les réactions chimiques dans les cellules. Exemple : la lactase aide à digérer le lactose. 19, 22, 28, 60, XIV

Exon Séquence de nucléotides au sein d'un gène condant généralement une partie de la protéine. Contrairement aux introns, les exons sont, en général, conservés et traduits en séquence d'acides aminés lors de l'expression du gène. (Voir Box 1). 22, 67, XI

Feuille (d'un arbre phylogénétique) Élément d'une phylogénie qui représente les gènes ou les espèces étudiées. (Voir Box.6). 74, 79, 97, 99, 179, XIV

Fitness ou valeur sélective Concept central en biologie évolutive et écologie qui mesure la capacité d'un individu à se reproduire. 44, 184, X, XIII, XV

Fixisme Théorie selon laquelle les espèces n'évoluent pas dans le temps. Cette notion se place en opposition au transformisme. 10, 15, XVI

gBGC (= Conversion Génique Biaisée vers GC) Processus évolutif qui favorise l'utilisation d'un nucléotide G ou C plutôt que A ou T en cas de mésappariement lors de la recombinaison. 188, 191, 192

Gène Séquence d'ADN contenant une information génétique particulière qui code pour une protéine. La plupart des gènes eucaryotes contiennent des exons et des introns. (Voir Box 1). 18, 19, 22, 25, 30, 39, 106, IX, X, XII, XIV

Génome Ensemble de l'information génétique portée par un organisme. 35, 44, 58, 59, 64, 77, 83, 89, 121, 153, 176, 188

Génotype Correspond à l'ensemble des gènes portés par un individu avec toutes les variations alléliques qui lui sont propres. 18, 19, 23, 93, 184, XIV

Généalogie Représentation des relations de parenté entre différents membres d'une famille, ou ici plus souvent, entre différents membres d'une population. Une généalogie permet d'identifier les liens de filitation entre les individus (voir phylogénie). 13, 82, VII, XIV

Haplotype Groupe d'allèles situés sur un même chromosome et généralement hérités ensemble. Les haplotypes permettent d'identifier des régions du génome qui sont transmises de manière conservée à travers les générations, et sont souvent utilisés pour étudier les relations de parenté ou la diversité génétique au sein des populations. 83

Haploïde Se dit d'un organisme qui possède un unique jeu complet de chromosome, contrairement à un diploïde qui en possède deux. 62

Homologie ou Homologue Correspond à des éléments qui ont une histoire évolutive commune. L'homologie rend les éléments comparables entre eux. 10, 13, 30, 57, 67, 75, 84, XIII

Hybride Organisme issu du croisement entre deux individus d'espèces différentes. L'hybride possède des caractéristiques génétiques mixtes appartenant aux deux parents. 17

Hérédité Transmission d'un caractère d'un parent à sa descendance. 16, 17, 19, 24

Hétérozygotie Quantité de différence génétique entre les deux chromosomes homologues d'un individu diploïde. Dans ce manuscrit et notamment dans le chapitre 5, l'hétérozygotie est utilisée comme proxy du polymorphisme. 62, 71, 84, 88, 111, 121, 123, 154, 176

Indel INsertion ou DELétion d'une séquence nucléotidique dans un gène. 67

Intron Séquence de nucléotides dans un gène qui n'est pas traduite en une partie de la protéine. (Voir Box 1). 22, 36, XI

Monophylétique Se dit d'un groupe d'espèces qui contient toutes les espèces descendant d'un ancêtre commun. Exemple : les mammifères sont monophylétiques. 13

Mutation (avantageuse, délétère, fixée, neutre, non codante, synonyme) Changement dans la séquence ADN. Elle peut être provoquée notamment par des agents mutagènes (rayons X, etc.) et des erreurs de copie ou de réparation de l'ADN. Une mutation **synonyme** ne change pas la séquence de la protéine. Une mutation **délétère** diminue la fitness de l'individu tandis qu'une mutation **avantageuse** l'augmente. Une mutation **non-codante** correspond à une mutation qui s'est produite dans une région génomique qui ne contient pas de gène. Une mutation **neutre** est une mutation qui n'affecte pas la fitness de son porteur. Une mutation **fixée** est une mutation qui a totalement remplacé l'allèle initial (sa fréquence est égale à 1). 22, 23, 32, 58, 67, 77, 82, 92, 106, 176, 184, 192, XV

Nucléotide Molécule composant l'ADN et l'ARN. (Voir Box 1). 19, 22, 32, 36, 75, 192, IX–XIV

Néo-Darwinisme ou théorie synthétique de l'évolution Théorie des années 1940-60 qui intègre la théorie de l'hérédité mendélienne, la génétique des populations et la théorie darwinienne. 25, 28, 40, XV

Nœud (d'un arbre phylogénétique) Élément d'une phylogénie qui représente une unité taxonomique. (Voir Box.6). 74, 98

Orthologue Deux éléments (par exemple un gène) qui partagent un même ancêtre commun (homologue) et qui se sont différenciés par une spéciation de cet ancêtre en deux nouvelles espèces portant chacune l'élément. 57, 59, 62, 64, 65, 67, 70, 72, 101, 107, 117, 122, 176, 188, 190

Paire de base (pb) Unité de mesure qui désigne deux nucléotides complémentaires situés sur des brins opposés d'une molécule d'ADN. Par exemple, une séquence de 30 pb correspond à 60 nucléotides au total. 60, 67, 75

Paralogue Deux éléments (par exemple un gène) qui partagent un même ancêtre commun (homologue) et qui se sont différenciés par une duplication de l'élément dans le génome de l'ancêtre. 65, 190

Phylogramme Représentation phylogénétique dans laquelle les longueurs de branche sont informatives, contrairement à un cladogramme. Voir Box.6. 74, X

Phylogénie Graphique sous la forme d'un arbre qui représente les relations évolutives entre taxons. Une phylogénie permet d'identifier quelles espèces sont

évolutivement proches. (voir généalogie). 13, 29, 39, 59, 74, VII, IX, XI–XV

Phénotype Correspond à l'ensemble des traits observables d'un individu. C'est la réalisation du génotype dans un contexte environnemental particulier. 18, 19, 26, 28, 37, 93, 191

Polymorphisme Quantité de différence génétique accumulée entre deux individus d'une même population ou espèce. 29, 31, 32, 34, 36, 59, 70, 91, 110, 121, 176, 178, 192, XII

Polymérase Enzyme qui permet la synthèse de polymères (comme la molécule d'ADN), à partir de monomères (comme les nucléotides). Dans le cas de l'ADN, la polymérase assemble les nucléotides en suivant le modèle du brin d'ADN existant, permettant ainsi la réplication de l'ADN. 60, 195

Population Groupe d'individus d'une même espèce qui vivent dans une région géographique donnée, interagissant entre eux et partagant un pool génétique commun. 23, 33, 41, 44, 47, 57, 62, 81, 106, 153, XI, XII

Protéine Produit de la transcription d'un gène en ARN, suivi de la traduction de cet ARN en une chaine d'acides aminés. (Voir Box 1). 22, 29, 30, 36, 39, 57, IX, XI, XII

Racine (d'un arbre phylogénétique) Élément basal d'une phylogénie. La racine représente l'ancêtre commun de toutes les feuilles de l'arbre. Elle donne un sens/une direction à l'évolution et permet de déterminer quel nœud est relativement plus ancien qu'un autre. (Voir Box.6). 74, 100

Recombinaison Processus biologique de réarrangement entre deux séquences d'ADN, créant de nouvelles combinaisons de gènes. Le taux de recombinaison est variable entre espèces et le long des génomes. 52, 83, 86, 176, 191

Saturation Phénomène qui se produit lorsqu'un système atteint sa capacité maximale pour réaliser une certaine fonction, entraînant une stabilisation ou une diminution de la réponse malgré une augmentation des données d'entrée. Lors d'une analyse phylogénétique, il y a saturation quand trop de substitutions se sont accumulées, menant au masquage de certaines d'entre elles par d'autres plus récentes. 31, 39, 75, 81

SFS (*Site Frequency Spectrum*) Spectre de fréquence allélique, représentation graphique permettant de visualiser en ordonnée le nombre de sites présents

dans une population en fonction de la fréquence donnée en abscisse. 86, 90, 94, 95

Signal phylogénétique Correspond à la ressemblance phénotypique ou moléculaire entre deux espèces, due à leur histoire évolutive commune. C'est une mesure de la force de la corrélation entre la parenté évolutive et la distribution de traits biologiques dans un groupe d'espèces. Plus le signal est fort et plus les espèces se ressemblent en raison de leur héritage partagé. Peut aussi être défini comme une mesure de la qualité de l'information pour reconstruire une phylogénie. 58, 69, 75, 77, 96, 187

Spéciation Processus évolutif engendrant l'apparition de deux nouvelles espèces à partir d'une autre qui devient une espèce ancestrale. 25, 79, 99, XI

Substitution Correspond à une mutation fixée. 29, 31, 33, 58, 67, 75, 76, 106, 192, XV

Systématique Domaine de la biologie qui traite de la classification et de la dénomination des organismes vivants. 9, 13

Sélection balancée Type de sélection naturelle où plusieurs allèles sont maintenus dans une population à des fréquences intermédiaires, car chacun confère un avantage dans certaines conditions. 90–92

Sélection positive Type de sélection qui favorise la fixation de mutations qui augmentent la fitness de l'individu. 41, 77, 90, 91, 93, 94, 116, 191, 192

Sélection purifiante Type de sélection qui favorise l'élimination de mutations qui diminuent la fitness de l'individu. 36, 44, 78, 90, 94

Taux de mutation (μ) Nombre de mutations par site par unité de temps (génération ou année). 26, 34, 77, 92, 178

Taux de substitution Nombre de substitutions et donc de différences entre espèces, par site par unité de temps (génération ou années). 41, 75, 76, 101, 185

Taxon Groupe d'organismes classés ensemble en fonction de leurs caractéristiques partagées. Un taxon peut correspondre à une espèce, un genre, un ordre, un phylum etc. 38, 43, 66, 79, 95, 100, XIII

Théorie synthétique de l'évolution Voir néo-darwinisme. 25

Topologie Caractérise la forme d'une phylogénie. C'est la structure de l'arbre qui montre comment les différents taxons ou espèces sont reliés entre eux par des ancêtres communs. 58, 74

Trait d'histoire de vie Trait biologique d'un organisme qui caractérise son cycle de vie ou sa reproduction comme le temps de génération, l'âge à la maturité sexuelle ou la longévité. 49, 79, 95, 106, 107, 116, 121, 176, 178

Transformisme Théorie du début du XIX^{e} siècle selon laquelle les espèces évoluent dans le temps par un changement morphologique progressif des individus en réponse à une problématique environnementale. Cette notion se place en opposition au fixisme et a été particulièrement portée par Lamarck. 11, 15, XI

Première partie

Introduction

Préambule : Pourquoi la biologie évolutive ?

J'ai toujours éprouvé un intérêt particulier pour les notions d'origine et de temporalité. Cet intérêt, je le dois en partie au fait d'avoir grandi à 300 mètres d'un planétarium. Pendant les vacances scolaires, j'y passais mes après-midi à explorer en détail les expositions temporaires et à assister aux différentes projections à propos du Soleil, de l'univers ou des voyages spatiaux. En plus d'apprendre à me repérer dans le ciel, j'ai particulièrement développé une curiosité concernant l'origine de notre univers. Je voulais comprendre comment une petite singularité, telle que décrite par les scientifiques, avait pu aboutir au monde si grand et complexe que nous connaissons aujourd'hui. Je voulais également comprendre comment les scientifiques étaient parvenus, au fil des siècles, à décrire, avec plus ou moins de certitudes, des événements aussi anciens que le Big Bang.

En grandissant, et parce que je croyais naïvement que « nous » avions terminé de résoudre l'énigme de l'origine de l'univers et de la formation de notre planète, j'ai redirigé ma curiosité sur les questions d'origine de la vie. Je voulais comprendre comment une telle complexité avait pu émerger et comment les scientifiques avaient étudié ce phénomène. Plus encore, je voulais savoir pourquoi nous sommes tous à la fois si différents et si semblables, comment les premières molécules d'ADN s'étaient formées et, surtout, comment elles fonctionnent. Contrairement à l'astrophysique, je trouvais, face à mes questions, des réponses moins claires.

À la fin du lycée, je savais que je voulais étudier la biologie pour accéder à une plus grande connaissance scientifique sur le sujet [1]. Je n'avais pas pour ambition de faire partie un jour de ces personnes qui créent cette connaissance, tellement la tâche me semblait ardue et réservée à une élite intellectuelle, à laquelle je ne m'identifiais pas. J'ai tenté d'abord une classe préparatoire BCPST où j'ai été plusieurs fois frustrée par des cours de biologie auxquels il m'était interdit de poser la question «Pourquoi ?» C'est seulement lors de mon premier semestre de licence en science de la vie que j'ai pris conscience de l'étendue de la diversité du vivant. J'ai notamment été marquée

1. Merci Mme Colson-Proch !

par l'organisation systématique de la vie en groupes imbriqués les uns dans les autres, chacun répondant à des caractéristiques qui leur sont propres. J'ai appris qu'on pouvait reconstruire l'évolution des espèces le long de ce qu'on appelle un arbre de la vie, les espèces actuelles correspondant aux feuilles et les branches aux liens entre ces espèces. C'est à ce moment-là que j'ai su que je voulais comprendre en détail comment reconstruire ces arbres et à travers eux reconstruire l'évolution des organismes vivants, de leur origine à aujourd'hui.

À la fin de ma licence, j'ai dû choisir entre un master de paléontologie à Paris [2] et le master de biodiversité, écologie et évolution de Lyon. Il me fallait alors choisir mon objet d'étude : les fossiles ou les molécules d'ADN ? Les espèces passées ou actuelles ? Après des discussions éclairantes avec mes professeurs [3], j'ai compris que les molécules d'ADN contiennent l'histoire des individus qui les portent, mais également l'histoire de leurs ancêtres, et mon choix fut fait.

C'est avec cette aspiration à comprendre les origines de la vie et son évolution à travers le temps que j'ai commencé ma thèse en 2021 et c'est dans cette continuité que j'ai décidé de l'écrire. C'est pourquoi je consacrerai dans cette introduction une place importante à l'histoire des idées et découvertes scientifiques menant à l'étude de l'évolution des espèces à travers leurs génomes. Pour ce faire, je souhaite enraciner mon récit au commencement de l'étude de l'évolution des espèces et raconter son évolution via différentes mutations majeures comme le passage d'une considération fixiste du vivant à une perspective évolutive, ou encore le passage de l'étude des caractères morphologiques à celle des séquences puis des génomes. Je traverserai aussi quelques exemples de spéciations qui ont donné naissance à des disciplines tels que la génétique des populations ou la phylogénie. J'aborderai également quelques domaines aux feuilles de cet « arbre de la biologie évolutive » qui font l'objet de mes recherches comme le concept de taille efficace de population, de théorie neutre et quasi-neutre de l'évolution, mais aussi de traitement de données à l'échelle génomique et de méthode comparative intégrative.

Afin que les lecteurs et lectrices puissent se rendre compte de la temporalité des

2. J'avais également une passion pour les dinosaures et fossiles...
3. Merci M. Douady !

événements, de la coexistence de certaines théories avant que l'une d'elle soit fixée jusqu'à nouvelle mutation et de l'effet de certains bouleversements méthodologiques ou théoriques sur la quantité de découvertes qui s'ensuivent, j'ai tenté de reconstruire une frise chronologique relatant les différents événements que je compte aborder dans mon manuscrit.

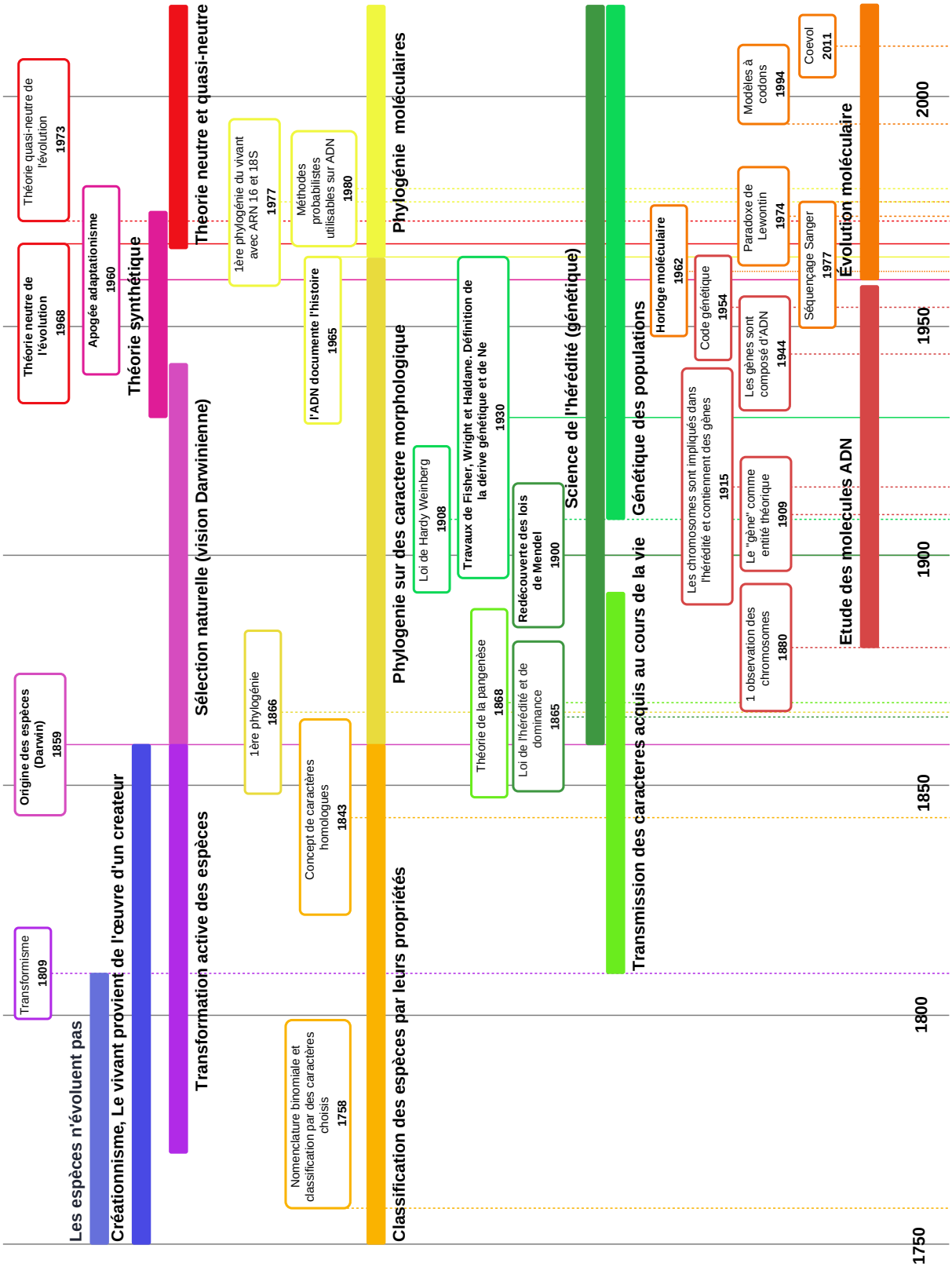


Figure 1 – *Frise chronologique non-exhaustive relatant différents faits marquants dans le domaine de la biologie évolutive. Les dates des périodes sont approximatives.*

1

Origine de la biologie évolutive : De Darwin à la théorie synthétique de l'évolution

1.1 De l'histoire naturelle à l'étude de l'évolution des espèces

1.1.1 Classer et nommer les espèces pour mieux les reconnaitre

L'histoire naturelle, ancêtre de la biologie, est un domaine d'étude qui consiste à décrire la nature par une approche scientifique. Cette discipline fait son apparition en Grèce antique au IV^{e} siècle av. J.-C avec le philosophe grec Aristote et son élève Théophraste qui ont respectivement développé les prémisses de la zoologie et de la botanique. Ils ont chacun décrit des centaines d'espèces de plantes et d'animaux [1]. Aristote a également mis en évidence la nécessité de savoir classer et ordonner ces espèces via des caractères observables pour mieux appréhender la diversité du monde vivant. Il développe ainsi la première classification du vivant dans *Histoire des Animaux*. Celle-ci contient dix groupes d'animaux dont les oiseaux, les poissons, les cétacés et les mollusques. Certains autres groupes comme les serpents ou les crocodiles ne sont pas inclus dans cette classification, ce qui montre une volonté de regrouper certaines espèces similaires pour des raisons pratiques de classification, plutôt qu'exhaustives. Les groupes sont ordonnés de façon linéaire dans une série continue via leur complexité morphologique et anatomique. Pour Aristote, plus nous avançons dans la série et plus l'espèce est dite « supérieure » à la précédente. On retrouve ainsi d'abord les êtres inanimés, puis les plantes, les animaux sans sang, les animaux avec sang et enfin les humains.

Par la suite, on note peu d'avancées majeures dans le domaine de la classification avant le XVIII^{e} siècle tant les écrits d'Aristote feront foi. Toutefois, avec les explorations du monde telles que celles de Marco Polo (1296) et Christophe Colomb (1492), le monde occidental se rend progressivement compte de l'immensité de la diversité de la faune et de la flore à travers le monde. De plus, la médecine se développant, il devient nécessaire de savoir décrire et identifier correctement les différentes plantes médicinales. La description de la nature par les scientifiques

1. Théophraste, *Histoire des plantes*, Aristote, *Histoire des animaux*, Aristote, *Parties des animaux* et Aristote, *Génération des animaux*.

reprend alors, avec pour mission de réunir ce qui a déjà été décrit et d'y ajouter les nouvelles observations, avec un système efficace. En effet, beaucoup d'espèces ont, à ce moment-là, des noms confus contenant des mélanges de langues. Certains noms étant parfois attribués à plusieurs espèces et certaines espèces étant, elles aussi, décrites par plusieurs noms. C'est Linné qui reforme la classification par une standardisation des noms. Il propose d'utiliser une nomenclature binomiale en latin de type «*Genre espèce*» tel que *Homo sapiens* pour l'humain ou *Felis silvestris* pour le chat sauvage. Dans *Species Plantarum* (1753) il classifie 8000 plantes en 24 classes en fonction de leur système sexuel. Il divise également les animaux en six classes dans *Systema naturae* (1758, 10^e édition) qui correspondent globalement aux classes déjà définies par Aristote, à la différence que Linné réfute la classification en série linéaire continue et la notion d'échelle des êtres comme principe qui ordonne la nature. Cette classification par un nombre limité de caractères choisis a, bien sûr, des opposants. Buffon, un zoologiste, est l'un des plus connus. Pour lui, les espèces d'animaux sont moins nombreuses et plus distinguables que les plantes, ce qui les dispense d'une classification qui, dans ce cas, complexifie l'apprentissage naturaliste. Il défend qu'il ne faut pas réduire les espèces à quelques caractères, mais plutôt les étudier sous toutes leurs facettes. Buffon est un personnage très important de son siècle, car il participe à développer la démarche scientifique par l'observation et l'expérimentation. C'est notamment lui qui met à jour la notion d'espèce comme regroupant des individus qui sont en capacités de se reproduire, plutôt que des individus qui se ressemblent.

Peu de temps après, Lamarck, dans *Flore françoise* (1779), développe la distinction entre un système de classification qui sert à distinguer la nature et qui utilise tous les caractères, et un système naturel qui n'utilise que des caractères restreints. Il est le premier à tenter de calculer une proximité entre différents genres pour les classer objectivement, et donc, à apporter des mathématiques dans le domaine de la systématique. Lamarck contribue fortement à la classification des invertébrés qu'il a divisés en sept classes, aidé par les études en anatomie comparée de Cuvier.

Entre 1806 et 1812, Saint-Hilaire développe le concept de caractère analogue en décrivant des organes similaires entre espèces différentes qui sont toujours connectés

aux mêmes autres organes. Ce concept est repris par Owen, en 1843, dans *Lectures on the comparative anatomy and physiology of the invertebrate animals* (Owen, 1843), qui décrit un caractère homologue comme « un même organe dans différents animaux sous toutes les variétés possibles de forme et de fonction ». Ces caractères homologues deviennent des points de comparaison pour les futures études (sous-section 3.1.2).

1.1.2 Les espèces évoluent : du fixisme au transformisme

Bien qu'Aristote développe l'idée d'organisation linéaire du vivant dans une série continue, il ne décrit aucune possibilité d'évolution ou de passage d'une espèce à une autre le long dudit gradient. Pour lui, chaque individu est un variant imparfait du modèle parfait de l'espèce, lui-même fixe dans le temps. Cette notion de fixisme des espèces, autrement dit de non-évolution dans le temps, fut très présente pendant plusieurs siècles, car elle est en adéquation avec le créationnisme, une vision biblique de la nature.

Au XVIII^{e} siècle, le fixisme est de plus en plus rejeté à mesure que l'église perd de son emprise sur le peuple. On retrouve dès lors des écrits faisant l'hypothèse d'une métamorphose des animaux aquatiques vers le milieu terrestre (*Telliamed*, Maillet, 1755) ou de transmission de variations aux générations suivantes. Par exemple, Duschesne écrit : « les espèces paraissent fixes et immuables, mais les accidents qui font varier certains individus procurent à d'autres des changements assez considérables pour qu'ils se perpétuent dans leur postérité » (*Histoire naturelle des fraisiers*, 1766).

Dans les années 1800, Cuvier, le père de la paléontologie actuelle, décrit différentes strates géologiques et observe un changement de composition en espèces fossiles dans le temps avec des couches profondes qui correspondent à des espèces plus anciennes et des couches récentes qui contiennent des fossiles ressemblant aux espèces actuelles. Il détermine ainsi que l'étude successive de ces strates, jusqu'à la surface, pourrait permettre d'étudier l'histoire des espèces passées. Cependant, Cuvier ne voit pas de lien entre les différentes strates. Étant un défenseur du fixisme, il croit en la théorie du catastrophisme selon laquelle il y aurait des cycles de vies créés par génération spontanée et qui se terminent par des catastrophes planétaires.

Lamarck est le premier à formuler une théorie de l'évolution impliquant le transformisme (*Philosophie zoologique*, 1809). Il reprend la hiérarchisation continue des espèces en fonction de leur complexité (comme celle d'Aristote) mais autorise les espèces à évoluer de façon continue d'un maillon à l'autre de la chaîne. Dans sa théorie, les organismes se transforment activement sous l'effet de leur milieu afin de survivre et transmettent les caractères acquis lors de leur vie à la génération suivante. Il développe même un essai pour expliquer le fonctionnement de la transformation par la théorie de la dynamique de fluides organiques, qu'il appelle « Biologie ». Cette théorie postule qu'il y aurait un mouvement de fluides dans des tissus cellulaires souples, engendrant la création de canaux qui sculptent des organes de plus en plus complexes.

1.2 Décrire le mécanisme de l'évolution par la sélection naturelle

Bien qu'il y ait eu quelques essais pour expliquer le mécanisme de transformation des espèces, il faut attendre encore quelques années pour arriver à une théorie qui fera progressivement consensus [2]. C'est en 1831 que Darwin démarre son célèbre voyage de quatre ans sur le *Beagle*. À chaque escale, il étudie les couches sédimentaires, recherche des fossiles et décrit ou recueille des échantillons végétaux et animaux. Bien qu'il espère devenir pionnier dans la géologie de l'Amérique du Sud, ce sont ses communications scientifiques, portant principalement sur la zoologie et la paléontologie, qui le feront connaitre auprès de la société philosophique de Cambridge, avant même son retour.

Le voyage terminé, il entame la rédaction de ses observations et son hypothèse de « transmutation » des espèces. Il est certainement l'un des premiers à faire un lien entre les faits géologiques, géographiques et biologiques, en examinant la répartition des espèces dans l'espace et le temps. Il décrit trois ensembles d'observations majeurs issues de son voyage. Premièrement, il distingue une forte ressemblance entre des organismes vivants, sur des îles proches. C'est le fameux exemple des pinsons des îles de l'archipel des Galápagos. Darwin note une différence entre ces

2. Dans cette section, ce seront les travaux de Darwin qui seront mis en avant, mais il est bon de garder en mémoire que des conclusions analogues ont été proposées par Alfred Wallace, un naturaliste britannique, dans la même période.

oiseaux concernant leur bec qui semble, pour chacun, adapté au régime alimentaire de l'île. Deuxièmement, il constate des similitudes morphologiques, entre des espèces fossiles et actuelles présentes dans une même zone. Enfin, il remarque une similitude entre des espèces éloignées géographiquement, mais ayant des habitats aux propriétés similaires. De ces observations, Darwin fait l'hypothèse d'une modification progressive des espèces, unies par des liens de parenté, en réponse à des contraintes environnementales. La transformation des espèces ayant déjà été décrite (sous-section 1.1.2), il est plutôt question de déterminer à quand remontent ces transformations et de proposer un mécanisme les expliquant.

1.2.1 Les espèces évoluent, mais comment ?

Lutter pour son existence dans un environnement aux ressources limitées

Pour Darwin dans *L'origine des espèces*, (1859, voir figure 1), les individus reçoivent passivement des variations suite à un mécanisme aléatoire [3]. Elles peuvent ainsi être néfastes comme avantageuses en fonction des contextes. Cette variabilité entre individus est déjà connue et exploitée par les agriculteur.rice.s et éleveur.euse.s. Lorsqu'un caractère apparait avantageux, ils et elles favorisent la reproduction des individus les portant pour augmenter son occurrence dans la génération suivante. Ainsi, les caractères démontrent d'une certaine sélectionnabilité et transmissibilité entre individus. En parallèle, dans la nature, les espèces se reproduisent plus rapidement que ne se renouvellent les ressources alimentaires. Cette situation implique une limitation des ressources et, par conséquent, une « lutte » entre les individus d'une même espèce pour leur existence. C'est dans ce contexte que la variabilité entre les individus ainsi que sélectionnabilité des caractères va engendrer un phénomène de tri passif des différents variants. Les combinaisons de variants les plus aptes à exploiter leur environnement vont plus survivre que les autres et ainsi fournir plus de descendants portant ces caractères. Au fil des générations, si les contraintes environnementales restent les mêmes, tous les individus de l'espèce finissent par posséder ces caractères avantageux. Cette forme de sélection dans la nature a des effets semblables à celle engendrée

3. Contrairement à Lamarck qui pense qu'elles survenaient directement en réponse aux besoins de l'individu.

artificiellement par les agriculteur.rice.s et les éleveur.euse.s, ainsi, Darwin la nommera « sélection naturelle » en analogie à la sélection artificielle. Cette théorie de la sélection naturelle implique que si l'environnement venait à changer, alors les caractères sélectionnés changeraient aussi. C'est le cas du bec des oiseaux des Galápagos dont la taille varie en fonction des îles et donc des environnements.

Les espèces ont des histoires et des ancêtres en communs

Pour expliquer l'existence de caractères communs entre des espèces qui ne se reproduisent pas, Darwin émet l'hypothèse que les ancêtres de ces espèces se reproduisaient ensemble par le passé. Elles auraient donc une ascendance commune. Cette hypothèse, intitulée la « descendance avec modification », implique que chacune des deux espèces, une fois séparée de cet ancêtre commun, a accumulé des variations. Cela explique le fait qu'elles soient toutes deux distinguables aujourd'hui (voir figure 1.1.A). De plus, cette théorie explique un certain degré d'emboitement dans la répartition des caractères, plutôt connu des naturalistes. Par exemple, toutes les espèces qui ont des poils ont des vertèbres. Ainsi, les structures hiérarchiques trouvées dans les classifications précédentes peuvent s'interpréter par une généalogie des espèces où les caractères homologues décrits par Owen sont en fait hérités des ancêtres communs (figure 1.1.B).

Darwin explique que pour constituer les groupes de la classification, il faut considérer leurs caractères commun et non les modifications apparues sur une seule lignée. Ainsi, si une espèce semble différer d'un groupe, mais qu'elle possède le caractère discriminant dudit groupe, alors l'espèce doit être incluse dans celui-ci. C'est ainsi que l'on a pu réunir les reptiles et les oiseaux dans un même ensemble dit monophylétique. De fait, il existe une unique classification des espèces représentant une unique histoire évolutive (voir figure 1.1). L'objectif en systématique change alors et devient celui de retracer l'histoire du vivant et de comprendre le lien entre les organismes et leur évolution. Après un premier schéma hypothétique de Darwin (voir figure 1.1.B), c'est Haeckel, en 1866, dans *Generelle morphologie des organischen* qui met en forme une première classification des espèces basée sur la généalogie. Il place, en plus, dans celle-ci, des observations paléontologiques (voir figure 1.1.C). Il invente également le terme « phylogénie » bien connu aujourd'hui

(voir figure 1). Il faudra cependant attendre un siècle pour obtenir les premières phylogénies basées sur des séquences ADN (voir figure 1, figure 1.1.D, sous-section 2.1.3).

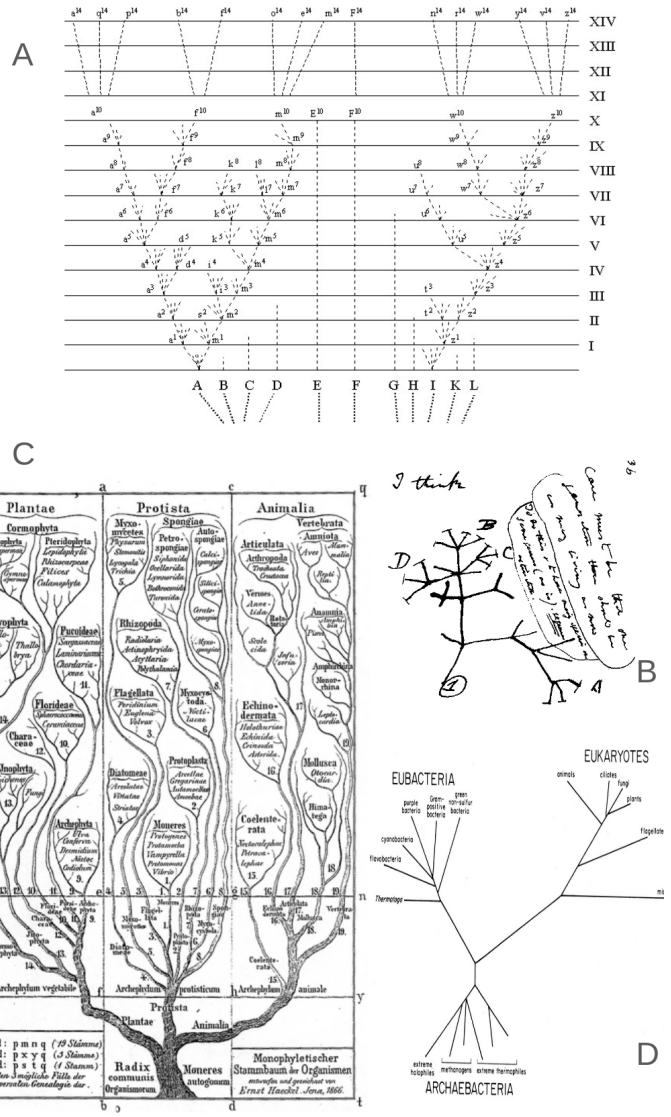


Figure 1.1 – ***A***, *Un schéma théorique de la forme que devrait prendre la généalogie du vivant (1859). En ordonnée, différentes échelles de temps et en abscisse, la variabilité entre individus ou espèces. On retrouve la schématisation de différentes unités, qui se modifient continuellement. Certaines unités s'éteignent tandis que d'autres survivent et se scindent en deux nouvelles unités (entrainant, par conséquence, l'extinction de l'unité initiale, devenue ancêtre) qui vont elles-mêmes évoluer.* ***B***, *Une ébauche célèbre de 1837, retrouvée dans un des carnets de Darwin, qui inspirera plus tard le développement des phylogénies.* ***C***, *Première classification du vivant basée sur la généalogie par Haeckel (1866).* ***D***, *Première phylogénie du vivant en utilisant l'ARN 16s et 18s par Woese (1987).*

1.2.2 Les controverses et les mauvaises interprétations

La sélection naturelle fait preuve d'une très forte puissance explicative. En effet, elle a réuni des observations supposées indépendantes telles que la distribution géographique des espèces, les adaptations morphologiques, les observations paléontologiques et les extinctions. Si bien qu'elle est, aujourd'hui, identifiée comme une révolution scientifique majeure.

Bien que Darwin prend particulièrement la peine de formuler ses explications de façon à prévenir les différentes critiques, sa théorie connait tout de même plusieurs obstacles tant elle remet en cause un grand nombre de croyances et théories déjà établies. Parmi ces croyances, on retrouve les notions de créationnisme et de fixisme, portées par l'Église ainsi que d'autres scientifiques comme Cuvier. En effet, avec sa théorie de la sélection naturelle, Darwin, comme d'autres, remplace le fixisme par le transformisme, mais également, le créationnisme par le hasard et la sélection naturelle. De plus, l'idée d'un lien généalogique entre les êtres vivants crée une rupture avec l'échelle des êtres se perfectionnant, décrite par Aristote. Ici, il n'est plus question de donner un sens et une intention à l'évolution des espèces qui ne tendent finalement pas vers une perfection absolue. Dans la figure 1.1 panel A et B, on peut justement voir que Darwin représente de multiples espèces coexistantes et évoluant toutes en parallèle, de façon buissonnante et non pas linéaire.

Pour autant, il ne faut pas prêter à Darwin des propos qu'il n'a pas tenus, quand bien même ces propos sont aujourd'hui des idées reçues populaires concernant la sélection naturelle. Dans *La filiation de l'Homme* (Darwin), il inclut l'humain dans son explication et enracine son origine parmi les primates. Cependant, plusieurs de ses détracteurs, ont tourné son travail en dérision en transformant cela en « l'Homme descend du singe », idée fausse qui perdure encore aujourd'hui. Il est important de rappeler que toutes les espèces actuelles d'un groupe sont descendantes d'un ancêtre commun qui n'est pas actuel (voir figure 1.1 A). L'humain est ainsi un cousin des autres singes actuels, pas leur descendant.

Parmi les différentes mécompréhensions, on peut également mentionner la glorification de la « loi du plus fort ». On confond ici une aptitude supérieure à la

survie dans un contexte donné à une supériorité en terme de puissance physique. Pourtant, Darwin a dédié dans son livre de 1859 un chapitre expliquant que les animaux les plus puissants par la force sont les premiers à se raréfier et à s'éteindre car cette force est énergivore et donc sensible aux pénuries alimentaires.

Enfin, il n'est pas possible de mentionner les fausses idées sur la sélection naturelle sans mentionner le « darwinisme social » qui consiste à naturaliser les inégalités présentes dans les sociétés humaines. Pour Spencer, un partisan de ce concept, une société est un organisme qui est soumis, de fait, à la loi sélective et donc à la survie du plus apte (Spencer, 1864). Il justifie ainsi qu'il n'est pas nécessaire d'essayer de préserver les « moins aptes » (« ou moins méritant »), de l'espèce humaine et que c'est la nature qui réalise, elle-même, le tri nécessaire permettant une amélioration naturelle et continue de la société. Certains, comme Galton, vont encore plus loin et développent l'idée d'une intervention artificielle sur la société pour compenser les déficits apportés par la survie « contre-nature » des « moins aptes ». C'est ainsi que se développe l'eugénisme (Galton, 1865, 1891) qui entraine au XX^{e} siècle, aux États-Unis et en Europe, par exemple, des milliers de stérilisations humaines sur des individus[4] présentant des déficits mentaux ou des caractéristiques jugées non souhaitables comme l'homosexualité. Galton est un fort défenseur de la théorie de Darwin et de l'évolution continue face au mutationnisme mendélien (voir sous-section 1.3.2). Il développe d'ailleurs le domaine de la biométrie. Mais cependant, il se retrouve en forte opposition avec Darwin sur son interprétation erronée de la théorie de la sélection naturelle.

1.3 Mécanismes de transmission des caractères à la descendance

Un point important des théories de Darwin et de Lamarck est qu'elles décrivent une transmission des caractères biologiques des parents aux descendants que l'on nomme hérédité. Au même moment, plusieurs théories de l'hérédité coexistent, impliquant, pour la plupart, la transmission aux descendants, par les parents, des

4. Dont Alan Turing qui aurait peut-être mis fin à ses jours à cause d'un traitement chimique de stérilisation auquel il a été soumis de force. Il est l'inventeur de l'ordinateur, sans lequel ce travail de thèse n'aurait pas été possible.

caractères qu'ils ont acquis au cours de leur vie. Une des théories les plus connues, d'ailleurs développée par Darwin[5], est la théorie de la pangénèse. Cette théorie soutient que chaque partie de l'organisme participe à l'hérédité en transmettant des particules porteuse de caractères, tout au long de la vie. Ces différentes théories concernant la transmission des caractères acquis au cours de la vie restent longtemps dominantes jusqu'à ce qu'elles soient réfutées par Weissman en 1892 lorsqu'il présente sa théorie du plasma germinatif séparant les cellules germinales (porteuse des mutations qui seront transmises) des cellules somatiques (Weismann, 1893).

Le sujet de la transmission des caractères aux descendants devient particulièrement important dans le contexte de croissance démographique et économique quand il fut question d'augmenter le rendement agronomique. Il faut alors être capable de détecter efficacement les caractères les plus pertinents (par exemple, une meilleure résistance aux parasites) et de les sélectionner.

1.3.1 Des croisements hybrides aux propriétés particulières

Plusieurs botanistes comme Sageret et Gärtner (Sageret, 1826; Gärtner, 1849) ont réalisé des expériences d'hybridations de plantes ayant chacune un caractère d'intérêt, mais aucun n'en formula de loi permettant de prédire les observations.

Entre 1856 et 1863, Mendel étudie le croisement de milliers de plantes, notamment des pois, et élabore des lois statistiques décrivant l'hérédité de caractères discrets qu'il nomme « loi de dominance » et « loi d'appariement » (Mendel, 1865, 1870) (voir figure 1). Il est le premier à inventer un protocole expérimental reproductible n'impliquant que deux caractères à la fois et permettant de quantifier les différents hybrides observés, par génération. Dans ses publications, Mendel décrit que la première génération d'un croisement hybride donne des descendants tous semblables puis, la deuxième génération forme un ensemble non homogène d'individus, mais qui présentent une répartition très particulière des caractères d'intérêt selon une proportion de un pour quatre. Par exemple, lors du croisement d'une fleur rouge et d'une autre blanche, la première génération obtenue présente le

5. Dans le chapitre *Pangenesis* de *The variation of animals and plants under domestication*, 1868.

caractère rouge. Le croisement de deux individus de cette première génération donne, en proportion, un individu blanc sur quatre. En effet, il existe deux variants du « facteur » déterminant la couleur dont un est dominant, ici le rouge. Chaque individu porte deux fois l'information de la couleur (il est diploïde) ce qui mène à quatre combinaisons possibles (appelées plus tard des génotypes) : rouge-rouge, blanc-rouge, rouge-blanc, et blanc-blanc, menant au phénotype rouge pour les trois premiers cas et blanc pour le dernier.

L'étude simultanée de caractères différents a permis à Mendel de montrer que chaque caractère est héréditairement indépendant et déterminé par un « facteur » qui lui est propre. Plus tard, ces facteurs seront renommés « gènes » par Johannsen (Johannsen, 1909) et on découvrira qu'ils sont portés par des chromosomes présents dans les noyaux de chaque cellule.

Ces résultats restent ignorés jusqu'en 1900, à l'exception d'une unique publication mentionnant ces travaux en 1881 par Focke[6].

1.3.2 L'évolution est-elle continue ou graduelle ?

En parallèle des études de Mendel, des développements mathématiques sont proposés afin d'analyser la transmission héréditaire des variations entre générations. On relève notamment les concepts, aujourd'hui communs, de droite de régression (Galton, 1877), de coefficient de corrélation et de test de Khi-2 (Pearson, 1900). Ces études s'inscrivent dans le domaine naissant de la biométrie qui correspond à l'étude statistique de la variabilité des traits entre individus. Les biométricien.ne.s font, comme Darwin, l'hypothèse que l'évolution se produit de façon continue par action de la sélection naturelle sur des variations minimes.

Dans les années 1900, les conclusions de Mendel sont redémontrées de façon indépendante par plusieurs scientifiques (De Vries, 1900; Correns, 1900; Tschermak, 1900) et reliées à leur premier découvreur grâce à la citation de Focke mentionnée plus haut. D'autres scientifiques testent avec succès ces lois sur des

6. « Mendel croyait avoir trouvé des relations numériques constantes entre les types produits par hybridation », Focke 1881

animaux (Cuénot, 1902; Bateson and Gregory, 1905; Morgan *et al.*, 1915). La science de l'hérédité, telle qu'imaginée par Mendel, est alors renommée « génétique » en 1906, et on distingua les notions de génotype (combinaison de gènes dans l'individu) et de phénotype (réalisation observable du génotype) (Johannsen, 1909; Churchill, 1974). Petit à petit se développe un rejet de l'évolution continue par changements minimes défendue par Darwin et les biométricien.ne.s (voir l'article de Provine (1970) pour une description du contexte). On mentionne notamment De Vries qui décrit dans sa « théorie mutationelle » que les nouvelles espèces se forment par bonds mutationnels (De Vries, 1909). Il explique notamment que « la sélection naturelle peut expliquer la survivance, mais non l'apparition du plus apte » (De Vries, 1909). Des expériences sont également menées sur des lignées pures, montrant l'inefficacité de la sélection naturelle sur ces lignées (Johannsen, 1909; Provine, 1970).

Par la suite, Fisher montre qu'une évolution continue et sa transmission héréditaire ne sont pas incompatibles avec la théorie mendélienne (Fisher, 1918, 1919). Ce qui permet d'associer les deux disciplines sous le nom de « biométrie génétique » (Mather, 1964; Visscher and Goddard, 2019).

1.3.3 Du concept de gène théorique au code génétique

Les notions de gènes et de génotype ont été postulées avant d'être réellement observées. Pour autant, il n'a pas fallu longtemps pour relier l'observation de chromosomes (initialements définis comme des corps filamenteux colorés présents dans les noyaux des cellules), faites dans les années 1880, avec les processus liés à l'hérédité. C'est ainsi que, en 1915, Morgan énonce que les chromosomes sont impliqués dans l'hérédité et sont le support des gènes (Morgan *et al.*, 1915). Cette découverte n'est que la première pierre à l'édifice de ce qui deviendra l'étude des séquences moléculaires en lien avec l'histoire évolutive des espèces (section 2.1 et figure 1). On montre que les gènes, maintenant identifiés, sont en lien avec la production d'enzymes particulières et sont composés d'ADN lui-même fait de quatre types de nucléotides nommés « A », « T », « G » et « C »[7] (Beadle and Tatum, 1941; Avery *et al.*, 1944; Portin, 2002) (Voir Box 1). Ainsi, si l'ADN est modifié, on

7. Pour Adénine, Thymine, Guanine et Cytosine, respectivement.

suppose que les caractères observés le seront aussi.

Par la suite, des chercheurs et chercheuses mettent en évidence la complémentarité entre les paires de nucléotides A T et les paires G C (Chargaff, 1950). On a fait le lien entre cet alphabet nucléotidique de quatre lettres, composant l'ADN, et les séquences protéiques, faites à partir d'une liste de vingt acides aminés différents (Box 1). En effet, les séquences ADN se lisent par trois pour former des codons, ensuite décodés en acides aminés formant les protéines, via le code génétique (Gamow, 1954; Di Giulio, 2005). Cependant, la combinaison par trois des différents nucléotides donne 64 propositions possibles tandis que les protéines ne présentent que 20 acides aminés différents. Ainsi, plusieurs associations de trois nucléotides peuvent donner le même acide aminé. On dit qu'ils sont synonymes et que le code génétique est dégénéré. En général, les codons synonymes différent entre eux par leur troisième lettre (figure 1.2), impliquant que la troisième position d'un codon peut être mutée sans qu'il n'y ait de conséquences pour la séquence protéique.

L'analyse des séquences nucléotidiques et protéiques dans les années 1960-80 a permis d'observer que la majorité des différences entre les séquences de divers individus, correspond à des mutations synonymes qui n'impactent donc pas la fonction de la protéine. On dit de ces mutations qu'elles sont « neutres » et qu'elles ne sont pas soumises à la sélection naturelle.

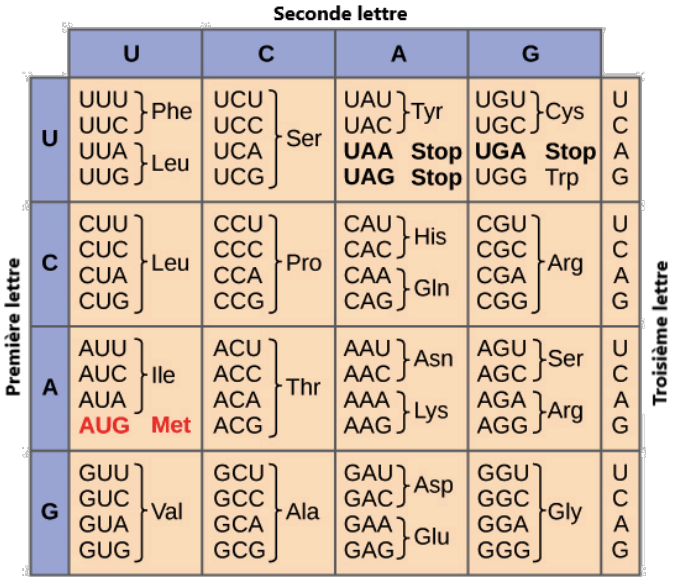


Figure 1.2 – *Le code génétique. La majorité des troisièmes position des codons peut muter sans que l'acide aminé ne change. Les codons qui codent un même acide aminé sont dits synonymes.*

Box 1 : De la molécule d'ADN à la séquence protéique

La molécule d'ADN (Acide DésoxyriboNucléique) est faite de deux brins complémentaires enroulés en hélice. Un brin d'ADN contient une suite de 4 molécules, A (Adénine), T (Thymine), G (Guanine) et C Cytosine, appelés **nucléotides** et qui sont complémentaires 2 à 2. Ainsi, quand il y a une base A, on retrouve un T sur le brin complémentaire.

Un **gène** correspond à une séquence particulière de nucléotides qui code pour une **protéine** spécifique faite d'**acides aminés**. Il contient à la fois des **exons**, qui sont des séquences renfermant le code de la protéine, et des **introns**, des séquences non codantes.

Quand une protéine est nécessaire pour l'organisme, le gène est **transcrit** en un brin d'**ARNm** (Acide RiboNucléique messager). Après un processus de maturation appelé épissage, qui élimine les introns, l'ARNm mature est traduit par les **ribosomes**, qui sont constitués d'**ARNr** (ARN ribosomique). Des molécules d' **ARNt** (ARN de transfert) apportent les acides aminés nécessaires, permettant ainsi la synthèse d'une protéine par une **traduction** de l'ARNm. Pour ce faire, les nucléotides sont lus trois par trois et traduits en acides aminés via le **code génétique**, contenant 20 sortes d'acides aminés différents (voir Fig.1.2). Parmi les protéines, on retrouve, par exemple, des enzymes et des hormones.

Une **mutation** est un changement dans la séquence ADN. Elle peut prendre la forme d'un remplacement d'un nucléotide par un autre ou bien d'un changement structurel comme une suppression d'une partie de la séquence ADN.

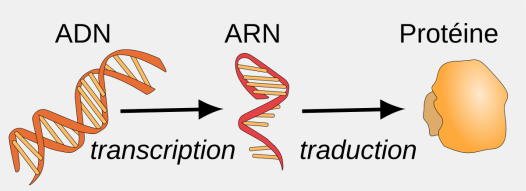


Figure 1.3 : Le dogme central de la biologie moléculaire

1.4 Modéliser mathématiquement le devenir des variants dans une population

En parallèle de l'étude de la transmission des caractères des parents aux descendants, se développe un nouveau champ disciplinaire concernant la dynamique évolutive des variants (maintenant nommés allèles) à l'échelle d'une population. On l'appelle « génétique des populations » (voir figure 1). Dans ce contexte d'étude, des chercheurs et chercheuses réalisent des expérimentations sur de grandes populations de drosophiles (Morgan, 1910; Roberts, 2006), afin d'éviter tout effet potentiel dû au hasard. Ces études sont menées en milieu contrôlé, pour ne garder que quelques facteurs explicatifs dont la variation est contrôlée lors de l'expérience. Le développement des méthodes mathématiques, initié par les biométricien.ne.s, fournit un socle solide pour analyser les résultats de ces expériences, mais aussi, pour développer des modèles permettant de simuler le réel et de fournir des prédictions concernant le devenir des différents variants.

Un premier développement important est proposé par Hardy et Weinberg en 1908 (Hardy, 1908; Weinberg, 1908) lorsqu'ils développent le lien entre la fréquence des génotypes et celle des allèles. Leurs travaux ont mené à l'écriture de la « loi de Hardy Weinberg » (Dobzhansky, 1951), aujourd'hui bien connue des étudiant.e.s en génétique. Néanmoins, c'est dans les années 1930 que la génétique des populations se développe réellement avec trois ouvrages majeurs qui sont ceux de Fisher, Haldane et Wright (Fisher, 1930; Haldane, 1932; Wright, 1931) et qui ont permis de proposer une vraie synthèse commune du darwinisme, du mendélisme et de la biométrie. On postule, dans les modèles développés, que la mutation introduit de la variabilité en créant de nouveaux allèles, et que la sélection naturelle agit sur la dynamique de ces allèles dans la population.

Fisher propose notamment des modèles stochastiques tenant compte de la fluctuation aléatoire des gamètes, sans sélection, dans lesquels les fréquences alléliques fluctuent avec une variance de 1/2N avec N le nombre d'individus (initialement 1/4N (Fisher *et al.*, 1922) puis corrigé (Fisher, 1931)). Il décrit que sans l'arrivée de nouvelles mutations ni sélections dans le système, les allèles sont progressivement éliminés ou fixés aléatoirement, ce qui diminue la variabilité entre

individus. Cependant, il en conclut que cette diminution de variabilité serait très lente et donc que l'effet aléatoire (aussi appelé dérive génétique, voir sous-section 2.2.2) serait négligeable, en raison de la très grande taille des populations (Fisher, 1931). Il a notamment illustré son propos en développant le concept aujourd'hui bien connu en écologie de « capture-recapture », lui permettant d'estimer la taille d'une population de *Panaxia dominula* (Fisher *et al.*, 1947) et d'en conclure que celle-ci est effectivement trop grande pour que les fluctuations alléliques puissent se faire par des processus aléatoires. Ainsi, même si Fisher inclut, en plus de la sélection naturelle, différents mécanismes pouvant impacter la variation des fréquences alléliques dans ses modèles, tels que la dérive génétique et la mutation, il leur attribue des rôles négligeables dus aux grandes tailles de populations. Fisher a une grande influence en Angleterre, si bien que l'étude de la dérive génétique sera longtemps négligée parmi ses descendants scientifiques.

En parallèle, Wright, qui lui, est en Écosse, développe la théorie des glissements de terrains, décrivant qu'une grande population subdivisée en petites sous populations montre une évolution rapide en trois phases : dérive, sélection dans la sous-population et sélection entre sous-populations (Wright *et al.*, 1932; Wright, 1970, 1984). Il insiste particulièrement sur l'importance de la dérive génétique dont il développera une définition dans plusieurs articles (Wright, 1931; Wright *et al.*, 1932; Wright, 1970) (voir sous-section 2.2.2, section 2.5 et Box 3). Sa position concernant l'importance de la dérive génétique dans l'évolution des espèces mène à de vifs débats contre Fisher et son école (Fisher and Ford, 1950; Wright, 1951).

On peut mentionner également Haldane qui propose une approche déterministe de l'évolution (sans processus stochastique dans ses modèles) et défend la nécessité d'une théorie quantitative concernant la sélection naturelle (Haldane, 1926, 1932).

1.5 Mise à jour de la théorie de l'évolution

À la moitié du XX^{e} siècle, plusieurs avancées importantes dans l'étude de l'évolution des espèces sont faites (voir figure 1). Premièrement, la description de facteurs héréditaires transmis indépendamment permet de définir les lois de l'hérédité et de la dominance des caractères (Mendel, 1865). Deuxièmement, la

formalisation du concept de gène, d'abord théorique puis observé biologiquement (Johannsen, 1909; Morgan *et al.*, 1915), donne un support physique concret à ces facteurs héréditaires. Ensuite, le débat entre une évolution continue ou ponctuée stimule le développement de méthodes mathématiques pour étudier la transmission des caractères (Fisher, 1918; Provine, 1970). Cela favorise l'émergence de principes et de modèles mathématiques en génétique des populations, intégrant divers facteurs explicatifs de la variation des fréquences des allèles dans une population. Enfin, d'autres avancées, non détaillées ici, comme les travaux de Simpson en paléontologie ou de Mayrs sur l'étude de la spéciation, viennent compléter ces découvertes.

Ces progrès nécessitent une mise à jour de la théorie de l'évolution formulée 100 ans plus tôt par Darwin. On nomme cette nouvelle théorie, le néo-darwinisme (Angleterre) ou théorie synthétique de l'évolution (États-Unis). Pour Julian Huxley (1942), un des co-fondateurs de la théorie synthétique [8], l'évolution d'une espèce se produit lorsqu'un variant apparait par mutation dans la population, et que sa fréquence change par l'action principale de la sélection naturelle, dans un environnement particulier. Les tailles de populations étant supposées très grandes, on néglige l'effet de la dérive génétique (Fisher, 1930). Les mutations sont considérées comme ayant un fort effet sur les individus et sont donc rapidement fixées ou éliminées par la sélection. Même si on suppose que les mutations génétiques sont majoritairement défavorables, on s'attend à ce que la plupart des mutations fixées observées soient avantageuses puisque les mutations défavorables ont été éliminées.

Dans les années 1960, la théorie synthétique est généralisée dans les milieux scientifiques et la sélection naturelle est largement reconnue comme étant le principe majeur régissant l'évolution. Dans ces années-là, toute variation génotypique et phénotypique s'explique par l'adaptation d'une espèce à son environnement et on suggère même de retirer la mention de la dérive génétique dans les discours concernant les causes de l'évolution (Mayr (1963), chap 8). Néanmoins, des scientifiques comme Waddington et Mayr (Waddington, 1957; Mayr, 1955) mettent en lumières certaines lacunes dans cette théorie comme le fait que la génétique des

8. Eugéniste et petit fils de Thomas Huxley lui-même connu sous le nom de « Bouledogue de Darwin » tant il participa aux débats soulevés par la théorie de la sélection naturelle dans les années 1860.

populations ne permet aucune quantification concrète puisqu'elle repose sur des paramètres inconnus, comme le taux de mutation ou la taille des populations. De plus, la génétique des populations classique a besoin d'être réformée pour considérer l'effet d'une mutation particulière sur l'individu relativement à d'autres facteurs, comme les interactions entre gènes et la taille de population, plutôt que fixe et absolue comme initialement envisagé (Mayr, 1955). Enfin, tout ce développement reste très théorique. En effet, on parle de fréquences alléliques dans une population et de ses modifications dans le temps, dans un contexte où on ne sait pas encore relier directement un gène à un changement particulier de phénotype (Fisher, 1918; Visscher and Goddard, 2019). La plupart des observations se font, en réalité, uniquement sur des changements phénotypiques sans vraies connaissances des structures internes des gènes. De fait, à cette époque, la théorie synthétique ne peut que fournir des prédictions théoriques concernant la réponse d'un trait à la sélection, sans possibles vérifications empiriques directement sur les séquences moléculaires.

2

Comprendre l'évolution et ses mécanismes par l'étude des séquences moléculaires

2.1 (R)évolution moléculaire

2.1.1 Les attendus de la théorie synthétique concernant l'évolution moléculaire

Comme développé dans le premier chapitre de cette introduction, la biologie évolutive des années 1960 s'inscrit dans un contexte où la théorie synthétique de l'évolution (ou néo-darwinisme) met particulièrement en avant le rôle de la sélection naturelle et minimise celui de la dérive génétique pour expliquer l'évolution des organismes. Des modèles mathématiques prenant en compte plusieurs facteurs liés à l'évolution sont développés, mais ils ne peuvent pas encore être confrontés directement aux données empiriques moléculaires. Ainsi, à la lumière de ce qui a déjà été observé sur les phénotypes, on présuppose que les gènes sont proches d'un optimum évolutif atteint par sélection des mutations favorables, et qu'ils peuvent potentiellement s'en éloigner si l'environnement change, entrainant un nouveau cycle d'adaptations.

L'étude par migration sur gel des enzymes, qui sont des protéines particulières, dans les années 1960, puis l'arrivée du séquençage nucléique par les méthodes Sanger (1977) (sous-section 3.1.1), ont permis de tester empiriquement ces attendus en comparant des séquences d'individus de mêmes espèces ou d'espèces différentes.

2.1.2 Comparaisons de séquences et contradictions de la théorie synthétique

En comparant les séquences d'une même enzyme de deux ou plusieurs individus d'une même espèce (ou en s'intéressant à l'hétérozygotie d'un unique individu diploïde), on observe qu'elles présentent bien plus de différences entre elles qu'attendu selon les préconceptions entretenues à l'époque (Harris, 1966; Lewontin

and Hubby, 1966). On appelle ces différences entre individus du polymorphisme et il est assez étonnant d'en observer une si grande quantité alors que la variation phénotypique entre les individus d'une même espèce est en comparaison plutôt faible.

Ces premières observations montrent une incohérence entre les attendus néo-darwiniens basés sur l'évolution phénotypique, et les premières données moléculaires. À cela, viennent s'ajouter d'autres observations majeures qui seront abordées plus en détail dans les sections suivantes, comme (1) l'identification d'un nombre important de substitutions neutres dans les séquences codantes (sous-section 2.2.1), (2) la présence de régions génomiques plus ou moins sensibles à la mutation (sous-section 2.2.3) et (3) l'apparent lien entre le nombre de mutations dans les protéines et le temps qui les sépare de leur ancêtre commun (horloge moléculaire, section 2.3). Toutes ces observations concernant la variation moléculaire intra et inter espèces vont venir questionner la légitimité de la sélection des mutations avantageuses comme mécanisme majeur de l'évolution au niveau génomique et stimuler l'émergence d'un nouveau paradigme (section 2.2).

Dans la mesure où cette théorie tente d'expliquer simultanément les variations intra et inter spécifiques, et avant de l'exposer plus en détail, il est nécessaire d'introduire les avancées qui ont lieu au même moment sur le front de ce qui deviendra la phylogénie moléculaire.

2.1.3 L'ADN, matière première pour reconstruire l'histoire évolutive

L'arrivée des données moléculaires nécessite un travail de réflexion concernant le choix des molécules à étudier pour retracer l'histoire évolutive des organismes. C'est Zuckerkandl et Pauling qui proposent la molécule d'ADN comme élément à privilégier pour retracer cette histoire (Zuckerkandl and Pauling, 1965b). Ils identifient ainsi ce qui deviendra, et demeure encore aujourd'hui, le matériel de base de l'étude de l'évolution et la reconstruction des phylogénies moléculaires (Zuckerkandl and Pauling, 1965b) (voir figure 1). Dans leur article « *Molecules as a document of evolutionary history* », Zuckerkandl et Pauling (1965b)

détaillent trois types de molécules : les sémantides, des molécules qui portent l'information du gène ou de ses transcrits (ADN, ARN, protéine), les épisémantides, qui sont synthétisées sous le contrôle des sémantides de type protéines (les glucoses, par exemple), et les asémantides, des molécules qui ne seraient pas produites par l'organisme (les vitamines, par exemple). Ils font également l'hypothèse que plus une molécule est complexe, plus elle est intéressante pour la phylogénie. Pour eux, le type de molécule dont il faut privilégier l'étude est les sémantides, car en plus d'être partagées par tous les organismes vivants sur Terre, elles présentent une variation abondante fournissant du matériel pour la reconstruction phylogénétique, tout en ayant des régions communes entre espèces, ce qui facilite l'établissement des relations d'homologies entre séquences. Zuckerkandl et Pauling développent également que, parmi les sémantides, toutes ne portent pas la même information. En effet, la dégénérescence du code génétique (Weisblum *et al.*, 1962) (figure 1.2 et sous-section 1.3.3) mène à des changements dans l'ADN et l'ARN qui ne sont pas toujours observés dans les séquences protéiques. Ces substitutions sont dites « isosémantiques » et aujourd'hui on les appelle « synonymes ». Du fait de l'existence de ces mutations synonymes, les molécules d'ADN semblent plus pertinentes à étudier que les protéines, car elles portent l'information, selon les auteurs, de « ce qui a été réalisé et de ce qui est potentiel » (Zuckerkandl and Pauling, 1965b).

À partir de cet article affirmant que l'ADN documente l'histoire évolutive et des phylogénies reconstruites sur la base des séquences d'hémoglobines, montrées par Zuckerkandl et Pauling (résumé par Morgan 1998), plusieurs développements méthodologiques sont poursuivis afin de reconstruire des phylogénies, non plus sur la base de caractères morphologiques, mais plutôt en utilisant l'information historique présente dans les séquences protéiques et nucléiques. C'est le début de la phylogénie moléculaire (section 3.2). On crée des « matrices de distance », dans lesquels on comptabilise le nombre de différences entre séquences homologues, afin de reconstruire l'arbre décrivant les relations évolutives des espèces (Sneath, 1977; Sokal *et al.*, 1965; Schlee, 1975). En parallèle, Cavalli-Sforza et Edwards développent différentes méthodes, comme celle des moindres carrés, celle du maximum de vraisemblance et celle de l'évolution minimale (Edwards and Cavalli-Sforza, 1963, 1965; Edwards, 1972; Cavalli-Sforza and Edwards, 1967). Chaque méthode choisit, parmi l'ensemble des arbres possibles, celui qui est le plus approprié selon des

critères de discrimination qui leur sont propres. Cavalli-Sforza et Edwards déterminent que la méthode la plus rigoureuse est la méthode de maximum de vraisemblance, bien que celle-ci est, du moins à cette époque, difficilement utilisable. Par la suite, Saitou et Nei (1987) développent l'algorithme de Neighbour Joining, qui reconstruit directement l'arbre d'évolution minimum sur la base d'une matrice de distance donnée en entrée.

Les séquences moléculaires, bien que très informatives, apportent aussi leurs limitations. Il existe par exemple un phénomène de saturation où une substitution ancienne peut être recouverte par une autre plus récente, rendant la première invisible (voir review : Moreira and Philippe (2000); Philippe *et al.* (2011)). On parle alors de substitutions multiples. Plus la divergence entre deux séquences est grande et plus il y a de substitutions multiples. Il est alors nécessaire de créer des modèles probabilistes d'évolution qui, au lieu de comparer les différences effectivement observées entre séquences, fournissent des probabilités de substitution. Ils permettent ainsi d'estimer le nombre réel de substitutions à partir du nombre de différences observées dans les séquences (voir section 3.2) (Håstad and Björklund, 1998; Strimmer *et al.*, 2003; Pupko and Mayrose, 2020). Cependant, comme mentionné plus haut, l'utilisation de méthodes probabilistes est très couteuse computationellement et ainsi, elles ne seront pas réellement appliquées avant les années 1980, lorsque les ordinateurs seront suffisamment puissants. Ce sera d'ailleurs à cette période que Felsenstein (1981) proposera le premier modèle sur ADN.

2.2 La théorie neutre de l'évolution

Kimura, dans son livre *The neutral theory of molecular evolution* (1983), explique avoir eu, en 1967, l'idée que la majorité de la divergence et du polymorphisme observés correspondent à des mutations neutres (sous-section 2.2.1) qui ont été fixées (ou sont en train de se fixer) par l'action d'un processus de dérive génétique aléatoire (Kimura, 1968a)(sous-section 2.2.2). Pour lui, la dégénérescence du code génétique (figure 1.2, sous-section 1.3.3) ainsi que l'effet, finalement plutôt faible, d'une fraction potentiellement importante de mutations non-synonymes, rendent l'action de la dérive génétique plus importante que celle de la sélection (Kimura, 1968a,b). Il mentionne également que d'autres équipes de recherche sont

indépendamment arrivées aux mêmes conclusions que lui, bien que ces équipes prêtent plus attention à la divergence entre espèces qu'à l'étude du polymorphisme (King and Jukes, 1969).

La théorie neutre de l'évolution est largement débattue et mise en opposition à la théorie de la sélection darwinienne via ce qu'on appelle le débat neutralistes/sélectionnistes. Il existe de nombreux articles qui défendent l'une et/ou l'autre des théories, dont ceux de Crow (1981) et Lewontin (1974) par exemple.

2.2.1 Une majorité de mutations délétères et de substitutions neutres

Dans sa théorie neutre de l'évolution, Kimura découpe les mutations en trois classes. Une première classe regroupe les mutations à effet délétère. Ces mutations sont assez rapidement purifiées par la sélection naturelle, les rendant souvent non visibles dans le polymorphisme ou la divergence. La deuxième classe comprend les mutations à effet avantageux. Ces mutations sont rapidement fixées, ce qui implique qu'elles sont observables dans la divergence, mais pas (ou peu) dans le polymorphisme (ces propriétés pourront être utilisées plus tard, voir section 3.4). Les séquences protéiques étant proches de leur optimum adaptatif, il y a plus de chance qu'une mutation réduise l'optimalité (mutation délétère) plutôt qu'elle ne l'améliore (mutation avantageuse), rendant la première classe plus abondante que la deuxième. la sélection est alors plutôt vue alors comme une force purificatrice. Enfin, la dégénérescence du code génétique (figure 1.2, sous-section 1.3.3) implique qu'une part non négligeable des mutations dans les gènes sont synonymes (Kimura, 1977; Jukes, 1978), formant la troisième classe de mutations décrite par Kimura. On fait l'hypothèse que ces mutations synonymes sont neutres, c'est-à-dire qu'elles n'ont pas d'effet sur les individus qui les portent [1]. Elles sont donc à priori non impactées par les processus de sélection.

Il est important de préciser ici qu'une mutation correspond à un changement de nucléotide dans la séquence d'un individu. Si celle-ci se répand et se fixe dans la

1. Kimura (1968b) précise que les mutations synonymes ne sont « probablement pas toutes neutres » et que la plupart sont plutôt « presque neutre ».

population, alors on parle d'une substitution. On peut ainsi dire que la plupart des mutations sont délétères, mais qu'on observe surtout des substitutions neutres puisque les premières sont éliminées par la sélection.

2.2.2 Les mutations neutres dérivent aléatoirement

La dérive génétique, initialement formulée par Fisher (1930) et Wright (1931), correspond à la fluctuation au hasard des fréquences alléliques dans une population sexuée de taille finie, due à l'échantillonnage aléatoire des gamètes (voir Kimura (1983), chapitre 3.2 pour une explication détaillée). Quand la population correspond à une population idéale (voir Box 3), l'intensité de cette fluctuation est inversement proportionnelle à la taille de population (N) (voir figure 2.1). Si la population n'est pas idéale, notamment si les appariements ne sont pas aléatoires, on remplace N par la taille efficace de la population, N_e, qui est une unité qui synthétise, dans un cadre simple, les modalités de reproductions entre espèces et permet de comparer l'intensité de la dérive génétique entre espèces (voir section 2.5 pour une discussion plus approfondie du concept). En contexte de dérive génétique, des mutations neutres peuvent envahir la population et se fixer dans celle-ci par simple action du hasard (Kimura, 1968a; King and Jukes, 1969). Il n'est alors plus nécessaire qu'une mutation soit avantageuse pour qu'elle soit fixée. On parle d'évolution non-adaptative, en opposition à la sélection naturelle qui est un processus adaptatif.

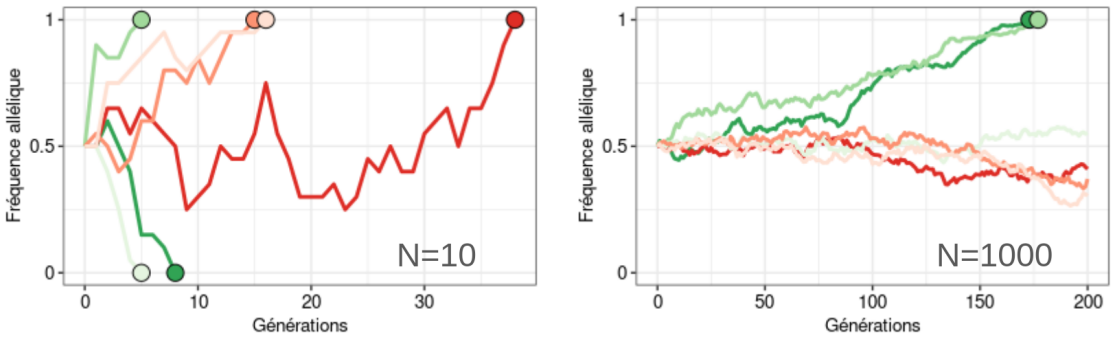


Figure 2.1 – *Simulation de l'évolution aléatoire de variants alléliques neutres dans une population de 10 et 1000 individus. Les allèles dans la petite population sont perdus (fréquence = 0) ou fixés (fréquence = 1) plus rapidement que dans la grande population.*

Étant donné que la majorité des substitutions observées dans les séquences sont neutres et que le mécanisme qui peut engendrer leur fixation est la dérive génétique, Kimura (1968a) développe que la majorité des changements évolutifs inter ou intra espèces, sont dus à l'action de la dérive génétique sur ces mutations neutres. Dans leur article de 1969b, Kimura et Ohta calculent, avec des équations de diffusions, que le temps moyen pour qu'une mutation neutre se fixe dans une population diploïde de taille efficace N_e est $4N_e$ générations, si la mutation n'est pas perdue avant (voir figure 2.2). Ainsi, le processus de dérive génétique aboutissant à fixation d'un allèle neutre se fait sur un temps long et dépend de la taille des populations. L'action de la sélection est, elle, plus rapide. Kimura et Ohta estiment également que le temps moyen entre la fixation de deux substitutions neutres est de 1/v générations avec v le taux de mutation par site par génération (voir figure 2.2).

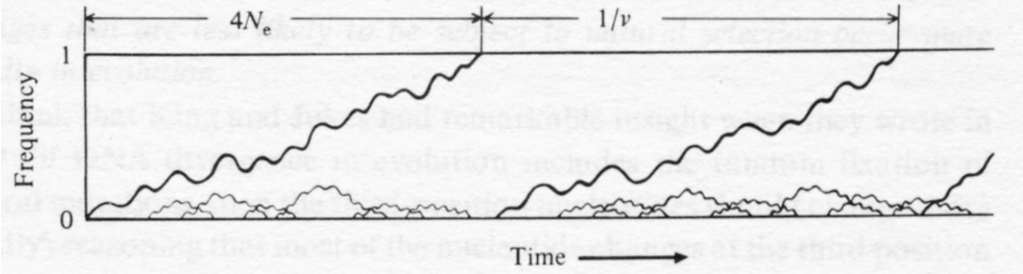


Figure 2.2 – *Représentation schématique de l'évolution par dérive génétique de la fréquence dans la population de différents mutants neutres. Un allèle met en moyenne $4N_e$ générations à se fixer s'il n'est pas perdu avant, avec N_e qui représente la taille efficace de population. Il faut attendre en moyenne $1/v$ générations entre deux fixations de mutation neutre. On observe qu'une grande proportion des mutations sont perdues dû au hasard (Fisher, 1930; Kimura and Ohta, 1969a). Figure extraite de Kimura (1983), figure 3.1.*

La dérive génétique permet d'expliquer le fort taux de polymorphisme observé. En effet, le polymorphisme est le lieu de la transition aléatoire lente d'une mutation neutre vers sa fixation ou sa perte. Il est ainsi une pré-étape de l'évolution moléculaire, qui permet d'expliquer les divergences observées entre espèces (Kimura and Ohta, 1971). Par extension, la divergence et le polymorphisme sont deux aspects d'un même phénomène causé par la dérive aléatoire des allèles neutres en population finie. Autrement dit, la macro-évolution est la résultante de la micro-évolution au sein des populations.

À noter que la théorie neutre n'a pas vocation à remplacer la théorie de la sélection naturelle. Elle diminue seulement son rôle au profit d'autres mécanismes

non adaptatifs.

Plus tard, le séquençage du génome humain a révélé que la proportion de gènes dans notre ADN est très faible (cf. Box 2). Ainsi, au moins chez les mammifères et à l'échelle du génome entier, la plupart des mutations sont neutres car elles touchent peu fréquemment un gène, ce qui laisse beaucoup de terrain à la dérive génétique. Inversement, les mutations impactantes pour la survie de l'individu sont plutôt rares.

Box 2 : Seulement 1 % du génome humain est codant !

Le séquençage du génome humain a pris un demi-siècle (1950-2003) et a apporté son lot de surprises.

En effet, en héritage des pensées d'Aristote (section 1.1), on imagine que plus les organismes sont complexes, plus leur génome est grand et contient beaucoup de gènes. Dans ce contexte, l'humain est vu comme le plus complexe des organismes.

Or, on a déterminé que le génome humain contient environ 21 000 gènes, ce qui est peu en comparaison aux 6,7 millions imaginés en 1964 et moins que les 37 500 gènes du riz.

Mais surtout, ces 21 000 gènes correspondent en réalité à seulement 1 % du génome ! On découvre ainsi que l'ADN n'est pas fait d'un enchainement de gènes, mais plutôt d'une grande portion de séquences qui ne codent pas de caractères phénotypiques, parmi lesquels sont dispersés les gènes.

Si les processus adaptatifs n'ont lieu que dans les séquences codantes, alors ces régions non-codantes, encore appelées « *junk DNA* » ou « ADN poubelle », ne sont soumises qu'à des processus non-adaptatifs tels que la dérive génétique. Ces régions peuvent alors servir de base de comparaison pour détecter l'effet de la sélection dans des séquences codantes, soumises à dérive et sélection.

2.2.3 Les positions les moins contraintes évoluent plus vite

Dans un contexte où l'évolution est basée, pour l'essentiel, sur la conservation de ce qui est important, par sélection purifiante, et, l'évolution aléatoire du reste, on peut faire l'hypothèse que les positions les moins importantes pour les séquences protéiques évoluent plus vite, car elles ne sont pas soumises à la sélection qui les stabilise (ont dit qu'elles sont moins contraintes). On suppose donc une relation négative entre le taux d'évolution et le niveau de contrainte des séquences. On s'attend notamment à ce que les nucléotides en troisième position de codon changent plus fréquemment (King and Jukes, 1969).

Les données moléculaires ont permis de confirmer, à l'échelle du polymorphisme et de la divergence, ces attendus théoriques. On observe également que les nucléotides en deuxième position de codon évoluent moins vite que ceux en première position, la troisième position étant, effectivement, celle qui évolue le plus vite. L'évolution plus rapide des premières positions par rapport aux deuxièmes est due au fait que certaines mutations aux premières positions de codon sont, elles aussi, parfois synonymes tandis qu'aucune mutation en deuxième position ne l'est (figure 1.2). De plus, les substitutions d'acide aminé semblent se produire moins fréquemment dans les régions fonctionnelles importantes des protéines que dans les régions moins importantes (Margoliash and Smith, 1965; Zuckerkandl and Pauling, 1965a).

Plus particulièrement, chez les eucaryotes, on découvre que les gènes contiennent des introns (Crick, 1979), c'est-à-dire des morceaux de séquences qui ne participent par à la formation de la protéine. Ceux-ci présentent un taux d'évolution particulièrement élevé (Van den Berg *et al.*, 1978).

En élargissant les observations aux régions non-codantes de l'ADN et aux pseudogènes (des anciens gènes qui ne sont plus fonctionnels), on réalise également que le taux d'évolution des pseudogènes est plus élevé que celui des gènes et comparable à celui des troisièmes positions de codons. Ainsi, à l'échelle du génome entier, moins une position est contrainte (c'est-à-dire, moins elle est essentielle pour la survie), plus elle évolue vite. Cette affirmation est en contradiction avec les

attendus néo-darwiniens qui affirment que les régions évoluant le plus vite sont celles qui sont soumises à la sélection la plus forte.

2.3 L'évolution moléculaire semble se faire à vitesse constante

Dans la section précédente, nous avons pu constater que différents compartiments génomiques peuvent évoluer à des vitesses différentes en fonction du niveau de contrainte appliqué. Plus largement, l'estimation et l'explication des vitesses d'évolution observées au niveau phénotypique et moléculaire ont largement alimenté les débats entre sélectionnistes et neutralistes et ont permis de renforcer la théorie neutre de l'évolution.

2.3.1 Des vitesses d'évolution phénotypique liées aux variations de l'environnement

Au niveau du phénotype, on observe que l'évolution semble se faire en réponse à des changements d'environnement. On constate notamment plusieurs cas d'évolution convergente quand il est nécessaire de répondre à une même problématique environnementale (par exemple, l'apparition multiple de la capacité de voler ou de l'écholocation [2]). Ces cas de convergence semblent montrer que l'évolution des formes et des fonctions est principalement gouvernée par la sélection naturelle qui engendre l'adaptation des organismes à leur environnement (mais voir Losos (2011)).

L'environnement étant variable [3], on attend également une forte variation des taux d'évolution phénotypiques dans le temps. Cette variation avait déjà été observée par Darwin dans l'*Origine des espèces* (chapitre 10, 1859) où il décrit notamment que les espèces terrestres semblent se transformer plus vite que les espèces marines. On proposera plus tard que cela pourrait être dû au fait que les environnements marins sont moins soumis aux variations périodiques de facteurs,

2. D'autres exemples : https://fr.wikipedia.org/wiki/Convergence_%C3%A9volutive, consultÃ Ile18/09/24

3. Variable en lui-même, mais aussi variable dans son intensité de variation et sa prédictibilité (variations cycliques ou non)

comme la température ou l'humidité, et sont donc moins variables, nécessitant moins d'adaptation, ce qui entraine une vitesse d'évolution plus faible (Rensch, 1959). De façon particulièrement impressionnante, on note le cas des organismes appelés « fossiles vivants » qui ressemblent très fortement à des organismes déjà présents il y a des millions d'années. Leur taux d'évolution phénotypique est donc très faible. Par exemple, la *Lingula* est un brachiopode marin qui ne semble pas avoir changé de morphologie par rapport à son ancêtre il y a 400 millions d'années.

De nombreuses analyses des taux d'évolutions phénotypiques sont réalisées par Simpson (1944; 1953) à partir de fossiles. Il utilise par exemple la longueur des dents d'une lignée de chevaux, de leurs ancêtres dans l'Éocène inférieur à aujourd'hui, pour calculer la vitesse moyenne d'évolution de cette lignée. Ces métriques morphologiques ont le défaut d'être taxon-spécifique et réservées à des experts[4]. Pour autant, une partie des scientifiques sont d'accord sur la nécessité d'une échelle moins subjective et applicable à une plus grande étendue d'espèces, en témoigne Simpson (1953) qui écrit qu'un « taux génétique » serait une mesure idéale de vitesse évolutive.

2.3.2 Des vitesses d'évolution moléculaires aussi constantes qu'une horloge

Au niveau de l'évolution moléculaire, les travaux de Pauling et son élève Zuckerkandl sur les molécules d'hémoglobine sont d'une grande importance (Zuckerkandl, 1987; Morgan, 1998). Initialement, Zuckerkandl et Pauling essaient simplement de trouver un marqueur moléculaire en supposant que ce marqueur évolue en accord avec les taux morphologique précédemment définis. Pour cela, ils ont observé des profils de migration sur gel de molécules d'hémoglobine extraite de plusieurs mammifères. Ils constatent qualitativement un lien entre la similarité des profils et la distance phylogénétique des espèces étudiées (Zuckerkandl *et al.*, 1960). Une fois les différentes hémoglobines séquencées (Zuckerkandl and Schroeder, 1961), ils ont dénombré quantitativement les différences entre chaque séquence et ont utilisé cette quantité pour estimer le temps de divergence entre

4. Simpson écrit en 1953 qu' « il est possible qu'une différence taxonomique fondée sur le jugement de chercheurs expérimentés soit une indication plus sûre de la totalité des changements évolutifs que toutes les collections de mesures des différences de caractères morphologiques ». (Simpson, 1953)

chaque paire de séquences (Zuckerkandl, 1962). Ils appliquent ainsi la première « horloge moléculaire » de l'histoire, sans vraiment la nommer (voir figure 1). En effet, utiliser les différences moléculaires pour estimer le temps de divergence entre deux séquences revient à supposer que le taux d'évolution est constant et régulier, tel le tic-tac d'une horloge. Par analogie avec une horloge qui mesure le temps par intervalles constants, l'horloge moléculaire, elle, mesure le temps de divergence entre séquences par l'accumulation progressive de mutations dans l'ADN. Pour calibrer cette horloge, Zuckerkandl et Pauling utilisent les chaines alpha de l'hémoglobine du cheval et de l'humain et la datation paléontologique de l'ancêtre de ces deux espèces estimée entre 100 et 160 millions années. Les deux protéines présentant en tout 18 substitutions pour 150 acides aminés, ils estiment qu'il se produit en moyennes pour ces séquences une substitution par acide aminé, tous les 14.5 millions d'années. Ils comparent ensuite les séquences d'hémoglobine de l'humain et du gorille et utilisent l'horloge précédemment calibrée pour dater la divergence entre ces deux espèces qu'ils estiment à 11 millions d'années, ce qui est consistant avec les relevés paléontologiques.

Dans *The neutral theory of molecular evolution*, Kimura (1983) montre également le lien, visiblement constant, entre le pourcentage d'acides aminés différents d'une séquence d'hémoglobine à une autre, chez six vertébrés et leur âge de divergence (figure 2.3).

Les mêmes études sont réalisées sur d'autres alignements de gènes (Dayhoff, 1978), et permettent de confirmer l'existence d'un taux d'évolution moléculaire constant le long de la phylogénie, mais qui est, en plus, différent pour chaque famille de gènes. Par exemple, les globines évoluent à un taux uniforme de 10^{-9} substitutions par site et par an, en revanche les insulines ont un taux de 0.44×10^{-9} substitutions par site et par an (Dayhoff, 1978). La découverte que certains gènes évoluent plus lentement que d'autre permet de réaliser des phylogénies sur différentes échelles de temps sans être trop impacté par les phénomènes de saturation. Par exemple, le gène de l'ARN 18s, évolue plutôt lentement et permet de reconstruire la phylogénie des eucaryotes (voir le premier arbre du vivant reconstruit à partir de l'ARN 16s et 18s par Woese (1987), figure 1.1), tandis que le cytochrome b évolue plus rapidement et permet de reconstruire l'histoire des mammifères.

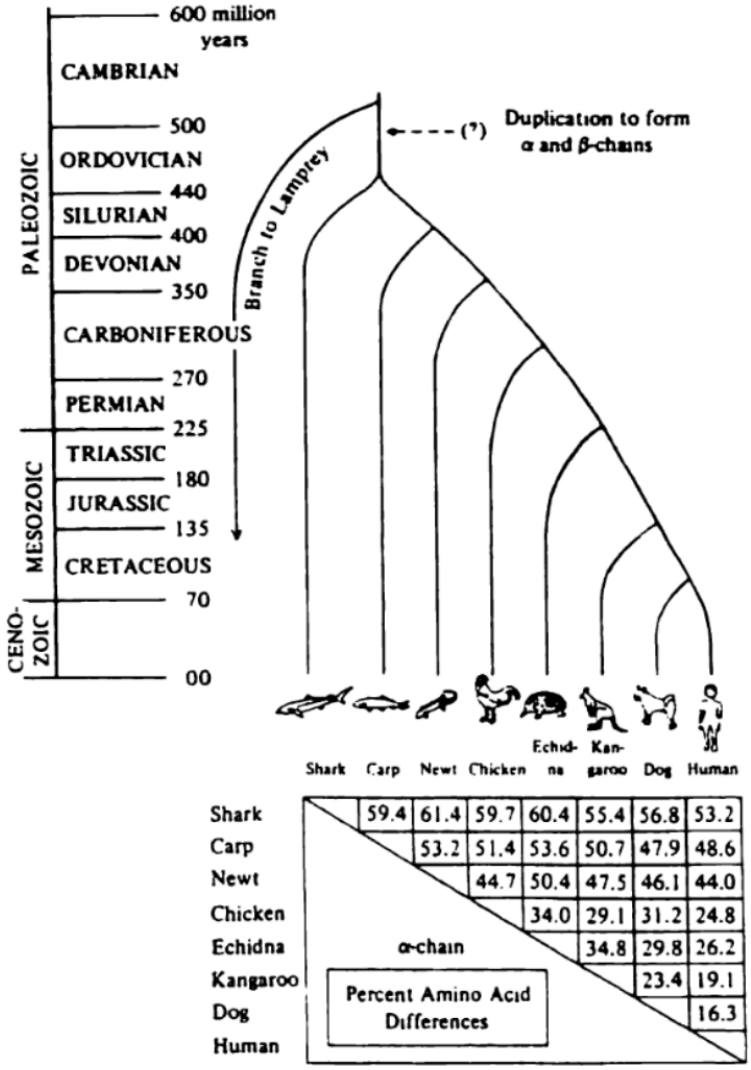


Figure 2.3 – *Relation entre le taux de substitution des alpha-globines et leur phylogénie. La phylogénie représente les relations évolutives entre 8 espèces de vertébrés. Le tableau comptabilise le pourcentage d'acides aminés différents entre chaque paire d'espèces, pour les séquences alpha d'hémoglobine. Plus deux espèces sont éloignées sur la phylogénie et plus le pourcentage est élevé. Figure modifiée à partir de Kimura (1983).*

Le développement de l'horloge moléculaire par Zuckerkandl et Pauling (Zuckerkandl, 1987) est réalisé dans un contexte de validation et démonstration du néo-darwinisme au niveau moléculaire. Pour autant, les résultats obtenus sont en contradiction avec l'évolution non-constante observée au niveau des phénotypes. De plus, on découvrira peu après que ces travaux ont plutôt permis de consolider les bases de la théorie neutre de l'évolution [5] (Kimura, 1983) (voir

5. Dans une interview Zuckerkandl dit « Pour moi, la sélection naturelle n'a jamais été en contradiction avec l'horloge moléculaire, mais pour Kimura l'horloge était un des fondements de la dérive neutre ».

section 2.2 et sous-section 2.3.3).

Quoi qu'il en soit, la découverte de la constance apparente des taux d'évolution et donc de l'horloge moléculaire sera décrite comme « *one of the simplest and most powerful concept in the field of evolution* » [6] (Lewin, 1996).

2.3.3 Expliquer l'horloge moléculaire par la théorie neutre

L'horloge moléculaire, et donc la constance des taux d'évolution moléculaires, semble incompatible avec une sélection adaptative dominante au niveau phénotypique où le taux d'évolution est lié aux changements environnementaux qui ne se font pas à taux constant. À l'inverse, la théorie neutre semble plus à même, en tout cas en première approximation, d'expliquer cette observation empirique. Dans ce qui suit, ce point est formalisé plus précisément en suivant le raisonnement de Kimura (1983).

Si l'on développe mathématiquement le calcul du taux de substitution (noté k) en utilisant un modèle à nombre infini de sites (chaque mutation est unique) (Kimura, 1971), on trouve que :

$$k = 2Nv \cdot P_{fix} \tag{2.1}$$

Avec N la taille de la population, $2Nv$ le taux de mutation par gamète et par an, à l'échelle de la population et P_{fix} la probabilité de fixation d'une mutation dans la population. Dans ce développement, le taux de substitution s'écrit en nombre de substitutions par site et par an. Dans un contexte où la sélection positive est majoritaire, la probabilité de fixation d'une mutation avec un coefficient de sélection (noté s), tel que $4N_e s \gg 1$, est $P_{fix} = 2s.Ne/N$ (Kimura and Crow, 1964; Kimura, 1983) avec $\frac{N_e}{N}$ qui correspond à la probabilité d'apparition d'une mutation dans un individu reproducteur parmi tous les individus et N_e la taille efficace de la population (section 2.5). Ainsi, le taux de substitution correspond à :

$$k = 2Nv.\frac{2s.Ne}{N} = S.v \tag{2.2}$$

6. « Un des plus simples et des plus puissants concepts dans le domaine de l'évolution »

Où $S = 4N_e s$. Sachant qu'il existe d'autres mutations que les mutations avantageuses, on définit f_a la proportion de mutations avantageuses ayant un coefficient de sélection moyen s_a. Les mutations délétères sont présentes en proportion $1 - f_a$ et ont une probabilité de fixation nulle. On écrit alors que :

$$\begin{aligned} k_a &= 2Nv.P_{fix} \\ &= 2Nv.[(1-f_a).0 + f_a.\frac{2s_aN_e}{N}] \\ &= v.f_a.S_a \end{aligned} \tag{2.3}$$

Dans ce cas, obtenir k constant, comme observé empiriquement dans les séquences, impliquerait que le produit de $v.f_a.S_a$ est constant, ce qui est fortement improbable. En effet, même si v est constant (également supposé constant en théorie neutre), N_e et f_a varient dès qu'on change d'environnement, impliquant un k plus élevé pour une espèce dont l'environnement vient de changer, par rapport à une espèce dans un environnement stable.

En revanche, en contexte neutre, la probabilité de fixation d'une mutation neutre est $\frac{1}{2N}$. Une mutation est considérée comme neutre tant que son coefficient de sélection, s est inférieur à $\frac{1}{2N_e}$ (Kimura, 1968b). On dénote directement la proportion de mutations neutres, par f_0. Ainsi le taux de substitution neutre k_0 s'écrit :

$$k_0 = 2Nv.\frac{1}{2N}.f_0 = v.f_0 \tag{2.4}$$

Le taux de substitution neutre par an, dans ce cas, correspond uniquement au taux de mutation, qui est supposé constant, multiplié à la fraction de mutations neutres. Sous l'hypothèse que la fonction biochimique ou enzymatique de la plupart des gènes ne change que très peu entre espèces, il est raisonnable de supposer que cette fraction de mutations neutres est elle-même contrainte et donc plutôt constante. Cet argument, développé sur les séquences codantes implique un taux total de substitution dans les gènes tel que $k = k_0 + k_a = v.f_0 + v.f_a.S_a$ (équation 2.3 et équation 2.4). Cependant, sous l'hypothèse que les mutations adaptatives sont négligeables et/ou ne sont pas soumises à une sélection très forte, et qu'en dehors des gènes la plupart des substitutions sont neutres, alors on peut considérer que $f_a.S_a \ll f_0$ et donc que $k \approx k_0$. Finalement, la théorie neutre semble proposer une

explication plus satisfaisante de l'horloge moléculaire. Par ailleurs, il devient clair que l'évolution phénotypique et moléculaire répondent à des mécanismes très différents. Dans le premier cas, c'est l'adaptation à des environnements changeants qui domine, tandis que dans le second cas, l'évolution se fait, avant tout, par conservation des séquences les plus importantes et évolution aléatoire par dérive du reste à un taux constant (Kimura, 1983).

À ce stade, une incohérence persiste. En effet, l'ensemble du raisonnement repose sur l'hypothèse que le taux de mutations par année reste constant. Or, rien ne permet de penser que cette hypothèse est valide. Ce point sera rediscuté plus tard par Ohta (voir sous-section 2.4.3). Globalement, la théorie neutre permet une avancée dans la compréhension des mécanismes de l'évolution par rapport à la théorie synthétique, mais elle ne résout cependant pas toutes les questions.

Il faut également préciser que l'explication de l'horloge moléculaire par la théorie neutre de l'évolution ne supprime pas la part d'évolution adaptative qui a eu lieu dans les gènes. Par exemple, les séquences de globines ont certainement vécu une forte pression de sélection adaptative pour passer d'une forme monomérique à une forme tétramérique fonctionnelle partagée par un grand ensemble de taxons (Dickerson and Geis, 1969). Cependant, ces substitutions avantageuses sont présentes en quantité inférieure par rapport aux substitutions neutres qui expliquent le « tic » régulier de l'horloge moléculaire.

2.4 Une mutation délétère peut-être considérée comme (quasi)-neutre

2.4.1 Neutre ou délétère ? Plutôt un continuum d'effet

Au fur et à mesure du développement de la théorie neutre de l'évolution, Kimura et Ohta se rendent compte qu'il existe plus de classes de mutations que seulement les trois classes discrètes « avantageuses », « neutres » et « délétères » précédemment définies. Il existe en réalité une continuité dans les effets des mutations. Cela permet l'existence de mutations, par exemple, faiblement délétères (Ohta, 1973). Ces dernières, bien que

soumises à la sélection naturelle, mettent néanmoins plus de temps à être détectées par celle-ci et peuvent, entre temps, être fixées ou éliminées par la dérive, comme les mutations neutres. Kimura a d'ailleurs entamé cette réflexion en mentionnant que pour qu'un allèle dépende principalement de la dérive génétique, celui-ci n'a pas besoin d'être strictement neutre mais seulement, que son coefficient de sélection corresponde à $|s| \ll 1/2N_e$ (Kimura, 1968b).

Poursuivant cette réflexion, Ohta propose une extension de la théorie neutre en théorie quasi-neutre de l'évolution (Ohta, 1973, 1974), dans laquelle des mutations faiblement délétères peuvent passer sous le radar de la sélection purifiante et donc, évoluer dans la population comme des mutations neutres. On appelle ces mutations, des mutations « quasi-neutres ». Ohta formalise mathématiquement sa théorie en incluant les mutations faiblement délétères dans ses équations (Ohta, 1973).

À ce stade, on suppose que la répartition des coefficients de sélection se fait suivant une distribution nommée DFE (*Distribution of Fitness Effect* ou distribution des effets sur la fitness) qui représente, pour chaque coefficient de sélection, la fréquence associée de mutations dans les génomes (figure 2.4). Cette distribution a tendance à former un L qu'on modélise par une loi gamma entièrement spécifiée par un paramètre de forme, β, et une moyenne (Keightley and Eyre-Walker, 2007; Eyre-Walker and Keightley, 2007). Cette forme leptokurtique décrit que la majorité des substitutions sont à faible, ou a très faible, effet et qu'une minorité sont à fort, voire à très fort, effet. La fraction de mutations à effet intermédiaire est, quant à elle, assez faible (voir figure 2.4 courbe bleue).

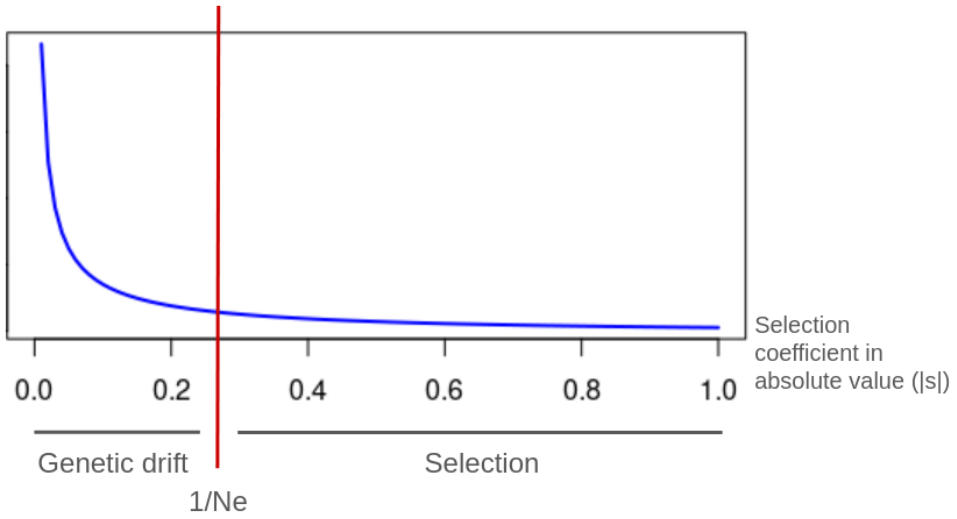


Figure 2.4 – *Distribution théorique de la fréquence des coefficients de sélection en valeur absolue (délétères et avantageuses sont mélangées). La barre verticale rouge représente le seuil 1/Ne qui délimite l'effet dominant de la sélection ou de la dérive sur ces mutations.*

2.4.2 L'efficacité de la sélection purifiante est déterminée par N_e

Une fois la distribution des coefficients de sélection établie, on détermine, pour chaque mutation, la force évolutive majeure qui va régir son évolution. Kimura (1968b) puis Ohta (1973) décrivent que, par défaut, une mutation est soumise à l'action de la dérive génétique, mais si son coefficient de sélection est supérieur à $1/N_e$, alors la sélection purificatrice sera plus efficace que la dérive et pourra éliminer la mutation potentiellement délétère avant sa fixation potentielle [7] (voir figure 2.4). Ainsi, la dérive aléatoire ou non de mutations faiblement à moyennement délétères va dépendre de la taille efficace des populations, N_e (voir section 2.5). Dans ce contexte, N_e devient une mesure quantitative directe de la dérive génétique. Celle-ci diminue en intensité si N_e augmente (figure 2.1). On fait donc l'hypothèse d'une relation entre efficacité de la sélection et taille efficace de population. Pour une mutation faiblement délétère avec un coefficient de sélection donné, les espèces à grand N_e pourront l'éliminer, tandis que les plus petites n'en auront pas le temps, car la dérive génétique sera plus efficace. C'est pour cela que l'on observe plus de

7. Ou la fixer par sélection positive avant sa perte potentielle, mais ce sont des cas rares

substitutions chez des groupes d'espèces à faible N_e comme les mammifères, à l'inverse des groupes d'espèces a fort N_e comme les drosophiles. Les espèces à petite taille de population vont ainsi cumuler un fardeau de mutation faiblement délétère [8]. À noter que, puisque la fraction des mutations à effet intermédiaire est assez petite (voir ci-dessus), la théorie prédit que ces variations restent assez faibles en fonction de N_e. Autrement dit, la théorie quasi-neutre ne fait qu'apporter une correction plutôt modérée à la théorie neutre initialement proposée par Kimura.

2.4.3 L'horloge moléculaire en péril

À l'origine, la théorie quasi-neutre vise à résoudre l'incohérence mentionnée en sous-section 2.3.3 selon laquelle le taux de mutation par année (v) n'est en réalité pas constant, et donc à sauver l'horloge moléculaire présentée en sous-section 2.3.2. L'argument utilisé est que le taux de mutation par année est plus faible chez les espèces à temps de génération long (le nombre de divisions de cellules sexuelles par an est plus faible) mais que ces mêmes espèces ont un N_e plus petit et donc une proportion de substitution neutre, f_0, plus grande. L'un dans l'autre, v et f_0 se compensent, ce qui permet au taux de substitution par site et par an de rester constant.

Cependant, cette compensation n'est que partielle chez les mammifères. On observe un fort effet du temps de génération (qui n'est peut-être même pas lui-même constant) sur le taux de substitution des positions synonymes (d_S) (Ohta, 1993). L'horloge semble plus stable en utilisant le taux de substitution des positions non-synonymes (d_N) (Ohta, 1993). Plus tard, l'hypothèse de l'horloge moléculaire sera révisée pour développer « l'horloge moléculaire relaxée » (voir sous-section 3.2.2) [9].

2.5 What the hell is N_e ?

8. Les études de conservation montrent que les espèces en danger ont un fort fardeau mutationnel et un petit N_e.

9. Finalement, les théories neutre et quasi neutre sont nées d'une tentative d'explication d'un phénomène, l'horloge moléculaire, qui finalement n'existe pas.

2.5.1 Les différentes « flavours » de N_e

Dans un article de 2022 [10], Waples (2022) témoigne d'un échange qu'il a eu avec un collègue à propos de la définition de la taille efficace de population, N_e. Pour Waples, N_e détermine le taux de dérive génétique dans une population (Waples, 2016). Pour son collègue, c'est l'inverse, le taux de dérive génétique détermine N_e. Cet échange témoigne d'une confusion générale dans le domaine à propos de la définition de N_e.

N_e est une entité conceptuelle qui a été introduite par Wright (1931) dans son modèle de population idéale (cf Box 3). Dans ce contexte, N_e est définie comme la taille d'une population idéale panmictique, de taille constante, non soumise à la sélection naturelle et qui expérimente la même quantité de dérive que la population réelle. C'est donc un nombre proche de N, la taille de population, qui essaie de paramétriser la dynamique d'un processus populationnel complexe (Waples, 2022). On distingue l'usage de N et de N_e dans les équations quand on aborde respectivement des processus écologiques ou évolutifs. Ainsi, on utilise N_e dans chaque équation impliquant de la dérive génétique. La taille efficace de population offre la possibilité de comparer des populations ou des espèces dans un même cadre concernant la quantité de dérive génétique attendue, sous un modèle de Wright-Fisher (cf Box 3). Cependant, il y a différentes façons de mettre N_e en équation qui impliquent chacune différentes hypothèses et donc, par extension, différentes sous définitions de N_e qui ne sont pas toujours compatibles. C'est cet ensemble de définitions qui rend le concept de N_e si difficile à saisir.

Les N_e contemporain

Parmi l'ensemble des formalisations possibles de N_e, on retrouve les équations basées sur l'évolution de populations sur une à quelques générations. On dit alors qu'on estime des N_e contemporains. Crow et Kimura définissent deux types de N_e contemporains qui sont le *inbreeding* N_e (ou N_e^i) et le *variance* N_e (ou N_e^v) (Crow *et al.*, 1954; Kimura and Crow, 1963).

10. Le titre de cette section est très inspiré du « What is N_e, anyway ? » de Waples (2022), de même que pour les « flavours » de N_e.

N_e^i se distingue par la considération du nombre d'individus dans la génération parentale. En terme d'équation, on définit la probabilité P qu'une paire de gènes homologues vienne du même parent tel que $P = 1/N_e^i$, et donc :

$$N_e^i = 1/P \tag{2.5}$$

En ce qui concerne N_e^v, on considère plutôt le nombre d'individus dans la génération des descendants, noté p_0. On obtient mathématiquement que la variance des fréquences alléliques V correspond à $V = p_0(1 - p_0)/2N_e^v$ et donc :

$$N_e^v = \frac{p_0(1 - p_0)}{2V} \tag{2.6}$$

Comme l'expliquent Crow et Kimura, N_e est la taille d'une population idéale qui expérimente la même quantité de consanguinité - d'*inbreeding* - (N_e^i) ou de dérive (N_e^v) que la population en question. Ainsi N_e^i et N_e^v sont supposés être identiques dans un contexte de population de taille constante et isolée (Crow *et al.*, 1954; Kimura and Crow, 1963). Cependant, il est assez facile d'imaginer des cas où la consanguinité et la variance reproductive sont partiellement découplées, et où ces deux N_e contemporains prennent alors des valeurs différentes. Par exemple, dans une population contenant un unique individu hétérozygote qui s'auto-féconde et produit beaucoup de descendants, on s'attend à ce que N_e^i soit plus petit que N_e^v (Waples, 2022). De plus, on sait que N_e^v est plus grand que N_e^i quand la population est en expansion (Crow and Morton, 1955).

Les N_e basés sur le coalescent

En plus des N_e contemporains, on distingue des N_e de plus long terme, nommés ici N_e coalescents [11] (ou N_e^c). Celui-ci est associé à l'équation

$$\theta = 4N_e^c\mu \tag{2.7}$$

avec θ qui représente la quantité de diversité dans une population idéale (voir section 3.3). C'est plutôt ce type de N_e qui est utilisé en évolution moléculaire (Sjodin *et al.*, 2005) et qui sera utilisé dans mes travaux de thèse lorsqu'il sera question d'étudier les mécanismes évolutifs à l'échelle des populations

11. Je préfère réserver l'utilisation de « long terme » à un autre type de N_e, même si l'usage est plutôt d'appeler ce N_e coalescent un N_e de long terme.

(ou « micro-évolution ») (section 3.3, chapitre 5 et chapitre 6).

Les N_e de "long terme"

Enfin, on peut tenter de définir un N_e à l'échelle de la divergence[12] (N_e^l). Ce N_e^l ne résulte pas d'une équation décrivant la dérive génétique, mais découle plutôt de l'observation d'une relation entre les traits d'histoire de vie et la taille des populations. Par exemple, les petits mammifères, tels que les mulots, ont une longévité et un temps de génération plus courts que les grands mammifères, comme les éléphants, mais aussi des populations plus grandes. La taille de population est, elle-même, liée à N_e. Ainsi, on peut utiliser les variations des traits d'histoire de vie le long d'une phylogénie pour approximer les variations de N_e^l le long de cette même phylogénie. Dans mes travaux de recherche (chapitre 5), j'utilise ce N_e^l lorsqu'il est question d'étudier les mécanismes évolutifs à l'échelle de la comparaison des séquences entre espèces (ou « macro-évolution »).

Quoi qu'il en soit, comme la grande quantité de définitions de N_e le montre, il est légitime qu'il y ait une confusion latente concernant sa « vraie » signification et il ne semble pas y en avoir une meilleure qu'une autre. Un garde-fou efficace est de prendre la peine de définir le N_e que l'on utilise pour éviter toute confusion, et de ne pas oublier son lien avec le modèle de Wright-Fisher[13]. En revanche, il est incorrect de dire que N_e correspond simplement au nombre d'individus qui se reproduisent. C'est une trop forte simplification du concept général de N_e qui ne peut être vraie qu'en population de Wright-Fisher, et qui laisse à penser, à tort, que N_e peut être vu comme une entité physique correspondant à un nombre réel d'individus.

12. C'est celui-ci auquel je me réfère comme N_e de long terme car il représente un temps évolutif encore plus long que N_e^c.
13. 'It would be more indicative of the meaning of the concept if the adjective « effective » were replaced by « in some given respect Wright-Fisher model equivalent » (Ewens 2004, p38)

Box 3 : Le modèle de population idéale de Wright-Fisher

Idéalement, si une population est infiniment grande et ne subit pas de sélections, mutations ou migrations, alors la fréquence des variants est censée rester constante dans le temps (Hardy, 1908; Weinberg, 1908). En revanche, dans un contexte de taille de population finie (N), il y a fluctuation stochastique des fréquences alléliques due à la dérive génétique dont l'effet est inversement proportionnel à la taille de population.

Le modèle mathématique, proposé par Wright (1931) et Fisher (1930), décrit comment les individus d'une population, dite idéale, se remplacent d'une génération à l'autre. La population idéale doit pour cela correspondre aux critères suivants :

- la taille de population est finie et constante ;
- tous les individus ont une même chance de se reproduire ;
- la reproduction est aléatoire avec remise ;
- chaque individu possède deux copies de chaque gène ;
- tous les individus d'une génération sont remplacés par la génération suivante ;
- il n'y a ni sélection, ni mutation, ni migration.

Ce modèle permet de comprendre l'impact de la dérive génétique dans des populations de petite taille. Il sert de base pour des modèles plus complexes qui incorporent des mécanismes comme la sélection.

L'échantillonnage des allèles dans la population suit une loi binomiale $B(2N, i/2N)$ avec $2N$ le nombre de copies de séquences et $i/2N$ la fréquence d'un des deux allèles d'un gène bi-allélique. Plus N est petit et plus les fréquences alléliques vont varier rapidement, ce qui augmente l'intensité de la dérive génétique.

2.5.2 Variation de N_e entre espèces : le paradoxe de Lewontin

Comme abordé dans les sections précédentes, la diversité neutre présente dans les populations est le résultat d'un équilibre entre l'arrivée de variants neutres par mutation (proportionnelle à μ, le taux de mutation neutre par site, par génération) et leur élimination par dérive génétique (inversement proportionnelle à N_e). Cette dynamique est décrite par la formulation [14] présentée en équation 2.7. On estime de fait que la variation de la quantité de diversité synonyme entre espèces est une approximation de la variation de N_e entre ces mêmes espèces (tant que μ reste constant). Par exemple, en population de grande taille efficace, la diversité neutre est plus grande, car la dérive génétique est moins efficace. On le voit notamment dans la figure 2.1 où il y a plus de variants présents en même temps, plus longtemps, dans la population de 1000 individus par rapport à celle de 10 individus. On s'attend également à ce que l'amplitude des variations de la diversité neutre (et donc de N_e) reflète celle des tailles de population (N).

Pour autant, Lewontin and Hubby (1966); Lewontin (1974) observent que le niveau de variation de la diversité neutre dans les populations est bien inférieur à celui de N. Cette observation est, depuis, appelée le *Paradoxe de Lewontin*. Dans une étude récente, des chercheurs et chercheuses ont compilé les estimations de diversité génétique neutre de 167 espèces, réparties sur 14 phylums, et ont observé une diversité neutre, bornée entre 0,01 % pour *Lynx lynx* et 8,01 % pour *Ciona savignyi* (Leffler *et al.*, 2012). Or, on sait que les tailles de populations, même si difficiles à estimer, varient bien plus entre espèces (Buffalo, 2021), ce qui confirme l'existence d'un tel paradoxe. Ainsi, ajouter plus d'espèces ou améliorer les techniques d'études ne résout pas le paradoxe de Lewontin (Leffler *et al.*, 2012). C'est seulement une compréhension plus fine des différents mécanismes qui en sera la clé.

Les prochains paragraphes n'ont pas pour but de proposer une revue complète des avancées concernant ce paradoxe formulé par Lewontin, d'autant plus qu'il en existe déjà plusieurs dont celle de Leffler *et al.* (2012), mais aussi celles dans les

14. On utilise ici le N_e^c de la section précédente.

introductions des articles de Romiguier *et al.* (2014a); Corbett-Detig *et al.* (2015); Galtier and Rousselle (2020); Buffalo (2021); Charlesworth and Jensen (2022). L'objectif est plutôt de tirer parti de ce contexte pour énumérer rapidement les différents mécanismes qui pourraient contraindre le niveau de variation de la diversité neutre (largement réutilisé dans les prochains chapitres) et ainsi rendre compte de la complexité des processus liés à N_e.

Des exemples de processus qui réduisent la diversité neutre

Selon l'équation 2.7, la mutation augmente la diversité génétique par ajout de variants dans la population tandis que la dérive la diminue en fixant ou éliminant ces mêmes variants. Par conséquence, plus la taille de population est faible et plus il y a de dérive génétique et donc de réduction de la diversité (Fisher, 1930; Wright, 1931; Kimura, 1983). De plus, on évoque parfois une corrélation potentiellement négative entre N et μ, menant à une réduction de diversité pour les populations de grande taille due à un afflux diminué de mutations nouvelles (Lynch, 2010, 2011). Cependant, cette relation entre N et μ, reste discutable (Yoder and Tiley, 2021; Krasovec *et al.*, 2020).

Il existe également des effets indirects de la sélection qu'on appelle sélection liée ou effet auto-stop (Hill and Robertson, 1966; Smith *et al.*, 1974; Gillespie, 1991, 2000). Concrètement, quand une mutation avantageuse est sélectionnée positivement, celle-ci entraine indirectement, avec elle, les variants neutres qui lui sont liés physiquement dans la séquence (Smith *et al.*, 1974). Il en est de même pour les mutations délétères qui sont purifiées et qui entrainent, avec elles, l'élimination des mutations neutres autour (Charlesworth *et al.*, 1993). Cet effet auto-stop engendre ainsi une diminution de la diversité neutre autour des sites sous sélection (Aguade *et al.*, 1989; Begun and Aquadro, 1992; Charlesworth *et al.*, 1993; Corbett-Detig *et al.*, 2015; Kern and Hahn, 2018). En revanche, la recombinaison génétique est un processus qui entraine le découplage physique des différentes positions nucléotidiques entre elles et donc casse les liaison entre sites. On observe, de fait, dans les régions du génome ou la recombinaison est de forte intensité, une diversité neutre plus importante qu'ailleurs (Sella *et al.*, 2009; Cai *et al.*, 2009). Certaines théories vont plus loin et proposent que l'intensité de la sélection liée est proportionnelle à la taille de population (Chen *et al.*, 2020). Ce qui mène à une

relation négative entre N et diversité neutre expliquant le découplage observé entre ces deux métriques [15].

La diversité neutre est bornée

En plus des processus qui diminuent la diversité neutre, il peut exister des contraintes fonctionnelles ou structurelles qui empêchent le niveau de diversité d'être trop élevé ou trop faible. Par exemple, une diversité trop forte pourrait empêcher les appariements entre chromosomes lors de la méiose (Stephan and Langley, 1992) ou bien engendrer une incompatibilité lors de la reproduction (Seidel *et al.*, 2008). Dans l'autre sens, en dessous d'un certain seuil de diversité, la population ne pourrait potentiellement plus maintenir la variabilité nécessaire pour répondre aux changements environnementaux, ce qui mènerait à une extinction rapide de l'espèce (Lynch and Lande, 1998).

Toutes les propositions précédemment évoquées s'accordent pour dire qu'il existe des mécanismes qui contraignent la diversité neutre à une borne supérieure et inférieure, en plus de la dérive génétique, et qui permettent d'expliquer le paradoxe de Lewontin. Aujourd'hui, la question est de savoir lesquels de ces processus ont un impact décisif sur le niveau de variation de la diversité neutre. Pour ce faire, différentes études empiriques prenant en compte les effets évoqués précédemment ou bien des études utilisant des simulations sont menées (Corbett-Detig *et al.*, 2015; Buffalo, 2021; Charlesworth and Jensen, 2022). Cependant, elles ne traitent souvent que d'un effet à la fois, alors que la question nécessiterait l'utilisation de modèles incorporant tous les mécanismes soupçonnés d'avoir un impact sur le niveau de diversité neutre, pour quantifier l'impact de chacun.

Bien qu'il reste encore de nombreuses choses à comprendre concernant l'évolution des organismes, l'usage des données moléculaires a permis des avancées majeures dans le domaine. L'observation directe des séquences moléculaires a révélé l'abondance de substitutions neutres dans les séquences codantes et mis en lumière l'importance de processus non adaptatifs, tels que la dérive génétique, ce qui a

15. Cela sous-entend que la sélection est suffisamment intense pour impacter la diversité neutre sur plusieurs ordres de grandeur, ce qui ne semble pas être le cas (Andolfatto, 2001; Hernandez *et al.*, 2011)

contribué au développement de la théorie neutre de l'évolution. Aussi, l'existence d'une distribution continue des effets des mutations, développé par la théorie quasi-neutre, permet de faire le lien entre l'efficacité de la sélection naturelle à éliminer des mutations plus ou moins délétères, et la taille (efficace) des populations. Ce lien peut aujourd'hui être étudié plus en détails grâce à la grande quantité de données génomique dont nous disposons et l'avancée continuelle des méthodes d'études de l'évolution.

3

Mesurer l'évolution aujourd'hui : l'ère de la génomique

Comme présenté dans le chapitre 2, l'étude de l'évolution des organismes et des mécanismes sous-jacents a largement progressé au cours de ces dernières décénnies grâce à l'utilisation des données moléculaires en complément des données phénotypiques. Nous nous trouvons aujourd'hui dans un contexte d'accumulation de données moléculaires en tout genre (ADN, ARN, protéines, etc), ce qui nous permet de confronter plus en détail les attendus théoriques, évoqués dans les chapitres précédents, aux observations empiriques abondantes. Plus concrètement, nous sommes aujourd'hui en mesure de questionner les attendus de la théorie quasi-neutre concernant le rôle de la dérive génétique dans l'évolution des séquences codantes. Nous pouvons par exemple étudier les variations d'intensité de la dérive entre espèces et les confronter, de façon comparative le long d'une phylogénie, aux variations d'intensité de la sélection naturelle ; et ce, à différentes échelles évolutives. Une première échelle est l'échelle macro-évolutive, ou phylogénétique, dans laquelle on compare des séquences homologues provenant d'espèces différentes. La seconde échelle correspond à l'échelle micro-évolutive qui se place au niveau de la comparaison de séquences homologues d'individus d'une même espèce ou d'une même population. Ces deux échelles, prises indépendamment, sont individuellement pertinentes pour tester les prédictions de la théorie quasi-neutre. Leur confrontation l'est tout autant, car elles présentent chacune des sensibilités différentes face aux processus évolutifs à l'œuvre (voir si dessous).

Cependant, avant de pouvoir réaliser de telles études comparatives, il faut être en mesure de savoir agréger les données nécessaires puis, de les analyser correctement. Dans ce chapitre, je propose donc une revue technique des différentes étapes du traitement des données, allant du séquençage à l'écriture d'alignements de gènes orthologues (nécessaire à l'étude macro-évolutive), en passant par l'identification de variants dans les populations (nécessaires à l'étude micro-évolutive) (section 3.1). Je détaille ensuite quelques outils d'étude des données à l'échelle macro-évolutive (section 3.2) puis micro-évolutive (section 3.3), avant d'aborder une méthode capable de tirer parti conjointement de ces deux compartiments évolutifs (section 3.4). Enfin, je présente des méthodes statistiques d'analyses comparatives, qui permettent d'étudier l'évolution de traits d'intérêt le long de la phylogénie, ainsi que leur potentielle relation avec d'autres traits, dans un cadre statistique robuste (section 3.5).

3.1 L'enfer des données : la quantité au détriment de la qualité ?

Lorsqu'on veut étudier l'évolution des espèces et les mécanismes impliqués, il faut être capable de reconstituer la phylogénie de ces espèces à partir des informations portées par leurs séquences moléculaires. Pour ce faire, nous devons, en amont, reconstruire les génomes des espèces étudiées, ainsi que les événements qui ont abouti, aux séquences observées. Dans l'idéal, on souhaite développer un modèle de reconstruction phylogénétique que l'on applique aux données brutes et qui est capable d'articuler tous les niveaux de processus qui ont pu générer les différences entre espèces. Par exemple, le modèle prendrait en compte des événements comme les duplications de régions génomiques, les pertes de gènes, les insertions et délétions d'éléments, et les substitutions ponctuelles. On appelle ce type de modèle, un modèle intégratif. Cependant, aujourd'hui, bien que des modèles existent pour ces différents niveaux (par exemple : Tavaré (1986) pour les mutations ponctuelles, Holmes and Bruno (2001) pour les insertions et délétions), construire un modèle qui les prend tous en compte simultanément est encore computationnellement hors de portée (Simion *et al.*, 2020). De plus, le modèle intégratif idéal doit également être capable de modéliser directement les diverses sources d'erreurs issues des processus d'acquisition des données. Or ceux-ci sont plutôt complexes et pas toujours bien contrôlés. De fait, l'approche actuelle d'acquisition et exploitation des données se doit d'être séquentielle, autrement dit, étape après étape (Simion *et al.*, 2020).

Dans ce contexte séquentiel, chaque étape du processus produit un certain nombre d'erreurs qui impactent les étapes suivantes, potentiellement avec un effet boule de neige. À première vue, on observe que la plupart des pipelines produisent des topologies similaires quand on leur donne les mêmes données d'entrées [1]. Ils font donc preuve d'une certaine robustesse malgré les erreurs et approximations propres à chaque outil. En effet, les séquences contiennent un riche signal phylogénétique de type additif, qui désigne la manière dont les traits ou caractéristiques d'une espèce s'accumulent le long de la phylogénie. Ce type de signal ne peut qu'être concurrencé par des erreurs, elles aussi, de type additif (Simion *et al.*, 2020). Une erreur de type

1. Toutefois, des différences peuvent apparaitre sur des questions plus délicates concernant le placement de certains groupes dans les phylogénies (Lartillot and Philippe, 2008).

additive est une erreur qui produit un même type de biais entre des marqueurs ou lignées, ce qui engendre un signal non-phylogénétique et donc un arbre faux. Par exemple, des lignées avec une évolution particulièrement rapides présentent en général une forte homoplasie (de la similarité non liée à un ancêtre commun), ce qui a pour effet de les rapprocher artefactuellement dans les phylogénies. Il existe également des erreurs non-additives qui sont reparties aléatoirement et donc, ne changent pas le signal phylogénétique. On dit, que ces erreurs produisent du bruit, car elles réduisent la robustesse statistique des modèles (Simion *et al.*, 2020). Il est donc essentiel de savoir détecter ces deux types d'erreurs afin de les limiter ou, au moins, de les garder à l'esprit lors de l'analyse des données.

Une première étape importante dans le processus est celle du séquençage des génomes (sous-section 3.1.1). Souvent, les biologistes de l'évolution ne séquencent pas eux-mêmes leurs données, mais les récupèrent en accès libre. Cette pratique, tout à fait légitime dans le principe, crée néanmoins une distance avec les technologies utilisée, et qui peut entrainer une méconnaissance des biais que celles-ci peuvent engendrer. Dans une seconde étape, les génomes sont annotés (sous-section 3.1.2), souvent dans le but d'en extraire des séquences orthologues permettant de comparer les espèces entre elles. Les séquences orthologues sont alignées (sous-section 3.1.3) avant d'être utilisées pour reconstruire une phylogénie. Quand plusieurs individus d'une même espèce sont séquencés, il est également possible d'extraire le polymorphisme entre séquences par des méthodes de *variant calling* (sous-section 3.1.4).

3.1.1 Séquencer les génomes

Aujourd'hui, nous avons un accès libre à une quantité grandissante de génomes (ou gènes) (Box 4). Par exemple, en 1985, aucun génomes n'était séquencé, puis deux génomes l'ont été en 1995, 346 en 2005 et 20 500 en 2015. Cependant, le séquençage massif se fait parfois au détriment de la qualité et du suivi de ce qui est déposé dans les bases de données (Box 4). Il est donc, plus que jamais, nécessaire d'être particulièrement vigilant concernant la qualité des données utilisées.

Le séquençage Sanger et Illumina

Le séquençage de type Sanger (Sanger *et al.*, 1977) est l'une des premières méthodes permettant de lire une séquence ADN nucléotide par nucléotide. On utilise pour cela un mélange de nucléotides « normaux » et de nucléotides terminateurs de chaine fluorescents (avec quatre types de fluorescence pour distinguer les quatre types de nucléotides). Les nucléotides, normaux ou terminateurs, sont incorporés aléatoirement par une polymérase qui copie la séquence d'intérêt. Lorsqu'un terminateur est incorporé, la lecture s'arrête, générant des « lectures », ou « fragments », de tailles variés que l'on trie par taille croissante pour reconstituer la séquence par lecture des signaux lumineux (figure 3.1). L'efficacité de cette méthode a été augmentée dans les années 1990 grâce au développement de la technique PCR (*Polymerase Chain Reaction*) (Saiki *et al.*, 1985) et d'enzymes thermorésistantes. Ces enzymes, plus robustes aux changements de température nécessaires à la réaction, permettent de réaliser plusieurs cycles de séquençage et donc d'augmenter la quantité de lectures d'ADN obtenues à partir du séquençage d'une même séquence. Le séquençage Sanger est rapidement remplacé par le « *Next Generation Sequencing* » (NGS). Avec cette méthode, l'ADN est découpé en petits fragments de 200 à 500 paires de bases et amplifié par PCR en millions de copies identiques. Ensuite, chaque copie est lue par une polymérase qui ajoute des nucléotides, cette fois tous fluorescents, ce qui permet d'identifier chaque position de la séquence.

Les méthodes Sanger comme NGS utilisent des polymérases susceptibles de faire des erreurs. Il y a donc des lectures qui sont fausses. Cependant, ces erreurs sont généralement réparties aléatoirement et sont peu fréquentes pour chaque position. On peut donc les détecter et les corriger en séquençant plusieurs fois le même site (au moins 30 fois pour le NGS). De plus, ces méthodes ne produisent que des lectures courtes. Ces très petites séquences présentent des difficultés lorsqu'il est question, par la suite, de les réassembler en une plus grande séquence (voir plus bas). Récemment, des méthodes « *long-read* », comme PacBio et Nanopore (Rhoads and Au, 2015; Wang *et al.*, 2021), ont été développées pour obtenir des lectures plus longues (environ 15 000) ce qui facilite l'assemblage de génomes complexes en intégrant mieux les séquences répétées et les variations structurelles.

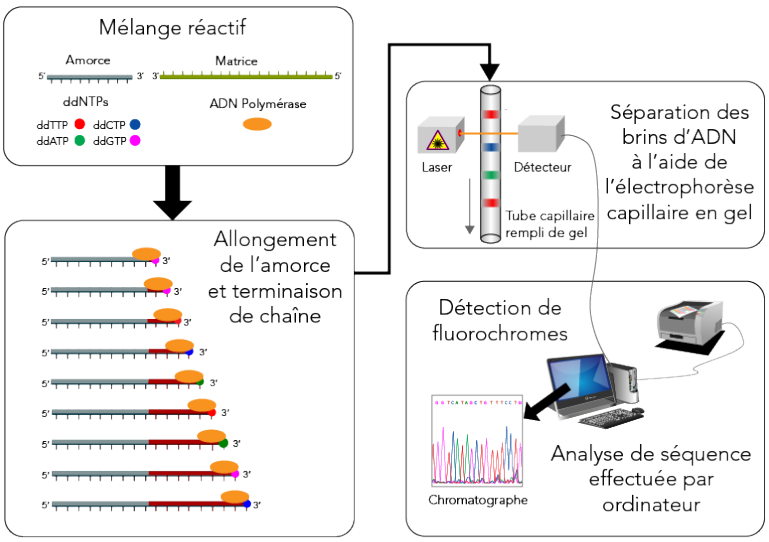


Figure 3.1 – *Schéma des différentes étapes du séquençage Sanger. (2020 Parlons sciences à partir d'une image d'Estevezj).*

Une fois l'ADN séquencé, quelle que soit la méthode utilisée, on obtient un fichier de type « fastq » qui contient les différentes lectures ainsi qu'un score de qualité par position.

Un puzzle à reconstruire

À la fin de l'étape de séquençage, on obtient des morceaux plus ou moins grands de séquences de l'ADN d'intérêt. Ces morceaux doivent être assemblés pour former progressivement une plus grande séquence[2] [3]. Quand le génome de l'espèce étudié, ou bien celui d'une espèce sœur, est déjà séquencé, on peut apposer les lectures issues du séquençage à ce génome qui devient alors une référence. On dit, dans ce cas, qu'on fait du *mapping*[4]. Dans ce contexte, un changement structurel, comme une insertion dans le génome séquencé, ne pourra pas être mappé sur le génome de référence, ce qui réduit la variation observée. De plus, les génomes de références sont produits de différentes façons en utilisant, par exemple, plusieurs génomes mélangés

2. Les lectures sont assemblées en « contigs », eux-mêmes assemblés en « scaffolds » puis parfois en chromosomes.

3. Cela revient à reconstruire un livre complet et cohérent à partir de morceaux de pages mélangés et issus de plusieurs exemplaires du même livre.

4. En poursuivant l'analogie avec le livre découpé, avoir une référence revient à s'aider d'une autre version du même livre, potentiellement un peu différente ou dans une autre langue.

d'individus d'une même population. Ils sont également souvent représentés comme haploïdes dans les bases de données, ce qui élimine toute l'hétérozygotie naturellement présente chez les individus diploïdes et peut entrainer le mauvais mapping d'une position polymorphe. Les génomes de référence, choisis comme base pour reconstruire de nouvelles séquences, sont donc des versions idéalisées des vrais génomes. Après le mapping, on élimine ou on masque les régions du génome reconstruit qui présentent un nombre anormal de lectures mappées dessus. Elles peuvent être signes de régions répétées qui se sont toutes identifiées au même endroit ou d'erreurs de séquençage.

Dans le cas où il n'y a pas de génome de référence, il faut réaliser un assemblage dit « *de novo* ». Pour ce faire, on utilise des algorithmes capables de détecter les superpositions entre lectures et de les fusionner (Zerbino, 2010; Li *et al.*, 2012; Koren *et al.*, 2017). Cette opération nécessite une certaine redondance des lectures et donc une couverture suffisante (30 à 100 lectures par position). Les séquenceurs de types Sanger ou NGS peuvent, dans ce cas, être problématiques car ils fournissent des lectures très petites qui demandent beaucoup de mémoire aux algorithmes.

La qualité d'un assemblage peut être examinée de plusieurs façons. En général, on utilise des mesures comme le N50 et L50 qu'on calcule en ordonnant les contigs par taille décroissante puis en formant un ensemble contenant 50 % des données. Le N50 correspond à la taille du plus petit contigs dans cet ensemble et le L50 correspond au nombre de contig dans l'ensemble. Un génome est ainsi de bonne qualité quand son N50 est grand et son L50 est petit. On peut également rechercher des gènes qui sont censés être présents dans le génome. L'outil BUSCO (Simão *et al.*, 2015) est capable de rechercher les gènes orthologues établis comme communs à un groupe donné. On peut, dès lors, calculer la fraction de ces gènes qui sont retrouvés dans le génome reconstruit, sans utiliser de génome de référence. En revanche, BUSCO ne recherche que les gènes orthologues appartenant à un groupe donné, il ne détectera donc pas la présence de gènes plus spécifiques à l'espèce. De plus, son score de qualité se base sur la présence de gènes très contraints par la sélection et donc souvent plus faciles à séquencer. De fait, le bon séquençage de ces gènes particuliers n'est parfois pas représentatif de la qualité globale du génome séquencé, surtout quand on se souvient que seulement 1 % du génome humain est fait de gènes (Box 2).

Box 4 : Que séquence-t-on ?

Beaucoup de génomes ont été séquencés depuis les premiers essais à la fin des années 1990. On a d'abord séquencé des espèces dites « modèles » comme l'humain, la souris ou la drosophile puis, des espèces d'intérêt agronomique, social ou scientifique. Aujourd'hui, nous séquençons de plus en plus d'espèces inconnues pour compléter l'arbre du vivant, qui présente une répartition inégale des génomes séquencés (figure 3.2). Bien que cela puisse être passionnant de collectionner les génomes de toutes les espèces vivantes, il n'est pas nécessaire de tous les connaitre pour comprendre les mécanismes évolutifs. D'autant plus que le séquençage nécessite souvent de prélever, et parfois de tuer, des individus, mais aussi d'interférer avec leur habitat, ce qui peut être plus préjudiciable qu'une simple méconnaissance de leur génome.

De plus, peu de ces nouvelles séquences sont controlées une fois déposées dans les bases de données. Par exemple, parmi les 245 500 000 séquences actuelles dans la base de données Uniprot, seulement 2,3 % ont été vérifiées [a] et il existe une fracture entre le nombre de génomes ou de gènes séquencés et leur annotation.
Ainsi, le séquençage massif se fait souvent au détriment de la qualité des données, rendant leur exploitation complexe. Peut-être qu'un modèle de séquençage plus parcimonieux, mais de meilleure qualité, pourrait être un modèle à privilégier à l'avenir.

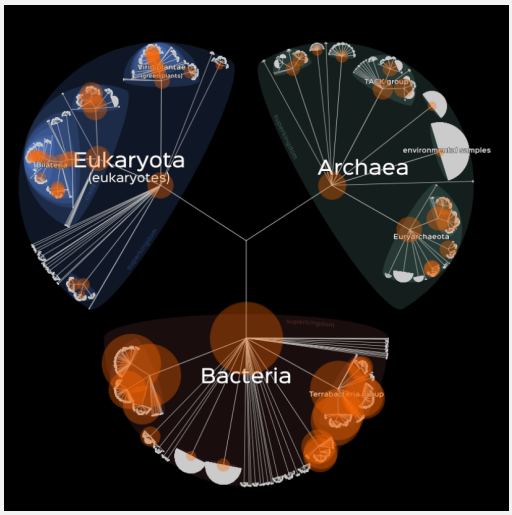


Figure 3.2 : Capture d'écran de LifeMap, une application permettant de visualiser l'arbre du vivant. Les zones orangées correspondent aux espèces pour lesquelles le génome est entièrement séquencé. lien vers le site : https://lifemap-ncbi.univ-lyon1.fr/

a. https://www.uniprot.org/uniprotkb/statistics

3.1.2 Annoter les éléments d'intérêt

Une fois le génome séquencé, on souhaite transformer ces données brutes en connaissances biologiques. Pour ce faire, on peut tenter de quantifier et de localiser certains éléments d'intérêt comme des gènes. Avant cela, il est souvent nécessaire de masquer les régions auxquelles on ne souhaite pas faire confiance. On peut par exemple, comme mentionné en sous-sous-section 3.1.1, utiliser la couverture fournie pour chaque position ou bien utiliser des programmes plus spécifiques comme Repeat Masker (Nishimura, 2000) dont l'objectif est de masquer les régions répétées.

Dans le contexte de ce manuscrit, je souhaite travailler avec des séquences orthologues communes à un groupe d'espèces. Pour ce faire, il faut d'abord annoter les gènes puis sélectionner ceux qui sont orthologues pour un groupe donné (voir Box 5). Comme pour le mapping présenté en sous-sous-section 3.1.1, si un génome de la même espèce que celle séquencée, ou proche de l'espèce d'intérêt, a déjà été séquencé et annoté, on peut utiliser cette référence pour annoter le génome sur lequel on travaille. Si le génome de l'espèce en question n'a jamais été séquencé, il est possible d'utiliser des méthodes dites *ab initio* qui ont été entrainées à reconnaitre un gène dans une séquence (par exemple, Augustus (Stanke and Morgenstern, 2005)). Une fois un gène identifié, on peut rechercher ses orthologues dans d'autres espèces en effectuant une recherche par similarités à l'aide de bases de données regroupant les gènes identifiés à ce jour et de programmes spécialisés, comme BLAST (Altschul *et al.*, 1990).

On peut aussi combiner annotation et recherche de gènes orthologues en détournant BUSCO (Simão *et al.*, 2015) de son usage habituel (sous-sous-section 3.1.1). Pour ce faire, on profite de sa recherche de gènes orthologues à un groupe donné pour récupérer les positions de ces gènes dans le génome évalué. Cette dernière approche, qui sera utilisée dans le travail présenté dans cette thèse (chapitre 5), offre ainsi une annotation à minima des gènes codants orthologues à un groupe taxonomique prédéfinis, que l'on peut recruter facilement et appliquer à un grand nombre de génomes. À noter qu'il existe une autre méthode, récemment développée, qui permet, elle aussi, d'identifier des gènes orthologues sans

que le génome soit annoté au préalable[5] (Kirilenko *et al.*, 2023). Cette méthode, nommée TOGA, s'appuie sur le machine learning et utilise des alignements entre un génome de référence annoté et des fragments du génome d'intérêt pour détecter les orthologues. TOGA est capable de détecter la majorité des orthologues déjà répertoriés dans la base de données Ensembl[6] mais également ceux qui ont été mal identifiés (par exemple, des faux positifs qui sont non fonctionnels) ou non identifiés.

Bien sûr, une annotation n'est jamais parfaite et peut engendrer plusieurs biais. Par exemple, on peut identifier un gène comme étant orthologue alors qu'il est en réalité paralogue (cf Box 5). Ce gène va alors être particulièrement divergent et présenter une histoire évolutive différente de celles des espèces. L'annotation peut également manquer un morceau du gène. Les erreurs sont d'autant plus fréquentes que le génome est de mauvaise qualité (fragmenté ou insuffisamment couvert).

5. Cette méthode étant très récente et ayant été publiée après le début de ma thèse, je n'ai pas pu l'utiliser

6. 97.6% détectés pour les orthologues du rat en utilisant le génome humain en référence (Kirilenko *et al.*, 2023)

Box 5 : Orthologue versus paralogue

L'homologie est un concept introduit dans les années 1840 par Owen. Ce concept est aujourd'hui très utilisé en génomique et plus largement en biologie. Il fait référence à une ressemblance entre structures ou gènes, présents dans différents taxons, qui proviennent d'un même ancêtre commun. On dit alors que ces éléments sont homologues. Par exemple, le gène de l'insuline du chimpanzé est homologues au gène de l'insuline de la vache.

Parmi les séquences homologues, on discerne plusieurs sous-catégories introduites par Zuckerkandl (1962) puis Fitch (1970) dont :

— l'**orthologie** : les deux séquences sont issues d'un événement de **spéciation**,

— la **paralogie** : les deux séquences sont issues d'un événement de **duplication**.

En évolution moléculaire, ce sont les séquences orthologues qui sont particulièrement intéressantes car elles permettent de reconstruire l'historique des spéciations et donc l'histoire des espèces. On s'attend à ce que plusieurs familles de gènes orthologues retracent la même histoire.

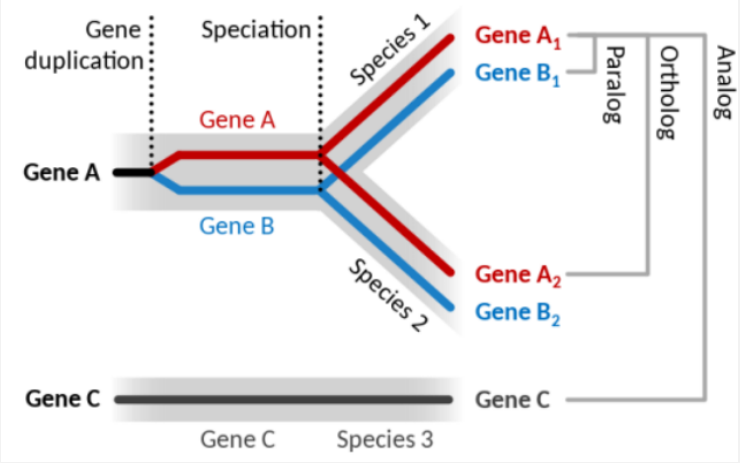


Figure 3.3 : Schéma conceptuel concernant la différence entre orthologue et paralogue.

3.1.3 Aligner les séquences orthologues

Bien que similaire, une même version d'un gène orthologue chez plusieurs espèces peut présenter des différences en termes de longueur (une des espèces peut avoir perdu un exon ou subi un indel, par exemple) ou de séquence (un nucléotide a pu être substitué par un autre) (voir la partie gauche de la figure 3.4). Ces différences sont, en principe, pertinentes pour étudier l'histoire évolutive des espèces portant ces gènes. De fait, il existe des modèles qui essaient d'inférer l'histoire détaillée des insertions, des délétions et des substitutions pour une famille de séquences homologues le long d'un arbre (par exemple, Baliphy (Suchard and Redelings, 2006)). Toutefois, ces méthodes sont complexes et coûteuses. Dans la pratique, on préfère aligner les séquences en ajoutant des « *gaps* » (représentés par des tirets), afin d'obtenir un alignement dans lequel chaque colonne représente une position homologue entre séquences (figure 3.4, partie droite). Il existe plusieurs méthodes d'alignement plus ou moins sophistiquées et adaptées à de grandes séquences ou à un nombre important d'espèces. La liste qui s'ensuit ne se prétend donc pas exhaustive.

Parmi les méthodes les plus simples, on retrouve l'algorithme CLUSTAL (Higgins and Sharp, 1988). Avec cette méthode, les séquences sont comparées par paires afin de reconstruire une matrice de distance qui, elle-même, permet de reconstruire un arbre guide. Les séquences proches, identifiées grâce à l'arbre, sont alignées entre elles puis fusionnées au reste de l'alignement jusqu'à construire progressivement un alignement global. Ce type de méthode est adapté pour des alignements de taille moyenne. CLUSTAL aligne généralement quelques dizaines de séquences de centaines de paires de bases en quelques minutes avec peu de ressources computationnelles.

MACSE (Ranwez *et al.*, 2011) est une méthode plus complexe, spécialement optimisée pour traiter les séquences codantes et gérer les mutations qui introduisent des décalages de cadre de lecture ou des codons stop prématurés. Les séquences ADN codantes sont converties en séquences d'acides aminés pour faciliter l'alignement tout en maintenant le cadre de lecture. Un alignement initial est effectué via des algorithmes standards comme CLUSTAL. Les séquences d'acides aminés sont ensuite reconverties en séquences nucléiques avec une attention particulière concernant les cadres de lecture via l'introduction de gaps ou la correction des indels. MACSE est

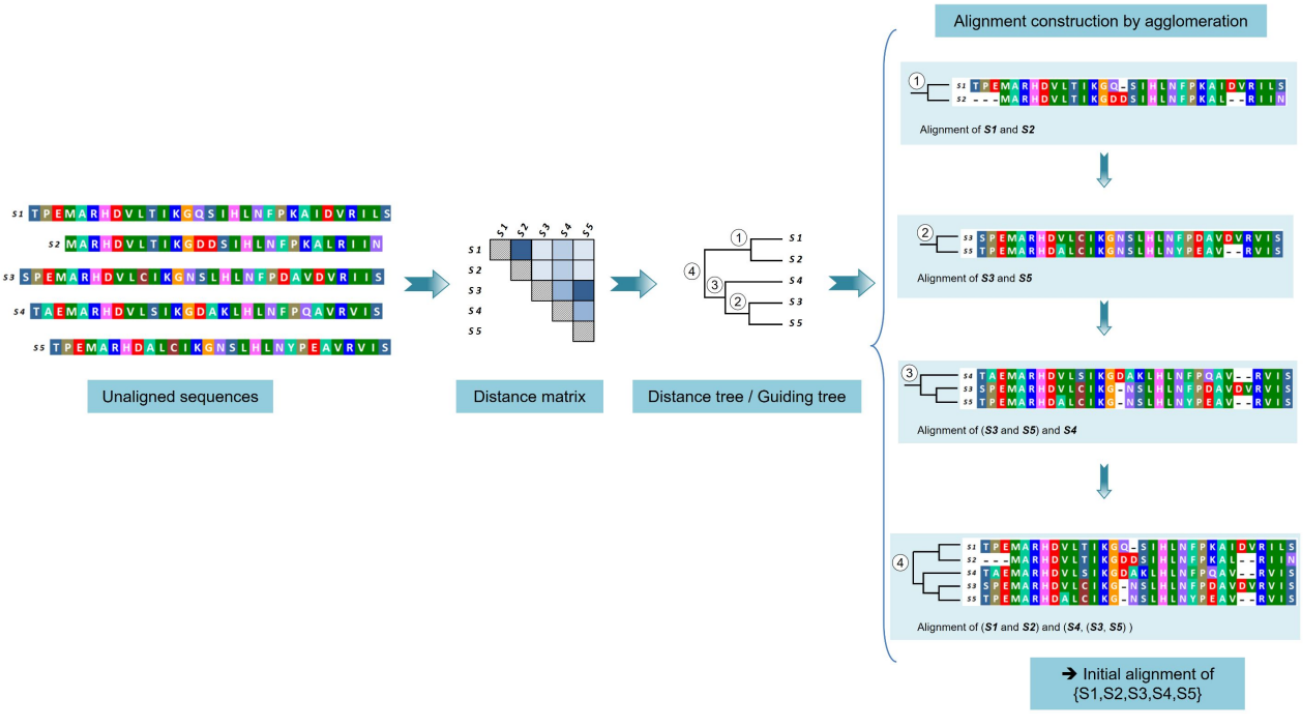


Figure 3.4 – *Différentes étapes d'un alignement utilisant un arbre guide. Une matrice de distance est reconstruite à partir des séquences afin de former un arbre guide qui conditionne l'alignement de sous groupes de séquences. Figure adaptée de Ranwez and Chantret (2020).*

capable d'intégrer de quelques dizaines à quelques centaines de séquences pouvant aller jusqu'au millier de paires de base. Il lui faut environ entre 15 et 30 minutes pour une centaine de séquences. Le programme est plutôt gourmand en termes de ressources computationnelles et peut nécessiter l'utilisation d'un cluster de calcul.

Enfin, PRANK (Löytynoja and Goldman, 2008; Löytynoja, 2014) est un programme qui prend en compte les processus évolutifs via l'utilisation d'une phylogénie guide, fournie avec les séquences. Cela lui permet de mieux intégrer les gaps causés par différentes insertions et déletions. Il a été montré que PRANK fournit des alignements de meilleure qualité que MACSE ou CLUSTAL (Löytynoja and Goldman, 2009; Fletcher and Yang, 2010; Dessimoz and Gil, 2010).

Ces différentes procédures d'alignement peuvent induire plusieurs types d'erreurs dans les données qui sont susceptibles de perturber les inférences phylogénétiques effectuées en aval (Wong *et al.*, 2008). L'alignement par paires engendre la formation de séquences consensus qui sur-simplifient les relations entre séquences. De plus, l'homologie établie à chaque étape n'est pas réévaluée : les erreurs faites dans les

étapes initiales de l'alignement vont avoir un fort impact sur le résultat final. Après alignement, on peut retrouver quelques régions problématiques pouvant venir de séquences localement ou globalement trop divergentes, de motifs répétés, ou de longues insertions ou délétions (voir des exemples figure 3.5). Pour pallier à cela, il existe différentes méthodes de filtrage des alignements. On cherche alors le meilleur équilibre entre la perte de signal phylogénétique et la réduction du bruit dans les données. Une façon de procéder consiste à filtrer les régions contenant trop de gaps, car celles-ci sont, par nature, moins fiables. Certaines méthodes filtrent des colonnes entières de l'alignement si elles contiennent plus d'un certain taux de gaps [7] tandis que d'autres filtrent, ou masquent, des régions dans certaines séquences sans impacter les autres élements de la même colonne [8]. En général, il est bienvenu d'utiliser ces deux méthodes l'une après l'autre pour profiter de leur complémentarité (Ranwez and Chantret, 2020).

Figure 3.5 – *Quelques exemples d'alignements problématiques issus de Ranwez and Chantret (2020).*

7. Méthode TILI, « Take It or Leave It », par exemple : BMGE (Criscuolo and Gribaldo, 2010) ou Gblocks (Castresana, 2000; Talavera and Castresana, 2007)
8. Méthode « picky », par exemple : HMMCleaner (Di Franco *et al.*, 2019)

Il existe également de nouvelles méthodes de filtrage qui comparent les arbres de plusieurs familles de gènes orthologues, chacun reconstruit à partir de son alignement, et l'arbre des espèces reconstruit à partir de l'ensemble des alignements (Comte *et al.*, 2023). On détecte ainsi les gènes entiers ou les séquences qui ne correspondent pas à l'histoire évolutive de l'espèce (qui ne sont donc pas orthologues). Dans le chapitre 5, les séquences nucléiques sont alignées avec l'outil PRANK (Löytynoja, 2014) puis, converties en séquences protéiques par un outil interne de Seaview (Gouy *et al.*, 2010), filtrées par HmmCleaner (Di Franco *et al.*, 2019), reconverties en séquences nucléiques par MACSE (Ranwez *et al.*, 2011) et enfin filtrées par BMGE (Criscuolo and Gribaldo, 2010) puis, par Phylter (Comte *et al.*, 2023).

3.1.4 Identifier des variants populationnels

Si plusieurs individus d'une même population ont été séquencés, on peut obtenir des données de polymorphisme. Pour ce faire, on confronte un génome de référence de l'espèce à celui des autres individus séquencés et on note les positions qui sont variables entre génomes. Très souvent, on se focalise sur les polymorphismes ponctuels, qu'on appelle des SNP, « *Single Nucleotide Polymorphism* ». On nomme cette procédure du *variant calling*.

L'analyse des variants dans une population engendre quelques difficultés. Il faut être capable de discerner les vrais SNP, des erreurs de séquençage ou de mapping. En effet, un site polymorphe observé peut être mal identifié (il n'est pas polymorphe) ou mal génotypé (on identifie mal le variant). Ces deux types d'erreur doivent, dans l'idéal, être considérés différemment. Lors du *variant calling*, chaque site identifié reçoit plusieurs scores de qualité qui dépendent de l'outil utilisé. Il existe toutefois une réelle difficulté concernant la compréhension de ces scores et leur diversité, ce qui ne facilite pas le filtrage des SNP, surtout dans un contexte où les fichiers VCF (« *Variant Calling Format* », les fichiers qui résument l'identification des SNP et leurs scores de qualité) sont eux-mêmes très volumineux et donc difficiles à manipuler.

Un mauvais filtrage des SNP peut biaiser le reste des analyses et non simplement les bruiter. Par exemple, du fait de la structure du code génétique, les erreurs de détection des SNP dans les séquences codantes sont plus fréquemment

non-synonymes que synonymes, créant un biais sur l'estimation du rapport π_N/π_S, rapport qui sera abordé en section 3.3.

Il est important de noter que, sur la base d'un seul individu diploïde recombinant, il est déjà possible de mesurer une hétérozygotie qui, elle-même, résume le niveau de polymorphisme de la population (voir section 3.3). Pour ce faire, on emploie à nouveau les lectures utilisées pour l'assemblage que l'on réaligne sur le même génome, afin de détecter les sites hétérozygotes. Dans ce cas particulier, étant donné que chaque chromosome est hérité de l'un des deux parents et représenté de manière équivalente, les deux allèles devraient normalement être détectés avec une fréquence égale, c'est-à-dire 50 % pour l'un et 50 % pour l'autre. Cependant, dans le premier travail de thèse (chapitre 5 et annexe section 9.1) nous observons plusieurs types de déviation de cette distribution. Certaines peuvent être le signe d'amplification inégale des deux chromosomes de l'individu par la méthode PCR, ce qui engendre une représentation biaisée des deux allèles dans les lectures. Ces observations soulignent, une nouvelle fois, l'importance et la difficulté du filtrage des données et de la vérification systématique de celles-ci, y compris pour certains attendus basiques.

3.1.5 La nécessité de l'informatique

En raison du volume et de la complexité des données évoquées précédemment, l'utilisation de ressources informatiques est devenue, à ce jour, indispensable. Les outils bio-informatiques font aujourd'hui partie du quotidien du chercheur et de la chercheuse en biologie moléculaire et les nouvelles générations d'étudiants et étudiantes en biologie sont de plus en plus formées à la programmation et à l'utilisation de ces outils.

Face à la gestion et à la formalisation d'énormes quantités de données, obtenues après différentes étapes de traitement, les chercheurs et chercheuses se retrouvent régulièrement à jongler avec une diversité de types, formats et contenus de fichiers. Ces fichiers, souvent très volumineux, peuvent atteindre plusieurs gigaoctets, comme les fichiers BAM qui portent l'information de la couverture à chaque position, ou les fichiers FASTA contenant les séquences des génomes[9]. Ainsi, l'analyse complète

9. Par exemple, dans mon étude, les fichiers bruts concernant la girafe (*Giraffa tippelskirchi*) font 2,6 Giga pour le génome, 900 Megabytes pour le VCF et 64 Gigabytes pour le fichier BAM.

d'un génome et des informations associées est très couteuse en espace de stockage [10] ce qui génère une forte inertie.

L'amélioration des capacités informatiques et surtout la constitution de cluster de calcul, ou super-ordinateurs, a permis d'augmenter, non seulement, la capacité de stockage mais aussi la complexité des modèles utilisés pour les analyses, la mémoire requise et la quantité de données traitables [11]. Cependant, il reste des défis à surmonter, notamment pour trouver un équilibre entre le temps de calculs, pouvant demander de quelques jours à plusieurs semaines la consommation des ressources computationnelles (collectives et polluantes (voir discussion sous-section 7.3.3) et la qualité du traitement. C'est pour ces raisons qu'il est crucial de publier les données produites pour éviter de les générer inutilement à nouveau et pour garantir la transparence des méthodes employées. Il existe des bases de données pour cela, comme Ensembl (Harrison *et al.*, 2023) pour les génomes, Uniprot (The UniProt Consortium, 2022) pour les protéines ou encore Orthomam (Scornavacca *et al.*, 2019) pour les gènes orthologues de mammifères.

Une fois les données agrégées, nous pouvons désormais nous concentrer sur les méthodes permettant d'analyser les différents processus évolutifs à l'œuvre dans l'évolution des génomes. Plus encore, nous pouvons étudier leurs variations d'intensité le long de la phylogénie et ainsi tester les prédictions de la théorie quasi neutre concernant le lien entre taille efficace de population et intensité de la sélection, et ce, aux échelles macro et micro-évolutives.

10. Dans mon travail sur 250 génomes, mon espace principal de stockage contient 2 Terabytes de données, les fichiers BAM sont stockés ailleurs et occupent un espace de 8 Terabytes tandis que le stockage des génomes occupes 1 à 2 Terabytes.

11. Par exemple, au LBBE, nous travaillons aujourd'hui avec un cluster de 1 400 CPU et 10 Terabytes de mémoire

3.2 Phylogénomique et macro-évolution

Dans un premier temps, nous pouvons mesurer l'intensité de la sélection à l'échelle phylogénétique, ou macro-évolutive, en comparant des séquences provenant de différentes espèces. En effet, initialement, la phylogénétique est une discipline qui a pour vocation de reconstruire les relations évolutives entre espèces (Darwin, 1859), d'abord à partir de caractères phénotypiques, puis moléculaires. Cependant, les méthodes de reconstructions phylogénétiques modernes utilisent des modèles probabilistes d'évolution des séquences qui peuvent également être employés pour étudier plus directement les mécanismes d'évolution de ces séquences. Ces différents modèles ainsi que leur utilisation pour mesurer les variations d'intensité de la sélection le long d'une phylogénie vont être détaillés dans les prochaines sections. Le vocabulaire associé aux différents éléments d'une phylogénie est répertorié en Box 6.

Box 6 : Vocabulaire en phylogénie

Une phylogénie est un graphique sous forme d'arbre qui représente les relations évolutives entre différentes espèces. La forme de la phylogénie est nommée une **topologie**.

Les **nœuds** représentent des unités taxonomiques (les différentes lettres de l'arbre A en Figure 3.6). Ils peuvent être externes comme les nœuds de A à E. Ont dit que ce sont les **feuilles** de l'arbre. Les autres nœuds (ou nœuds internes) représentent des unités taxonomiques hypothétiques.

Les nœuds sont reliés entre eux par des **branches** externes (qui relie un nœud externe à un nœud interne) ou internes (qui relie deux nœuds internes). Le nœud le plus ancien, celui a la base de tous, correspond à la **racine** de l'arbre. La racine représente l'ancêtre commun le plus récent à toutes les feuilles. Toutes les topologies ne sont pas racinées. Dans la Figure 3.6, seuls les topologies A, B et C sont racinées.

Si la topologie possède des longueurs de branche proportionnelles à la distance évolutive entre les séquences, ont dit que c'est un **phylogramme** (Figure 3.6, topologie C et D). Sinon, c'est un **cladogramme** et toutes les branches ont la même longueur (Figure 3.6, topologie A, B et E).

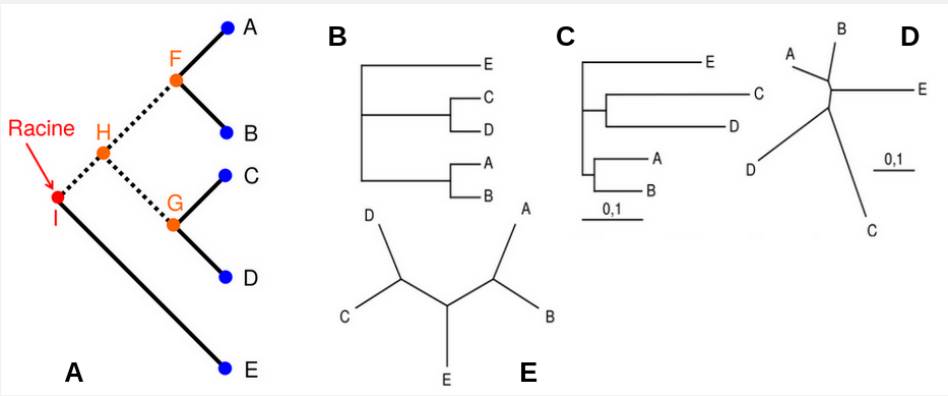


Figure 3.6 : Différentes formes de phylogénie.

3.2.1 Modèles de substitution nucléotidique

Comme mentionné plus haut, l'utilisation des séquences ADN pour la reconstruction phylogénétique à très vite nécessité le développement de modèles probabiliste d'évolution (ou modèles de substitution), initialement appliqués au niveau des nucléotides. En effet, lorsque deux séquences homologues sont séparées depuis un temps long, certaines substitutions ont pu être remplacées par d'autres, menant à une saturation du signal phylogénétique (Smith and Smith, 1996). Pour prendre en compte cette saturation, les modèles de substitution estiment la probabilité d'une substitution d'un nucléotide par un autre. Certaines substitutions étant moins probables, leur observation peut indiquer soit un long intervalle de temps écoulé, soit une substitution intermédiaire. Ces modèles permettent ainsi de reconstruire l'évolution des séquences, position par position, et peuvent nous renseigner sur les mécanismes moléculaires à l'œuvre. Par exemple, ils permettent de déterminer la distance évolutive entre des séquences homologues, ou sont à la base des méthodes de reconstruction phylogénétique par maximum de vraisemblance ou inférence bayésienne.

Les modèles de substitution nucléotidique sont entièrement caractérisés par leur matrice de taux de substitution (notée Q), de taille 4x4, qui décrit le taux de substitution par unité de temps entre toutes les paires de nucléotides. Il existe différents niveaux de complexité concernant la matrice de substitution avec des hypothèses plus ou moins réalistes (figure 3.7). Le modèle le plus simple est le modèle Jukes Cantor à un paramètre (Jukes, 1969). Il considère que chaque type de substitution à la même probabilité de se produire. Le modèle de Kimura (1980) dissocie les taux de transition et transversions [12] en désignant deux paramètres qui leur sont propres. Parmi les modèles les plus complexes, on retrouve le modèle GTR (*General Time Reversible*) (Tavaré, 1986) qui distingue un taux de substitution différent pour chaque paires de bases (soit 6 paramètres), ainsi qu'une proportion des quatre nucléotides à l'équilibre différent pour chacun (Il n'y a pas forcément autant de T que de G). Le modèle GTR combine ainsi neuf paramètres ajustables et estimables (figure 3.7), ce qui permet une grande flexibilité, mais nécessite plus de

12. Il existe quatre types de transversion (A↔T, T↔G, G↔C et C↔A) et deux types de transition (A↔G et C↔T). Il y a donc deux fois plus de transversion que de transition, dans une même unité de temps.

données et de ressources computationnelles.

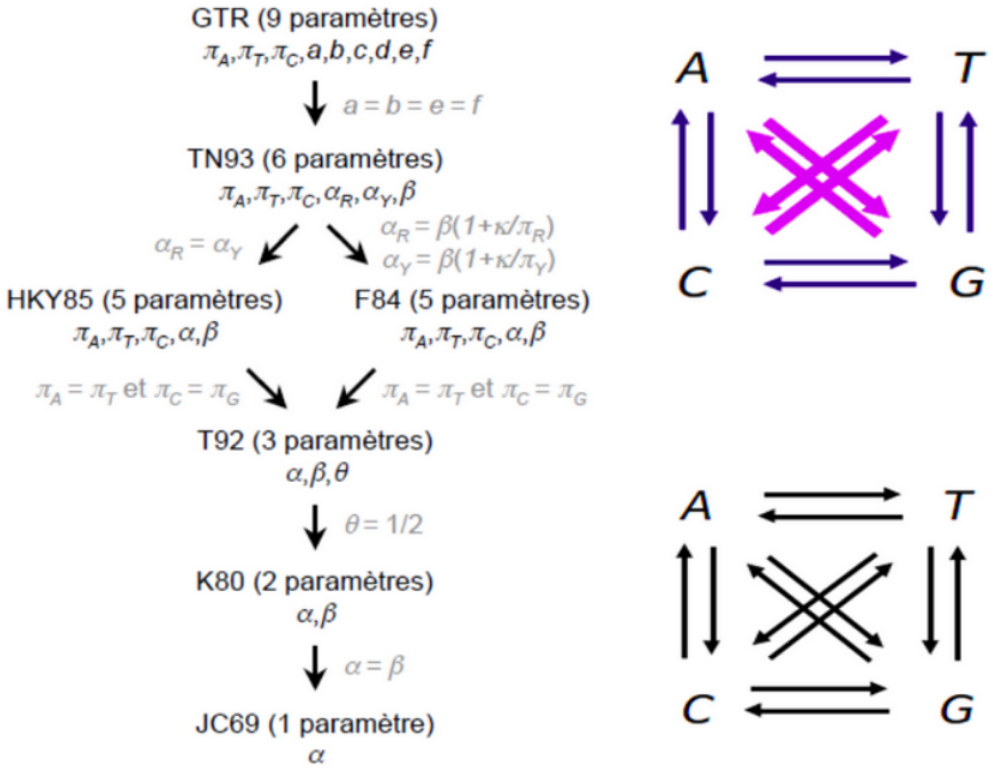


Figure 3.7 – ***À gauche*** *: Différents modèles de substitution ordonnés par leur complexité.* ***À droite en bas*** *: le modèle Jukes Cantor à un paramètre.* ***À droite en haut*** *: le modèle de Kimura à deux paramètres distingués par deux types de flèches.*

3.2.2 Modèles à codons

Les modèles de substitution peuvent être détournés de leur usage historique, dans la reconstruction phylogénétique, et être utilisé pour mesurer l'intensité de la sélection naturelle et sa variation entre espèces. Pour cela, on tire parti de la redondance du code génétique et de la conformation en codon des séquences codantes. En effet, en supposant que les substitutions synonymes sont neutres et donc uniquement influencées par la dérive génétique, on peut estimer un taux d'évolution neutre. Ce taux d'évolution neutre représente un bruit de fond homogène, non adaptatif, dans l'évolution des séquences. Comme vu au chapitre précédent, ce taux d'évolution correspond au taux de mutation. Il peut être comparé au taux de substitution des positions non-synonymes, soumises à la fois à la

mutation, la sélection et à la dérive, afin de mesurer un taux d'évolution adaptatif dans les gènes.

Pour que le nombre absolu de substitutions synonymes et non-synonymes dans une même séquence soient comparables (il y a environ deux fois plus de position non-synonymes que synonyme dans un gène), il est nécessaire de normaliser le nombre de substitutions observées (notées K) par le nombre de cibles potentielles (notées L). Ainsi, on calcule $d_N = K_N/L_N$, le taux de substitution non-synonyme et $d_S = K_S/L_S$ le taux de substitution synonyme. Si le ratio d_N/d_S est inférieur à 1, il y a un déficit de substitutions non-synonymes, signe d'une sélection globalement purifiante. Le cas d_N/d_S supérieur à 1 est signe d'une sélection positive appliquée sur les sites non-synonymes. Dans les faits, ce résultat est très rarement observé à l'échelle du génome ou à l'échelle des gènes. En effet, les substitutions avantageuses étant rares, leur signal est masqué par les substitutions délétères, plus nombreuses (75 à 85 % des mutations non-synonymes sont délétères (Eyre-Walker, 2002)). Il est donc rare de détecter de la sélection positive de cette façon. Globalement, la valeur moyenne de d_N/d_S pour une espèce, varie entre 0.15 et 0.25 tandis que la valeur de d_N/d_S à l'échelle des gènes varie de 0 à parfois plus de 2 (RA, 2004; Clark *et al.*, 2007; Yang and Gaut, 2011).

Pour pouvoir déterminer le nombre de cibles synonyme et non synonyme nécessaire au calcul de d_S et d_N, on peut choisir de simplement considérer qu'une substitution sur trois est synonyme (car les mutations aux troisièmes positions de codons sont synonymes) ou bien utiliser des modèles d'évolution à codons qui estiment directement le d_N/d_S. Ces modèles ont été développés pour la première fois dans les années 1990 et ont permis de grandes avancées dans la compréhension des processus évolutifs à partir du signal phylogénétique (voir figure 1). Un premier type de modèles, développés par Muse and Gaut (1994) puis Nielsen and Yang (1998), modélise le passage d'un codon i à un codon j via la substitution nucléotidique, dans une matrice de substitution, Q, cette fois de taille 61x61. Dans ces modèles, les codons i et j sont toujours séparés par une seule mutation et le taux de mutation (μ_{ij}) entre codons correspond au taux de mutation entre nucléotides. On peut alors distinguer les taux de substitution synonymes et non synonymes tels que :

$$\begin{cases} Q_{ij} = \mu_{ij} & \text{si i et j sont synonymes} \\ Q_{ij} = \mu_{ij} \cdot \omega & \text{si i et j sont non-synonymes} \end{cases} \tag{3.1}$$

Ici ω est un facteur d'échelle qui quantifie l'intensité de la sélection et qui peut être assimilé au d_N/d_S. Dans ce contexte, toutes les mutations non-synonymes sont considérés comme équivalentes, indépendamment de leur effet sélectif. Le modèle inclut l'estimation d'un paramètre de fréquence d'équilibre par nucléotide, soit indépendamment de sa position dans le codon (modèle 1x4 paramètres), soit en tenant compte des trois positions dans le codon (modèle 3x4 paramètres). Il peut également considérer la fréquence du codon final dans la séquence, ce qui mène à un modèle contenant 61 paramètres (Goldman and Yang, 1994)

À partir de ce modèle de base, plusieurs variantes cherchent à décliner un ω qui serait variable le long d'une séquence ou à travers les branches de la phylogénie, ce qui permet d'étudier plus finement les variations empiriques du d_N/d_S et notamment de rechercher de la sélection positive avec plus de puissance (Yang and Bielawski, 2000). Ces modèles sont résumés de façon complète et précise dans l'article de Anisimova and Kosiol (2008). Je propose ici une revue plus brève qui permet de se faire une idée du fonctionnement de ces modèles et des possibilités d'étude qu'ils offrent.

Modèles à codons par sites

Concernant les modèles par sites (Nielsen and Yang, 1998), ceux-ci ont l'intérêt de permettre de faire ressortir les sites sous sélection positive par rapport à ceux sous sélection purifiante. On estime ici un ω pour chaque site dans un contexte où le ω global du gène aurait certainement été inférieur à 1. Une approche parmi d'autre consiste à utiliser des modèles qui invoquent un petit nombre de catégories de ω, comme PAML (Yang *et al.*, 1997; Yang, 2007) et MrBayes (Huelsenbeck and Ronquist, 2001; Ronquist *et al.*, 2012). Chaque site est alors libre de s'associer à une catégorie de ω prédéfinie. On estime ensuite la probabilité pour chaque site d'appartenir à chacune des catégories et, de fait, la proportion de sites dans chaque catégorie. Parmi les catégories de ω, on peut en fixer une pour qu'elle soit supérieure à un et confronter la vraisemblance du modèle à un autre modèle dans lequel une telle catégorie n'est pas autorisée. Ces modèles montrent que les gènes contenant des sites sous sélection positive représentent 5% de l'ensemble des gènes des mammifères et que parmi ces gènes, seulement 10% des sites sont concernés (Kosiol *et al.*, 2008). Ces résultats sont en deça des estimations de modèles plus puissants comme les modèles McDonald-

Kreitman (McDonald and Kreitman, 1991) (détaillés plus tard en section 3.4). Cela témoigne d'une potentielle faiblesse statistique des modèles à codons par site, qui en réalité ne détectent que les sites sous sélection positive récurrente.

Modèles à codons par branches

En plus de ne pas être homogène le long de la séquence, l'intensité de la sélection ne l'est pas non plus le long de la phylogénie. De fait, des modèles à codons par branche ont été développés afin de tenir compte de cette hétérogénéité (Yang and Nielsen, 1998). Certains de ces modèles permettent seulement à un sous-ensemble de taxons d'avoir un ω différent afin de détecter une potentielle action de la sélection positive sur ce groupe après un événement de spéciation. Par exemple, dans la figure 3.8, on différencie le ω des primates (en vert) de celui des rongeurs (en noir). On peut également appliquer un ω différent à une unique branche, comme la branche des hominidés, en rouge dans la figure 3.8. A chaque fois, la vraisemblance de ces modèles est comparée à celle d'un modèle impliquant un unique ω. D'autres modèles permettent à chaque branche d'avoir soit son propre taux d'évolution (Yang, 1998; Galtier, 2001), soit d'appartenir à une des catégories prédéfinies de ω, comme pour les modèles par sites.

Ces modèles à codons par branche forment une sous-catégorie de modèles à codons particulièrement intéressante dans un contexte d'étude empirique de la relation, à l'échelle macro-évolutive, entre intensité de la sélection et N_e. En effet, les modèles à codons par branche permettent d'obtenir un d_N/d_S pour chaque branche terminale de la phylogénie que l'ont peu confronter dans un deuxième temps à d'autres traits reflétant N_e comme les traits d'histoire de vie. Il existe également des modèles capables de modéliser ω comme un trait continu qui varie le long de la phylogénie via un processus stochastique log-Brownien (Huelsenbeck and Rannala, 2003; Guindon *et al.*, 2004; Lartillot and Poujol, 2011) (voir sous-section 3.5.3 et chapitre 5). On peut alors obtenir un ω aux feuilles de la phylogénie, ce qui est d'autant plus pertinent pour la confrontation avec des traits d'histoire de vie.

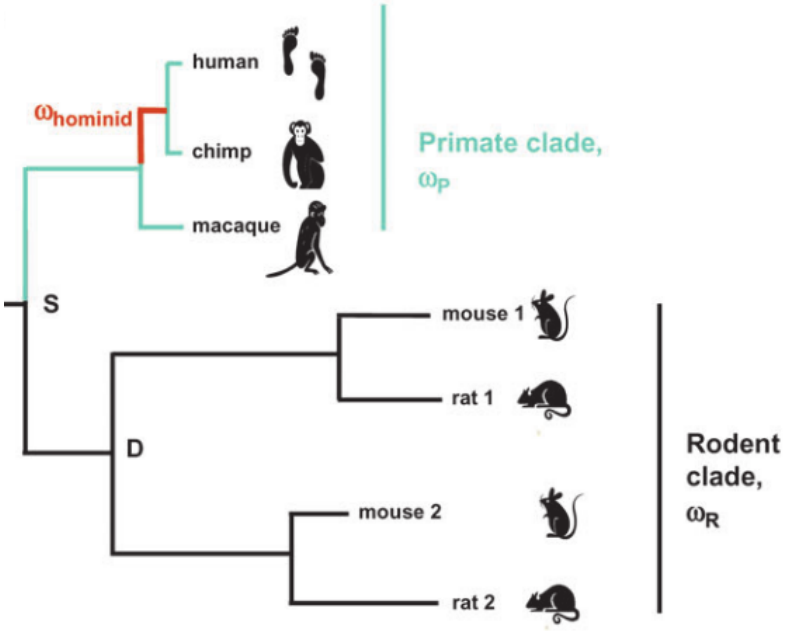


Figure 3.8 – *Exemple hypothétique de l'usage d'un modèle à codons par branche. Le nœud S représente une spéciation séparant les primates et les rongeurs et le nœud D représente une duplication au sein des rongeurs. Dans cet exemple, le modèle teste 1) si le sous-groupe des primates à un ω différent de celui des rongeurs (distinction vert/noir), 2) si la branche des hominidés (en rouge) à un ω différent du reste de l'arbre. Figure adaptée depuis l'article de Anisimova and Kosiol (2008).*

3.2.3 Autres types de modèles de substitution

Il existe d'autres modèles de substitutions plus sophistiqués tels que les modèles branche-site qui combinent la variation du d_N/d_S par branche et par site (Yang and Nielsen, 2002; Zhang *et al.*, 2005; Kosakovsky Pond *et al.*, 2011; Murrell *et al.*, 2012), ou encore des modèles qui permettent à différents sous ensemble des données d'évoluer sous un ensemble de paramètres différents (Lanfear *et al.*, 2017). On retrouve également des modèles qui prennent en compte la structure tridimensionnelle des protéines (Choi *et al.*, 2007).

Enfin, les modèles présentés ici s'appliquent à des séquences nucléotidiques, mais il existe également des modèles basés sur l'évolution des acides aminés. Ceux-ci présentent des matrices de substitution de taille 20x20. En général, on utilise des matrices d'échangeabilité entre acides aminés qui sont précalculées à partir d'un large jeu de données protéiques consensus (Dayhoff, 1978; Jones *et al.*, 1992; Whelan and Goldman, 2001), auxquels on multiplie une matrice contenant les fréquences des

acides aminés provenant des données étudiées. L'utilisation de ces modèles semble pertinente pour l'étude de séquences d'espèces distantes afin d'éviter la saturation des sites synonymes (Smith and Smith, 1996). Cependant, ils ne permettent plus de contraster le taux de substitution non-synonymes par le taux de substitution synonyme et donc d'inférer un d_N/d_S et ses variations le long de la phylogénie.

3.3 Génomique des population et micro-évolution

À l'échelle micro-évolutive, l'étude des séquences génétiques permet de comparer des individus au sein d'une même espèce. Contrairement à l'analyse des mutations fixées le long des branches d'une phylogénie, qui se concentre sur des événements évolutifs plus anciens, la génétique des populations, ou micro-évolution, s'intéresse aux mutations qui ségrègent actuellement dans les populations. Cela permet d'étudier les mécanismes évolutifs sur des échelles de temps plus courtes et offre ainsi des perspectives complémentaires à celles de la macro-évolution.

Avant de développer comment étudier la distribution des mutations dans une population, il est bon de définir les termes « population » et « espèce », qui prêtent souvent à confusion. Déjà, la définition d'espèces ne fait pas consensus (Mallet, 1995; Mayr, 1996). Globalement, on pourrait dire qu'une espèce correspond à l'ensemble des individus en capacité de se reproduire entre eux[13]. Cependant, pour d'autres, l'espèce devrait être redéfinie d'un point de vue génétique (Mallet, 1995). Concernant les populations, je dirais simplement que ce sont différents sous groupes d'une espèce avec des individus qui interagissent réellement entre eux. On observe souvent des processus de migration entre population ainsi que de la hiérarchie entre populations, formant des méta-populations. La comparaison de deux populations dans une espèce ne sera pas le sujet de ce chapitre. Ici, je m'intéresse plutôt à une population en sous entendant qu'elle représente son espèce. Cette hypothèse est discutée dans le chapitre 6 du manuscrit présentant mon deuxième travail de thèse.

3.3.1 Mesurer la diversité génétique en population idéale

13. Il existe quelques espèces qui peuvent s'hybrider, mais en général les descendants ne sont pas fertiles.

En comptant le nombre de sites polymorphes

La diversité génétique est souvent définie comme une mesure de la variation moléculaire dans une population. Plus concrètement, Nei and Li (1979) ont défini la diversité nucléotidique (notée π) comme la proportion moyenne de différences nucléotidiques entre deux séquences, échantillonnées aléatoirement dans une population.

À partir de cette définition idéale, on peut introduire différents estimateurs de la diversité génétique. En pratique, π peut-être estimé en échantillonnant n chromosomes dans une population. On l'écrit alors :

$$\hat{\pi} = \frac{\sum k_{ij}/L}{n(n-1)/2} \tag{3.2}$$

avec L, la taille de la séquence et k_{ij} le nombre de différences entre les séquences i et j. C'est un estimateur peu sensible aux allèles rares. Il est plutôt influencé par les allèles fréquents, présents depuis longtemps dans la population, qu'il comptabilisera plusieurs fois.

Une autre estimateur de la diversité génétique à partir des données de polymorphisme observées est le « θ de Watterson » ou θ_W (Watterson, 1975). Cette mesure prend en compte le nombre de sites qui montrent de la variation dans l'échantillon (s) divisé par la longueur L de la séquence et l'espérance de la taille totale de la généalogie de la population, (noté $E[T]$), tel que :

$$\hat{\theta_W} = \frac{s/L}{E[T]} = \frac{s/L}{\sum_{i=1}^{n-1} 1/i} \tag{3.3}$$

Ici, à chaque séquence ajoutée, on ne compte que les variants non observés précédemment, ce qui donne autant de poids aux variants rares que fréquents.

Sous l'hypothèse que la population ainsi étudiée corresponde à une population idéale à reproduction aléatoire (voir Box 3), on s'attend à ce que la mesure de $\hat{\pi}$ ou de $\hat{\theta_W}$ reflète un équilibre entre mutation et dérive génétique. Un équilibre mutation/dérive correspond à un équilibre entre l'arrivée de variants à un taux μ et une perte de variants via la dérive génétique, dont l'intensité est inversement

proportionnelle à N (Fisher, 1930). Si cet équilibre est respecté, on peut montrer que

$$\pi = 4N\mu = \theta \tag{3.4}$$

où l'on définit $\theta = 4N\mu$ comme le taux de mutation normalisé par la taille de population. Les deux estimateurs défini plus haut sont, en principe, non biaisés, ce qui implique que :

$$E[\hat{\pi}] = E[\hat{\theta_W}] = \theta \tag{3.5}$$

On notera également que la variance de $\hat{\theta_W}$ est plus faible que celle de $\hat{\pi}$ (Pluzhnikov and Donnelly, 1996; Tajima, 1983) ce qui fait de cette seconde mesure, une mesure à privilégier.

Cependant, $\hat{\pi}$ et $\hat{\theta_W}$ sont deux mesures imparfaites, car elles ont été définies dans un contexte où un site ne peut muter qu'une fois. Dans la réalité, il peut se produire plusieurs mutations par site, ce qui engendre des positions avec plus de deux allèles différents dans la population. Quand un site n'est pas bi-allélique, $\hat{\theta_W}$ est sous-estimé, car le nombre total de sites polymorphes observés est plus petit que le nombre de mutations qui ont eu lieu. Cela est d'autant plus impactant quand la population est grande, car elle possède plus de position tri-alléliques. De plus, les erreurs de séquençage augmentent artefactuellement le nombre de sites polymorphes, ce qui augmente $\hat{\theta_W}$. Concernant $\hat{\pi}$, le troisième allèle d'un site tri-allélique, tout comme les erreurs de séquençage, ne seront présents qu'en fréquence très faible et donc contribuent peu à la différence moyenne entre paires de séquences. Ainsi, en condition réelle, on favorise l'utilisation de $\hat{\pi}$ à celle de $\hat{\theta_W}$ (Tajima, 1996; Achaz, 2008; Johnson and Slatkin, 2008).

Enfin, mesurer π implique de disposer de données populationnelles, ce qui n'est pas toujours le cas en fonction du groupe étudié. En revanche, il est plus facile d'accéder à des données génomiques concernant un unique individu. Il se trouve, que dans le génome d'un individu diploïde recombinant, on retrouve des loci qui peuvent présenter des histoires généalogiques différentes. En effet, la recombinaison engendre un mélange des haplotypes issus des deux parents qui sont eux-mêmes des mélanges des haplotypes issus de chacun de leurs parents. On peut remonter comme cela jusqu'à l'ancêtre de la

population, ce qui implique que l'individu séquencé contient en réalisé un échantillon des séquences génétiques provenant de toute la population. Ainsi, si on mesure la fraction de différences nucléotidique entre les deux séquences homologues présentes dans l'individu, le long du génome, on obtient une mesure représentant la diversité moyenne de la population. On nomme cette mesure l'hétérozygotie (ou H). Dans le cas panmictique, H est strictement équivalent à $\hat{\pi}$ avec n=2. Dans un contexte d'étude comparative des variations de la diversité génétique entre espèces, on privilégiera l'usage de l'hétérozygotie, peut être moins précis que $\hat{\pi}$ mais moins limitant pour le choix des populations ou espèces étudiées. L'estimation de l'hétérozygotie et la proposition de son usage à la place de π sont discutés dans le chapitre 6.

En reconstruisant la généalogie de la population par coalescence

Reconstruire les liens de parenté entre les individus d'une même population peut être très informatif concernant son évolution démographique ainsi que les processus évolutifs en œuvre. Pour ce faire, on utilise un échantillon de séquences provenant de cette population et on remonte dans le temps en regroupant progressivement celles partageant un ancêtre commun à chaque étape. On dit qu'il y a eu un événement de coalescence quand deux séquences, à une génération donnée, se réunissent en une seule à la génération précédente (figure 3.9). Petit à petit, toutes les lignées coalescent jusqu'à converger vers une séquence ancestrale unique, qui est l'ancêtre commun le plus récent (nommé MRCA pour *Most Recent Common Ancestor*, figure 3.9). Ce processus produit un arbre de coalescence sur lequel différentes métriques peuvent être calculées. Par exemple, on peut estimer la probabilité qu'un événement de coalescence réduise le nombre de lignées de i à i-1. Pour ce faire, on considère la probabilité que deux individus à la génération i partagent le même ancêtre commun parmi les 2N présents dans la population à la génération i-1.

$$P(i \to i-1) = \frac{i(i-1)}{4N} \tag{3.6}$$

Dans ce contexte, le temps écoulé entre deux événements de coalescence correspond à l'inverse de cette probabilité, ce qui s'exprime par :

$$T(i \to i-1) = \frac{4N}{i(i-1)} \tag{3.7}$$

Ainsi, lorsque la probabilité qu'un événement de coalescence se réalise augmente, le temps entre deux événements diminue. Ce phénomène est particulièrement marqué lorsque le nombre de lignées est élevé, c'est-à-dire au début du processus de coalescence (figure 3.9).

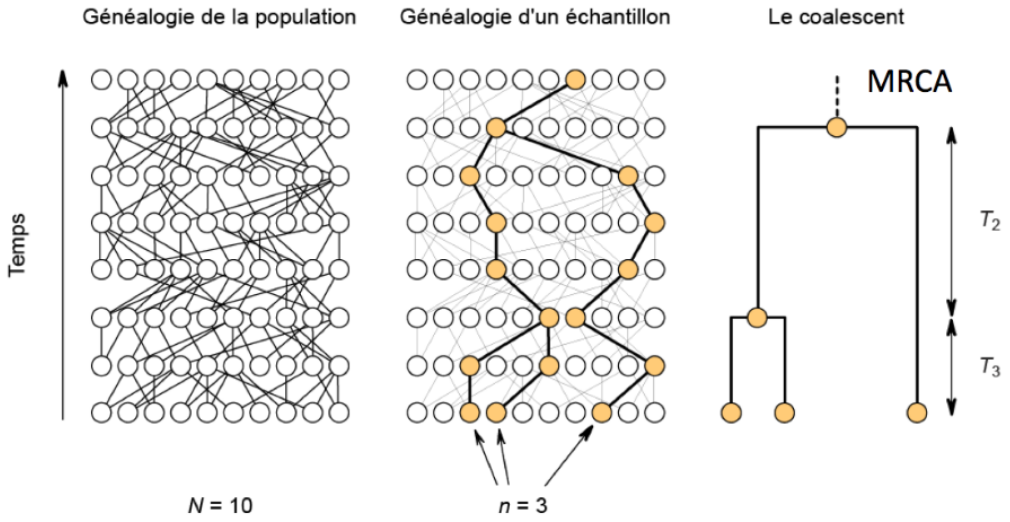


Figure 3.9 – *Reconstruction du coalescent à partir d'un échantillon de n=3 individus dans une population de N=10 individus. Quand deux des trois individus ont coalescés, le temps pour rencontrer un nouvel événement de coalescence est plus long (T2 > T3).*

Le coalescent permet de dériver assez simplement tout un ensemble de résultats théoriques. Ainsi, on peut montrer la relation donnée ci-dessus (équation 3.4) en utilisant le fait que la diversité nucléotidique moyenne, π, est censée correspondre au temps nécessaire pour remonter jusqu'à l'ancêtre commun de deux séquences, $2N$, multiplié par le taux de mutations par génération (μ), multiplié par 2 (les 2 branches de la généalogie) :

$$\pi = 2 \cdot 2N \cdot \mu = \theta \tag{3.8}$$

Le processus de coalescence peut également être utilisé pour simuler efficacement des données populationnelles (Kingman, 1982a,b; Hudson, 1983; Tajima, 1983). Une fois le coalescent simulé sous un certain modèle, on peut confronter les mesures de coalescence précédemment décrites à la quantité de différences observées entre séquences. En revanche, l'utilisation de la coalescence peut devenir inefficace pour

simuler de grandes régions génomiques ou des régions avec un fort taux de recombinaison. Pour pallier ces difficultés, des méthodes ont été développées, notamment les méthodes SMC (Sequential Markovian Coalescent) (Wiuf and Hein, 1999; McVean and Cardin, 2005). Celles-ci permettent de modéliser la corrélation entre les arbres de coalescence à des positions voisines, localisant ainsi les événements de recombinaison et segmentant les séquences en conséquence (figure 3.11).

Par l'usage d'un spectre de fréquence allélique

Le spectre de fréquence allélique (SFS) est également une bonne ressource pour évaluer la diversité dans une population et caractériser plus finement les effets conjoints de la démographie et de la sélection. C'est un graphique sous forme d'histogramme dont la hauteur des barres représente le nombre de sites dérivé observés à une fréquence donnée (en abscisse) (figure 3.10). Quand on ne sait pas quel allèle est dérivé et lequel est ancestral, on utilise l'allèle minoritaire et la fréquence du SFS est par construction bornée entre 0 et 0.5. On appelle ce type de spectre, un MAF (*minor allele frequency*). La forme du SFS ou MAF va renseigner sur les dynamiques évolutives présentes dans la population.

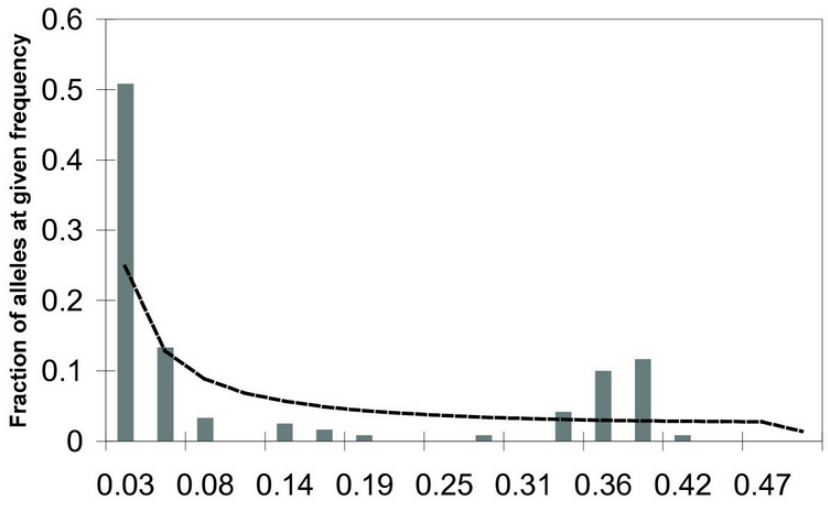


Figure 3.10 – *Spectre de fréquence allélique sous forme de MAF. En ordonnée, le nombre de sites observés à la fréquence présentée en abscisse. Par exemple, ici environ 50% des sites polymorphes sont observés à une fréquence 0.03 dans l'échantillon, pour l'allèle minoritaire. Figure adaptée de Geiler and Harrison (2010).*

3.3.2 Mesurer la diversité génétique en population non panmictique

La sous-section 3.3.1 définie des mesures de diversité génétique en population idéale. Cependant, en réalité, les populations s'éloignent plus ou moins fortement de la définition d'une population idéale, notamment par une reproduction qui n'est pas aléatoire. Comme détaillé dans la section 2.5, on peut remplacer l'usage de N dans l'équation 3.4 par la taille efficace de population (N_e) qui se voit alors définie comme la taille d'une population idéale qui aurait la même diversité nucléotidique que la population réelle. Dans ce contexte, π s'écrit :

$$\pi = 4N_e\mu = \theta \tag{3.9}$$

Puisque le travail réalisé lors de cette thèse est motivé par l'étude comparative des variations d'intensité de la dérive génétique, et donc de N_e, cette nouvelle définition de θ devient particulièrement intéressante. En effet, si on connait μ, θ donne un accès directe à N_e et donc à l'intensité de la dérive génétique. Cependant, dans ce contexte de population non-idéale, le lien entre les mesures $\hat{\pi}$, $\hat{\theta_W}$ et H avec θ est moins évident, ce qui complexifie l'estimation de N_e.

3.3.3 Intégrer la variation démographique

En plus de s'écarter du schéma idéal concernant la panmixie, une population n'a pas une taille constante dans le temps. De fait, les mesures de N_e proposées par un θ approximé par $\hat{\theta_W}$, $\hat{\pi}$ ou H en population non idéale correspondent en réalité à une mesure des N_e moyens dans le temps. Étant donné que les différents estimateurs de diversité reposent sur des hypothèses distinctes, chacun d'eux va proposer une moyenne de N_e avec une pondération différente des valeurs successives de N_e à travers le temps. Cela signifie que certains estimateurs peuvent accorder plus d'importance à des périodes spécifiques de l'évolution de la population, tandis que d'autres pourraient couvrir l'ensemble de l'histoire démographique. Par exemple $\hat{\theta_W}$ va donner plus de poids aux mesures de N_e dans le temps récent que $\hat{\pi}$ ou H. Finalement, il n'existe pas de définition unique de N_e et cela montre encore une fois la nécessité de préciser quelle définition est utilisée lors d'une analyse.

De façon plus optimiste, les différences conceptuelles entre les divers estimateurs de diversité peuvent être exploitées pour obtenir des informations complémentaires sur l'évolution démographique de la population. Par exemple, on peut calculer un « D de Tajima » (Tajima, 1989) tel que :

$$D = \hat{\pi} - \hat{\theta_W} \tag{3.10}$$

Si D est négatif, alors il y a un surplus de sites rares, ce qui peut être signe d'une expansion de la population. Si D est positif alors, il y a un déficit en sites rares, signe d'une diminution de la taille de la population.

De plus, dans une population avec des variations de taille efficace au cours du temps, le temps entre deux événements de coalescence va varier proportionnellement à $\frac{1}{N_e}$ (équation 3.7). Par exemple, une population en expansion présente un taux de coalescence plus lent actuellement qu'à des périodes passées, ce qui se traduit par des branches internes plus courtes que les branches externes. Cela entraîne également une augmentation des variants à faible fréquence, déplaçant ainsi le spectre de fréquence allélique (SFS) vers la gauche. À l'inverse, dans une population en déclin, les branches internes de l'arbre de coalescence sont plus longues. On observe alors une plus grande proportion de variants à fréquence intermédiaire, car les variants rares disparaissent plus rapidement. Ainsi, la reconstruction de l'arbre de coalescence permet d'estimer les tailles effectives de population (N_e) à différentes époques.

Parmi les méthodes SMC évoquées plus haut, Une méthode particulièrement intéressante est la méthode PSMC (Pairwise Sequentially Markovian Coalescent) (Li and Durbin, 2011), qui permet de reconstruire les variations de N_e dans le temps à partir du génome d'un unique individu diploïde (figure 3.11). Cela offre un avantage majeur, notamment pour des études basées sur l'hétérozygotie, car le temps de coalescence entre deux loci représente ici le temps écoulé depuis l'ancêtre commun le plus récent (TMRCA). Cette méthode permet de reconstruire les variations de N_e dans le temps à partir d'un seul génome, ce qui, comme pour l'hétérozygotie, est très avantageux (voir chapitre 6).

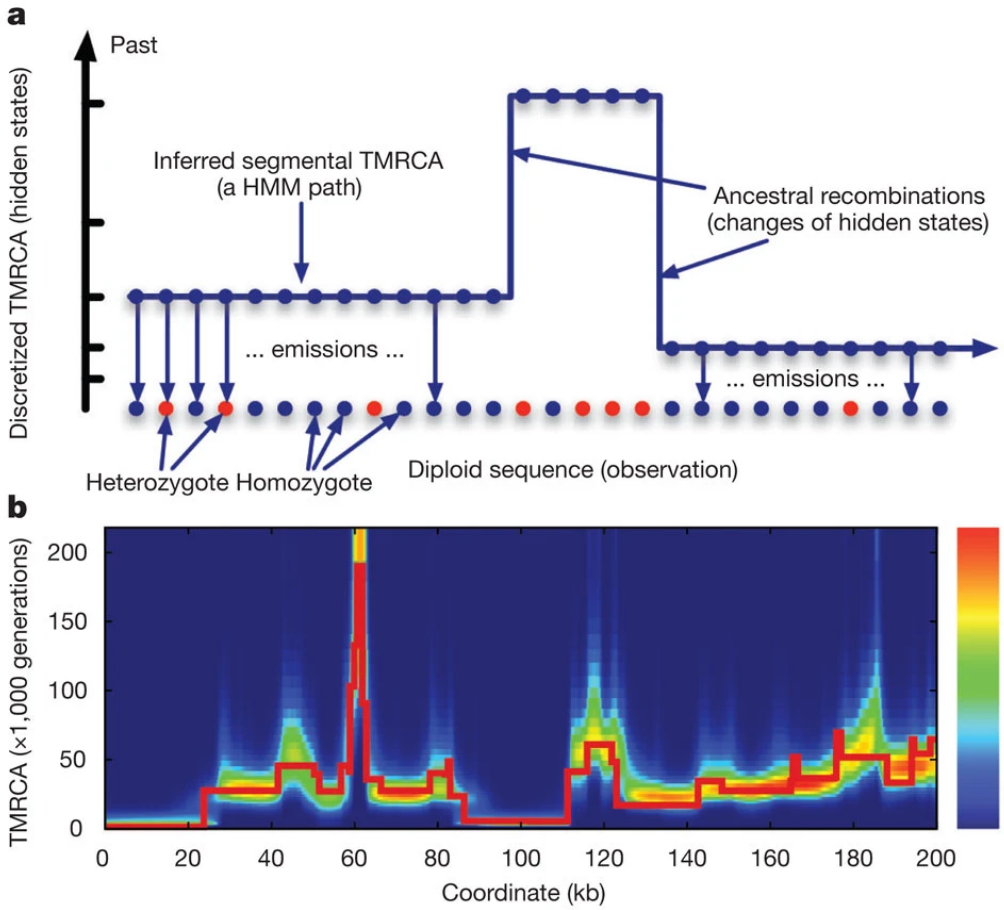


Figure 3.11 – *Le modèle et la méthode PSMC.* ***A :*** *inférence du temps de coalescence pour chaque paire de séquences dans un segment non recombinant du génome. Comme il y a seulement deux chromosomes, le temps de coalescence représente également le temps depuis l'ancêtre commun le plus récent de chaque segment. Les points rouges représentent les sites hétérozygotes.* ***B :*** *Simulation de l'estimation du temps depuis l'ancêtre commun le plus récent (TMRCA) sur une région de 200kb (ligne rouge) avec une méthode « ms » et inférence sur tout le génome avec une méthode PSMC (heat map). Figure adaptée de Li and Durbin (2011).*

3.3.4 Confronter les attendus quasi-neutres aux données populationnelles

Il est important de noter que les conclusions des sous-sections précédentes sont valables en l'absence de sélection. En réalité, les populations ne sont pas exemptées de processus sélectifs et il existe différents outils pour mesurer et caractériser le type de sélection à l'œuvre.

On peut notamment utiliser le spectre de fréquences alléliques et plus précisément sa forme. En effet, en absence de sélection et à démographie constante, le SFS possède une forme simple avec un nombre de mutations qui ségrègent à une fréquence f_i (avec i allant de 1 à n-1 individus échantillonnés), qui est proportionnelle à $1/i$ (histogramme gris dans la figure 3.12). En présence de sélection positive, on s'attend à une sur-représentation des mutations à forte fréquence (histogramme bleu dans la figure 3.12). En revanche, en présence de sélection purifiante, on devrait observer plutôt une sur-abondance de mutation à faible fréquence, car elles sont en cours d'élimination (histogramme rouge dans la figure 3.12). Enfin, dans le cas de sélection balancée, on observe une sur-représentation des mutations à des fréquences intermédiaires (histogramme jaune dans la figure 3.12)

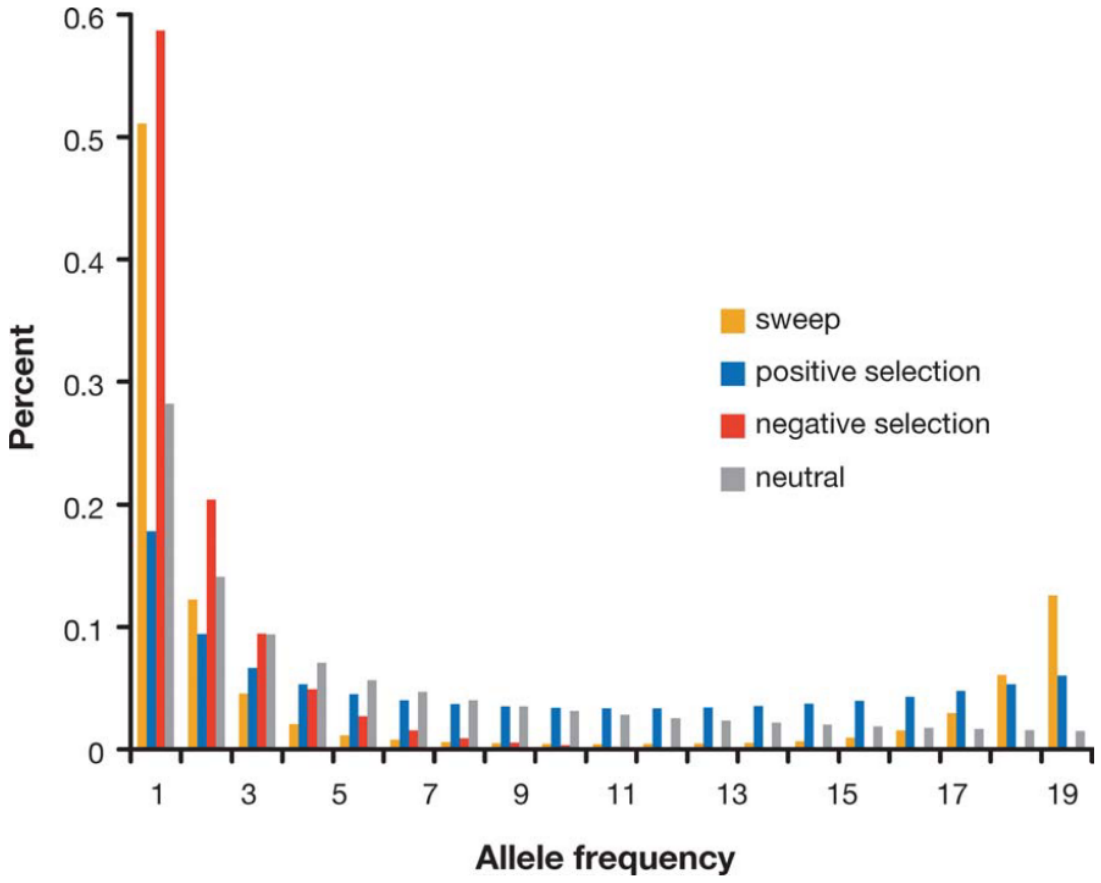


Figure 3.12 – *Les différentes formes d'un SFS en présence de différents mécanismes évolutif. Figure issue de Nielsen (2005).*

Par ailleurs, dans les séquences codantes, on peut calculer séparément la diversité synonyme (π_S) et la diversité non-synonyme (π_N). Si les mutations synonymes sont neutres alors :

$$\pi_S \simeq \pi = \theta = 4N_e\mu \tag{3.11}$$

De plus, on peut définir le rapport π_N/π_S comme une mesure de l'efficacité de la sélection, de la même façon qu'avec le d_N/d_S (sous-section 3.2.2). Comme pour d_N/d_S, un π_N/π_S inférieur à 1 marque un déficit de polymorphisme non-synonyme et donc l'action de sélection purifiante. En revanche, un π_N/π_S supérieur à 1 marque un excès de polymorphisme non-synonyme, ce qui est le signe de la sélection balancée. Les mutations avantageuses, quant à elles, se fixent rapidement et contribuent peu au polymorphisme. En théorie quasi neutre, on s'attend à ce que $\hat{\pi_S}$ corrèle négativement avec π_N/π_S . À noter que π_N/π_S et π_S peuvent également être estimés en utilisant un seul génome diploïde recombinant, à partir de l'hétérozygotie synonyme et non-synonyme. Cette mesure, plus facilement réalisable sur un ensemble large et homogène d'espèces, permet de tester les attendus quasi neutres de façon comparative à large échelle.

3.4 Croiser les deux échelles : modèle McDonald-Kreitman

3.4.1 McDonald-Kreitman classique

Dans les sections section 3.2 et section 3.3, j'ai proposé un tour d'horizon des méthodes macro et micro-évolutive pour étudier les séquences codantes, en lien avec l'évaluation de l'efficacité de la sélection dans ces séquences. Cependant, ces deux domaines méthodologiques sont rarement confrontés ou utilisés conjointement pour répondre à une même question. Une exception notable existe cependant avec les méthodes de type McDonald-Kreitman qui permettent de contraster des données de divergence et de polymorphisme pour tenter de détecter des traces de sélection positive dans un gène (McDonald and Kreitman, 1991).

Pour ce faire, on utilise les séquences nucléotidiques de plusieurs individus d'une espèce et au moins une séquence d'un individu d'une autre espèce, puis on comptabilise

le nombre de positions polymorphes et fixées, synonymes et non synonymes. Il n'est pas nécessaire de quantifier les cibles mutationnelles, seuls les comptages sont nécessaires.

Sous la théorie neutre, on sait que la quantité de polymorphisme entre deux séquences correspond à $4N_e\mu$ (section 3.3) et que la quantité de divergence correspond à l'accumulation de mutations le long des branches qui séparent les deux espèces soit $2t\mu$, avec μ le taux de mutation effectivement neutre et t le temps. Si on considère que la majorité des mutations synonymes sont neutres, alors le taux de mutations synonymes effectivement neutres μ_s correspond à $\mu_s = \mu.K_s$ avec K_s le nombre de sites synonymes. Concernant les mutations non-synonymes, il en existe une fraction f qui sont effectivement neutres, tel que le taux de mutation non-synonymes effectivement neutres μ_{ns} correspond à $\mu_{ns} = \mu.f.K_{ns}$. Dans ce contexte, on s'attend à ce que $\frac{d_N}{d_S} = \frac{\pi_N}{\pi_S}$ car :

$$\begin{cases} \frac{d_N}{d_S} = \frac{2t \cdot \mu_{ns}}{2t \cdot \mu_s} = \frac{2t \cdot \mu \cdot fK_{ns}}{2t \cdot \mu \cdot K_s} & = f\frac{K_{ns}}{K_s} \\ \frac{\pi_N}{\pi_S} = \frac{4N_e \cdot \mu_{ns}}{4N_e \cdot \mu_s} = \frac{4N_e \cdot \mu \cdot fK_{ns}}{4N_e \cdot \mu \cdot K_s} & = f\frac{K_{ns}}{K_s} \end{cases} \tag{3.12}$$

On retrouve également que :

$$\frac{\pi_N}{d_N} = \frac{\pi_S}{d_S} = \frac{4N_e}{2t} \tag{3.13}$$

Le test de McDonald and Kreitman (1991) permet de mesurer l'écart des données observées par rapport aux prédictions de l'équation 3.12. Dans un cas de sélection positive, les sites concernés sont rapidement fixés et donc ne contribuent pas au polymorphisme (Kimura, 1969; Smith and Eyre-Walker, 2002). En revanche, ils apparaissent dans la divergence. On observe alors un excès de divergence non-synonyme tel que $\frac{d_N}{d_S} > \frac{\pi_N}{\pi_S}$. Dans l'autre sens, un excès de polymorphisme non-synonyme est un signe de sélection balancée. Enfin, le test ne sera pas significatif s'il y a à la fois un excès de mutations non-synonymes dans le polymorphisme et dans la divergence.

3.4.2 Indice de neutralité et taux de substitution non-adaptatif

On peut utiliser l'équation 3.13 et le test de McDonald-Kreitman sur différents gènes pour les comparer via un index de neutralité [14] (Rand and Kann, 1996) tel que :

$$NI = \frac{\pi_N / d_N}{\pi_S / d_S} \tag{3.14}$$

S'il y a un excès de polymorphisme non-synonyme, signe de sélection balancée, alors NI sera supérieur à 1. Si NI est inférieur à 1, c'est un cas d'excès de fixation non-synonyme, signe de sélection positive. À partir de NI, on peut calculer la proportion de substitutions non-synonymes fixées par de la sélection positive (Smith and Eyre-Walker, 2002) tel que :

$$\alpha = 1 - NI \tag{3.15}$$

On obtient alors le taux de substitutions adaptatives :

$$\omega_a = \alpha \cdot \frac{d_N}{d_S} \tag{3.16}$$

Et le taux de substitutions non-adaptatives :

$$\omega_{na} = (1 - \alpha) \cdot \frac{d_N}{d_S} \tag{3.17}$$

3.4.3 McDonald-Kreitman moderne et difficultés

Les tests de McDonald-Kreitman classiques, tels que décrits ci-dessus, se placent dans le cadre de la théorie neutre stricte de l'évolution. C'est-à-dire qu'ils font l'hypothèse que les mutations sont soit fortement délétères soit strictement neutres. Cependant, on sait maintenant par la théorie quasi-neutre (Ohta, 1973) et des validations empiriques (Eyre-Walker, 2002; Loewe and Charlesworth, 2006;

14. Dans un projet auquel j'ai participé en parallèle de ma thèse (Latrille *et al.* (2024) et annexe section 9.3), nous avons utilisé une analogie à cet index pour détecter de la sélection diversifiante sur un trait à partir du contraste de données de génotype et phénotype (plutôt que divergence et polymorphisme).

Eyre-Walker and Keightley, 2007) qu'il existe également des mutations faiblement délétères qui sont plus lentement éliminées par la sélection et donc qui contribuent transitoirement au polymorphisme, sans nécessairement se fixer. Ces mutations faiblement délétères peuvent engendrer un ratio π_N/π_S supérieur au d_N/d_S, ce qui a pour effet de réduire la puissance du test de McDonald-Kreitman.

Étant donné que les mutations faiblement délétères sont caractérisables par une fréquence faible dans la population, on peut utiliser un spectre de fréquence allélique (SFS) pour les détecter et les éliminer. On peut choisir d'éliminer les singletons (mutation portée par une seule séquence) (Templeton, 1996) ou plus largement, les mutations à fréquences inférieures à 10% (Fay *et al.*, 2002; Smith and Eyre-Walker, 2002).

Une autre approche consiste à estimer directement la distribution des effets sélectifs (DFE) à partir des données de polymorphisme et divergence (synonymes et non synonymes) (Charlesworth and Eyre-Walker, 2008; Eyre-Walker and Keightley, 2009). La DFE peut comprendre une partie négative, qui correspond à de la sélection purifiante, et une partie positive, qui correspond à de la sélection positive. Quantifier la sélection positive revient alors à estimer la proportion de la partie positive. À noter toutefois que l'interprétation de cette contribution positive en tant qu'adaptation est sujette à débats (Latrille *et al.*, 2023). Cependant, ces approches sont plus coûteuses computationnellement. De plus, il existe des débats concernant la forme de la DFE et on ne sait pas à quel point le choix de cette forme peut impacter les conclusions (Eyre-Walker, 2002; Loewe and Charlesworth, 2006; Eyre-Walker and Keightley, 2007). Souvent, on utilise une distribution gamma réfléchie pour la partie négative.

Dans le cas d'une DFE gamma réfléchie avec un paramètre de forme β, Welch *et al.* (2008) dérivent un attendu simple :

$$\begin{cases} \pi_N/\pi_S \propto N_e^{-\beta} \\ d_N/d_S \propto N_e^{-\beta} \end{cases} \tag{3.18}$$

Sous l'hypothèse que la DFE est constante entre espèces, ces relations permettent en

principe d'étudier quantitativement la relation entre N_e et l'intensité de la sélection aux échelles micro et macro-évolutives, moyennant l'estimation du paramètre β qui dépend des SFS synonymes et non synonymes. Cette piste sera brièvement discutée dans la sous-section 4.4.2 et la section 7.1.

3.5 Méthode statistique comparative

Les méthodes exposées dans ce chapitre permettent de mesurer l'efficacité de la sélection à différentes échelles évolutives. Souvent, on souhaite comparer cette mesure à d'autres traits en lien avec des questionnements particuliers. Par exemple, dans le contexte de cette thèse, à l'échelle micro-évolutive, il est pertinent de comparer les variations entre espèces de π_N/π_S, mesure d'éfficacité de la sélection, à celles de π_S, proxy d'un N_e de court terme. On s'attend avec la théorie neutre à ce que π_N/π_S soit négativement corrélé à π_S. Dans la même logique, à l'échelle macro-évolutive, on peut s'intéresser aux variations de d_N/d_S qu'on suppose positivement corrélées à celles des traits d'histoire de vie qui sont cette fois des proxy indirectes d'un N_e de long terme. On appelle ce type d'analyse, une « analyse comparative ». Une analyse comparative correspond à l'ensemble des études de relations entre traits phénotypiques sur un ensemble de taxons et qui ont pour objectif de déceler, entre autres, des compromis évolutifs entre traits[15] ou des contraintes mécanistes. Un point méthodologique important soulevé par Felsenstein (1985) est que les relations observées entre traits ne sont pas indépendantes en raison de l'histoire évolutive commune des espèces. Il est donc nécessaire, en étude comparative, de prendre en compte la phylogénie.

Dans cette section, j'aborderai l'importance de la prise en compte de la phylogénie dans la comparaison de différents traits évolutifs entre espèces, puis je présenterai brièvement quelques méthodes comparatives de base utilisées dans les analyses comparatives classiques, avant de détailler une approche comparative « intégrative » que j'ai utilisée lors de mes travaux de recherche.

15. Par exemple, il y a un compromis évolutif entre produire beaucoup de descendants et vivre longtemps.

3.5.1 Prendre en compte la non-indépendance phylogénétique

Avant de pouvoir étudier la potentielle corrélation entre différents traits d'intérêt, mesurés chez différentes espèces, il faut en amont être en mesure de dissocier la corrélation réelle entre traits de celle induite par leur relation évolutive. En effet, les espèces se ressemblent d'autant plus qu'elles sont proches phylogénétiquement. On appelle cette ressemblance due à la phylogénie, un signal phylogénétique. Celui-ci peut être plus ou moins fort en fonction du trait étudié. Par exemple, dans la figure 3.13.a, les grands individus sont regroupés au même endroit dans la phylogénie, il y a donc un signal phylogénétique fort, tandis que dans la figure 3.13.b, la répartition de la taille des individus est indépendante de la phylogénie. Le signal phylogénétique créé ce qu'on appelle une « inertie phylogénétique » qu'on souhaite distinguer explicitement de la covariation potentielle entre traits d'intérêts.

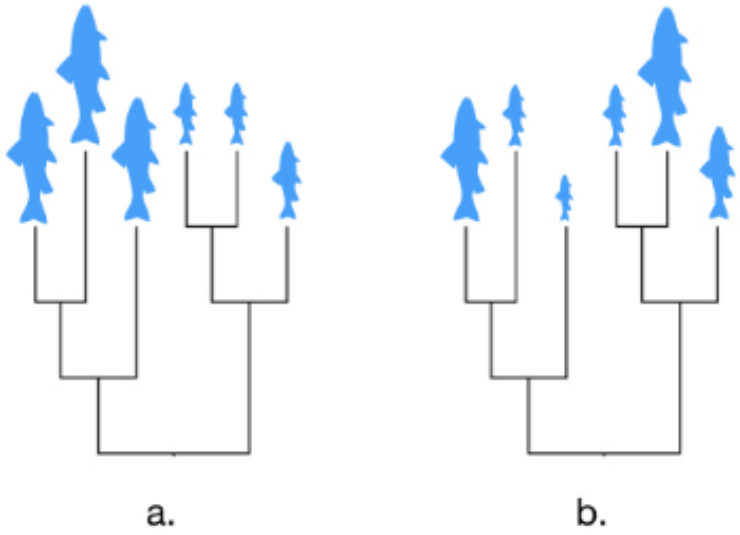


Figure 3.13 – *Schématisation d'un signal phylogénétique fort (a) ou absent (b). Figure issue de la figure 4 de Desdevises (2018).*

Il arrive encore aujourd'hui, que l'inertie phylogénétique ne soit pas prise en compte dans les analyses de corrélations entre traits. Pour Felsenstein (1985), c'est d'ailleurs un « *serious statistical problem* » [16] . En effet, lorsqu'on étudie la relation

16. « un problème statistique sérieux »

entre deux traits et qu'on utilise une seule valeur par espèce, tracer directement une droite de régression linéaire pour calculer la corrélation suppose que chaque point est indépendant Dans son article, Felsenstein (1985) présente le cas hypothétique de deux traits ayant évolué indépendamment le long d'un arbre formé de deux sous clades en étoile contenant chacun 20 espèces, reliés par une longue branche (figure 3.14 A). Il montre que si on utilise les valeurs observées à chaque feuille de l'arbre, on observera une droite de régression avec une pente significativement non nulle, laissant supposer une corrélation entre les deux traits X et Y étudiés (figure 3.14 B). Cependant, si on figure sur ce même graphique, la distinction entre les deux groupes de 20 espèces, on observe alors que la pente de régression intra-groupe est nulle, signe d'une absence de corrélation entre X et Y (figure 3.14 C).

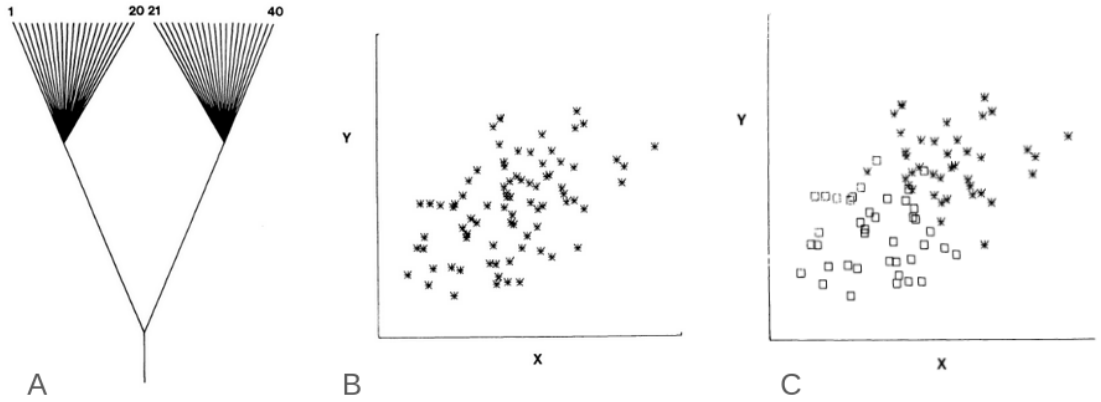


Figure 3.14 – *Figure et légende adaptée de Felsenstein (1985).* ***A :*** *Phylogénie de 40 espèces contenant deux groupes de 20 espèces proches.* ***B :*** *Un jeu de données typiques qui pourrait être généré par la phylogénie en utilisant un modèle de mouvements Brownien indépendant sur chaque caractère. On observe une relation positive entre X et Y* ***C :*** *le même jeu de données avec les points distingués pour représenter les deux groupes de 20 espèces. La relation présentée en B devient inexistante (« illusory »).*

Plus concrètement, dans le cas de l'étude du rôle de N_e dans l'évolution des génomes, certaines études ne prenant pas en compte l'inertie phylogénétique ont montré une relation significative entre $N_e\mu$ et la taille des génomes (Lynch and Conery, 2003). Cependant, cette relation n'est plus significative quand la phylogénie est prise en compte (Whitney and Garland Jr, 2010; Whitney *et al.*, 2011). Ainsi, ne pas prendre en compte la phylogénie dans une étude de corrélation peut mener à des conclusions fausses. C'est pour cela qu'il est nécessaire de soustraire l'effet de la

phylogénie lorsqu'on cherche à estimer une corrélation entre traits d'intérêt. En cela, les méthodes comparatives phylogénétiques, présentée ci-après, cherchent à rendre compte des liens de covariance entre traits tout en prenant en compte la part de relation expliquée par la phylogénie.

3.5.2 Méthodes comparatives phylogénétiques classiques

Modéliser l'évolution d'un trait le long de la phylogénie

Il existe tout un ensemble de méthodes comparatives phylogénétiques (détaillées dans Desdevises (2018)). La plupart de ces méthodes intègrent explicitement la modélisation de l'évolution d'un trait phénotypique le long d'une phylogénie. Le modèle le plus simple est le modèle Brownien qui simule l'évolution de traits comme une marche aléatoire, sans mémoire ni direction, et ce, de façon indépendante le long de chaque branche de la phylogénie (figure 3.15.a). En moyenne, la valeur du trait ne change pas de sa valeur initiale et donc correspond à la valeur du nœud précédent. La variance du trait en revanche va augmenter avec l'accumulation de micro décalages au cours du temps. De fait, plus la branche est longue et plus la variance est élevée. Par construction, la divergence attendue entre les valeurs du trait pour les deux espèces est en moyenne proportionnelle au temps de divergence entre les deux espèces (figure 3.15.b).

Ce type de modèle est utile dans une situation de dérive génétique au niveau phénotypique. Il a cependant été critiqué en cas d'évolution adaptative (Leroi *et al.*, 1994; Martins, 2000). Un autre modèle a ainsi été développé pour mieux prendre en compte les processus adaptatifs. C'est le modèle Ornstein-Uhlenbeck qui, en plus d'un mouvement brownien, considère une force de rappel (représentant l'effet de la sélection stabilisante) qui conduit la valeur des traits de différentes espèces à osciller autour d'un optimum, ce qui est plus réaliste dans un contexte adaptatif (Hansen, 1997) (figure 3.15.c). Le modèle Ornstein-Uhlenbeck fait toutefois l'hypothèse que l'optimum adaptatif est constant. En pratique, un mouvement brownien simple peut finalement bien représenter une sélection stabilisante pour un optimum qui lui-même change au cours du temps.

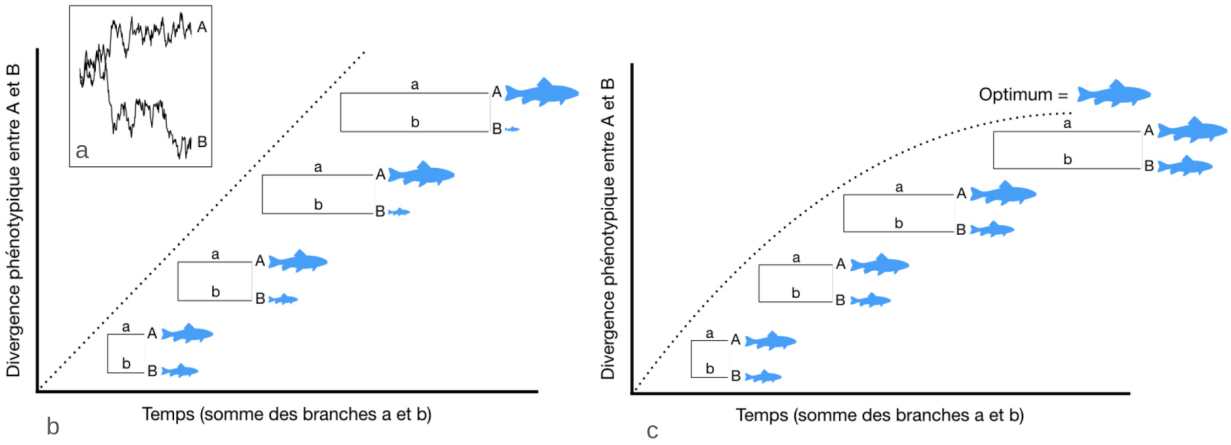


Figure 3.15 – *Modele Brownien et Ornstein-Uhlenbeck d'évolution de traits phénotypiques.* ***a.*** *Modèle Brownien : simulation de l'évolution d'un caractère quantitatif chez deux espèces A et B après spéciation. Les caractères évoluent aléatoirement mais leur divergence tend à augmenter avec le temps. La moyenne du trait reste la même, sa variance augmente proportionnellement avec la divergence.* ***b.*** *Divergence phénotypique entre A et B en fonction du temps.* ***c.*** *Modèle de Ornstein-Uhlenbeck. Le phénotype des espèces A et B converge vers un optimum, et la divergence phénotypique après spéciation tend à se stabiliser. Figure et légende adaptées des figures 2 et 3 de Desdevises (2018).*

La méthode des contrastes indépendants

Une fois les modèles d'évolution de traits phénotypiques définis, on peut les intégrer dans les méthodes d'analyses comparatives phylogénétiques. Dans son article de 1985, Felsenstein propose la méthode des contrastes indépendants qui reconstruit l'évolution des traits en utilisant un mouvement Brownien. Un contraste indépendant correspond à la comparaison de la valeur du trait X pour deux feuilles i et j qui ont un ancêtre commun au nœud k. On écrit alors que le contraste entre X_i et X_j correspond à la différence entre X_i et X_j, standardisée par la quantité d'évolution qui les sépare de k, tel que :

$$C(X_i, X_j) = \frac{X_i - X_j}{\sqrt{l_{ik} + l_{jk}}} \tag{3.19}$$

Avec l_{ik} et l_{jk} qui correspondent aux longueurs de branches allant respectivement de k à i et k à j. La valeur de X_k n'est pas observée, mais le modèle d'évolution Brownien permet de l'inférer en prenant la moyenne des valeurs observées pour ses descendants X_i et X_j, pondérées par l'inverse des quantités d'évolution tel que :

$$X_k = \frac{X_i/l_{ik} + X_j/l_{jk}}{1/l_{ik} + 1/l_{jk}} = \frac{l_{jk}X_i + l_{ik}X_j}{l_{ik} + l_{jk}} \tag{3.20}$$

En présence de deux traits X et Y à comparer, on peut calculer le contraste $C(X_i, X_j)$ et $C(Y_i, Y_j)$. Chaque paire de contraste correspond à un point indépendant par rapport aux autres paires, car celui-ci ne dépend que des événements évolutifs sur les branches spécifiques aux feuilles considérées, sans être influencé par les branches partagées par ces feuilles. Felsenstein (1985) développe un algorithme qui permet de reconstruire n-1 contrastes en parcourant une phylogénie de n feuilles. Une fois tous les contrastes calculés, on peut utiliser une régression linéaire sur ces contrastes et calculer un coefficient de corrélation qui cette fois aura pris en compte l'inertie phylogénétique.

La méthode PGLS

La méthode PGLS (*Phylogenetic Generalised Least Square*) (Grafen, 1989; Pagel, 1993; Martins, 1994) est une autre une méthode statistique qui permet de comparer des traits phénotypiques en prenant en compte l'inertie phylogénétique et en s'appuyant sur l'hypothèse que les traits évoluent de façon brownienne. En PGLS, la structure de l'arbre est représentée sous forme d'une matrice de variance-covariance reliant les différentes unités taxonomiques. La diagonale de la matrice correspond à la variance du trait X chez les différents taxons (la quantité d'évolution accumulée entre la racine et l'unité évolutive concernée). En dehors de la diagonale, on retrouve la covariance entre chaque paire de taxon (la quantité d'évolution de la racine jusqu'à leur ancêtre commun).

Cette approche peut ensuite être généralisée pour prendre en compte une éventuelle contribution non-phylogénétique aux traits observés. Pour ce faire, on multiplie tous les éléments en dehors de la diagonale par un paramètre λ compris entre 0 et 1 (Pagel, 1997). La valeur de λ donne directement le poids du signal phylogénétique dans la corrélation des caractères. L'estimation de λ et des valeurs des traits dans les nœuds internes se fait par utilisation de la méthode des moindres carrés généralisée (GLS) (Grafen, 1989) qui minimise la somme des carrés des écarts entre les valeurs observées et les valeurs prédites par le modèle.

3.5.3 La méthode intégrative Coevol

Dans chacune des deux méthodes présentées ci-dessus, on restreint les analyses comparatives aux traits écologiques et phénotypiques alors qu'il existe d'autres traits continus d'intérêt comme le taux de substitution ou le d_N/d_S. Or l'étude de la corrélation entre ces paramètres moléculaires et les paramètres écologiques est d'un grand intérêt pour mieux comprendre le fonctionnement de l'évolution et tester empiriquement les attendus de la théorie neutre et quasi neutre. De plus, avec ces méthodes, on calcule des contrastes indépendants puis, dans un second temps, on réalise une analyse de régression sur les données corrigées. On dit, de fait, que ce sont des méthodes d'analyses séquentielles.

Face à ces limitations, des méthodes comparatives dites « intégratives » ont été développées, capables d'invoquer tous les éléments nécessaires à l'analyse, en une seule étape plutôt que séquentiellement. Ici, je vais particulièrement présenter la méthode intégrative Coevol (Lartillot and Poujol, 2011) qui permet de faire évoluer ensemble, plutôt que séparément, des traits écologiques et moléculaires d'intérêts le long d'une phylogénie, par l'usage d'un mouvement Brownien multivarié.

Coevol utilise une méthode bayésienne d'estimation de la covariance entre traits phénotypiques (Huelsenbeck and Rannala, 2003) et y ajoute la considération de paramètres du processus de substitution moléculaire eux-mêmes susceptibles de varier au cours du temps et entre espèces (Lartillot and Poujol, 2011). Pour que les taux de substitutions puissent être comparés avec les caractères phénotypiques, il convient de fournir un taux de substitution instantané (Welch and Waxman, 2008). En général, on approxime ce taux par la moyenne du taux de substitution sur les branches internes de l'arbre. Une approche plus efficace consiste à modéliser les taux comme un paramètre continu qui évolue le long de la phylogénie par un processus de diffusion brownien (Thorne *et al.*, 1998; Seo *et al.*, 2004; Rannala and Yang, 2007). Ces modèles, dits à horloge moléculaire relaxée, permettent de reconstruite les taux de substitution instantanés pour chaque nœud de l'arbre à partir de l'utilisation d'un alignement de gènes orthologues. Tous les traits d'intérêt sont modélisés dans un unique processus de diffusion brownien multivarié noté X(t), dont la dimension correspond au nombre de paramètres du modèle de substitution ainsi qu'aux traits

phénotypiques considérés. Ainsi, Coevol combine les contrastes indépendants de Felsenstein (1985) et les méthodes de Huelsenbeck and Rannala (2003) et de Seo *et al.* (2004) pour estimer conjointement dans un cadre bayésien, une matrice de covariance entre traits, la reconstruction des traits et des taux le long de la phylogénie et les temps de divergence (figure 3.16).

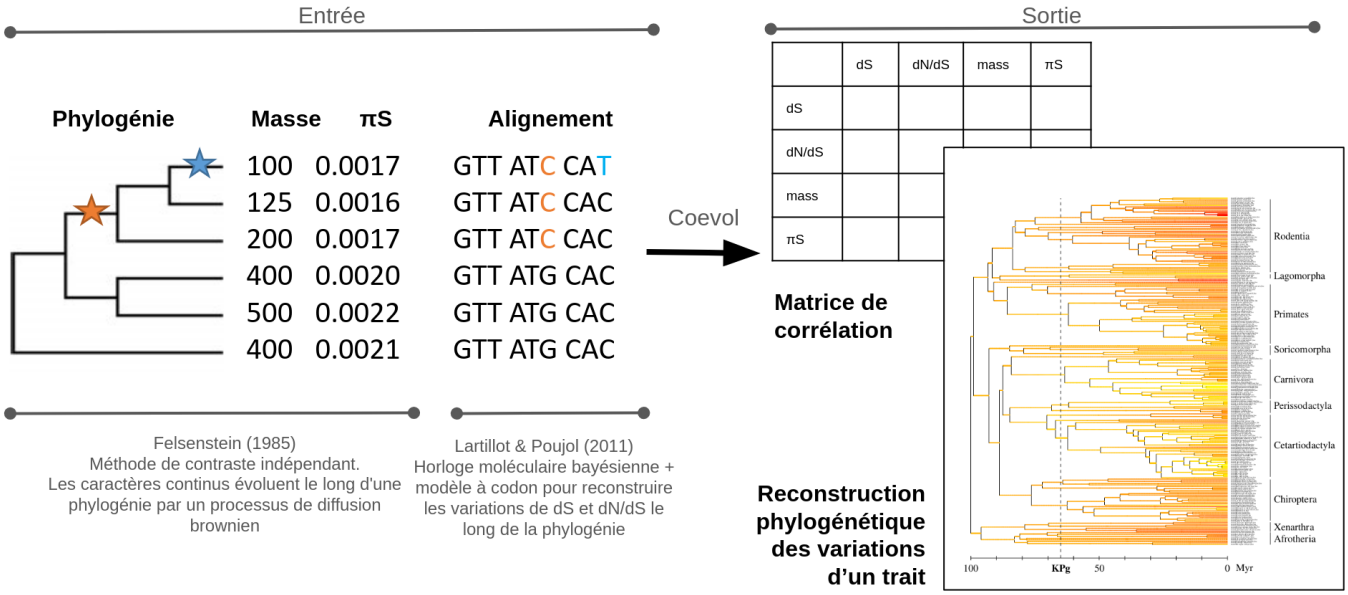


Figure 3.16 – *Le modèle intégratif Coevol et ces différents composants.*

Bien que prometteuse, la méthode Coevol reste coûteuse en temps de calcul et est difficilement applicable sur de grands ensembles de gènes. En effet, l'utilisation d'une inférence de type Monte Carlo (MCMC), requise par la combinaison du mouvement brownien et d'une vraisemblance calculée sur un alignement de séquences génétiques, est computationellement très lourde. Un goulot d'étranglement majeur réside dans la reconstitution des événements de substitution le long de l'arbre, qui est répétée à chaque itération du processus (Lartillot, 2006; Mateiu and Rannala, 2006). Lors de ma thèse, Nicolas Lartillot a développé la méthode FastCoevol [17], une version plus rapide de Coevol, qui utilise l'approximation dite des mappings de substitution (Romiguier *et al.*, 2012; Lemey *et al.*, 2012). Cette approximation consiste à reconstruire préalablement cette histoire des substitutions sous un modèle de référence plus simple et à l'utiliser pour calculer les corrélations. Dans ce contexte, on définit K_{ij}^S et K_{ij}^N le nombre de substitutions synonymes ou

17. C'est cette méthode qui est utilisée dans mon principal travail de thèse (chapitre 5).

non-synonymes pour le gène i sur la branche j ainsi que L_{ij}^S et L_{ij}^N le nombre d'opportunités mutationelles synonymes et non-synonymes pour les mêmes gènes. La vraisemblance du modèle d'approximation des mapping pour chaque gène et branche de longueur l_{ij} et chaque d_N/d_S mesuré par ω_{ij}, correspond alors à :

$$K_{ij}^S \sim Poisson(l_{ij} L_{ij}^S) \tag{3.21}$$

$$K_{ij}^N \sim Poisson(l_{ij} \omega_{ij} L_{ij}^N) \tag{3.22}$$

On peut ensuite obtenir des estimations des l_{ij} et des ω_{ij} :

$$\hat{l_{ij}} = \frac{K_{ij}^S}{L_{ij}^S} \tag{3.23}$$

$$\hat{\omega_{ij}} = \frac{\frac{K_{ij}^N}{L_{ij}^N}}{\frac{K_{ij}^S}{L_{ij}^S}} \tag{3.24}$$

En plus de cela, FastCoevol contient un niveau additionnel de stochasticité qui permet de prendre en compte des déviations de court terme du d_S et d_N/d_S par rapport à la tendance de long terme capturée par le mouvement Brownien. Cette implémentation est similaire à celle introduite dans Lartillot *et al.* (2016).

4

Étude de la relation entre Ne et intensité de la sélection

Dans le chapitre précédent, j'ai détaillé les processus d'acquisition de données micro et macro-évolutives (section 3.1) puis j'ai abordé différentes méthodes permettant de les analyser, soit en séparant les deux échelles (section 3.2 et section 3.3), soit en les utilisant conjointement (section 3.4). J'ai également introduit les méthodes statistiques qui permettent d'évaluer les relations évolutives entre différents traits d'intérêt (section 3.5).

Ces différents types de données et méthodes permettent d'évaluer empiriquement les attendus de la théorie quasi-neutre de Ohta (1992), évoqués dans le chapitre 2. En résumé, Ohta et Kimura ont suggéré que la plupart des mutations dans les gènes correspondent à des mutations délétères, rapidement éliminées par sélection naturelle, et que les autres mutations, que l'on observe dans le polymorphisme et la divergence, correspondent majoritairement à des mutations neutres, soumises à dérive génétique. Entre ces deux catégories, il existe un continuum d'effet des mutations incluant des mutations faiblement délétères. Ces dernières vont mettre plus de temps à être détectées par la sélection naturelle et peuvent, selon le contexte, être fixées entre temps par la dérive génétique. Pour Ohta et Kimura, c'est la taille efficace, N_e, qui détermine ce contexte. Plus N_e est petit et plus la dérive génétique est intense et va potentiellement fixer des mutations non-synonymes déletères qui n'ont pas encore été éliminées par la sélection. A l'échelle des populations, cela se traduit par un fardeau mutationel plus important pour les petites populations, qui, à l'échelle de la divergence, engendre une accumulation de substitutions délétères limitant l'adaptation des organismes à leur environnement. Cette relation entre N_e et l'intensité de la sélection et son impact aux échelles micro et macro-évolutives reste à vérifier empiriquement.

Concrètement, à l'échelle macro-évolutive, on peut mesurer l'intensité de la sélection par le d_N/d_S (voir sous-section 3.2.2). Comme N_e n'est pas mesurable directement, on utilise des proxies pour s'en approcher comme les traits d'histoire de vie qui seraient des proxies indirects de N_e. On attend alors une relation positive entre d_N/d_S et les traits d'histoire de vie. À l'échelle micro-évolutive, on utilise π_N/π_S comme mesure de l'intensité de la sélection et π_S, comme proxy, cette fois direct, de N_e (voir sous-section 3.3.4). Ici, on s'attend à une relation négative entre π_N/π_S et π_S.

Plusieurs études empiriques ont exploré ces deux types de corrélations à différentes échelles taxonomiques et avec des méthodologies variées. Dans ce chapitre, je compte présenter un bref état des lieux des études portant sur la relation entre proxies de N_e et intensité de la sélection. Je commencerai par examiner celles qui abordent séparément les échelles macro et micro-évolutives en étudiant d'une part les relations entre d_N/d_S et les traits d'histoire de vie, et d'autre part celles entre π_N/π_S et π_S. Je discuterai également des quelques travaux qui ont tenté de confronter les deux échelles micro et macro-évolutives dans leurs mesures respectives d'intensité de la sélection et N_e. Ensuite, j'aborderai les études qui, au lieu d'approximer N_e, ont comparé des mesures de d_N/d_S et/ou π_N/π_S en s'appuyant sur des contrastes écologiques connus pour influencer N_e, ainsi que les études qui ont tenté de reconstruire directement N_e le long de la phylogénie. Enfin, je présenterai les objectifs de mon travail de thèse et comment celui-ci s'inscrit de façon décisive dans l'étude des relations entre N_e et intensité de la sélection, chez les mammifères, en combinant, de façon intégrative, les échelles micro et macro-évolutives.

4.1 d_N/d_S versus traits d'histoire de vie

Concernant la relation présupposée positive entre d_N/d_S et traits d'histoire de vie, il existe plusieurs études à des échelles taxonomiques variées. Une première étude d'intérêt est celle de Figuet *et al.* (2016) à l'échelle des amniotes. Les auteurs et autrices [1] ont utilisé un jeu de données de 1077 gènes orthologues à environ 100 amniotes afin d'estimer un d_N/d_S par espèce. Iels ont ensuite testé la corrélation entre ces mesures de d_N/d_S et trois traits d'histoire de vie (masse, maturité et longévité), sans corriger pour l'inertie phylogénétique. A l'aide d'une régression linéaire, les auteurs et autrices observent une relation positive entre le d_N/d_S et la longévité avec un coefficient de corrélation de 0.69 pour les mammifères et 0.68 pour les reptiles (figure 4.1.A, points rouges et verts). Les autres traits d'histoire de vie réagissent de la même façon. En revanche, iels n'observent pas de relation significative concernant le groupe des oiseaux (figure 4.1.A, points bleus) et suspectent alors que les traits d'histoire de vie ne représentent pas N_e pour ce

1. Dans ce chapitre, contrairement aux précédents, j'ai choisi de porter une attention particulière à la présence de femmes dans les co-auteurs des articles cités et à les rendre visibles quand il y en a en utilisant l'écriture inclusive. Bien sûr, les femmes sont présentes dans tous les autres chapitres.

groupe (ce qui sera ensuite démenti dans le même article, voir ci-dessous).

A l'échelle plus restreinte des mammifères, on retrouve l'étude de Popadin *et al.* (2007) où les auteurs et autrices utilisent les séquences de 13 protéines mitochondriales issues de 110 espèces de mammifères pour reconstruire un d_N/d_S par espèces (nommé k_a/k_s). Iels confrontent ces mesures à des données de masses par une régression linéaire et observent une corrélation positive avec un coefficient de corrélation de 0.4 (figure 4.1.B). Popadin *et al.* (2007) ont également appliqué à leurs données un modèle qui prend en compte la non-indépendance phylogénétique et observent que, bien que la corrélation soit plus faible, celle-ci reste significative. Une autre étude, qui utilise cette fois un d_N/d_S reconstruit à partir de séquences nucléaires pour 17 espèces, observe une relation positive entre d_N/d_S et le temps de génération (r=0.73) (Nikolaev *et al.*, 2007).

Comme montré dans la sous-section 3.5.3, le même type d'étude peut être réalisé de façon intégrative, plutôt que séquentielle, ce qui permet de prendre en compte directement la phylogénie ainsi que l'incertitude sur l'estimation du d_N/d_S. Dans l'article qui présente la méthode Coevol (Lartillot and Poujol, 2011), les auteurs ont testé leur modèle en utilisant un alignement de séquences mitochondriales issues de 410 thériens dont 29 marsupiaux et 381 placentaires, et trois traits d'histoire de vie (maturité sexuelle, masse et longévité), tiré de Nabholz *et al.* (2008). A l'échelle des 410 thériens, les auteurs observent une relation positive significative entre d_N/d_S et la masse (r=0.58) ainsi que la longévité (r=0.13) mais pas la maturité. A l'échelle restreinte du groupe des carnivores, il n'y a que la relation entre d_N/d_S et masse qui est retrouvée significative (r=0.9). Étonnamment, dans une autre étude utilisant la même méthode, mais sur un jeu de données plus riche de 17 gènes codant nucléaires partagés par 73 mammifères placentaires, Lartillot and Delsuc (2012) obtiennent une relation positive entre d_N/d_S et la maturité (r=0.36) et la longévité (r=0.27) mais pas avec la masse.

Enfin, à l'échelle des primates, Brevet and Lartillot (2021) utilisent Coevol sur un alignement de 54 gènes présents chez 61 primates. Cependant, ils n'observent pas de corrélation entre d_N/d_S et les traits d'histoire de vie (masse, maturité et longévité) à l'exception d'une faible corrélation avec la longévité (r=0.49).

Bien sûr, beaucoup d'autres études ont tenté de mesurer la relation entre les traits d'histoire de vie et d_N/d_S à différentes échelles taxonomiques (Martin and Palumbi, 1993; Ohta, 1993; Bromham *et al.*, 1996; Lanfear *et al.*, 2007; Nabholz *et al.*, 2013). Cependant, ces études montrent, elles aussi, des résultats contrastés. Ces divergences mettent en lumière l'importance de facteurs méthodologiques comme la méthode d'étude (séquentielle ou intégrative), le type de données (mitochondrial ou nucléaire), la quantité de données, et l'échelle taxonomique, des aspects qui restent difficiles à démêler pour l'instant. Néanmoins, globalement, ces différentes études semblent confirmer l'existence d'une relation négative entre intensité de la sélection et N_e, à l'échelle macro-évolutive.

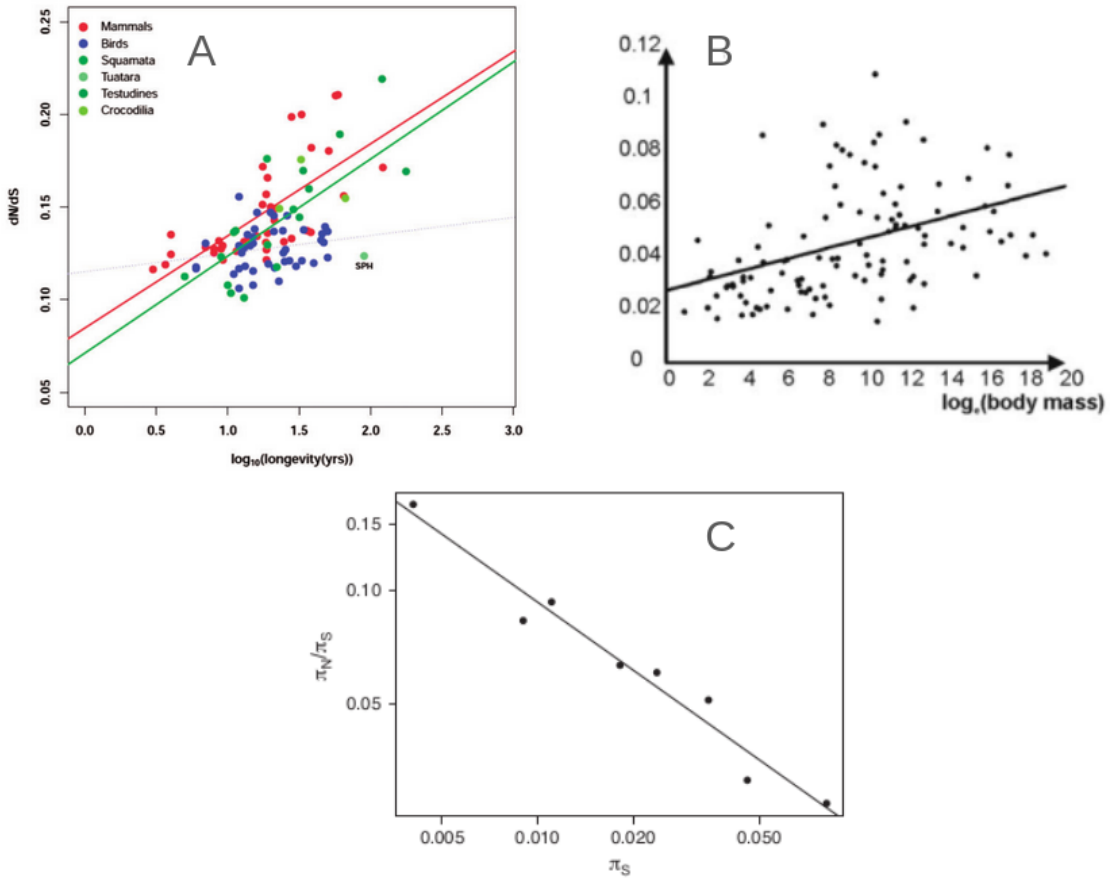


Figure 4.1 – *3 exemples d'étude de corrélations entre intensité de la sélection et proxies de N_e.* ***A*** *: Etude de Figuet* et al. *(2016) concernant la relation entre d_N/d_S et la longévité. On retrouve une corrélation significative pour les mammifères et reptiles mais pas pour les oiseaux. L'inertie phylogénétique n'est pas prise en compte.* ***B*** *: Etude de Popadin* et al. *(2007) concernant les mammifères avec une relation positive entre d_N/d_S et la masse. L'inertie phylogénétique n'est pas prise en compte.* ***C*** *: Etude de James* et al. *(2017) concernant les mammifères avec une relation négative entre π_N/π_S et π_S, en prenant en compte l'inertie phylogénétique.*

4.2 π_N/π_S versus π_S

Concernant la relation attendue négative entre π_S et π_N/π_S, un ensemble plus restreint d'études ont été réalisées. Parmi elles, on retrouve celle de James *et al.* (2017) à l'échelle des mammifères, qui utilise un jeu de données de 751 espèces contenant des données de polymorphisme mitochondrial. Les autrices et auteurs ont également développé une méthode qui permet de prendre en compte la non-indépendance phylogénétique en regroupant les espèces en paires d'espèces sœurs ou en groupes plus larges. Leur analyse révèle une forte corrélation négative entre les mesures en log de π_S et π_N/π_S (r=-0.98) (figure 4.1.C). Iles ont également calculé la pente de la corrélation et ont observé que celle-ci varie entre groupes d'espèces, ce qui est un point très intéressant pour les études concernant la relation entre le paramètre de forme de la DFE et la pente de la relation entre N_e et l'efficacité de la sélection (Welch *et al.*, 2008) (voir section 7.1).

À l'échelle des primates, Brevet and Lartillot (2021) utilisent des données transcriptomique pour estimer un π_S et un π_N/π_S et observent une relation négative entre les deux entitée (r=-0.78).

D'autres études ont montré cette même relation à différentes échelles taxonomiques (Piganeau and Eyre-Walker, 2009; Chen *et al.*, 2017) et permettent de confirmer une relation entre intensité de la sélection et N_e à l'échelle micro-évolutive.

4.3 Contraster les deux échelles évolutives

Étant donné que la relation entre intensité de la sélection et N_e semble globalement validée de façon indépendante aux échelles micro et macro-évolutive, on s'attend à ce que les différents estimateurs de N_e et d'intensité de la sélection, corrèlent également entre eux.

À ma connaissance, peu d'études se sont penchées sur la relation entre π_N/π_S et d_N/d_S à part celles de Lefébure *et al.* (2017) et Brevet and Lartillot (2021), qui ne montrent cependant pas de relation significative entre les deux traits.

Concernant la relation supposée négative entre π_S et traits d'histoire de vie, des articles comme celui de Figuet *et al.* (2016) ont permis de l'observer. En effet, dans leur article précédemment évoqué, concernant les amniotes, Figuet *et al.* (2016) n'observent pas de relation significative entre d_N/d_S et traits d'histoire de vie chez les oiseaux ce qui les pousse à questionner l'usage des traits d'histoire de vie comme proxy de N_e pour ce groupe taxonomique. Iels mesurent alors, pour 45 espèces d'oiseaux, une hétérozygotie sur l'ensemble du génome et observent une relation négative significative entre cette hétérozygotie et les trois traits d'histoire de vie. Ce résultat vient contredire leur hypothèse que les traits d'histoire de vie ne représentent pas N_e chez les oiseaux et valide la relation entre ces deux proxies de N_e issus d'échelle évolutive différente. Iels posent cependant la question des raisons pour lesquelles le d_N/d_S ne semble pas corrélé à N_e chez les oiseaux et explorent ce point dans un article plus tardif (Botero-Castro *et al.*, 2017).

Dans une autre étude, Romiguier *et al.* (2014a) montrent une relation négative entre π_S, cette fois mesurée à partir de données populationnelles, et des traits d'histoires de vie chez 76 animaux. Iels observent notamment un fort effet de la taille des propagules (r=0.66). Les auteurs et autrices ne prennent pas en compte l'inertie phylogénétique dans leurs résultats finaux, mais montrent que son intégration ne change pas la significativité des résultats observés.

À noter que d'autres études ne retrouvent pas de corrélation entre π_S et les traits d'histoire de vie comme celle de Brevet and Lartillot (2021) à l'échelle des primates.

Ces résultats, contrastés concernant la relation entre mesures d'intensité de la sélection ou de N_e à différentes échelles temporelles, montrent un potentiel découplage entre les données de court et long terme. Ce découplage pourrait s'expliquer par l'effet du groupe taxonomique étudié, le type de méthode statistique utilisée, la qualité variable des données, ou encore refléter un véritable phénomène biologique.

4.4 Estimer N_e autrement

Afin de s'affranchir de l'estimation quantitative de N_e par un proxy plus ou moins direct, plusieurs études ont mis en œuvre des protocoles permettant un accès

plus direct à N_e en contrastant les séquences de groupes d'espèces proches dont on pense, avant tout pour des raisons écologiques, qu'ils ont un N_e différent, ou bien en reconstruisant directement N_e le long de la phylogénie.

4.4.1 Entre taille de population contrastée

Parmi les groupes d'espèces proches que l'on peut contraster pour leur N_e, on retrouve notamment les espèces qui ont vécu des changements d'habitats, vers un habitat plus contraignant. C'est le cas par exemple des espèces continentales qui colonisent des iles ou bien des espèces vivant en surface qui colonisent des milieux souterrains. Dans les deux cas, les espèces devenues insulaires ou souterraines ont des tailles efficaces de population plus petites, car elles ont récemment vécu un événement de colonisation réduisant leur taille de population, en plus d'être soumises à un environnement plus contraint. On peut ensuite calculer un d_N/d_S et/ou un π_N/π_S par groupe et vérifier si leurs variations coïncident avec les attendus concernant les groupes de statuts écologiques ou biogéographiques distincts. On s'attend notamment à un plus grand d_N/d_S et π_N/π_S pour les espèces continentales par rapport aux insulaires, et de surface par rapport aux souterraines.

Concernant le contraste entre espèces insulaires et continentales, dans leur article, Leroy *et al.* (2021) ont utilisé des données populationnelles concernant 25 espèces de passereaux dont 14 espèces insulaires et 11 continentales représentant 4 événements indépendants d'insularisation. La taille efficace de chaque population a été estimée en utilisant un individu par espèce par une méthode PSMC (Li and Durbin, 2011), puis moyennée sur le dernier million d'années. Celle-ci a permis de confirmer, par une analyse prenant en compte la phylogénie (PGLS), que les espèces insulaires ont un N_e plus faible que les espèces continentales. Les auteurs et autrices confirment également les attendus quasi-neutre en montrant un π_N/π_S plus élevée pour les espèces insulaires ainsi qu'un π_S plus faible.

Pour le contraste entre espèces souterraines et de surfaces, Lefébure *et al.* (2017) utilisent le transcriptome de 22 isopodes qui représentent 11 changements indépendants d'un milieu de surface à souterrain. Iels calculent, pour chaque espèce,

un d_N/d_S, un π_N/π_S et un « $\hat{\theta_W}$ »[2], lequel représente le taux de mutation populationnel. Ce taux est proportionnel au produit de N_e et μ, et reflète donc N_e à l'échelle micro-évolutive. Au niveau macro-évolutif, les auteurs et autrices observent que les espèces souterraines ont un plus grand d_N/d_S pour 8 paires d'espèces sur 11 et montrent que ce résultat ne semble pas lié à de l'adaptation suite à un changement d'environnement, mais bien à une diminution de l'efficacité de la sélection. A l'échelle micro-évolutive, les auteurs et autrices ont utilisé un modèle de régression linéaire qui prend en compte l'inertie phylogénétique (un modèle PGLS) et observent une corrélation négative entre $\hat{\theta_W}$ et π_N/π_S (r=0.66). Concernant le π_N/π_S, 6 groupes sur 11 présentent un polymorphisme plus élevé pour les espèces souterraines. Ce dernier point est en accord avec les observations déjà faites concernant le d_N/d_S. À noter qu'il n'y a pas de corrélation significative entre le d_N/d_S et le π_N/π_S. Les résultats sont cependant moins clairs à propos du $\hat{\theta_W}$, pour lequel on s'attend à observer une valeur plus faible pour les espèces souterraines. En effet, parmi les 11 couples d'espèce, 6 présentent un $\hat{\theta_W}$ inférieur pour les espèces souterraines et tandis que 3 présentent un résultat inverse. Iels concluent donc qu'il n'y a pas de différences significatives de $\hat{\theta_W}$, proxy de N_e, entre espèce souterraines et de surface, ce qui est plutôt surprenant. Lefébure *et al.* (2017) expliquent ces résultats contrastés entre les échelles micro et macro-évolutives par un N_e de court terme qui fluctuerait rapidement autour d'un N_e moyen, ce qui impacte le $\hat{\theta_W}$ et le π_N/π_S qui deviennent des proxies bruités de N_e, contrairement au d_N/d_S. Iels supposent aussi que les mesures de polymorphismes sont plus impactées par les erreurs de mesures que le d_N/d_S.

2. Qui n'est pas le $\hat{\theta_W}$ présenté en section 3.3.

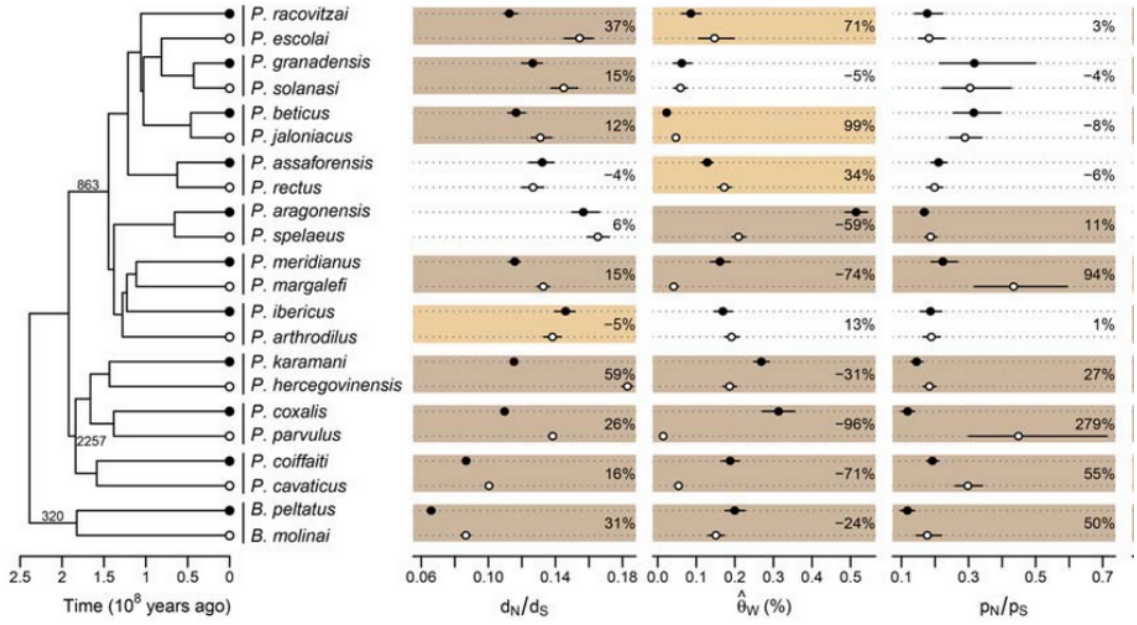


Figure 4.2 – *Mesure de d_N/d_S, π_N/π_S et de $\hat{\theta_W}$ (proxy de N_e court terme) pour onze paires d'isopodes de surface (cercles blanc) et souterrain (cercle noir). Les boxes marron foncé indiquent une corrélation entre une des trois mesures avec l'etat surface/souterain, en accord avec les attendus quasi-neutre. Les boxes marron clair indiquent une corrélation dans le sens opposées. Les boxes non colorées indiquent qu'ils n'y a pas de différences significatives entre les espèces de la paire. Etude menée par Lefébure* et al. *(2017).*

Enfin, il existe également des études qui contrastent l'effet de la sélection chez les insectes eusociaux, à plus petit N_e, par rapport aux non eusociaux (Romiguier *et al.*, 2014b; Weyna and Romiguier, 2021). Ces études continuent de confirmer la théorie quasi-neutre.

4.4.2 Reconstruire directement N_e le long de la phylogénie

Dans leur étude, Brevet and Lartillot (2021) ont reconstruit directement N_e à partir des estimations de π_S et du taux de mutation par site et par génération (μ) en utilisant la relation $N_e = \pi_s/4\mu$. Cela permet de dissocier les effets de μ qui agit lui aussi sur le π_S et donc l'éloigne de sa représentation de N_e. Pour reconstruire les variations de μ le long de la phylogénie, les auteurs utilisent une horloge moléculaire relaxée qui fournit des taux de mutation par site et par an, ainsi que des données de temps de génération (τ). Brevet and Lartillot (2021) ont ainsi développé deux façons de modéliser la reconstruction de N_e et ses relations avec les autres traits. Une

première approche est l'approche phénoménologique. Dans ce contexte, l'évolution de différents traits (π_S, π_N/π_S, traits d'histoire de vie, d_S et d_N/d_S) est reconstruite conjointement le long d'une phylogénie, dans un même mouvement brownien multivarié, puis μ est extrait par :

$$\mu = d_S \cdot \tau \tag{4.1}$$

$$ln\mu = lnd_S + ln\tau \tag{4.2}$$

quant à N_e,

$$\pi_S = 4N_e\mu \tag{4.3}$$

et donc,

$$lnN_e = ln\pi_S - ln\mu - ln4 \tag{4.4}$$

Les corrélations entre chaque trait sont ensuite étudiées. Le modèle a été appliqué sur un jeu de données de primates (54 gènes, 61 espèces), déja présenté plus haut, et montre une relation négative entre d_N/d_S et N_e (r=-0.58) ainsi qu'entre N_e et les données de polymorphisme. Cette dernière relation, est plutôt attendue puisque N_e est reconstruit à partir des données de polymorphisme. Cependant, sachant que les auteurs n'ont pas observé de relation directe entre d_N/d_S et π_S ou μ, les briques composant le N_e, observer une corrélation entre d_N/d_S et N_e est plutôt surprenant. On note qu'il n'y a toujours pas de corrélation observée entre l'intensité de la sélection aux deux échelles évolutives et les traits d'histoire de vie.

Ces différentes relations sont difficilement interprétables, car le modèle reste agnostique concernant les relations entre d_N/d_S et π_N/π_S avec N_e. Or, il se trouve que certaines études ont démontré une relation entre la forme de la DFE et l'impact des différentes forces évolutives sur la divergence et le polymorphisme (Eyre-Walker and Keightley, 2007; Welch *et al.*, 2008). On a notamment décrit que, sous l'hypothèse d'une DFE constante entre espèces, la pente de la relation entre d_N/d_S (ou π_N/π_S) en fonction de N_e correspond paramètre de forme de la DFE (β). Ceci faisant, Brevet and Lartillot (2021) ont développé un deuxième modèle, cette fois-ci mécaniste, qui utilise β pour prédire les variations entre espèces de d_N/d_S et π_N/π_S

en fonction de N_e via :

$$\begin{aligned} \pi_N/\pi_S(t) &= e^{\epsilon_1(t)} N_e(t)^{-\beta}, \\ d_N/d_S(t) &= e^{\epsilon_2(t)} N_e(t)^{-\beta} \end{aligned} \tag{4.5}$$

avec $\epsilon_i(t)$ qui correspondent aux mouvements browniens utilisés dans le modèle (Brevet and Lartillot, 2021). Ce modèle permet d'obtenir des résultats plus facilement interprétables dans un contexte de théorie quasi neutre (voir discussion section 7.1). Toutefois, les auteurs font un certain nombre d'hypothèses fortes, à savoir que la DFE est constante entre espèces, qu'il n'y a pas de sélection positive, que les N_e aux échelles micro et macro-évolutives sont identiques et que d_N/d_S et π_N/π_S sont littéralement des proxies de N_e.

4.5 Objectifs de thèse

Que ça soit à l'échelle micro ou macro-évolutive, les résultats des études empiriques concernant les relations entre intensité de la sélection et proxies de N_e sont plutôt contrastés. Au niveau macro évolutif, un lien entre d_N/d_S et traits d'histoire de vie est retrouvé (Figuet *et al.*, 2016; Popadin *et al.*, 2007; Lartillot and Delsuc, 2012; Lartillot and Poujol, 2011) mais il semble dépendre de l'échelle taxonomique étudiée (par de corrélation pour les oiseaux Figuet *et al.* (2016) ni pour les primates (Brevet and Lartillot, 2021)) et peut varier entre traits d'histoire de vie ou jeu de données. Au niveau micro-évolutif, bien que moins d'études aient été menées, on retrouve des relations fortes entre π_S et π_N/π_S (James *et al.*, 2017; Brevet and Lartillot, 2021). Concernant les mesures de N_e, la relation entre les deux échelles est plutôt instable avec des relations retrouvées parfois significatives dans certaines études (Romiguier *et al.*, 2014a; Figuet *et al.*, 2016) mais absentes dans d'autres (Lefébure *et al.*, 2017; Brevet and Lartillot, 2021). L'ensemble de ces études, qui varient considérablement en termes de méthodologies (séquentielles ou intégratives), de types et de quantité de données (influant sur la puissance de détection des relations), ainsi que dans la prise en compte de l'inertie phylogénétique, ne permet pas de conclure de manière décisive sur le lien supposé fort entre N_e et l'intensité de la sélection, tel que décrit par la théorie quasi-neutre de l'évolution. Il manque donc d'une étude complète qui articule de manière

cohérente les échelles micro et macro-évolutive sur un jeu de données suffisamment riche à la fois en gènes et en espèces, à une échelle taxonomique appropriée et qui prend en compte l'inertie phylogénétique. En particulier, concernant le génome nucléaire des mammifères, on retrouve principalement des études qui se placent à l'échelle macro-évolutive. À ce jour, on trouve peu d'études à l'échelle micro-évolutive (à l'exception de celle de James *et al.* (2017), mais qui utilise des séquences mitochondriales), et aucune ne traite de la jonction entre micro et macro-évolution. Pourtant, la relation entre les traits d'histoire de vie et π_S est particulièrement importante. En effet, l'idée de prendre les traits d'histoire de vie comme proxies de N_e repose avant tout sur l'idée reçue que les gros animaux ont des plus petits N_e. Or, on ne connait aucune évidence indépendante et quantitative concernant cette relation chez les mammifères.

Lors de ma thèse, j'ai réalisé cette étude à l'échelle des mammifères en utilisant un jeu de données, que j'ai moi-même constitué, et qui contient de plus de 6000 gènes orthologues à environ 150 mammifères. Afin d'inclure des données populationnelles, propres à l'échelle micro-évolutive, j'ai utilisé le fait que l'hétérozygotie, mesurée sur un génome diploïde recombinant correspond à une mesure du polymorphisme de la population. Pour ce faire, je suis revenue aux lectures à partir desquelles les génomes ont été assemblés, et j'ai identifié les positions hétérozygotes synonymes et non synonymes présentes dans les gènes orthologues précédemment identifiés. J'ai ensuite utilisé FastCoevol, une version améliorée de méthode Coevol (Lartillot and Poujol (2011) et sous-section 3.5.3), pour reconstruire de façon intégrative les variations de d_N/d_S, π_N/π_S, π_S et des traits d'histoire de vie conjointement le long de la phylogénie. FastCoevol fournis également une matrice de corrélation des différents traits, ce qui m'a permis d'étudier les relations entre intensité de la sélection au niveau micro et macro-évolutif ainsi que le lien entre les différents proxies de N_e ou de l'efficacité de la sélection, entre échelles. La constitution du jeu de données et la vérification empirique des attendus quasi-neutres font l'objet de mon principal travail de thèse (chapitre 5).

Aussi, une hypothèse centrale de mon travail de thèse est que le polymorphisme d'une population peut être approximé par la mesure de l'hétérozygotie d'un seul

individu de cette population. Cette hypothèse est valide dans une population idéale, mais peut ne pas s'appliquer en dehors de ce cadre. Ainsi, dans une seconde partie de mon travail (chapitre 6), j'ai utilisé des données populationnelles provenant de plusieurs espèces de vertébrés pour estimer, pour chaque individu, deux mesures de diversité populationnelle à partir d'un unique génome. Ces mesures sont l'hétérozygotie ainsi qu'une mesure issue d'une méthode de type PSMC (Li and Durbin, 2011). La méthode PSMC, comme évoquée en section 3.3, est une méthode qui utilise un génome diploïde pour reconstruire les variations de N_e dans le temps. Afin d'en obtenir une mesure ponctuelle, comparable à celle de l'hétérozygotie, j'ai d'abord converti le N_e reconstruit en le divisant par μ afin qu'il représente une mesure de diversité, puis j'ai réalisé une moyenne des variations de N_e dans le temps. Pour ces deux mesures, j'ai utilisé une décomposition de variance afin d'examiner l'effet de l'individu, de la population et de l'espèce sur la variance totale. Cela permet notamment de déterminer le bruit ajouté à l'estimation d'une mesure de diversité populationnelle lorsqu'on utilise un seul génome. La comparaison des deux mesures permet également de déterminer laquelle est la plus sensible aux individus aberrants (par exemple un individu exceptionnellement très consanguin) et aux variations individuelles et donc représente le mieux sa population.

Finalement, ce travail de thèse permet de confirmer les attendus quasi-neutre chez les mammifères, aux échelles micro et macro-évolutives et démontre que l'utilisation de l'hétérozygotie comme mesure du polymorphisme est appropriée, bien qu'elle introduise un certain bruit dans l'estimation de N_e.

Deuxième partie

Études

5

Validation empirique de la théorie neutre chez les mammifères aux échelles de la phylogénie et de la génétique des populations

Contexte :

Au début de mon stage de deuxième année de master, en janvier 2021, nous avons souhaité explorer la théorie quasi-neutre de l'évolution qui présuppose une relation entre intensité de la sélection et N_e (chapitre 2). Pour ce faire, nous nous sommes placés à l'échelle micro-évolutive et nous avons étudié la pente de la relation entre $log(\pi_N/\pi_S)$ et $log(\pi_S)$ à partir de séquences de cythochrome b de mammifères. Cela nous a permis d'obtenir des résultats comparables à ceux fournis par James *et al.* (2017) sur des séquences mitochondriales. Environ un mois après, nous avons pris connaissance du jeu de donnée publié par le Zoonomia Consortium (2020) qui présente un alignement des génomes complets de 240 espèces de mammifères et des fichiers VCF pour 120 d'entre elles. Ces fichiers VCF, conçus ici à partir d'un unique génome diploïde, permettent d'estimer l'hétérozygotie de l'individu, qui, sous des conditions restant à mieux caractériser (voir chapitre 6), pourrait servir de mesure du polymorphisme de son espèce. J'ai très vite décidé d'exploiter ce jeu de donnée d'une richesse exceptionnelle, car celui-ci permet d'estimer π_N/π_S et π_S pour la moitié des génomes de l'alignement, mais également de reconstruire le d_N/d_S le long de la phylogénie. Cette reconstruction du d_N/d_S avec l'ajout d'information sur les traits d'histoire de vie permet d'étudier sous un deuxième angle, celui de la macro-évolution, la question de la relation entre intensité de la sélection et N_e. Plus encore, disposer simultanément de données sur le polymorphisme et la divergence

des mêmes génomes, fournit l'occasion idéale de confronter les échelles micro- et macro-évolutives tout en étudiant leur réponse commune à un même mécanisme, la dérive génétique, ainsi que ses variations à travers différentes échelles de temps.

Pour réaliser l'ensemble de ces analyses, j'ai utilisé la méthode intégrative FastCoevol, une version plus rapide de Coevol (Lartillot and Delsuc, 2012), développée par Nicolas lors de ma thèse (sous-section 3.5.3). Coevol comme FastCoevol permettent de reconstruire l'évolution de traits d'intérêt, moléculaires comme écologiques, le long d'une phylogénie et d'étudier les corrélations entre ces traits.

Concernant la constitution du jeu de données, les 240 génomes fournis par Zoonomia (2020) ont nécessité une quantité de traitement bio-informatique non négligeable. En effet, le jeu de données est constitué de génomes non annotés ou bien annotés avec des méthodes aux sensibilités différentes. Ainsi, j'ai annoté par moi-même les gènes orthologues des 240 génomes de façon consistante et reproductible en utilisant l'outil BUSCO (Simão *et al.*, 2015). Je les ai ensuite alignés et filtrés en ne conservant que ceux de meilleure qualité et communs à un maximum de mammifères. Après un an de thèse, et suite à différents essais de traitement des données, je me suis rendu compte que la qualité initiale des données utilisées était un facteur très important. Or, certains des génomes proposés par Zoonomia ont des scores BUSCO très faibles, ce qui indique une mauvaise qualité de séquençage pour ces génomes. Pour continuer à utiliser ces génomes, il est nécessaire de masquer les régions qui sont mal séquencées afin de ne pas les prendre en compte. Cela peut être réalisé en utilisant les mesures de couverture, position par position, pour détecter les zones qui dévient de la couverture moyenne du génome concerné. Cependant, les fichiers BAM, contenant cette information, ne sont pas fournis, bien qu'ils aient été générés pour au moins 120 des génomes séquencés par le consortium.

En parallèle, je suis entrée en contact avec David Enard, par l'intermédiaire de Laurent Duret, qui a fait le même constat concernant le manque de disponibilité de fichiers pourtant existants [1]. David était alors en train de remapper les lectures brutes

1. Nous avons pris contact via ce Tweet de novembre 2021 : " *The number of mammalian genomes out there with no submitted SRA is a little bit depressing. We are unable to get heterozygosity data for a lot of them because raw reads have not been submitted. Not good.*"

de séquençage sur les génomes correspondants pour 260 espèces de mammifères afin de réaliser du *variant calling* et donc écrire, pour chaque génome, des fichiers VCF et BAM. Les fichiers VCF permettent d'estimer une hétérozygotie par individu et les fichiers BAM permettent de masquer certaines régions des génomes en fonction de leur couverture. J'ai, par conséquent, collaboré avec lui pour bénéficier de ces fichiers VCF et BAM, et j'ai récupéré les génomes concernés pour les annoter de la même façon que pour les génomes de Zoonomia. Assez vite, en comparant les résultats obtenus avec le jeu de données Zoonomia, je me suis rendu compte que les génomes utilisés par David étaient de meilleure qualité, mais surtout, que l'usage de la couverture pour masquer les génomes augmentait fortement la qualité de mes analyses pilotes avec Coevol. De plus, le fait de disposer d'une mesure d'hétérozygotie pour chaque espèce a certainement permis de réduire les incertitudes lors de la reconstruction de π_S le long de la phylogénie. J'ai donc arrêté d'exploiter le jeu de données Zoonomia pour ne me consacrer qu'à celui fourni par David Enard.

C'est ainsi que j'ai progressivement développé un jeu de donnée de bonne qualité et complet contenant les alignements de 150 génomes de mammifères, chacun accompagné de données de polymorphisme. Par la suite, j'ai utilisé ce jeu de données pour examiner les attendus de la théorie neutre concernant la relation entre intensité de la sélection (représentée par d_N/d_S et π_N/π_S) et N_e (représenté par π_S et les traits d'histoire de vie). J'ai obtenu des résultats plutôt concluant avec une relation positive entre d_N/d_S et traits d'histoire de vie et une relation négative entre π_N/π_S et π_S, ce qui est congruent avec la théorie neutre de l'évolution, à l'échelle des mammifères. J'ai également observé des relations contrastées entre π_S et les traits d'histoire de vie ainsi qu'une relation positive étonnamment élevée entre π_N/π_S et d_N/d_S. Ces deux derniers points sont discutés dans l'article et dans la discussion générale (section 7.1).

Dans cet article, je me suis beaucoup posé la question des choix de filtrage de mes données et de la limite floue avec le *p-hacking* (discuté en sous-section 7.3.2) car ceux-ci ont progressivement été affinés et ajustés au cours même de l'analyse. C'est pour cela que j'ai choisi d'y présenter les matrices de corrélation entre traits en fonction de plusieurs types d'action sur les données, plutôt que de choisir les résultats qui me complaisaient le plus dans le contexte d'une validation de la théorie

neutre de l'évolution.

Quoi qu'il en soit, je pense que ce travail poursuit celui déjà entamé par mes confrères et consœurs concernant la validation empirique de la théorie quasi-neutre de l'évolution (chapitre 4). J'ai pour espoir qu'il apporte un éclaircissement décisif sur la relation entre intensité de la sélection et N_e, en proposant une emphase sur la nécessité de l'utilisation de données riches et de bonne qualité ainsi que de méthodes comparatives intégratives.

Contributions :

J'ai co-signé cet article avec **David Enard** et **Nicolas Lartillot**. David Enard a généré les fichiers VCF et BAM de l'étude et Nicolas a développé les outils Coevol et FastCoevol. J'ai mis en place l'ensemble du jeu de données et Nicolas et moi avons analysé ensemble les résultats obtenus. J'ai écrit la première version de l'article avant que nous la retravaillions tous les deux. Tout au long de ce travail, j'ai bénéficié des retours de **Laurent Duret**, **Carina Mugal**, **Thibault Latrille** et **Julien Joseph**. Merci à elles et eux.

Empirical validation of the nearly neutral theory at divergence and population genomic scale using 150 mammals genomes

M. Bastian[1], **D. Enard**[2], **N. Lartillot**[1]

[1]Laboratoire de Biométrie et Biologie Evolutive, UMR5558, Université Lyon 1, Villeurbanne, France

[2]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85719, USA

Corresponding author : melodie.bastian@univ-lyon1.fr

January 1, 2025

Abstract

The nearly-neutral theory predicts that the efficacy of selection against random drift is proportional to the effective population size (N_e). Efficacy of selection in coding sequences can be measured at the macro-evolutionary scale using d_N/d_S, and at the micro-evolutionary scale using π_N/π_S. Thus, in coding sequences, we expect a positive relation between d_N/d_S and life history traits (N_e proxies at macro scale) and a negative relation between π_N/π_S and π_S (N_e proxy at micro scale). In mammalian nuclear genomes, studies about these relations are scarce and mainly focused on macro-evolution, showing a positive relation between life history traits and d_N/d_S, although with some inconsistencies between studies as to the strength and significance of these correlations. At the micro-evolutionary scale, only mitochondrial genomes have been studied, revealing a negative relation between π_S and π_N/π_S(James *et al.*, 2017). There are no studies addressing N_e proxies or selection efficacy measures across evolutionary scales. Here we propose to test the nearly-neutral prediction in an integrative manner using the genomes of 150 mammals species and around 6000 orthologous genes. Our investigation spans two timescales: between species (phylogenetic) and within species (polymorphism), using for the latter a measure of heterozygosity in coding sequences, such as estimated by variant calling on those individuals used for genome assembly. For both scales, we compute measures of selection efficacy (d_N/d_S and π_N/π_S, respectively) and proxies of N_e (life history traits and π_S, respectively) and analyse their correlations. We confirm the positive correlation between d_N/d_S and life history traits. Notably, we also observe, for the first time in mammalian nuclear genomes, a relationship between π_N/π_S and π_S. These observations validate the prediction of the nearly-neutral theory separately for the micro- and macro-evolutionary scales in mammals. Across time scales, we infer a correlation of π_S and π_N/π_S with life history traits and with d_N/d_S ,

although the correlation between π_S and life-history traits appears to be weaker and more condition-dependent. Taken together, these observed correlations are globally consistent with the nearly-neutral expectations and can be explained by invoking that all variables are in fact correlated with a single hidden variable, which is N_e.

***K*eywords** population genomic, effective population size, selection efficacy, mammals, phylogenomic, micro-macro evolution, genetic drift

1 Introduction

1.1 Phylogeny and population genomic, two independent discipline ?

Recent methodological progress in phylogenetics and molecular evolution (reviewed in Russo *et al.* (2024)) makes it possible to study evolution while integrating complex mechanism like the heterogeneity of processes along genomes (Lartillot and Philippe, 2004) or GC-biased gene conversion (Bolívar *et al.*, 2018; Galtier and Duret, 2007). Moreover, the growing quantity of data together with methodological advances in the field offer the opportunity to study molecular evolution at the whole-genome scale (Begun *et al.*, 2007; Wolfe and Li, 2003) and across a wide range of species (Feng *et al.*, 2020; Zoonomia, 2020).

In parallel, theoretical and methodological progress in population genomics gives us the opportunity to disentangle the drivers of molecular diversity inside a population like demographic bottleneck, migration or population structure (Boitard *et al.*, 2016; Li and Durbin, 2011; Mathew and Jensen, 2015). Empirically, population genomics was initially limited by the lack of sequence data, restricting its application to model species like humans (Siva, 2008), fruit flies (Lack *et al.*, 2015), or domestic species like horses and mice (Al Abri *et al.*, 2020; Sherry *et al.*, 2001). However, this has been changed in recent years, and population genomics now data exist for a wide range of species like birds (Burri *et al.*, 2015) or invertebrates (Gayral *et al.*, 2013).

Phylogenomics and population genomics have thus been the subject of extensive research for several years. However, they are rarely used in tandem or even confronted (for exceptions, see: Brevet and Lartillot (2021); De Maio *et al.* (2015); Latrille *et al.* (2023)). Yet, it would be beneficial to include them within the same framework, as they both attempt to understand the same evolutionary mechanisms, although at different time scales. In fact, the evolutionary history of a species or a group of species, also known as the macro-evolutionary process, is the result of micro-evolutionary processes (population genetics) over a long period of time (Rolland *et al.*, 2023). Considering this, it becomes crucial to incorporate micro-evolutionary processes into the study of macro-evolution. Such an integration raises several challenges, both conceptual and methodological. We first need to arrive at a global understanding of how the population genetic processes modulate genetic diversity in the short term and impact genome evolution in the long term. We then have to identify the key parameters and how they vary between species, with a clear idea of how this variation is supposed to impact the observable patterns of intra and inter-specific variation. Finally, we need to recruit a robust comparative methodology to test those predictions against empirical data across a clade of interest.

1.2 The nearly-neutral theory and its predictions for protein-coding sequences

On the conceptual side, genome evolution is the result of the interplay between different evolutionary processes. Mutational processes create a pool of variants which then segregate in the population and can be fixed mainly under the combined action of selection and genetic drift. Among these evolutionary processes, genetic drift represents a particularly interesting one because it is a stochastic process which imposes a limit on adaptation at the genomic level, implying that many genomic features may in fact be non-adaptative (Lynch *et al.*, 2011; Lynch and Walsh, 2007).

A global unifying perspective on molecular evolution was first proposed by the neutral theory (Kimura, 1979; King and Jukes, 1969), which states that the majority of mutations are either deleterious and purified by natural selection or neutral. As a result, observed variation between and within species is essentially neutral. The nearly-neutral theory (Ohta, 1973, 1992) is a refinement of this neutral perspective, which relies on a more quantitative account of the impact of new mutations. Here, the fitness effects of mutation are likely to vary continuously, covering a broad range of effects, from neutral to strongly deleterious. This translates into a continuum of selection coefficients over all possible mutations, which is commonly referred to as the Distribution of Fitness Effects (DFE) (Eyre-Walker and Keightley, 2007; Welch *et al.*, 2008). In this context, the effective population size (N_e) acts like a threshold on the DFE. For example, the fate of a given mutation will be mainly determined by genetic drift if its selection coefficient is lower than $1/N_e$. Thus, since N_e differs between species, the nearly neutral theory predicts that the efficacy of selection against random genetic drift should also vary between species.

In the case of protein coding genes, mutations can be distinguished as either synonymous or non-synonymous. Synonymous mutations are often considered as neutral (or nearly-neutral) in mammals (Ohta, 1995). Such mutations can provide a meaningful neutral background against which to evaluate the fate of non-synonymous mutations, which on the other hand have typically a broad range of fitness effects, because they change the encoded amino acid (Eyre-Walker *et al.*, 2006; Eyre-Walker and Keightley, 2007). In the context of a continuous DFE, non-synonymous mutations whose fitness effect is too low to be efficiently detected by selection can be randomly lost or fixed by drift, even if this mutation is deleterious.

Using this contrast between the non-synonymous and the synonymous compartments leads to two key molecular quantities that can be measured or inferred based on coding sequences: at the population (or micro-evolutionary) scale, the ratio of non-synonymous over synonymous polymorphism (or π_N/π_S) and at the phylogenetic (or macro-evolutionary) scale, the ratio of non-synonymous over synonymous divergence (or d_N/d_S). These two metrics represent key measures of the efficacy of selection against random drift (or for short, in the following, *selection efficacy*). According to the nearly neutral theory, they are expected to decrease (i.e. more efficient selection against random drift) as a function of N_e at their respective scale.

1.3 Defining N_e (and its proxies)

Testing these predictions, however, raises the question of how to estimate N_e and more fundamentally, how to define it, depending on the time scale considered. There are different ways to define N_e (recently reviewed on Waples (2022)). Some will define it from a demographic point of view, when others invoke the effective

population size as a measure of genetic drift. Some authors argue for a more complex definition of N_e and derive the concept at different time scales (Nadachowska-Brzyska *et al.*, 2022; Müller *et al.*, 2022).

In our work, we will consider two types of N_e proxies, each defined at a different evolutionary scale. At micro-evolutionary scale, we use π_S as proxy of N_e based on the formula from coalescent theory : $\pi_S = 4N_e\mu$ where μ is the mutation rate per site, per generation (Charlesworth, 2009; Wright *et al.*, 1939). Of note, π_S depends on both N_e and μ, although it appears that most of the variation in π_S in contributed by the N_e component (Lynch *et al.*, 2023). At the macro-evolutionary scale, the molecular evolutionary theory doesn't provide a direct proxy of N_e. Furthermore, the exact meaning of a long-term N_e, which would stand for a measure of random drift at the macro-evolutionary scale, is not so well formalized. Nevertheless, a reasonable intuition would be that, if all populations are subject to recurrent demographic fluctuations, some species, because of their ecological or life-history characters, will have structurally lower genetic carrying capacities than other species. Particularly in mammals, large long-lived mammalian species such as primates, present lower structural genetic carrying capacities than small-bodied species such as rodents (Waples *et al.*, 2013). It is therefore reasonable to assume that, on average, large mammals tend to be characterized by a bigger impact of random drift in the evolution of their genomes (i.e. smaller long-term N_e). As in previous studies (Popadin *et al.*, 2007; Figuet *et al.*, 2016; Nabholz *et al.*, 2013), we thus opt for the use of life history traits as a proxy of long-term N_e, acknowledging that its relation with N_e is somewhat indirect and less well theoretically formalized compared to its short term equivalent, π_S.

These two proxies of N_e now defined, we want to interrogate the nearly-neutral predictions at the macro- and micro-evolutionary scales, by empirically examining the relations between π_S and π_N/π_S, which is expected to be negative, on the one hand, and the relation between life-history traits and d_N/d_S, which is expected to be positive, on the other hand. Moreover, considering that macroevolution is the result of the accumulation of microevolutionary processes over longer timescale, we want to question the congruence between these two scales.

A number of previous study have already attempted to test those predictions. At the macro-evolutionary scale, a positive correlation between life history traits and d_N/d_S as often been observed (Nabholz *et al.*, 2013; Popadin *et al.*, 2007; Figuet *et al.*, 2016; Romiguier *et al.*, 2014) which is often interpreted as a confirmation of the indirect correlation of long-term N_e with life-history traits. Actually, the literature on this subject is somewhat contrasted, with both positive and negative results, depending on the exact traits and on the taxonomic scale (Lartillot and Delsuc, 2012; Nabholz *et al.*, 2013; Figuet *et al.*, 2016; Brevet and Lartillot, 2021). At the micro-evolutionary scale, some studies have reported an empirical relation between π_N/π_S and π_S (James *et al.*, 2017; Brevet and Lartillot, 2021). Across time-scales, only the relation between micro-evolutionary process metrics and life-history traits in birds and across metazoan has been reported (Figuet *et al.*, 2016; Romiguier *et al.*, 2014). Equivalent results have never been reported in the case of mammals. This incomplete picture of the relation between the different micro and macro-evolutionary metrics across mammals calls for further investigation.

1.4 Comparative methods

The globally incomplete and undecisive picture offered by previous work is further affected by the use of disparate methods for estimating the molecular quantities and testing for their correlations. In this respect, one important aspect is to properly account for phylogenetic non-independance (Felsenstein, 1985), and this, in a context where some of the quantities of interest (in particular d_N/d_S) are not, strictly speaking, observed in extant species (as is the case for life history traits or π_S and π_N/π_S), but are only indirectly inferred along the branches of the phylogeny.

To deal with this issue, in the previous literature, two main avenues have been explored: a sequential approach, and an integrative one. The sequential approach consists in first estimating d_N/d_S on the terminal branches of the phylogeny and tabulating those estimates along with the other relevant variables. Then this table and a phylogeny are analysed using a standard methods based on independent contrasts (PGLS) (Pagel, 1997). The integrative approach, on the other hand, proposes to directly integrate the comparative test (the phylogenetic regression) into the phylogenetic codon model used for inferring the patterns of synonymous and non-synonymous substitutions over the tree. This method operates a fusion between the comparative method based on independent contrasts (Felsenstein, 1985), phylogenetic codons models and the relaxed molecular clock to simultaneously integrate molecular and ecological parameters in a unique Brownian motion along the tree (Lartillot and Poujol, 2011).

1.5 Aligning genome-wide data across time-scales

In addition to relying on a robust methodology as clarified above, a proper testing of the nearly neutral predictions bridging the two times-scales also requires adequate data at the genomic scale and globally across species of a large clade.

In a previous work, Brevet and Lartillot (2021) used an integrative approach to examine these questions in primates. They found a significant negative correlation between π_S and π_N/π_S but didn't observe any other correlation between life history traits (except among themselves), d_N/d_S, π_S or π_N/π_S. However, their dataset was limited in several aspects. Only 9 species out of the 61 studied present data about polymorphism. Moreover, these polymorphism data were computed from transcriptomic sources using different genes that did not match those used for the d_S and d_N/d_S reconstruction. At the scale of primates, Brevet and Lartillot (2021) relied on a restricted range of variation for the different metrics studied, which could explain the globally unconclusive results.

This is a general problem with this type of study. The currently available data from population genomics projects are still relatively heterogeneous along the living tree, making it difficult to gather data on intra-specific polymorphism for all species across a clade, that would be matched with the genes included in the multiple sequence alignment used for reconstruct the macro-evolutionary patterns over the phylogeny. Today, wild populational data just started to arrive but in the meantime, the increasing availability of whole genomes sequences in some taxa Zoonomia (2020); Feng *et al.* (2020), suggests an alternative approach not depending on population genomics projects. At least, for taxonomic groups such as mammals, we can rely on the fact that all published genomes are from diploid individuals. As such, they contain information about

the heterozygosity of those individuals, which itself is informative about the genetic diversity of the population (Li and Durbin, 2011). Although published genome sequences typically do not provide this information, it is nevertheless possible to obtain it by returning to the original sequencing reads (Zoonomia, 2020; Figuet *et al.*, 2016). Compared to a true population genomics project, this way to proceed provides more limited information, potentially affected by the fact that some individuals may happen to be inbred. On the other hand, it gives access to intra-specific diversity for all species for which complete genomes are available. Thus, it makes it possible to obtain a nearly complete data matrix, genome wide and over an entire clade, with which intra and inter-specific information can be globally confronted in the context of a comparative analysis.

In this article, we propose to expand Brevet and Lartillot (2021) analyses by combining whole-genome orthologous genes and heterozygosity data across a wide range of mammals. For this purpose, we annotated the complete one to one orthologous genes from 144 whole genomes. We then did variant calling on them to obtain the heterozygosity on the annotated genes for all species that are present in the phylogeny. This heterozygosity can then be confronted with the rich information currently available in mammals about life history traits (De Magalhaes *et al.*, 2009). We performed our correlation study using both an integrative and sequential approach, to do a global consistent test of the nearly-neutral theory.

2 Materials and Methods

Figure 1 summarises the pipeline developed and used here from the acquisition of the data to the final analyses. Below, we describe in detail the different steps of this pipeline.

2.1 Data acquisition

In our work, we used four types of input data (depicted in colour in Figure 1): complete genomes in fasta format, their associated reads, a BUSCO database and life-history traits. The reference genome assemblies were downloaded from NCBI Genomes in July 2021. At this time, we selected the assemblies of eutherian mammals using different criteria. First, we selected assemblies with median contig size (N50) of at least 30kb to avoid overly truncated genes. We then select assemblies with at least one individual with sequencing depth higher than 20x according to NCBI Genomes statistics. In addition, we excluded the assemblies from domestic species whose genetic variation is well-known to be reduced by the domestication process and the lab individuals that are usually highly inbred.

For each of the 197 assemblies that met all these requirements, we selected the latest reference assembly available in July 2021. Because the entire analysis relies on heterozygosity in single individuals, when multiple individuals were sequenced for a genome assembly project, we only selected and downloaded the reads from the individual with the highest sequencing depth. Sequencing reads were then mapped on their corresponding assembly with bwa_mem2 v2.2.1 used with default parameters (Vasimuddin *et al.*, 2019). Variant calling and the generation of whole-genome sequencing depth files were done with samtools bcftools v1.13, using default parameters (Danecek *et al.*, 2021) (https://samtools.github.io/bcftools/howtos/variant-calling.html).

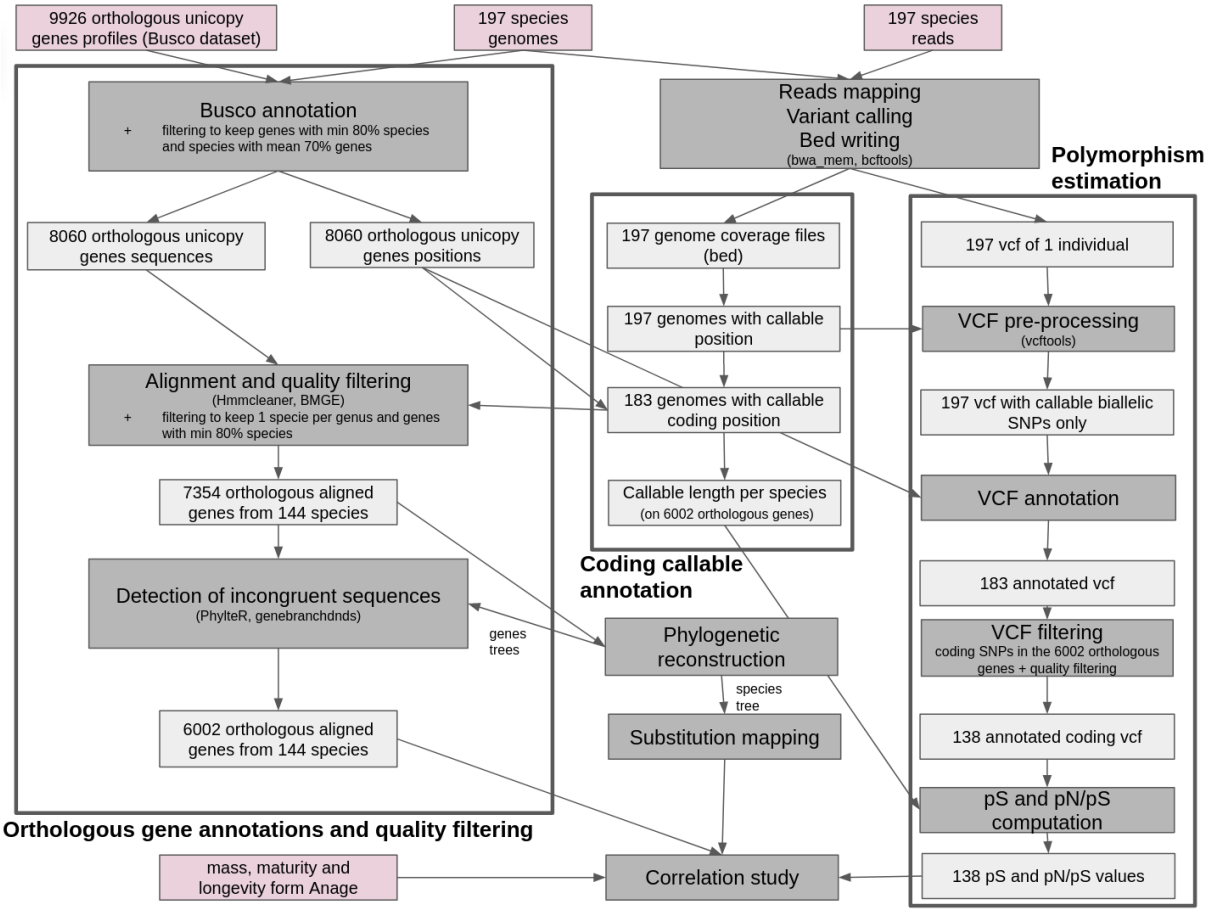


Figure 1: **Simplified summary of the pipeline from whole genomes acquisition and vcf writing to correlation study**. Red boxes correspond to the input data, light boxes correspond to the intermediate data and grey boxes correspond to different process in the pipeline.

2.2 Orthologous gene annotation and filtering

We developed a pipeline starting from the 197 assembled complete genomes, producing in output a data matrix of aligned and filtered single copy orthologous genes. At different steps of the process, quality controls and filters were applied (detailed below), leading to the removal of some species or genes from the analysis. Additionally, closely related species can share ancestral polymorphism. This shared polymorphism can affect the estimation of the d_N/d_S. To mitigate this issue, we chose to retain only one species per genus. The reduced dataset at the end of the pipeline contain 144 species and 6002 single-genes alignments.

2.2.1 Identification of orthologs

Not all the genomes are annotated, and those that are were annotated using different procedures. Since our focus is on orthologous coding sequences, and to ensure a homogeneous treatment across the entire species set, we (re-)annotated the genomes using BUSCO (Simão *et al.*, 2015). BUSCO is a commonly used bioinformatics tool for assessing genome assembly quality based on the identification of orthologous

genes assumed universal across the taxonomic group specified. Here we relied on the mammals database (orthodb.10) containing 9226 gene profiles supposed orthologous unicopy across all mammals. The program provides a percentage of recovered orthologous genes wich is meant as an indicator of the genome quality. BUSCO also provides intermediate output files specifying the coordinates of the orthologous genes identified in the genome being analysed. These intermediate files were processed so as to generate, for each genome, a fasta file containing all coding sequences and a bed file with their position along the genome. From our initial dataset of 197 species, we only kept species with BUSCO complete single-copy score higher than 70%, meaning that at least, 70% of the searched genes were found. In a second step, we removed genes present in less than 80% of the species to ensure their universality in the studied clade. After this step, we are left with 183 mammalian species and 8060 BUSCO one-to-one orthologous complete genes.

2.2.2 Alignment and quality filtering

The sequences of each orthologous gene were aligned using PRANK (Löytynoja, 2014) and then filtered using HMMcleaner (Di Franco *et al.*, 2019) and BMGE (Criscuolo and Gribaldo, 2010) with default parameters. These two filtering methods are complementary: BMGE filter alignments by columns, while HMMCleaner filters small segments of sequence on a per species basis (Ranwez and Chantret, 2020). Upon filtering, some alignments ended up empty or with a lot of gapps. We removed orthologous genes for which the alignment has less than 10 nucleotides or more than 75% gaps, resulting in 8056 aligned fasta files. Next, our dataset was reduced to only one species per genus leading to a set of 144 species. When multiple species occur for a genus, the one with the higher whole-genome coverage was chosen. We then removed the genes with less than 80% occurrence in this species set, resulting in a list of 7726 genes.

2.2.3 Detection of incongruent gene using gene tree

Although BUSCO aims at identifying single-copy orthologs, some alignments may nevertheless contain sequences with incorrect orthology assignments. Gene-level alignments for which the true evolutionary history differs from that of the species may introduce noise (Boussau and Scornavacca, 2020) and have a strong adverse input on the study and in particular on the species-level d_N/d_S estimates. Other sources of errors may also be present, such as sequencing errors resulting in compensated frameshifts, incorrect exon identification, etc.

To identify such deviant sequences and genes, we used a combination of two approaches. First, we reconstructed the 7726 gene trees using Iqtree2 with defaults parameters (Minh *et al.*, 2020). The trees where then given as an input to PhylteR (Comte *et al.*, 2023), which analyses all gene trees and provides a list of sequences and genes exhibiting outlier behaviour. Second, we developed a Bayesian shrinkage model (see Supplementary Methods) to identify sequences showing large deviations from the global pattern of synonymous branch lengths across the tree, which are expected to be very similar across genes.

PhylteR detected 20 genes that required complete removal and 1565 genes that contain at least one sequence to be removed. The Bayesian shrinkage model identified 1695 genes with at least one sequence displaying a deviant synonymous branch length. In total, 2015 genes out of the initial 7726 were flagged for at least one sequence, by either PhylteR or the Bayesian shrinkage model, with 1074 genes identified by both

methods. We removed all the flagged sequences from the alignments and also genes containing less than 80% of the species after this filtering step, resulting in a dataset set of 7354 gene alignments.

2.2.4 Callable fraction of the genome

Using genomes and their associated reads mapped onto them, we computed bam files containing the sequencing depth at each position of the genomes using bcftools (Danecek *et al.*, 2021). Given the large size of the resulting files, we converted the sequencing depth per position into a sequencing depth per window of 100 bp. We then computed for each species a mean genome sequencing depth and identified the 100 bp windows for which sequencing depth is 2 times higher or lower than this mean. These sites were considered as non-callable and were masked for the rest of the analysis. One species, *Neotragus schauinslandi*, for which more than 80% of the genome was masked, was removed from the analysis. Without *Neotragus schauinslandi*, the mean non-callable fraction of the genome is around 12% with a maximum of 40%.

Of note, some of the genomic regions considered as non-callable according to the criteria just described may overlap the coding sequence contained in the multiple sequence alignment across placentals (see above). To address this potential issue, we came back to the 7354 genes and replaced the non-callable coding sites by gaps directly in their Prank output. We then reran the filtering pipeline described above from Hmmcleaner to the exclusion of the too short or gapped sequences and of the genes represented in less than 80% of species. This resulted in a final high-quality dataset of 6002 genes with few missing data.

2.3 Alternative species subsampling schemes

To investigate the impact of closely related species in our correlation analysis, we subsampled the 144 species to 89 species as follows. Using the dated tree reconstructed by the integrative analysis on the 144 species (See results, tree in Figure 3), we identified all the subclades whose most recent common ancestor is younger than 10% of the total tree depth (corresponding to approximatively 10 My, assuming a date of 100My for the most recent common ancestor of placental mammals) and chose only one species per subclade, prioritizing species with higher sequencing depth and with all three life history traits informed. This leads to a reduced dataset of 89 species (listed in Supplementary Material). The 6002 single-genes alignments where restricted to these 89 species without realigning. At this step, we cannot remove genes present in less than 80% of species without removing too many genes.

2.4 Phylogenetic reconstruction

For the 89 and 144 species dataset, the species tree was estimated using Iqtree2 (Minh *et al.*, 2020) with the GTR+G4 model, on a concatenation of 1000 genes. The trees are in Supplementary Material (Fig.S4). For gene trees, required for Phylter (Comte *et al.*, 2023), we use Iqtree2 too with default parameters and model search options.

2.5 VCF annotation and polymorphism

Variant calling was performed using default parameters of samtools bcftools v1.13 (https://samtools.github.io/bcftools/howtos/variant-calling.html). We obtained VCF files containing all variant sites detected. We used vcftools (Danecek *et al.*, 2011) to keep only bi-allelic SNPs and remove the SNPs located in the non-callable regions such as previously defined.

SNPs were annotated as synonymous, non-synonymous or non-coding, using a home-made Python script that maps the SNPs and the coding orthologous annotation on the genomes. The script also labels each SNP by the name of the gene it belongs to. We then removed the non-coding SNPs and the ones not in the restricted list of 6002 orthologous genes. Furthermore, SNPs with GQ<150 and QUAL < 125 were removed (see Supplementary Material for the treshold definition) as well as SNPs with an allelic frequency higher than 0.8 or lower than 0.2. After this step, we decided not to further consider the VCF of six species because they present a global abnormal allelic frequency distribution (*Acomys cahirinus*, *Przewalskium albirostris*, *Mastomys coucha*, *Litocranius walleri* ,*Cheirogaleus medius* and *Cephalophus harveyi*) (see Supplementary Material, FigureS1).

After all these filters, six species ended up with less than 1000 SNPs (*Muscardinus avellanarius*, *Beatragus hunteri*, *Sigmodon hispidus*, *Diceros bicornis*, *Sousa chinensis* and *Alouatta palliata*). These species also have a low genome quality based on the N50 and L50 metrics. Below, we will show correlation analyses with and without these 6 species to evaluate their impact on the analysis.

Estimates of π_S and π_N were obtained by counting the total number of synonymous (S) and non-synonymous SNPs (N) and then normalizing by a rough estimate of the number of synonymous and non-synonymous mutational opportunities. Thus, denoting by L the total number of coding and callable position across all genes, K_S and K_N the number of heterozygote synonymous and non-synonymous positions, and assuming that point mutations in coding sequences are twice more likely to be non-synonymous than synonymous:

$$\pi_S = \frac{K_S}{L}$$
$$\pi_N = \frac{K_N}{2L}$$

We then compute π_N/π_S using our π_S and π_N measures.

The sampling variance of our π_S and π_N/π_S estimates was estimated by bootstrap at the gene level (i.e. drawing 6002 genes from the list with replacement and recomputing π_S and π_N/π_S). A 95% confidence interval was computed.

2.6 Life history traits

Data about adult body mass, longevity and age at female sexual maturity (called maturity for short in the following) across mammals were obtained from the Anage database (De Magalhaes *et al.*, 2009). To mitigate lack of data for some species, we systematically compute the mean of the trait per genus, even if the trait exists for the species. We obtain respectively for the 144 and 89 species datasets, 127 and 81 data points for mass, 118 and 73 for maturity and 118 and 74 for longevity.

2.7 Phylogenetic correlation analysis

Phylogenetic correlation analyses between life history traits and molecular variables such as π_N/π_S or d_N/d_S (here after, collectively referred to as "traits"), were conducted using two alternative methods, both of which account for phylogenetic inertia.

2.7.1 Bayesian integrative approach: FastCoevol

First, we developed a fast implementation of the Coevol model (Lartillot and Poujol, 2011). Coevol is an integrative Bayesian approach for analysing the joint evolutionary patterns over the phylogeny of observable traits (such as body mass) and substitution rates (here, dS, dN/dS). The underlying model recruits the conceptual basis of the independent contrasts approach (Felsenstein, 1985), by modeling the evolution of continuous characters (traits) via a multivariate Brownian diffusion process, while coupling it with a Brownian molecular clock for dS and dN/dS (rates). The coupling implied by the joint Brownian process for traits and rates enables gene alignments to be incorporated into the study, along with life-history traits and data about heterozygosity in extant species, eventually leading to an estimate of their joint correlation patterns, while automatically accounting for phylogenetic inertia. Fitting the model to the data produces as an output a dated tree, an estimate of the covariance matrix between traits, as well as a reconstruction of traits along the branches of the tree (see Figure 2).

Compared to the original Coevol model, the version used here, named FastCoevol entails an additional level of stochasticity, meant to account for short-term deviations in d_S and d_N/d_S from the long-term trends implied by the Brownian process. This two-level model is a direct generalization of the mixed relaxed clock model introduced in Lartillot *et al.* (2016) in the context of a codon model.

In its original version, Coevol is computationally intensive, precluding applications on empirical data sets with more than a few tens of genes. To overcome this limitation, here, a fast approximate version was developed, which is based on substitution mapping approximation (Romiguier *et al.*, 2012; Lemey *et al.*, 2012). Specifically, the detailed substitution history for all genes is inferred under a simpler reference model (assuming a constant gene-level d_N/d_S and shared synonymous branch lengths across genes). This substitution history is then collapsed into posterior mean numbers of synonymous and non-synonymous substitutions, and posterior mean numbers of synonymous and non-synonymous mutational opportunities, and this separately for each branch of the tree. Those numbers are then used as an approximate sufficient statistic for the molecular component of the multivariate Brownian model of Coevol. The use of this computation reduces the computational time from one week, using the original Coevol version, to 1-2 days for the substitution mapping and less than an hour for the fitting of the Brownian model on the data over the 6000 genes aligments.

For computational reasons, we perform the substitution mapping under the reference model on lists of 1000 genes. We then summed statistics across all gene sets, on a per-branch basis, and used this as an input for the FastCoevol model. With FastCoevol, we produced five independent MCMC runs per dataset to check the reproducibility of the study. We stopped their progression after stable convergence (around 30000 points) and used a large burn-in (around 10000 points). We computed the results using 1 point to 5 after the burn-in

in order to have independents points. The values of the five covariance matrices were averaged, even though they are very similar to each other. We considered a positive or negative correlation as significant when the posterior probability of the correlation is respectively higher than 0.975 or lower than 0.025, as in Lartillot and Poujol (2011).

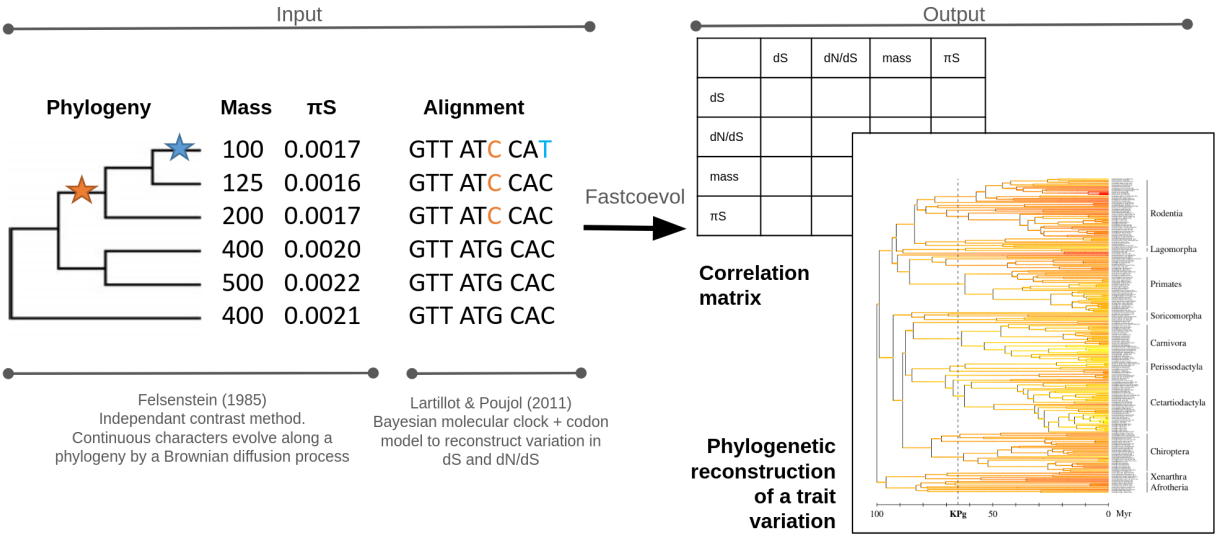


Figure 2: **FastCoevol pipeline**. Input : a rooted phylogeny, a table containing continuous traits of interest (here mass and π_S) and a multi-alignment of genes. Output: a correlation matrix with correlation coefficients and posterior probabilities for each couple of traits, including d_S and d_N/d_S, and a phylogenetic reconstruction figuring the evolution of each trait.

2.7.2 Sequential approach: PGLS

As an alternative to the integrative approach of FastCoevol, we used a sequential approach based on the standard PGLS method (Pagel, 1997). Specifically, the mapping statistics just described were summed over all genes and then used to compute an empirical d_N/d_S for each branch. Then, a linear regression analysis with correction for phylogenetic inertia was conducted using a phylogenetic covariance matrix derived from the evolutionary relationships among the species such as inferred by IQTree. For the PGLS analysis, we consider a correlation significant when the p-value is lower than 0.05.

3 Results

Starting from 197 complete mammalian genomes and after several steps of data processing and filtering, we gathered a dataset of 6002 single copy orthologous genes for 144 species covering all placental orders. Species and genes were chosen to ensure sufficient assembly quality, and maximise completeness of the data matrix (each gene represents at least 80% of species and conversely each species represents at least 70% of genes). Genes were aligned and trimmed, and a phylogeny was reconstructed using a sample of 1000 of these genes. This provides the macro evolutionary backbone of our analysis, which was complemented with

information about intra-specific polymorphism. Specifically, for each species represented in the dataset, the original reads that were used for genome assembly were mapped onto the genome to detect heterozygous sites. We focussed on the SNPs that were called in the coding sequences of the 6002 genes. In the end, this yields a set of synonymous and non-synonymous SNPs from which a π_S and π_N/π_S were inferred.

Starting from this reference dataset, we considered two additional species subsampling schemes. First, we trimmed the dataset to keep only species that are separated by at least 10 My of divergence, resulting in a dataset of 89 species. Second, we conducted the analysis without the data about polymorphism for the six species for which less than 1000 SNPs were called (*Muscardinus avellanarius*, *Beatragus hunteri*, *Sigmodon hispidus*, *Diceros bicornis*, *Sousa chinensis* and *Alouatta palliata*). This leads to four version of the datasets, which were then systematically analysed using both standard (PGLS) and integrative (FastCoevol) approaches.

3.1 Visualisation of d_N/d_S variation along the tree

The history of the changes in d_N/d_S over the phylogeny, such as inferring by FastCoevol, is shown in Figure 3. The d_N/d_S ranges from 0.15 to 0.34 over the tips. Taxonomic groups with the lowest d_N/d_S are Rodentia (i.e. *Peromyscus maniculatus*, d_N/d_S =0.148), Lagomorpha (i.e. *Ochotona princeps*, d_N/d_S = 0.151), and Caniforms in the Carnivora groups (i.e. *Vulpes vulpes*, $d_N/d_S = 0.157$). At the other end of the range, taxonomic groups with the highest d_N/d_S correspond to Cetacea (i.e. *Monodon mococeros* d_N/d_S = 0.296) and Primates (i.e. *Homo sapiens* $d_N/d_S = 0.258$). This general observations from Figure 3 is globally suggestive of a pattern of a high d_N/d_S in large and long lived mammals. Of note, some isolated lineages show more unexpected patterns, such as *Mustela putorius*, a Caniform which presents the highest d_N/d_S (0.341) while some other species within the same group exhibit the lowest d_N/d_S. At this stage, we have no explanation for this particular observation.

We now focus on the correlation analyses between the molecular and ecological traits at different evolutionary scales. First, we briefly analyse the correlation among life history traits and between d_S and life history traits. Although not the purpose of the present study, these correlations represent a useful sanity check. Then, we successively examine the macro and micro-evolutionary relations separately and finally confront the two time-scales.

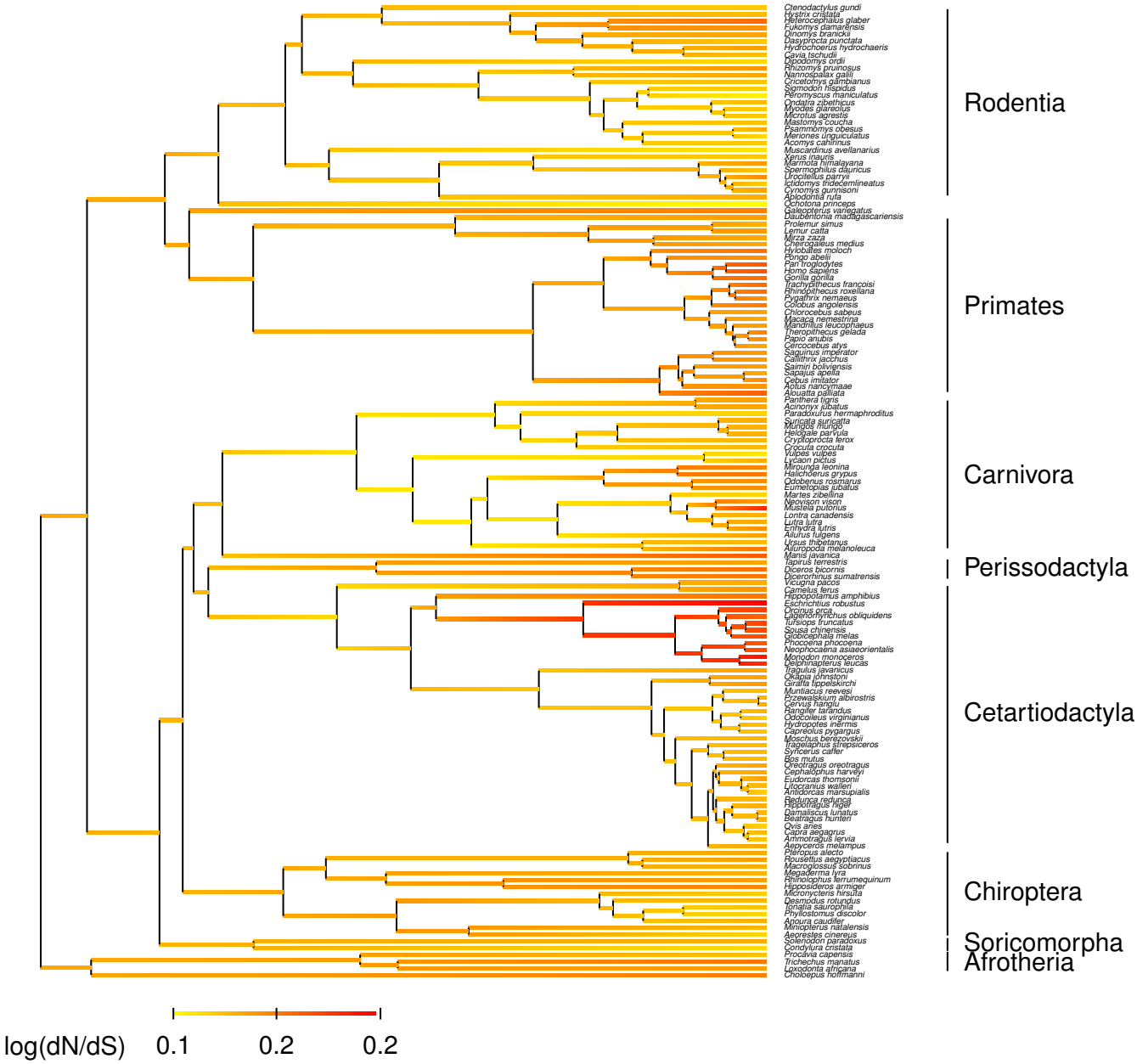


Figure 3: Reconstruction of $\log(d_N/d_S)$ along the 144 mammalian species tree by FastCoevol

3.2 Macro evolutionary scale : life history traits versus d_S and d_N/d_S

At the macro-scale, life history traits are known to be strongly correlated with each other. This well known result is recovered in our analysis using both PGLS and FastCoevol on the 144 and 89 species tree (Fig.4 A-C). Moreover, a highly significant negative correlation is observed between d_S and life history traits (Fig.4

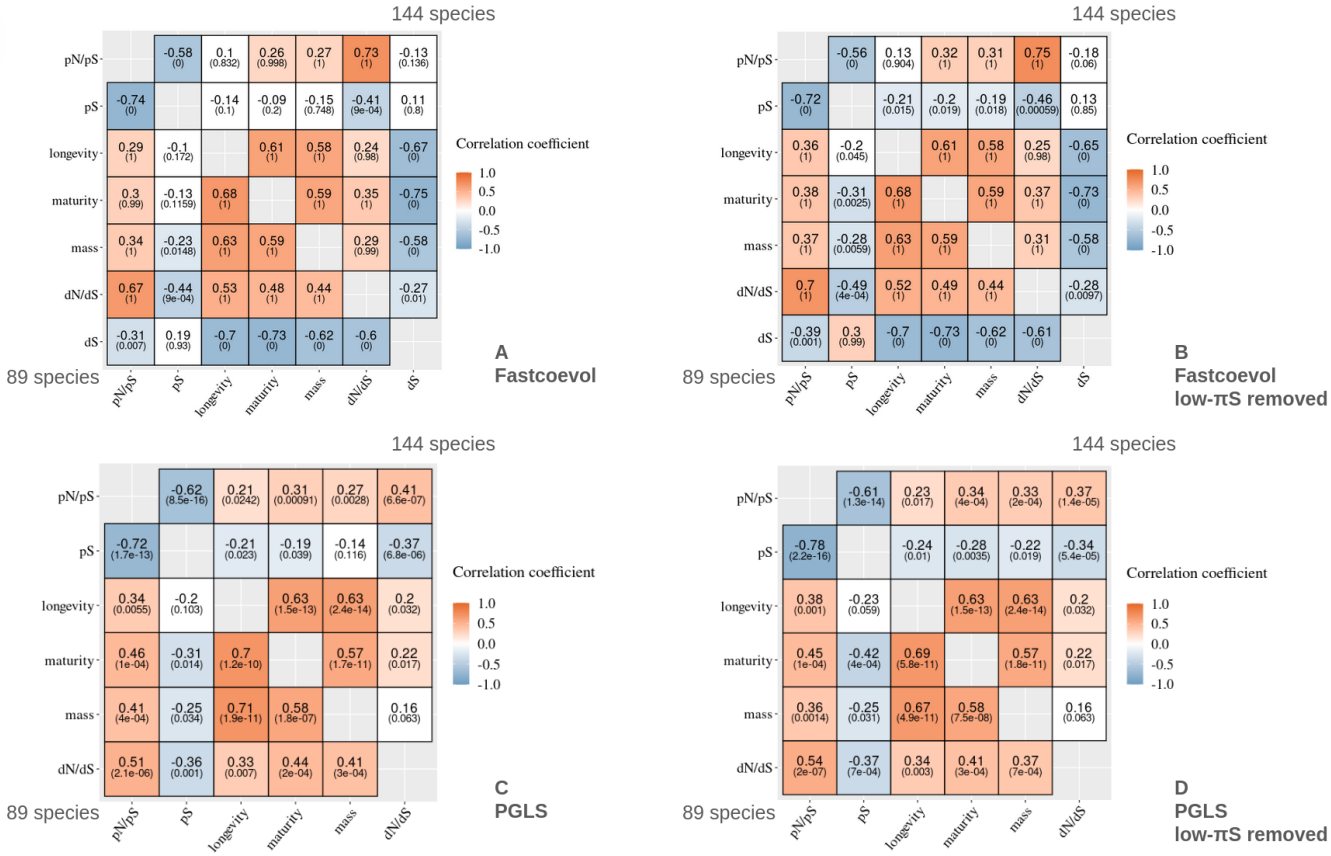


Figure 4: **Correlation coefficients from the FastCoevol and PGLS analyse**. **A**: FastCoevol analysis with the full dataset. **B**: FastCoevol analysis with the dataset excluding polymorphism data for the six low π_S species. **C**: PGLS analysis with the full dataset. **D**: PGLS analysis with the dataset excluding polymorphism data for the six low π_S species. **Upper diagonal**: the 144-species dataset. **Bottom diagonal**: the 89-species dataset excluding short branches. Colours correspond to the magnitude of the correlation coefficient (r). Numbers in brackets correspond to the posterior probability for FastCoevol analyses (which must be higher than 0.975 or lower than 0.025 for the correlation to be considered significant) or p-value for PGLS analyses (which must be lower than 0.05 for the correlation to be considered as significant). Non-coloured estimates correspond to non-significant correlations.

A). This negative relation between d_S and life history traits is congruent with what is already known and generally interpreted as a generation time effect (Thomas *et al.*, 2010). Altogether, these results support our dataset and methods.

Using the 144 species tree, the FastCoevol analysis shows a significant positive correlation between d_N/d_S and all life history traits, with correlation coefficients varying from 0.24 to 0.35 depending on the trait (Fig.4 A). Similarly, the PGLS analysis shows significant correlations ($\simeq 0.21$), with a non-significant correlation concerning d_N/d_S versus mass (Fig.4 C). The correlations look slightly stronger with FastCoevol, which can be explained by the fact that the additional variance contributed by the short-term overdispersion of d_N/d_S,

which is modelled by FastCoevol, has been discounted. Interestingly, using the 89 species tree, we observe stronger correlations under both methods. With the 89 species, correlations between d_N/d_S and mass also became significant in the PGLS analysis.

It may seem paradoxical that removing data points increases the strength of the inferred correlations. A reasonable explanation is that the independent contrasts between closely related species are dominated by estimations errors or short-term effects and that reducing the number of species reduces this source of noise This effect is also presents comparing the 89 versus 144 species correlations matrix from FastCoevol, which suggests that the d_N/d_S over-dispersion along branches is not totally captured by the method.

Altogether, both PGLS and FastCoevol analyses agree on a significant positive correlation between proxies of long-term N_e and selection efficacy at the macro-evolutionary scale. These results are generally interpreted as an indirect correlation of both variable with N_e (Ohta, 1973, 1992). These conclusions are compatible with a nearly-neutral interpretation, which states that selection is less efficient in long-lived, large-bodied mammals, presumably characterised by lower N_e over the tree.

3.3 Micro-evolutionary scale : π_S versus π_N/π_S

At the micro-scale, the efficacy of selection is reflected by π_N/π_S, while π_S provides a direct proxy of N_e. The nearly-neutral theory then predicts a negative correlation between these two measures. In the phylogenetic correlation analysis, we indeed observe a negative correlation using both methods (Fig.4 A-C). Again, we note stronger correlation coefficients with the 89 species analysis than in the one with 144 species (r=-0.74 versus r=-0.58 with FastCoevol, r=-0.72 versus r = -0.62 with PGLS).

However, plotting π_N/π_S against π_S (Fig.5) reveals some species showing a departure from the general pattern. Those species harbour a low π_S value given their π_N/π_S. This could be the sign of recent inbreeding, although these species don't seem to be outliers in their runs of homozygosity (see Supplementary Material). Low π_S could also reflect recent demographic bottleneck, which would more strongly impact π_S than π_N/π_S. More probably, however, incorrect normalisation of the synonymous and non-synonymous counts, and more specifically systematic overestimation of the number of callable positions when coverage is low, could lead to a downward bias in the π_S (and π_N) estimates, without impacting π_N/π_S. In Figure 5, we coloured the points using the species genome depth (see Material and Methods) and we marked with squares (instead of dots) the six species already identified for their smaller number of high-quality coding SNPs. We note that these flagged species correspond to the outliers in the graph and present a lower sequencing depth than the other species in the core of the general pattern. Altogether, this suggests more a quality issue for these outlying points rather than a low real π_S. Accordingly, in the following, we systematically analyse the correlation with (Fig.4 A-C) and without (Fig.4 B-D) polymorphism data for these six species with small SNP counts.

In the end, our analysis shows a robust correlation between selection efficacy and π_S as a proxy of N_e at the micro-evolutionary scale. The correlation at this scale appears stronger than at the macro scale. One possible reason for this is that π_S is a more direct proxy of N_e in comparison to life history traits.

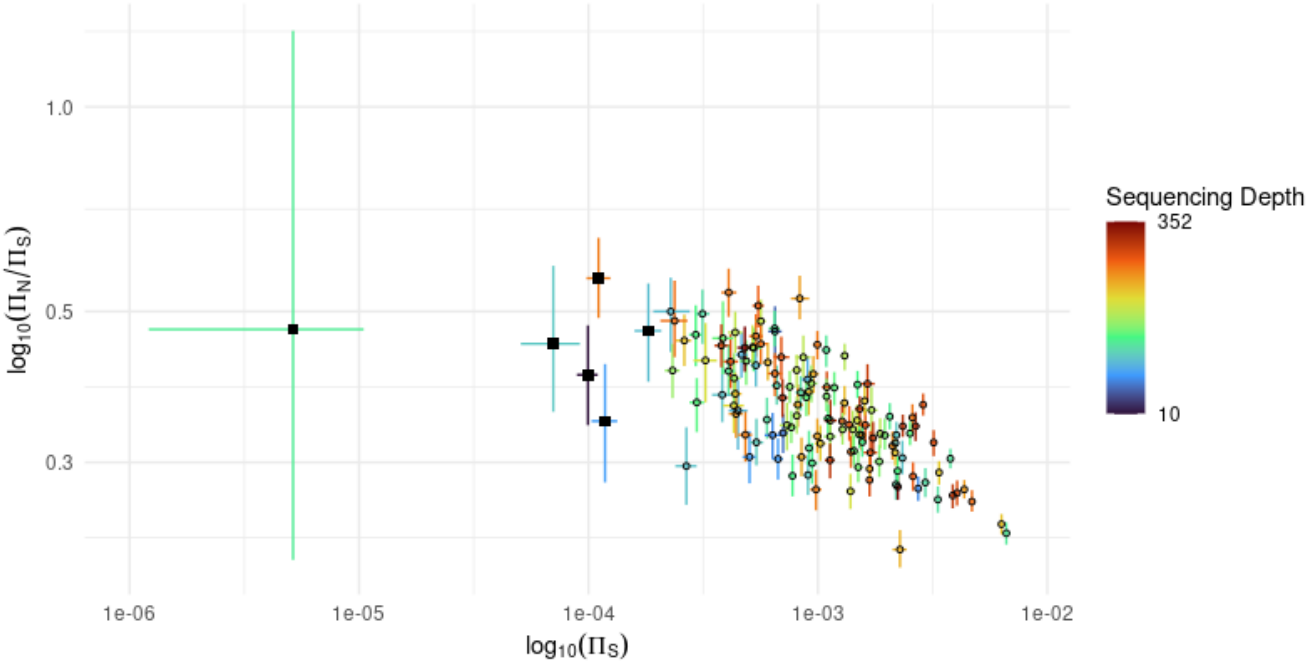


Figure 5: Distribution of the log10 transformed π_S and π_N/π_S values for each species (bars: 95% CI). Colours correspond to genome sequencing depth. Square points correspond to the six species with less than 1000 coding SNPs after filtering. The small dot points correspond to the other species

3.4 Comparing the micro- and macro-evolutionary scales

If macro-evolutionary patterns are the long term result of micro-evolutionary processes, we expect to observe correlations of our metrics of interest across timescales. However, this depends on the extent to which short-term N_e (such as measured by π_S) reflects long-term N_e (such as captured by life history traits).

In this direction, with FastCoevol, we found significant negative correlations between π_S and life history traits although only in the analysis without the six low-π_S species. Even then, the correlation coefficients are weak and below the significant threshold for mass in the 89 species analysis (Figure 4 A-B). The, PGLS analysis also shows weak correlations coefficients and the statistical support vary depending on the life history traits and the taxon sampling (Figure 4 C-D).

On the other hand, π_N/π_S shows a robust negative correlation with life history traits, under the four data settings and with the two methods. The most straightforward explanation for this correlation is a nearly-neutral effect. Here also, as above for d_N/d_S versus life history traits, the correlations are stronger and more consistent when the dataset is thinned to remove closely-related species. Again, this could be due to the negative impact of non-phylogenetic variation. In the present case, short-term fluctuations in N_e may represent an important contribution (see discussion).

Taken together, these correlation patterns of π_S and π_N/π_S with life history traits, which are congruent in their sign, but contrasted in their strength and significance, are consistent with the hypothesis of a connection between short- and long-term Ne.

Conversely, and finally, d_N/d_S correlates both negatively with π_S and positively with π_N/π_S. The correlation coefficient are globally stable across the different analyses for d_N/d_S versus π_S, but more variable for d_N/d_S versus π_N/π_S.

4 Conclusion and Discussion

The nearly neutral theory states that, the higher the effective population size is, the more the selection is able to purge mildly deleterious mutations (Ohta, 1973). In this study, we aim to confirm this hypothesis at two time scales : micro-evolutionary scale (using π_S and π_N/π_S) and macro-evolutionary scale (using life history traits and d_N/d_S), based on a large dataset of 6002 genes on 144 placental mammals.

The correlations observed in our study are recapitulated in Figure 6. Overall, all correlations are globally compatible with a nearly neutral interpretation. More specifically, all could be explained by all variables being correlated with a single hidden variable, which is N_e. The results are robust to alternative subsampling schemes, except for the correlation between π_S and life history traits which is perhaps the weakest and least convincing among all observed correlations.

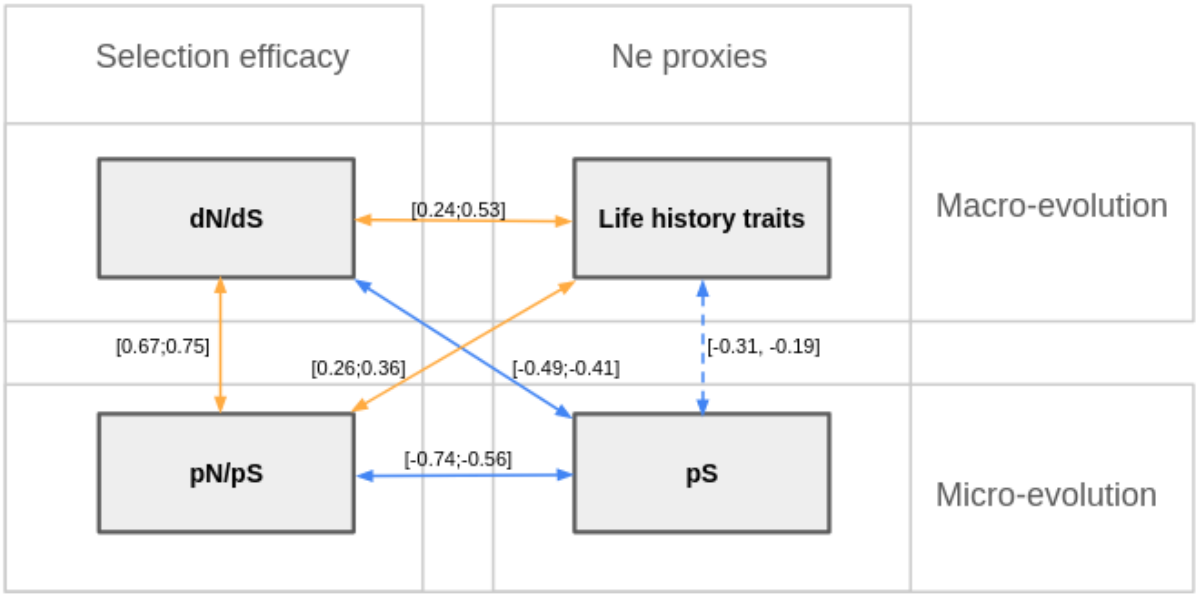


Figure 6: **Schematic view of the relation between the studied quantities.**In blue, a significant negative correlation, in orange, a significant positive correlation. Values in bracket correspond to the significant correlation coefficient range obtained from the four different analysis using FastCoevol

Several prior studies have already examined some of these correlations (between d_N/d_S, life history traits, π_S, and π_N/π_S). In Table 7, we present a sample from the published literature concerning both the significant

correlations and those sought but not found to be significant (labelled as "NS"). We also reported our results in this table. Regarding the macroevolutionary scale, the relation between d_N/d_S and life history traits has been extensively explored, and published results are not always consistent but are nevertheless globally in agreement with a nearly neutral effect on d_N/d_S (Figuet *et al.*, 2016; Nabholz *et al.*, 2013; Popadin *et al.*, 2007; Lartillot and Delsuc, 2012; Lartillot, 2012; Brevet and Lartillot, 2021). At the microevolutionary scale, some studies have identified a correlation between π_S and π_N/π_S in mammals (for mitochondial data in mammals (James *et al.*, 2017), or nuclear data in primates (Brevet and Lartillot, 2021)). Across scales, on the other hand, much less has been reported except for the relation observed between π_S or π_N/π_S with life history traits in birds (Figuet *et al.*, 2016) and metazoans (Romiguier *et al.*, 2014), a correlation not found in primates (Brevet and Lartillot, 2021). Notably, the correlation between π_S and life history traits has never been investigated thus far across mammals.

It is worth mentioning that only one study (Brevet and Lartillot, 2021) has examined all of these correlations within the same framework, although with insufficient data to draw decisive conclusion. The other are limited to either micro or macro scales or were focused on estimating N_e without estimating its correlation with selection efficacy.

		dN/dS	LHT			pS	pN/pS
			mass	maturity	longevity		
dN/dS			Figuet (2016) r=0,38* Nabholz (2013) r=0,47** Lartillot (2012) r=NS Lartillot Delsuc (2012) r=NS This study r=[0,29**;0,44**] 2/8 NS	r=0,71** r=0,38** r=0,27** r=0,36** r=[0,22*;0,49**]	r=0,69** r=0,51** r=0,27** r=0,27* r=[0,2*;0,53**]	This study r=[-0,49** ;-0,37**]	This study r=[0,37**;0,75**]
LHT	mass	Brevet (2021) primates r=NS Figuet (2016) birds and reptiles r=NS Nabholz (2013) birds r=NS				This study r=[-0,28** ;-0,19**] 2/8 NS	This study r=[0,27**;0,41**]
LHT	maturity	Brevet (2021) primates r=NS Figuet (2016) birds r=NS ; reptiles r=0,52* Nabholz (2013) birds r=NS				This study r=[-0,42** ;-0,19*] 2/8 NS	This study r=[0,26**;0,46**]
LHT	longevity	Brevet (2021) primates r=NS Nabholz (2013) birds r=0,63** Figuet (2016) birds r=NS ; reptiles r=0,68**				This study r=[-0,24* ;-0,21**] 5/8 NS	This study r=[0,21*;0,38**] 2/8 NS
pS		Brevet (2021) primates r=NS	Brevet (2021) primates r=NS Figuet (2016) birds r=-0,39*	r=-0,52**	r=-0,45*		James (2017) r=-0,96** This study r=[-0,72** ;-0,56**]
pN/pS		Brevet (2021) primates r=NS	Brevet (2021) primates r=NS Figuet (2016) birds r=0,46*	r=0,5**	r=0,45*	Brevet (2021) primates r=0,78**	

Figure 7: **Table summarising the state of the art concerning a sample of the different studies of correlation concerning d_N/d_S, life history traits, π_S and π_N/π_S.** Upper diagonal : Studies on placental mammals. Lower diagonal : Studies on other taxonomic groups. "*" mean a significant correlation (p-value < 0.05, pp> 0.95 or pp< 0.05), "**" mean a strongly significant correlation (p-value < 0.001, pp> 0.975 or pp< 0.025), "NS" mean a non-significant correlation. Results in blue correspond to our study. The values are under bracket as they represent the min and max of the correlation coefficient from the eight alternative sampling and method scheme. When a correlation is not significant in one of the eight modalities, we emphasize it by giving the number of non statistically significant observations over the eight.

4.1 Robustness to data errors

All the successive steps in our pipeline, from genome annotation to the alignment of gene sequences and the reconstruction of a phylogeny, are subject to different types of errors which could accumulate and could impact our conclusions (Simion *et al.*, 2020). Moreover, genomes of poor quality could be particularly problematic. Here, we have paid attention to genome assembly quality, as measured by BUSCO and a sufficiently large coverage. Nevertheless, metrics like N50 or L50 (minimal number of scaffold necessary to cover 50% of the genome), which are indicators of a fragmented assembly are also useful to detecting poor quality genomes, even if their Busco score or coverage looks correct. We note that species already flagged by a suspiciously low π_S or removed due to abnormal VCF (see material and methods) are species with a low N50 and a high L50.

Given the variable quality of the genome assemblies used, a lot of work has been done to control for data errors. We took care of the completeness of the dataset, refined the orthologous sequences and SNPs with different filters and masked conjointly the non-callable position on the divergence and polymorphism files. Some errors may still be present in some amount. It thus raises the question of whether they could drive or compromise some of the correlations observed here.

Two main sources of errors, susceptible to bias the correlation analyses, can be identified. First, alignment errors tends to inflate the apparent d_N/d_S. This effect is stronger for short branches (measured in synonymous evolutionary divergence) which could create a false correlation between d_N/d_S and d_S. If long-living species tend to have a low d_S, this could induce an artefactual correlation between d_N/d_S and life history traits. Second, and similarly, SNP calling errors will tend to inflate the apparent π_N/π_S. This effect will be stronger in species with a low π_S and can create an artefactual correlation between π_S and π_N/π_S. By extension, if long-living mammals tend to have a low π_S, this will induce an artefactual correlation between π_N/π_S and life history traits. It is still difficult, as it stands, to completly rule out the possibility that some of the correlation that we observe here are driven by these potential biases. Of note, concerning π_N/π_S, this would still imply a strong correlation between π_S and life-history traits.

Additionally, we cannot exclude the possibility of improper normalisation of π_S (and π_N), due to incorrect normalisation of the number of callable positions. Here, the number of callable positions isn't readjusted after the filtering of low quality SNPs. This is particularly an issue concerning lower quality genomes because they present less raw SNPs detected by the variant calling and lower SNPs quality overall, inducing a stronger reduction of the total number of final SNPs, without removing them from the callable set. Of note, this bias should affect π_S and π_N but should cancel in their ratio, and thus is not expected to impact π_N/π_S. Such normalisation problems could explain the mismatch between π_S and π_N/π_S in the low-π_S species of Figure 4, but also the fact that π_S has much weaker correlation than π_N/π_S with life history traits.

4.2 Impact of species subsampling

To examine the impact of closely related species in the correlation analysis, we conducted in parallel the correlation analysis using the full 144 species set and a reduced set of 89 species separate by at least by 10Mya. We observed that the reduced dataset gives globally stronger correlation. As explained above, this

can be due to the non-phylogenetic signal present in the variables being analysed. Several factors could contribute to this non-phylogenetic signal, like measurement errors (already discussed above), intra-specific variation and short-term fluctuation, in particular in N_e.

Concerning intra-specific variation, the use of the heterozygosity of a single individual as a proxy for population-level diversity can be problematic if the individual used does not reflect its population diversity (due to population structure, inbreeding...). In any case, using a single individual will always contribute some additional variance, that remains to be quantified. Of note, here, since π_S and π_N/π_S are calculated on the same individual, their correlates should not be subject to this additional variance.

The increased correlation coefficients between π_S and life history traits or d_N/d_S, when we shift from the 144-species to the 89-species tree, can be an indication for a possible short-term fluctuation in N_e and thus a mismatch between short-term and long-term N_e. The tree with longer branches will be less impacted by this short term fluctuation. However, this short-term fluctuation in N_e should not impact so strongly π_S compared to π_N/π_S, except if π_S is more sensitive or more responsive than π_N/π_S to short-term demographic effects, which remains to be theoretically investigated. Nevertheless, it could explain the increased correlation between the micro-evolutionary measures and life history traits. This mismatch between short-term and long-term N_e is certainly an interesting aspect to further investigate and quantify.

In the case of d_N/d_S, in addition to short-term fluctuations (which are detected and are inferred to be substantial by the over-dispersed model of FastCoevol), another important factor could be at play, namely segregating polymorphism. Segregating polymorphism imply that the apparent d_N/d_S contains a certain proportion of π_N/π_S in its estimation. This proportion is higher in the terminal branches of the 144-species tree because of their shorter length. Thus, supposing that the real d_N/d_S correlates more strongly with life-history traits (both being long-term quantities) than does π_N/π_S, then the π_N/π_S part present in the observed d_N/d_S, would reduce the correlation coefficient between d_N/d_S and life history traits for the 144-species tree. This hypothesis is supported by a stronger correlation between π_N/π_S and d_N/d_S using the 144 species tree.

4.3 Perspectives on micro versus macro-evolution

In spite of these potential weaknesses, in the end, our analysis confirms and consolidates the correlations among long-term quantities (d_N/d_S, life history traits), and among short-term quantities (π_N/π_S, π_S). More importantly, it provides a more decisive entry point on micro-macro connections. Nevertheless, our results still remain to be confirmed in particularly by implementing a more robust normalisation of π_S and by quantifying the contribution of the polymorphism to the d_N/d_S estimates.

About the methods, in the end, the use of an integrative (FastCoevol) rather than sequential (PGLS) method doesn't seem to make a big difference. They both gives similar results. However, integrative methods can be furthered developed to reconstruct N_e. In this article, we make the assumption that π_S represents N_e, using the relation $\pi_S = 4N_e\mu$. However, μ itself varies between species, and is known to correlate with life history traits. Mutations rate became a crucial factor which is important to take into account in our N_e reconstruction. A classical approach to correct for μ is to estimate it by using the synonymous

divergence between close species and fossil calibration with a generation time. In the integrative method, we can incorporate a decomposition of π_S in is N_e and μ components by the reconstruction of μ along the topology. This was already implemented in Brevet and Lartillot (2021) and could be recruited and applied to the dataset obtained here for mammals. However, this would first require explicitly modelling the mismatch between short-term and long-term N_e in the context of this integrative approach. Such improvement in the model could provide a more quantitative idea about the intensity of the short term N_e fluctuations.

Finally, here we focused on nearly neutral evolution, but our dataset could be used to investigate other aspects, as it presents the opportunity to confront two genomic compartiments which respond differently to certain evolutionary processes such as positive selection (McDonald and Kreitman, 1991) or mutation bias and genetic biased conversion toward GC (Duret and Galtier, 2009).

Acknowledgements

Funding: Agence Nationale de la Recherche (ANR-20-CE02-0008-01 "NeGA")
Author contributions: Original idea: N.L M.B; Model conception: N.L; Code: M.B, N.L; Data analyses: M.B, D.E ; Interpretation: M.B, N.L; First draft: M.B; Editing and revisions: M.B, N.L, D.E; Project management and funding: NL.
Competing interests: The authors declare no conflicts of interest.
Data and materials availability: .

References

Al Abri, M. A., Holl, H. M., Kalla, S. E., Sutter, N. B., and Brooks, S. A. 2020. Whole genome detection of sequence and structural polymorphism in six diverse horses. *PLoS One*, 15(4): e0230899.

Begun, D. J., Holloway, A. K., Stevens, K., Hillier, L. W., Poh, Y.-P., Hahn, M. W., Nista, P. M., Jones, C. D., Kern, A. D., Dewey, C. N., *et al.* 2007. Population genomics: whole-genome analysis of polymorphism and divergence in drosophila simulans. *PLoS biology*, 5(11): e310.

Boitard, S., Rodríguez, W., Jay, F., Mona, S., and Austerlitz, F. 2016. Inferring population size history from large samples of genome-wide molecular data-an approximate bayesian computation approach. *PLoS genetics*, 12(3): e1005877.

Bolívar, P., Mugal, C. F., Rossi, M., Nater, A., Wang, M., Dutoit, L., and Ellegren, H. 2018. Biased inference of selection due to gc-biased gene conversion and the rate of protein evolution in flycatchers when accounting for it. *Molecular biology and evolution*, 35(10): 2475–2486.

Boussau, B. and Scornavacca, C. 2020. Reconciling gene trees with species trees. *Phylogenetics in the genomic era*, pages 3–2.

Brevet, M. and Lartillot, N. 2021. Reconstructing the history of variation in effective population size along phylogenies. *Genome Biology and Evolution*, 13(8): evab150.

Burri, R., Nater, A., Kawakami, T., Mugal, C. F., Olason, P. I., Smeds, L., Suh, A., Dutoit, L., Bureš, S., Garamszegi, L. Z., *et al.* 2015. Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of ficedula flycatchers. *Genome research*, 25(11): 1656–1665.

Charlesworth, B. 2009. Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics*, 10(3): 195–205.

Comte, A., Tricou, T., Tannier, E., Joseph, J., Siberchicot, A., Penel, S., Allio, R., Delsuc, F., Dray, S., and de Vienne, D. M. 2023. Phylter: efficient identification of outlier sequences in phylogenomic datasets. *bioRxiv*, pages 2023–02.

Criscuolo, A. and Gribaldo, S. 2010. Bmge (block mapping and gathering with entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC evolutionary biology*, 10: 1–21.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., *et al.* 2011. The variant call format and vcftools. *Bioinformatics*, 27(15): 2156–2158.

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., *et al.* 2021. Twelve years of samtools and bcftools. *Gigascience*, 10(2): giab008.

De Magalhaes, J., Costa, and J 2009. A database of vertebrate longevity records and their relation to other life-history traits. *Journal of evolutionary biology*, 22(8): 1770–1774.

De Maio, N., Schrempf, D., and Kosiol, C. 2015. Pomo: an allele frequency-based approach for species tree estimation. *Systematic biology*, 64(6): 1018–1031.

Di Franco, A., Poujol, R., Baurain, D., and Philippe, H. 2019. Evaluating the usefulness of alignment filtering methods to reduce the impact of errors on evolutionary inferences. *BMC Evolutionary Biology*, 19: 1–17.

Duret, L. and Galtier, N. 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annual review of genomics and human genetics*, 10: 285–311.

Eyre-Walker, A. and Keightley, P. D. 2007. The distribution of fitness effects of new mutations. *Nature Reviews Genetics*, 8(8): 610–618.

Eyre-Walker, A., Woolfit, M., and Phelps, T. 2006. The distribution of fitness effects of new deleterious amino acid mutations in humans. *Genetics*, 173(2): 891–900.

Felsenstein, J. 1985. Phylogenies and the comparative method. *The American Naturalist*, 125(1): 1–15.

Feng, S., Stiller, J., Deng, Y., Armstrong, J., Fang, Q., Reeve, A. H., Xie, D., Chen, G., Guo, C., Faircloth, B. C., *et al.* 2020. Dense sampling of bird diversity increases power of comparative genomics. *Nature*, 587(7833): 252–257.

Figuet, E., Nabholz, B., Bonneau, M., Mas Carrio, E., Nadachowska-Brzyska, K., Ellegren, H., and Galtier, N. 2016. Life history traits, protein evolution, and the nearly neutral theory in amniotes. *Molecular biology and evolution*, 33(6): 1517–1527.

Galtier, N. and Duret, L. 2007. Adaptation or biased gene conversion? extending the null hypothesis of molecular evolution. *TRENDS in Genetics*, 23(6): 273–277.

Gayral, P., Melo-Ferreira, J., Glemin, S., Bierne, N., Carneiro, M., Nabholz, B., Lourenco, J. M., Alves, P. C., Ballenghien, M., Faivre, N., *et al.* 2013. Reference-free population genomics from next-generation transcriptome data and the vertebrate–invertebrate gap. *PLoS genetics*, 9(4): e1003457.

James, J., Castellano, D., and Eyre-Walker, A. 2017. Dna sequence diversity and the efficiency of natural selection in animal mitochondrial dna. *Heredity*, 118(1): 88–95.

Kimura, M. 1979. The neutral theory of molecular evolution. *Scientific American*, 241(5): 98–129.

King, J. L. and Jukes, T. H. 1969. Non-darwinian evolution: Most evolutionary change in proteins may be due to neutral mutations and genetic drift. *Science*, 164(3881): 788–798.

Lack, J. B., Cardeno, C. M., Crepeau, M. W., Taylor, W., Corbett-Detig, R. B., Stevens, K. A., Langley, C. H., and Pool, J. E. 2015. The drosophila genome nexus: a population genomic resource of 623 drosophila melanogaster genomes, including 197 from a single ancestral range population. *Genetics*, 199(4): 1229–1241.

Lartillot, N. 2012. Interaction between selection and biased gene conversion in mammalian protein-coding sequence evolution revealed by a phylogenetic covariance analysis. *Molecular biology and evolution*, 30(2): 356–368.

Lartillot, N. and Delsuc, F. 2012. Joint reconstruction of divergence times and life-history evolution in placental mammals using a phylogenetic covariance model. *Evolution*, 66(6): 1773–1787.

Lartillot, N. and Philippe, H. 2004. A bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular biology and evolution*, 21(6): 1095–1109.

Lartillot, N. and Poujol, R. 2011. A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters. *Molecular biology and evolution*, 28(1): 729–744.

Lartillot, N., Phillips, M. J., and Ronquist, F. 2016. A mixed relaxed clock model. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 371(1699).

Latrille, T., Rodrigue, N., and Lartillot, N. 2023. Genes and sites under adaptation at the phylogenetic scale also exhibit adaptation at the population-genetic scale. *Proceedings of the National Academy of Sciences*, 120(11): e2214977120.

Lemey, P., Minin, V. N., Bielejec, F., Kosakovsky Pond, S. L., and Suchard, M. A. 2012. A counting renaissance: Combining stochastic mapping and empirical Bayes to quickly detect amino acid sites under positive selection. *Bioinformatics*.

Leroy, T. and Nabholz, B. 2022. Response to kratochvíl and rovatsos. *Current Biology*, 32(1): R30–R31.

Li, H. and Durbin, R. 2011. Inference of human population history from individual whole-genome sequences. *Nature*, 475(7357): 493–496.

Löytynoja, A. 2014. Phylogeny-aware alignment with prank. *Multiple sequence alignment methods*, pages 155–170.

Lynch, M. and Walsh, B. 2007. *The origins of genome architecture*, volume 98. Sinauer associates Sunderland, MA.

Lynch, M., Bobay, L.-M., Catania, F., Gout, J.-F., and Rho, M. 2011. The repatterning of eukaryotic genomes by random genetic drift. *Annual review of genomics and human genetics*, 12: 347–366.

Lynch, M., Ali, F., Lin, T., Wang, Y., Ni, J., and Long, H. 2023. The divergence of mutation rates and spectra across the tree of life. *EMBO reports*, 24(10): e57561.

Mathew, L. A. and Jensen, J. D. 2015. Evaluating the ability of the pairwise joint site frequency spectrum to co-estimate selection and demography. *Frontiers in genetics*, 6: 268.

McDonald, J. H. and Kreitman, M. 1991. Adaptive protein evolution at the adh locus in drosophila. *Nature*, 351(6328): 652–654.

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Von Haeseler, A., and Lanfear, R. 2020. Iq-tree 2: new models and efficient methods for phylogenetic inference in the genomic era. *Molecular biology and evolution*, 37(5): 1530–1534.

Müller, R., Kaj, I., and Mugal, C. F. 2022. A nearly neutral model of molecular signatures of natural selection after change in population size. *Genome Biology and Evolution*, 14(5): evac058.

Nabholz, B., Uwimana, N., and Lartillot, N. 2013. Reconstructing the phylogenetic history of long-term effective population size and life-history traits using patterns of amino acid replacement in mitochondrial genomes of mammals and birds. *Genome biology and evolution*, 5(7): 1273–1290.

Nadachowska-Brzyska, K., Konczal, M., and Babik, W. 2022. Navigating the temporal continuum of effective population size. *Methods in Ecology and Evolution*, 13(1): 22–41.

Ohta, T. 1973. Slightly deleterious mutant substitutions in evolution. *Nature*, 246(5428): 96–98.

Ohta, T. 1992. The nearly neutral theory of molecular evolution. *Annual review of ecology and systematics*, 23(1): 263–286.

Ohta, T. 1995. Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. *Journal of molecular evolution*, 40: 56–63.

Pagel, M. 1997. Inferring evolutionary processes from phylogenies. *Zoologica Scripta*, 26(4): 331–348.

Popadin, K., Polishchuk, L. V., Mamirova, L., Knorre, D., and Gunbin, K. 2007. Accumulation of slightly deleterious mutations in mitochondrial protein-coding genes of large versus small mammals. *Proceedings of the National Academy of Sciences*, 104(33): 13390–13395.

Ranwez, V. and Chantret, N. N. 2020. Strengths and limits of multiple sequence alignment and filtering methods.

Rolland, J., Henao-Diaz, L. F., Doebeli, M., Germain, R., Harmon, L. J., Knowles, L. L., Liow, L. H., Mank, J. E., Machac, A., Otto, S. P., *et al.* 2023. Conceptual and empirical bridges between micro-and macroevolution. *Nature Ecology & Evolution*, 7(8): 1181–1193.

Romiguier, J., Figuet, E., Galtier, N., Douzery, E. J. P., Boussau, B., Dutheil, J. Y., and Ranwez, V. 2012. Fast and robust characterization of time-heterogeneous sequence evolutionary processes using substitution mapping. *PLoS One*, 7(3): e33852.

Romiguier, J., Gayral, P., Ballenghien, M., Bernard, A., Cahais, V., Chenuil, A., Chiari, Y., Dernat, R., Duret, L., Faivre, N., *et al.* 2014. Comparative population genomics in animals uncovers the determinants of genetic diversity. *Nature*, 515(7526): 261–263.

Russo, C. A., Eyre-Walker, A., Katz, L. A., and Gaut, B. S. 2024. Forty years of inferential methods in the journals of the society for molecular biology and evolution. *Molecular Biology and Evolution*, 41(1): msad264.

Sherry, S. T., Ward, M.-H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., and Sirotkin, K. 2001. dbsnp: the ncbi database of genetic variation. *Nucleic acids research*, 29(1): 308–311.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. 2015. Busco: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19): 3210–3212.

Simion, P., Delsuc, F., and Philippe, H. 2020. To what extent current limits of phylogenomics can be overcome?

Siva, N. 2008. 1000 genomes project. *Nature biotechnology*, 26(3): 256–257.

Thomas, J. A., Welch, J. J., Lanfear, R., and Bromham, L. 2010. A generation time effect on the rate of molecular evolution in invertebrates. *Molecular biology and evolution*, 27(5): 1173–1180.

Vasimuddin, M., Misra, S., Li, H., and Aluru, S. 2019. Efficient architecture-aware acceleration of bwa-mem for multicore systems. In *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, pages 314–324.

Waples, R. S. 2022. What is n e, anyway? *Journal of Heredity*, 113(4): 371–379.

Waples, R. S., Luikart, G., Faulkner, J. R., and Tallmon, D. A. 2013. Simple life-history traits explain key effective population size ratios across diverse taxa. *Proceedings of the Royal Society B: Biological Sciences*, 280(1768): 20131339.

731 Welch, J. J., Eyre-Walker, A., and Waxman, D. 2008. Divergence and polymorphism under the nearly neutral
732 theory of molecular evolution. *Journal of molecular evolution*, 67: 418–426.

733 Wolfe, K. H. and Li, W.-H. 2003. Molecular evolution meets the genomics revolution. *Nature genetics*, 33(3):
734 255–265.

735 Wright, S., Teissier, G., *et al.* 1939. Statistical genetics in relation to evolution. *(No Title)*.

736 Zoonomia 2020. A comparative genomics multitool for scientific discovery and conservation. *Nature*,
737 587(7833): 240–245.

6

Un génome individuel comme mesure de la diversité génétique dans des populations de vertébrés

Contexte :

Régulièrement, quand je présente les résultats de mon principal travail de thèse (chapitre 5), il me faut préciser une hypothèse forte qui est de considérer que le génome d'un individu diploïde représente la diversité génétique de sa population. Cette hypothèse a toujours soulevé des interrogations sur son impact concernant les résultats que je présente. Pourtant, en population idéale, le génome d'un individu diploïde recombinant est censé correspondre a un echantillonnage des sequences genetiques de ses ancêtres, eux-mêmes formant un échantillon représentatif de la population ancestrale (sous-section 3.3.1). Cependant s'il y a bien une chose que j'ai apprise lors de ma thèse, c'est que les données empiriques se fient bien des attendus théoriques et que les modèles de population idéale sont souvent bien loin de représenter la réalité. Il m'est ainsi apparu de plus en plus urgent de questionner concrètement l'impact de l'usage du génome d'un individu diploïde pour représenter la diversité de sa population et notamment la variance que cela ajoute aux mesures de polymorphisme.

Il se trouve que Thibault Latrille travaillait au même moment sur la constitution d'un jeu de données contenant les génomes de plusieurs espèces de mammifères avec pour chaque espèce une à plusieurs populations elles-mêmes pourvues du génome complet de plusieurs individus. Nous avons travaillé ensemble pour augmenter ce jeu de données avec des données issues de deux espèces de vertébrés non-mammifères (la

mésange et le saumon) puis nous l'avons utilisé pour étudier l'impact des mesures individuelles de diversité génétique sur les mesures de polymorphisme.

Initialement, il n'avait été question de ne regarder que la distribution des hétérozygoties par individus et par population afin de mesurer un écart individuel à la moyenne populationnelle. Cela nous a permis de questionner l'importance de cet écart et sa variabilité entre espèces et populations. Nous avons notamment pu constater que l'usage d'un unique génome diploïde pour mesurer la diversité génétique d'une espèce ajoute environ 30% de variance à cette mesure. Par la suite, et au vu des résultats, nous avons envisagé que certains écarts individuels importants par rapport l'hétérozygotie moyenne d'une population, pourraient être due à des individus consanguins. Pour approfondir cette idée, nous avons analysé les génomes du jeu de données par une méthode pSMC (Li and Durbin, 2011), qui est capable, elle aussi, de fournir une mesure de diversité génétique à partir d'un seul génome. Plus encore, la méthode pSMC propose une reconstruction des variations du taux de coalescent, et donc du N_e instantané, au cours du temps. Ce N_e peut ensuite être divisé par le taux de mutation afin d'obtenir une mesure de diversité génétique. Pour comparer les mesures de pSMC avec l'hétérozygotie précédemment estimée, nous avons dérivé deux métriques individuelles : l'une correspondant à la moyenne de la diversité génétique sur l'ensemble de la période couverte par la méthode, et l'autre en se limitant au tiers le plus ancien de la courbe. Cette dernière mesure est attendue comme moins impactée par la consanguinité, car nous faisons l'hypothèse que les régions consanguines dans les génomes sont interprétées, par la méthode pSMC, comme des événements récents de perte de diversité.

L'usage d'une méthode pSMC a nécessité la mise en place d'un protocole de traitement des données plutôt complexe, en Snakemake, qui intègre un grand nombre de fichiers intermédiaires de types différents et dont le volume total génère beaucoup d'inertie. Cependant, nous avons persévéré dans ce projet avec l'attente que l'estimation d'une mesure de diversité génétique n'utilisant qu'un seul génome via une méthode type pSMC, fournirait un signal moins bruité qu'une simple mesure d'hétérozygotie.

Il se trouve que nous n'avons pas obtenu de résultats permettant de dire que la

méthode pSMC fournis des mesures de diversité génétique moins variables à l'échelle individuelle que l'hétérozygotie. L'utilisation d'outil pSMC pour ce genre de question, au vu de nos résultats, ne semble donc finalement pas pertinente. De plus, nous avons réalisé une analyse de variance permettant de quantifier la part de variation dans ces mesures individuelle de diversité qui est portée au niveau de l'espèce, de la population et de l'individu. Nous observons que la majorité de la variance est définie au niveau de l'espèce, comme attendus, puis au niveau de l'individu. Cela implique qu'on ne peut pas distinguer deux populations issues d'une même espèce en utilisant ces méthodes. Autrement dit, les mesure de diversité individuelle reflètent la diversité génétique de l'espèce (en ajoutant environ 30% de variance), mais pas de celle de la population d'où provient l'individu, ce qui est cependant suffisant pour justifier l'hypothèse faite dans le premier article.

Idéalement, nous aurions aimé confronter ces deux mesures de diversité utilisant un unique génome diploïde, aux mesures plus classiques de diversité tel que les π et θ_W définie en section 3.3. Cependant, dans un contexte plus réaliste, et non idéale, où les tailles de populations varient au cours du temps, chacun de ces quatre estimateurs réagit différemment et donc approxime différemment la vraie diversité génétique de la population (pour peu qu'il y en ait une). En effet, dans ce contexte, chaque mesure de diversité correspond en réalité à une moyenne des variations de la diversité génétique dans le temps, avec des pondérations propres à chaque mesure. Par exemple, π et l'hétérozygotie alloue plus de poids aux mesures de diversité dans le temps ancien, tandis que θ_W donne plus de poids aux temps récents. Il devient de fait assez vite difficile de déterminer des attendus concernant les relations entre ces différents estimateurs populationnels et individuels et donc de les comparer. Nous argumentons que chacun de ces estimateurs représentent une réalité qui peut correspondre à une mesure de diversité génétique, en fonction de la définition qu'on l'on appose à celle-ci. Nous ne pouvons donc pas vérifier directement si les mesures à génome unique représentent la diversité d'une population puisque cette diversité est elle-même inconnue.

Contributions :

Cet article est co-signé avec **Thibault Latrille** et **Nicolas Lartillot**. Thibault a principalement eu un rôle dans l'élaboration préliminaire du projet, dans la structuration des données et dans la mise en place du pipeline d'analyse. Il m'a d'ailleurs transmis un savoir précieux en matière de code Python et Snakemake[1]. Nicolas a davantage contribué à l'analyse des résultats et a particulièrement participé à l'analyse de variance présentée dans l'article. Pour ma part, j'ai analysé les résultats et rédigé la première version de l'article, que nous avons ensuite retravaillée avec Nicolas et Thibault.

1. Merci Thibault !

Individual genome as a proxy for nucleotide diversity in vertebrate

M. Bastian[1], **T. Latrille**[2], **N.Lartillot**[1]

[1]Laboratoire de Biométrie et Biologie Evolutive, UMR5558, Université Lyon 1, Villeurbanne, France

[2] Department of Computational Biology, Université de Lausanne, Lausanne, Switzerland

corresponding author: melodie.bastian@univ-lyon1.fr

October 18, 2024

Abstract

Nucleotide diversity is a key quantity in population genetic studies, which is widely used in fields such as conservation biology to assess genetic health and biodiversity or evolutionary biology to study adaptation and speciation. However, this measure ideally requires a large number of sequenced individuals randomly sampled from the population. This type of data is not often available, genome-wide, except for targeted population genomic projects. Single-genome metrics, such as genome-wide heterozygosity, can be used to overcome this issue as a proxy for population-level diversity. Also using the genome of a single diploid individual, tools such as pSMC can reconstruct population histories, offering additional metrics to estimate nucleotide diversity over time. Nevertheless, to what extent the genome of a single individual represents the diversity of its population remain unquantified. In this study, we evaluate three single-genome metrics meant as proxies of population-level nucleotide diversity, using 2,517 genomes from 56 populations across seven species. These include whole-genome heterozygosity, as well as two pSMC-based measures of nucleotide diversity. We decompose the variance of each metric into three components, corresponding to the individual, population and species levels, and compare them to identify the most stable. Our analysis quantifies the deviations of diversity measures based on individual genomes from species-level diversity, providing insights into the accuracy of single-genome metrics in evolutionary and conservation studies. We observe that estimates based on single-diploid genomes instead of population data are reasonable estimators of their species diversity (but not of their population). We also conclude that the heterozygosity and the pSMC harmonic mean measures are similar in their magnitude of variation inside a population and a species, suggesting that the use of pSMC-derived measures, which are more costly, is not necessary for this kind of purpose.

***K*eywords** population genomic, pSMC, heterozygosity, single-genome, genetic diversity

1 Introduction

Genetic diversity can be defined as the genetic variability within species. More concretely, Nei and Li (1979) defined the nucleotide diversity (named π) as the average fraction of pairwise nucleotide differences between randomly sampled genomes from the population. In an idealised population of constant size N, panmixia, no selection and non-overlapping generations (Fisher, 1930; Wright, 1931), nucleotide diversity is theoretically equal to the scaled mutation rate (named θ). This entity combines the arrival of new variants in the population through mutation (quantified by the mutation rate, μ) and the loss of these variants through genetic drift (quantified by $1/2N$), as $\theta = 4N\mu = \pi$. Nucleotide diversity thus reflects the size of the population, which in turn determines the intensity of genetic drift (Fisher, 1930).

In reality, however, populations deviate from this idealized situation. We therefore define the effective population size (named N_e), as the size of an ideal population that has the same nucleotide diversity as the observed population, and hence the same drift intensity. Then the nucleotide diversity measured from a sample of individuals in a non-ideal population corresponds to $\hat{\pi} = 4N_\mathrm{e}\mu$. In addition, the population size is not constant over time, which leads to variations in N_e and thus in the intensity of genetic drift. In this case, nucleotide diversity reflects the mean of the different N_e over time. Note that, by construction, π, as defined by Nei and Li (1979), gives more weight in this mean to ancient nucleotide diversity because it can count the same ancestral variant multiple times, present in different pairs of haplotypes. In this context, it is of interest to better understand what determines N_e, genetic drift, nucleotide diversity and the relationship between them when the population deviates from an ideal model.

Estimating diversity, and therefore N_e, require population genetic data, ideally from a sufficiently large number of individual randomly sampled from the population, which is a limited resource. Although we have sequenced more and more at the population level in recent years (Ellegren, 2014), this effort is biased towards species of scientific, economic or agronomic interest, and covers only a small proportion of all genera, even in well-studied clades such as mammals. This is particularly limiting, for example, in the case of comparative analyses that seek to investigate correlations between nucleotide diversity and the ecological drivers of N_e (e.g. life-history traits) and thus require diversity estimates for a large number of species. Some studies overcome this issue by using the genome-wide heterozygosity of a diploid individual to represent the nucleotide diversity of its population (Figuet *et al.*, 2016; Zoonomia, 2020; Brevet and Lartillot, 2021). This approximation is relevant because diploid recombinant individuals are a mosaic of loci with different histories and genealogies. In fact, the genome-wide heterozygosity of a diploid individual is already an average of many independent genealogies from the population.

A more detailed analysis of the distribution of these genealogies can also be performed using pSMC (Li and Durbin, 2011) which is able to reconstruct the demographic history of the entire population based on a single individual. The pSMC method consists in reconstructing the coalescent tree at each position of the sequence and identifying the segments of consistent coalescent tree (i.e. haplotypes), which can then be analysed to determine N_e variation over time. Since genetic diversity depends on N_e, pSMC methods are very useful for estimating genetic diversity over time. Interestingly, some studies also use the pSMC

methods to output a single metric (the harmonic mean) of genetic diversity (Wilder *et al.*, 2023), instead of the heterozygosity.

Although these single-genome metrics look promising for studying variation in genetic diversity, we don't know their individual variance and the impact of this additional source of variance on the comparative analyses mentioned above. In addition, the heterozygosity and pSMC measures both assume that comparing two sequences from a diploid individual is equivalent to comparing two randomly sampled haploid sequences from the population. This is only relevant if the reproductive matches are random (i.e. if the population is panmictic). Otherwise, the diploid individual may be inbred or outbred which may lead to under- or overestimation of nucleotide diversity.

In this study, we investigate three single-genome metrics of nucleotide diversity using 2517 individuals genomes from 56 population and 7 species (*Equus caballus*, *Bos taurus*, *Ovis aries*, *Chlorocebus sabaeus*, *Homo sapiens*, *Salmo salar*, *Parus Major*). The first metric corresponds to whole genome non-coding heterozygosity (θ_H). The other make use of the pSMC approach and its computation of nucleotide diversity variation over time. To do this, we average the pSMC curve by using harmonic mean to obtain a single estimate, comparable to θ_H. We calculate this mean either on the entire diversity curve or on the older third of this curve. The latter measure should be less sensitive to inbreeding, assuming that inbreeding only affects recent time.

For each metric, we decompose the variance into an individual, population and species component, so as to quantify the variance potentially added to a study when using a single-genome nucleotide diversity metric. We also compare the three measures to determine which one has the least individual-level variation. Finally, we compute a root mean square deviation (RMSD) and a maximum deviation to quantitatively examine the deviation of individual genome measurements from their species mean and potentially detect some outlier individuals.

2 Materials and Methods

2.1 Population data

Population level polymorphism data were obtained from previously published population genomic projects for the following species : *Equus caballus* (EquCab2 assembly in the EVA study PRJEB9799 (Al Abri *et al.*, 2020)), *Bos taurus* (UMD3.1 assembly in the NextGen project: `https://projects.ensembl.org/nextgen/`), *Ovis aries* (Oar_v3.1 assembly in the NextGen project), *Chlorocebus sabaeus* (ChlSab1.1 assembly in the EVA project PRJEB22989 (Svardal *et al.*, 2017)), *Homo sapiens* (GRCh38 assembly in the 1000 Genomes Project (Zheng-Bradley *et al.*, 2017)), *Salmo salar* (ICSASG_v2 assembly in the EVA project PRJEB34225 (Gao *et al.*, 2020)), *Parus Major* (Parus_major1.1 assembly in the EVA project PRJEB24964 (Laine *et al.*, 2016)).

2.2 Genetic diversity

Three alternative estimators of the genetic diversity θ, based on a single individual, were considered: the heterozygosity θ_H as well as two metrics derived from the pSMC (noted θ_C^{All} and θ_C^{Old} in the following).

For each species, we masked the coding part of the genome using gene annotation files (GTF format) downloaded from Ensembl (ensembl.org). Also, we masked the ambiguous positions not assigned to a base nucleotide (A, C, G, T) in the reference genome using the reference genome file (FASTA format) from Ensembl (ensembl.org) to mitigate the effect of sequencing errors.

The heterozygosity, θ_H, is obtained by simply counting the number of heterozygous sites across the whole genome, divided by the total number of callable sites. For the two estimates based on the pSMC, we used the SMC++ software (Terhorst *et al.*, 2017). The pSMC methods use the genome from one individual and the position of the variants to estimate the coalescent times between pairs of alleles at each position of the genome. In other words, they rely not just on the number of heterozygous sites but also on their distribution along the genome, to determine the ancestral recombination points and the coalescent time between them. When an estimate of μ is given as an additional input, the method uses it to rescale its output and thus provides a reconstruction of N_e, itself approximated as a piecewise constant function (Li and Durbin, 2011). In this study, however, we revert this rescaling such that all measures derived from pSMC are in units of θ. This allows us to emancipate ourselves from the problem that μ is not precisely known in all species.

Even if the pSMC methods provide a reconstruction of the variation in θ_C over time, our interest here is to derive a single measure representing the diversity of the whole population. As such, we computed the weighted harmonic mean of θ_C, as in Wilder *et al.* (2023). This define our first estimator derived from pSMC, which we call θ_C^{All}:

$$\theta_C^{\mathrm{All}} = \frac{\sum_{i=1}^{T} t(i)}{\sum_{i=1}^{T} \frac{t(i)}{\theta_C(i)}}, \tag{1}$$

where T is the number of time segments in the SMC++ output and $\theta_C(i)$ is the value of θ_C for segment i, and $t(i)$ is the length of the segment (in generations).

Alternatively, and as an attempt to filter out the input of recent inbreeding that would have impacted the specific individual being considered, we reasoned that recent inbreeding should be interpreted by pSMC as a very recent decrease in N_e. In contrast, the more ancient history of N_e should be more representative of the whole population. Accordingly, we define θ_C^{Old} as the weighted harmonic mean of θ_C over the oldest third of the inferred time period.

2.3 Variance decomposition

For each estimator, the variance across the collection of estimates obtained for all individuals of all species was decomposed into three components: individual, population, and species. We account for the variable number of populations per species and the variable number of individuals per population.

Specifically, we define p as the number of species in the dataset. For each species i ($1 \leq i \leq p$), we define m_i as the number of populations in this focal species, and n_i as the total number of individuals in the species. Finally, n_{ij} is the number of individuals in the population j ($1 \leq j \leq m_i$) of the species i.

The total number of individuals in species i is given by $n_i = \sum_{j=1}^{m_i} n_{ij}$ and the total number of individuals across all species is given by $n = \sum_{i=1}^{p} n_i$.

Noting y_{ijk} the metric of interest for individual k in population j and species i, the total variance is given by:

$$SS_{\text{tot}} = \sum_{i=1}^{p} \sum_{j=1}^{m_i} \sum_{k=1}^{n_{ij}} (y_{ijk} - \overline{y})^2. \tag{2}$$

We decompose SS_{tot} into three components:

$$SS_{\text{tot}} = SS_{\text{ind}} + SS_{\text{pop}} + SS_{\text{sp}}. \tag{3}$$

with

$$SS_{\text{ind}} = \sum_{i=1}^{p} \sum_{j=1}^{m_i} \sum_{k=1}^{n_{ij}} (y_{ijk} - \overline{y}_{ij})^2, \text{ where } \overline{y}_{ij} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} y_{ijk}, \tag{4}$$

$$SS_{\text{pop}} = \sum_{i=1}^{p} \sum_{j=1}^{m_i} n_{ij} (\overline{y}_{ij} - \overline{y}_i)^2, \text{ where } \overline{y}_i = \frac{1}{n_i} \sum_{j=1}^{m_i} \sum_{k=1}^{n_{ij}} y_{ijk}, \tag{5}$$

$$SS_{\text{sp}} = \sum_{i=1}^{p} n_i (\overline{y}_i - \overline{y})^2, \text{ where } \overline{y} = \frac{1}{n} \sum_{i=1}^{p} \sum_{j=1}^{m_i} \sum_{k=1}^{n_{ij}} y_{ijk}. \tag{6}$$

We then compute the relative contribution of each of these component to the total variation.

2.4 Root mean square deviation and maximum deviation

To assess the variation of a given estimator, we compute two metrics: the root mean square deviation (noted RMSD in the following) and the maximum deviation, both in logarithmic scale. The use of a logarithmic scale is important since diversity values can vary over several orders of magnitude across species, such that the variation on the natural scale is dominated by species with the largest diversity.

The RMSD is a measure of the average deviation of an individual measure from the mean of its population in log scale, thus formally defined as:

$$RMSD(y) = \sqrt{\frac{\sum_{i=1}^{n} \left(\log(y_i) - \overline{\log(y)} \right)^2}{n}}, \tag{7}$$

where y_i is the individual measure, $\overline{\log(y)}$ is the mean of the log of the individual measures and n is the number of individuals. The maximum deviation is a measure of the highest deviation of an individual measure from the mean of its population in log scale, as:

$$\text{max\_norm}(y) = \max_{1 \leq i \leq n} \left(\left| \log(y_i) - \overline{\log(y)} \right| \right). \tag{8}$$

3 Results

We collected population polymorphism data from seven population genomics projects representing seven species, 56 populations and 2517 individuals. We used them to obtain measures of genetic diversity based on a single individual, such as heterozygosity and estimates based on pSMC analysis (hereafter referred to as θ_H and θ_C, respectively).

3.1 Individual heterozygosity

To assess the practical value of individual heterozygosity estimates as a proxy for the population or species diversity, we computed the whole genome non-coding heterozygosity on the 2517 individuals of the study. These estimates were plotted against the population mean (Figure 1, in log scale). We then compute a root mean squared deviation (named RMSD in the following) and a maximum deviation for each population (Figure 2).

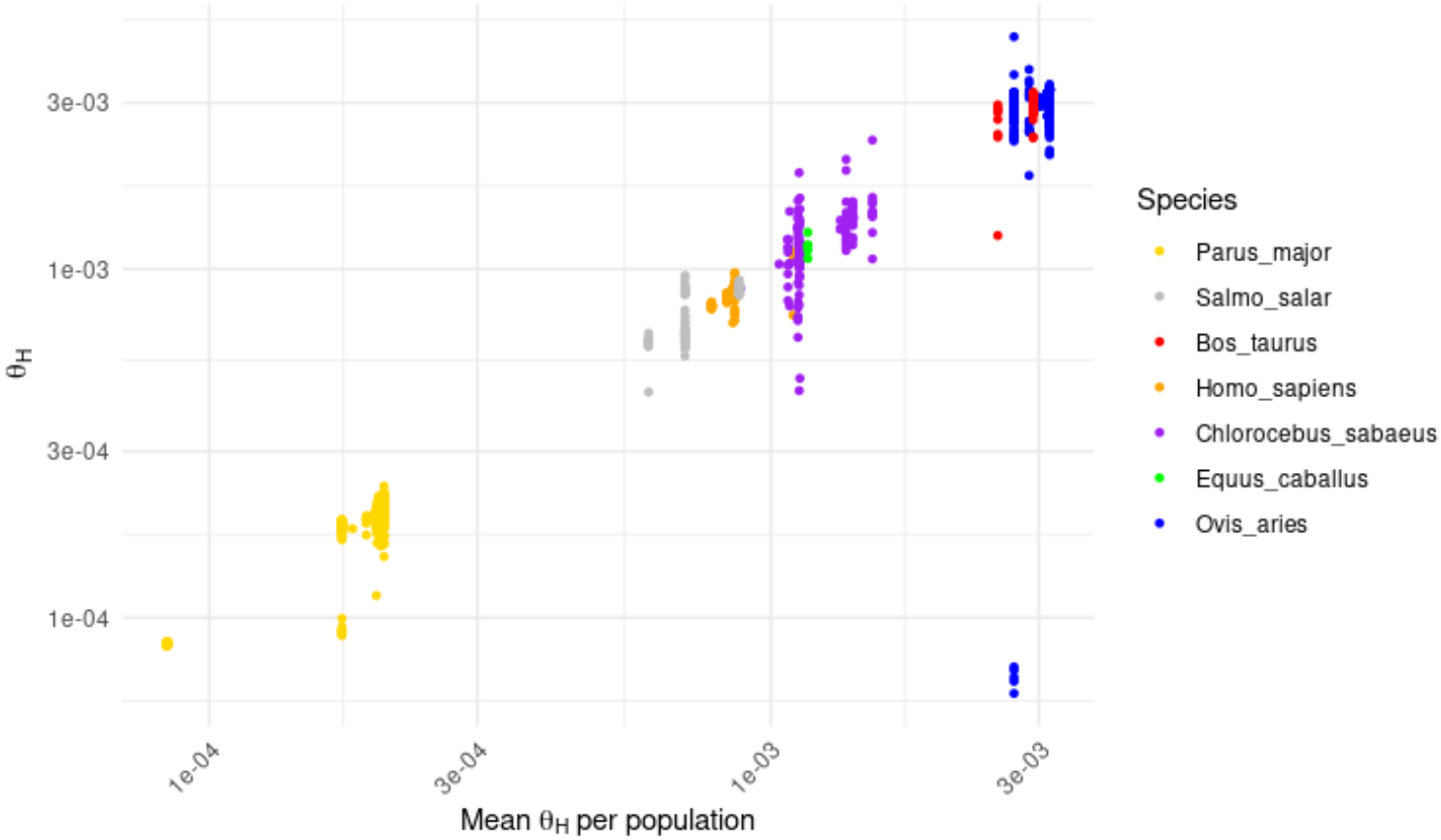


Figure 1: Distribution of individual heterozygosity (θ_H) compared to the mean heterozygosity of its population (natural values). Scales are expressed in log10. Each point represents an individual and each column in represents a population. Species are indicated by colour.

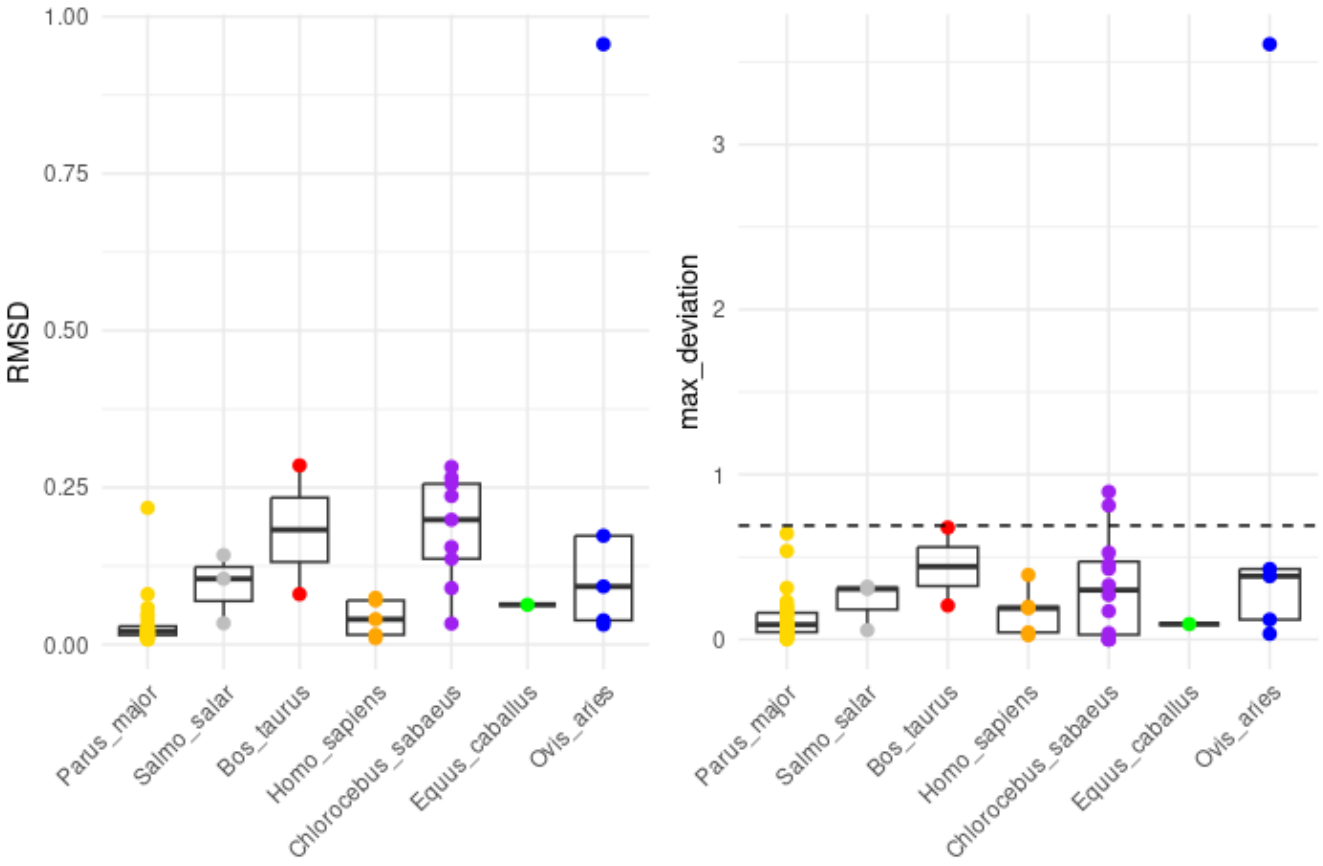


Figure 2: Distribution of the RMSD (left panel) and the maximum deviation (right panel) of the $\log(\theta_H)$ measures in each population, grouped by species in a boxplot. The black dotted line in the right panel corresponds to ln(2), which indicates a deviation of twice the population mean.

Individual estimates show some variability around the population mean (Figure 1). There are a few extreme cases, notably for one population of *Ovis aries* (ISGC population), where individuals were observed with either very high or very low diversity measures compared to the population mean. These extreme points presumably correspond to outbred and inbred individuals, respectively. Except for these exceptional cases, however, the variation is relatively moderate, with a mean RMSD per population on average equal to 0.074 and never exceeding 0.30 (Figure 2). As an example, the *Equus caballus* mean RMSD is 0.063, indicating that the diversity of an individual is typically within 6.3% of the population mean. Regarding the maximum deviation, the dashed line in Figure 2, right panel, represents a cutoff at ln(2)=0.69, corresponding to a diversity two times higher or lower than the population mean. Here we observe 3 populations with at least one individual above this cut-off. By removing the *Ovis aries* ISGC population, we observe an average maximum deviation per population of 0.2, corresponding to a diversity 0.53 times higher or lower than the population mean.

Figure 1 gives a visual representation of the variation of the population mean around the species mean, for each species. It shows that different populations of the same species tend to have very similar mean diversity, with a dispersion that looks smaller or equal to that of different individuals from the same population. In contrast, the variation between species is much larger and nearly over two order of magnitude. This suggests that, if it does not allow one to discriminate between populations, most often, the heterozygosity of an individual will still be a good proxy of the diversity of the species it belongs to.

To investigate this question more quantitatively, we performed an analysis of variance to partition the total variance of individual log-heterozygosity estimates into individual, population and species effects. We found that the individual effect accounts for 3.4% of the total variance, the population effect for 1.2% and the species effect for 95.4%. However, in the case of *Parus major*, there is a very large number of individuals (1845 individuals, while *Ovis aries*, the second most abundant species in the dataset, presents 273 individuals). Many of these individuals turn out to have very similar levels of diversity, inducing a strong influence of this species in our analysis. Excluding *Parus major* from the analysis of variance gives an individual effect of 25.2%, a population effect of 3.6% and a species effect of 71.2%. Thus, relying on this last analysis, 71.2% of the total variation is attributable to the species, which gives a more quantitative evaluation of how much information about species-level diversity can be gained based on individual heterozygosity estimates.

Overall, even though it contributes a non-negligible amount of additional variance (up to 28.8% when adding individual and species variance, and about 30% based on Figure 2), measuring the heterozygosity of a single individual can be practically useful for species-level analyses. However, as seen here in the case of the *Ovis aries* ISGC population, if the sampled individual is inbred or outbred, as suspected in this case, this can lead to incorrect conclusions about species-level diversity.

3.2 Estimates based on the pSMC measures

In an attempt to derive less variable estimates of the species-level diversity, we further explored the structure of the heterozygosity profiles of each individual using the pSMC approach. Since, in practice, N_e always varies over time, any diversity measure of a population implicitly corresponds to a mean of N_e values (or, equivalently, knowing μ, of $4N_e\mu$) over time. On the other hand, the pSMC provide a detailed reconstruction of the history of N_e through time. This allows direct observation of some events, such as population expansion, but also individual aspects, such as recent inbreeding, which can then be filtered out if necessary. Here, we applied SMC++ to the 2517 individuals in the study and produced a plot for each population representing the variation over time of $\theta_C = 4N_e\mu$ (Figure 3 and Supplementary Material).

The populations in Figure 3 show different variation schemes, and in each population the individuals appear to have a common pattern of variation, except for a few outliers. We observe that the recent pSMC estimates are more variable between individuals.

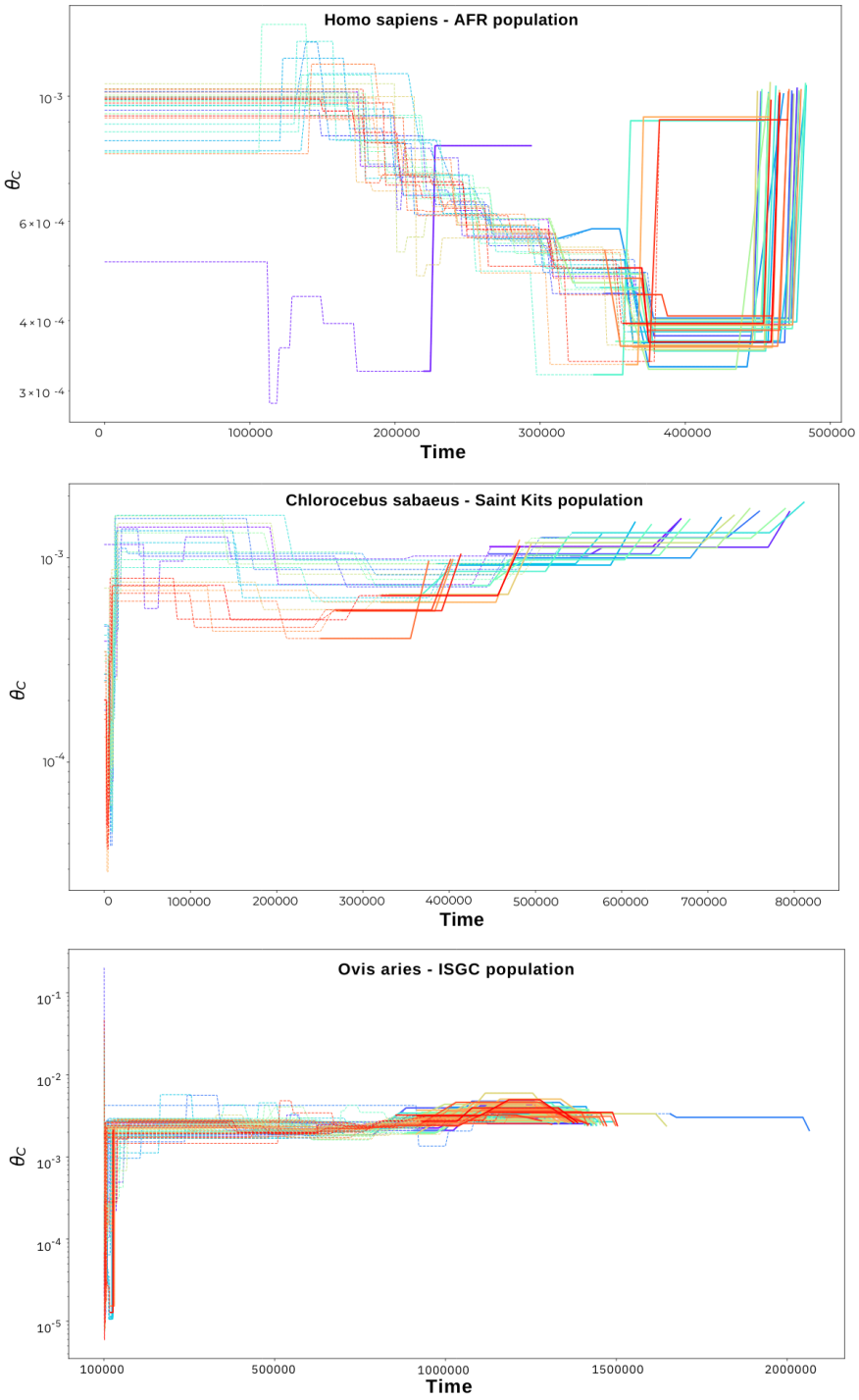


Figure 3: Reconstruction of θ_C across time for 3 of the 56 populations by SMC++ (Terhorst *et al.*, 2017) (a sample of other populations graph is provided in Supplementary Material). Each line represents an individual and the full line represents the oldest part of the reconstruction (used for θ_C^{Old}). The colours are only used to distinguish the lines, they have no meaning.

Based on the output of pSMC, we now want to derive a discrete estimate of θ_C. Theoretically, the raw heterozygosity of an individual in a panmictic population is expected to reflect the harmonic mean of the variation of N_e through the history of its population. Accordingly, we first computed this harmonic mean over the entire profile returned by pSMC (noted θ_C^{All}). Thus, in principle, θ_C^{All} should be closed to θ_H. Alternatively, and to filter out the observed variability in the recent time, which could be due to inbreeding or outbreeding events, we compute a harmonic mean over the older third of each curve (noted θ_C^{Old}) (which corresponds to the full line in Figure 3). Contrary to our expectation, the three measures show a more contrasting relationship, largely depending on the species studied (Figure 4). In general, heterozygosity values tend to be higher than estimates based on pSMC (although not universally, as seen in *Equus caballus* and *Bos taurus*) and more aligned with θ_C^{Old} than θ_C^{All} (with exceptions such as *Parus major*). The reasons for this pattern remain unclear.

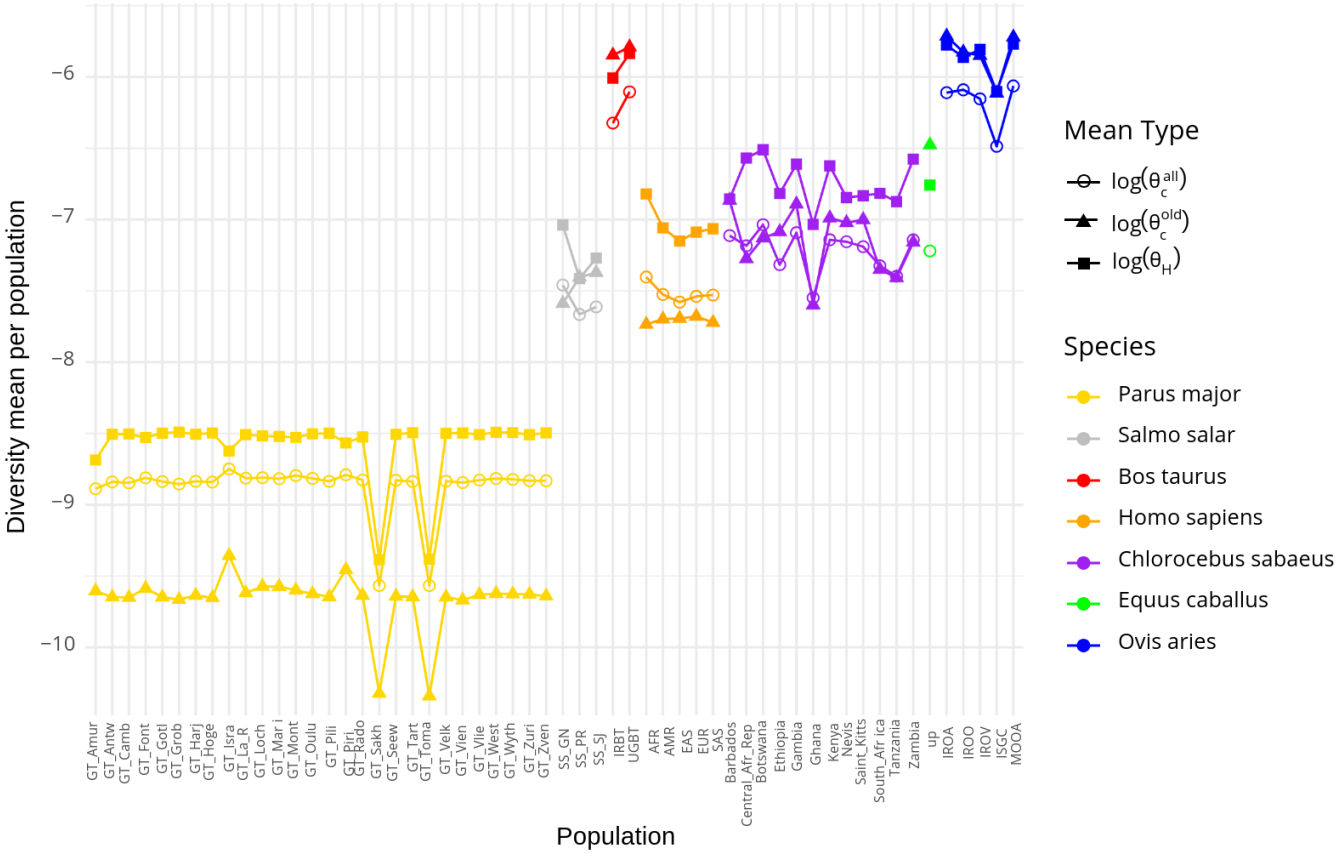


Figure 4: Mean value of the $\log(\theta_H)$, $\log(\theta_C^{\text{All}})$ and $\log(\theta_C^{\text{Old}})$ per population. The colours differentiate the species, and the shape of the points represent the three metrics.

As done above for heterozygosity, we then performed a variance decomposition analysis for θ_C^{All} and θ_C^{Old}, into individual, population and species effects (Table 1). As already observed with heterozygosity, the pSMC metrics also show a higher level of variance at the species level. It is worth noting that θ_C^{All} shows a slightly higher level of variation at the individual level than the other two metrics. Conversely, θ_C^{Old} appears to be more stable, with a slightly larger proportion of variance explained at the species level.

We observe a similar trend in the results when the *Parus major* contribution is removed from the analysis (Table 2), although the proportion of variance at the individual level is higher. However, θ_C^{Old} appears to be less affected than the other two estimates.

Thus, using a filtered estimate based on the pSMC seems to provide some improvement when the most recent part of the curve has been filtered out. This improvement is small, however, raising the question whether pSMC represents a useful alternative to heterozygosity, given its computational complexity.

	individual effect	population effect	species effect
θ_H	3.4	1.2	95.4
θ_C^{All}	5	1	94
θ_C^{Old}	3.5	0.5	96

Table 1: Decomposition of variance in individual, populational and species effect, for the three single-genome diversity metrics (H, θ_C^{All} and θ_C^{Old})

	individual effect	population effect	species effect
θ_H	25.2	3.6	71.2
θ_C^{All}	28.7	3.1	68.2
θ_C^{Old}	25.9	2	72.1

Table 2: Decomposition of variance in individual, populational and species effect, for the three single-genome diversity metrics (θ_H, θ_C^{All} and θ_C^{Old}) without *Parus major*

We compute the RMSD and maximum deviation metrics for the pSMC measures and compare them with the previous heterozygosity results (Fig.5).

The results are quite variable depending on the species but globally, the pSMC metrics don't exhibit lower RMSD or maximum deviation compared to the heterozygosity. They also appear to be as sensitive to outliers as raw heterozygosity (Fig.5). Moreover, the θ_C^{All} RMSD is higher than the θ_C^{Old} only for one of the 7 species (*Bos taurus*). It shows that the use of θ_C^{Old} does not seem to be able to filter out recent inbreeding or other aspect of the heterozygosity profile of the individual that does not reflect the population.

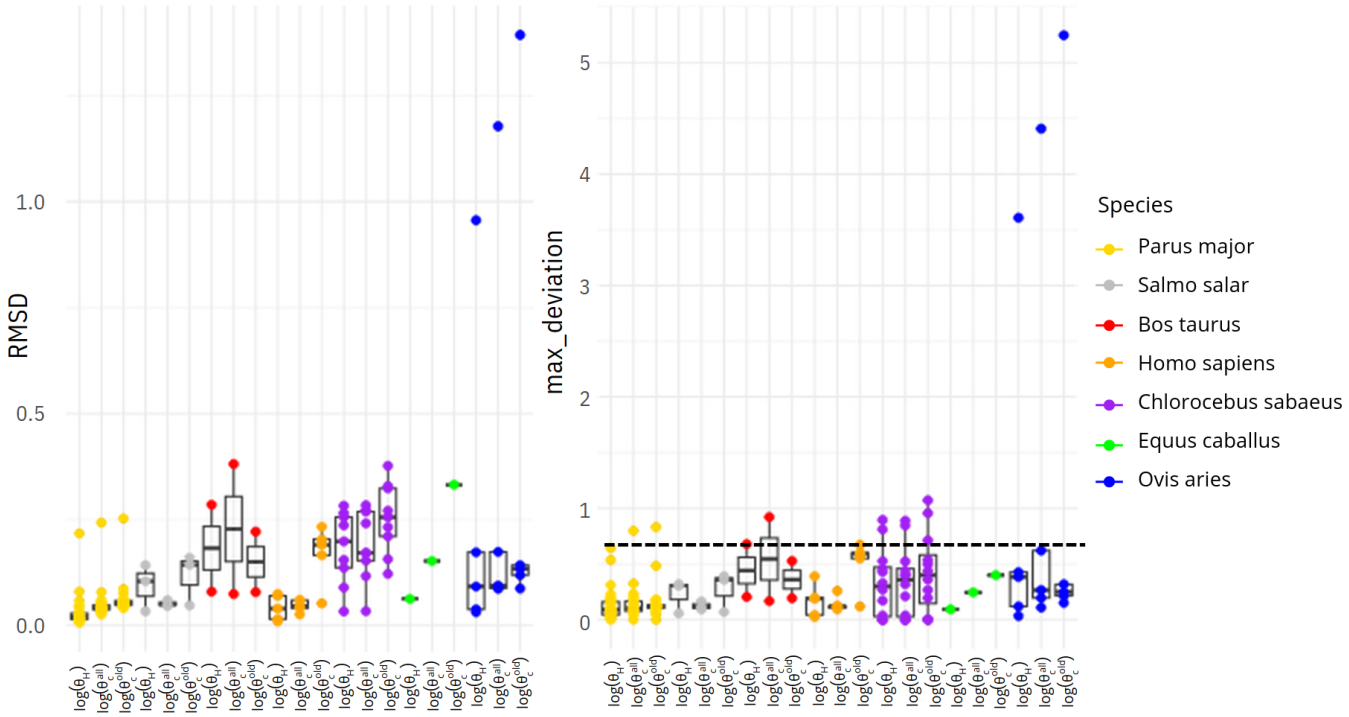


Figure 5: Distribution of the root mean square deviation (left panel) and the maximum deviation (right panel) of the three single-genome diversity measures in each population, grouped by specie. The black doted line in right panel correspond to ln(2) which indicate a deviation of twice the mean population value

3.3 Discussion

Coalescent theory states that the genomes of a diploid individual from a sexual and recombining species is a mosaic of the genomes of the different past individuals in its population. In turn, if the population is at least approximately panmictic, the genetic ancestors of the individual form a representative sample of the population throughout its history. This suggests that using the genome of a single individual would be sufficient to reconstruct estimates of population diversity.

In this study, we used population data to first produce a whole-genome heterozygosity estimation, independently for each individual, and decompose its variance into what is attributable to the individual, the population and the species. The idea was to get a rough quantitative estimate of how much additional variance would be contributed to a species-level analysis by using these individual-based proxies. Using seven vertebrate species, we find that most of the variance in observed heterozygosity is explained at the species level (Table 1). On the other hand, we cannot clearly conclude that two individuals from the same population are more similar than two individuals of the same species but from different populations. This is even more striking when *Parus major* is removed from the variance analysis (Table 2). These results lead us to conclude that the genome-wide heterozygosity of a single individual represents the diversity of its species, but not so clearly that of its particular population. This last observation can be approached from at least two directions. First, there is a lack of consistency in the definition of a population across population genetic projects,

making it difficult to distinguish between them. Second, populations can be admixed, making it difficult to separate their signals. However, practically, our analysis provides some insights for the species-level genomic studies. For example, if we want to correlate single-individual diversity estimates with life-history traits or other measures putatively reflecting a long-term N_e, then we know that the use of a single individual already entails up to 30% of variance (including population level) in addition to other factors such as the discrepancy between short-term and long-term N_e.

Regarding our results based on heterozygosity measures, we suspected inbreeding to be a major source of variation between individuals on the genome-wide heterozygosity, as inbreeding creates long stretches of homozygosity in the genomes, which reduces the diversity. In an attempt to reduce the effect of individual variance due to inbreeding, we tried another diversity estimate based on a single individual genome by processing the output of the pSMC analyses. The pSMC reconstructs temporal variation in effective population size, which we can convert into diversity estimates using the relationship $\theta_C = 4N_e\mu$ and a known and fixed mutation rate (μ). We expected pSMC to interpret the potential long runs of homozygosity as a recent and sharp decrease in N_e. To compare pSMC diversity measures with heterozygosity, we averaged the different curves in two ways, either over the entire period covered by the method (named θ_C^{All}), or using the older third of the curves (named θ_C^{Old}), assumed to be more stable and representative of the species if the individual effects correspond to recent times. In Figure 3, however, we sometimes see a more global shift of certain curves over time, rather than recent declines in diversity (see *Homo sapiens* AFR population in Figure 3) or also both patterns simultaneously (see *Ovis aries* ISGC population in Figure 3). For example, in the plot of the problematic ISGC population of *Ovis aries*, some curves start at a more recent or older time than others. Thus, our attempt to use pSMC to filter out individual effects, based on the assumption that they should only be visible in the recent past, was not successful, as we continue to observe outliers individuals and the same distribution in the variance decomposition (Figure 5 and Table 1 and 2). This suggests that inbreeding in the outliers may be more ancestral than thought, rendering the pSMC method ineffective in reducing individual variance.

Therefore, prior to diversity analyses, we still need to find a way to detect and remove outliers. For example, analyses to detect long segments of homozygosity could be implemented prior to heterozygosity or pSMC measurements in order to detect certain inbred individuals. These measures can be performed on single individuals, provided that we have good quantification criteria for considering an individual to be inbred on the basis of such measures (at what size are homozygosity runs characteristic of inbreeding? How many are required?) Depending on the case, individuals can then be removed from the analysis or part of their genome can be masked in inbred regions. This would certainly reduce the number of outliers.

Among the issues that could be examined is the question of the specific pSMC implementation chosen for this study. In our analysis, we used the SMC++ software (Liu and Fu, 2015; Terhorst *et al.*, 2017) rather than the original pSMC implementation (Li and Durbin, 2011) due to computational constraints (i.e. more than 2000 individuals). In a article, Patton and colleagues (Patton *et al.*, 2019) examine the behaviour of three SMC-like methods, which are pSMC (Li and Durbin, 2011), MSMC (Schiffels and Durbin, 2014) and SMC++, an hybrid between SMC and SFS-like method (Liu and Fu, 2015). They used genomes of different

quality (estimated by their level of fragmentation) to assess the impact of the method or implementatio on N_e reconstructions. Since the SMC-like methods use the length of homogeneous heterozygosity blocks to infer N_e, they should be more affected by fragmentation than the SFS-like methods. The SMC-like methods should also be more robust to ancient time estimation than the SFS-like methods and vice versa (Patton *et al.*, 2019). Comparing the three methods, the authors found similar results when the genome contains less than 5000 scaffolds. The SMC++ methods appear to be less sensitive to fragmentation than others, but show false inference of demographic variation (Patton *et al.*, 2019). Indeed, SMC++ is an hybrid of SFS and SMC methods, it is supposed to provide a more accurate reconstruction of ancient (from SMC) and recent (from SFS) N_e signals, but it also has the drawbacks of both, creating a possible conflicting signal. This results can explain why we don't observe a strong congruence between our θ_H and θ_C measures (Figure 4). The authors conclude that the pSMC method (Li and Durbin, 2011) is the most precise and accurate method. In regard to this study, our analysis should be further explored using the pSMC methods from Li and Durbin (2011) on a sample of our individuals.

Globally, with these contrasted results between the three diversity measures, we cannot say which one is best, as they are varying in the same order of magnitude and show similar dispersion to their species mean. However, the heterozygosity computation is by far less costly. We therefore suggest that it is not necessary to use a pSMC method for a point estimate of θ_C or N_e (as done in Wilder *et al.* (2023) study), as θ_H seems to be equally good. On the other hand, pSMC methods remain relevant for studying θ_C and N_e dynamics over time.

References

Al Abri, M. A., Holl, H. M., Kalla, S. E., Sutter, N. B., and Brooks, S. A. 2020. Whole genome detection of sequence and structural polymorphism in six diverse horses. *PLoS ONE*, 15(4): e0230899.

Brevet, M. and Lartillot, N. 2021. Reconstructing the history of variation in effective population size along phylogenies. *Genome Biology and Evolution*, 13(8).

Ellegren, H. 2014. Genome sequencing and population genomics in non-model organisms. *Trends in ecology & evolution*, 29(1): 51–63.

Figuet, E., Nabholz, B., Bonneau, M., Mas Carrio, E., Nadachowska-Brzyska, K., Ellegren, H., and Galtier, N. 2016. Life history traits, protein evolution, and the nearly neutral theory in amniotes. *Molecular Biology and Evolution*, 33(6): 1517–1527.

Fisher, R. A. 1930. The distribution of gene ratios for rare mutations. *Proceedings of the Royal Society of Edinburgh*, 50: 205–220.

Gao, G., Pietrak, M. R., Burr, G. S., Rexroad, C. E., Peterson, B. C., and Palti, Y. 2020. A New Single Nucleotide Polymorphism Database for North American Atlantic Salmon Generated Through Whole Genome Resequencing. *Frontiers in Genetics*, 11.

Laine, V. N., Gossmann, T. I., Schachtschneider, K. M., Garroway, C. J., Madsen, O., Verhoeven, K. J. F., de Jager, V., Megens, H.-J., Warren, W. C., Minx, P., Crooijmans, R. P. M. A., Corcoran, P., Sheldon,

B. C., Slate, J., Zeng, K., van Oers, K., Visser, M. E., and Groenen, M. A. M. 2016. Evolutionary signals of selection on cognition from the great tit genome and methylome. *Nature Communications*, 7(1): 10474.

Li, H. and Durbin, R. 2011. Inference of human population history from individual whole-genome sequences. *Nature*, 475(7357): 493–496.

Liu, X. and Fu, Y.-X. 2015. Exploring population size changes using snp frequency spectra. *Nature genetics*, 47(5): 555–559.

Nei, M. and Li, W.-H. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences*, 76(10): 5269–5273.

Patton, A. H., Margres, M. J., Stahlke, A. R., Hendricks, S., Lewallen, K., Hamede, R. K., Ruiz-Aravena, M., Ryder, O., McCallum, H. I., Jones, M. E., *et al.* 2019. Contemporary demographic reconstruction methods are robust to genome assembly quality: a case study in tasmanian devils. *Molecular biology and evolution*, 36(12): 2906–2921.

Schiffels, S. and Durbin, R. 2014. Inferring human population size and separation history from multiple genome sequences. *Nature genetics*, 46(8): 919–925.

Svardal, H., Jasinska, A. J., Apetrei, C., Coppola, G., Huang, Y., Schmitt, C. A., Jacquelin, B., Ramensky, V., Müller-Trutwin, M., Antonio, M., Weinstock, G., Grobler, J. P., Dewar, K., Wilson, R. K., Turner, T. R., Warren, W. C., Freimer, N. B., and Nordborg, M. 2017. Ancient hybridization and strong adaptation to viruses across African vervet monkey populations. *Nature Genetics*, 49(12): 1705–1713.

Terhorst, J., Kamm, J. A., and Song, Y. S. 2017. Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nature Genetics*, 49(2): 303–309.

Wilder, A. P., Supple, M. A., Subramanian, A., Mudide, A., Swofford, R., Serres-Armero, A., Steiner, C., Koepfli, K.-P., Genereux, D. P., Karlsson, E. K., Lindblad-Toh, K., Marques-Bonet, T., Munoz Fuentes, V., Foley, K., Meyer, W. K., Zoonomia Consortium, Ryder, O. A., and Shapiro, B. 2023. The contribution of historical processes to contemporary extinction risk in placental mammals. *Science*, 380(6643): eabn5856.

Wright, S. 1931. Evolution in Mendelian populations. *Genetics*, 16(2): 97–159.

Zheng-Bradley, X., Streeter, I., Fairley, S., Richardson, D., Clarke, L., Flicek, P., and the 1000 Genomes Project Consortium 2017. Alignment of 1000 Genomes Project reads to reference assembly GRCh38. *GigaScience*, 6(7): gix038.

Zoonomia, C. 2020. A comparative genomics multitool for scientific discovery and conservation. *Nature*, 587(7833): 240–245.

Troisième partie

Discussion et Perspectives

7

Discussion et perspectives

À la fin du XX^{e} siècle, Kimura et Ohta ont formulé la théorie neutre et quasi neutre de l'évolution qui met en lumière le rôle majeur de la dérive génétique, par rapport à celui de la sélection naturelle, dans la structuration des génomes (Kimura, 1968a,b; Ohta, 1973, 1974). C'est notamment Ohta qui décrit une continuité d'effet des mutations allant d'une mutation sans effet sur l'individu à fortement impactante (Ohta, 1973). Dans un contexte où les mutations avantageuses sont suffisamment rares pour être négligées, la théorie quasi-neutre met en avant le fait que l'efficacité de la sélection naturelle à éliminer des mutations faiblement délétères dans les génomes, avant leur fixation par dérive, dépend de la taille efficace des populations, N_e. Cette théorie, bien que d'abord très critiquée par les partisans de la théorie néo-darwinienne, a trouvé un appui considérable dans l'essor des données moléculaires qui ont permis de la vérifier à l'échelle de quelques gènes et espèces (Ohta, 1973; Kimura, 1983).

Aujourd'hui, avec la quantité grandissante de données génomiques, nous pouvons tenter de fournir une deuxième salve de validation empirique de cette théorie, et ce, à deux échelles évolutives. En effet, à l'échelle de la macro-évolution, nous pouvons mesurer l'efficacité de la sélection à l'aide du d_N/d_S en utilisant, à cette fin, des modèles à codons (sous-section 3.2.2). À l'échelle de la micro-évolution, nous pouvons mesurer l'efficacité de la sélection à l'aide du π_N/π_S (sous-section 3.3.4). Cependant, une difficulté persiste quant à la mesure des tailles efficaces de population, qu'il faut approximer via des estimateurs plus ou moins directs. Chez les mammifères, nous supposons une relation négative entre N_e et les traits d'histoire de vie. Ces traits forment alors des proxies indirects de N_e à confronter au d_N/d_S. En revanche, à l'échelle de la génétique des populations, le polymorphisme synonyme est connu pour être un proxy plus direct de N_e via la relation $\pi_S = 4N_e\mu$ que l'on peut donc confronter au π_N/π_S.

Lors de mon travail de thèse (chapitre 5), j'ai mis en forme un jeu de données contenant les alignements de 6000 gènes orthologues appartenant à 150 espèces de mammifères, ce qui m'a permis de reconstruire le d_N/d_S le long de leur phylogénie. J'ai également tiré parti de l'hétérozygotie de chacun des individus et du fait qu'en présence de recombinaison et dans des populations raisonnablement panmictiques, le génome d'un individu représente la diversité des génomes de sa population, pour

estimer π_S et π_N/π_S. Le jeu de données a été analysé avec une méthode développée dans mon laboratoire d'accueil, qui permet de reconstruire de façon intégrative, via un seul mouvement brownien multivarié, l'évolution des différents traits précédemment évoqués, le long de la phylogénie des mammifères, ainsi que d'estimer leur matrice de variance-covariance (sous-section 3.5.3).

Au niveau macro-évolutif comme micro-évolutif, j'ai observé des corrélations entre proxies de N_e et les mesures d'efficacité de la sélection, ce qui permet de confirmer la théorie quasi-neutre, à l'échelle génomique, chez les mammifères. J'ai également noté une relation positive entre les deux estimateurs d'efficacité de la sélection, d_N/d_S et π_N/π_S. Au niveau, des proxys de N_e, la relation entre échelles micro et macro-évolutive reste plus incertaine. Globalement, les relations impliquant π_S et les autres traits sont plutôt faibles. Une explication envisageable pour ce résultat serait que l'utilisation de l'hétérozygotie n'est finalement pas appropriée pour estimer le polymorphisme synonyme.

Cette possibilité a motivé mon deuxième travail de thèse visant à vérifier, à l'aide de données populationnelles, si l'hétérozygotie d'un génome représente ou non le polymorphisme de sa population. J'ai[1] établi que, globalement, la mesure individuelle est imparfaite et bruitée, mais qu'elle ne dévie pas déraisonnablement de la moyenne populationnelle (chapitre 6). J'ai cependant détecté quelques individus potentiellement consanguins qui peuvent, s'ils sont utilisés seuls, fausser l'estimation du polymorphisme. Cette observation a motivé la construction d'un autre indice de diversité, à partir d'un seul génome, par filtrage des profils de variation de N_e au cours du temps, estimés par une méthode PSMC (Li and Durbin, 2011; Terhorst *et al.*, 2017), espérée plus robuste à la consanguinité. Je n'ai cependant pas observé d'impact significatif de cette méthode sur les estimées obtenues sur ces individus suspects, ni de réduction de l'écart à la moyenne populationnelle. De plus, après une analyse de variance sur ces deux indicateurs individuels, j'ai pu constater que la majorité de la variance est portée par le niveau espèces, ce qui est attendu, puis par le niveau individus. Cela m'a permis de conclure que l'hétérozygotie d'un individu est un proxy du polymorphisme de son espèce, mais pas de sa population. Plus finement, cela a permis d'effectuer une évaluation un peu plus quantitative de l'excès

1. Avec l'aide de Thibault Latrille et Nicolas Lartillot

de variance apportée par cette approche : de l'ordre de 30 %.

Ainsi, concernant les relations plus faibles entre π_S et les autres traits étudiés, je ne crois pas que celles-ci puissent être entièrement, ni même majoritairement dues à l'approximation du polymorphisme par l'hétérozygotie, bien qu'il faille définitivement faire très attention aux individus consanguins. En revanche, elle peut s'expliquer par d'autres facteurs développés plus loin en discussion comme l'impact du taux de mutation sur la relation entre π_S et N_e ou l'incertitude concernant l'estimation des cibles mutationelles synonymes et non-synonymes. Quoi qu'il en soit, les résultats principaux de ma thèse semblent confirmer la théorie quasi-neutre de l'évolution aux deux échelles évolutives chez les mammifères. Quelques perspectives sont envisagées dans la suite de la discussion, comme la mise à l'épreuve de cette théorie à d'autres échelles taxonomiques, la reconstruction directe de N_e le long de la phylogénie, ou l'exploitation du jeu de données de 6000 gènes orthologues de mammifères et leur hétérozygotie pour répondre à d'autres questions concernant les processus évolutifs. S'ensuivra également une discussion un peu plus élargie concernant le travail avec des données empiriques et l'impact environnemental de ce type d'études, ainsi que quelques pensées personnelles concernant le domaine de la recherche et l'importance de la vulgarisation scientifique.

7.1 Un peu plus de N_e

Dans mon travail principal de thèse (chapitre 5), j'observe une faible relation entre π_S et les autres traits et notamment avec les traits macro-évolutifs comme le d_N/d_S et les traits d'histoire de vie. Ce résultat pourrait être attendu si les échelles micro et macro-évolutive sont découplées. Or, j'observe une forte relation entre π_N/π_S, un autre trait lié à la micro-évolution, et les traits macro-évolutifs, ce qui semble contredir cette hypothèse. Comme discuté dans le chapitre 5, la forte relation entre d_N/d_S et π_N/π_S peut s'expliquer par une contamination du d_N/d_S par le polymorphisme résiduel Cette contamination agirait de sorte que le d_N/d_S apparent est en réalité une valeur intermédiaire entre le vrai d_N/d_S et le π_N/π_S, ce qui augmente artefactuellement le coefficient de corrélation entre ces deux entités. Cependant, cette hypothèse de contamination ne permet pas d'expliquer la forte relation entre π_N/π_S et les traits d'histoire de vie qui, elle, suggère qu'il y a bien un

couplage entre les N_e micro et macro-évolutifs. En revanche, la faible corrélation entre π_S et traits d'histoire de vie pourrait s'expliquer via deux autres effets. Le premier serait que π_S est mal estimé et notamment que les cibles mutationelles synonymes sont mal estimées. Dans l'étude, π_S correspond au nombre de sites polymorphes synonymes observés dans les gènes orthologues utilisés, divisé par le nombre de sites appelables dans ces gènes. Or, il est possible que le nombre de sites appelable soit mal estimé. À noter que, π_N/π_S est exempté de la potentielle mauvaise approximation du nombre de cibles synonymes et non-synonymes, qui se simplifie dans le rapport entre π_N et π_S, ce qui pourrait expliquer ses plus fortes relations avec les autres traits. Pour adresser ce problème, d'autres méthodes plus robustes de variant calling pourraient être utilisées, comme GATK (McKenna *et al.*, 2010). On pourrait également tenter de réaliser une meilleure estimation des sites appelables.

Une seconde hypothèse, plus indépendante du traitement et de la qualité des données, pourrait être que le taux de mutation μ est positivement corrélé aux traits d'histoire de vie (Lanfear *et al.*, 2014; Bergeron *et al.*, 2023; Lynch *et al.*, 2023). Dans ce cas, quand N_e augmente, μ diminue et π_S est réduit dans sa variation, par les effets compensatoires de N_e et μ et ne corrèle que partiellement avec les traits d'histoire de vie. Une façon de vérifier cela est de corriger π_S par le taux de mutation afin d'obtenir directement N_e tel que $N_e = \pi_S/4\mu$.

7.1.1 Découpler π_S et μ

Des travaux en vue de démêler les effets du taux de mutations et de N_e sur π_S ont déjà été entamés par Brevet and Lartillot (2021) via leur modèle CoevolNe (voir sous-section 4.4.2). Pour rappel, Brevet and Lartillot (2021) ont développé deux modèles dans leur article. Le premier est un modèle phénoménologique où N_e, μ, le temps de génération (noté τ), d_N/d_S et π_N/π_S sont tous des variables libres, évoluant conjointement le long de la phylogénie par un processus brownien multivarié. N_e et μ permettent de déterminer le π_S aux feuilles tandis que μ et τ déterminent le d_S sur les branches. Le deuxième modèle est un modèle mécaniste où N_e, μ et τ sont libres et déterminent π_S et d_S comme précédemment. Cependant, π_N/π_S et d_N/d_S sont des fonctions allométriques de N_e (figure 7.1). Dans les deux

cas, la valeur de N_e aux feuilles est directement utilisé dans la relation $\pi_S = 4N_e\mu$. Il n'y a donc pas de distinction court/long terme sur N_e. De plus, le modèle CoevolNe est computationellement très lent.

En parallèle, lors de mon travail de thèse, j'ai utilisé FastCoevol, une version améliorée de Coevol (Lartillot and Poujol, 2011), qui est applicable, grâce à l'approximation des mappings, de façon efficace sur l'ensemble des 6000 alignements de gènes orthologues précédemment décris (sous-section 3.5.3). De plus, FastCoevol ajoute une surdispersion au d_S et d_N/d_S permettant de prendre en compte un bruit supplémentaire induit soit par les fluctuations de court terme du d_N/d_S, non capté par l'approximation brownienne, soit par les erreurs sur le d_N/d_S induites par les erreurs d'alignement.

Dans l'idée de poursuivre plus loin le travail de Brevet et Lartillot sur la reconstruction phylogénétique de N_e, nous avons développé une nouvelle version de leur modèle, nommée FastCoevolNe, qui recrute les avantages computationnels de FastCoevol pour les utiliser dans le cadre du modèle de Brevet et Lartillot. Nous nous sommes basés sur la version mécaniste de CoevolNe qui fait des hypothèses concernant les relations entre π_N/π_S, π_S et d_N/d_S car nous les avons observées dans le chapitre 5 avec FastCoevol. Nous avons également ajouté à ce modèle, une distinction de N_e en un N_e macro-évolutif (noté N_e^l) et un N_e micro-évolutif (noté N_e^s)(figure 7.1).

Le modèle considère maintenant un N_e^l, μ et τ qui évoluent de façon brownienne le long de l'arbre avec toujours μ et τ qui servent à déterminer le d_S le long des branches. Le N_e^s est modélisé par une surdispersion du N_e^l aux feuilles puis il est utilisé avec μ pour reconstruire π_S. Le d_N/d_S est, comme dans CoevolNe, défini comme une fonction allométrique de N_e^l tandis que le π_N/π_S, lui, n'est pas pris en compte pour ajuster le modèle. Le modèle fait ainsi une hypothèse mécaniste forte sur la relation entre d_N/d_S et N_e où il devient un proxy direct de N_e. De plus, le modèle ignore les problèmes d'estimation de π_S. Par conséquent, il fait l'hypothèse que π_S reflète fidèlement le N_e de court terme (et potentiellement les effets confondants de μ). Les quatre modèles (Coevol, FastCoevol, CoevolNe mécaniste et FastCoevolNe) sont résumés dans la figure 7.1.

Il est important de souligner que le modèle est à ce jour dans sa version préliminaire et que les analyses réalisées à partir de celui-ci le sont tout autant.

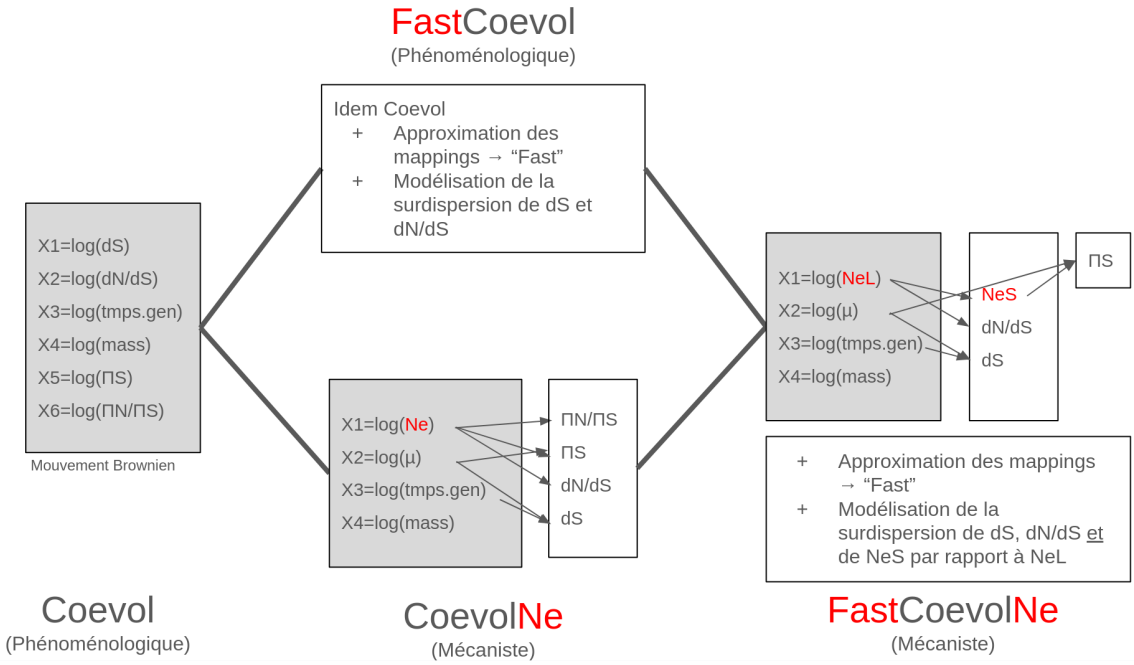


Figure 7.1 – *Schematisation de **Coevol** puis de son évolution indépendante en **FastCoevol**, intégrant une approximation des mappings et une modélisation de la surdispersion du d_N/d_S, et en **CoevolNe** (mécaniste) intégrant la reconstruction de μ et N_e à partir de π_S, d_S et le temps de génération. Les deux modèles ont ensuite été réunis en **FastCoevolNe** avec en plus une distinction du N_e de long terme (noté ici NeL) et du N_e de court terme (noté ici NeS). Les zones grises correspondent aux paramètres auquels est appliqué le mouvement Brownien multivarié. À chaque fois, le modèle prend en entrée un arbre, des données de traits d'histoire de vie, des alignements de gènes (ou statistique de mapping) et des données de polymorphisme.*

Nous avons appliqué FastCoevolNe aux données du chapitre 5, soit les 6000 alignements de gènes orthologues provenant de 144 espèces de mammifères. Nous n'avons pas masqué les données de polymorphismes des espèces qui nous semblent suspectes, comme cela a pu être fait dans le chapitre 5. La figure 7.2 présente la reconstruction de $\log(N_e^l)$ le long de la phylogénie des mammifères par FastCoevolNe. On retrouve les mêmes groupes à plus faible N_e que dans le chapitre 5 ce qui est un résultat plutot encourageant pour la suite de l'analyse.

Le modèle FastCoevolNe fait l'hypothèse forte que le d_N/d_S est un proxy parfait du N_e^l et que $\pi_S = 4N_e^s\mu$. Ceci implique tout un ensemble de relations déterministes qui sont présupposées, *a priori*, entre certains paramètres comme la relation entre le N_e^l

reconstruit et le d_N/d_S empirique ou entre π_S et N_e^s. Ainsi, je m'abstiens, *a posteriori*, d'analyser ces relations. En revanche, FastCoevolNe reste naïf concernant la relation des différents traits avec π_N/π_S puisque celui-ci est exclu du modèle. Le π_N/π_S est donc une mesure indépendante idéale pour l'analyse *a posteriori* des résultats.

Maintenant que N_e^s a été reconstruit à partir de $\pi_S = 4N_e^s\mu$, nous pouvons vérifier si c'est l'effet de μ sur π_S qui engendrait de faibles relations observées entre π_S et les autres traits. En effet, si tel est le cas (et si notre estimée de μ est fiable), alors notre estimée de N_e^s devrait, elle aussi, être fiable, et alors la relation entre N_e^s et π_N/π_S devrait être plutôt forte, en tout cas, plus qu'entre N_e^l et π_N/π_S. Lors de l'analyse avec les données de mammifères, nous obtenons un coefficient de corrélation de 0.29 pour la relation entre N_e^s et π_N/π_S contre 0.24 pour N_e^l avec π_N/π_S. L'effet va dans le bon sens, mais reste plutôt faible ce qui n'est pas suffisant pour confirmer nos attendus. Ainsi, même en corrigeant pour l'effet du taux de mutation, la taille efficace de court terme reconstruite par FastCovolNe sur la base d'un π_S jugé fiable, semble avoir un lien faible ou inexistant avec π_N/π_S et plus généralement avec les traits d'histoire de vie et d_N/d_S. De plus, on observe que N_e^s varie très fortement entre espèces, sur plus de trois ordres de grandeur (x1250) contre seulement un ordre de grandeur (x12) pour N_e^l. La variance de ces deux mesures nous montre que seulement 15% du N_e^l permet d'expliquer le N_e^s. Cette très forte dispersion du N_e^s pourrait être en partie due à une mauvaise estimation de π_S et notamment des cibles mutationelles synonymes et non synonymes comme mentionné précédemment. On ne peut donc pas exclure, encore une fois, que l'absence de corrélations pourrait être due à l'utilisation de données empiriques imparfaites plutôt qu'une réalité biologique. Nous sommes aujourd'hui plutôt convaincus que l'utilisation d'un modèle évolutif pour estimer ces cibles ainsi qu'un meilleur masquage des génomes pourraient être une clé dans cette étude.

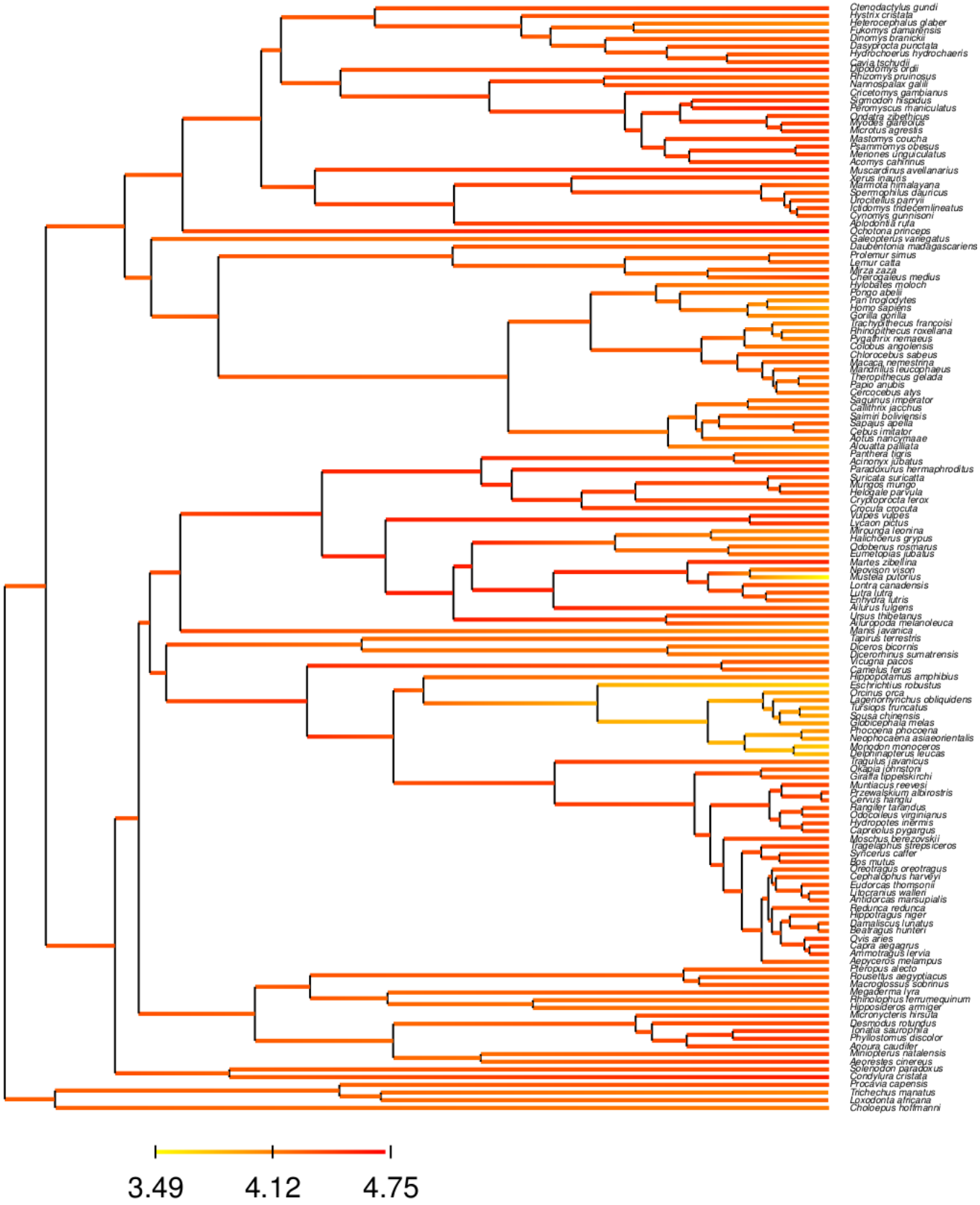


Figure 7.2 – *Reconstruction du log(N_e) de long terme le long de la phylogénie des mammifères par FastCoevoNe. On retrouve des groupes à plus faible N_e déjà observés dans le chapitre 5 comme les primates et les cétacés.*

7.1.2 Incorporer les paysages de fitness dans les modèles à codons

Comme mentionné ci-dessus, le modèle mécaniste de Brevet and Lartillot (2021) fait l'hypothèse que d_N/d_S et π_N/π_S sont impliqués dans une relation allométrique (log-linéaire) avec N_e. Les auteurs justifient l'usage de cette relation par le travail de Welch *et al.* (2008) qui montre que, si la DFE est gamma avec un paramètre de forme β, alors $d_N/d_S \propto N_e^{-\beta}$ et $\pi_N/\pi_S \propto N_e^{-\beta}$. Toutefois, utiliser cette relation dans un cadre phylogénétique revient à faire l'hypothèse que la DFE est constante à travers les clades. Or, des travaux comme ceux de Castellano *et al.* (2019) ou Goldstein (2013) ont montré que des changements de N_e peuvent changer la DFE. Ainsi, supposer une DFE constante dans le long terme n'est pas tenable.

Dans son article, Castellano *et al.* (2019) écrit notamment que « l'effet des mutations est dynamique, même entre espèces proches, et dépend du contexte génétique dans lequel elles arrivent ». Ce contexte génétique correspond au paysage de fitness, décrit pour la première fois par Wright *et al.* (1932). C'est un concept qui permet de faire un lien entre les génotypes et leur fitness. Plus concrètement, pour chaque site d'une séquence, on décrit un vecteur de fitness de taille 20 correspondant à la valeur de fitness des 20 acides aminés possibles à ce site (figure 7.3). On peut alors, pour chaque espèce, quantifier pour chaque site et pour chaque mutation, quel en sera l'impact sur la fitness de l'individu et donc estimer une DFE propre à cette espèce. Modéliser à partir des paysages de fitness permet d'intégrer plus directement les concepts et mécanismes qui définissent la DFE. Ainsi, au lieu de supposer une DFE constante, comme dans le modèle de Brevet and Lartillot (2021), on peut considérer que le paysage de fitness est constant, et qu'il influence une DFE qui, elle, peut varier au cours de l'évolution des espèces.

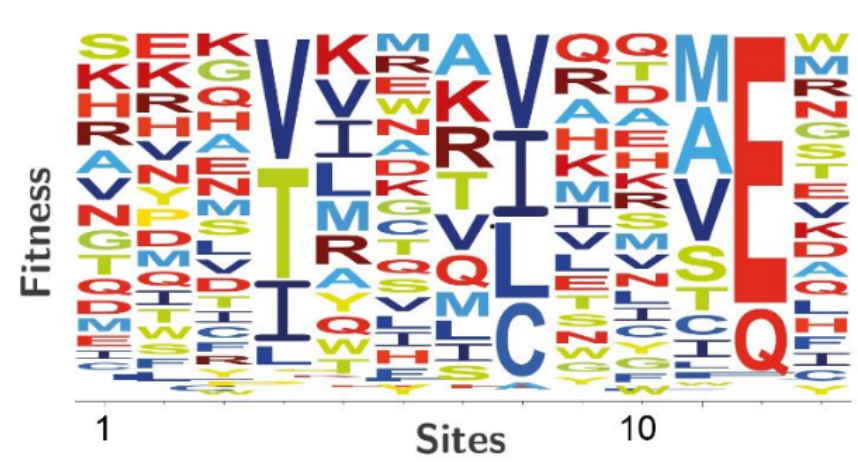


Figure 7.3 – *Exemple de paysage de fitness. A chaque site, en abscisse, on associe un vecteur de fitness contenant les 20 acides aminés représentés ici par des lettres de taille variables rendant compte de la fitness de l'acide aminé. Pour certains sites, plusieurs acides aminés sont également optimaux (premier site) tandis que pour d'autres, seulement un à deux acides aminés sont envisageables (douzième position). Figure adaptée de Latrille (communication orale).*

Il existe une grande quantité de modèles permettant d'estimer un paysage de fitness (Blanquart *et al.*, 2014). Plus particulièrement, les modèles de type Mutation-Sélection (ou MutSel) (Yang and Nielsen, 2008; Rodrigue *et al.*, 2010) sont capables d'allier les paysages de fitness à des modèles à codons (sous-section 3.2.2). Comme pour les modèles à codon classique, les mutations synonymes sont considérées comme neutres et uniquement soumises à dérive génétique, tandis que les mutations non-synonymes sont soumises à dérive et sélection naturelle. De fait, chaque mutation non-synonyme, à un site spécifique, a un effet différent sur la fitness (Halpern and Bruno, 1998), dont on peut spécifier le coefficient de sélection. On détermine donc le taux de substitution d'un codon i à j par :

$$Q_{ij} = \mu_{ij}.2N.P_{fix}(s_{ij}) \tag{7.1}$$

avec $P_{fix}(s_{ij})$ la probabilité de fixation du codon sachant le coefficient de sélection s_{ij}, lui-même dépendant de la fitness (notée f), de chacun des deux codons tel que :

$$s_{ij} = f_j - f_i \tag{7.2}$$

Pour distinguer les changements synonymes et non-synonymes, on utilise une probabilité de fixation différente menant à (Halpern and Bruno, 1998; Yang and

Nielsen, 2008) :

$$\begin{cases} Q_{ij} = \mu_{ij} & \text{si i et j sont synonymes} \\ Q_{ij} = \mu_{ij} \dfrac{4Ne(f_j - f_i)}{1 - e^{-4Ne(f_j - f_i}} & \text{si i et j sont non synonymes} \end{cases} \tag{7.3}$$

Cependant, dans leur version initiale, les modèles MutSel font l'hypothèse que N_e est constant le long de la phylogénie. Or sait aujourd'hui que c'est une hypothèse à relaxer. Des chercheurs ont ainsi développé le modèle MutSelNe (Latrille *et al.*, 2021) dans lequel la taille efficace devient un paramètre qui peut varier le long de la phylogénie et qui correspond plutôt à un N_e de long terme (figure 7.4). Ce modèle permet donc de reconstruire explicitement N_e (à un facteur d'échelle près), le long de la phylogénie et ainsi d'intégrer explicitement la relation entre d_N/d_S et N_e, en fonction du paysage de fitness.

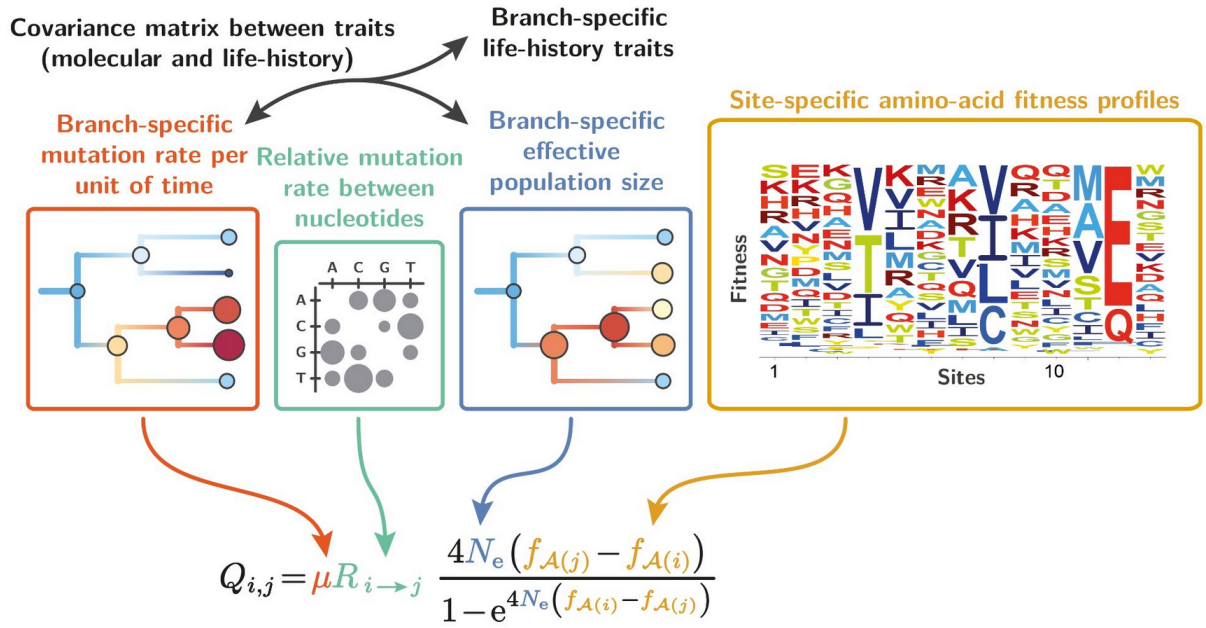


Figure 7.4 – *Le modèle MutSelNe. Le modèle prend en entrée des traits d'histoire de vie, un alignement de gènes et une phylogénie. Il donne en sortie une reconstruction de la variation des traits d'histoire de vie, de N_e et du taux de mutation le long de l'arbre, une matrice de covariance entre traits moléculaires et écologiques, un paysage de fitness et la matrice de taux de mutation. Figure adaptée de Latrille* et al. *(2021).*

Dans leur analyse Latrille *et al.* (2021) obtiennent une corrélation négative significative entre N_e, tel que reconstruit par ce modèle, et les traits d'histoire de vie, ce qui est très encourageant. Toutefois, la magnitude des variations de N_e semble sous-estimée. Les auteurs mettent en cause la non prise en compte des interactions entre sites (l'épistasie) qui pourraient diminuer la réponse du processus de substitution aux variations de N_e.

Il pourrait être intéressant de confronter les estimations des variations du N_e^l reconstruit par FastCoevolNe à celle du N_e issu de MutSelNe, tant ces deux modèles fonctionnent différemment, pour reconstruire une même entité. Cependant, il faudrait avant cela en développer une version « fast » pour l'utiliser avec mon jeu de données.

7.1.3 Reconstruire les variations de Ne dans le temps

Lors de ma thèse, j'ai principalement utilisé des estimations discrètes de N_e et les ai classifiées en court et long terme. Pourtant, comme déjà exploré dans la section 2.5, il existe toute une panoplie de N_e chacun valides à différentes échelles de temps qui ne se réduisent pas à la simple dichotomie micro/macro-évolutif. De plus, dans le chapitre 5 et chapitre 6, nous avons pu constater respectivement que N_e varie le long des branches et dans l'histoire des populations. À mon sens, la prochaine étape cruciale dans les réflexions autour de N_e sera d'apprendre à intégrer ces variations ensembles dans un continuum plutôt que de les confronter. Par exemple, cela pourrait passer par l'intégration d'un module de PSMC (Li and Durbin, 2011) dans FastCoevolNe.

7.1.4 Etudier N_e à une autre échelle taxonomique ou régime sélectif

L'ensemble des résultats présentés dans ce manuscrit correspondent à l'étude des séquences de génomes de mammifères. Cette échelle taxonomique semble idéale, d'une part parce qu'elle est assez large pour contenir suffisamment de signal phylogénétique et d'autre part suffisamment restreinte pour ne pas être trop saturée. Cependant, il existe d'autres groupes présentant ces caractéristiques comme les

lépidoptères ou les oiseaux et pour lesquels nous avons du matériel génomique avec un échantillonnage taxonomique adéquat (Janzen *et al.*, 2009; Feng *et al.*, 2020). De plus, les mammifères répondent à quelques caractéristiques génomiques particulières comme une forte gBGC (Duret and Galtier, 2009), un faible N_e (Leffler *et al.*, 2012) et des mutations synonymes plutôt neutres (Tomoko, 1995) (mais voir Chamary *et al.* (2006)). Or, d'autres groupes comme les drosophiles ou les lépidoptères ne présentant pas forcément l'ensemble de ces traits simultanément. Ainsi, tester empiriquement la théorie quasi neutre sur d'autres groupes taxonomiques serait pertinent.

7.2 Perspectives pour le jeu de données

Une des contributions majeure de ma thèse au domaine de la génomique est selon moi la constitution de ce jeu de données de plus de 6000 gènes orthologues alignés, couvrant plus de 140 espèces de mammifères et où le polymorphisme et les processus de substitution sont estimés sur les mêmes gènes et individus. Le jeu de données contient très peu de données manquantes, et un soin particulier concernant sa qualité a été appliqué. Bien qu'il existe déjà de tels jeux de données concernant les mammifères, comme ceux du Zoonomia consortium (Zoonomia, 2020) ou Orthomam (Scornavacca *et al.*, 2019), aucun ne contient soit autant d'espèce de mammifères, autant de gènes ou de données de polymorphisme. Dans les prochains paragraphes, je vais rappeler brièvement les étapes clés de l'écriture de ce jeu de données, puis je vais présenter quelques perspectives d'exploitation.

Initialement, nous avons récupéré via NCBI (en juillet 2021) les génomes de 197 mammifères placentaires avec un N50 supérieur à 30kb et une couverture supérieure à 20X. Nous avons évité les espèces domestiques ou de laboratoire. Notre collaborateur, David Enard, a réalisé le variant calling sur chacun des génomes avec l'outil samtools-bcftools (Li *et al.*, 2009; Danecek *et al.*, 2021). J'ai annoté les gènes orthologues communs aux mammifères et simple copie de ces 197 génomes à l'aide de l'outil BUSCO (Simão *et al.*, 2015) présenté en section 3.1. Je n'ai sélectionné que les gènes présents chez plus de 70% des espèces, puis les espèces contenant plus de 80% des gènes me menant à 183 espèces et 8060 gènes. Les gènes ont été alignés à l'aide de l'outil PRANK (Löytynoja, 2014) puis filtrés successivement par

HMMCleaner (Di Franco *et al.*, 2019), BMGE (Criscuolo and Gribaldo, 2010), Phylter (Comte *et al.*, 2023) et un modèle bayésien expliqué dans les annexes du chapitre 5. J'ai ensuite éliminé les gènes trop gappés où qui contenait à présent moins de 80 % des espèces, menant à un jeu de données de 7354 gènes et 144 espèces. Dans un deuxième temps, je me suis rendu compte qu'il était nécessaire de masquer certaines positions des génomes en fonction de leur couverture afin de ne plus travailler qu'avec les positions qui me semblent de meilleure qualité et que je considèrerais par la suite comme « appelables ». J'ai appliqué ce masquage aux alignements puis filtré à nouveau les données manquantes afin d'obtenir le jeu de données final de 6002 gènes. En parallèle de cela, j'ai récupéré les positions des différents gènes dans les génomes pour annoter mes fichiers de variant calling et ne récupérer que le polymorphisme qui concerne les gènes orthologues identifiés. J'ai appliqué différents filtres plutôt stringeants sur les VCF afin de minimiser la présence de SNPs faux positifs. C'est par l'usage de l'ensemble de ces filtres aux différentes étapes et l'attention portées aux données manquantes que je pense pouvoir affirmer avec confiance de la qualité de ce jeu de données. Néanmoins, toutes choses étant perfectibles, quelques améliorations pourraient être proposées comme l'utilisation d'un outil de variant calling faisant plus consensus dans la communauté ou un meilleur masquage des génomes concernant, par exemple, les éléments transposables ou les régions consanguines.

7.2.1 6000 gènes orthologues de mammifères

Annoter les gènes orthologues : BUSCO versus TOGA

Comme présenté en introduction (sous-section 3.1.2), il existe un outil plutôt prometteur, nommé TOGA, qui permet d'annoter des gènes orthologue à partir d'alignement de génomes (Kirilenko *et al.*, 2023). Les auteurs et autrices de cet outil ont notamment montré que TOGA améliore la détection des gènes orthologues, l'annotation des gènes conservés, réunit des régions fragmentées et donne une meilleure estimation de la qualité des génomes que celle de BUSCO (Simão *et al.*, 2015). En effet, Kirilenko *et al.* (2023) ont compilé un set de 18000 gènes qu'iels décrivent comme ancestraux aux mammifères placentaires, puis ces gènes sont recherchés dans les génomes testés. On obtient ainsi une mesure de complétion

similaire à celle de BUSCO. Cependant, il semblerait que la mesure fournie par BUSCO rencontre des problèmes de saturation autour de 97% pour les génomes très complets, tandis que la mesure de TOGA montrerait une plus large dynamique et donc un score de qualité avec une meilleure résolution. Il semblerait également que BUSCO ai tendance à surestimer la complétion des génomes, menant à des scores par exemple de 96% quand TOGA ne montre que 90% de complétion (Kirilenko *et al.*, 2023).

Cet outil ayant été développé après mon annotation des génomes par BUSCO, j'ai choisi de ne pas l'utiliser lors de mon projet de recherche. Cependant, les auteurs de TOGA ont réalisé une annotation du jeu de données Zoonomia (2020), proche de la composition en espèces de mon jeu de données, que je pourrais utiliser comme outil de comparaison pour mettre à l'épreuve mon annotation de gènes orthologues [2]. Je pourrais également utiliser directement l'outil pour annoter les génomes de mon étude, d'autant plus qu'il présente une façon différente de détecter les paralogues et orthologues inactifs ce qui pourrait me permettre d'encore améliorer mon jeu de données. Cependant, TOGA s'appuie sur des programmes d'intelligence artificielle très couteux énergétiquement et je pense que le jeu de données tel que je le propose a déjà assez participé au réchauffement climatique (voir sous-section 7.3.3).

A quoi correspondent ces 6000 gènes ?

Finalement, dans ma thèse, l'identité des espèces ou des gènes m'importe assez peu tant ce sont les relations entre ces entités qui m'intéresse. Ainsi, je n'ai pas vraiment idée de ce à quoi correspondent ces 6000 gènes communs à un ensemble de mammifères. Est-ce que ce sont des gènes liés au développement ? Au métabolisme cellulaire ? Sont-ils spécifiques aux mammifères ? Puisque ce sont des gènes très conservés depuis des millions d'années, il semble raisonnable d'imaginer que ce sont des gènes particulièrement importants pour ces organismes. En tout cas, des analyses de gene ontology sur ce jeu de donnée pourraient être très intéressantes.

Aussi, il existe des outils comme Pelican (Duchemin, 2023) qui sont capables

2. J'ai d'abord été très frustrée quand, après 2 ans à développer mon jeu de données, cet outil a été publié et appliqué sur des données très similaires aux miennes. Puis, après la lecture de l'article de Kirilenko *et al.* (2023), j'ai surtout été envieuse de ne pas avoir participé à son développement.

d'associer des sites protéiques à des phénotypes. Je trouve très tentant de tester ce genre d'outil sur mon jeu de données avec des phénotypes comme le régime alimentaire, le fait d'être plutôt terrestre ou aquatique, la longévité, l'écholocation ou bien l'appartenance à un clade particulier.

7.2.2 Continuer de questionner la théorie neutre

Bien sûr, le jeu de données peut permettre d'investiguer encore quelques questions concernant la théorie neutre chez les mammifères. Par exemple, nous savons que les génomes de mammifères sont soumis à un mécanisme non-adaptatif qui est la conversion génique biaisée vers GC (gBGC) (Duret and Galtier, 2009). Ce mécanisme a pour effet d'augmenter le taux de G et C dans les séquences à fort taux de recombinaison. Ainsi, les gènes avec un fort taux de GC dans leurs positions synonymes sont possiblement des gènes expérimentant une plus forte gBGC que les autres. Il serait donc possible de tester, avec FastCoevol, l'effet de la gBGC sur le lien entre efficacité de la sélection et N_e via ces différents compartiments de gènes.

Aussi, pour l'instant, j'ai toujours fait l'hypothèse que la sélection positive était négligeable dans mes analyses. Hypothèse qui me semble réaliste au vu du niveau de conservation des gènes que j'utilise. Cependant, si finalement, il y a une quantité non négligeable de positions sous sélection positive dans mes alignements, celle-ci pourrait impacter le d_N/d_S en l'augmentant et impacter mon étude confrontant l'efficacité de la sélection et N_e. Il serait tout à fait possible d'utiliser des modèles à codon pour détecter les sites sous sélection positive et potentiellement les masquer.

7.2.3 Poursuivre la jonction micro/macro évolution en dehors de la théorie neutre

Un des avantages à disposer de données correspondant à deux échelles temporelles différentes est de pouvoir tirer parti de leur sensibilité propre à divers processus évolutifs. Lors de mon travail de thèse, j'ai particulièrement travaillé à réaliser une jonction entre micro et macro-évolution et questionner les effets de la théorie quasi neutre à chaque échelle, mais en utilisant un jeu de données et un outil

d'analyse qui les réunissent. Cette jonction avait déjà été entamée avec le développement de CoevolNe (Brevet and Lartillot, 2021) et elle existe aussi en dehors du cadre de la théorie quasi-neutre comme dans les développements de modèles de type McDonald Kreitman (voir section 3.4). Pour moi, cette idée de confronter les compartiments micro et macro-évolutifs ouvre de nouvelles portes aux travaux concernant l'étude de l'évolution des génomes, tant au niveau théorique qu'empirique. De plus, cette confrontation peut se faire à propos des processus adaptatifs comme non-adaptatif.

Un processus non-adaptatif particulièrement d'intérêt chez les mammifères est la gBGC, comme évoqué plus haut. Par défaut, il existe un biais mutationel vers AT qui conduit à une augmentation de la proportion de nucléotides A et T dans le génome. A l'inverse, la gBGC est un mécanisme qui induit un biais de fixation conduisant à une augmentation de la proportion de nucléotides G et C dans les régions à fort taux de recombinaison du génome. Ainsi, la mutation engendre, de manière biaisée, des variants plutôt de type AT tandis que la gBGC favorise l'augmentation en fréquence des variants GC. De ce fait, le biais mutationel observé est sous-estimé en présence de recombinaison et donc de gBGC, et ce, d'autant plus quand on l'observe dans la divergence que dans le polymorphisme, quand le processus de gBGC est encore en cours. On prévoit donc une différence entre le biais de mutation et le biais de substitution. Pourtant, dans les modèles précédemment abordés comme MutSel (Rodrigue *et al.*, 2010), on fait l'hypothèse que le biais de substitution est une estimation directe du biais de mutation. En combinant des données de polymorphisme et de divergence, on peut tirer parti du fait que ces deux compartiments réagissent différemment aux biais de mutation et de fixation. On peut alors démêler les effets de ces deux quantités, que l'on peut estimer simultanément, et vérifier cette hypothèse. De plus, il se pourrait également que le biais mutationel ne soit pas lui-même constant le long de la phylogénie, mais plutôt lié positivement au temps de génération. Avec un développement particulier de FastCoevol, nous pourrions faire évoluer des taux de mutation et de substitution vers AT, le long de la phylogénie, pour ensuite les confronter.

Concernant les processus adaptatifs comme la sélection positive, nous savons que les mutations avantageuses sont rapidement fixées dans les génomes, les rendant

invisibles, ou en tout cas peu présentes, dans le polymorphisme. Ainsi, détecter un excès de fixation de divergence non-synonyme par rapport au polymorphisme non-synonyme est un indice pour détecter de la sélection positive. C'est ce que testent les modèles de type McDonald Kreitman (voir section 3.4). Cependant, ces modèles nécessitent l'utilisation de données de polymorphisme collectées sur un nombre suffisant d'individus par espèces, et ce, afin d'estimer la DFE. Des développements pour adapter ces modèles à mon jeu de données, qui n'utilise qu'un individu par espèces, pourraient être envisagés et permettraient de quantifier efficacement la présence de sélection positive dans celui-ci. Ces développements nécessiteraient toutefois de faire des hypothèses sur la manière dont la DFE varie entre espèces. Dans cette direction, il pourrait être pertinent de développer des modèles type McDonald-Kreitman qui prennent directement en compte le paysage de fitness, comme dans les modèles MutSel, afin de relaxer les hypothèses concernant la DFE.

7.3 Complexités et enjeux dans le domaine de la bio-informatique

Dans cette section, je souhaite élargir un peu plus la discussion autour des domaines de la bio-informatique et de la biologie évolutive. J'ai l'impression qu'il y a des sujets importants à aborder, comme l'usage de données et d'outils imparfaits pour les analyses, la tentation du *p-hacking* et l'impact environnemental de ce genre de travaux. Chacun de ces trois sujets, et bien d'autres, ont eu, à un moment ou un autre, un fort impact sur ma façon de mener ma recherche, engendrant parfois de fortes prises de conscience et des remises en question, et ont mis à l'épreuve ma résilience et ma patience. C'est pourquoi je souhaite les développer ici, car ils font partie intégrante de mon expérience de thèse.

7.3.1 L'enfer des données (encore...)

Dans la section 3.1 de l'introduction, je détaille la suite des processus qui permettent d'obtenir, à partir de données de séquençage, un alignement de gènes et des données de polymorphismes associées. Je répertorie également quelques erreurs

qui peuvent se glisser à chaque étape et leur potentiel impact sur les analyses menées en bout de processus. À plusieurs reprises, lors de ce travail de thèse, nous nous sommes retrouvés à observer des résultats qui nous ont semblé aberrants, sans comprendre s'ils étaient porteurs d'une vraie affirmation biologique ou la résultante d'une accumulation d'erreurs dans les données. À mon sens, il existe plusieurs sources d'erreurs qui proviennent de l'usage des outils bio-informatique. En effet, il est parfois difficile de comprendre comment les utiliser, dans quel contexte et comment interpréter leur sortie. Par exemple, lors du filtrage des alignements mentionnés dans le chapitre 5, j'ai mis plusieurs jours, voire plusieurs semaines, à comprendre certains détails. Tout d'abord, les alignements réalisés avec l'outil PRANK (Löytynoja, 2014) devaient être traduits en acides aminés pour pouvoir être correctement filtrés par HMMCleaner (Di Franco *et al.*, 2019) qui ne sait pas prendre en compte le format codon des séquences nucléiques. Ensuite, ces alignements d'acides aminés devaient être retraduits en nucléotides pour BMGE (Criscuolo and Gribaldo, 2010) qui est capable de filtrer en prenant en compte les codons. Tout cela a nécessité l'usage d'outils intermédiaires pour garantir une traduction en arrière qui prenne en compte le fait qu'un acide aminé puisse correspondre à plusieurs codons. Également, lors du filtrage des VCF, je me suis rendu compte qu'il existe une panoplie d'indices de qualité (QUAL, GQ, PL, GT, AD, DP, etc) qu'il faut d'abord savoir décrypter pour ensuite pouvoir les recruter comme indice d'intérêt pour le filtrage des données, et ce, dans un contexte où les fichiers sont très lourds, ce qui entrave leur exploration. De plus, tous les variant-caller n'utilisent pas les mêmes indices de qualité et quand bien même certains indices sont en communs, ceux-ci ont parfois des unités ou des significations différentes. C'est notamment à cause de ces difficultés que j'ai utilisé un filtrage très strict sur mes fichiers VCF.

J'ai l'impression qu'il faudrait être un.e expert.e de chaque outil pour pouvoir les utiliser en connaissant leurs hypothèses et leurs défauts, ce qui me semble difficilement envisageable quand on sait qu'un projet comme le mien implique des dizaines d'outils de ce type. Néanmoins, j'admets que l'expertise peut venir avec l'usage, pour peu que le paysage des outils les plus efficaces n'évolue pas trop vite. Personnellement, j'ai pris l'habitude de vérifier, après chaque étape du processus de traitement des données, si l'outil avait correctement fait son travail (pas de décalage du cadre de lecture, pas de fichier tronqué, les éléments à filtrer sont filtrés, etc). Ces contrôles, bien que

nécessaires, sont rapidement chronophages. Cela a fini par engendrer chez moi une certaine paranoïa, si bien que, lorsque c'était possible, je préférais coder moi-même les étapes de traitement des données pour m'assurer de leur fiabilité. Cependant, que se passe-t-il pour les non bio-informaticien.ne.s lorsqu'ils ou elles utilisent ponctuellement ces outils ? Ou bien lorsqu'une étude doit être réalisée rapidement, sous pression des dates de fin de projet ou fin de financement (voir section 7.4), empêchant de vérifier le bon fonctionnement de ces outils ? Quoi qu'il en soit, ces mésaventures avec mes données m'ont poussée à développer une certaine vigilance quant à la lecture des articles publiés et m'ont permis de tirer une des plus grandes leçons liées à ma thèse : toujours vérifier que les résultats obtenus ne peuvent pas être expliqués par des erreurs dans les données avant d'en tirer une conclusion biologique.

Aussi, au-delà du mauvais usage des outils bio-informatiques, ceux-ci présentent des limites algorithmiques ou théoriques qui engendrent des erreurs dans les données et donc diminuent leur qualité. Je fais ici référence aux outils ou aux modèles qui utilisent des hypothèses erronées ou trop simplistes, créant ainsi des artefacts dans les données. Par exemple, les outils d'alignements comme CLUSTAL (Thompson *et al.*, 2003) prennent mal en compte les contraintes sélectives agissant sur les insertions et délétions dans les séquences codantes et peuvent engendrer des décalages de cadre de lecture (sous-section 3.1.3). Dans cet exemple, on peut contourner ces erreurs, plutôt simplement, en les filtrants (si on sait les reconnaitre). On peut également aligner les séquences traduites puis revenir aux nucléotides ou bien utiliser des outils plus adaptés (comme MACSE (Ranwez *et al.*, 2011)). Dans des cas plus complexes, il n'est pas toujours possible de filtrer ou de corriger les erreurs a *posteriori*. Il est donc crucial de choisir, dès le départ, les outils et les modèles les plus appropriés en fonction des données et des questions de recherche. Or, dans la pratique, le compromis entre précision et faisabilité (temps de calcul, accessibilité) mène souvent à utiliser des outils trop simplistes.

Par ailleurs, les erreurs liées à ces outils peuvent être augmentées par une mauvaise qualité initiale des données, qui a elle-même pour source un mauvais équilibrage entre qualité et quantité d'informations (voir Box.4). En effet, les outils de séquençage peuvent inclure des erreurs dans les données (des erreurs de la polymérase, un mauvais assemblage, etc). Cependant, ces erreurs sont censées être, en partie, compensées par

un nombre suffisant de cycles PCR (ce qui augmente la couverture) et une quantité et une qualité suffisante d'ADN fourni. L'optimisation de ces différents paramètres étant couteuse, on fait souvent le choix de séquencer plusieurs génomes avec une qualité moyenne plutôt qu'un seul génome de haute qualité. C'est ici que commence la propagation des erreurs par effet boule de neige.

À mon sens, l'ensemble de ces mésusages des outils bio-informatiques ainsi que les erreurs qui leur sont propres et l'utilisation de données de mauvaise qualité, peuvent expliquer qu'une cohorte d'études, comme celles sur les relations entre N_e et efficacité de la sélection chez les mammifères, présentent des résultats si contrastés. En effet, nous avons montré, dans le chapitre 5, que l'usage d'une méthode séquentiel (PGLS) ou intégrative (Coevol) ne change pas la teneur des résultats, mais qu'en revanche, la qualité des données peut être impactante (voir les corrélations avec et sans les six espèces aux mesures de polymorphisme suspectes).

J'argumente ainsi qu'un levier qui me semble pertinent pour essayer de réduire cet « enfer des données » serait de développer des analyses comparatives sur chacune des familles d'outils et en répertorier leurs qualités et leurs défauts pour que leur usage puisse se faire en conscience. Cela soulève cependant le problème de l'absence d'une référence fiable à laquelle se comparer lors du test des outils. De plus, rien ne peut se faire tant que la qualité initiale des données reste si hétérogène. J'espère qu'une fois ce boom du séquençage passé, la communauté prendra un virage vers la génération de données de meilleure qualité.

7.3.2 Attention au *p-hacking*

Quoi qu'il en soit, une fois les données générées, on les analyse dans l'objectif de tester certaines hypothèses prédéfinies. Pour savoir si un résultat confirme ou infirme une hypothèse, nous utilisons des tests statistiques d'hypothèses qui permettent de démêler un effet d'intérêt d'un bruit de fond (Fisher, 1928). On définit notamment H0 comme l'hypothèse nulle selon laquelle il n'y a pas de relation entre les variables étudiées ou pas d'effet d'un traitement. Puis, on calcule une probabilité p de trouver un effet dans les données, au moins supérieur à celui attendu sous H0. Un usage très répandu dans la communauté scientifique est de décider que l'hypothèse H0 est

rejetée, donc qu'il y a bien une relation entre les variables, quand p est supérieur à 0,05 (Nickerson, 2000; Ridley *et al.*, 2007). Cependant, cette façon de procéder peut mener à quelques déviances comme le « *p-hacking* » qui consiste à sélectionner uniquement les données ou les analyses qui fournissent des valeurs p inférieures à 0,05 (Gadbury and Allison, 2012). Cette pratique très courante parmi les chercheurs et chercheuses, et parfois inconsciente, est encouragée par les journaux scientifiques qui publient préférentiellement les articles présentant des résultats significatifs (Head *et al.*, 2015) (voir section 7.4). Elle peut mener à des effets impressionnants comme lorsqu'un groupe de travail a proposé à tout un ensemble de chercheurs et chercheuses en écologie de réaliser une même analyse sur un même jeu de données. Les résultats retournés étaient alors tous aussi différents les uns des autres, avec certains, présentant des corrélations positives entre les traits étudiés tandis que d'autres montraient des corrélations négatives pour la même question (Gould *et al.*, 2023).

Lors de ma thèse, je me suis régulièrement demandée si les choix que je faisais dans le traitement de mes données relevaient du *p-hacking* ou non. En effet, le *p-hacking* se manifeste par trois facteurs : l'utilisation d'un grand nombre de variables réponses pour ensuite ne sélectionner que les variables produisant des relations satisfaisantes, le filtrage des données déviantes et l'arrêt de l'exploration des données seulement une fois que les résultats sont significatifs (Head *et al.*, 2015). Or, dans mon premier article de thèse, lors de l'écriture de mon jeu de données, j'ai à plusieurs reprises retiré des espèces ou des gènes dont la qualité me semblait douteuse. J'ai également fait des allers-retours entre des analyses FastCoevol et le filtrage des données quand j'obtenais des résultats inattendus. Bien que, après coup, j'ai toujours trouvé des raisons pertinentes d'ajouter un type de filtrage pour augmenter la qualité de mes données en fonction d'un point soulevé par mon analyse, je ne peux m'empêcher de me demander quelle est la part de *p-hacking* dans tout cela. Néanmoins, parce que nous avions conscience de ce biais potentiel, nous avons décidé de répertorier, dans l'article du chapitre 5, tous les filtrages réalisés avec leurs justifications, ainsi que différents lots de résultats mis sur le même pied d'égalité alors qu'ils représentent des niveaux de filtrage différent. Par exemple, certaines espèces nous ont semblé particulièrement déviantes concernant leur π_S. Nous avons alors réalisé une analyse avec et sans ces espèces et nous avons présenté les deux résultats car nous n'avons pas trouvé de raison pertinente de les supprimer. De plus, lorsque nous avons voulu réaliser une analyse FastCoevol avec

moins d'espèces, nous avons également décidé, avant l'analyse, de montrer les deux résultats (jeu de donnée entier ou restreint) quoiqu'ils puissent nous raconter. Ces précautions rendent l'analyse plus complexe pour nous et plus difficile à lire pour les lecteurs et lectrices, mais elles sont, selon moi, essentielles et me permettent de diffuser mes résultats avec plus de sérénité au regard de cette problématique du *p-hacking*.

7.3.3 Empreinte carbone de ce genre de travaux

Au LBBE, nous disposons d'un cluster de calcul mettant à disposition plus de 1400 CPU et 10 TB de mémoire vive. Cela a été un net avantage pour mes travaux de thèse. Une étude interne au laboratoire menée en 2018 [3] a montré que ce cluster représente 17,5 % des couts énergétiques du laboratoire avec 39 tonnes d'équivalent CO2 dépensées par an. C'est le deuxième poste le plus énergivore après les transports en avion (45 % des dépenses).

Lors de ma thèse, j'ai moi-même largement usé du cluster de calcul et j'ai souhaité questionner la dépense en carbone que cela a engendré. Pour ce faire, nous disposons au laboratoire d'une page web qui recense, pour chaque mois de l'année, le nombre d'heures de calculs utilisant un CPU (hCPU) réalisés sur le cluster, par utilisateur.ice. Cette page met aussi en relation le nombre d'hCPU de chacun et chacune avec le total d'heures consommées par l'ensemble des membres du laboratoire. Dans la figure 7.5, on peut y voir l'évolution de mon utilisation du cluster. Sans surprise, entre janvier 2021 et août 2024, j'ai figuré 21 fois parmi les 10 utilisateur.ice.s les plus important.e.s du cluster, qui compte environ 70 à 80 usagers différent.e.s chaque mois. En tout, pour développer mon jeu de données et l'analyser, j'ai consommé 1 837 367 heures CPU, soit 209 années de calcul sur un ordinateur portable avec 1 CPU. Cela équivaut à 6,6 tonnes d'équivalent CO2 (Berthoud *et al.*, 2020) rejetées.

3. https://ferme.yeswiki.net/Empreinte/?PagePrincipale

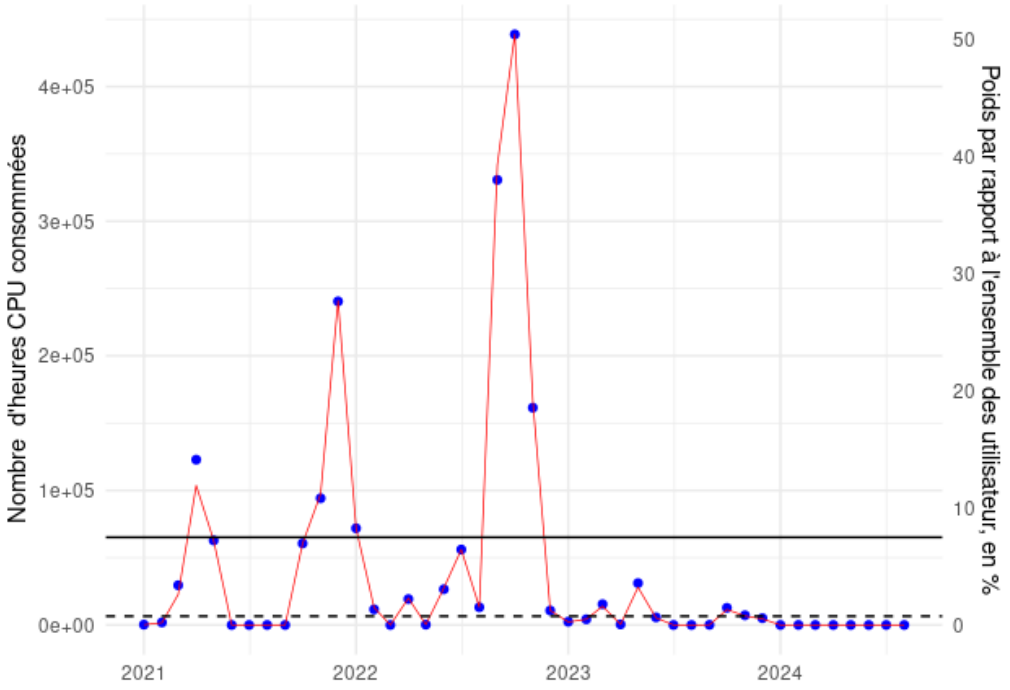


Figure 7.5 – *Nombre d'heures CPU consommé par mois pour ce projet, depuis janvier 2021. Les points bleus et l'axe à gauche représentent le nombre d'heures CPU. La ligne rouge et l'axe à droite représentent mon taux de consommation par rapport à l'ensemble des utilisateurs du laboratoire (en moyenne 75 personnes). La barre horizontale noire pointillée montre la consommation d'une personne si la répartition des heures CPU était équilibrée, la barre horizontale noire représente la consommation de 10 personnes. De mémoire, le premier pic de consommation correspond à l'analyse BUSCO faite en stage de M2 sur les données de Zoonomia, le deuxième pic, l'analyse BUSCO sur le nouveau jeu de données et je suppose que le troisième pic correspond aux alignements de gènes ainsi qu'à l'analyse Coevol et FastCoevol.*

J'ai essayé de comprendre à quoi correspond réellement ces 6 tonnes de CO2 en estimant, via différentes sources, mon impact carbone annuel, en dehors de la thèse. Celui-ci a été estimé entre 5,7 et 7 tonnes par an[4]. Ainsi, mon travail de thèse correspond à environ une année en plus de ma vie en termes d'impact carbone. Selon des outils de comparaison du Monde[5], il faudrait 9 ans de véganisme ou 4,5 ans d'abandon total de la voiture pour compenser l'impact carbone de cette thèse.

Je dois en plus souligner que le cluster du laboratoire a été ma principale source computationnelle mais pas l'unique. Parfois, quand les ressources n'étaient pas

4. source : https://nosgestesclimat.fr/empreinte-climat et https://www.footprintcalculator.org/home/fr

5. https://www.lemonde.fr/les-decodeurs/article/2023/12/05/vous-voulez-compenser-votre-vol-en-avion-par-des-ecogestes-voici-combien-de-temps-cela-vous-prendra_6204046_4355770.html

disponibles, j'ai bénéficié de machines virtuelles de l'institut français de bio-informatique et d'un autre cluster de calcul de l'université. Aussi, l'entièreté de l'analyse du deuxième article a été réalisée sur mon ordinateur portable (d'où le faux plat fin 2023-2024 dans la figure 7.5). Mon nombre d'heures CPU réalisées lors de ma thèse est donc certainement sous-estimé.

Avant cette réalisation de l'impact de mes analyses il y a environ un an et demi, je ne réfléchissais pas forcément à utiliser le cluster de façon raisonné d'un point de vue environnemental. Si une nouvelle étape dans le pipeline me semblait pertinente pour améliorer mon jeu de données, alors je l'implémentais dans la foulée. Il est très facile d'appuyer sur un bouton à distance sans en percevoir les conséquences et il est très tentant de relancer des analyses à cause d'une petite erreur de format plutôt que de reformater soi-même un fichier de résultat. Cependant, certaines étapes couteuses de mon traitement de données n'ont pas pu être évitées, car elles nécessitaient soit des données qui par ailleurs existent mais qui n'ont pas été rendues accessibles, ce qui nous a obligé à les régénérer [6], soit elles nécessitaient un traitement lourd pour compenser leur mauvaise qualité. À ce titre, je regrette aujourd'hui de ne pas avoir été formée à un usage raisonné des ressources numériques (comment optimiser un code, estimer la RAM nécessaire, quels bénéfices pour quels couts, ce que coute une heure CPU, etc). Je pense que c'est un point qui doit être abordé plus souvent au démarrage de ce genre de projet avec, peut-être, un seuil d'impact écologique à fixer puis à ne pas dépasser. Je sais désormais que mes futurs choix de projets de recherche seront soumis à un filtrage écologique et que je m'assurerai de gérer mes ressources avec beaucoup plus d'attention. Par exemple, pour une analyse similaire à celle de ma thèse, je travaillerai en augmentant progressivement la taille de mon ensemble de données. Il serait alors tout à fait possible d'utiliser plusieurs sous-ensembles de données simultanément et de cesser d'ajouter des gènes lorsque la variance entre ces sous-groupes devient suffisamment faible [7].

Enfin, si même quand on réduit au maximum l'impact écologique d'un projet de recherche fondamentale, on aboutit tout de même un cout énergétique élevé, alors

6. C'est le cas des fichiers de couverture des génomes que David Enard a dû régénérer à partir des reads initiaux.

7. Cette réflexion est largement inspirée de mes discussions avec Nicolas, qui soutient également qu'un modèle raisonnable doit pouvoir fonctionner sur un ordinateur portable.

peut-être que nous devrions remettre en question la réalisation de ce projet. Un grand nombre de chercheurs et chercheuses ont pour objectif très noble de décoder le monde, mais je pense qu'aujourd'hui, il est plutôt nécessaire de se demander à quel prix ce décodage est possible. Préférons-nous tout connaitre de notre monde plutôt que le préserver ?

7.4 Recherche académique et diffusion des connaissances

Cette dernière section a pour objectif de proposer une prise de recul, encore plus large cette fois-ci, sur le métier de chercheur et chercheuse en explorant divers aspects qui le composent comme la culture de la recherche, la santé mentale en thèse et la diffusion du savoir. Ce sont des sujets importants pour moi, et j'ai fait le choix conscient de leur accorder une place significative lors de mes trois années de thèse, au détriment d'une plus grande production de résultats scientifiques. Ainsi, si vous cherchez de nouvelles pistes concernant l'étude des séquences moléculaires, vous ne trouverez plus d'informations à partir de ce point. En revanche, si vous souhaitez explorer des sujets qui me passionnent profondément, les prochaines sections vous intéresseront peut-être.

Aussi, j'ai longtemps réfléchi à la manière de rendre cette section inclusive, sans alourdir la lecture avec des formulations comme « le/la chercheur.euse ». En tant que femme (apprentie) chercheuse, il est important pour moi de ne pas reléguer la présence du féminin au second plan lorsque je mentionne les acteurs et actrices du domaine de la recherche. C'est pourquoi, dans les prochains paragraphes, j'ai choisi de m'écarter des règles du français classique et d'adopter une approche où « le féminin l'emporte sur le masculin ». Ainsi, lorsque je mentionne « les chercheuses », je fais référence à toutes les personnes qui pratiquent la recherche, quel que soit leur genre.

7.4.1 « *Research culture* », encadrement en thèse et santé mentale

« Publish or perish » et « PhD mental health crisis »

On le sait, le domaine de la recherche académique est plutôt singulier. Il suffit d'aborder le système de publications pour montrer qu'il y a un problème. En effet, la chercheuse doit d'abord trouver des financements pour sa recherche, puis payer pour accéder à d'autres articles sur le sujet. Après avoir réalisé ses travaux, elle rédige un article pour présenter ses résultats, et paye à nouveau des frais pour pouvoir le publier. Pendant ce temps, les scientifiques qui relisent l'article pour vérifier sa qualité, doivent le faire gratuitement. Tout cela s'inscrit dans un contexte où c'est le nombre d'articles publiés qui fait la renommée de la chercheuse[8]. On entre ainsi dans un système vicieux où il faut constamment trouver plus d'argent et publier toujours davantage. C'est le cœur même de la notion de « *publish or perish* » (publier ou périr), qui entraine une forte culture du surmenage et une hyper-compétition dans le milieu académique (Hall, 2023).

Des sondages rapportent que 73 % des chercheuses[9] témoignent travailler dans un environnement hyper-compétitif où il est impératif de ne montrer aucune faiblesse pour obtenir un poste permanent ou des financements. Certaines études montrent que 71 % des étudiantes en thèse travaillent plus de 41 heures par semaine, tandis que 5 % d'entre elles dépassent les 80 heures (Woolston, 2019).

Pour les doctorantes, cette pression exercée par le milieu académique peut être exacerbée par différents facteurs, dont la relation d'encadrement avec leur directrice de thèse. En effet, l'encadrante détient souvent un pouvoir décisionnel important sur le déroulement de la thèse. De plus, un manque de soutien de la part des personnes censées jouer un rôle de guide lors de la thèse peut s'avérer particulièrement préjudiciable. On observe même une corrélation négative entre la qualité de l'encadrement et le niveau d'anxiété et de dépression chez les étudiantes (Evans *et al.*, 2018). De plus, certains encadrements peuvent être qualifiés d'abusifs (Moss and Mahmoudi, 2021) sans que l'encadrante ne soit réellement inquiétée. Plus

8. Forcément dans ce contexte, des déviances permettant d'accéder à plus de visibilité sont observées (Seeber *et al.*).

9. Ici encore, « chercheuse » fait référence à toutes les personnes qui font de la recherche. Je ne proposerais pas d'analyses propres aux femmes dans ces sections. Idem pour « doctorantes », « étudiantes », « encadrantes », « maitresses de thèse », etc.

généralement, des études montrent que 24 % des doctorantes interrogées présentent des symptômes concordant avec un diagnostic de dépression tandis que 17 % présentent des symptômes d'anxiété clinique (Satinsky *et al.*, 2021). En effet, on estime que les doctorantes ont six fois plus de risque de développer un trouble psychique que la population générale (Evans *et al.*, 2018). Comparées à un groupe dit « très éduqué », les doctorantes se retrouvent face à douze facteurs de risque supplémentaires tels que le sentiment d'être sous contrainte, la difficulté à prendre des décisions et la dévalorisation (Levecque *et al.*, 2017). Ces études soulignent l'existence de ce qu'on appelle aujourd'hui le « *PhD mental health crisis* » [10] (Evans *et al.*, 2018) ainsi que l'urgence de la promotion de programmes et d'interventions concernant la santé mentale en thèse (Evans *et al.*, 2018; Satinsky *et al.*, 2021).

S'adapter, informer et dénoncer

Un des leviers connus pour diminuer l'impact de ce « *PhD mental health crisis* » sur les doctorantes, ainsi que l'effet du milieu académique sur les chercheuses en général, consiste à informer et à préparer les membres du domaine pour qu'elles puissent être le moins impactées possible par les contraintes précédemment évoquées [11]. Ainsi, ce paragraphe présente quelques solutions personnelles que j'ai mises en œuvre durant ma thèse pour atténuer l'impact de cette crise de la santé mentale, ainsi que mes tentatives d'actions pour agir, à mon échelle, sur ces questions.

À mon arrivée en thèse, j'ai assez vite senti qu'il était plutôt valorisé, chez les doctorantes, de réaliser de gros volumes horaires, travailler le week-end et prendre peu de vacances [12]. Dans l'imaginaire collectif, une thèse est souvent considérée comme devant être éprouvante pour être qualitative. Pour me protéger de cette pression et éviter de me sentir impostrice (Clance and Imes, 1978) si je ne travail pas le week-end, j'ai choisi de comptabiliser quotidiennement mes heures de travail. Cela m'a permis d'évaluer objectivement si j'en faisais trop ou pas assez pour ma thèse, sur une base de normalité fixée à 7-8h de travail par jour [13].

10. « Crise de la santé mental en thèse »
11. À ce titre, je conseille vivement la lecture de *Managing your mental health during your PhD* de Ayres (2022).
12. Cette idée est particulièrement renforcée par le comportement des doctorantes elles-mêmes.
13. J'ai réalisé ces mesures durant mes trois années de thèses, les curieux et curieuses pourrons aller voir les graphiques en annexe (section 9.4).

J'ai également bénéficié d'un encadrement particulièrement humain, non-sexiste et bienveillant [14]. En revanche, cela a créé un fort contraste avec ce que j'ai pu observer par ailleurs. Cette expérience m'a même plutôt encouragée à m'engager auprès des autres étudiantes du laboratoire pour défendre leurs intérêts lors des réunions mensuelles de type CSO (Conseil Stratégique d'Orientation) et conseils d'UMR (Unité mixte de Recherche) du laboratoire. En plus de mon rôle de représentante des étudiantes, j'ai également été, pendant deux ans, vice-présidente de l'association des *Pinsons migRateurs*, hébergée au laboratoire. Cette association a pour objectif de favoriser les interactions scientifiques et sociales au sein du laboratoire, que ça soit entre étudiantes ou entre toutes les membres du laboratoire. Ces deux casquettes m'ont permis, avec l'aide d'autres étudiantes [15], d'organiser des temps d'échanges avec les étudiantes et permanentes autour des questions d'encadrement en thèse, de la vision de ce qu'est une thèse et de la culture de la recherche. J'ai également pu réaliser trois séminaires de préventions en santé mentale à l'attention des étudiantes du laboratoire [16]. Aujourd'hui, avec l'association des *Pinsons migRateurs*, nous défendons et diffusons activement l'idée de choisir ses futures encadrantes avec autant d'attention que son sujet de thèse, voir davantage.

Cette prise de conscience de l'importance du bien-être et de l'accompagnement en thèse n'est qu'une des réalisations marquantes de mon parcours doctoral. Celui-ci m'a également conduite à réfléchir plus largement à la façon dont nous, scientifiques, diffusons nos découvertes et à l'impact réel que notre travail peut avoir sur la société.

7.4.2 Diffuser la science

Lors de ma thèse, je me suis à plusieurs reprises confrontée à la question de l'utilité de mes travaux. En effet, bien que je trouve cela passionnant, je ne crois pas que connaitre les mécanismes liés à la variation d'intensité de la dérive génétique chez les mammifères, puisse permettre fondamentalement de résoudre les problèmes actuels auquel l'humanité est confrontée. Je me suis même plusieurs fois convaincue de rechercher des offres de post-doctorat dans des domaines de recherche plus

14. Merci Nicolas !
15. Merci Léa !
16. J'ai suivi, lors de ma thèse, une formation de premiers secours en santé mentale, sur laquelle je me suis basée pour pouvoir ensuite faire de la prévention

appliqués, comme l'étude des maladies génétiques ou des épidémies. Moins excitant, à mon avis, mais plus utile. Cette réflexion arrive dans un contexte où notre recherche à un impact écologique non négligeable (sous-section 7.3.3).

En parallèle de ces réflexions, j'ai pu constater le nombre grandissant de théories du complot (Kużelewska and Tomaszuk (2022) mais voir Uscinski *et al.* (2022)), une défiance envers les consensus scientifiques (Van der Linden *et al.*, 2015; Cobern, 2000) et l'impact de la diffusion d'intox, toutes plus énormes les unes que les autres [17]. C'est une observation que je ne suis pas la seule à faire (Hunter, 2016). À mon sens, tout cela a, en partie, pour origine une mécompréhension de ce qu'est la démarche et le raisonnement scientifique. Pour moi, il n'est plus possible de continuer à produire de la connaissance toujours de plus en plus complexe, si on ne fournit pas aux non-initiées intéressées, les clés pour la comprendre correctement. J'ai donc finalement trouvé du sens à mes travaux et à la recherche fondamentale en général, quand j'ai commencé à enseigner à l'université et à vulgariser la science au grand publique. Pour moi, c'est devenu une mission fondamentale qui fait partie intégrante du travail de chercheuse et qui ne pouvait pas ne pas être mentionnée dans ce manuscrit.

Par l'enseignement

Lors de ma thèse, j'ai bénéficié d'un contrat ACE (Activité Complémentaire d'Enseignement) de trois ans. J'ai eu l'occasion d'enseigner des notions de bio informatique, d'évolution et de phylogénie de base à des étudiantes en troisième année de licence et en deuxièmes années de master. Ce deuxième groupe d'étudiantes était particulièrement intéressante, car elles provenaient de différents masters plutôt éloignés des notions d'évolution et de bio-informatique [18], ce qui m'a poussée à aller les chercher dans leurs domaines respectifs pour leur montrer l'utilité de ce que j'avais à leur apprendre. J'ai également eu l'occasion, à trois reprises, de donner un cours magistral sur ce qu'est une phylogénie et les méthodes de reconstruction associées. Cette expérience enrichissante m'a permis de prendre confiance en ma

17. J'ai d'ailleurs participé à l'écriture d'une tribune concernant certaines défaillances médiatiques dans la diffusion scientifique : https://sfecologie.org/regard/ro24-juin-2024-jb-andre-et-al-evolution-et-medias/

18. Master biologie de la peau, cancérologie, épidémiologie, etc

qualité d'enseignante [19] et d'apprendre à construire *de novo* un cours complet, intéressant et accessible. Globalement, j'ai trouvé l'enseignement à l'université particulièrement stimulant, car il m'a permis d'être face à des étudiantes matures, qui ont envie d'apprendre et qui sont riches de questions. Je trouve qu'enseigner lors de sa thèse est particulièrement formateur, car on apprend, par l'expérience, à parler de science, à structurer un propos et à le reformuler de plusieurs façons.

Par la vulgarisation

Bien que j'ai particulièrement apprécié enseigner à l'université, je me suis rarement sentie plus utile que lors de mes activités de vulgarisation scientifique. Pour remettre les choses en contexte, je n'ai pas baigné dans le monde de la recherche dès ma naissance et j'ai longtemps aspiré à faire des métiers de scène ou d'images, avant de réaliser qu'il était possible d'explorer d'autres voies. Même si j'ai été préservée de cela, j'ai également grandi dans une banlieue où le taux d'échec scolaire est plutôt important et le niveau sociale faible. C'est pourquoi, il est essentiel pour moi, aujourd'hui, d'aller expliquer à des jeunes, et particulièrement à des jeunes filles, qu'elles peuvent et sont capables de faire de longues études, de la science, et notamment de l'informatique, si elles le souhaitent [20].

Afin de souligner mon propos concernant l'inclusion des femmes dans les représentations scientifiques et comment j'ai personnellement défendu cela, je reprends, ici, une écriture dans un français traditionnel et « correct », où le « masculin l'emporte sur le féminin » [21].

Lors de ma première année de thèse, j'ai eu l'opportunité de faire partie du projet « *Mon collège sur Mars* ». J'ai ainsi encadré une classe de CM1-CM2 dans mon quartier de naissance, à Vaulx-en-Velin, avec pour objectif de les faire réfléchir sur une potentielle implantation de leur futur collège sur la planète Mars. Nous avons donc, sur plusieurs séances, appris à faire des recherches documentaires, en

19. Merci Annabelle !
20. Voir ici la relation entre l'intensité des stéréotypes de genres et le nombre de femmes poursuivant des études en science : Miller *et al.* (2015).
21. Mais quand même, allez voir ces études : Miller and Swift (2001); Gaston (2015); Gygax *et al.* (2019); Kricheli-Katz and Regev (2021)

tirer des problématiques (Comment respirer sur Mars ? Comment se chauffer ? Comment créer de l'énergie ? etc.) et proposer des solutions. Les enfants ont ensuite réalisé une maquette de leur collège et présenté leurs résultats à d'autres classes lors d'un vrai séminaire scientifique. Pour moi, cet accompagnement a été l'occasion de leur fournir des outils pour un raisonnement rigoureux et de leur montrer que tout le monde peut faire de la science à son échelle.

En troisième année de thèse, j'ai eu l'occasion de prendre la parole dans un format relaxé de « *Ma thèse en 180 secondes* » lors de la journée des femmes et des filles de sciences, organisée par l'université Lyon 1. J'ai aussi donné une conférence, suivie d'une table ronde, lors d'un atelier de l'association « *Girls Can Code* », regroupant uniquement des jeunes filles auxquelles on apprend gratuitement à programmer en python de façon ludique. Lors de ces deux événements, ma volonté a été d'apporter un témoignage féminin de parcours scientifique et de montrer une voie, peut-être pas encore envisagée, à des jeunes filles en questionnement sur leur avenir.

Enfin, en 2024, j'ai participé à « *Pint of Science* , un festival de vulgarisation scientifique à l'échelle mondiale, lors duquel j'ai eu carte blanche pour parler de science au grand public et notamment de théorie neutre de l'évolution[22]. Ces différentes expériences m'ont fortement donné envie de poursuivre mon engagement dans la diffusion des sciences après ma thèse et m'ont encore plus convaincue de la nécessité de le faire.

7.5 Conclusion et réflexions pour l'avenir

J'ai l'impression qu'il existe une multitude de façons de conclure une thèse. Certain.e.s restent factuels et rappellent leurs résultats majeurs, d'autres proposent des perspectives ou abordent leurs futurs travaux. Pour ma part, j'ai choisi de me ressaisir du préambule de ce manuscrit qui met en avant mon envie, présente dès mon adolescence, de comprendre le monde, sa création et sa transformation. C'est une façon pour moi de vérifier si, à travers ma thèse, j'ai pu apporter quelques réponses ou satisfaction à la personne que j'étais il y a 10-15 ans.

22. Merci Damien !

Concernant mon attrait pour les questions d'origine et de temporalité, le fait de travailler avec des données représentant deux échelles temporelles différentes a été particulièrement stimulant pour moi. J'ai pu comprendre par moi-même, plutôt que de le lire dans des livres, que la variation génétique entre individus d'une même population peut, sur le long terme, conduire aux différences génétiques que l'on observe entre les espèces. J'ai également appris qu'on ne terminera jamais d'en apprendre sur des questions aussi larges que l'origine de la Terre ou de la vie, et qu'en science, il est à la fois sain et nécessaire de se poser fréquemment la question « Pourquoi ? » afin de déchiffrer les mécanismes qui sous-tendent nos observations. Enfin, contrairement à ce que j'ai pu penser, j'ai aujourd'hui le sentiment de faire partie de ces personnes qui participent à la création de connaissances. Cependant, je me suis beaucoup questionnée sur la nécessité de tout connaitre de notre monde et sur le cout de cette connaissance, qu'il soit financier, écologique, éthique ou social. Je pense qu'aujourd'hui, l'argument seul de la « beauté de la science » ne suffit plus, à mes yeux, pour justifier l'utilisation d'énormes clusters de calcul, le dérangement d'animaux pour leur séquençage, ou les nombreux voyages à travers le monde pour des congrès ou pour de la collecte de données. Il devient plus que nécessaire de repenser nos questions de recherche en fonction de leurs utilités réelles pour le monde et leurs impacts sur celui-ci. Pour moi, cela représente un dilemme entre le cœur (la recherche fondamentale en génomique, très énergivore et abstraite) et la raison (faire de la science, en utilisant des ressources raisonnées, appliquée au médical ou à la conservation), dilemme que je n'ai pas encore résolu et qui est central dans ma recherche de futurs contrats post-doctorats. En revanche, je n'ai aucun doute quant à la nécessité d'accompagner le développement de la connaissance par le devoir de la diffuser au plus grand nombre afin d'éviter qu'elle soit oubliée ou qu'elle devienne source de fractures sociales, tout en veillant à ne pas l'imposer aux individus comme une forme de colonisation intellectuelle.

Quoi qu'il en soit, ces dernières années ont été particulièrement épanouissantes pour moi, certainement parce que j'ai pu nourrir chacune de mes aspirations mentionnées dans le préambule de ce manuscrit, mais aussi m'en découvrir d'autres. À ce jour, mes envies scientifiques sont multiples. Jongler entre la phylogénie et la génétique des populations m'a permis de saisir le lien temporel entre ces deux disciplines et d'éviter de les découpler complètement. Cependant, naviguer entre ces

deux domaines, chacun très dense, demande une forte métaconnaissance en biologie évolutive de manière générale. Cela est parfois très éprouvant et créé un terrain propice au syndrome de l'imposteur. Aujourd'hui, je ressens plutôt le besoin d'approfondir mes connaissances dans ces deux domaines de façon indépendante afin de les appréhender plus sereinement à l'avenir. La question restante est de savoir par quoi commencer. De plus, mon expérience acquise avec l'usage de données empiriques me motive à poursuivre leur exploration. Cependant j'ai souvent regretté, ces dernières années, de ne pas savoir éprouver mes hypothèses théoriques par le biais de simulation avant de les confronter aux données empiriques. J'ai également, à plusieurs reprises, été très motivée à l'idée de prendre un virage méthodologique vers l'intelligence artificielle, mais je me questionne encore sur son impact écologique et sa pertinence par rapport à mes questions de recherche. J'ai pourtant l'impression que savoir manipuler ces outils « à la mode » sera crucial pour l'avenir.

Au vu de ces différentes aspirations parfois contradictoires et floues, je ne développerai pas sur le terme « mes questions de recherche », évoqué juste ci-dessus, tant le champ des possibles me semble à la fois vaste et confus. Je sais seulement que je souhaite continuer d'étudier les mécanismes des différents processus évolutifs, adaptatifs ou non, quelle que soit l'échelle et le type de données avec, pour objectif, de mieux comprendre notre monde tout en le respectant. Je sais également que je souhaite participer le moins possible à ce système de *publish or perish* où le taux de publication par an et notre capacité à tout quitter et changer de ville ou pays tous les 3-4 ans, conditionne l'accès à des postes permanents. Ainsi, même si j'ai adoré ma thèse et faire de la recherche, je suis consciente qu'il existe d'autres voies professionnelles attrayantes si nécessaire [23].

23. Et je ne suis pas la seule : https://www.lemonde.fr/sciences/article/2024/10/03/bernadette-bensaude-vincent-beaucoup-de-chercheurs-ont-envie-de-tout-plaquer-ou-d-aller-vers-des-actions-militantes_6342759_1650684.html

Remerciements

Il est temps pour moi de remercier les personnes qui m'ont aidée et soutenue tout au long de ce projet. D'abord, je remercie celles et ceux, et particulièrement les membres du jury, qui auront pris le temps de lire tout ou partie de ce manuscrit. J'espère que la lecture aura été plaisante et intéressante.

Ce manuscrit ne serait pas ce qu'il est sans l'aide d'un grand nombre de personnes. Un **merci** plein de gratitude à **Anouk** et à **Philippe** qui m'ont permis de rédiger l'introduction de cette thèse dans un cadre particulièrement apaisant. **Merci** également à toutes les personnes qui m'ont aidée dans la relecture et la correction de ce manuscrit : **Ana**, **Matthieu**, **Papa**, **Flow**, **Théo**, **Julien** et **Nicolas**. **Merci** à **Théo** qui a été comme un technicien 24/7 pour mon document Overleaf, avec un délai de réponse quasi-immédiat à mes « je crois que j'ai encore tout cassé ». Tu m'as évité tout le stress de la mise en page et de la gestion d'un document en LaTeX. Sans toi ce manuscrit ne serait pas aussi agréable à parcourir (pour moi en tout cas !). Je remercie aussi **Léa. B** et **Fanette** qui ont créé cette si jolie page de garde.

Un grand **merci** à **Nicolas** de m'avoir offert l'opportunité de faire cette thèse. En acceptant de travailler avec moi, d'abord en stage puis en thèse, tu m'as permis de prendre plus confiance et de m'imaginer un avenir concret dans la science. **Merci** pour ta pédagogie, le partage de ton immense connaissance, ta disponibilité, ta bienveillance et ton humanité. J'estime avoir été chanceuse de travailler à tes côtés. J'ai remarqué assez tard qu'il y a peu de thèse avec un unique encadrant. Bien que cela ne convienne peut-être pas à tout le monde, j'ai particulièrement apprécié ce mode de fonctionnement en duo. **Merci** de ne jamais m'avoir freinée dans tous mes engagements au sein du laboratoire ni dans mes multiples formations et activités de vulgarisation et d'enseignement.

Merci également aux membres de l'**équipe BPGE**, et notamment à **Laurent. D**, **Carina, Laure** et **Anouk**, pour les interactions scientifiques, l'aide apportée et le soutien moral. **Merci** à mon comité de suivi : **Thibault**, **Tristan**, **Benoit** et **Laurent. G**. Plus largement, je remercie le **LBBE** pour son accueil, la richesse des animations proposées, la place qui est faite à l'écoute des étudiant.e.s et la bonne humeur autour d'une machine à café parfois capricieuse.

Un **merci** particulier à mes co-bureaux, **Marie**, **Annabelle** et **Damien**. J'ai l'impression que c'est une autre originalité que de choisir de partager son bureau avec des permanent.e.s plutôt qu'avec d'autres étudiant.e.s. En tout cas, j'ai beaucoup appris à vos côtés. **Merci** pour l'entraide, les discussions et le bazar familier sur chacun de nos bureaux. **Merci** à **Annabelle** pour tous les cours que tu m'as permis de donner et notamment les CM. **Merci** à **Damien** pour *Pint of Science*. **Merci** à **Marie** d'être la femme solaire et inspirante que tu es.

En parlant de femmes inspirantes et puisqu'une partie de la discussion de cette thèse se veut teintée de messages féministes, je ne pouvais pas ne pas remercier les femmes inspirantes

que j'ai pu rencontrer au laboratoire. Celles qui n'ont pas peur de prendre la parole en réunions ou séminaires, d'imposer leurs désaccords ou de dénoncer des inégalités. **Merci** **Anouk**, **Laure**, **Carina**, **Nathalie**, **Marie**, **Léa**, **Amandine**, **Pauline** et pleins d'autres.

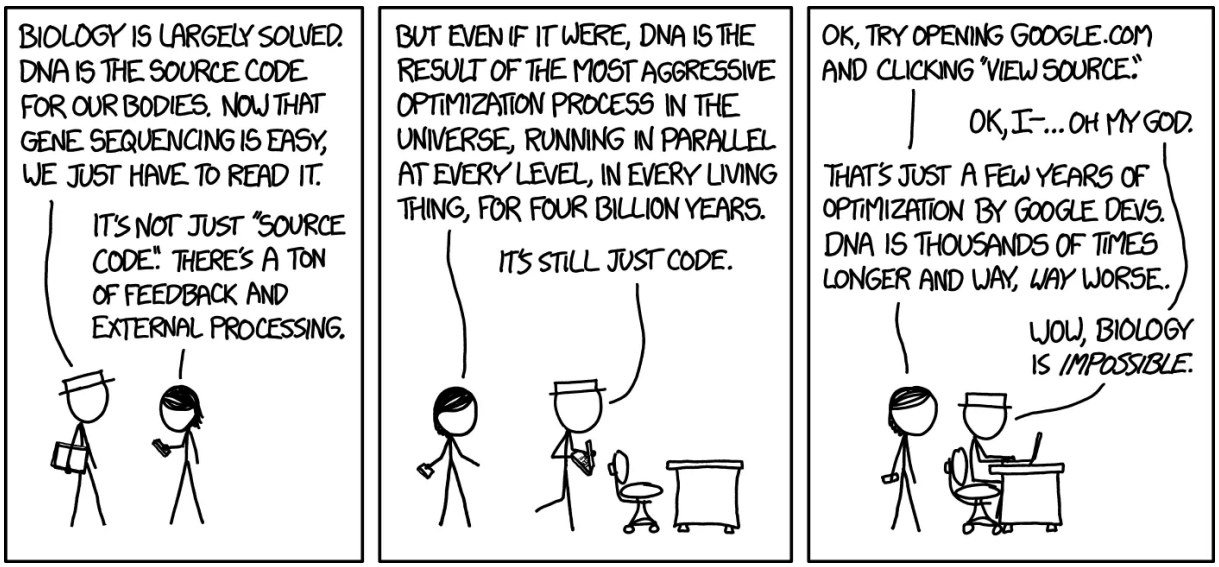
Par ailleurs, comme beaucoup de femmes, j'ai mis un pied dans la science par la biologie et j'ai appris la bio-informatique sur le tas, avec l'aide des membres du laboratoire. Un grand **merci** à **Théo** qui m'a surtout appris le Bash et à utiliser le cluster, ainsi qu'a **Thibault**, **Philippe** et **Florian. B** qui m'ont principalement appris le Python et Snakemake. **Merci** à **Flow** pour l'aide en R et ggplot, et au **pôle informatique** pour leurs formations.

Un **merci** chaleureux aux membres présent.e.s et passé.e.s du bureau des *Pinsons migRateurs* : **Léa**, **Emma. A**, **Chloé**, **Emma. C**, **Sasha**, **Solène**, **Lucas**, **Blandine** et **Alice**. J'ai été ravie de porter cette association avec vous et je suis fière de ce qu'on en a fait. Un **merci** aussi aux représentant.e.s des étudiant.e.s actuel.le.s et passé.e.s : **Amandine**, **Sasha** et **Julien**. J'ai adoré être aux CSO avec vous à tenter de décoder tous les acronymes. **Merci** au **LBBE** de nous permettre d'y participer.

Mon séjour au LBBE a été le théâtre d'un grand épanouissement personnel et je le dois, en partie, aux amitiés qui s'y sont crées. **Merci** à **Théo** qui est devenu mon *partner in crime* au labo. **Merci** pour ton accueil dès mon stage, d'avoir partagé ta somptueuse cuisine avec moi ainsi que ta musique, si chère à tes oreilles. **Merci** pour l'aide et le soutien que tu m'as apportés jusqu'au bout lors de la rédaction de ce manuscrit. **Merci** à **Léa** d'être entrée si fort dans ma vie. Avec toi, j'ai découvert la vraie signification du mot « sororité » et j'ai pu enfin comprendre que l'amitié peut être un sentiment aussi important que l'amour (Power in sisterhood !). Tu m'as tellement appris, sans forcément t'en rendre compte, et je ne suis pas assez bonne littéraire pour te le retranscrire avec des mots. Alors, je t'adresse simplement un énorme **MERCI** pour tout. **Merci** à **Flow**, avec qui notre amitié ne s'est, certes, pas formée au laboratoire mais remonte à notre licence. Tu as été une des personnes présente dès le début qui m'a soutenue tout du long et je suis très fière qu'on soit tous les deux arrivés ou nous en sommes. **Merci**, plus généralement, à tous les autres collègues de laboratoire qui sont devenus des amis : **Julien. J**, **Victor**, **Marie. M**, **Florian. B**, **Adrian**, **Emma.A**, **Rémi**, **Lisa**, **Mary**, **Gaspard**, **Amandine**, **Thibault**, **Blandine**, **Alexandre** et pleins d'autres.

Merci également à mes proches ; à mes **parents** qui m'ont toujours acheté des livres pour peu qu'ils soient nécessaires à mes études, aux **copains des maths** qui acceptent mon exotisme dans leur monde de chiffres et à **Mme Colson-Proch** de m'avoir fait entrevoir le champ des possibles dès le début du lycée. **Merci**, enfin, à **Matthieu**, mon amoureux et mon soleil. Tu as été d'un soutien constant lors de mes études et lors de cette thèse. L'admiration à mon égard que je vois dans tes yeux est ma plus grande force. Je t'aime.

Je me suis promise de terminer ces remerciements en me remerciant moi-même. Alors **merci** **moi**, de m'être autorisée à faire les études qui me plaisaient sans trop m'ennuyer avec des choix stratégiques d'orientation. **Merci** à **moi** d'avoir concrétisé mes envies d'être actrice de la science et de m'être lancée dans cette aventure si enrichissante.



Pour Nicolas, crédit xkcd (https://xkcd.com/1605/)

Quatrième partie

Bibliographie

Bibliographie

[1] Achaz, G. 2008. Testing for neutrality in samples with sequencing errors. *Genetics*, 179(3) : 1409–1424. *Cité à la page 83*

[2] Aguade, M., Miyashita, N., and Langley, C. H. 1989. Reduced variation in the yellow-achaete-scute region in natural populations of drosophila melanogaster. *Genetics*, 122(3) : 607–615. *Cité à la page 52*

[3] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. 1990. Basic local alignment search tool. *Journal of molecular biology*, 215(3) : 403–410. *Cité à la page 64*

[4] Andolfatto, P. 2001. Adaptive hitchhiking effects on genome variability. *Current opinion in genetics & development*, 11(6) : 635–641. *Cité à la page 53*

[5] Anisimova, M. and Kosiol, C. 2008. Investigating Protein-Coding Sequence Evolution with Probabilistic Codon Substitution Models. *Molecular Biology and Evolution*, 26(2) : 255–271. *Cité à la page 78, 80*

[6] Aristote -325. *Génération des animaux*. *Cité à la page 8*

[7] Aristote -343. *Histoire des animaux*. *Cité à la page 8*

[8] Aristote 1 av. J.-C. *Parties des animaux*. *Cité à la page 8*

[9] Avery, O. T., MacLeod, C. M., and McCarty, M. 1944. Induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type iii. *Die Entdeckung der Doppelhelix*, 97. *Cité à la page 19*

[10] Ayres, Z. J. 2022. *Managing your mental health during your PhD : A survival guide*. Springer. *Cité à la page 203*

[11] Bateson, W. and Gregory, R. P. 1905. On the inheritance of heterostylism in primula. *Proceedings of the Royal Society of London. Series B, Containing Papers of a Biological Character*, 76(513) : 581–586. *Cité à la page 19*

[12] Beadle, G. W. and Tatum, E. L. 1941. Genetic control of biochemical reactions in neurospora. *Proceedings of the National Academy of Sciences*, 27(11) : 499–506. *Cité à la page 19*

[13] Begun, D. J. and Aquadro, C. F. 1992. Levels of naturally occurring dna polymorphism correlate with recombination rates in d. melanogaster. *Nature*, 356(6369) : 519–520. *Cité à la page 52*

[14] Bergeron, L. A., Besenbacher, S., Zheng, J., Li, P., Bertelsen, M. F., Quintard, B., Hoffman, J. I., Li, Z., St. Leger, J., Shao, C., *et al.* 2023. Evolution of the germline mutation rate across vertebrates. *Nature*, 615(7951) : 285–291. *Cité à la page 179*

[15] Berthoud, F., Bzeznik, B., Gibelin, N., Laurens, M., Bonamy, C., Morel, M., and Schwindenhammer, X. 2020. *Estimation de l'empreinte carbone d'une heure. coeur de calcul*. Ph.D. thesis, UGA-Université Grenoble Alpes ; CNRS ; INP Grenoble ; INRIA. *Cité à la page 198*

[16] Blanquart, F., Achaz, G., Bataillon, T., and Tenaillon, O. 2014. Properties of selected mutations and genotypic landscapes under fisher's geometric model. *Evolution*, 68(12) : 3537–3554. *Cité à la page 185*

[17] Botero-Castro, F., Figuet, E., Tilak, M.-K., Nabholz, B., and Galtier, N. 2017. Avian genomes revisited : hidden genes uncovered and the rates versus traits paradox in birds. *Molecular biology and evolution*, 34(12) : 3123–3131. *Cité à la page 111*

[18] Brevet, M. and Lartillot, N. 2021. Reconstructing the history of variation in effective population size along phylogenies. *Genome Biology and Evolution*, 13(8) : evab150. *Cité à la page 108, 110, 111, 114, 115, 116, 179, 184, 192*

[19] Bromham, L., Rambaut, A., and Harvey, P. H. 1996. Determinants of rate variation in mammalian dna sequence evolution. *Journal of molecular evolution*, 43 : 610–621. *Cité à la page 109*

[20] Buffalo, V. 2021. Quantifying the relationship between genetic diversity and population size suggests natural selection cannot explain lewontin's paradox. *Elife*, 10 : e67509. *Cité à la page 51, 52, 53*

[21] Cai, J. J., Macpherson, J. M., Sella, G., and Petrov, D. A. 2009. Pervasive hitchhiking at coding and regulatory sites in humans. *PLoS genetics*, 5(1) : e1000336. *Cité à la page 52*

[22] Castellano, D., Macià, M. C., Tataru, P., Bataillon, T., and Munch, K. 2019. Comparison of the full distribution of fitness effects of new amino acid mutations across great apes. *Genetics*, 213(3) : 953–966. *Cité à la page 184*

[23] Castresana, J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular biology and evolution*, 17(4) : 540–552. *Cité à la page 69*

[24] Cavalli-Sforza, L. L. and Edwards, A. W. 1967. Phylogenetic analysis. models and estimation procedures. *American journal of human genetics*, 19(3 Pt 1) : 233. *Cité à la page 30*

[25] Chamary, J.-V., Parmley, J. L., and Hurst, L. D. 2006. Hearing silence : non-neutral evolution at synonymous sites in mammals. *Nature Reviews Genetics*, 7(2) : 98–108. *Cité à la page 188*

[26] Chargaff, E. 1950. Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. *Experientia*, 6(6) : 201–209. *Cité à la page 20*

[27] Charlesworth, B. and Jensen, J. D. 2022. How can we resolve lewontin's paradox ? *Genome biology and evolution*, 14(7) : evac096. *Cité à la page 52, 53*

[28] Charlesworth, B., Morgan, M., and Charlesworth, D. 1993. The effect of deleterious mutations on neutral molecular variation. *Genetics*, 134(4) : 1289–1303. *Cité à la page 52*

[29] Charlesworth, J. and Eyre-Walker, A. 2008. The mcdonald–kreitman test and slightly deleterious mutations. *Molecular biology and evolution*, 25(6) : 1007–1015. *Cité à la page 94*

[30] Chen, J., Glémin, S., and Lascoux, M. 2017. Genetic diversity and the efficacy of purifying selection across plant and animal species. *Molecular biology and evolution*, 34(6) : 1417–1428. *Cité à la page 110*

[31] Chen, J., Glémin, S., and Lascoux, M. 2020. From drift to draft : how much do beneficial mutations actually contribute to predictions of ohta's slightly deleterious model of molecular evolution ? *Genetics*, 214(4) : 1005–1018. *Cité à la page 52*

[32] Choi, S. C., Hobolth, A., Robinson, D. M., Kishino, H., and Thorne, J. L. 2007. Quantifying the impact of protein tertiary structure on molecular evolution. *Molecular biology and evolution*, 24(8) : 1769–1782. *Cité à la page 80*

[33] Churchill, F. B. 1974. William johannsen and the genotype concept. *Journal of the History of Biology*, pages 5–30. *Cité à la page 19*

[34] Clance, P. R. and Imes, S. A. 1978. The imposter phenomenon in high achieving women : Dynamics and therapeutic intervention. *Psychotherapy : Theory, research & practice*, 15(3) : 241. *Cité à la page 203*

[35] Clark, A. G., Eisen, M. B., Smith, D. R., Bergman, C. M., Oliver, B., Markow, T. A., Kaufman, T. C., Kellis, M., Gelbart, W., Iyer, V. N., *et al.* 2007. Evolution of genes and genomes on the drosophila phylogeny. *Nature*, 450(7167) : 203–18. *Cité à la page 77*

[36] Cobern, W. W. 2000. The nature of science and the role of knowledge and belief. *Science & Education*, 9 : 219–246. *Cité à la page 205*

[37] Comte, A., Tricou, T., Tannier, E., Joseph, J., Siberchicot, A., Penel, S., Allio, R., Delsuc, F., Dray, S., and de Vienne, D. M. 2023. Phylter : efficient identification of outlier sequences in phylogenomic datasets. *Molecular Biology and Evolution*, 40(11) : msad234. *Cité à la page 70, 189*

[38] Corbett-Detig, R. B., Hartl, D. L., and Sackton, T. B. 2015. Natural selection constrains neutral diversity across a wide range of species. *PLoS biology*, 13(4) : e1002112. *Cité à la page 52, 53*

[39] Correns, C. 1900. G. mendel's regal uber das verhalten der nach-kommenschaft der rassenbasterde. *Ber Deutsch Bot Ges*, 18 : 158–168. *Cité à la page 18*

[40] Crick, F. 1979. Split genes and rna splicing. *Science*, 204(4390) : 264–271. *Cité à la page 36*

[41] Criscuolo, A. and Gribaldo, S. 2010. Bmge (block mapping and gathering with entropy) : a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC evolutionary biology*, 10 : 1–21. *Cité à la page 69, 70, 189, 194*

[42] Crow, J. 1981. The neutralist-selectionist controversy : an overview. *Population and Biological Aspects of Human Evolution. Academic Press, New York*, pages 3–14. *Cité à la page 32*

[43] Crow, J. F. *et al.* 1954. Breeding structure of populations. ii. effective population number. *Statistics and mathematics in biology*, 543 : 556. *Cité à la page 47, 48*

[44] Crow, J. F. and Morton, N. E. 1955. Measurement of gene frequency drift in small populations. *Evolution*, pages 202–214. *Cité à la page 48*

[45] Cuénot, L. 1902. La loi de mendel et l'hérédité de la pigmentation chez les souris. *Archives de Zoologie Expérimentale et Générale 3 Series*. *Cité à la page 19*

[46] Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., *et al.* 2021. Twelve years of samtools and bcftools. *Gigascience*, 10(2) : giab008. *Cité à la page 188*

[47] Darwin, C. 1859. *On the origin of species*. *Cité à la page 12, 14, 37, 73*

[48] Darwin, C. 1868. *The variation of animals and plants under domestication*. *Cité à la page 17*

[49] Darwin, C. 1871. *La filiation de l'homme*. *Cité à la page 15*

[50] Dayhoff, M. 1978. Atlas of protein sequence and structure. *(No Title)*, page 345. *Cité à la page 39, 80*

[51] De Vries, H. 1900. Sur la loi do disjonction des hybrides. *CR Acad Sci*, 130 : 845–847. *Cité à la page 18*

[52] De Vries, H. 1909. *Espèces et variétés : leur naissance par mutation*, volume 111. Alcan. *Cité à la page 19*

[53] Desdevises, Y. 2018. Introduction générale aux méthodes comparatives phylogénétiques. *Biosystema*, 31 : 23–43. *Cité à la page 96, 98, 99*

[54] Dessimoz, C. and Gil, M. 2010. Phylogenetic assessment of alignments reveals neglected tree signal in gaps. *Genome biology*, 11 : 1–9. *Cité à la page 68*

[55] Di Franco, A., Poujol, R., Baurain, D., and Philippe, H. 2019. Evaluating the usefulness of alignment filtering methods to reduce the impact of errors on evolutionary inferences. *BMC Evolutionary Biology*, 19 : 1–17. *Cité à la page 69, 70, 189, 194*

[56] Di Giulio, M. 2005. The origin of the genetic code : theories and their relationships, a review. *Biosystems*, 80(2) : 175–184. *Cité à la page 20*

[57] Dickerson, R. E. and Geis, I. 1969. The structure and action of proteins. *(No Title)*. *Cité à la page 43*

[58] Dobzhansky, T. 1951. Genetics and the origin of species. *Cité à la page 23*

[59] Duchemin, L. 2023. *Détection phylogénétique de sites protéiques associés à un phénotype, à l'échelle génomique*. Ph.D. thesis, Lyon 1. *Cité à la page 190*

[60] Duret, L. and Galtier, N. 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annual review of genomics and human genetics*, 10 : 285–311. *Cité à la page 188, 191*

[61] Duschesne 1766. *Histoire naturelle des fraisiers*. *Cité à la page 10*

[62] Edwards, A. W. and Cavalli-Sforza, L. L. 1963. The reconstruction of evolution. *Cité à la page 30*

[63] Edwards, A. W. and Cavalli-Sforza, L. L. 1965. A method for cluster analysis. *Biometrics*, pages 362–375. *Cité à la page 30*

[64] Edwards, A. W. F. 1972. Likelihood. In *Time Series and Statistics*, pages 126–129. Springer. *Cité à la page 30*

[65] Evans, T. M., Bira, L., Gastelum, J. B., Weiss, L. T., and Vanderford, N. L. 2018. Evidence for a mental health crisis in graduate education. *Nature biotechnology*, 36(3) : 282–284. *Cité à la page 202, 203*

[66] Ewens, W. J. 2004. *Mathematical population genetics : theoretical introduction*, volume 27. Springer. *Cité à la page 49*

[67] Eyre-Walker, A. 2002. Changing effective population size and the mcdonald-kreitman test. *Genetics*, 162(4) : 2017–2024. *Cité à la page 77, 93, 94*

[68] Eyre-Walker, A. and Keightley, P. D. 2007. The distribution of fitness effects of new mutations. *Nature Reviews Genetics*, 8(8) : 610–618. *Cité à la page 44, 94, 115*

[69] Eyre-Walker, A. and Keightley, P. D. 2009. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Molecular biology and evolution*, 26(9) : 2097–2108. *Cité à la page 94*

[70] Fay, J. C., Wyckoff, G. J., and Wu, C.-I. 2002. Testing the neutral theory of molecular evolution with genomic data from drosophila. *Nature*, 415(6875) : 1024–1026. *Cité à la page 94*

[71] Felsenstein, J. 1981. Evolutionary trees from dna sequences : a maximum likelihood approach. *Journal of molecular evolution*, 17 : 368–376. *Cité à la page 31*

[72] Felsenstein, J. 1985. Phylogenies and the comparative method. *The American Naturalist*, 125(1) : 1–15. *Cité à la page 95, 96, 97, 99, 100, 102*

[73] Feng, S., Stiller, J., Deng, Y., Armstrong, J., Fang, Q., Reeve, A. H., Xie, D., Chen, G., Guo, C., Faircloth, B. C., *et al.* 2020. Dense sampling of bird diversity increases power of comparative genomics. *Nature*, 587(7833) : 252–257. *Cité à la page 188*

[74] Figuet, E., Nabholz, B., Bonneau, M., Mas Carrio, E., Nadachowska-Brzyska, K., Ellegren, H., and Galtier, N. 2016. Life history traits, protein evolution, and the nearly neutral theory in amniotes. *Molecular biology and evolution*, 33(6) : 1517–1527. *Cité à la page 107, 109, 111, 116*

[75] Fisher, R. A. 1918. Xv.—the correlation between relatives on the supposition of mendelian inheritance. *Earth and Environmental Science Transactions of the Royal Society of Edinburgh*, 52(2) : 399–433. *Cité à la page 19, 25, 26*

[76] Fisher, R. A. 1919. The causes of human variability. *The Eugenics Review*, 10(4) : 213. *Cité à la page 19*

[77] Fisher, R. A. 1928. *Statistical methods for research workers*. Number 5. Oliver and Boyd. *Cité à la page 196*

[78] Fisher, R. A. 1930. The genetical theory of natural selection oxford. *UK : Clarendon*. *Cité à la page 23, 25, 33, 34, 50, 52, 83*

[79] Fisher, R. A. 1931. Xvii.—the distribution of gene ratios for rare mutations. *Proceedings of the Royal Society of Edinburgh*, 50 : 204–219. *Cité à la page 23, 24*

[80] Fisher, R. A. *et al.* 1922. 024 : On the dominance ratio. *Cité à la page 23*

[81] Fisher, R. A. *et al.* 1947. 219 : The spread of a gene in natural conditions in a colony of the moth panaxia dominula l. *Cité à la page 24*

[82] Fisher, R. A. and Ford, E. B. 1950. The" sewall wright effect. *Cité à la page 24*

[83] Fitch, W. M. 1970. Distinguishing homologous from analogous proteins. *Systematic zoology*, 19(2) : 99–113. *Cité à la page 66*

[84] Fletcher, W. and Yang, Z. 2010. The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. *Molecular biology and evolution*, 27(10) : 2257–2267. *Cité à la page 68*

[85] Gadbury, G. L. and Allison, D. B. 2012. Inappropriate fiddling with statistical analyses to obtain a desirable p-value : tests to detect its presence in published literature. *Cité à la page 197*

[86] Galtier, N. 2001. Maximum-likelihood phylogenetic analysis under a covarion-like model. *Molecular Biology and Evolution*, 18(5) : 866–873. *Cité à la page 79*

[87] Galtier, N. and Rousselle, M. 2020. How much does ne vary among species ? *Genetics*, 216(2) : 559–572. *Cité à la page 52*

[88] Galton, F. 1865. Hereditary character and talent. *Macmillan's Magazine*, 12(1865) : 157–166. *Cité à la page 16*

[89] Galton, F. 1877. *Typical laws of heredity*. William Clowes and Sons. *Cité à la page 18*

[90] Galton, F. 1891. *Hereditary genius*. D. Appleton. *Cité à la page 16*

[91] Gamow, G. 1954. Possible relation between deoxyribonucleic acid and protein structures. *Nature*, 173(4398) : 318–318. *Cité à la page 20*

[92] Gärtner, C. F. v. 1849. Versuche und beobachtungen über die bastarderzeugung im pflanzenreich. *Manuscript material*. *Cité à la page 17*

[93] Gaston, N. 2015. *Why science is sexist*, volume 34. Bridget Williams Books. *Cité à la page 206*

[94] Geiler, K. and Harrison, R. 2010. A 11 desaturase gene genealogy reveals two divergent allelic classes within the european corn borer (ostrinia nubilalis). *BMC evolutionary biology*, 10 : 112. *Cité à la page 86*

[95] Gillespie, J. H. 1991. *The causes of molecular evolution*, volume 2. Oxford University Press, USA. *Cité à la page 52*

[96] Gillespie, J. H. 2000. Genetic drift in an infinite population : the pseudohitchhiking model. *Genetics*, 155(2) : 909–919. *Cité à la page 52*

[97] Goldman, N. and Yang, Z. 1994. A codon-based model of nucleotide substitution for protein-coding dna sequences. *Molecular biology and evolution*, 11(5) : 725–736. *Cité à la page 78*

[98] Goldstein, R. A. 2013. Population size dependence of fitness effect distribution and substitution rate probed by biophysical model of protein thermostability. *Genome biology and evolution*, 5(9) : 1584–1593. *Cité à la page 184*

[99] Gould, E., Fraser, H. S., Parker, T. H., Nakagawa, S., Griffith, S. C., Vesk, P. A., Fidler, F., Hamilton, D. G., Abbey-Lee, R. N., Abbott, J. K., *et al.* 2023. Same data, different analysts : variation in effect sizes due to analytical decisions in ecology and evolutionary biology. *Cité à la page 197*

[100] Gouy, M., Guindon, S., and Gascuel, O. 2010. Seaview version 4 : a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular biology and evolution*, 27(2) : 221–224. *Cité à la page 70*

[101] Grafen, A. 1989. The phylogenetic regression. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 326(1233) : 119–157. *Cité à la page 100*

[102] Guindon, S., Rodrigo, A. G., Dyer, K. A., and Huelsenbeck, J. P. 2004. Modeling the site-specific variation of selection patterns along lineages. *Proceedings of the National Academy of Sciences*, 101(35) : 12957–12962. *Cité à la page 79*

[103] Gygax, P., Gabriel, U., and Zufferey, S. 2019. Le masculin et ses multiples sens : Un problème pour notre cerveau... et notre société. 10. *Cité à la page 206*

[104] Haeckel, E. 1866. *Generelle morphologie des organischen.* *Cité à la page 13*

[105] Haldane 1932. *The causes of evolution*, volume 3. Longmans, Green & Co. *Cité à la page 23, 24*

[106] Haldane, J. 1926. A mathematical theory of natural and artificial selection. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 23, pages 363–372. Cambridge University Press. *Cité à la page 24*

[107] Hall, S. 2023. A mental-health crisis is gripping science—toxic research culture is to blame. *Nature*, 617(7962) : 666–668. *Cité à la page 202*

[108] Halpern, A. L. and Bruno, W. J. 1998. Evolutionary distances for protein-coding sequences : modeling site-specific residue frequencies. *Molecular biology and evolution*, 15(7) : 910–917. *Cité à la page 185*

[109] Hansen, T. F. 1997. Stabilizing selection and the comparative analysis of adaptation. *Evolution*, 51(5) : 1341–1351. *Cité à la page 98*

[110] Hardy, G. H. 1908. Mendelian proportions in a mixed population. *Science*, 28(706) : 49–50. *Cité à la page 23, 50*

[111] Harris, H. 1966. C. genetics of man enzyme polymorphisms in man. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 164(995) : 298–310. *Cité à la page 28*

[112] Harrison, P. W., Amode, M. R., Austine-Orimoloye, O., Azov, A. G., Barba, M., Barnes, I., Becker, A., Bennett, R., Berry, A., Bhai, J., Bhurji, S. K., Boddu, S., Branco Lins, P. R., Brooks, L., Ramaraju, S. B., Campbell, L. I., Martinez, M. C., Charkhchi, M., Chougule, K., Cockburn, A., Davidson, C., De Silva, N. H., Dodiya, K., Donaldson, S., El Houdaigui, B., Naboulsi, T. E., Fatima, R., Giron, C. G., Genez, T., Grigoriadis, D., Ghattaoraya, G. S., Martinez, J. G., Gurbich, T. A., Hardy, M., Hollis, Z., Hourlier, T., Hunt, T., Kay, M., Kaykala, V., Le, T., Lemos, D., Lodha, D., Marques-Coelho, D., Maslen, G., Merino, G. A., Mirabueno, L. P., Mushtaq, A., Hossain, S. N., Ogeh, D. N., Sakthivel, M. P., Parker, A., Perry, M., Piližota, I., Poppleton, D., Prosovetskaia, I., Raj, S., Pérez-Silva, J. G., Salam, A. I. A., Saraf, S., Saraiva-Agostinho, N., Sheppard, D., Sinha, S., Sipos, B., Sitnik, V., Stark, W., Steed, E., Suner, M.-M., Surapaneni, L., Sutinen, K., Tricomi, F. F., Urbina-Gómez, D., Veidenberg, A., Walsh, T. A., Ware, D., Wass, E., Willhoft, N., Allen, J., Alvarez-Jarreta, J., Chakiachvili, M., Flint, B., Giorgetti, S., Haggerty, L., Ilsley, G., Keatley, J., Loveland, J., Moore, B., Mudge, J., Naamati, G., Tate, J., Trevanion, S., Winterbottom, A., Frankish, A., Hunt, S. E., Cunningham, F., Dyer, S., Finn, R., Martin, F., and Yates, A. 2023. Ensembl 2024. *Nucleic Acids Research*, 52(D1) : D891–D899. *Cité à la page 72*

[113] Håstad, O. and Björklund, M. 1998. Nucleotide substitution models and estimation of phylogeny. *Molecular Biology and Evolution*, 15(11) : 1381–1389. *Cité à la page 31*

[114] Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., and Jennions, M. D. 2015. The extent and consequences of p-hacking in science. *PLoS biology*, 13(3) : e1002106. *Cité à la page 197*

[115] Hernandez, R. D., Kelley, J. L., Elyashiv, E., Melton, S. C., Auton, A., McVean, G., Project, . G., Sella, G., and Przeworski, M. 2011. Classic selective sweeps were rare in recent human evolution. *science*, 331(6019) : 920–924. *Cité à la page 53*

[116] Higgins, D. G. and Sharp, P. M. 1988. Clustal : a package for performing multiple sequence alignment on a microcomputer. *Gene*, 73(1) : 237–244. *Cité à la page 67*

[117] Hill, W. G. and Robertson, A. 1966. The effect of linkage on limits to artificial selection. *Genetics Research*, 8(3) : 269–294. *Cité à la page 52*

[118] Holmes, I. and Bruno, W. J. 2001. Evolutionary hmms : a bayesian approach to multiple alignment. *Bioinformatics*, 17(9) : 803–820. *Cité à la page 58*

[119] Hudson, R. R. 1983. Testing the constant-rate neutral allele model with protein sequence data. *Evolution*, pages 203–217. *Cité à la page 85*

[120] Huelsenbeck, J. P. and Rannala, B. 2003. Detecting correlation between characters in a comparative analysis with uncertain phylogeny. *Evolution*, 57(6) : 1237–1247. *Cité à la page 79, 101, 102*

[121] Huelsenbeck, J. P. and Ronquist, F. 2001. Mrbayes : Bayesian inference of phylogenetic trees. *Bioinformatics*, 17(8) : 754–755. *Cité à la page 78*

[122] Hunter, P. 2016. The communications gap between scientists and public : More scientists and their institutions feel a need to communicate the results and nature of research with the public. *EMBO reports*, 17(11) : 1513–1515. *Cité à la page 205*

[123] Huxley, J. 1942. Evolution : The modern synthesis. *Cité à la page 25*

[124] James, J., Castellano, D., and Eyre-Walker, A. 2017. Dna sequence diversity and the efficiency of natural selection in animal mitochondrial dna. *Heredity*, 118(1) : 88–95. *Cité à la page 109, 110, 116, 117, 121*

[125] Janzen, D. H., Hallwachs, W., Blandin, P., Burns, J. M., CADIOU, J.-M., Chacon, I., Dapkey, T., Deans, A. R., Epstein, M. E., Espinoza, B., *et al.* 2009. Integration of dna barcoding into an ongoing inventory of complex tropical biodiversity. *Molecular ecology resources*, 9 : 1–26. *Cité à la page 188*

[126] Johannsen, W. 1909. Elemente der exakten erblichkeitslehre (1913). *et seq*, page 254. *Cité à la page 18, 19, 25*

[127] Johnson, P. L. and Slatkin, M. 2008. Accounting for bias from sequencing error in population genetic estimates. *Molecular biology and evolution*, 25(1) : 199–206. *Cité à la page 83*

[128] Jones, D. T., Taylor, W. R., and Thornton, J. M. 1992. The rapid generation of mutation data matrices from protein sequences. *Bioinformatics*, 8(3) : 275–282. *Cité à la page 80*

[129] Jukes, T. 1969. Evolution of protein molecules. *Mammalian Protein Metabolism*, 3. *Cité à la page 75*

[130] Jukes, T. H. 1978. Neutral changes during divergent evolution of hemoglobins. *Journal of Molecular Evolution*, 11 : 267–269. *Cité à la page 32*

[131] Keightley, P. D. and Eyre-Walker, A. 2007. Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics*, 177(4) : 2251–2261. *Cité à la page 44*

[132] Kern, A. D. and Hahn, M. W. 2018. The neutral theory in light of natural selection. *Molecular biology and evolution*, 35(6) : 1366–1371. *Cité à la page 52*

[133] Kimura, M. 1968a. Evolutionary rate at the molecular level. *Nature*, 217(5129) : 624–626. *Cité à la page 31, 33, 34, 176*

[134] Kimura, M. 1968b. Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. *Genetics research*, 11(3) : 247–270. *Cité à la page 31, 32, 42, 44, 45, 176*

[135] Kimura, M. 1969. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, 61(4) : 893. *Cité à la page 92*

[136] Kimura, M. 1971. Theoretical foundation of population genetics at the molecular level. *Theoretical population biology*, 2(2) : 174–208. *Cité à la page 41*

[137] Kimura, M. 1977. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature*, 267(5608) : 275–276. *Cité à la page 32*

[138] Kimura, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of molecular evolution*, 16 : 111–120. *Cité à la page 75*

[139] Kimura, M. 1983. *The neutral theory of molecular evolution*. Cambridge University Press. *Cité à la page 31, 33, 34, 39, 40, 41, 43, 52, 176*

[140] Kimura, M. and Crow, J. F. 1963. The measurement of effective population number. *Evolution*, pages 279–288. *Cité à la page 47, 48*

[141] Kimura, M. and Crow, J. F. 1964. The number of alleles that can be maintained in a finite population. *Genetics*, 49(4) : 725. *Cité à la page 41*

[142] Kimura, M. and Ohta, T. 1969a. The average number of generations until extinction of an individual mutant gene in a finite population. *Genetics*, 63(3) : 701. *Cité à la page 34*

[143] Kimura, M. and Ohta, T. 1969b. The average number of generations until fixation of a mutant gene in a finite population. *Genetics*, 61(3) : 763. *Cité à la page 34*

[144] Kimura, M. and Ohta, T. 1971. Protein polymorphism as a phase of molecular evolution. *Cité à la page 34*

[145] King, J. L. and Jukes, T. H. 1969. Non-darwinian evolution : Most evolutionary change in proteins may be due to neutral mutations and genetic drift. *Science*, 164(3881) : 788–798. *Cité à la page 32, 33, 36*

[146] Kingman, J. F. 1982a. On the genealogy of large populations. *Journal of applied probability*, 19(A) : 27–43. *Cité à la page 85*

[147] Kingman, J. F. C. 1982b. The coalescent. *Stochastic processes and their applications*, 13(3) : 235–248. *Cité à la page 85*

[148] Kirilenko, B. M., Munegowda, C., Osipova, E., Jebb, D., Sharma, V., Blumer, M., Morales, A. E., Ahmed, A.-W., Kontopoulos, D.-G., Hilgers, L., *et al.* 2023. Integrating gene annotation with orthology inference at scale. *Science*, 380(6643) : eabn3107. *Cité à la page 65, 189, 190*

[149] Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., and Phillippy, A. M. 2017. Canu : scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome research*, 27(5) : 722–736. *Cité à la page 62*

[150] Kosakovsky Pond, S. L., Murrell, B., Fourment, M., Frost, S. D., Delport, W., and Scheffler, K. 2011. A random effects branch-site model for detecting episodic diversifying selection. *Molecular biology and evolution*, 28(11) : 3033–3043. *Cité à la page 80*

[151] Kosiol, C., Vinař, T., da Fonseca, R. R., Hubisz, M. J., Bustamante, C. D., Nielsen, R., and Siepel, A. 2008. Patterns of positive selection in six mammalian genomes. *PLoS genetics*, 4(8) : e1000144. *Cité à la page 78*

[152] Krasovec, M., Rickaby, R. E., and Filatov, D. A. 2020. Evolution of mutation rate in astronomically large phytoplankton populations. *Genome biology and evolution*, 12(7) : 1051–1059. *Cité à la page 52*

[153] Kricheli-Katz, T. and Regev, T. 2021. The effect of language on performance : do gendered languages fail women in maths ? *NPJ science of learning*, 6(1) : 9. *Cité à la page 206*

[154] Kużelewska, E. and Tomaszuk, M. 2022. Rise of conspiracy theories in the pandemic times. *International Journal for the Semiotics of Law-Revue internationale de Sémiotique juridique*, 35(6) : 2373–2389. *Cité à la page 205*

[155] Lamarck, J.-B. 1779. *Flore françoise*. *Cité à la page 9*

[156] Lamarck, J.-B. 1809. *Philosophie zoologique*. *Cité à la page 11*

[157] Lanfear, R., Thomas, J. A., Welch, J. J., Brey, T., and Bromham, L. 2007. Metabolic rate does not calibrate the molecular clock. *Proceedings of the National Academy of Sciences*, 104(39) : 15388–15393. *Cité à la page 109*

[158] Lanfear, R., Kokko, H., and Eyre-Walker, A. 2014. Population size and the rate of evolution. *Trends in ecology & evolution*, 29(1) : 33–41. *Cité à la page 179*

[159] Lanfear, R., Frandsen, P. B., Wright, A. M., Senfeld, T., and Calcott, B. 2017. Partitionfinder 2 : new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Molecular biology and evolution*, 34(3) : 772–773. *Cité à la page 80*

[160] Lartillot, N. 2006. Conjugate Gibbs sampling for Bayesian phylogenetic models. *J. Comput. Biol.*, 13(10) : 1701–1722. *Cité à la page 102*

[161] Lartillot, N. and Delsuc, F. 2012. Joint reconstruction of divergence times and life-history evolution in placental mammals using a phylogenetic covariance model. *Evolution*, 66(6) : 1773–1787. *Cité à la page 108, 116, 122*

[162] Lartillot, N. and Philippe, H. 2008. Improvement of molecular phylogenetic inference and the phylogeny of bilateria. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 363(1496) : 1463–1472. *Cité à la page 58*

[163] Lartillot, N. and Poujol, R. 2011. A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters. *Molecular biology and evolution*, 28(1) : 729–744. *Cité à la page 79, 101, 108, 116, 117, 180*

[164] Lartillot, N., Phillips, M. J., and Ronquist, F. 2016. A mixed relaxed clock model. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 371(1699). *Cité à la page 103*

[165] Latrille, T., Lanore, V., and Lartillot, N. 2021. Inferring long-term effective population size with mutation–selection models. *Molecular Biology and Evolution*, 38(10) : 4573–4587. *Cité à la page 186, 187*

[166] Latrille, T., Joseph, J., Hartasánchez, D., and Salamin, N. 2023. Mammalian protein-coding genes exhibit widespread beneficial mutations that are not adaptive. *Cité à la page 94*

[167] Latrille, T., Bastian, M., Gaboriau, T., and Salamin, N. 2024. Detecting diversifying selection for a trait from within and between-species genotypes and phenotypes. *Journal of Evolutionary Biology*, page voae084. *Cité à la page 93*

[168] Lefébure, T., Morvan, C., Malard, F., François, C., Konecny-Dupré, L., Guéguen, L., Weiss-Gayet, M., Seguin-Orlando, A., Ermini, L., Der Sarkissian, C., *et al.* 2017. Less effective selection leads to larger genomes. *Genome Research*, 27(6) : 1016–1028. *Cité à la page 110, 112, 113, 114, 116*

[169] Leffler, E. M., Bullaughey, K., Matute, D. R., Meyer, W. K., Ségurel, L., Venkat, A., Andolfatto, P., and Przeworski, M. 2012. Revisiting an old riddle : what determines genetic diversity levels within species ? *Cité à la page 51, 188*

[170] Lemey, P., Minin, V. N., Bielejec, F., Kosakovsky Pond, S. L., and Suchard, M. A. 2012. A counting renaissance : Combining stochastic mapping and empirical Bayes to quickly detect amino acid sites under positive selection. *Bioinformatics*. *Cité à la page 102*

[171] Leroi, A. M., Rose, M. R., and Lauder, G. V. 1994. What does the comparative method reveal about adaptation ? *The American Naturalist*, 143(3) : 381–402. *Cité à la page 98*

[172] Leroy, T., Rousselle, M., Tilak, M.-K., Caizergues, A. E., Scornavacca, C., Recuerda, M., Fuchs, J., Illera, J. C., De Swardt, D. H., Blanco, G., *et al.* 2021. Island songbirds as windows into evolution in small populations. *Current Biology*, 31(6) : 1303–1310. *Cité à la page 112*

[173] Levecque, K., Anseel, F., De Beuckelaer, A., Van der Heyden, J., and Gisle, L. 2017. Work organization and mental health problems in phd students. *Research policy*, 46(4) : 868–879. *Cité à la page 203*

[174] Lewin, R. 1996. *Patterns in evolution : the new molecular view. (p107 for the quote cited)*. (No Title). *Cité à la page 41*

[175] Lewontin, R. C. 1974. *The genetic basis of evolutionary change*. *Cité à la page 32, 51*

[176] Lewontin, R. C. and Hubby, J. L. 1966. A molecular approach to the study of genic heterozygosity in natural populations. ii. amount of variation and degree of heterozygosity in natural populations of drosophila pseudoobscura. *Genetics*, 54(2) : 595. *Cité à la page 28, 51*

[177] Li, H. and Durbin, R. 2011. Inference of human population history from individual whole-genome sequences. *Nature*, 475(7357) : 493–496. *Cité à la page 88, 89, 112, 118, 154, 177, 187*

[178] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Subgroup, . G. P. D. P. 2009. The sequence alignment/map format and samtools. *bioinformatics*, 25(16) : 2078–2079. *Cité à la page 188*

[179] Li, Z., Chen, Y., Mu, D., Yuan, J., Shi, Y., Zhang, H., Gan, J., Li, N., Hu, X., Liu, B., *et al.* 2012. Comparison of the two major classes of assembly algorithms : overlap–layout–consensus and de-bruijn-graph. *Briefings in functional genomics*, 11(1) : 25–37. *Cité à la page 62*

[180] Linné, C. v. 1753. *Species plantarum*. *Cité à la page 9*

[181] Linné, C. v. 1758. *Systema naturae*. 10 edition. *Cité à la page 9*

[182] Loewe, L. and Charlesworth, B. 2006. Inferring the distribution of mutational effects on fitness in drosophila. *Biology Letters*, 2(3) : 426–430. *Cité à la page 93, 94*

[183] Losos, J. B. 2011. Convergence, adaptation, and constraint. *Evolution*, 65(7) : 1827–1840. *Cité à la page 37*

[184] Löytynoja, A. 2014. Phylogeny-aware alignment with prank. *Multiple sequence alignment methods*, pages 155–170. *Cité à la page 68, 70, 188, 194*

[185] Löytynoja, A. and Goldman, N. 2008. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *science*, 320(5883) : 1632–1635. *Cité à la page 68*

[186] Löytynoja, A. and Goldman, N. 2009. Uniting alignments and trees. *Science*, 324(5934) : 1528–1529. *Cité à la page 68*

[187] Lynch, M. 2010. Evolution of the mutation rate. *TRENDS in Genetics*, 26(8) : 345–352. *Cité à la page 52*

[188] Lynch, M. 2011. The lower bound to the evolution of mutation rates. *Genome biology and evolution*, 3 : 1107–1118. *Cité à la page 52*

[189] Lynch, M. and Conery, J. S. 2003. The origins of genome complexity. *science*, 302(5649) : 1401–1404. *Cité à la page 97*

[190] Lynch, M. and Lande, R. 1998. The critical effective size for a genetically secure population. In *Animal Conservation forum*, volume 1, pages 70–72. Cambridge University Press. *Cité à la page 53*

[191] Lynch, M., Ali, F., Lin, T., Wang, Y., Ni, J., and Long, H. 2023. The divergence of mutation rates and spectra across the tree of life. *EMBO reports*, 24(10) : e57561. *Cité à la page 179*

[192] Maillet 1755. *Telliamed*. *Cité à la page 10*

[193] Mallet, J. 1995. A species definition for the modern synthesis. *Trends in Ecology & Evolution*, 10(7) : 294–299. *Cité à la page 81*

[194] Margoliash, E. and Smith, E. L. 1965. Structural and functional aspects of cytochrome c in relation to evolution. In *Evolving genes and proteins*, pages 221–242. Elsevier. *Cité à la page 36*

[195] Martin, A. P. and Palumbi, S. R. 1993. Body size, metabolic rate, generation time, and the molecular clock. *Proceedings of the National Academy of Sciences*, 90(9) : 4087–4091. *Cité à la page 109*

[196] Martins, E. P. 1994. Estimating the rate of phenotypic evolution from comparative data. *The American Naturalist*, 144(2) : 193–209. *Cité à la page 100*

[197] Martins, E. P. 2000. Adaptation and the comparative method. *Trends in Ecology & Evolution*, 15(7) : 296–299. *Cité à la page 98*

[198] Mateiu, L. and Rannala, B. 2006. Inferring complex DNA substitution processes on phylogenies using uniformization and data augmentation. *Syst. Biol.*, 55 : 259–269. *Cité à la page 102*

[199] Mather, K. 1964. Ra fisher's work in genetics. *Biometrics*, 20(2) : 330–342. *Cité à la page 19*

[200] Mayr, E. 1955. Integration of genotypes : synthesis. *Cité à la page 25, 26*

[201] Mayr, E. 1963. *Animal species and evolution*. Harvard University Press. *Cité à la page 25*

[202] Mayr, E. 1996. What is a species, and what is not ? *Philosophy of science*, 63(2) : 262–277. *Cité à la page 81*

[203] McDonald, J. H. and Kreitman, M. 1991. Adaptive protein evolution at the adh locus in drosophila. *Nature*, 351(6328) : 652–654. *Cité à la page 79, 91, 92*

[204] McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., *et al.* 2010. The genome analysis toolkit : a mapreduce framework for analyzing next-generation dna sequencing data. *Genome research*, 20(9) : 1297–1303. *Cité à la page 179*

[205] McVean, G. A. and Cardin, N. J. 2005. Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 360(1459) : 1387–1393. *Cité à la page 86*

[206] Mendel, G. 1865. Versuche uber pflanzen-hybriden. *Vorgelegt in den Sitzungen*. *Cité à la page 17, 24*

[207] Mendel, G. 1870. *Versuche über pflanzenhybriden*. Springer. *Cité à la page 17*

[208] Miller, C. and Swift, K. 2001. *The handbook of nonsexist writing*. IUniverse. *Cité à la page 206*

[209] Miller, D. I., Eagly, A. H., and Linn, M. C. 2015. Women's representation in science predicts national gender-science stereotypes : Evidence from 66 nations. *Journal of Educational Psychology*, 107(3) : 631. *Cité à la page 206*

[210] Moreira, D. and Philippe, H. 2000. Molecular phylogeny : pitfalls and progress. *International Microbiology*, 3(1) : 9–16. *Cité à la page 31*

[211] Morgan, G. J. 1998. Emile zuckerkandl, linus pauling, and the molecular evolutionary clock, 1959-1965. *Journal of the History of Biology*, pages 155–178. *Cité à la page 30, 38*

[212] Morgan, T. H. 1910. Sex limited inheritance in drosophila. *Science*, 32(812) : 120–122. *Cité à la page 23*

[213] Morgan, T. H., Sturtevant, A. H., Muller, H. J., and Bridges, C. B. 1915. The mechanism of mendelian heredity. *Cité à la page 19, 25*

[214] Moss, S. E. and Mahmoudi, M. 2021. Stem the bullying : an empirical investigation of abusive supervision in academic science. *EClinicalMedicine*, 40. *Cité à la page 202*

[215] Murrell, B., Wertheim, J. O., Moola, S., Weighill, T., Scheffler, K., and Kosakovsky Pond, S. L. 2012. Detecting individual sites subject to episodic diversifying selection. *PLoS genetics*, 8(7) : e1002764. *Cité à la page 80*

[216] Muse, S. V. and Gaut, B. S. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular biology and evolution*, 11(5) : 715–724. *Cité à la page 77*

[217] Nabholz, B., Glémin, S., and Galtier, N. 2008. Strong variations of mitochondrial mutation rate across mammals—the longevity hypothesis. *Molecular biology and evolution*, 25(1) : 120–130. *Cité à la page 108*

[218] Nabholz, B., Uwimana, N., and Lartillot, N. 2013. Reconstructing the phylogenetic history of long-term effective population size and life-history traits using patterns of amino acid replacement in mitochondrial genomes of mammals and birds. *Genome biology and evolution*, 5(7) : 1273–1290. *Cité à la page 109*

[219] Nei, M. and Li, W.-H. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences*, 76(10) : 5269–5273. *Cité à la page 82*

[220] Nickerson, R. S. 2000. Null hypothesis significance testing : a review of an old and continuing controversy. *Psychological methods*, 5(2) : 241. *Cité à la page 197*

[221] Nielsen, R. 2005. Molecular signatures of natural selection. *Annu. Rev. Genet.*, 39(1) : 197–218. *Cité à la page 90*

[222] Nielsen, R. and Yang, Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the hiv-1 envelope gene. *Genetics*, 148(3) : 929–936. *Cité à la page 77, 78*

[223] Nikolaev, S. I., Montoya-Burgos, J. I., Popadin, K., Parand, L., Margulies, E. H., of Health Intramural Sequencing Center Comparative Sequencing Program, N. I., and Antonarakis, S. E. 2007. Life-history traits drive the evolutionary rates of mammalian coding and noncoding genomic elements. *Proceedings of the National Academy of Sciences*, 104(51) : 20443–20448. *Cité à la page 108*

[224] Nishimura, D. 2000. Repeatmasker. *Biotech Software & Internet Report*, 1(1-2) : 36–39. *Cité à la page 64*

[225] Ohta, T. 1973. Slightly deleterious mutant substitutions in evolution. *Nature*, 246(5428) : 96–98. *Cité à la page 43, 44, 45, 93, 176*

[226] Ohta, T. 1974. Mutational pressure as the main cause of molecular evolution and polymorphism. *Nature*, 252(5482) : 351–354. *Cité à la page 44, 176*

[227] Ohta, T. 1992. The nearly neutral theory of molecular evolution. *Annual review of ecology and systematics*, 23(1) : 263–286. *Cité à la page 106*

[228] Ohta, T. 1993. An examination of the generation-time effect on molecular evolution. *Proceedings of the National Academy of Sciences*, 90(22) : 10676–10680. *Cité à la page 46, 109*

[229] Owen, R. 1843. *Lectures on the comparative anatomy and physiology of the invertebrate animals*. *Cité à la page 10*

[230] Pagel, M. 1993. Seeking the evolutionary regression coefficient : an analysis of what comparative methods measure. *Journal of theoretical Biology*, 164(2) : 191–205. *Cité à la page 100*

[231] Pagel, M. 1997. Inferring evolutionary processes from phylogenies. *Zoologica Scripta*, 26(4) : 331–348. *Cité à la page 100*

[232] Pearson, K. 1900. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302) : 157–175. *Cité à la page 18*

[233] Philippe, H., Brinkmann, H., Lavrov, D. V., Littlewood, D. T. J., Manuel, M., Wörheide, G., and Baurain, D. 2011. Resolving difficult phylogenetic questions : why more sequences are not enough. *PLoS biology*, 9(3) : e1000602. *Cité à la page 31*

[234] Piganeau, G. and Eyre-Walker, A. 2009. Evidence for variation in the effective population size of animal mitochondrial dna. *PloS one*, 4(2) : e4396. *Cité à la page 110*

[235] Pluzhnikov, A. and Donnelly, P. 1996. Optimal sequencing strategies for surveying molecular genetic diversity. *Genetics*, 144(3) : 1247–1262. *Cité à la page 83*

[236] Popadin, K., Polishchuk, L. V., Mamirova, L., Knorre, D., and Gunbin, K. 2007. Accumulation of slightly deleterious mutations in mitochondrial protein-coding genes of large versus small mammals. *Proceedings of the National Academy of Sciences*, 104(33) : 13390–13395. *Cité à la page 108, 109, 116*

[237] Portin, P. 2002. Historical development of the concept of the gene. In *The Journal of Medicine and Philosophy : A Forum for Bioethics and Philosophy of Medicine*, volume 27, pages 257–286. Journal of Medicine and Philosophy Inc. *Cité à la page 19*

[238] Provine, W. B. 1970. *Origin of the theoretical population genetics*. Ph.D. thesis, The University of Chicago. *Cité à la page 19, 25*

[239] Pupko, T. and Mayrose, I. 2020. A gentle introduction to probabilistic evolutionary models. *phylogenetics in the genomic era*, pages 1–1. *Cité à la page 31*

[240] RA, G. 2004. Genome sequence of the brown norway rat yields insights into mammalian evolution. *Nature*, 428(6982) : 493–521. *Cité à la page 77*

[241] Rand, D. M. and Kann, L. M. 1996. Excess amino acid polymorphism in mitochondrial dna : contrasts among genes from drosophila, mice, and humans. *Molecular biology and evolution*, 13(6) : 735–748. *Cité à la page 93*

[242] Rannala, B. and Yang, Z. 2007. Inferring speciation times under an episodic molecular clock. *Systematic biology*, 56(3) : 453–466. *Cité à la page 101*

[243] Ranwez, V. and Chantret, N. N. 2020. Strengths and limits of multiple sequence alignment and filtering methods. *Cité à la page 68, 69*

[244] Ranwez, V., Harispe, S., Delsuc, F., and Douzery, E. J. 2011. Macse : Multiple alignment of coding sequences accounting for frameshifts and stop codons. *PloS one*, 6(9) : e22594. *Cité à la page 67, 70, 195*

[245] Rensch, B. 1959. *Evolution above the species level*. Columbia University Press. *Cité à la page 38*

[246] Rhoads, A. and Au, K. F. 2015. Pacbio sequencing and its applications. *Genomics, Proteomics and Bioinformatics*, 13(5) : 278–289. *Cité à la page 60*

[247] Ridley, J., Kolm, N., Freckelton, R., and Gage, M. 2007. An unexpected influence of widely used significance thresholds on the distribution of reported p-values. *Journal of evolutionary biology*, 20(3) : 1082–1089. *Cité à la page 197*

[248] Roberts, D. B. 2006. Drosophila melanogaster : the model organism. *Entomologia experimentalis et applicata*, 121(2) : 93–103. *Cité à la page 23*

[249] Rodrigue, N., Philippe, H., and Lartillot, N. 2010. Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. *Proceedings of the National Academy of Sciences*, 107(10) : 4629–4634. *Cité à la page 185, 192*

[250] Romiguier, J., Figuet, E., Galtier, N., Douzery, E. J. P., Boussau, B., Dutheil, J. Y., and Ranwez, V. 2012. Fast and robust characterization of time-heterogeneous sequence evolutionary processes using substitution mapping. *PLoS One*, 7(3) : e33852. *Cité à la page 102*

[251] Romiguier, J., Gayral, P., Ballenghien, M., Bernard, A., Cahais, V., Chenuil, A., Chiari, Y., Dernat, R., Duret, L., Faivre, N., *et al.* 2014a. Comparative population genomics in animals uncovers the determinants of genetic diversity. *Nature*, 515(7526) : 261–263. *Cité à la page 52, 111, 116*

[252] Romiguier, J., Lourenco, J., Gayral, P., Faivre, N., Weinert, L. A., Ravel, S., Ballenghien, M., Cahais, V., Bernard, A., Loire, E., *et al.* 2014b. Population genomics of eusocial insects : the costs of a vertebrate-like effective population size. *Journal of Evolutionary Biology*, 27(3) : 593–603. *Cité à la page 114*

[253] Ronquist, F., Teslenko, M., Van Der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A., and Huelsenbeck, J. P. 2012. Mrbayes 3.2 : efficient bayesian phylogenetic inference and model choice across a large model space. *Systematic biology*, 61(3) : 539–542. *Cité à la page 78*

[254] Sageret, A. 1826. Considérations sur la production des hybrides, des variantes et des variétés en général, et sur celles de la famille des cucurbitacées en particulier. In *Annales des sciences naturelles*, volume 8, page 294. *Cité à la page 17*

[255] Saiki, R. K., Scharf, S., Faloona, F., Mullis, K. B., Horn, G. T., Erlich, H. A., and Arnheim, N. 1985. Enzymatic amplification of β-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science*, 230(4732) : 1350–1354. *Cité à la page 60*

[256] Saitou, N. and Nei, M. 1987. The neighbor-joining method : a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, 4(4) : 406–425. *Cité à la page 31*

[257] Sanger, F., Nicklen, S., and Coulson, A. R. 1977. Dna sequencing with chain-terminating inhibitors. *Proceedings of the national academy of sciences*, 74(12) : 5463–5467. *Cité à la page 28, 60*

[258] Satinsky, E. N., Kimura, T., Kiang, M. V., Abebe, R., Cunningham, S., Lee, H., Lin, X., Liu, C. H., Rudan, I., Sen, S., *et al.* 2021. Systematic review and meta-analysis of depression, anxiety, and suicidal ideation among ph. d. students. *Scientific Reports*, 11(1) : 14370. *Cité à la page 203*

[259] Schlee, D. 1975. Numerical taxonomy. the principles and practice of numerical classification. *Cité à la page 30*

[260] Scornavacca, C., Belkhir, K., Lopez, J., Dernat, R., Delsuc, F., Douzery, E. J., and Ranwez, V. 2019. Orthomam v10 : scaling-up orthologous coding sequence and exon alignments with more than one hundred mammalian genomes. *Molecular Biology and Evolution*, 36(4) : 861–862. *Cité à la page 72, 188*

[261] Seeber, M., Cattaneo, M., and Birolini, S. (). Academic publishing business models : self-citations and the selectivity-reputation trade-off. *(No Title)*. *Cité à la page 202*

[262] Seidel, H. S., Rockman, M. V., and Kruglyak, L. 2008. Widespread genetic incompatibility in c. elegans maintained by balancing selection. *science*, 319(5863) : 589–594. *Cité à la page 53*

[263] Sella, G., Petrov, D. A., Przeworski, M., and Andolfatto, P. 2009. Pervasive natural selection in the drosophila genome ? *PLoS genetics*, 5(6) : e1000495. *Cité à la page 52*

[264] Seo, T.-K., Kishino, H., and Thorne, J. L. 2004. Estimating absolute rates of synonymous and nonsynonymous nucleotide substitution in order to characterize natural selection and date species divergences. *Molecular biology and evolution*, 21(7) : 1201–1213. *Cité à la page 101, 102*

[265] Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. 2015. Busco : assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19) : 3210–3212. *Cité à la page 62, 64, 122, 188, 189*

[266] Simion, P., Delsuc, F., and Philippe, H. 2020. To what extent current limits of phylogenomics can be overcome ? *Phylogenetics in the genomic era*, pages 2–1. *Cité à la page 58, 59*

[267] Simpson, G. 1944. Tempo and mode in evolution columbia univ. *Press, New York.* *Cité à la page 38*

[268] Simpson, G. G. 1953. *The major features of evolution*. Columbia University Press. *Cité à la page 38*

[269] Sjodin, P., Kaj, I., Krone, S., Lascoux, M., and Nordborg, M. 2005. On the meaning and existence of an effective population size. *Genetics*, 169(2) : 1061–1070. *Cité à la page 48*

[270] Smith, J., Maynard, and Haigh, J. 1974. The hitch-hiking effect of a favourable gene. *Genetics Research*, 23(1) : 23–35. *Cité à la page 52*

[271] Smith, J. M. and Smith, N. 1996. Synonymous nucleotide divergence : what is "saturation" ? *Genetics*, 142(3) : 1033–1036. *Cité à la page 75, 81*

[272] Smith, N. G. and Eyre-Walker, A. 2002. Adaptive protein evolution in drosophila. *Nature*, 415(6875) : 1022–1024. *Cité à la page 92, 93, 94*

[273] Sneath, P. 1977. A method for testing the distinctness of clusters : a test of the disjunction of two clusters in euclidean space as measured by their overlap. *Journal of the International Association for Mathematical Geology*, 9 : 123–143. *Cité à la page 30*

[274] Sokal, R. R., Camin, J. H., Rohlf, F. J., and Sneath, P. H. A. 1965. Numerical Taxonomy : Some Points of View. *Systematic Biology*, 14(3) : 237–243. *Cité à la page 30*

[275] Spencer, H. 1864. *A system of synthetic philosophy*, volume 2. *Cité à la page 16*

[276] Stanke, M. and Morgenstern, B. 2005. Augustus : a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic acids research*, 33(suppl_2) : W465–W467. *Cité à la page 64*

[277] Stephan, W. and Langley, C. H. 1992. Evolutionary consequences of dna mismatch inhibited repair opportunity. *Genetics*, 132(2) : 567–574. *Cité à la page 53*

[278] Strimmer, K., von Haeseler, A., Salemi, A.-M., *et al.* 2003. Nucleotide substitution models. *The Phylogenetics Handbook A Practical Approach to DNA and Protein Phylogeny*, pages 72–100. *Cité à la page 31*

[279] Suchard, M. A. and Redelings, B. D. 2006. Bali-phy : simultaneous bayesian inference of alignment and phylogeny. *Bioinformatics*, 22(16) : 2047–2048. *Cité à la page 67*

[280] Tajima, F. 1983. Evolutionary relationship of dna sequences in finite populations. *Genetics*, 105(2) : 437–460. *Cité à la page 83, 85*

[281] Tajima, F. 1989. The effect of change in population size on dna polymorphism. *Genetics*, 123(3) : 597–601. *Cité à la page 88*

[282] Tajima, F. 1996. The amount of dna polymorphism maintained in a finite population when the neutral mutation rate varies among sites. *Genetics*, 143(3) : 1457–1465. *Cité à la page 83*

[283] Talavera, G. and Castresana, J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic biology*, 56(4) : 564–577. *Cité à la page 69*

[284] Tavaré, S. 1986. Some probabilistic and statistical problems on the analysis of dna sequence. *Lecture of Mathematics for Life Science*, 17 : 57. *Cité à la page 58, 75*

[285] Templeton, A. R. 1996. Contingency tests of neutrality using intra/interspecific gene trees : the rejection of neutrality for the evolution of the mitochondrial cytochrome oxidase ii gene in the hominoid primates. *Genetics*, 144(3) : 1263–1270. *Cité à la page 94*

[286] Terhorst, J., Kamm, J. A., and Song, Y. S. 2017. Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nature genetics*, 49(2) : 303–309. *Cité à la page 177*

[287] The UniProt Consortium 2022. UniProt : the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*, 51(D1) : D523–D531. *Cité à la page 72*

[288] Thompson, J. D., Gibson, T. J., and Higgins, D. G. 2003. Multiple sequence alignment using clustalw and clustalx. *Current protocols in bioinformatics*, (1) : 2–3. *Cité à la page 195*

[289] Thorne, J. L., Kishino, H., and Painter, I. S. 1998. Estimating the rate of evolution of the rate of molecular evolution. *Molecular biology and evolution*, 15(12) : 1647–1657. *Cité à la page 101*

[290] Théophraste -314. *Histoire des plantes*. *Cité à la page 8*

[291] Tomoko, O. 1995. Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. *Journal of molecular evolution*, 40 : 56–63. *Cité à la page 188*

[292] Tschermak, E. 1900. *Über künstliche Kreuzung bei Pisum sativum*. E. Tschermak. *Cité à la page 18*

[293] Uscinski, J., Enders, A., Klofstad, C., Seelig, M., Drochon, H., Premaratne, K., and Murthi, M. 2022. Have beliefs in conspiracy theories increased over time ? *PLoS One*, 17(7) : e0270429. *Cité à la page 205*

[294] Van den Berg, J., van Ooyen, A., Mantei, N., Schamböck, A., Grosveld, G., Flavell, R. A., and Weissmann, C. 1978. Comparison of cloned rabbit and mouse β-globin genes showing strong evolutionary divergence of two homologous pairs of introns. *Nature*, 276(5683) : 37–44. *Cité à la page 36*

[295] Van der Linden, S. L., Leiserowitz, A. A., Feinberg, G. D., and Maibach, E. W. 2015. The scientific consensus on climate change as a gateway belief : Experimental evidence. *PloS one*, 10(2) : e0118489. *Cité à la page 205*

[296] Visscher, P. M. and Goddard, M. E. 2019. From ra fisher's 1918 paper to gwas a century later. *Genetics*, 211(4) : 1125–1130. *Cité à la page 19, 26*

[297] Waddington, C. 1957. The strategy of the genes (london : Volume george allen and unwin). *Cité à la page 25*

[298] Wang, Y., Zhao, Y., Bollas, A., Wang, Y., and Au, K. F. 2021. Nanopore sequencing technology, bioinformatics and applications. *Nature biotechnology*, 39(11) : 1348–1365. *Cité à la page 60*

[299] Waples, R. S. 2016. Making sense of genetic estimates of effective population size. *Cité à la page 47*

[300] Waples, R. S. 2022. What is n e, anyway ? *Journal of Heredity*, 113(4) : 371–379. *Cité à la page 47, 48*

[301] Watterson, G. 1975. On the number of segregating sites in genetical models without recombination. *Theoretical population biology*, 7(2) : 256–276. *Cité à la page 82*

[302] Weinberg, W. 1908. Über vererbungsgesetze beim menschen. *Zeitschrift für induktive Abstammungs-und Vererbungslehre*, 1(1) : 377–392. *Cité à la page 23, 50*

[303] Weisblum, B., Benzer, S., and Holley, R. W. 1962. A physical basis for degeneracy in the amino acid code. *Proceedings of the National Academy of Sciences*, 48(8) : 1449–1454. *Cité à la page 30*

[304] Weismann, A. 1893. *The germ-plasm : a theory of heredity*. Scribner's. *Cité à la page 17*

[305] Welch, J. J. and Waxman, D. 2008. Calculating independent contrasts for the comparative study of substitution rates. *Journal of Theoretical Biology*, 251(4) : 667–678. *Cité à la page 101*

[306] Welch, J. J., Eyre-Walker, A., and Waxman, D. 2008. Divergence and polymorphism under the nearly neutral theory of molecular evolution. *Journal of molecular evolution*, 67 : 418–426. *Cité à la page 94, 110, 115, 184*

[307] Weyna, A. and Romiguier, J. 2021. Relaxation of purifying selection suggests low effective population size in eusocial hymenoptera and solitary pollinating bees. *Peer Community Journal*, 1. *Cité à la page 114*

[308] Whelan, S. and Goldman, N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular biology and evolution*, 18(5) : 691–699. *Cité à la page 80*

[309] Whitney, K. D. and Garland Jr, T. 2010. Did genetic drift drive increases in genome complexity ? *PLoS genetics*, 6(8) : e1001080. *Cité à la page 97*

[310] Whitney, K. D., Boussau, B., Baack, E. J., and Garland Jr, T. 2011. Drift and genome complexity revisited. *PLoS genetics*, 7(6) : e1002092. *Cité à la page 97*

[311] Wiuf, C. and Hein, J. 1999. Recombination as a point process along sequences. *Theoretical population biology*, 55(3) : 248–259. *Cité à la page 86*

[312] Woese, C. R. 1987. Bacterial evolution. *Microbiological reviews*, 51(2) : 221–271. *Cité à la page 14, 39*

[313] Wong, K. M., Suchard, M. A., and Huelsenbeck, J. P. 2008. Alignment uncertainty and genomic analysis. *Science*, 319(5862) : 473–476. *Cité à la page 68*

[314] Woolston, C. 2019. Phds : the tortuous truth. *Nature*, 575(7782) : 403–407. *Cité à la page 202*

[315] Wright, S. 1931. Evolution in mendelian populations. *Genetics*, 16(2) : 97. *Cité à la page 23, 24, 33, 47, 50, 52*

[316] Wright, S. 1951. Fisher and ford on" the sewall wright effect". *American Scientist*, 39(3) : 452–479. *Cité à la page 24*

[317] Wright, S. 1970. Random drift and the shifting balance theory of evolution. In *Mathematical topics in population genetics*, pages 1–31. Springer. *Cité à la page 24*

[318] Wright, S. 1984. *Evolution and the genetics of populations, volume 3 : experimental results and evolutionary deductions*, volume 3. University of Chicago press. *Cité à la page 24*

[319] Wright, S. *et al.* 1932. The roles of mutation, inbreeding, crossbreeding, and selection in evolution. *Cité à la page 24, 184*

[320] Yang, L. and Gaut, B. S. 2011. Factors that contribute to variation in evolutionary rate among arabidopsis genes. *Molecular Biology and Evolution*, 28(8) : 2359–2369. *Cité à la page 77*

[321] Yang, Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Molecular biology and evolution*, 15(5) : 568–573. *Cité à la page 79*

[322] Yang, Z. 2007. Paml 4 : phylogenetic analysis by maximum likelihood. *Molecular biology and evolution*, 24(8) : 1586–1591. *Cité à la page 78*

[323] Yang, Z. and Bielawski, J. P. 2000. Statistical methods for detecting molecular adaptation. *Trends in ecology & evolution*, 15(12) : 496–503. *Cité à la page 78*

[324] Yang, Z. *et al.* 1997. Paml : a program package for phylogenetic analysis by maximum likelihood. *Computer applications in the biosciences*, 13(5) : 555–556. *Cité à la page 78*

[325] Yang, Z. and Nielsen, R. 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *Journal of molecular evolution*, 46 : 409–418. *Cité à la page 79*

[326] Yang, Z. and Nielsen, R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Molecular biology and evolution*, 19(6) : 908–917. *Cité à la page 80*

[327] Yang, Z. and Nielsen, R. 2008. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Molecular biology and evolution*, 25(3) : 568–579. *Cité à la page 185*

[328] Yoder, A. D. and Tiley, G. P. 2021. The challenge and promise of estimating the de novo mutation rate from whole-genome comparisons among closely related individuals. *Molecular Ecology*, 30(23) : 6087–6100. *Cité à la page 52*

[329] Zerbino, D. R. 2010. Using the velvet de novo assembler for short-read sequencing technologies. *Current protocols in bioinformatics*, 31(1) : 11–5. *Cité à la page 62*

[330] Zhang, J., Nielsen, R., and Yang, Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Molecular biology and evolution*, 22(12) : 2472–2479. *Cité à la page 80*

[331] Zoonomia 2020. A comparative genomics multitool for scientific discovery and conservation. *Nature*, 587(7833) : 240–245. *Cité à la page 121, 122, 188, 190*

[332] Zuckerkandl, E. 1962. Molecular disease, evolution, and genic heterogeneity. *Horizons in biochemistry*, pages 189–225. *Cité à la page 39, 66*

[333] Zuckerkandl, E. 1987. On the molecular evolutionary clock. *Journal of molecular evolution*, 26 : 34–46. *Cité à la page 38, 40*

[334] Zuckerkandl, E. and Pauling, L. 1965a. Evolutionary divergence and convergence in proteins. In *Evolving genes and proteins*, pages 97–166. Elsevier. *Cité à la page 36*

[335] Zuckerkandl, E. and Pauling, L. 1965b. Molecules as documents of evolutionary history. *Journal of theoretical biology*, 8(2) : 357–366. *Cité à la page 29, 30*

[336] Zuckerkandl, E. and Schroeder, W. A. 1961. Amino-acid composition of the polypeptide chains of gorilla haemoglobin. *Nature*, 192(4806) : 984–985. *Cité à la page 38*

[337] Zuckerkandl, E., Jones, R. T., and Pauling, L. 1960. A comparison of animal hemoglobins by tryptic peptide pattern analysis. *Proceedings of the National Academy of Sciences*, 46(10) : 1349–1360. *Cité à la page 38*

Références générales et épistémologiques

En gras, les ouvrages particulièrement importants utilisés lors de ma thèse et pour sa rédaction

[1] Buican, D. (2008). *Mendel dans l'histoire de la génétique*. Ellipses.

[2] **Hahn, M. W. (2018). *Molecular Population Genetics*. Royaume-Uni : Sinauer Associates.**

[3] **Hervé, M. (2020). *Systématique animale : D'Aristote aux phylogénies moléculaires*. De Boeck Supérieur.**

[4] **Kimura, M. (1990). *Théorie neutraliste de l'évolution*. France : Flammarion.** Traduite de l'anglais : Kimura, M. (1985). *The neutral theory of molecular evolution*. Cambridge University Press.

[5] Lecointre, G. (2015). *Le Monde de Darwin*. (n.p.) : DLM.

[6] Miquel, P. (2008). *Biologie du XXIe siècle : évolution des concepts fondateurs*. Belgique : De Boeck Supérieur.

[7] Pigeaud, article sur le livre *Telliamed* de Maillet, 1755 dans la revue *Pour la science* 2023, https ://www.pourlascience.fr/sr/livres/telliamed-25921.php

[8] **Celine Scornavacca, Frédéric Delsuc, Nicolas Galtier. (2021). *Phylogenetics in the Genomic Era*. | Authors open access book, p.p. 1-568, 2020, 978-2-9575069-0-3. ⟨hal-02535070v3⟩**

[9] **Thomas, F., Lefevre, T., Raymond, M. (2016). *Biologie évolutive*. Belgique : De Boeck supérieur.**

Cinquième partie

Annexes

9 Annexes

9.1 Annexe article 1

5 Supplementary Material

5.1 VCF filtering

We annotated the 144 VCF files with a Python script that relates the coding exon positions of the non-filtered 7726 genes from the Busco annotation pipeline to the coordinates in the VCF files and in the genome assembly. We then obtained the coding synonymous, coding non-synonymous or non-coding nature of the SNP, the name of the corresponding Busco gene and some quality measures for each SNP. We focused on the GQ and QUAL metrics. The QUAL metrics gives the probability that a site has no variant (a false heterozygote). The GQ metric is about the probability that a variant is incorrect (the nucleotide assigned is wrong, but there is heterozygosity). Both metrics have the same unit : for example, QUAL=20: 1 % chance that there is no variant at the site. These two metrics are relevant to our need to having a true variant site with the good call, because of their attribution to synonymous or non-synonymous variant.

We choose a strong filtering on these metrics (QUAL$>$ 125 and GQ$>$ 150) and applied it to all species, regardless of their genome or calling quality, even if some of them end up conserving very few SNPs (6 species have less than 1000 SNPs: *Muscardinus avellanarius*, *Beatragus hunteri*, *Sigmodon hispidus*, *Diceros bicornis*, *Sousa chinensis* and *Alouatta palliata*).

The VCFs files also provides us with the number of ”reference” and ”alternative” reads per SNP. This allows us to have a look at the distribution of the frequency of the reference allele, after the VCF filtering above, computed by dividing the number of reference reads by the total number of reads for this SNP. In the case of our VCF, the reference genome correspond to the individual on which we perform variant calling. The ’reference’ allele called is simply the first read seen in the call and doesn’t have any real meaning. Therefore, we calculate the locus frequency using randomly either the ”reference” or the ”derived” allele. We expect the frequency distribution to be unimodal with a pic around 0.5 so we removed the SNP with a frequency lower to 0.2 or higher to 0.8.

We graphically compute this distribution of allelic balance and detect six species with a distribution different from the expected unimodal centred around 0.5. The six species are *Acomys cahirinus*, *Przewalskium albirostris*, *Mastomys coucha*, *Litocranius walleri*, *Cheirogaleus medius* and *Cephalophus harveyi*. These six VCFs are removed from the analysis. Figure S1 shows an example for a species with an allelic balance distribution that matches our expectation and two species with an abnormal distribution in two different manner. We compare these distributions to a binomial distribution where for each SNP we draw a value from 0 to 1 using the total number of reads for that SNP, centred at 0.5.

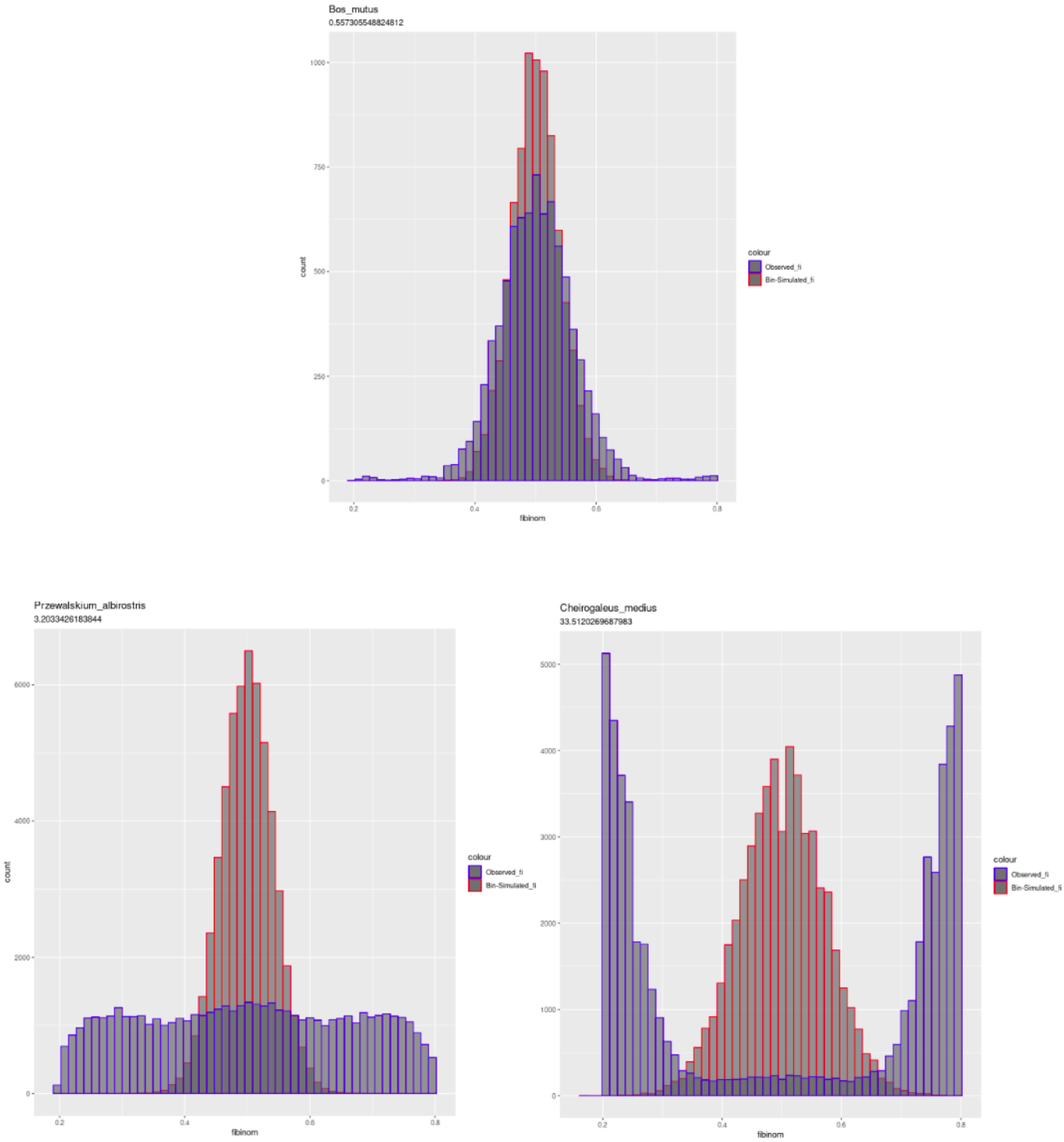


Figure S1: **Allelic balance distribution from three example**. In blue, distribution of locus frequencies (fi) for one species (*Bos mutus*, top) corresponding to our unimodal centred around 0.5 expectation and two outliers species (*Cheirogaleus medius*, bottom right and *Przewalskium albirostris*, bottom left). The red histograms correspond to what would be expected from a binomial sampling centred at 0.5.

5.2 d_N/d_S reconstruction by FastCoevol using the 89 species tree

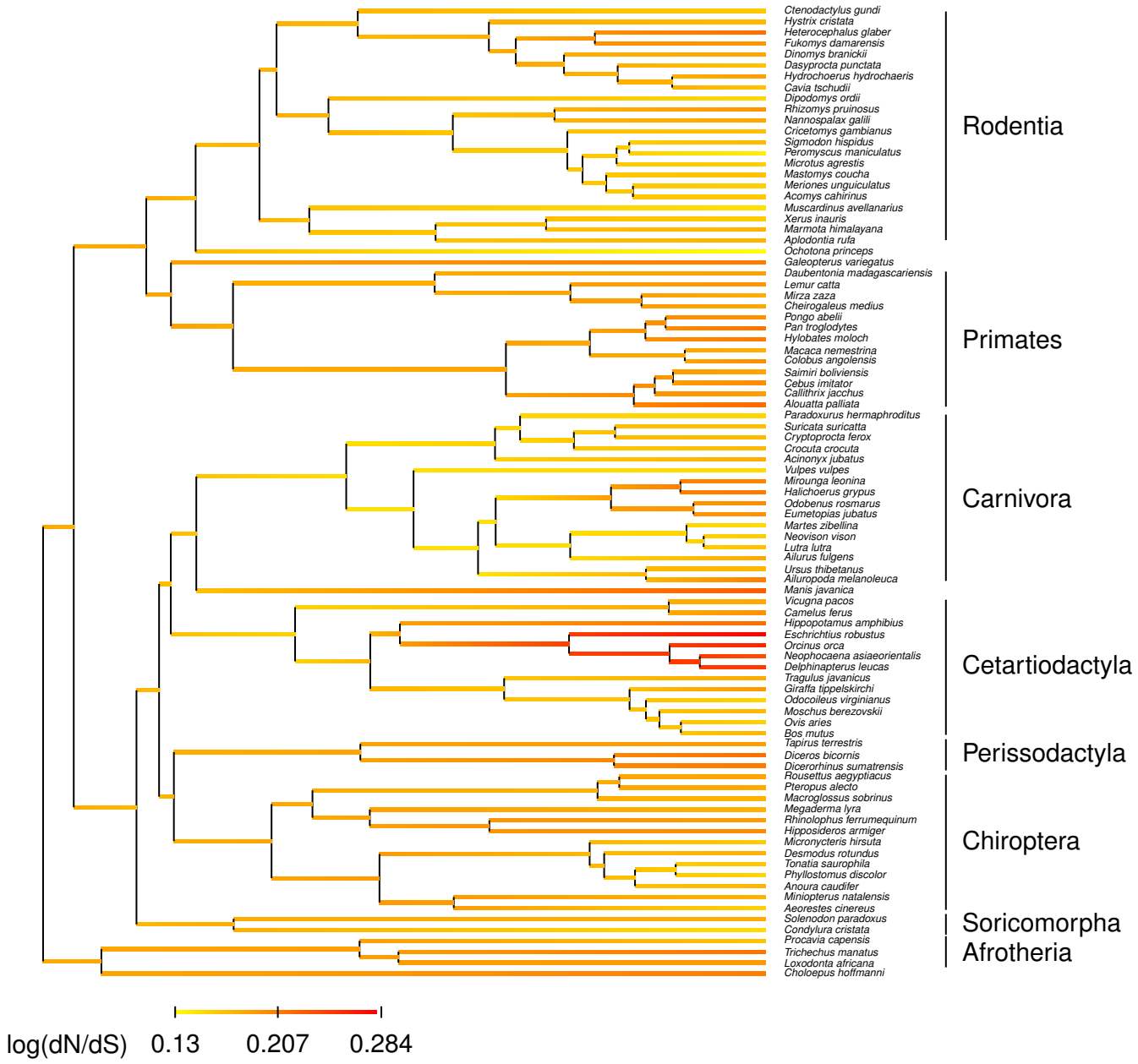


Figure S2: Reconstruction of $\log(d_N/d_S)$ along the reduced 89 mammalian species tree by FastCoevol

5.3 Phylogenies used as input for FastCoevol

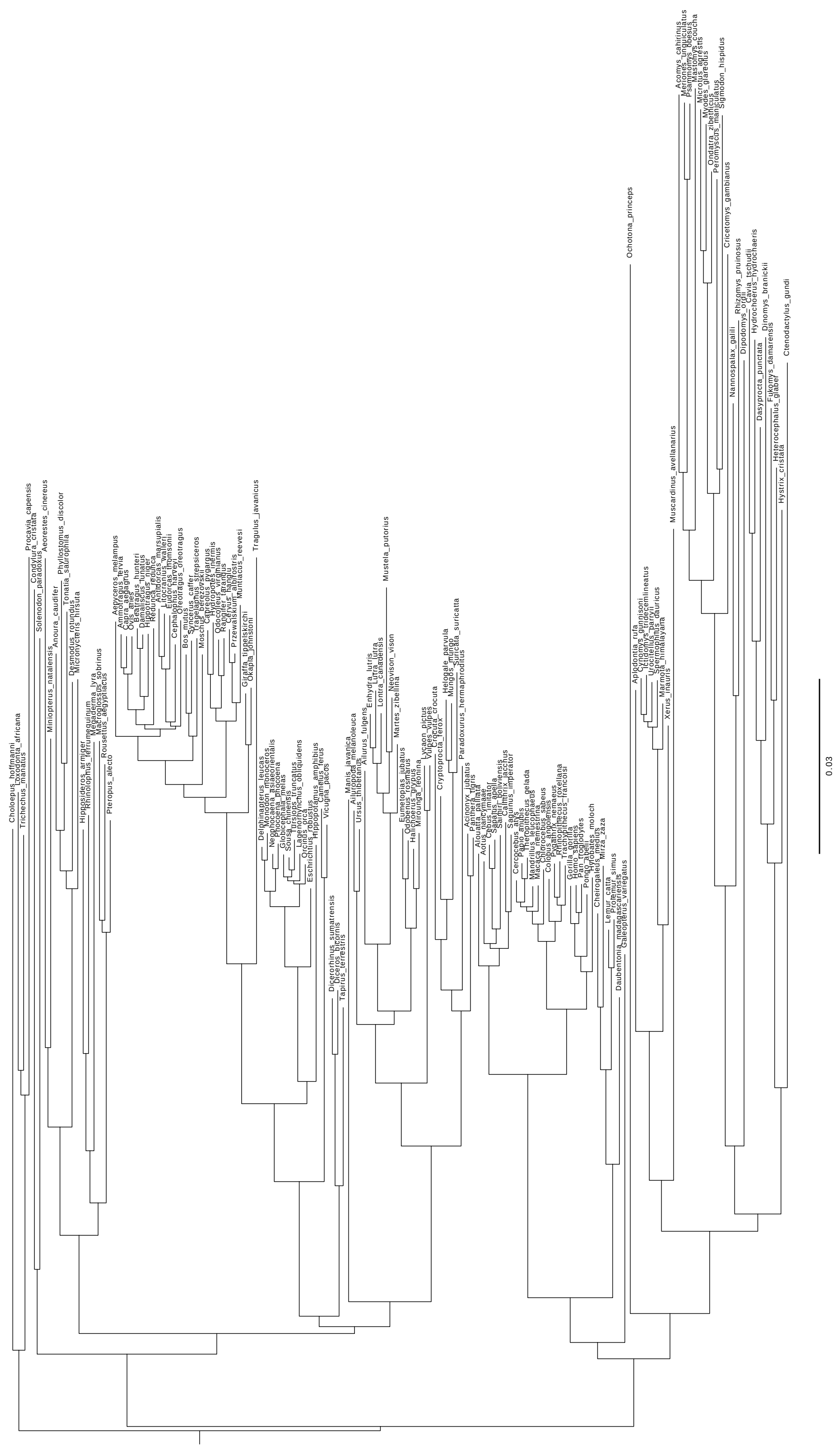


Figure S3: 144 mammals species phylogeny reconstructed using Iqtree

Figure S4: 89 mammals species phylogeny reconstructed using Iqtree

5.4 Species sampling effect on the d_S and d_N/d_S estimation

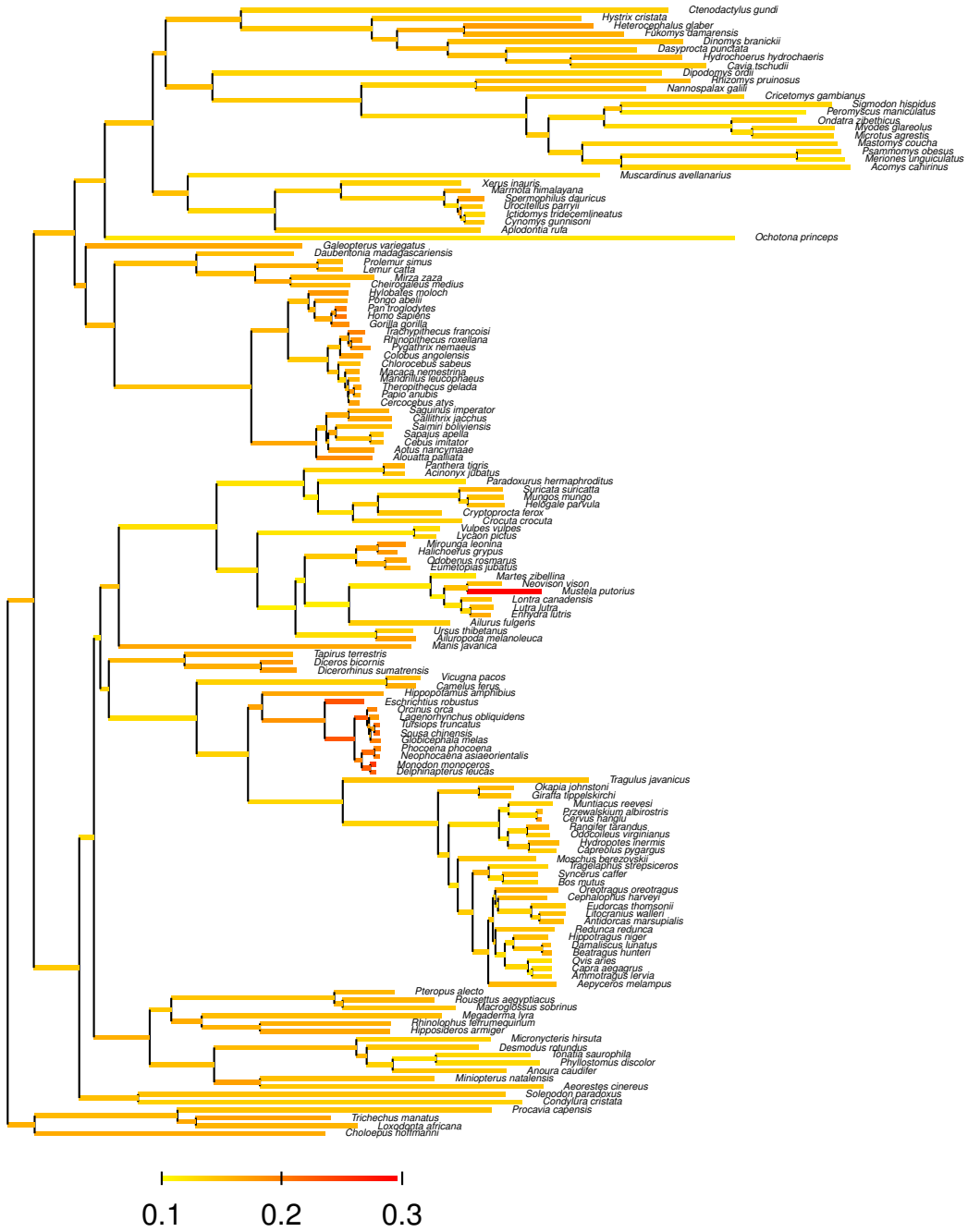


Figure S5: empirical d_S-tree with branches colored as a function of empirical d_N/d_S on the 144 species subset

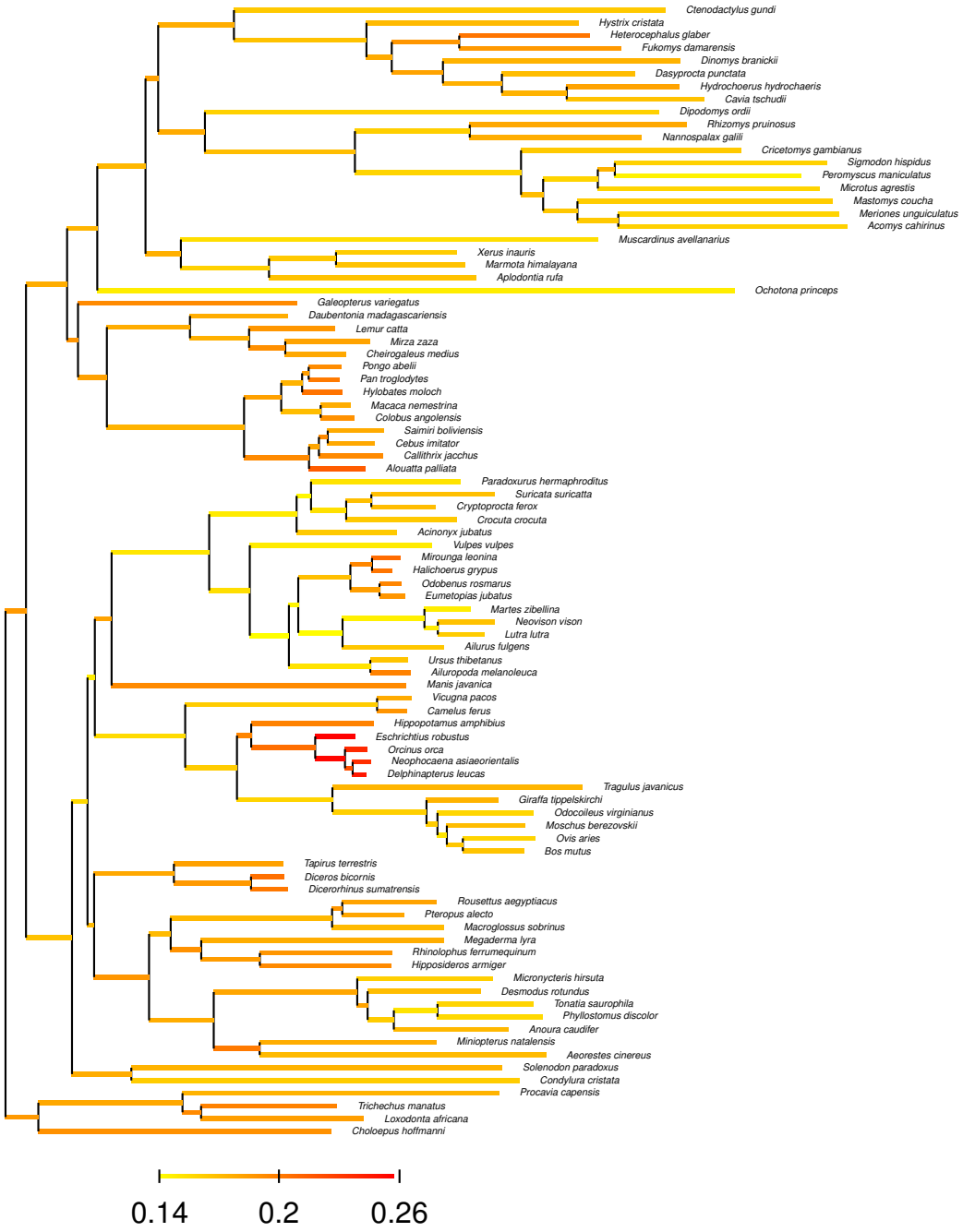


Figure S6: Empirical d_S-tree with branches colored as a function of empirical d_N/d_S on the 89 species subset

5.5 ROH study

For some species in our study, we note very few remaining SNP after quality filtering, resulting in homozygous regions in the genome. Long homozygous regions in a genome are signes of inbreeding level.

We used vcftools (Danecek *et al.*, 2011) to cut the genomes in 10kb windows and count the number of callable positions and the number of qualitative SNP. We consider only the windows containing more than 8kb callable over the 10kb and compute a density in SNP by windows. We did the same analysis on 50kb windows restricted to at least 40kb callable SNP. We then compute for each specie the mean and variance in the density distribution for the two windows size. We plot the distribution of those mean and, even if their density is low, we didn't observe a strong detachment of our six flagged species from the global trend (Figure S7, S8).

Moreover, we observe a high variance of the number of SNP per windows, even at 50kb windows scale, without any particular outlying of our suspected inbreeding species. These undecisives results didn't allow us to consider those species as inbred, and we decided not to remove them from our analysis for that reason.

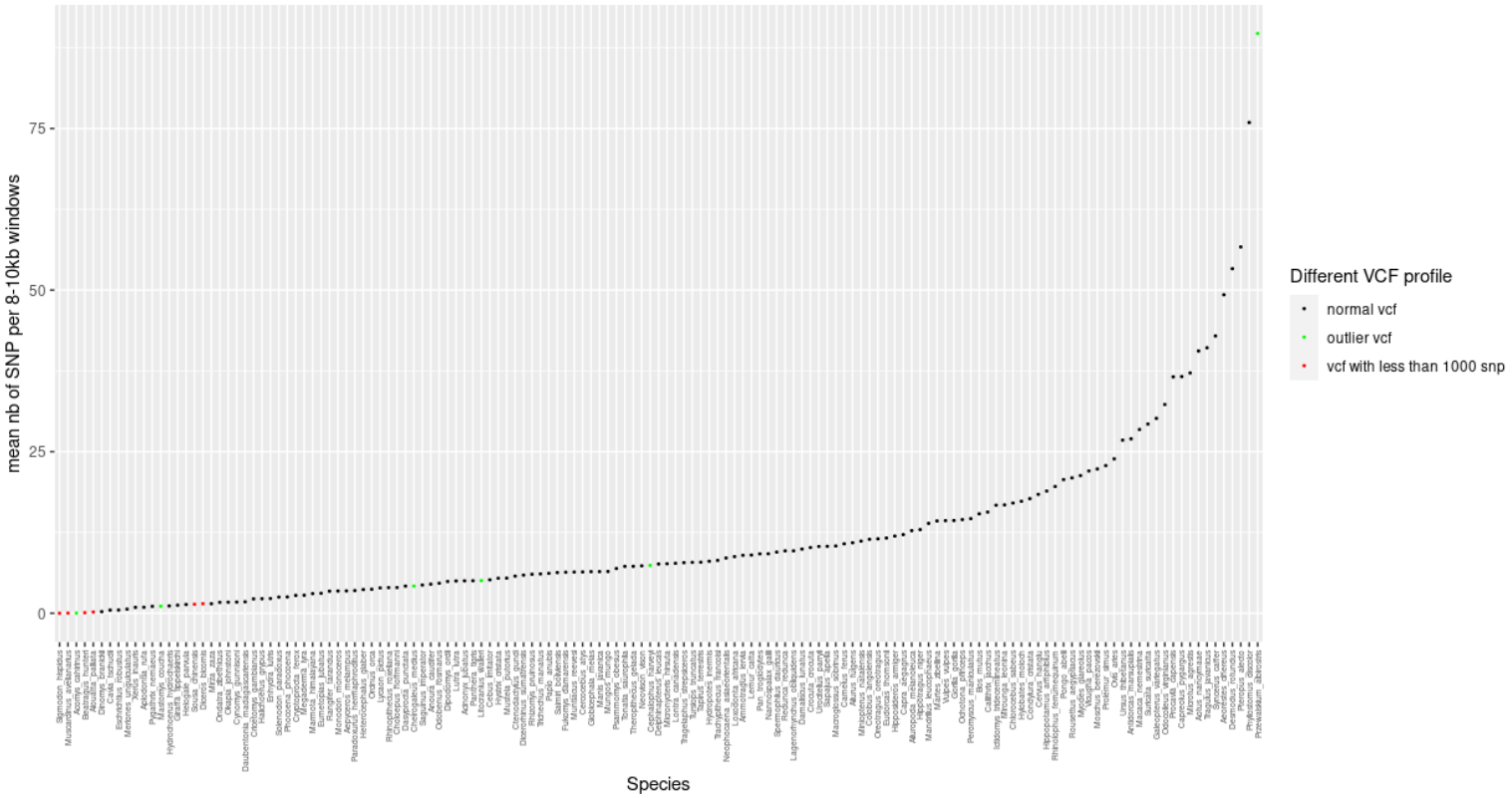


Figure S7: Mean number of callable qualitative SNP per windows of 10kb (with at least 8kb callable position) across the whole genome. Colors correspond to previously flaged species with either less than 1000 coding SNP or anormal distribution of allele frequencie

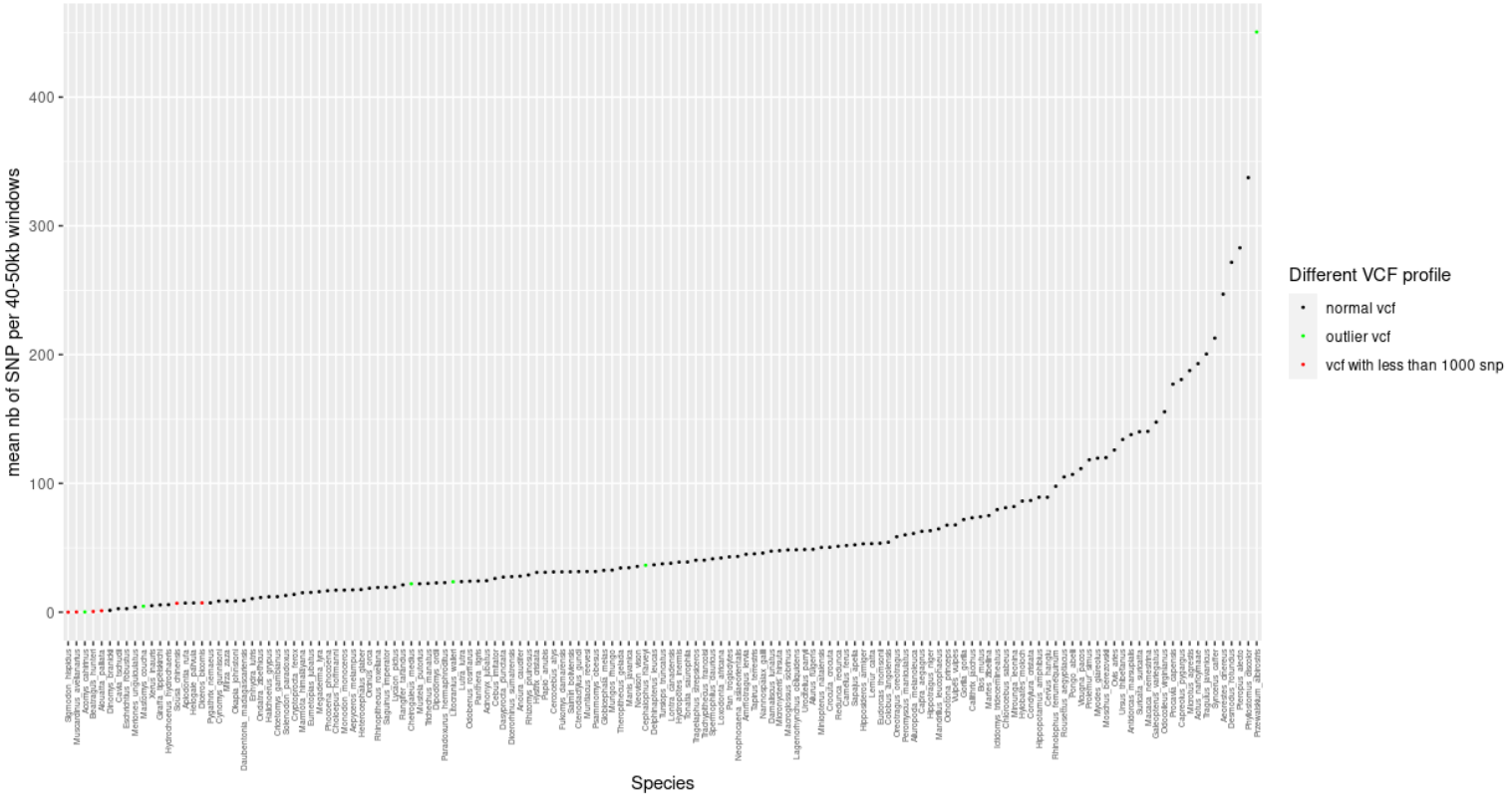


Figure S8: mean number of callable qualitative SNP per windows of 50kb (with at least 40kb callable position) across the whole genome. Colors correspond to previously flaged species with either less than 1000 coding SNP or anormal distribution of allele frequencie

5.6 Decoupled estimation of π_S and π_N/π_S

In our study, we estimate π_S and reused it as denominator for π_N/π_S estimation. This way to do has been controversial because this would mean studying the correlation between A and $\frac{B}{A}$, which are not independent measures, as it is suspected to create an artefactual negative correlation due to estimation errors on π_S.

We conducted a short analysis consisting in dividing randomly our 6002 genes list in two equal parts (named "1" and "2") and estimating π_S only, on one side and π_N and π_S to compute π_N/π_S on the other side. We did 500 times this random division of genes set and independant π_S and π_N/π_S computation in order to compute first, second and third quantile of our estimates.

We first compare the relation between the median π_N/π_S in our distribution with both type of π_S estimated ($\frac{B_2}{A_2}$ versus A_1 or A_2). Graphically, the linear regression didn't show high differences at log scale (which is the scale of our study) with identical slope and correlation coefficient whatever if the π_S is measured on the same genes for both π_S and π_N/π_S measures, or not. We then compare these slopes with the one obtained from the full dataset ($\frac{B}{A}$ versus A) used in the main results and didn't observe any difference.

With that analyse, we argue that using the same π_S for both π_N/π_S and π_S measure isn't a problem because of the consistence of π_S across the genome, due to the high quantity of SNPs used and the homogeneity of neutral selection on synonymous sites across the genome. Some authors also demonstrate that using ratio isn't misleading for this kind of study (Leroy and Nabholz, 2022).

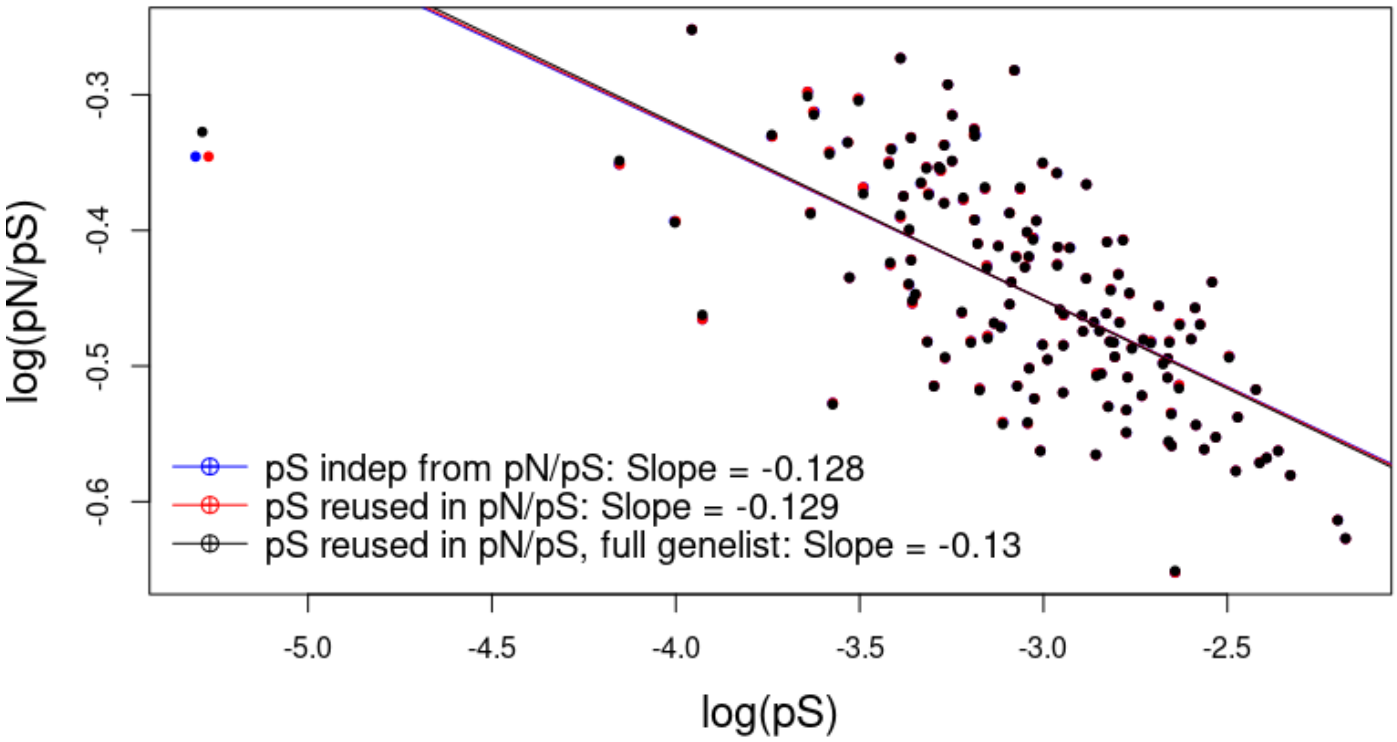


Figure S9: **Linear regression between log10(π_S) and log10(π_N/π_S)**. The gene set is divided in two part to estimating independently two π_S, one for the π_N/π_S estimation and the other for the π_S values in itself. Blue points and slope correspond to a linear regression analysis between π_S and π_N/π_S with independant π_S for the two measures. Red points and slope correspond to a π_S and π_N/π_S comparison using twice the same π_S as in the main results (but with half of the gene set). Black points and slope correspond to a π_S and π_N/π_S comparison from the main results, with the full gene list.

144spsummary_genomique_withn50

798	sp	accesion_nb	Busco score	Nb finalgenes	Coverage	n50	l50	10Matree
	Acinonyx_jubatus	GCF_003709585.1	95.9	5877	105	48500042	15	prst
	Acomys_cahirinus	GCA_004027535.1	72.59	4739	29	65411	10134	prst
	Aeorestes_cinereus	GCA_011751065.1	90.17	5486	37	35075546	19	prst
	Aepyceros_melampus	GCA_006408695.1	74.57	4873	250	344542	2451	none
	Ailuropoda_melanoleuca	GCF_002007445.1	93.66	4955	108	129245720	8	prst
	Ailurus_fulgens	GCA_002007465.1	93.77	5760	117	2983736	215	prst
	Alouatta_palliata	GCA_004027835.1	70.44	4204	27	72427	11068	prst
	Ammotragus_lervia	GCA_002201775.1	93.8	5849	94	1301762	604	none
	Anoura_caudifer	GCA_004027475.1	87.94	5483	48	185021	3282	prst
	Antidorcas_marsupialis	GCA_006408585.1	82.27	5171	192	694905	1300	none
	Aotus_nancymaae	GCF_000952055.2	93.05	5747	189	8268663	102	none
	Aplodontia_rufa	GCA_004027875.1	72.22	4768	24	37811	19816	prst
	Beatragus_hunteri	GCA_004027495.1	75.27	4994	21	69303	11525	none
	Bos_mutus	GCA_007646595.3	93.07	5794	218	16589160	48	prst
	Callithrix_jacchus	GCA_009663435.2	92.99	5663	56	98198953	12	prst
	Camelus_ferus	GCF_009834535.1	96.03	5790	128	76025729	11	prst
	Capra_aegagrus	GCA_000765075.1	94.53	5882	172	1750063	409	none
	Capreolus_pygargus	GCA_012922965.1	92.81	5736	94	6067221	126	none
	Cavia_tschudii	GCA_004027695.1	84.25	5369	28	91436	8528	prst
	Cebus_imitator	GCF_001604975.1	93.11	5789	161	5274112	149	prst
	Cephalophus_harveyi	GCA_006410635.1	85.94	5479	197	365466	2293	none
	Cercocebus_atys	GCF_000955945.1	94.15	5817	17	12849131	66	none
	Cervus_hanglu	GCA_010411085.1	92.71	5717	80	77688133	13	none
	Cheirogaleus_medius	GCA_004024725.1	80.91	5228	67	118572	5555	prst
	Chlorocebus_sabeus	GCF_000409795.2	93.78	5768	197	81825804	14	none
	Choloepus_hoffmanni	GCA_000164785.2	84.34	5279	66	366442	2423	prst
	Colobus_angolensis	GCF_000951035.1	91.38	5708	150	7840981	117	prst
	Condylura_cristata	GCF_000260355.1	89.52	5520	195	55520359	13	prst
	Cricetomys_gambianus	GCA_004027575.1	83.94	5339	29	110049	5340	prst
	Crocuta_crocuta	GCA_008692635.1	94.63	5861	119	7236831	105	prst
	Cryptoprocta_ferox	GCA_004023885.1	86.52	5569	41	173473	3830	prst
	Ctenodactylus_gundi	GCA_004027205.1	91.89	5706	42	354548	1820	prst
	Cynomys_gunnisoni	GCA_011316645.1	91.95	5675	80	824613	911	none
	Damaliscus_lunatus	GCA_006408505.1	91.35	5692	163	1166796	799	none
	Dasyprocta_punctata	GCA_004363535.1	70.4	4559	34	43703	17174	prst
	Daubentonia_madagascariensis	GCA_004027145.1	92.42	5813	69	379919	1894	prst
	Delphinapterus_leucas	GCF_002288925.2	93.94	5841	106	31183418	21	prst
	Desmodus_rotundus	GCF_002940915.1	96.12	5827	123	26869735	26	prst
	Dicerorhinus_sumatrensis	GCA_002844835.1	93.85	5795	69	614498	1268	prst
	Diceros_bicornis	GCA_004027315.2	96.75	5885	10	14664201	54	prst
	Dinomys_branickii	GCA_004027595.1	80.54	5218	27	77918	9284	prst
	Dipodomys_ordii	GCF_000151885.1	89.71	5507	146	11931245	56	prst
	Enhydra_lutris	GCF_002288905.1	96.6	5884	55	38751465	20	none
	Eschrichtius_robustus	GCA_004363415.1	78.56	5200	34	94414	7799	prst
	Eudorcas_thomsonii	GCA_006408755.1	92.69	5374	220	1581717	484	none
	Eumetopias_jubatus	GCF_004028035.1	92.62	5744	43	14018600	54	prst
	Fukomys_damarensis	GCF_012274545.1	91.36	5571	172	62586000	15	prst
	Galeopterus_variegatus	GCA_004027255.2	94.59	5784	27	7885395	91	prst
	Giraffa_tippelskirchi	GCA_001651235.1	85.1	5469	37	212164	3486	prst
	Globicephala_melas	GCF_006547405.1	93.13	5832	48	18102937	42	none
	Gorilla_gorilla	GCF_008122165.1	91.79	5646	37	26116462	35	none
	Halichoerus_grypus	GCA_012393455.1	92.27	4443	26	1033864	637	prst

799	Helogale_parvula	GCA_004023845.1	83.72	5399	30	179119	3602	none
	Heterocephalus_glaber	GCF_000247695.1	90.96	5552	150	20532749	42	prst
	Hippopotamus_amphibius	GCA_004027065.2	93.56	3253	63	4444377	171	prst
	Hipposideros_armiger	GCF_001890085.1	89.72	5552	186	2328177	272	prst
	Hippotragus_niger	GCA_006942125.1	94.58	5867	58	4586323	168	none
	Hydrochoerus_hydrochaeris	GCA_004027455.1	88.23	5514	25	202224	3717	prst
	Hydropotes_inermis	GCA_006459105.1	94.18	5773	21	13818975	55	none
	Hylobates_moloch	GCF_009828535.2	92.33	5617	64	125196221	9	prst
	Hystrix_cristata	GCA_004026905.1	79.49	5105	29	64768	10348	prst
	Ictidomys_tridecemlineatus	GCF_000236235.1	94.29	5692	64	8192786	80	none
	Lagenorhynchus_obliquidens	GCF_003676395.1	93.8	5845	52	28371583	22	none
	Lemur_catta	GCA_004024665.1	87.31	5509	66	215715	2945	prst
	Litocranius_walleri	GCA_006410535.1	91.79	5627	232	3126223	282	none
	Lontra_canadensis	GCF_010015895.1	95.53	5855	50	18460785	39	none
	Loxodonta_africana	GCF_000001905.1	93.03	5732	42	46401353	21	prst
	Lutra_lutra	GCA_902655055.1	95.56	5787	86	149004807	7	prst
	Lycaon_pictus	GCA_004216515.1	84.74	4993	45	7494581	76	none
	Macaca_nemestrina	GCF_000956065.1	94.0	5839	242	15219753	62	prst
	Macroglossus_sobrinus	GCA_004027375.1	91.85	5709	48	453401	1154	prst
	Mandrillus_leucophaeus	GCF_000951045.1	91.11	5720	104	3186748	285	none
	Manis_javanica	GCF_001685135.1	85.45	5464	94	204728	3399	prst
	Marmota_himalayana	GCA_005280165.1	91.47	5614	183	1497034	483	prst
	Martes_zibellina	GCA_012583365.1	95.0	5822	104	5199373	134	prst
	Mastomys_coucha	GCF_008632895.1	95.11	5706	56	118621203	8	prst
	Megaderma_lyra	GCA_004026885.1	88.14	5554	38	96489	6165	prst
	Meriones_unguiculatus	GCA_004026785.1	87.69	5581	28	100883	5504	prst
	Micronycteris_hirsuta	GCA_004026765.1	78.83	5104	35	68868	8795	prst
	Microtus_agrestis	GCA_902806755.1	89.98	5654	28	13349786	45	prst
	Miniopterus_natalensis	GCF_001595765.1	93.83	5733	115	4315193	118	prst
	Mirounga_leonina	GCF_011800145.1	93.62	5769	46	54232831	16	prst
	Mirza_zaza	GCA_008750895.1	72.62	4752	56	75843	8667	prst
	Monodon_monoceros	GCF_005190385.1	93.77	5772	42	107566389	9	none
	Moschus_berezovskii	GCA_006459085.1	91.82	5169	272	2509225	326	prst
	Mungos_mungo	GCA_004023785.1	88.37	5626	43	236501	2886	none
	Muntiacus_reevesi	GCA_008787405.2	93.81	5723	19	94101870	7	none
	Muscardinus_avellanarius	GCA_004027005.1	71.58	4712	29	59013	12037	prst
	Mustela_putorius	GCA_009859225.1	72.59	4568	157	33389539	24	none
	Myodes_glareolus	GCA_004368595.1	78.34	4760	195	1590265	413	none
	Nannospalax_galili	GCF_000622305.1	94.89	5809	62	3618479	238	prst
	Neophocaena_asiaeorientalis	GCF_003031525.1	93.45	5855	251	6341296	103	prst
	Neovison_vison	GCA_900108605.1	94.34	5320	64	6814223	103	prst
	Ochotona_princeps	GCF_000292845.1	91.43	5639	136	26863993	26	prst
	Odobenus_rosmarus	GCF_000321225.1	93.96	5818	35	2616778	269	prst
	Odocoileus_virginianus	GCF_002102435.1	91.35	5747	233	850721	758	prst
	Okapia_johnstoni	GCA_001660835.1	81.69	5302	32	111538	6512	none
	Ondatra_zibethicus	GCA_004026605.1	81.71	5304	39	89093	7754	none
	Orcinus_orca	GCF_000331955.2	93.98	5861	216	12735091	60	prst
	Oreotragus_oreotragus	GCA_006410675.1	79.14	4750	152	339390	3115	none
	Ovis_aries	GCA_000765115.1	94.1	5869	174	2217029	328	prst
	Pan_troglodytes	GCF_002880755.1	94.23	5787	71	53103722	19	prst
	Panthera_tigris	GCF_000464555.1	86.47	5477	131	8860407	87	none
	Papio_anubis	GCF_008728515.1	92.02	5654	54	140274886	9	none
	Paradoxurus_hermaphroditus	GCA_004024585.1	74.2	4833	32	71823	8704	prst

800	Peromyscus_maniculatus	GCF_000500345.1	94.88	5813	110	3760915	193	prst
	Phocoena_phocoena	GCA_004363495.1	82.16	5345	47	115969	6441	none
	Phyllostomus_discolor	GCF_004126475.1	96.15	5762	98	110241909	6	prst
	Pongo_abelii	GCF_002880775.1	94.97	5801	41	98475126	13	prst
	Procavia_capensis	GCA_004026925.2	94.7	5814	19	9071062	110	prst
	Prolemur_simus	GCA_003258685.1	94.91	5837	350	2710671	251	none
	Przewalskium_albirostris	GCA_006408465.1	89.79	5653	203	3769372	218	none
	Psammomys_obesus	GCA_002215935.2	95.83	5872	154	10472398	62	none
	Pteropus_alecto	GCF_000325575.1	95.59	5849	158	15954802	36	prst
	Pygathrix_nemaeus	GCA_004024825.1	73.31	4823	46	68569	12661	none
	Rangifer_tarandus	GCA_004026565.1	80.1	5259	39	89062	8918	none
	Redunca_redunca	GCA_006410935.1	71.3	4498	238	423407	1625	none
	Rhinolophus_ferrumequinum	GCF_004115265.1	96.13	5859	78	88025743	11	prst
	Rhinopithecus_roxellana	GCF_007565055.1	93.19	5714	100	144559847	9	none
	Rhizomys_pruinosus	GCA_009823505.1	89.82	5575	16	2203772	489	prst
	Rousettus_aegyptiacus	GCF_001466805.2	95.89	5826	50	2007187	297	prst
	Saguinus_imperator	GCA_004024885.1	76.11	5023	44	65636	12800	none
	Saimiri_boliviensis	GCF_000235385.1	92.62	5751	158	18744880	39	prst
	Sapajus_apella	GCF_009761245.1	92.71	5759	34	23742480	35	none
	Sigmodon_hispidus	GCA_004025045.1	82.44	5307	38	101373	7354	prst
	Solenodon_paradoxus	GCA_004363575.1	91.26	5692	27	407682	1301	prst
	Sousa_chinensis	GCA_007760645.1	93.44	5825	148	19436979	39	none
	Spermophilus_dauricus	GCA_002406435.1	88.97	5481	233	1761345	493	none
	Suricata_suricatta	GCF_006229205.1	93.85	5704	44	141453419	8	prst
	Syncerus_caffer	GCA_902825105.1	94.44	5804	198	69160875	13	none
	Tapirus_terrestris	GCA_004025025.1	84.83	5379	38	186384	3791	prst
	Theropithecus_gelada	GCF_003255815.1	93.5	5744	76	130230028	9	none
	Tonatia_saurophila	GCA_004024845.1	86.02	5410	42	165561	3525	prst
	Trachypithecus_francoisi	GCF_009764315.1	93.29	5746	39	130977661	10	none
	Tragelaphus_strepsiceros	GCA_006410795.1	84.0	5423	241	511483	1640	none
	Tragulus_javanicus	GCA_004024965.2	93.43	5730	36	14082842	49	prst
	Trichechus_manatus	GCF_000243295.1	94.26	5781	187	14442683	67	prst
	Tursiops_truncatus	GCF_011762595.1	93.3	5776	352	108430135	9	none
	Urocitellus_parryii	GCF_003426925.1	94.49	5752	122	3964291	175	none
	Ursus_thibetanus	GCA_009660055.1	95.24	5850	98	26803000	27	prst
	Vicugna_pacos	GCA_000767525.1	95.06	5797	137	5303709	107	prst
	Vulpes_vulpes	GCF_003160815.1	94.16	5771	145	12472085	55	prst
	Xerus_inauris	GCA_004024805.1	77.21	5064	21	83865	8399	prst

144spsummary_polymorphism

sp	nb_snp	pS	pN/pS	nbgeneswithsnp	vcf_type
Acinonyx_jubatus	3025	0.00044	0.37834	1789	normal
Acomys_cahirinus	1114	NA	NA	28	abnormal vcf
Aeorestes_cinereus	34281	0.00663	0.23602	4518	normal
Aepyceros_melampus	4815	0.00113	0.34563	1972	normal
Ailuropoda_melanoleuca	3966	0.00091	0.38094	1769	normal
Ailurus_fulgens	6003	0.00096	0.40461	2570	normal
Alouatta_palliata	732	0.00018	0.46825	410	too few snp
Ammotragus_lervia	6630	0.00103	0.31973	2208	normal
Anoura_caudifer	8364	0.0015	0.29506	3054	normal
Antidorcas_marsupialis	17281	0.00404	0.27044	4096	normal
Aotus_nancymaae	18646	0.00287	0.3646	4436	normal
Aplodontia_rufa	4514	0.0009	0.39706	2063	normal
Beatragus_hunteri	583	0.00012	0.34493	427	too few snp
Bos_mutus	8208	0.00127	0.34479	3183	normal
Callithrix_jacchus	6405	0.00094	0.39181	2492	normal
Camelus_ferus	5155	0.00082	0.36464	2264	normal
Capra_aegagrus	9007	0.00137	0.34068	3224	normal
Capreolus_pygargus	18079	0.00337	0.28982	4489	normal
Cavia_tschudii	2210	0.00038	0.37698	981	normal
Cebus_imitator	3999	0.00054	0.45991	1905	normal
Cephalophus_harveyi	8962	NA	NA	3011	abnormal vcf
Cercocebus_atys	4280	0.00071	0.33139	2270	normal
Cervus_hanglu	9298	0.00148	0.34573	2795	normal
Cheirogaleus_medius	52463	NA	NA	3546	abnormal vcf
Chlorocebus_sabeus	7449	0.00109	0.387	2586	normal
Choloepus_hoffmanni	4877	0.00086	0.42825	1682	normal
Colobus_angolensis	7013	0.00099	0.44668	2912	normal
Condylura_cristata	9197	0.00168	0.28259	3260	normal
Cricetomys_gambianus	4660	0.00091	0.28743	2188	normal
Crocuta_crocuta	6059	0.00099	0.32773	2582	normal
Cryptoprocta_ferox	4323	0.00077	0.28657	2393	normal
Ctenodactylus_gundi	6818	0.00111	0.34801	2913	normal
Cynomys_gunnisoni	2189	0.00032	0.42363	740	normal
Damaliscus_lunatus	7830	0.00139	0.31089	3073	normal
Dasyprocta_punctata	12324	0.00294	0.2802	2967	normal
Daubentonia_madagascariensis	3241	0.00044	0.46573	1834	normal
Delphinapterus_leucas	4387	0.0006	0.42086	2400	normal
Desmodus_rotundus	25734	0.00435	0.27372	4740	normal
Dicerorhinus_sumatrensis	5587	0.00081	0.40993	2261	normal
Diceros_bicornis	627	0.0001	0.40346	523	too few snp
Dinomys_branickii	1394	0.00023	0.5	743	normal
Dipodomys_ordii	2702	0.00048	0.32934	1419	normal
Enhydra_lutris	1657	0.00023	0.40944	1155	normal
Eschrichtius_robustus	2010	0.00031	0.49604	1404	normal
Eudorcas_thomsonii	7486	0.00165	0.3914	2439	normal
Eumetopias_jubatus	1978	0.0003	0.36754	1238	normal
Fukomys_damarensis	3862	0.00055	0.50994	1875	normal
Galeopterus_variegatus	13602	0.00218	0.32029	3350	normal
Giraffa_tippelskirchi	2584	0.00041	0.40858	1534	normal
Globicephala_melas	3475	0.00049	0.42273	1775	normal
Gorilla_gorilla	7285	0.00109	0.43879	2916	normal
Halichoerus_grypus	1437	0.00045	0.3574	807	normal

802	Helogale_parvula	2977	0.00054	0.32101	1321	normal
	Heterocephalus_glaber	3087	0.00041	0.53313	1403	normal
	Hippopotamus_amphibius	4762	0.00171	0.35802	1642	normal
	Hipposideros_armiger	9633	0.00161	0.34028	2886	normal
	Hippotragus_niger	7609	0.00128	0.33542	2993	normal
	Hydrochoerus_hydrochaeris	2928	0.00046	0.43189	1863	normal
	Hydropotes_inermis	3209	0.00067	0.30346	1790	normal
	Hylobates_moloch	8945	0.00131	0.43041	3391	normal
	Hystrix_cristata	10945	0.0022	0.32904	3537	normal
	Ictidomys_tridecemlineatus	7933	0.00142	0.33558	2361	normal
	Lagenorhynchus_obliquidens	5195	0.00075	0.38743	2488	normal
	Lemur_catta	11048	0.00187	0.33055	3296	normal
	Litocranius_walleri	2476	NA	NA	1248	abnormal vcf
	Lontra_canadensis	4998	0.00077	0.33775	2148	normal
	Loxodonta_africana	3685	0.00053	0.44197	1791	normal
	Lutra_lutra	2741	0.00043	0.36358	1187	normal
	Lycaon_pictus	2053	0.00039	0.45666	1070	normal
	Macaca_nemestrina	15490	0.00234	0.33893	4467	normal
	Macroglossus_sobrinus	12011	0.00196	0.32869	3273	normal
	Mandrillus_leucophaeus	8493	0.00131	0.36681	3101	normal
	Manis_javanica	9512	0.0016	0.36931	3159	normal
	Marmota_himalayana	2784	0.00042	0.42185	1619	normal
	Martes_zibellina	10149	0.00168	0.29351	3194	normal
	Mastomys_coucha	1708	NA	NA	181	abnormal vcf
	Megaderma_lyra	5321	0.00091	0.31485	2555	normal
	Meriones_unguiculatus	1489	0.00027	0.29626	599	normal
	Micronycteris_hirsuta	16030	0.00333	0.2645	3557	normal
	Microtus_agrestis	12976	0.00234	0.30446	3527	normal
	Miniopterus_natalensis	4841	0.00085	0.30549	2364	normal
	Mirounga_leonina	8972	0.00144	0.31195	3429	normal
	Mirza_zaza	5285	0.00113	0.32733	1873	normal
	Monodon_monoceros	1982	0.00029	0.46214	1429	normal
	Moschus_berezovskii	9356	0.00267	0.3391	3140	normal
	Mungos_mungo	9116	0.00152	0.32948	3244	normal
	Muntiacus_reevesi	3848	0.00063	0.32895	1292	normal
	Muscardinus_avellanarius	366	7E-05	0.44819	200	too few snp
	Mustela_putorius	1203	0.00024	0.48445	537	normal
	Myodes_glareolus	8242	0.00169	0.3101	2063	normal
	Nannospalax_galili	4842	0.00081	0.35127	2328	normal
	Neophocaena_asiaeorientalis	4578	0.0007	0.37333	2200	normal
	Neovison_vison	3381	0.00073	0.34021	1578	normal
	Ochotona_princeps	9286	0.00155	0.32911	3207	normal
	Odobenus_rosmarus	3954	0.0006	0.34668	2285	normal
	Odocoileus_virginianus	21995	0.00387	0.2682	3717	normal
	Okapia_johnstoni	3827	0.00066	0.38917	2139	normal
	Ondatra_zibethicus	4958	0.00094	0.29916	1746	normal
	Orcinus_orca	2780	0.00038	0.44558	1522	normal
	Oreotragus_oreotragus	10688	0.00258	0.34906	3038	normal
	Ovis_aries	15970	0.0026	0.28616	3750	normal
	Pan_troglodytes	4323	0.00056	0.48384	2338	normal
	Panthera_tigris	2656	0.00044	0.35347	1295	normal
	Papio_anubis	2880	0.00043	0.39888	1638	normal
	Paradoxurus_hermaphroditus	10174	0.00219	0.2783	2943	normal

803	Peromyscus_maniculatus	12272	0.00228	0.22333	1372	normal
	Phocoena_phocoena	6260	0.00109	0.37503	2858	normal
	Phyllostomus_discolor	34619	0.00631	0.24334	4797	normal
	Pongo_abelii	10236	0.00149	0.39024	3387	normal
	Procavia_capensis	15909	0.00274	0.27454	4318	normal
	Prolemur_simus	13305	0.00223	0.27616	3863	normal
	Przewalskium_albirostris	54210	NA	NA	5475	abnormal vcf
	Psammomys_obesus	3923	0.00056	0.44804	1771	normal
	Pteropus_alecto	26581	0.00469	0.26268	4852	normal
	Pygathrix_nemaeus	4686	0.00089	0.37393	2117	normal
	Rangifer_tarandus	8520	0.00156	0.32128	3118	normal
	Redunca_redunca	7403	0.00174	0.32567	2556	normal
	Rhinolophus_ferrumequinum	8204	0.00139	0.27207	2519	normal
	Rhinopithecus_roxellana	1893	0.00026	0.45317	1126	normal
	Rhizomys_pruinosus	4347	0.00065	0.46729	1857	normal
	Rousettus_aegyptiacus	11282	0.00185	0.30082	2681	normal
	Saguinus_imperator	6623	0.00118	0.38638	2543	normal
	Saimiri_boliviensis	4439	0.00065	0.40518	1802	normal
	Sapajus_apella	5700	0.00084	0.38017	2169	normal
	Sigmodon_hispidus	33	1E-05	0.47059	11	too few snp
	Solenodon_paradoxus	3513	0.00054	0.41675	1840	normal
	Sousa_chinensis	888	0.00011	0.55967	551	too few snp
	Spermophilus_dauricus	5946	0.00113	0.30221	2495	normal
	Suricata_suricatta	15597	0.00252	0.33095	4009	normal
	Syncerus_caffer	19532	0.0032	0.32081	4234	normal
	Tapirus_terrestris	11797	0.00206	0.3503	3667	normal
	Theropithecus_gelada	3820	0.00052	0.44321	1815	normal
	Tonatia_saurophila	12054	0.00223	0.29136	3415	normal
	Trachypithecus_francoisi	4952	0.00065	0.47212	2511	normal
	Tragelaphus_strepsiceros	7964	0.00152	0.35949	2904	normal
	Tragulus_javanicus	23253	0.00378	0.30377	4283	normal
	Trichechus_manatus	4850	0.00069	0.42841	1539	normal
	Tursiops_truncatus	3494	0.00048	0.44229	1626	normal
	Urocitellus_parryii	6172	0.00083	0.52253	1906	normal
	Ursus_thibetanus	13519	0.00217	0.31001	3724	normal
	Vicugna_pacos	11796	0.00211	0.31724	2905	normal
	Vulpes_vulpes	5541	0.00098	0.27367	2126	normal
	Xerus_inauris	2468	0.0005	0.30548	1292	normal

summary_filteringvcf_raw_indel+bi_call_cod_6002_hetero_qual

804 species	raw_snp	biallelic_snp	call_snp	coding_snp	genelist_snp	hetero_snp	qual_snp	allelic_balance
Choloepus_hoffmanni	3867911	3867392	3401371	7445	5236	5222	4878	4877
Loxodonta_africana	6348963	6343331	5438917	11318	7639	4154	3689	3685
Trichechus_manatus	3114185	3112578	2724745	7784	5188	5132	4871	4850
Procavia_capensis	16458011	16457830	16099833	32444	22215	22211	15919	15909
Condylura_cristata	3461962	3459798	3199650	13629	9405	9390	9225	9197
Solenodon_paradoxus	1687072	1686976	1561628	6435	4513	4513	3514	3513
Aeorestes_cinereus	13235902	13235152	12737699	101003	72671	72643	34326	34281
Miniopterus_natalensis	2249627	2249145	1989441	7515	4954	4928	4841	4841
Anoura_caudifer	3813786	3813688	3680239	12369	8575	8575	8365	8364
Phyllostomus_discolor	14766295	14765094	13909805	52198	35257	35049	34632	34619
Tonatia_saurophila	5250643	5250543	5129431	17856	12547	12547	12054	12054
Desmodus_rotundus	11742968	11738825	11182434	41464	27022	26028	25744	25734
Micronycteris_hirsuta	8425912	8425801	8240018	22843	16953	16951	16037	16030
Hipposideros_armiger	4436021	4434281	3943378	14878	10063	10026	9670	9633
Rhinolophus_ferrumequinum	4562708	4562551	4372112	12991	8815	8751	8217	8204
Megaderma_lyra	3680053	3679443	3115185	8266	5578	5578	5323	5321
Macroglossus_sobrinus	3887865	3887737	3778439	18082	12292	12292	12016	12011
Rousettus_aegyptiacus	13747426	13732212	13268215	53475	35118	11764	11290	11282
Pteropus_alecto	12191444	12188675	11253289	41356	27426	27012	26601	26581
Aepyceros_melampus	4256453	4254876	3443940	8655	6464	6430	4972	4815
Ammotragus_lervia	4458698	4457469	3969977	10376	6853	6840	6643	6630
Capra_aegagrus	5569313	5567129	5101678	14250	9313	9305	9057	9007
Ovis_aries	9419714	9416861	8827788	24397	16269	16244	16043	15970
Beatragus_hunteri	955550	955370	739422	1679	1247	1247	583	583
Damaliscus_lunatus	6341937	6340372	5393454	15971	10942	10804	7947	7830
Hippotragus_niger	4331751	4331187	4011095	12032	8116	8099	7611	7609
Redunca_redunca	6171205	6170311	5025385	11074	7639	7627	7439	7403
Antidorcas_marsupialis	15770605	15768692	11945184	24957	17765	17713	17301	17281
Litocranius_walleri	11124101	11122087	8387053	23735	14910	14859	3063	2476
Eudorcas_thomsonii	14262775	14251324	6489589	19774	13085	10102	7643	7486
Cephalophus_harveyi	6957545	6954378	4913687	16155	11879	11767	9089	8962
Oreotragus_oreotragus	13060574	13057358	9566685	19196	13865	13704	10804	10688
Bos_mutus	4839035	4838378	4270063	12873	8515	8484	8254	8208
Syncerus_caffer	20710227	20670757	18247129	52939	34999	26510	20026	19532
Tragelaphus_strepsiceros	6216438	6215254	4996545	13346	9538	9516	8065	7965
Moschus_berezovskii	15005110	14982232	9580723	25503	16825	11606	9768	9356
Capreolus_pygargus	11261623	11260278	9510421	27361	18491	18478	18088	18079
Hydropotes_inermis	6694661	6693334	5263569	16231	10985	10918	3346	3209
Odocoileus_virginianus	13887384	13882602	12888814	32642	22651	22590	22147	21995
Rangifer_tarandus	6752921	6751963	6235345	12112	8892	8892	8521	8520
Cervus_hanglu	5210533	5209493	4747395	14482	9706	9685	9303	9298
Przewalskium_albirostris	28139113	28132029	25944270	85893	59134	59092	56004	54210
Muntiacus_reevesi	4111903	4111509	3631262	10967	7306	7301	3904	3848
Giraffa_tippelskirchi	2390187	2389857	2106027	5510	4031	3928	2588	2584
Okapia_johnstoni	3818826	3817344	3182923	9056	6724	6612	3840	3827
Tragulus_javanicus	9572985	9572600	9157795	35322	24329	24329	23267	23253
Delphinapterus_leucas	2084051	2083198	1866270	6740	4478	4455	4392	4387
Monodon_monoceros	1229573	1229178	964606	3703	2472	2460	1985	1982
Neophocaena_asiaeorientalis	3954634	3951830	3500255	11097	7397	4895	4655	4578
Phocoena_phocoena	4242465	4241912	3806463	8581	6343	6343	6262	6260
Globicephala_melas	1888740	1888453	1711254	5627	3849	3840	3475	3475
Sousa_chinensis	1465607	1464423	1062513	3429	2225	1047	895	888

805	Tursiops_truncatus	3314718	3313772	3185866	9888	6538	3543	3494	3494
	Lagenorhynchus_obliquidens	2729131	2728839	2533884	8266	5569	5552	5197	5195
	Orcinus_orca	1502191	1501193	1171985	5265	2946	2822	2784	2780
	Eschrichtius_robustus	1683323	1682784	1311797	3226	2374	2374	2010	2010
	Hippopotamus_amphibius	6463386	6463116	4513924	8064	4879	4878	4762	4762
	Camelus_ferus	2785708	2784833	2337081	8135	5278	5277	5175	5155
	Vicugna_pacos	7676690	7671002	5841961	19031	12787	12017	11798	11796
	Dicerorhinus_sumatrensis	3559123	3558942	3192919	8901	6050	6049	5637	5587
	Diceros_bicornis	3852154	3851608	3087018	9619	6390	6335	655	627
	Tapirus_terrestris	7352432	7352142	7031044	17797	12657	12655	11798	11797
	Manis_javanica	6370097	6368844	5516276	14076	10012	9961	9548	9512
	Ailuropoda_melanoleuca	5814268	5809637	3766171	9177	5947	4229	4028	3966
	Ursus_thibetanus	6722212	6721541	6379954	20346	13840	13774	13520	13519
	Ailurus_fulgens	2935613	2934808	2550994	9292	6336	6317	6016	6003
	Enhydra_lutris	930695	930290	738019	2816	1829	1765	1657	1657
	Lutra_lutra	1789073	1788504	1521355	4930	3275	3195	2806	2741
	Lontra_canadensis	2305347	2305054	2102188	8110	5373	5344	4998	4998
	Mustela_putorius	6946880	6933923	6121941	34028	24489	2757	1274	1203
	Neovison_vison	4547129	4441163	3093604	8357	5291	4134	3508	3381
	Martes_zibellina	4975605	4973781	4447880	16159	10541	10472	10176	10149
	Eumetopias_jubatus	1118947	1118702	968190	3540	2331	2310	1978	1978
	Odobenus_rosmarus	2126365	2125178	1876071	7690	4682	4493	3957	3954
	Halichoerus_grypus	1983016	1982384	1222905	3119	2045	2001	1446	1437
	Mirounga_leonina	4172295	4172045	4054244	14006	9498	9493	8979	8972
	Lycaon_pictus	2106705	2105362	1688246	6549	4167	3779	2092	2053
	Vulpes_vulpes	4185824	4184669	3703756	9157	5926	5711	5547	5541
	Crocuta_crocuta	3391827	3388931	2881647	10556	7096	6506	6080	6059
	Cryptoprocta_ferox	2566789	2566564	2415647	6470	4712	4712	4323	4323
	Helogale_parvula	2175611	2175476	2059405	5131	3862	3860	2977	2977
	Mungos_mungo	4693201	4693082	4579866	13028	9406	9406	9117	9116
	Suricata_suricatta	11074252	11062917	10551371	38380	25701	16330	15615	15597
	Paradoxurus_hermaphroditus	7885209	7885086	7699389	14772	11188	11188	10174	10174
	Acinonyx_jubatus	1357617	1357341	1238946	4715	3155	3106	3025	3025
	Panthera_tigris	2490110	2486780	1934636	7975	5600	2826	2657	2656
	Alouatta_palliata	2109318	2108867	1590383	1547	1074	1073	737	732
	Aotus_nancymaae	12432061	12429888	11474271	28686	19078	19026	18684	18646
	Cebus_imitator	2683402	2680892	1992767	6817	4458	4317	4016	3999
	Sapajus_apella	4408932	4408079	4287174	11137	7493	7471	5727	5700
	Saimiri_boliviensis	2366752	2365351	2276134	8151	5347	5078	4491	4439
	Callithrix_jacchus	6982018	6976812	6436870	14788	9673	6706	6428	6405
	Saguinus_imperator	6940153	6939606	6915224	9486	7069	7069	6625	6623
	Cercocebus_atys	7231882	7227337	5885645	19716	11722	11207	4348	4280
	Papio_anubis	7305776	7301145	6468693	23092	15287	13444	3163	2880
	Theropithecus_gelada	2628228	2627470	2183835	5934	3992	3984	3822	3820
	Mandrillus_leucophaeus	6658637	6655822	5730681	13349	8931	8890	8557	8493
	Macaca_nemestrina	9845950	9837537	8694095	23707	15774	15614	15507	15490
	Chlorocebus_sabeus	6016217	5392043	4667211	11826	7670	7572	7455	7449
	Colobus_angolensis	4457083	4454753	3785790	11287	7356	7348	7103	7013
	Pygathrix_nemaeus	4252593	4252046	4190641	6750	4927	4927	4692	4686
	Rhinopithecus_roxellana	2326019	2325011	1875235	4565	2977	2042	1903	1893
	Trachypithecus_francoisi	4369518	4367252	4070721	11625	7482	5608	4975	4952
	Gorilla_gorilla	5681203	5679847	5040325	13577	8947	8924	7298	7285
	Pan_troglodytes	5139322	5135642	4589729	13282	8793	6694	4335	4323

806	Pongo_abelii	7447568	7445891	7047265	17979	11738	11661	10256	10236
	Hylobates_moloch	5911583	5911356	5262991	15500	10294	10294	8969	8945
	Cheirogaleus_medius	7652246	7652095	7455169	444243	333087	333087	78906	52463
	Mirza_zaza	3455259	3454926	3293758	7704	5513	5467	5285	5285
	Lemur_catta	10045555	10033595	9512017	25503	17584	11234	11055	11048
	Prolemur_simus	6459577	6457829	6066969	20307	13388	13376	13306	13305
	Daubentonia_madagascariensis	1470338	1470168	1361713	4811	3295	3295	3245	3241
	Galeopterus_variegatus	10155633	10155118	9439045	22523	14995	14995	13606	13602
	Ochotona_princeps	4125457	4123553	3518687	13886	9627	9597	9318	9286
	Aplodontia_rufa	5990566	5989663	5193284	7948	5990	5989	4515	4514
	Cynomys_gunnisoni	2962451	2961464	1743972	3601	2430	2411	2226	2189
	Ictidomys_tridecemlineatus	8372419	8363194	6245883	16012	10122	10081	7965	7933
	Urocitellus_parryii	5991789	5985440	3879617	11855	7766	7715	6668	6173
	Spermophilus_dauricus	6867570	6859530	4769752	13889	9337	6523	5954	5946
	Marmota_himalayana	3579162	2811360	1821584	4838	3245	3075	2797	2784
	Xerus_inauris	3216209	3215987	2990103	5764	4193	4193	2468	2468
	Muscardinus_avellanarius	523760	523679	386674	578	429	429	366	366
	Acomys_cahirinus	856064	855724	509494	2464	1528	1528	1160	1114
	Meriones_unguiculatus	1502162	1501971	1234032	2412	1745	1745	1489	1489
	Psammomys_obesus	3571352	3568492	2740708	11112	7517	7401	3927	3923
	Mastomys_coucha	2311275	2307558	988362	4242	2595	2590	1841	1708
	Microtus_agrestis	16092853	16063007	14704014	39199	25543	17721	13047	12976
	Myodes_glareolus	9773210	9548168	8387090	14023	8949	8782	8397	8242
	Ondatra_zibethicus	3597723	3597584	3415527	6925	5107	5107	4959	4958
	Peromyscus_maniculatus	6400141	6398315	5550936	18175	12776	12661	12282	12272
	Sigmodon_hispidus	699714	699415	372019	203	88	88	33	33
	Cricetomys_gambianus	3934123	3933512	3330123	8122	5995	5995	4660	4660
	Nannospalax_galili	7328757	7322576	5224805	9123	5957	5814	4862	4842
	Rhizomys_pruinosus	18346489	18317698	16326904	30593	20941	12915	4571	4347
	Dipodomys_ordii	2328055	2324478	1659079	4136	2799	2749	2702	2702
	Cavia_tschudii	2816202	2815610	2061480	3841	2781	2781	2212	2210
	Hydrochoerus_hydrochaeris	2084558	2084299	1770217	5668	3948	3948	2931	2928
	Dasyprocta_punctata	11616897	11616576	11341087	18602	13899	13899	12325	12324
	Dinomys_branickii	1329891	1329524	922130	2668	1980	1980	1394	1394
	Fukomys_damarensis	2172396	2171362	1682032	6303	4212	4146	3867	3862
	Heterocephalus_glaber	1828295	1827375	1309899	5287	3515	3388	3157	3087
	Hystrix_cristata	8063611	8063143	7527022	18754	13685	13684	10949	10945
	Ctenodactylus_gundi	3847485	3847306	3556424	10184	7033	7033	6841	6818

807

Gene-branch shrinkage model

Nicolas Lartillot
nicolas.lartillot@univ-lyon1.fr

October 17, 2024

The gene-branch shrinkage model is fundamentally an approximate version of the model originally introduced in Gobbo et al (2020), using the mapping approximation (Romiguier et al, 2012). It is used in the present context to detect outlier genes, i.e. genes that deviate in their synonymous branch lengths, compared to the average over all genes, and are thus suspected to be data errors (due to alignment errors or incorrect orthology assignment).

This model expresses the distribution of synonymous branch lengths and dN/dS across genes and branches as a combination of a branch effect, a gene effect, and a residual effect, following a Gamma-Poisson structure. Thus, for gene i on branch j, the local (gene-specific) branch length l_{ij} and dN/dS ω_{ij} are assumed to have a gamma prior:

$$\begin{aligned} l_{ij} &\sim Gamma(a_i b_j, \alpha) \\ \omega_{ij} &\sim Gamma(u_i v_j, \beta) \end{aligned}$$

where:

- a_i and b_j are the gene and branch effects on dS
- u_i and v_j are the gene and branch effects on dN/dS
- α and β tune the variance of the residual gene-branch deviations

In a hierarchical Bayes framework, the priors on a, b, u, v are all Gamma, with hyperpameters (mean and shape parameter) being themselves estimated. To make the model identifiable, the prior has mean of 1 for the a_i's and for the v_j's. With this convention, the mean over the b_j's represents the mean synonymous length of branch j, while the mean over the u_i's represents the mean dN/dS for gene i.

This hierarchical model can then be combined with the Poisson likelihood for the mapping statistics:

$$K^S_{ij} \sim Poisson(l_{ij}\, L^S_{ij}) \quad (1)$$

$$K^N_{ij} \sim Poisson(l_{ij}\, \omega_{ij}\, L^N_{ij}) \quad (2)$$

Here, K^S_{ij} and K^N_{ij} are the synonymous and non-synonymous counts, for gene i and branch j, and L^S_{ij} and L^N_{ij} are the corresponding numbers of mutational targets. These Poisson likelihoods can be seen as a link function, such that the overall procedure can be seen as a Bayesian hierarchical Poisson regression model, which, based on the 'observed' counts K and covariates L, estimates

808

gene- and branch-effects on dS and dN/dS. It can also be seen as a Bayesian shrinkage device, in the sense that the gene-branch parameters l_{ij} and ω_{ij} are shrunk towards their expected value (resp. $a_i b_j$ and $u_i v_j$) based on information-sharing in the two dimensions.

Shrinkage is particularly useful in the present context, as it will smooth out the stochastic errors on those gene-branch configurations that are characterized by low synonymous and non-synonymous counts (small gene length and/or short branch). As a result, and conversely, gene-branch configurations that in the end do show a large deviation between their posterior estimate for l_{ij} and the expectation based on other genes and branches, $a_i b_j$, can only do so by virtue of a strong empirical signal. Thus, the shrinkage model efficiently filters out stochastic errors, so as to more clearly single out the statistically significant outliers.

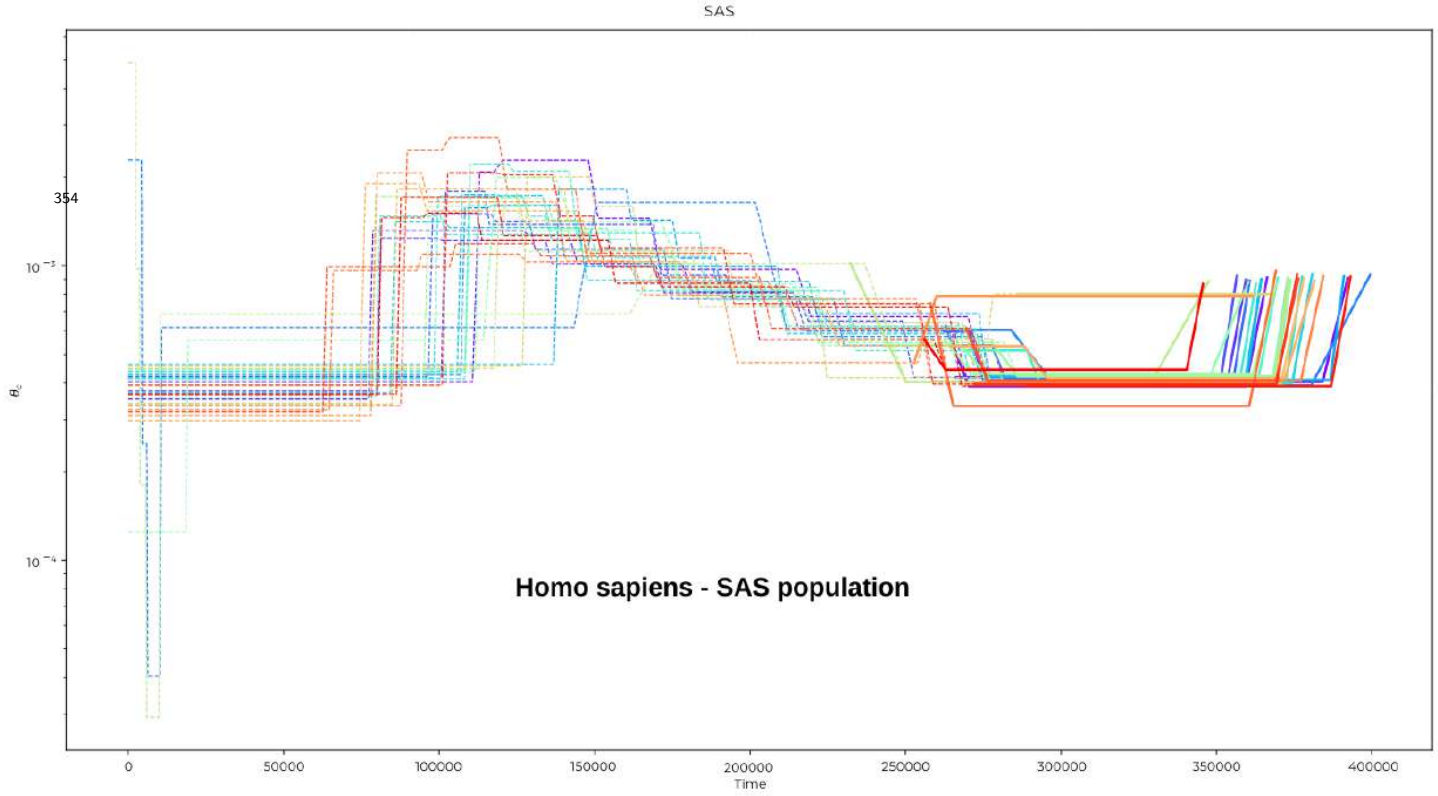
This model was implemented in a simple MCMC framework. After running the MCMC, for a given gene i and branch j, the dS and dN/dS deviations, defined as:

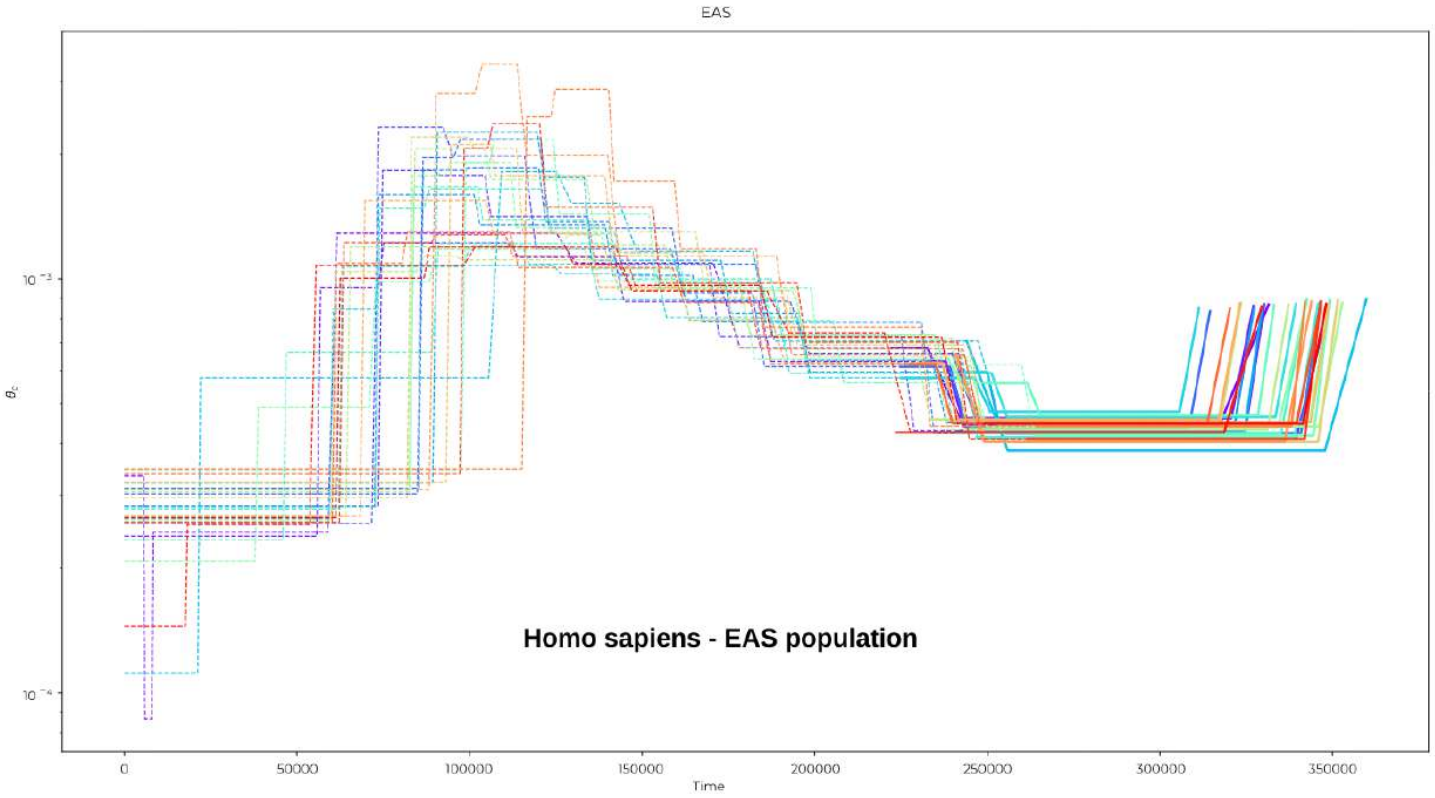
$$\begin{aligned} x_{ij} &= \frac{l_{ij}}{a_i b_j} \\ y_{ij} &= \frac{\omega_{ij}}{u_i v_j} \end{aligned}$$

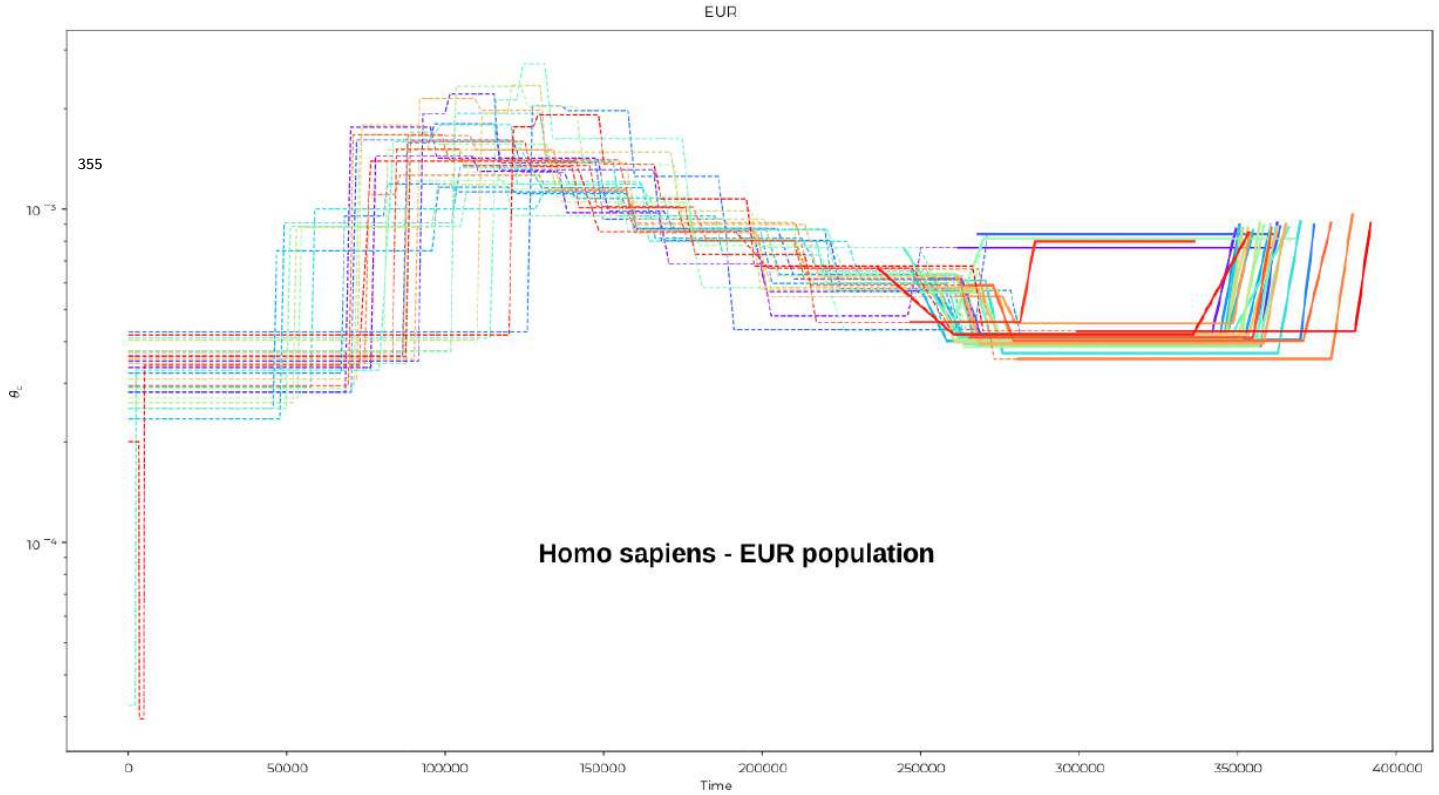
were averaged over the MCMC sample, giving posterior mean estimates of the deviations. Genes with a dS deviation larger than 3 were discarded.

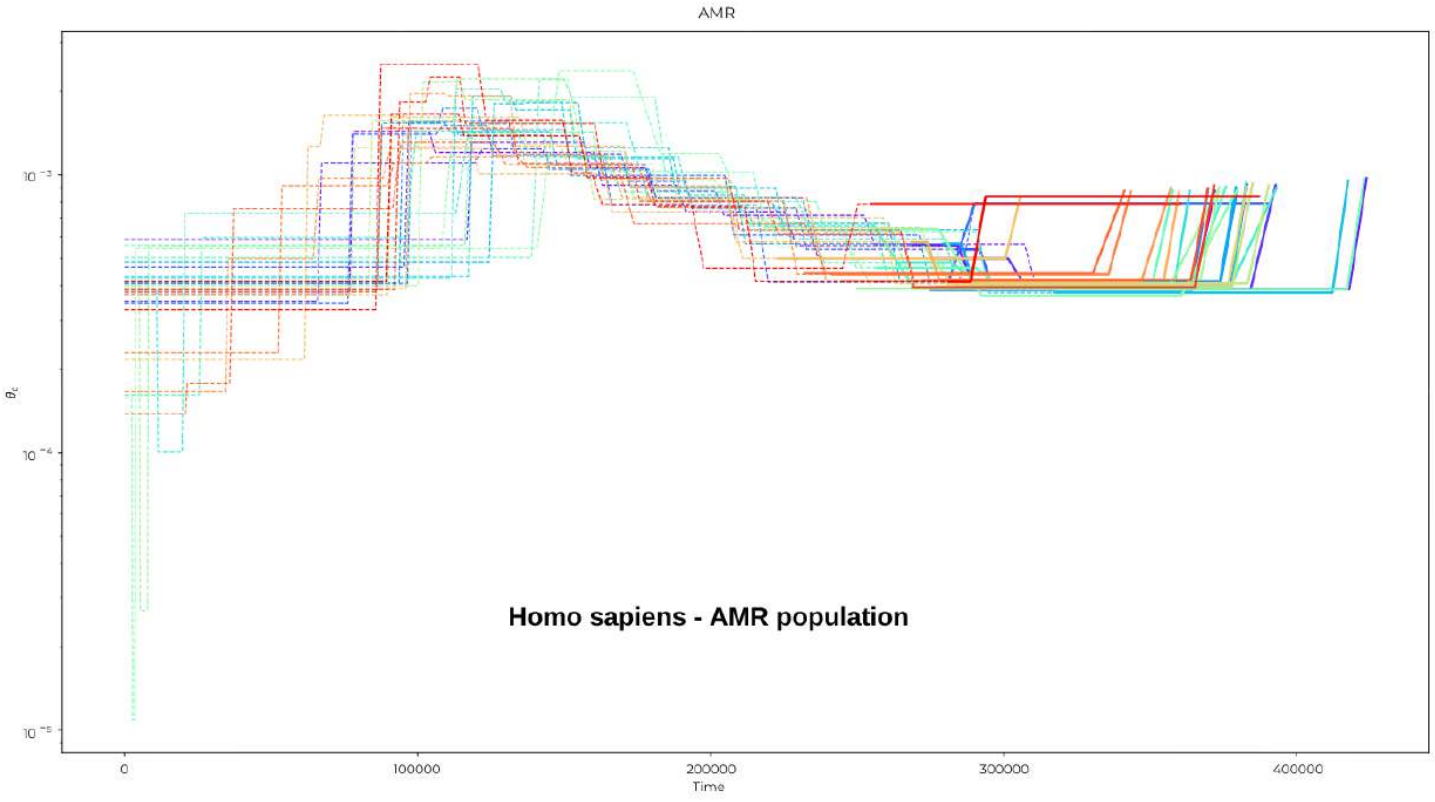
9.2 Annexe article 2

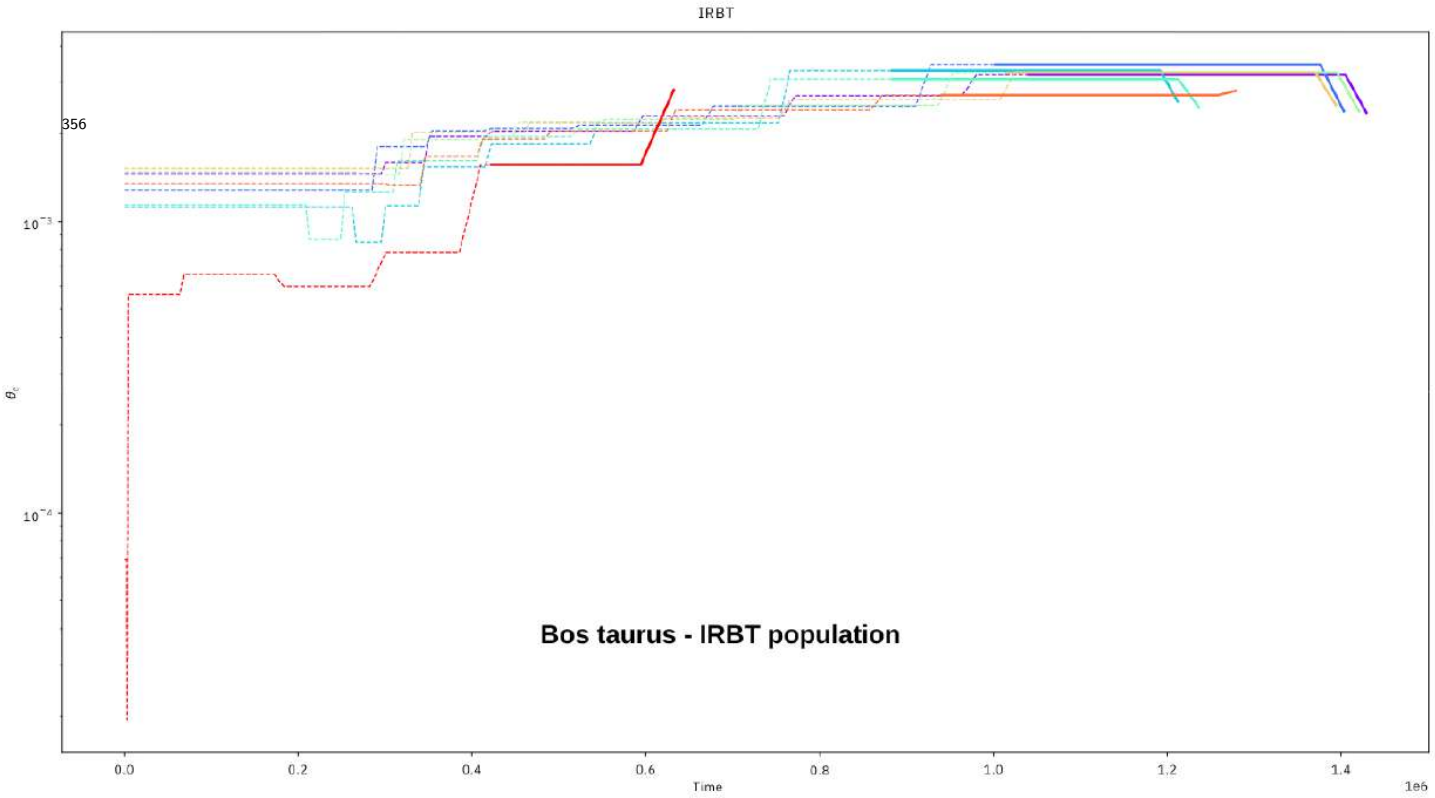
A sample of pSMC graphs, with a maximum of four graphs per species (randomly selected). Other graphs can be requested directly to the authors.

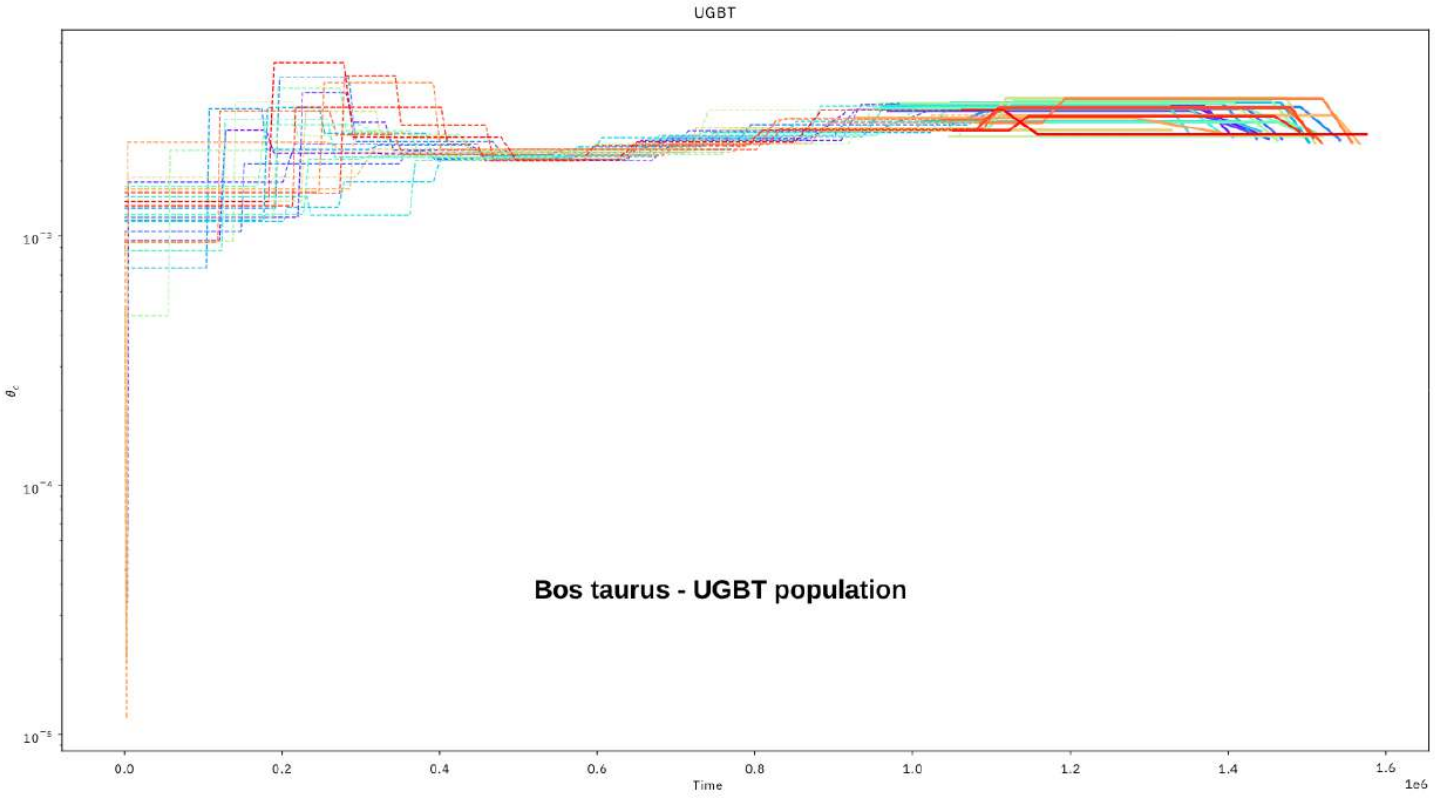

SAS
10^{-3}
10^{-4}
θ_S
Homo sapiens - SAS population
0
50000
100000
150000
200000
250000
300000
350000
400000
Time

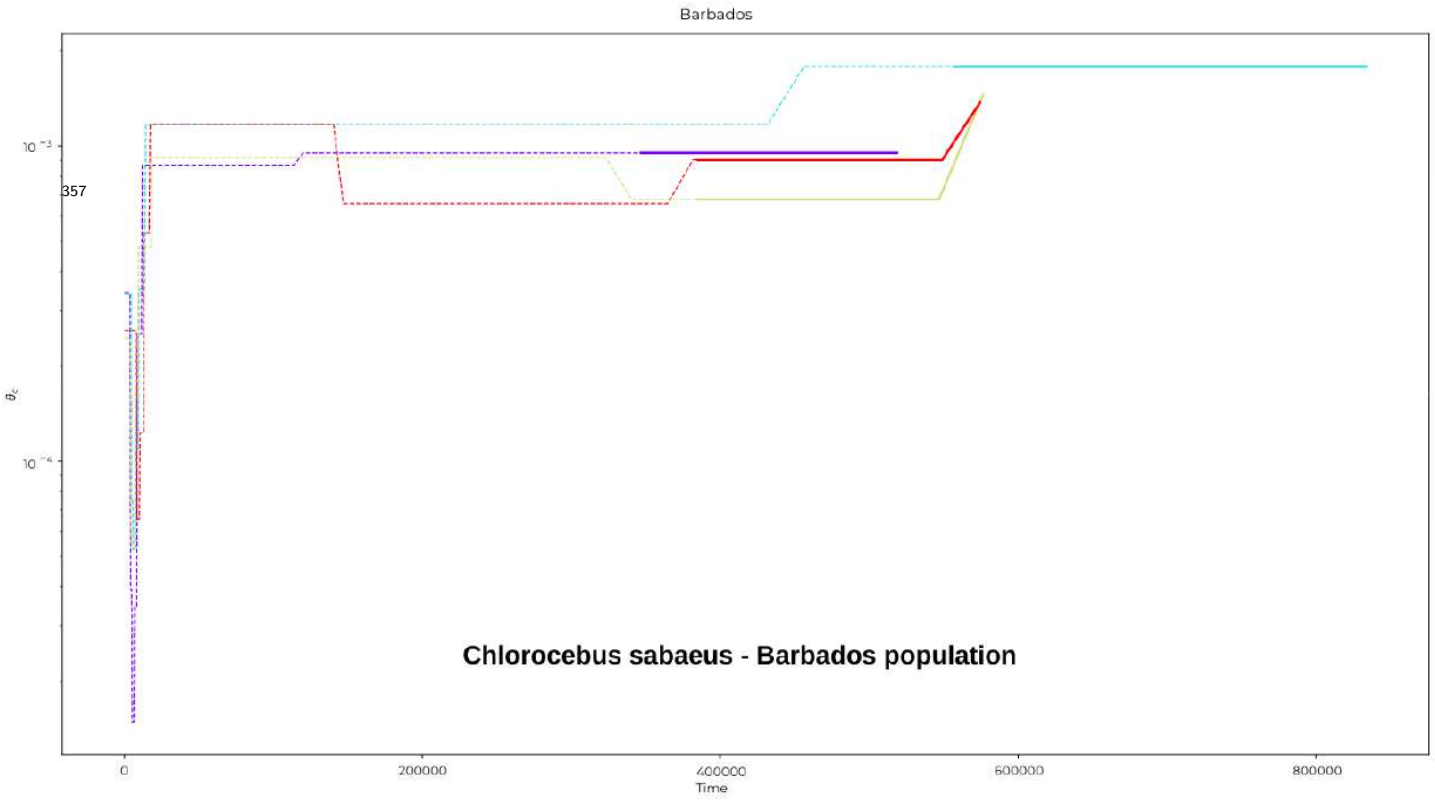

EAS
10^{-3}
10^{-4}
θ_E
Homo sapiens - EAS population
0
50000
100000
150000
200000
250000
300000
350000
Time

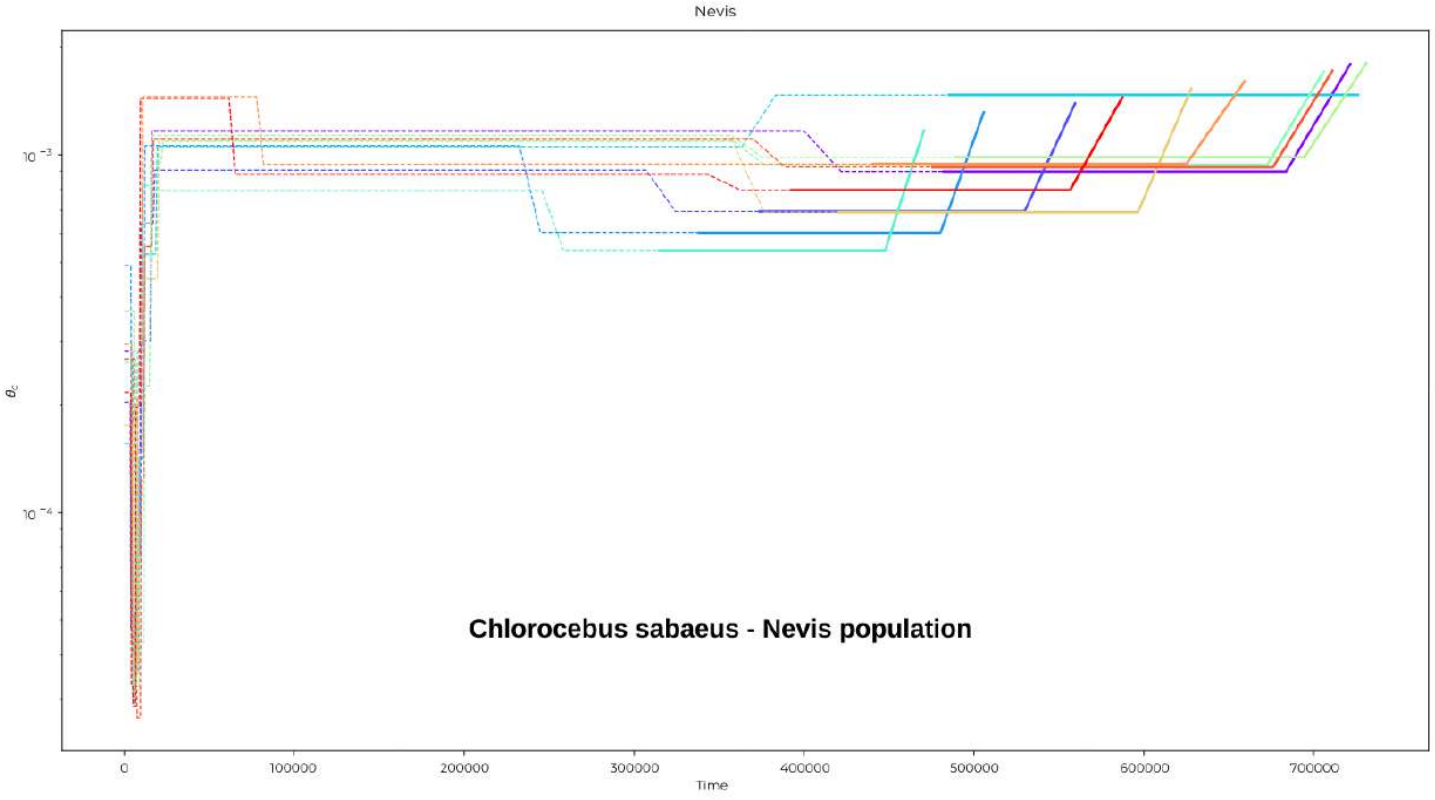

EUR
Homo sapiens - EUR population
Time
0
50000
100000
150000
200000
250000
300000
350000
400000

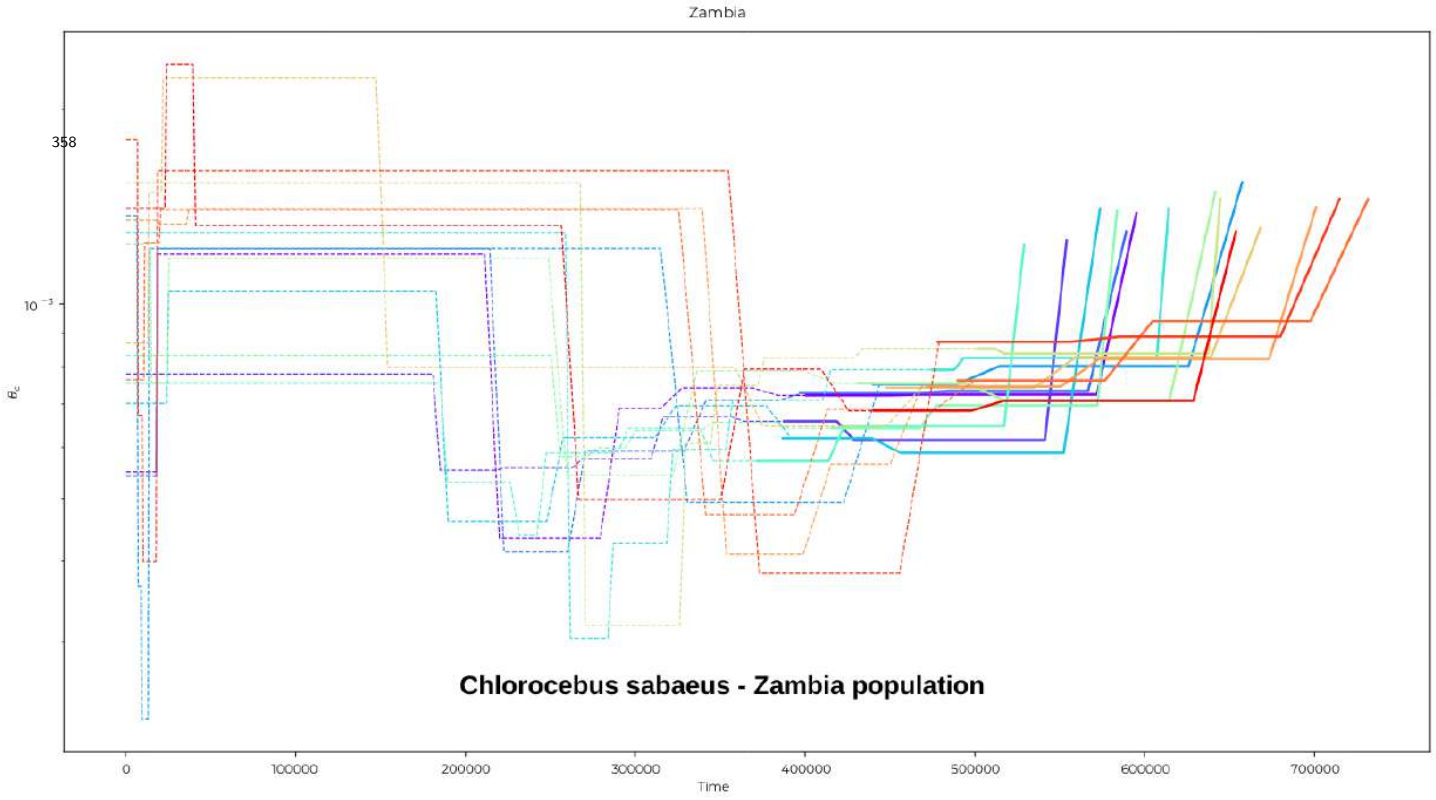

AMR
Homo sapiens - AMR population
Time
0
100000
200000
300000
400000

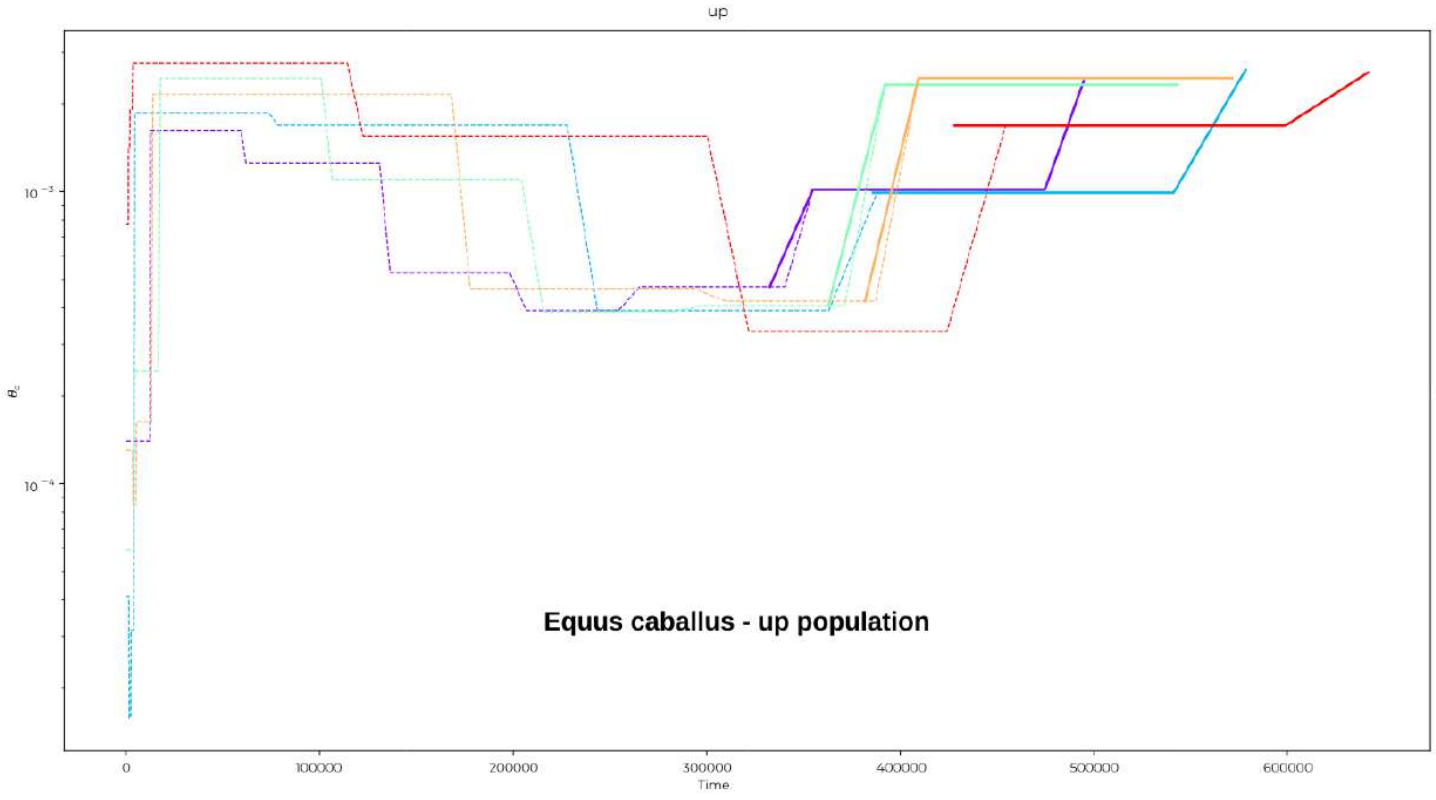

IRBT
θ_c
10^{-3}
10^{-4}
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
Time
1e6
Bos taurus - IRBT population

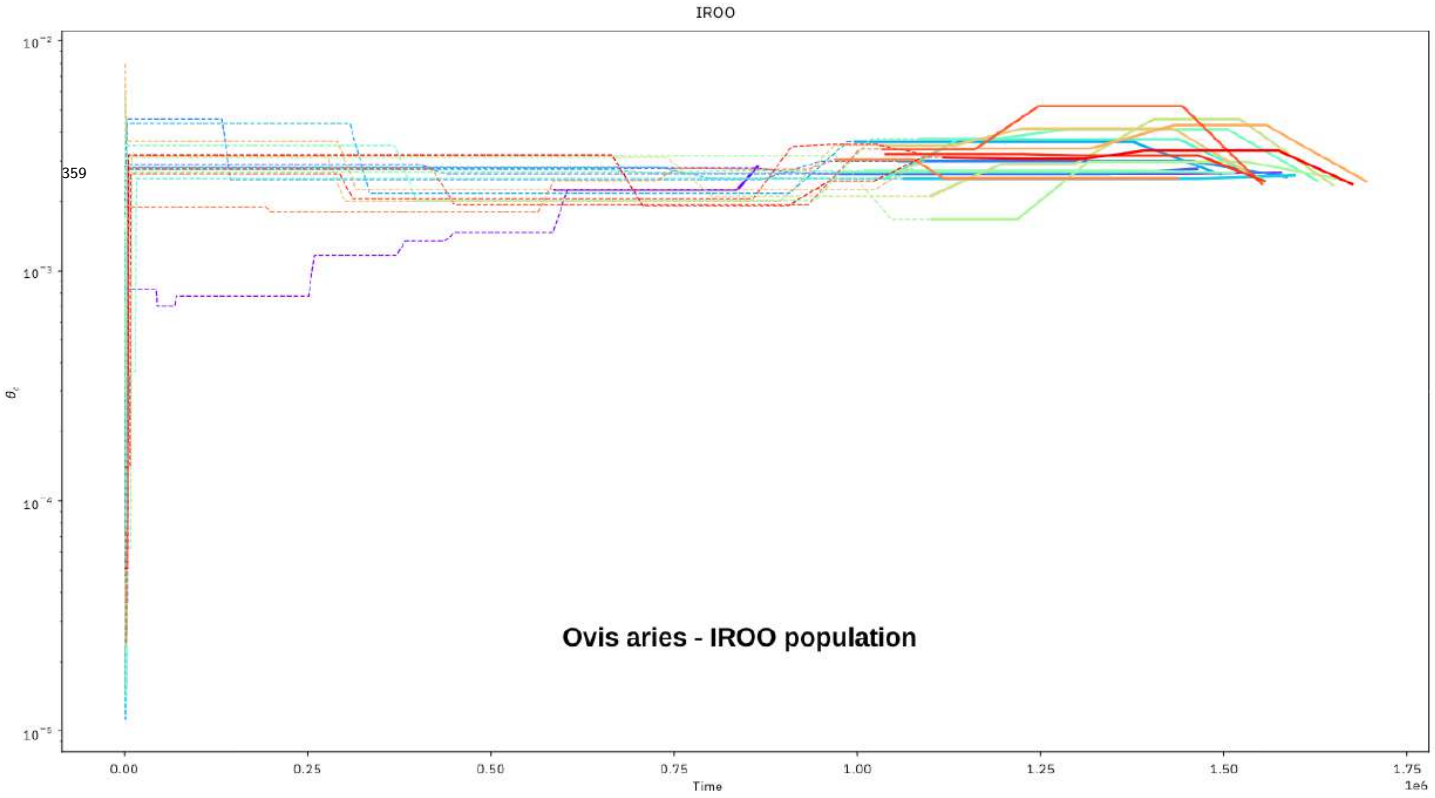

UGBT
θ_c
10^{-3}
10^{-4}
10^{-5}
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
Time
1e6
Bos taurus - UGBT population

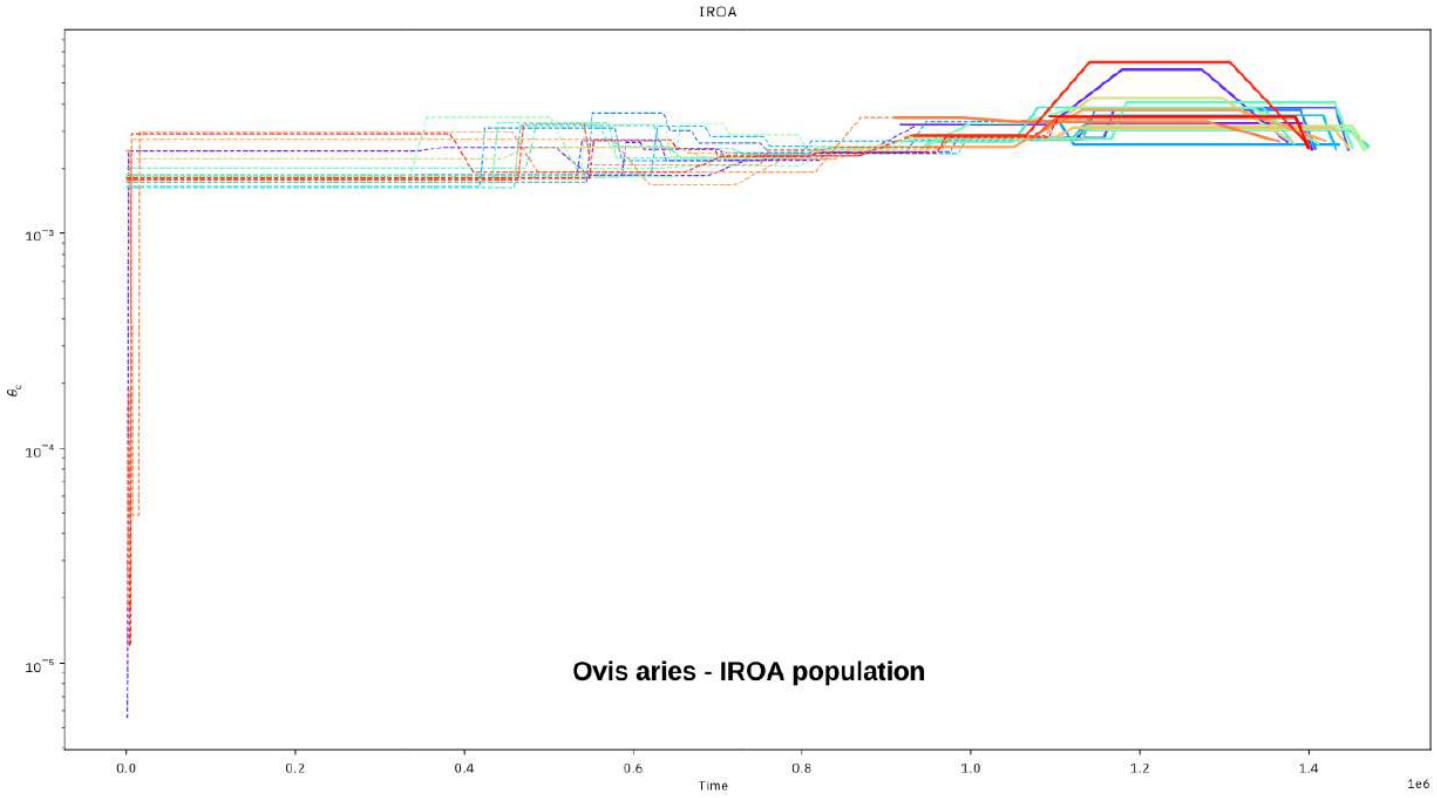

Barbados
10^{-3}
10^{-4}
θ_c
0
200000
400000
600000
800000
Time
Chlorocebus sabaeus - Barbados population

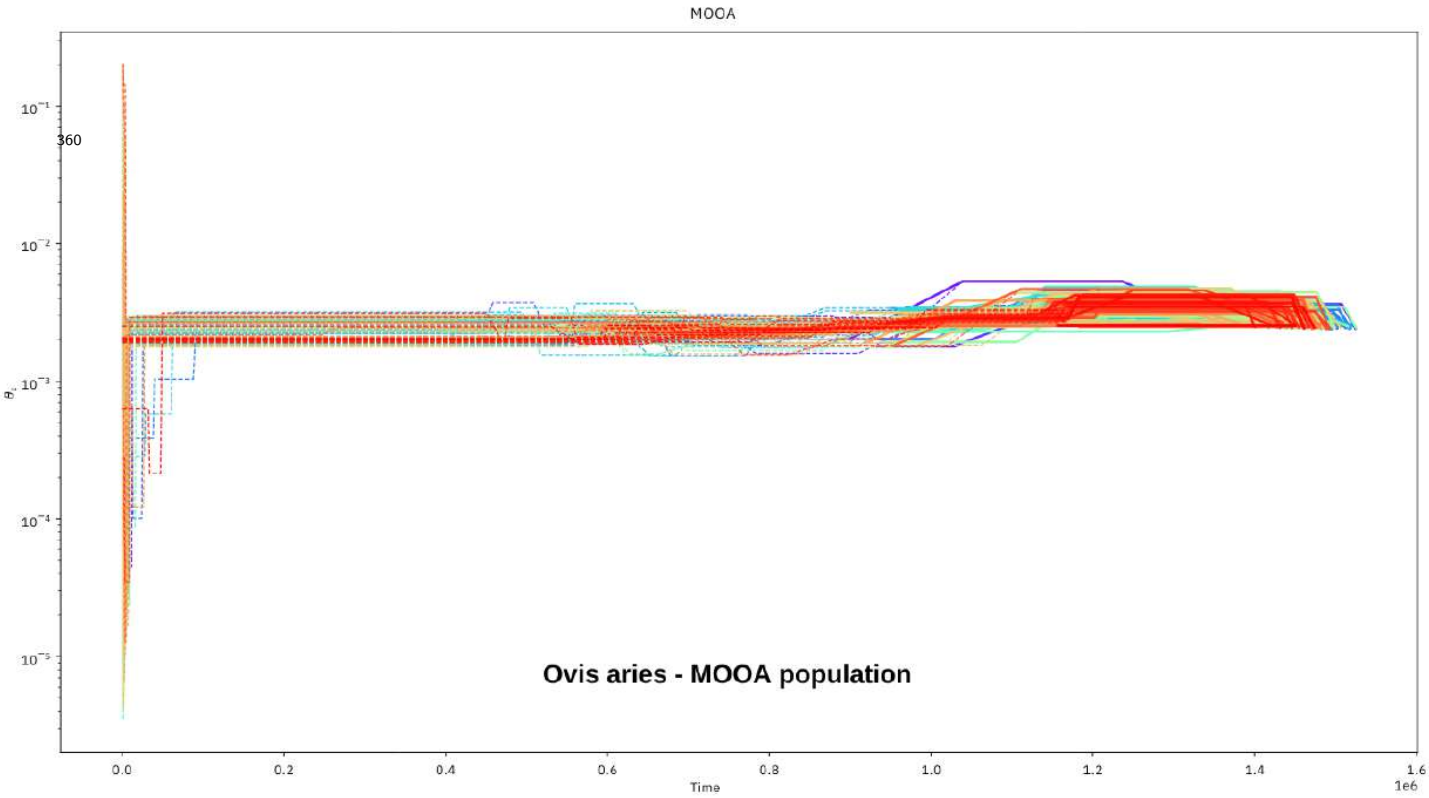

Nevis
10^{-3}
10^{-4}
θ_c
0
100000
200000
300000
400000
500000
600000
700000
Time
Chlorocebus sabaeus - Nevis population

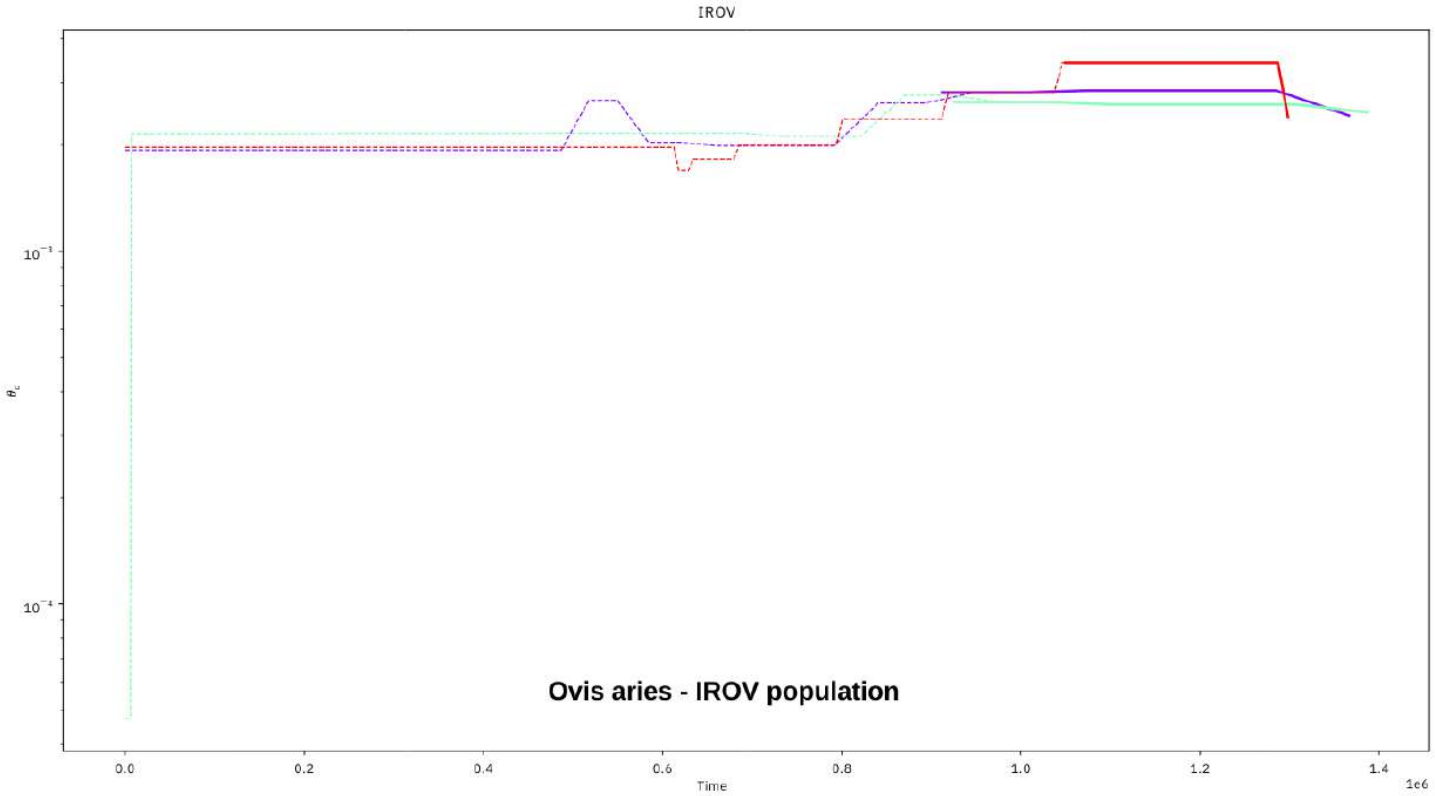

Zambia
10^{-3}
θ_E
Chlorocebus sabaeus - Zambia population
0
100000
200000
300000
400000
500000
600000
700000
Time

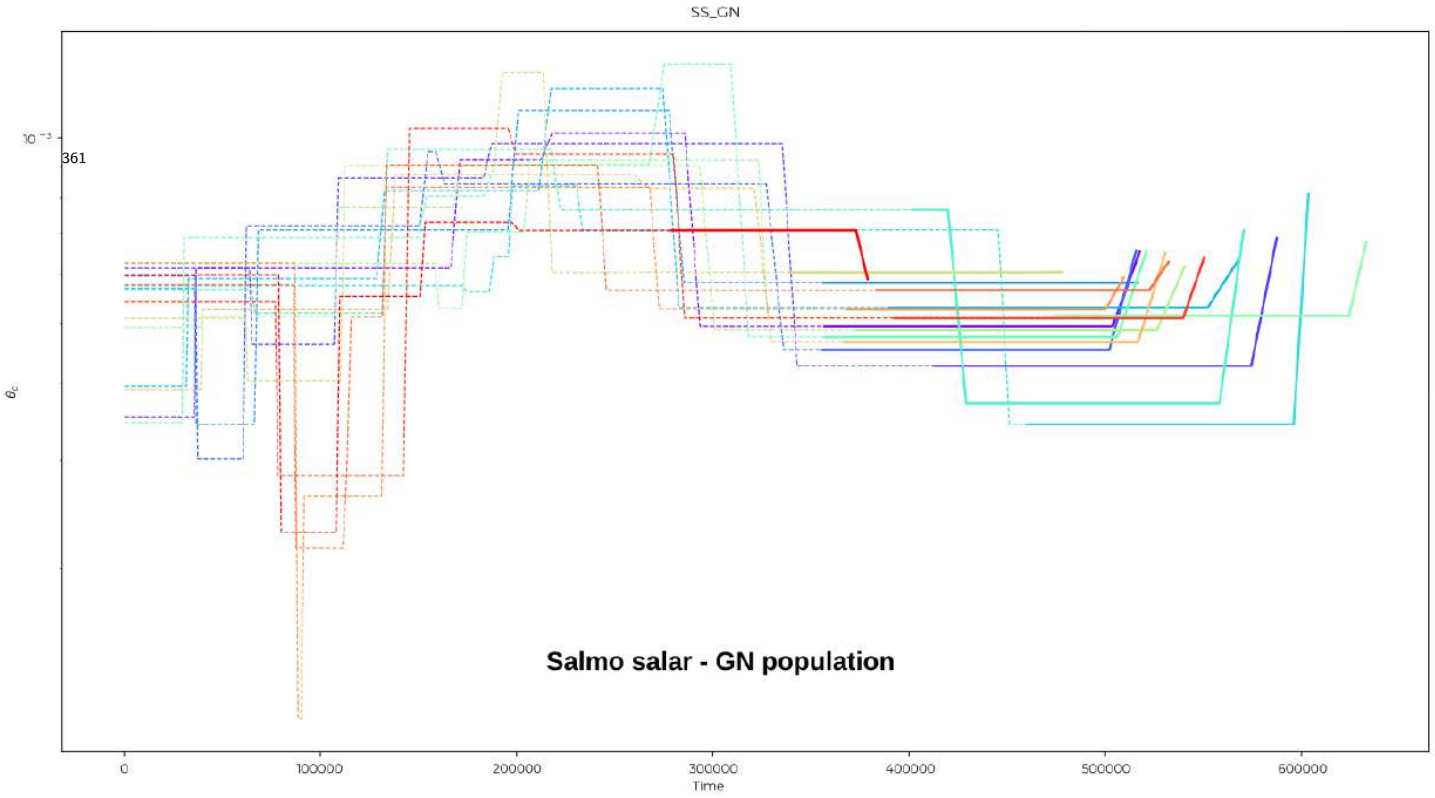

up
10^{-3}
10^{-4}
θ_E
Equus caballus - up population
0
100000
200000
300000
400000
500000
600000
Time

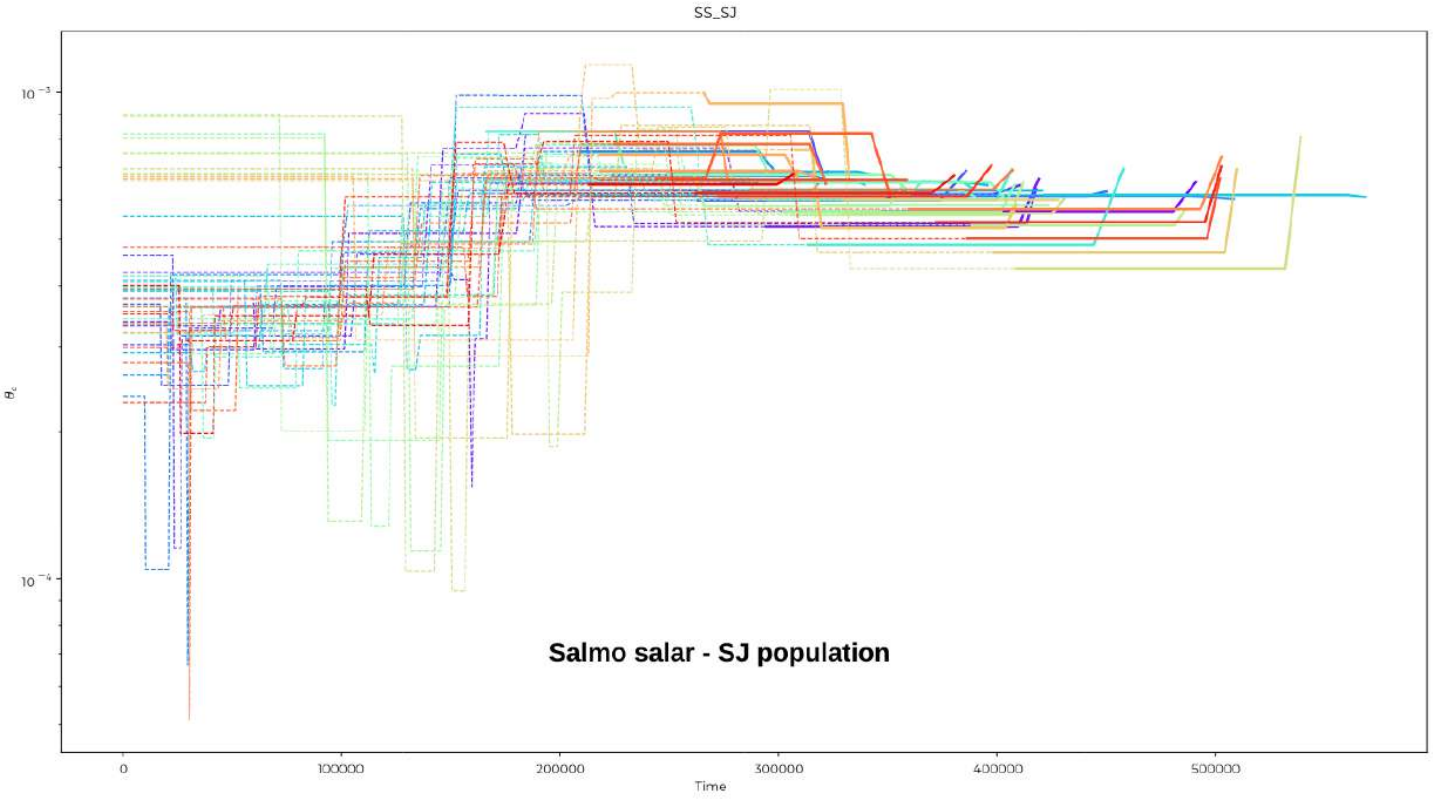

IROO
Ovis aries - IROO population
Time
1e6

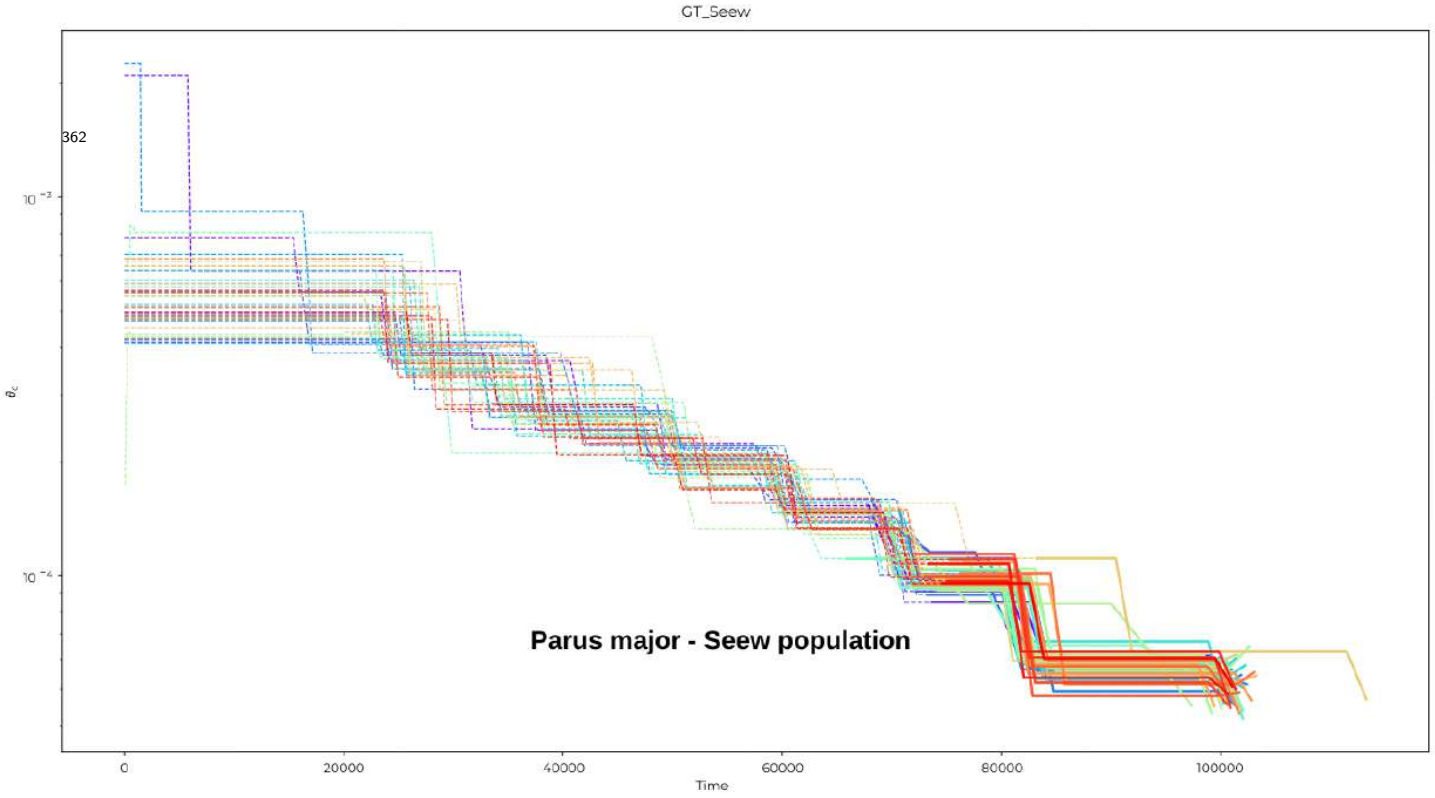

IROA
Ovis aries - IROA population
Time
1e6

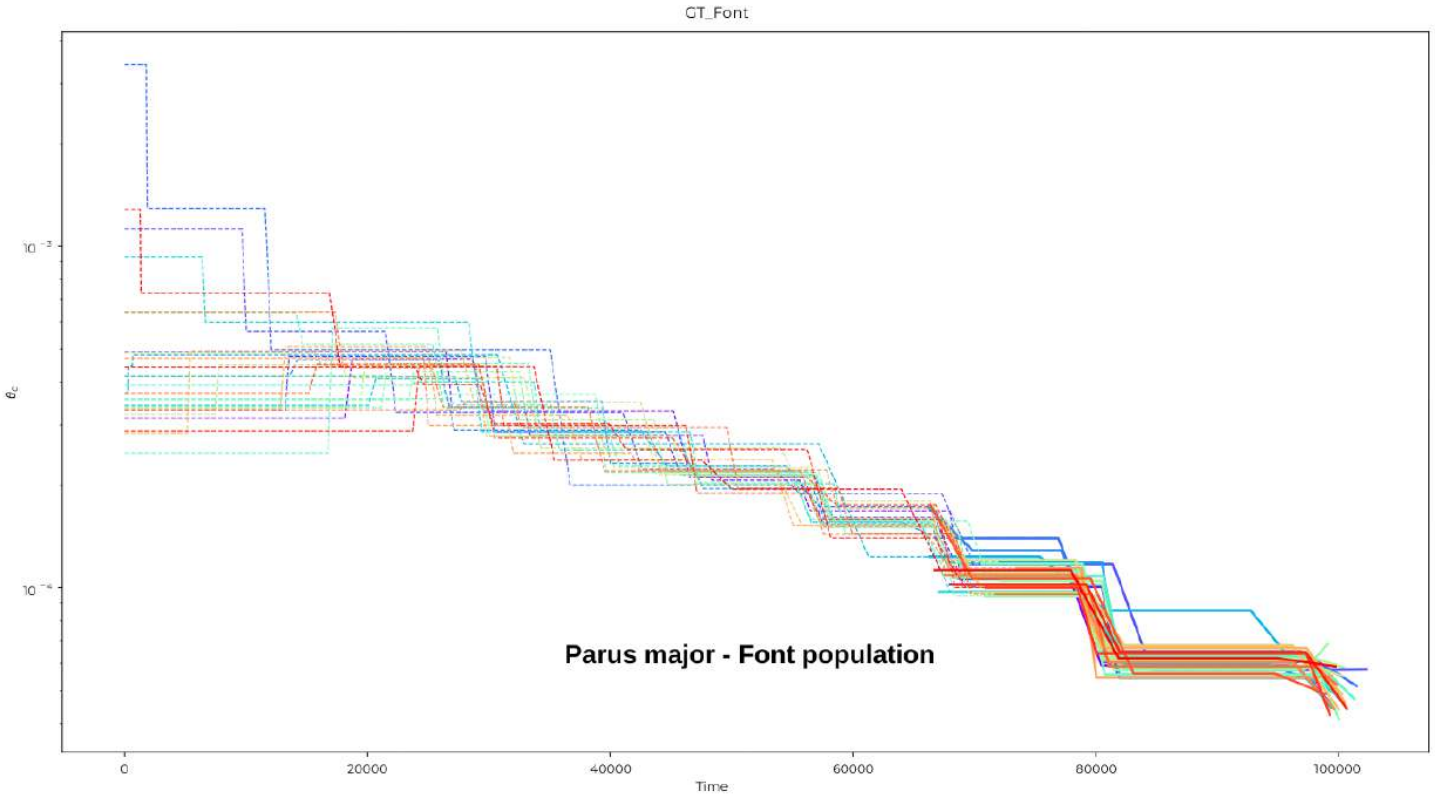

MOOA
360
Ovis aries - MOOA population
Time
1e6

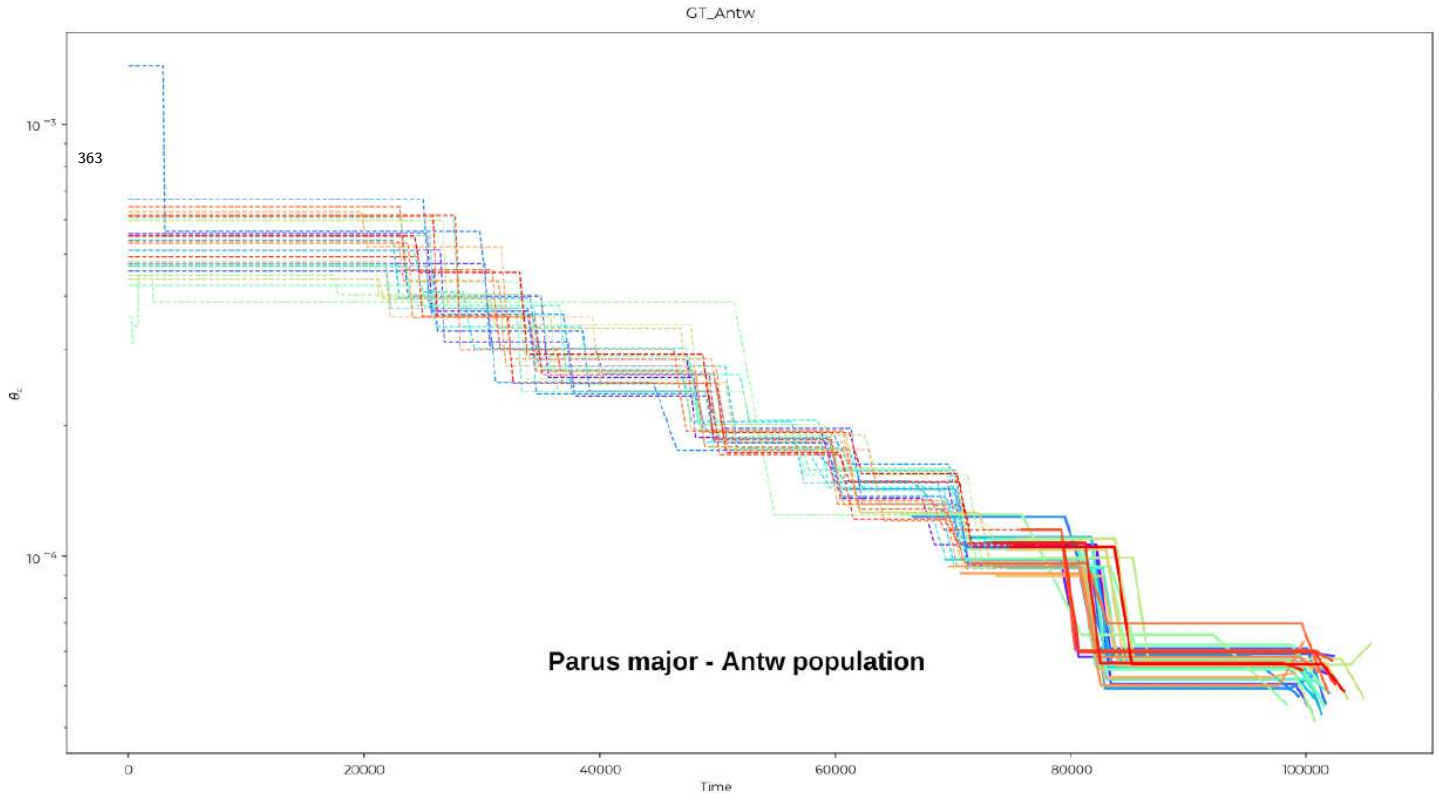

IROV
Ovis aries - IROV population
Time
1e6

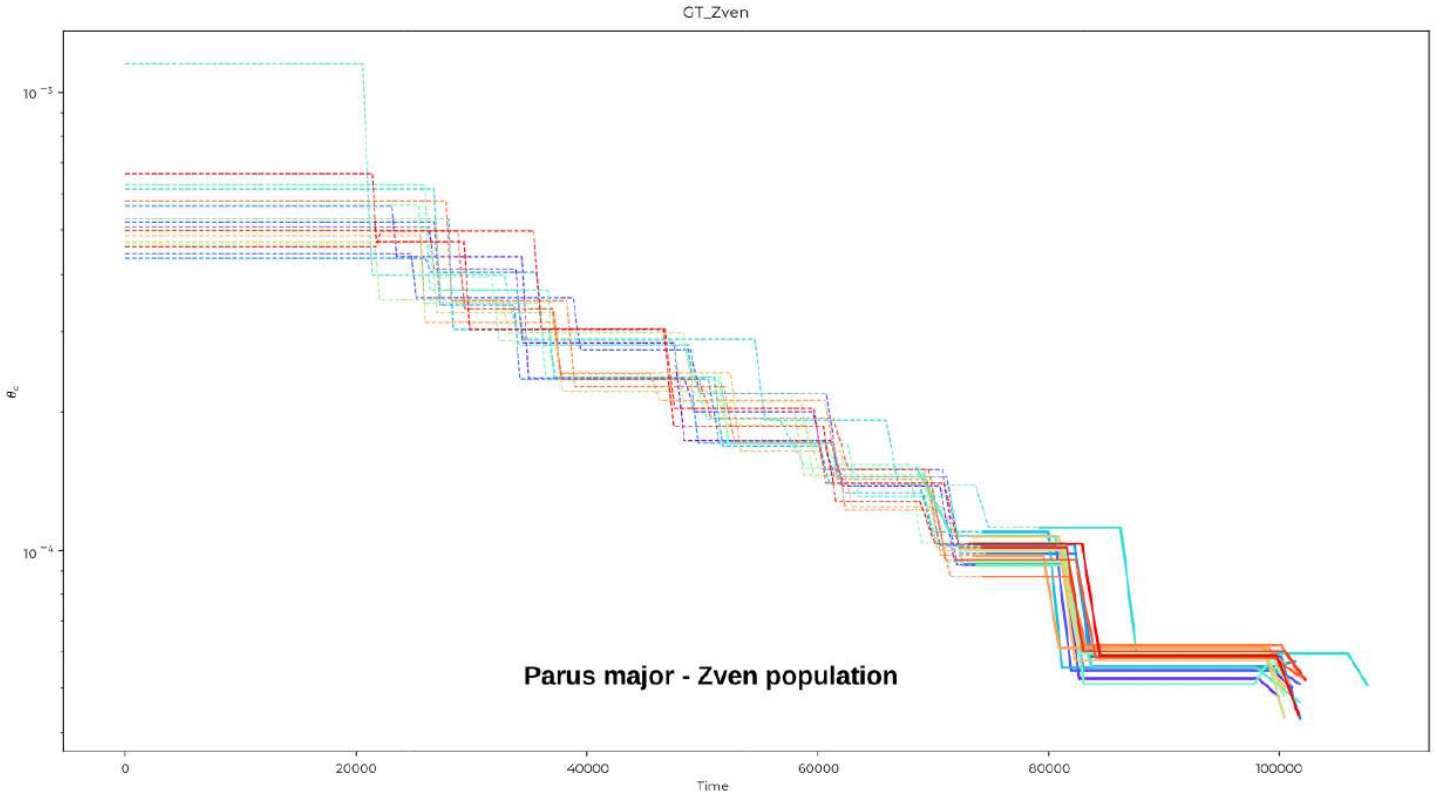

SS_GN
10^{-3}
θ_c
Salmo salar - GN population
0
100000
200000
300000
400000
500000
600000
Time


SS_SJ
10^{-3}
10^{-4}
θ_c
Salmo salar - SJ population
0
100000
200000
300000
400000
500000
Time


GT_Seew
10^{-3}
10^{-4}
θ_c
Parus major - Seew population
0
20000
40000
60000
80000
100000
Time


GT_Font
10^{-3}
10^{-4}
θ_c
Parus major - Font population
0
20000
40000
60000
80000
100000
Time


GT_Antw
363
10^{-3}
10^{-4}
θ_L
Parus major - Antw population
0
20000
40000
60000
80000
100000
Time


GT_Zven
10^{-3}
10^{-4}
θ_L
Parus major - Zven population
0
20000
40000
60000
80000
100000
Time

9.3 Annexe : Detecting diversifying selection for a trait from within and between-species genotypes and phenotypes

Journal of Evolutionary Biology, 2024, **XX**, 1–14
https://doi.org/10.1093/jeb/voae084
Advance access publication 11 July 2024
Research Article

JOURNAL OF
Evolutionary Biology
eseb

Detecting diversifying selection for a trait from within and between-species genotypes and phenotypes

T. Latrille[1], M. Bastian[2], T. Gaboriau[1], N. Salamin[1]

[1]Department of Computational Biology, Université de Lausanne, Lausanne, Switzerland
[2]Laboratoire de Biométrie et Biologie Evolutive, UMR5558, Université Lyon 1, Villeurbanne, France

Corresponding author: T. Latrille, Department of Computational Biology, Université de Lausanne, Lausanne, Switzerland.
Email: thibault.latrille@ens-lyon.org

Abstract

To quantify selection acting on a trait, methods have been developed using either within or between-species variation. However, methods using within-species variation do not integrate the changes at the macro-evolutionary scale. Conversely, current methods using between-species variation usually discard within-species variation, thus not accounting for processes at the micro-evolutionary scale. The main goal of this study is to define a neutrality index for a quantitative trait, by combining within- and between-species variation. This neutrality index integrates nucleotide polymorphism and divergence for normalizing trait variation. As such, it does not require estimation of population size nor of time of speciation for normalization. Our index can be used to seek deviation from the null model of neutral evolution, and test for diversifying selection. Applied to brain mass and body mass at the mammalian scale, we show that brain mass is under diversifying selection. Finally, we show that our test is not sensitive to the assumption that population sizes, mutation rates and generation time are constant across the phylogeny, and automatically adjust for it.

Keywords: quantitative genetics; trait evolution; selection; phylogenetics; population genetics

Introduction

Determining whether a trait is under a particular regime of selection has been a long-standing goal in evolutionary biology. Fundamentally, distinguishing neutral evolution from selection requires determining which selective regime is supported by the observed variation of traits or sequences. The variation of phenotypes (traits) and genotypes (sequences) can be observed at different scales, across different development stages at the individual level, across different individuals and populations at the species level, and finally across different species at the phylogenetic level. All these systems require different assumptions and methodologies, and the endeavor to determine the selective regime for a given trait has thus incorporated theories, methods, and developments across various fields of evolutionary biology such as quantitative genetics, population genetics, phylogenetics and comparative genomics (Lynch & Walsh, 1998; Walsh & Lynch, 2018).

Leveraging individual variations within the same species, genome-Wide association studies (GWAS) in humans have shown that traits are mostly polygenic (many loci associated with a given trait) and under stabilizing selection, while the loci affecting those traits are mostly pleiotropic (many traits associated with a given locus) with additive effects (Sella & Barton, 2019; Simons *et al.*, 2018). Given this genetic architecture of traits, from two diverging populations, it is possible to distinguish which traits have evolved under natural selection in controlled experimental settings, by performing genetic cross between individuals (Fraser, 2020). Across several populations, by contrasting both trait differentiation (Q_{ST}) and genetic differentiation (F_{ST}), so-called Q_{ST}–F_{ST} methods have been used to determine the selective regime and to quantify the strength of selection acting on a trait (Crnokrak & Merilä, 2001; Leinonen *et al.*, 2008). Q_{ST} higher than F_{ST} is interpreted as a signature of diversifying selection due to adaptation to different optimum trait values in the different populations. Contrarily, Q_{ST} lower than F_{ST} is interpreted as a signature of stabilizing selection (Lamy *et al.*, 2012). Other frameworks explicitly model genetic drift as a random process generating both trait and genetic differences between individuals and populations. This integrated framework can discriminate between selection and genetic drift as a cause of trait differentiation between populations of the same species (Ovaskainen *et al.*, 2011). However, regardless of the strengths and weaknesses of each method (Edelaar *et al.*, 2011; Ovaskainen *et al.*, 2011; Pujol *et al.*, 2008), tests of trait differentiation between populations are ultimately limited to recent local adaptation since they are based on the variation observed within a single species. To disentangle selection from neutral evolution, trait variation can also be observed at a larger time scale. For example, starting from the same ancestral population, divergent lineages accumulate phenotypic changes that will reach fixation in the population. These changes ultimately result in different mean trait values across lineages. Theoretically, the variance in mean trait value (between lineages) does increase linearly with time of divergence, and also proportionally to the trait variance at the population scale (Felsenstein, 1988; Lande, 1980a; Turelli, 1984). Empirically, this effect can be observed for genes with larger within-species variation in gene expression level, which exhibits a faster accumulation of divergence in mean expres-

Received July 10, 2024; revised June 14, 2024; accepted July 10, 2024

sion level (Khaitovich *et al.*, 2004). As an analogy, in the context of protein-coding DNA sequences, leveraging within species variation and divergence to a sister species is the crux of the McDonald and Kreitman (1991) test. In such a test, inflation of divergence to the sister species is compared to polymorphism within species, while neutral makers (usually synonymous sites) are used to determine the neutral expectation and thus are used for normalization. Altogether, both the trait variance and the evolution in mean value can be used to test for trait selection in a pair of species (Walsh & Lynch, 2018).

Alternatively, by accounting for the underlying relationships between several species, the selective regime for a quantitative trait can also be tested at the phylogenetic scale (Felsenstein, 1985). Under neutral evolution, the change in mean trait value along a given branch of the tree is normally distributed, with a variance proportional to divergence time (Felsenstein, 1985, 1988; Hansen & Martins, 1996). As a result, the mean trait value can be modeled as a Brownian process branching at every node of the tree (Harmon, 2018; Hansen & Martins, 1996). Reconstructing the trait variation along the whole phylogeny as a Brownian process can thus constitute a null model of neutral trait evolution. Deviations from the assumptions of the Brownian process are however well known. When trait variation is constrained because of optimum mean trait values across or between species, the pattern of evolution can be modeled by the Ornstein–Uhlenbeck processes, which is often interpreted as a signature of stabilizing selection (Catalán *et al.*, 2019; Hansen, 1997). Alternatively, a trend in the Brownian process (the tendency of a trait to evolve in a certain direction without fixed optimum) is interpreted as a signature of directional selection at the phylogenetic scale (Silvestro *et al.*, 2019). However, studies have shown that such comparative approaches are subject to different biases (Harmon, 2018). First, a trait under stabilizing selection for which the optimal trait value is also changing as a Brownian process will not deviate from a Brownian process, and thus be wrongly classified as neutral (Hansen & Martins, 1996). In other words, the better fit of a Brownian process does not necessarily constitute proof of the neutral model. Second, a better fit of a Brownian could be due to a trait evolving with a rate too low compared to the timespan on which it is measured (Grabowski *et al.*, 2023), and third, even for a trait evolving under a neutral regime, the Ornstein–Uhlenbeck process might sometimes be statistically preferred over a Brownian process due to sampling artifacts (Cooper *et al.*, 2016; Price *et al.*, 2022; Silvestro *et al.*, 2015). Those limitations, altogether with the use of mean trait estimates leaving out the variance in traits between individuals, easily generate misclassification of selection from methods at the phylogenetic scale.

At the frontier between micro and macro-evolution, comparative methods at the phylogenetic scale have acknowledged the importance of modeling within-species variation together with changes in mean trait value to either describe measurement errors (Hansen & Bartoszek, 2012; Lynch, 1991), incorporate values for individuals (Felsenstein, 2008) or to scale the rate of change in mean trait value (Gaboriau *et al.*, 2020, 2023; Kostikova *et al.*, 2016). Across many species, within-species variation has also been used to infer diversifying selection by estimating the ratio of between to within species variation of many traits and test for deviation from the average ratio across traits (Rohlfs & Nielsen, 2015; Rohlfs *et al.*, 2014). Here, our goal was again to use both variances between and within species to determine the selective regime of a quantitative trait. We build a novel framework that integrates trait variation at the phylogenetic and population scales together with estimates of nucleotide sequence variations at both scales. It allowed us to define an expected ratio of normalized variance between and within species while setting the threshold of this ratio for neutral, stabilizing, and diversifying selection. The ratio that we propose can be considered as a neutrality index for any quantitative trait (Lynch, 1990), while articulating trait and nucleotide variation within and between species. Importantly, our neutrality index also leverages nucleotide divergence and polymorphism to normalize trait variation at both scales, such that it does not require estimating population size (within-species) or speciation time (between species). From the field of population genetics, while Q_{ST}–F_{ST} methods and their derivatives ultimately seek trait differentiation among different populations from the same species (Ovaskainen *et al.*, 2011; Pujol *et al.*, 2008), our study can be seen as their macro-evolutionary analog to account for phylogenetic relationships between species. From the field of phylogenetics, our study can be seen as an alternative to the EVE model (Rohlfs & Nielsen, 2015; Rohlfs *et al.*, 2014) for a single trait, where we set a threshold for neutral evolution by leveraging species nucleotide polymorphism and divergence.

Materials and methods

Neutrality index for a quantitative trait

While observing trait variations across individuals of several species, we ask if the variation within species compared to variation between species is compatible with neutral evolution or not. In statistical terms, this can also be framed as: Is the variance of means equal to the mean of variances? The difficulty in such a study is that individuals are not independent samples, but are from species that diverged at different times. By reviewing theoretical expectations and leveraging nucleotide sequence variations, the goal of this section is thus to obtain normalized trait variation between and within species that are equal if the trait is neutral. Here we denote these normalized trait variations as respectively σ^2_{W} for within species and as σ^2_{B} for between species.

Within-species trait variations

For a given trait, the genetic architecture is mainly defined by the number of loci encoding the trait (L) and the random additive effect of a mutation on the trait (a). For a diploid individual, the mutational variance (V_{M}) is the rate at which new mutations contribute to the trait variance per generation. As shown in Lande (1979, 1980b), V_{M} is a function of the mutation rate per locus per generation (μ) and the genetic architecture of the trait as

$$V_{\mathrm{M}} = 2\mu \cdot L \cdot \mathbb{E}\left[a^2\right]. \quad (1)$$

While in an infinitesimal model mutations supply new genetic variants, random genetic drift depletes standing variation (Barton *et al.*, 2017; Sella & Barton, 2019; Turelli, 2017). For a neutral trait at equilibrium between mutation and drift (Lynch *et al.*, 1998), the additive genetic variance in

a species (V_A) is a function of the mutational variance (V_M) and the effective number of individuals in the population (N_e):

$$V_A = 2N_e \cdot V_M, \tag{2}$$
$$= 4N_e \cdot \mu \cdot L \cdot \mathbb{E}\left[a^2\right] \text{ from } \textit{Equation 1}. \tag{3}$$

For any neutral genomic region of interest, the nucleotide diversity, π, is the average number of differences between pairs of sequences drawn at random, which is also equal to the sum of expected heterozygosities over all nucleotide sites (Tajima, 1989). Any segregating mutations will eventually reach fixation or extinction due to random genetic drift and π is also at a balance between mutations and drift. As shown in Tajima (1989), π is a function of the mutation rate (u, per nucleotide site per generation) and the effective population size (N_e):

$$\pi = 4N_e \cdot u. \tag{4}$$

To remove the effect of N_e, we define σ^2_W as the ratio of additive genetic variance of the trait (V_A) over π of any neutral genomic region of interest. After simplification, σ^2_W is then solely a function of the underlying genetic architecture as

$$\sigma^2_W \overset{\text{def}}{=} \frac{V_A}{\pi}, \tag{5}$$
$$= \frac{4N_e \cdot \mu \cdot L \cdot \mathbb{E}\left[a^2\right]}{4N_e \cdot u} \text{ from } \textit{Equations 1} \text{ and } 4, \tag{6}$$
$$= \frac{\mu \cdot L \cdot \mathbb{E}\left[a^2\right]}{u}. \tag{7}$$

If V_A is not empirically accessible, it can be related to the observed phenotypic variance (V_P), multiplied by narrow-sense heritability of the trait (h^2), as (Hill *et al.*, 2008)

$$V_A = h^2 \cdot V_P. \tag{8}$$

Which leads to σ^2_W being a function of V_P and h^2 instead of V_A as

$$\sigma^2_W = \frac{h^2 \cdot V_P}{\pi} \text{ from definition } \textit{Equations 5} \text{ and } 8, \tag{9}$$
$$= \frac{\mu \cdot L \cdot \mathbb{E}\left[a^2\right]}{u} \text{ from } \textit{Equation 7} \tag{10}$$

Between-species trait variations

For a given species i, we denote by $\bar{P}_i$ the mean value of the trait across the individuals. If the trait is neutral and encoded by many loci as assumed by the infinitesimal model, $\bar{P}_i$ evolves as a Brownian process (Felsenstein, 1985; Hansen & Martins, 1996). Given a phylogenetic tree, for a pair of species i and j from this tree, we denote as $t_{i,j}$ the number of generations between the root of the tree and the most recent common ancestor of taxa i and j. Then, the covariance between $\bar{P}_i$ and $\bar{P}_j$ depends on $t_{i,j}$ as given by Hansen & Martins (1996)

$$\text{cov}\left(\bar{P}_i, \bar{P}_j\right) = \frac{V_A}{N_e} \cdot t_{i,j} \tag{11}$$
$$= 4t_{i,j} \cdot \mu \cdot L \cdot \mathbb{E}\left[a^2\right], \text{ from } \textit{Equation 3}. \tag{12}$$

Moreover, for any genomic region under neutral evolution, some mutations will eventually reach fixation due to random genetic drift, resulting in a substitution of a nucleotide at the species level. The probability of fixation ($\mathbb{P}_{\text{fix}}$) of a neutral mutation is $1/2N_e$ (Kimura, 1962). We can derive the substitution rate (q, per nucleotide site per generation) as the number of newly arisen mutations ($2N_e \cdot u$) multiplied by the probability of fixation for each newly arisen mutations $\mathbb{P}_{\text{fix}}$ (Kimura, 1968), giving:

$$q = 2N_e \cdot u \cdot \mathbb{P}_{\text{fix}}, \tag{13}$$
$$= 2N_e \cdot u \cdot \frac{1}{2N_e}, \tag{14}$$
$$= u. \tag{15}$$

That is, if mutations are neutral, the rate of substitution per generation within a genomic region equals the rate at which new mutations arise per generation for the same genomic region, reviewed by McCandlish and Stoltzfus (2014).

Next, we denote $d_{i,j}$ as the nucleotide divergence between the root of the tree and the most recent common ancestor of taxa i and j. In other words, $d_{i,j}$ is the expected number of substitutions per nucleotide site during the $t_{i,j}$ generations. Assuming that no multiple substitutions occurred at the same site, $d_{i,j}$ is the number of generations ($t_{i,j}$) multiplied by the nucleotide substitution rate per generation (q):

$$d_{i,j} = t_{i,j} \cdot q \tag{16}$$
$$= t_{i,j} \cdot u \text{ from } \textit{Equation 15}. \tag{17}$$

To remove the effect of the number of generations ($t_{i,j}$) first, and to also equate to σ^2_W (*Equation 7*), we define σ^2_B as the covariance in the mean trait value ($\text{cov}\left(\bar{P}_i, \bar{P}_j\right)$) normalized by 4 times the nucleotide divergence of any neutral genomic region ($4d_{i,j}$). After simplification, σ^2_B is also solely a function of the underlying genetic architecture as

$$\sigma^2_B \overset{\text{def}}{=} \frac{\text{cov}\left(\bar{P}_i, \bar{P}_j\right)}{4d_{i,j}}, \tag{18}$$
$$= \frac{4t_{i,j} \cdot \mu \cdot L \cdot \mathbb{E}\left[a^2\right]}{4t_{i,j} \cdot u} \text{ from } \textit{Equations 12} \text{ and } 17, \tag{19}$$
$$= \frac{\mu \cdot L \cdot \mathbb{E}\left[a^2\right]}{u}. \tag{20}$$

In *Equation 20*, we show that the covariance in mean trait value between a pair of species ($\text{cov}\left(\bar{P}_i, \bar{P}_j\right)$) does increase linearly with shared nucleotide divergence ($d_{i,j}$), if the trait and sequences are neutrally evolving and the genetic architecture of the trait has not changed. Importantly, since the number of generations is the ratio of time divided by generation time (average time between two consecutive generations), removing the effect of the number of generations in *Equation 20* also removes the effect of both time and generation time.

Neutrality index

The variability between either individuals or species can be obtained for both quantitative traits and genomic sequences. At the population level, the variability of the trait between individuals can be combined with the nucleotide diversity of any neutrally evolving genomic region to obtain σ^2_W. At the phylogenetic level, the variability of the mean trait value between species can be combined with the nucleotide divergence of any neutrally evolving genomic region to obtain σ^2_B.

If the trait is neutrally evolving and the genetic architecture of the trait has not changed along the phylogenetic tree, we thus have

$$\frac{\sigma^2_{\mathrm{B}}}{\sigma^2_{\mathrm{W}}} = \frac{\mathrm{cov}\left(\bar{P}_i, \bar{P}_j\right)}{4d_{i,j}} \cdot \frac{\pi}{h^2 \cdot V_P} \text{ by definition and } \textit{Equations 9 and 18}, \quad (21)$$

$$= \frac{\mu \cdot L \cdot \mathbb{E}\left[a^2\right]}{u} \cdot \frac{u}{\mu \cdot L \cdot \mathbb{E}\left[a^2\right]} \text{ from } \textit{Equations 10 and 20}, \quad (22)$$

$$= 1. \quad (23)$$

We define a neutrality index ρ as

$$\rho \overset{\text{def}}{=} \frac{\sigma^2_{\mathrm{B}}}{\sigma^2_{\mathrm{W}}}, \quad (24)$$

which will equal to 1 for a trait evolving neutrally. Both σ^2_{B} and σ^2_{W} can be estimated using quantitative trait and genomic sequences within and between species, while neither the mutation rates (μ and u), nor the effective population size (N_e), generation time or time of divergence ($t_{i,j}$) need to be estimated. Moreover, the nucleotide sequence from which π and $d_{i,j}$ are obtained should be neutrally evolving, but they are not necessarily linked to the quantitative trait under study.

Estimation

We hereby seek to obtain point estimates of σ^2_{B}, σ^2_{W} and ultimately ρ. For each species with data available, σ^2_{W} as defined in *Equation* 9 can be seen as a replicate sample. Thus, σ^2_{W} can be obtained by averaging out across all the sampled species. On the other hand, σ^2_{B} such as as defined in *Equation* 18 only refers to a pair of species, and thus must be generalized to account for different species divergence, as is done in the comparative framework (Felsenstein, 1985; O'Meara *et al.*, 2006). Generally, σ^2_{B} can thus be seen as an estimate of the rate of evolution of the quantitative trait along a phylogenetic tree, when the tree is measured in units of $4d$ (d being the nucleotide divergence). As such, any phylogenetic comparative methods that allow the estimation of phenotypic rates of evolution on a tree scaled by $4d$, instead of time as is usually the case, can be used to estimate σ^2_{W}. We provide a maximum likelihood estimate for ρ as well as a Bayesian estimate to derive posterior probabilities that the null model of neutrality (i.e. $\rho = 1$) is rejected.

Maximum likelihood estimate

At the phylogenetic scale, for n taxa in the tree, $\boldsymbol{D}$ ($n \times n$) is the symmetric distance matrix computed from the branch lengths and the topology of the phylogenetic tree. The diagonal $\boldsymbol{D}_{i,i}$ represents the total nucleotide divergence from the root of the tree to each taxon (i). The off-diagonal elements ($\boldsymbol{D}_{i,j} = d_{i,j}$) are the distances between the root and the most recent common ancestor of taxa i and j, as in Equation 17. The mean trait value at the root of the tree (ϕ) can be estimated from the $n \times 1$ vector of mean trait values $\bar{\boldsymbol{P}}$ at the tips of the tree using maximum likelihood (O'Meara *et al.*, 2006):

$$\phi = \left(\mathbf{1}^\intercal \times \boldsymbol{D}^{-1} \times \mathbf{1}\right)^{-1} \cdot \left(\mathbf{1}^\intercal \times \boldsymbol{D}^{-1} \times \bar{\boldsymbol{P}}\right), \quad (25)$$

where $\mathbf{1}$ is an $n \times 1$ column vector of ones.

Finally, between-species variation σ^2_{B} is estimated as (O'Meara *et al.*, 2006):

$$\sigma^2_{\mathrm{B}} = \frac{1}{4} \frac{\left(\bar{\boldsymbol{P}} - \phi \cdot \mathbf{1}\right)^\intercal \times \boldsymbol{D}^{-1} \times \left(\bar{\boldsymbol{P}} - \phi \cdot \mathbf{1}\right)}{n-1}. \quad (26)$$

For a given species i with inter-individual data available, additive genetic variance of a trait ($V_{\mathrm{A},i}$) is the product of heritability (h^2_i) and phenotypic variance ($V_{P,i}$). The ratio of $V_{\mathrm{A},i}$ over nucleotide diversity of neutrally evolving sequences (π_i) is a sample estimate of σ^2_{W}. Averaged across all species, we obtain the estimate σ^2_{W} as

$$\sigma^2_{\mathrm{W}} = \frac{1}{n}\sum_{i=1}^{n} \frac{V_{\mathrm{A},i}}{\pi_i} = \frac{1}{n}\sum_{i=1}^{n} \frac{V_{P,i} \cdot h^2_i}{\pi_i}. \quad (27)$$

As depicted in Figure 1, the neutrality index is estimated as

$$\rho = \frac{\sigma^2_{\mathrm{B}}}{\sigma^2_{\mathrm{W}}}. \quad (28)$$

Multivariate Brownian process

In the previous section, ρ is estimated independently for each trait of interest. Here we generalize to K traits co-varying along the phylogenetic tree, since simultaneously estimating all σ^2_{B} allows improving their estimation Adams & Collyer (2018). More specifically, trait variation along the phylogenetic tree is modeled as a K-dimensional Brownian process $\boldsymbol{\mathcal{B}}$ ($1 \times K$) starting at the root and branching along the tree topology (Huelsenbeck & Rannala, 2003; Lartillot & Poujol, 2011; Lartillot & Delsuc, 2012; Latrille *et al.*, 2021). The rate of change of the Brownian process is determined by the positive semi-definite and symmetric covariance matrix between traits Σ ($K \times K$). The branch lengths of the tree used to model the Brownian process runs is measured in units of $4d$ (d being the nucleotide divergence). The off-diagonal elements of Σ are the covariance between traits, and the diagonal elements are the variance of each trait when measured in $4d$ units, and thus equate to σ^2_{B} (see online supplementary material Section S2.1). Of note, modeling trait evolution as a multi-dimensional process is reliable only if $K \ll n$, meaning that the number of species is largely superior to the number of traits (Adams & Collyer, 2018). Thus, relying on a K-dimensional process should be reserved for a handful of allometric traits (e.g., brain mass and body mass). If K is large, the traits are better tested independently each with a 1-dimensional Brownian process, which is a specific case of the multi-dimensional process.

Bayesian estimate

The Bayesian framework allows obtaining the posterior distribution of neutrality index (ρ) for traits of interest. We used the *BayesCode* software to model K-dimensional Brownian processes along a phylogenetic tree (Latrille *et al.*, 2021). With an inverse Wishart distribution as the prior on the covariance matrix, the posterior on Σ, conditional on $\mathcal{B}$ is also an invert Wishart distribution (see online supplementary material Section S2.2). We used Metropolis-Hastings algorithm to sample $\boldsymbol{\mathcal{B}}$, while the posterior distribution of Σ is sampled using Gibbs sampling. For each trait and each species, the prior on heritability (h^2) for each species is set as a uniform distribution with user-defined boundaries. Heritability and phenotypic variance for each trait are combined with nucleotide

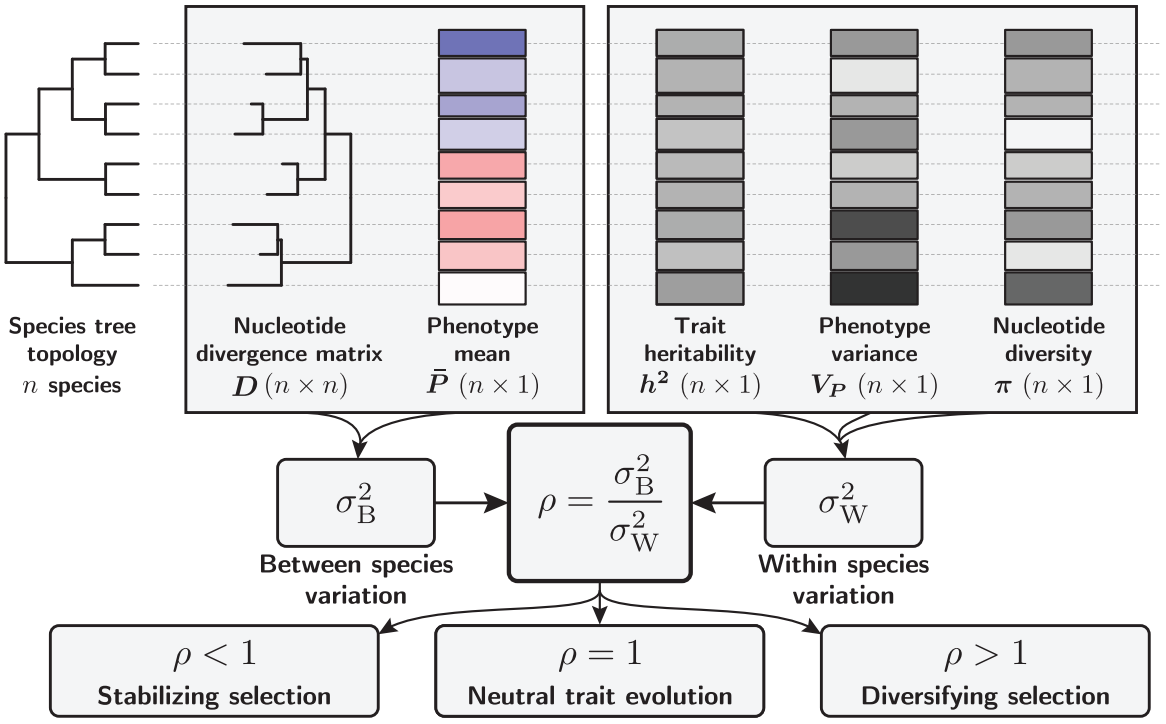


Figure 1. Between species, the change along the phylogeny of the mean phenotypic trait allows the estimation of between-species trait variation, σ^2_{B}, which is normalized by nucleotide divergence. Within species, the genetic variance allows the estimation of within-species trait variation, σ^2_{W}, which is normalized by nucleotide diversity. ρ is the ratio of σ^2_{B} over σ^2_{W}. Under neutral evolution, ρ is expected to be equal to one. Under diversifying selection, the trait is heterogeneous between species, but homogeneous within species, leading to ρ greater than one. Under stabilizing selection, the trait is homogeneous between species, leading to ρ smaller than one. Importantly, the sequence from which nucleotide diversity and divergence are estimated should be neutrally evolving, but they are not necessarily linked to the quantitative trait under study, they allow for discarding the confounding effect on mutation rate diversity, population size and divergence time.

diversity to compute σ^2_{W} for each species before being averaged across species (as in *Equation* 27). From σ^2_{W} estimated independently for each trait and the diagonal elements of $\boldsymbol{\Sigma}$ (i.e., the σ^2_{B} for each trait), the posterior distribution of ρ (as in *Equation* 28) is obtained for each trait. The posterior distribution of ρ thus allows testing for deviation from neutrality (Figure 1), for example, by computing $\mathbb{P}[\rho > 1]$ to test for evidence of diversifying selection and $\mathbb{P}[\rho < 1]$ to test for evidence of stabilizing selection.

Applicability to empirical data

Our method assumes that the narrow-sense heritability (h^2) of a trait is known such as to estimate additive genetic variance (V_{A}) from phenotypic variance (V_P) as $V_{\mathrm{A}} = h^2 \cdot V_P$. Fortunately, if heritability is not known, the test for diversifying selection can still be performed, although it is underpowered. Indeed, if the additive genetic variance is substituted by phenotypic variance, it is equivalent to assuming complete heritability ($h^2 = 1$). Because $h^2 \leq 1$ by definition, we overestimate the within-species variation and thus underestimate ρ. It is, however, possible to test for diversifying selection because testing for $\rho > 1$ while using phenotypic variance instead of additive genetic variance means that knowing the additive genetic variance would have only increased the evidence for diversifying selection. Similarly, using the broad-sense heritability (H^2) instead of narrow-sense heritability (h^2) results in an underestimation of ρ since $h^2 \leq H^2$ and thus can be used to detect diversifying selection if h^2 is not available. Additionally, empirical estimates of h^2 are surprisingly stable across species and fall within the range of 0.2-0.5 in a vast majority of phenotypic traits tested (Hansen *et al.*, 2011; Hansen & Pélabon, 2021). Thus, if available, such prior knowledge on h^2 can be leveraged instead of assuming complete heritability to increase the statistical power to detect diversifying selection.

In contrast to the test of diversifying selection, the test for stabilizing selection is invalid if ρ is underestimated. Several assumptions made by our test might not hold on empirical data and their consequences on the neutrality index and the test that can be performed are shown in Table 2.

Simulation

We tested the performance of our neutrality index (ρ) to detect selection on a quantitative trait using simulations. We performed simulations under different selective regimes (neutral, stabilizing, diversifying), different demographic histories (constant or fluctuating population size) and different evolution of the mutation rate (constant or fluctuating). Simulations were individual-based and followed a Wright–Fisher model with mutation, selection and drift for a diploid population including speciation along a predefined ultrametric phylogenetic tree (Figure 2). Each individual phenotypic value was the sum of genotypic value and an environmental effect. The environmental effect was normally distributed with variance V_{E}. We assumed that the genotypic value was encoded by $L = 5,000$ loci, with each locus contributing an additive effect that was normally distributed with standard deviation $a = 1$

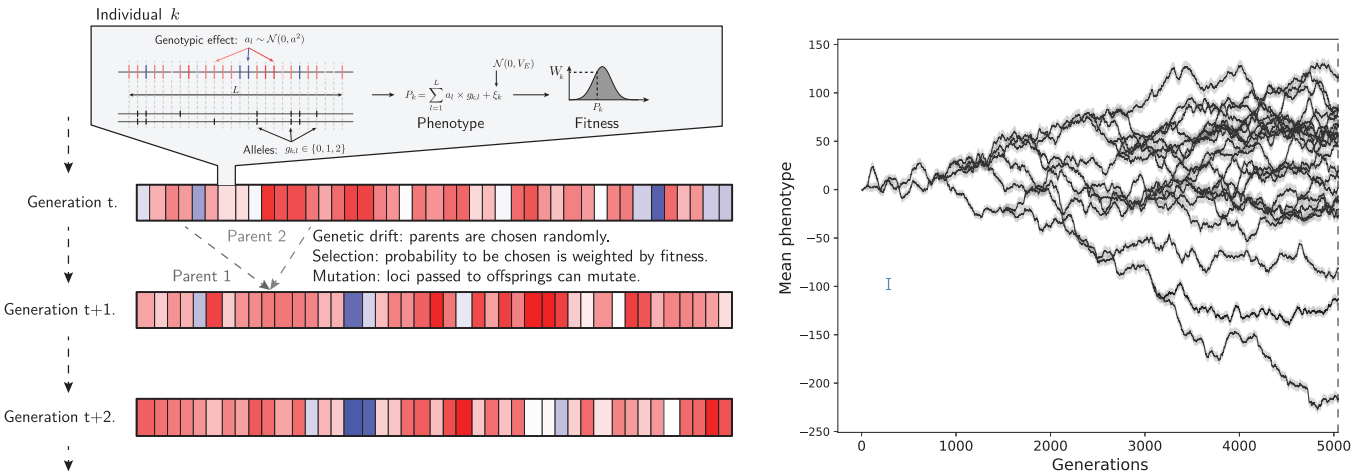


Figure 2. Wright–Fisher simulations with mutation, selection and drift. Left panel: For a given individual, the trait phenotypic value is the sum of genotypic value and a environmental effect (standard deviation V_E). The trait's genotypic value is encoded by L independent loci (meaning no linkage), with each locus contributing additively to the genotypic value. Parents are selected for reproduction to the next generation according to their phenotypic value, with a probability proportional to their fitness. Mutations are drawn from a Poisson distribution, with each locus having a probability μ to mutate. Drift is modeled by the resampling of parents. Right panel: examples of a trait evolving along a phylogenetic tree, with the mean phenotype (black line) and the variance of the trait genotypic value (gray area).

(Figure 2 and for the theoretical formulation see online supplementary material Section S1.1 and Figure S1). We assumed a trait with a narrow-sense heritability of $h^2 = 0.2$ and computed the theoretical V_E accordingly (see online supplementary material Section S1.1). Assuming a diploid panmictic population of size $N_e = 50$ at the root of the tree, and with non-overlapping generations, we simulated explicitly each generation along an ultrametric phylogenetic tree. For each offspring, the number of mutations was drawn from a Poisson distribution with mean $2 \cdot \mu \cdot L$, with the mutation rate per locus per generation μ. From the empirical mammalian dataset (see next section), we computed an average nucleotide divergence from the root to leaves of 0.18 and average genetic diversity of 0.00276. We scaled parameters in our simulations to fit plausible values for mammals. We thus used a nucleotide mutation rate of $u = 0.00276/4N_e = 1.38 \times 10^{-5}$ per site per generation and a total of $0.18/1.38 \times 10^{-5} = 13,500$ generations from root to leaves, and the number of generations along each branch was proportional to the branch length. We set $\mu = u$ without loss in generality since the genetic architecture (L and a) is assumed constant in the simulator.

The changes in μ and N_e along the lineages were both modeled by a Brownian process on the log scale (log-μ and log-N_e), leading to geometric Brownian motion on the linear scale (μ and N_e). These processes are parameterized as $\mathcal{B}\left(0, \sigma_\mu = 0.0086\right)$ and $\mathcal{B}\left(0, \sigma_{N_e} = 0.0086\right)$, which, if counted across $13,500$ generations, leads to a standard deviation of $0.0086 \cdot \sqrt{13,500} = 1.0$. In other words, the deviation in log-N_e and log-μ between the extant species and the root is 1.0. An Ornstein–Uhlenbeck process was overlaid to the instant value of log-N_e provided by the geometric Brownian process to account for short-term changes between generations (OU $\left(0, \sigma_{N_e} = 0.1, \theta_{N_e} = 0.9\right)$). The geometric Brownian motion accounted for long-term fluctuations (low rate of changes σ_{N_e} but unbounded), while the Ornstein–Uhlenbeck introduced short-term fluctuations (high rate of changes σ_{N_e} but bounded and mean-reverting). The simulation started from an initial sequence at equilibrium at the root of the tree and, at each node, the process was split until it finally reached the leaves of the tree. From a speciation process perspective, this was equivalent to an allopatric speciation over one generation.

At each generation, parents were randomly sampled with a weight proportional to their fitness (W). Selection was modeled as a one-dimensional Fisher's geometric landscape, with the fitness of an individual being a monotonously decreasing function of the distance between the individual and the optimal phenotype (Blanquart & Bataillon, 2016; Tenaillon, 2014). More specifically, the fitness of an individual was given by $W = e^{(P-\lambda)^2/\alpha}$, where P was the trait value of the individual, $\lambda = 0.0$ was the optimal trait value, and $\alpha = 0.02$ was the strength of selection. Mutations were considered as a displacement of the phenotype in the multidimensional space. Beneficial mutations moved the phenotype closer to the optimum, while deleterious mutations moved it further away. Stabilizing selection was implemented by fixing the optimum phenotype to a single value ($\lambda = 0.0$). Diversifying selection was implemented by allowing the optimum phenotype to move along the phylogenetic tree as a geometric Brownian process (Hansen, 1997) ($\lambda \sim \mathcal{B}\left(0, \sigma_\lambda = 1.0\right)$). Neutral evolution was implemented by flattening the fitness landscape ($W = 1$), which meant that each individual had the same probability of being sampled at each generation.

Nucleotide diversity (π) was measured as the heterozygosity of neutral markers that were simulated along the phylogenetic tree but not linked to the trait simulated. Nucleotide divergence (d) was measured as the number of substitutions per site of neutral markers along the branches of the phylogenetic tree. The additive genetic variance was measured

as phenotypic variance multiplied by heritability. Heritability was estimated from the slopes of the regression of offspring's phenotypic trait values on parental phenotypic trait values (Lynch & Walsh, 1998) averaged over the last 10 simulated generations. Heritability was thus not a given parameter of the simulations, but rather measured as it would be in empirical data.

Empirical dataset

We analyzed a dataset of body and brain masses from mammals. The log-transformed values of body and brain masses were taken from Tsuboi *et al.* (2018). We removed individuals not marked as adults and split the data into males and females due to sexual dimorphism in body and brain masses. We discarded species with only one representative since phenotypic variance cannot be estimated. The mammalian genomic data are gathered from the Zoonomia project (Genereux *et al.*, 2020). More specifically, nucleotide divergence is estimated on a set of neutral markers in Foley *et al.* (2023), and with nucleotide diversity measured as heterozygosity in Wilder *et al.* (2023).

We also analyzed a dataset of primate species, with the nucleotide variation obtained from Kuderna *et al.* (2023) and the quantitative trait variation also from Tsuboi *et al.* (2018), using the same filtering as for the mammalian dataset. However, the primate nucleotide divergence was not obtained on a set of neutral markers as for the mammalian dataset, but across the whole genome. As such, the evidence for $\rho > 1$ does not necessarily imply that the trait is evolving under diversifying selection since non-neutral markers included in the estimate of divergence can lead to a spurious $\rho > 1$ (see Table 2).

Results

Neutrality index

For a neutral trait, the genetic architecture, meaning the number of loci encoding the trait and the average effect of a mutation on the trait, is formally related to both within and between-species variation of the trait. We defined the neutrality index as $\rho = \sigma^2_B/\sigma^2_W$, which equals 1 for a neutral trait (see *Materials and methods*), suggesting that traits for which this relationship was not verified were putatively under selection. Under stabilizing selection, the variation between species is depleted because the mean trait is maintained toward similar values between different species, which leads to $\rho < 1$. In contrast, under diversifying selection, the variation between species is inflated because species will have potentially different trait values (Hansen, 1997), which leads to $\rho > 1$. Our neutrality index for a quantitative trait leveraged the data for any number of species, and took advantage of the signal over the whole phylogenetic tree, at the same time taking into account phylogenetic inertia and addressing the non-independence between species (Figure 1). This statistic was obtained as a maximum likelihood estimate from *Equations* 27 and 26. We also devised a Bayesian estimate to obtain the posterior distribution of the neutrality index, and test for diversifying selection as $\mathbb{P}[\rho > 1]$, and stabilizing selection as $\mathbb{P}[\rho < 1]$.

Our neutrality index made a series of assumptions that we described in details in *Material and methods*. Table 2 summarized these assumptions and outlined possible consequences for the neutrality test that we proposed.

Results against simulations

The inference framework was first tested on independently simulated datasets matching an empirically relevant mammalian empirical regime (see *Materials and methods*). Under constant population size (N_e) and constant mutation rates (μ and u) across the phylogenetic tree (Figure 3, top row), we found no false negative for simulations of stabilizing ($\mathbb{P}[\rho < 1] > 0.975$; blue in Figure 3) or diversifying ($\mathbb{P}[\rho > 1] > 0.975$; red in Figure 3) selection. For simulations under neutral evolution, 77% of those were correctly identified ($0.025 \leq \mathbb{P}[\rho > 1] \leq 0.975$; yellow in Figure 3), while 21% and 2% were wrongly detected as stabilizing or diversifying selection, respectively. Once we introduced fluctuating N_e, μ and u (Figure 3, bottom row), our ability to identify simulations under either diversifying or stabilizing selection remained the same with all cases detected correctly. For simulations under neutral evolution, 51% of the simulations were correctly detected ($0.025 \leq \mathbb{P}[\rho > 1] \leq 0.975$), while 49% were detected as stabilizing selection ($\mathbb{P}[\rho < 1] > 0.975$) and none as diversifying selection.

Results on empirical data

For mammalian body and brain mass, we obtained male (♂) and female (♀) trait variations. Combined with nucleotide diversity and divergence, we estimated ρ and posterior probabilities of diversifying selection under different assumptions for trait heritability as shown in Table 1. For body mass, assuming complete heritability led to zero posterior probabilities of diversifying selection for both males and females ($\mathbb{P}[\rho > 1] = 0.0$). If we assumed that heritability (h^2) of body mass was uniformly distributed between 20% and 40% (Hu *et al.*, 2022), posterior probabilities of diversifying selection became 0.635 for males and 0.324 for females. Mammalian brain mass was found to be under diversifying selection with posterior probabilities of 0.877 for males and 0.972 for females when complete heritability was assumed. Assuming a uniform distribution between 20% and 40% for heritability led to posterior probabilities of diversifying selection of 1.0 for both males and females.

We also analyzed a similar dataset for body mass focusing this time only at Primates (Table 1). For primates body mass, assuming complete heritability led to zero posterior probabilities of diversifying selection for both males and females, exactly as in the mammal dataset. However, we found posterior probabilities of diversifying selection of 1.0 for males and 0.914 for females when assuming a uniform distribution for the heritability of body mass between 20% and 40%. For brain mass, assuming complete heritability or not (between 20% and 40%) did not change the posterior probability of diversifying selection, which was 1.0. Evidence for diversifying selection on both brain and body mass was therefore more pronounced in Primates than in mammals. However, the genetic markers used to normalize trait variance with nucleotide divergence were not necessarily neutral, which could create spurious false positives by artificially inflating ρ (Table 2 and *Material and methods*).

Discussion

In this study, we proposed a neutrality index for a quantitative trait that can be used within a statistical framework to test for selection. Our neutrality index for a trait, ρ, is calculated as the ratio of the normalized within- to between-species

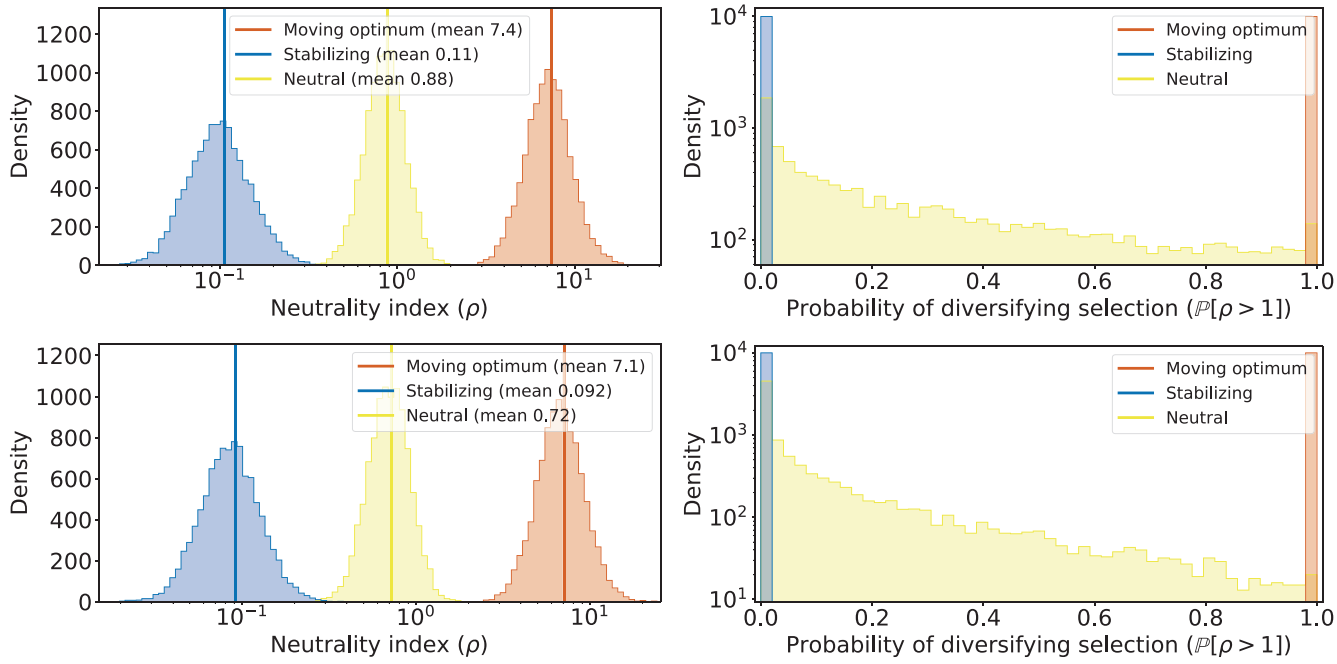


Figure 3. 10,000 simulations of trait evolution along a phylogenetic tree under different selection regimes. Traits simulated under stabilizing selection (blue), under a neutral evolution (yellow), and under a moving optimum (red). Histogram of ratio of between-species trait variation (σ^2_B) over within-species trait variation σ^2_W with $\rho = \sigma^2_B/\sigma^2_W$ estimated from each simulated data (left) and probabilities of ρ being greater than 1 (right). Effective population size (N_e) and mutation rates (μ and u) were either constant (top row), or fluctuating as a Brownian process along the phylogenetic tree (bottom row).

Table 1. Test of diversifying selection on a mammal and a primate dataset, by splitting males (♂) and females (♀). Traits considered were body mass or brain mass (log-transformed). Heritability (h^2) was either assumed complete ($h^2 = 1.0$) or uniformly distributed between 20% and 40% ($h^2 \sim \mathcal{U}(0.2, 0.4)$). n was the number of species in the dataset. ρ was the posterior estimate of our neutrality index, with the 95% credible interval (CI) for ρ also computed. $\mathbb{P}[\rho > 1]$ was the estimated posterior probability of diversifying selection.

Dataset	Trait	h^2	Sex	n	ρ	95% CI for ρ	$\mathbb{P}[\rho > 1]$
Mammals	Body mass	1.0	♂	36	0.340	0.217-0.523	0.000
Mammals	Body mass	1.0	♀	26	0.277	0.160-0.490	0.000
Mammals	Body mass	$\mathcal{U}(0.2, 0.4)$	♂	36	1.124	0.721-1.754	0.635
Mammals	Body mass	$\mathcal{U}(0.2, 0.4)$	♀	26	0.936	0.523-1.715	0.324
Mammals	Brain mass	1.0	♂	36	1.351	0.851-2.173	0.877
Mammals	Brain mass	1.0	♀	26	1.727	0.991-2.938	0.972
Mammals	Brain mass	$\mathcal{U}(0.2, 0.4)$	♂	36	4.527	2.831-7.091	1.000
Mammals	Brain mass	$\mathcal{U}(0.2, 0.4)$	♀	26	6.001	3.288-10.941	1.000
Primates	Body mass	1.0	♂	71	0.558	0.401-0.784	0.000
Primates	Body mass	1.0	♀	65	0.389	0.278-0.547	0.000
Primates	Body mass	$\mathcal{U}(0.2, 0.4)$	♂	71	1.875	1.288-2.695	1.000
Primates	Body mass	$\mathcal{U}(0.2, 0.4)$	♀	65	1.296	0.899-1.821	0.914
Primates	Brain mass	1.0	♂	71	1.929	1.395-2.616	1.000
Primates	Brain mass	1.0	♀	65	1.950	1.399-2.790	1.000
Primates	Brain mass	$\mathcal{U}(0.2, 0.4)$	♂	71	6.479	4.658-8.944	1.000
Primates	Brain mass	$\mathcal{U}(0.2, 0.4)$	♀	65	6.522	4.664-9.294	1.000

variation and it allowed the identification of the evolutionary regime of a quantitative trait. At the phylogenetic scale, trait variation between species was normalized by sequence divergence obtained from a neutral set of markers. Similarly, trait variation within species was normalized by sequence polymorphism obtained also from a neutral set of markers. Our estimate of ρ could be tested for deviation from the value of 1.0 expected under the null hypothesis of neutrality. Technically, the neutrality index can be estimated either as a maximum likelihood point estimate, or as a mean posterior estimate from a Bayesian implementation (see online supplementary material Section S3). The latter also enabled the estimation of the posterior credible interval to test for departure from a neutrally evolving trait (e.g., $\mathbb{P}[\rho > 1]$). We tested our statistical

Table 2. Assumptions breaks and their consequences on the estimation of within-species variation (σ^2_W), between-species variation (σ^2_B), and on the neutrality index $\rho = \sigma^2_B/\sigma^2_W$. The last two columns indicate whether the test for diversifying selection ($\rho > 1$) and for stabilizing selection $\rho < 1$ are conservative or invalid due to violated assumptions.

Broken assumption	Consequences	σ^2_W	σ^2_B	Test $\rho > 1$	Test $\rho < 1$
Trait encoded by few loci	Between-species trait variation is underestimated	–	Underestimated	Conservative	Invalid
Sexual dimorphism	Within-species trait variation is overestimated	Overestimated	–	Conservative	Invalid
Phenotypic plasticity	Trait responding to individual environments	Overestimated	–	Conservative	Invalid
Inbreeding	Nucleotide diversity (π) is underestimated	Overestimated	–	Conservative	Invalid
Markers for polymorphism are negatively selected	Nucleotide diversity (π) is underestimated	Overestimated	–	Conservative	Invalid
Markers for polymorphism are positively selected	Nucleotide diversity (π) is underestimated	Overestimated	–	Conservative	Invalid
Markers for divergence are positively selected	Nucleotide divergence (d) is overestimated	–	Underestimated	Conservative	Invalid
Markers for polymorphism under balanced selection	Nucleotide diversity (π) is overestimated	Underestimated	–	Invalid	Conservative
Markers for divergence are negatively selected	Nucleotide divergence (d) is underestimated	–	Overestimated	Invalid	Conservative
Multiple nucleotide substitutions at the same locus	Nucleotide divergence (d) is underestimated	–	Overestimated	Invalid	Conservative

procedure against simulated data and showed that our test was able to correctly detect simulations under diversifying selection (test of $\rho > 1$) or under stabilizing selection (test of $\rho < 1$). However, our test detected a spurious signal of stabilizing selection ($\rho < 1$) when we simulated the evolution of a neutral trait. An assumption of our test is that the neutral phenotypic trait is evolving as a Brownian process and is, therefore, unbounded. However, the phenotype may be bounded by what the genetic architecture can produce, and this could cause a slowdown of phenotypic divergence over time due to the erosion of possible phenotypic changes at the underlying loci. Typically, such an effect depends on the number of alleles per locus, whether new mutations are generating new alleles or instead reverting to previous alleles. Altogether, in our simulation setting under a constant genetic architecture with a fixed number of loci, such a slowdown of phenotypic divergence can result in a spurious signal of stabilizing selection ($\rho < 1$), especially for deeper phylogeny (see online supplementary material Figure S2 and Section S4). We thus argue that our method should be used to detect diversifying selection, but that it had low accuracy to detect stabilizing selection due to false positives.

Our results showed that our method significantly improved over currently available methods to detect selection acting on a trait at the phylogenetic scale. Current methods relying on evolution of the mean trait value between species also tend to statistically prefer a model of stabilizing selection over a Brownian process when the trait is neutral (Cooper *et al.*, 2016; Price *et al.*, 2022; Silvestro *et al.*, 2015). Our approach could in theory be applied to detect stabilizing selection at the phylogenetic scale, but we showed that it did not have the statistical power to identify those cases. In contrast, we showed that our method was able to identify correctly cases of diversifying selection, which is a clear improvement over current methods that model only mean trait value. Indeed, under diversifying selection, mean trait value will not deviate from a Brownian process, and thus cannot be distinguished from neutral evolution (Hansen & Martins, 1996; Harmon, 2018). For example, testing the selective regime in the expression level of the majority of genes led to the selection of a Brownian process as the prefered model and the interpretation that the expression was evolving neutrally (Catalán *et al.*, 2019). Instead, our diversity index has the advantage to discriminate the alternative model of diversifying selection from the neutral case by comparing within- and between-species variation while correctly normalizing them using nucleotide markers. Our approach is not the first one coupling between-species and within-species variations, and those approaches employ different strategies to detect selection. First, one empirical strategy is to compare the ratio of between to within variation across a pool of traits, which allow to identify outlier traits putatively under diversifying selection (Rohlfs *et al.*, 2014). However, this method does not formally allow testing for diversifying selection, and requires many traits such as expression level data to seek outliers genes (Gillard *et al.*, 2021; Rohlfs & Nielsen, 2015). Second, other methods leverage Lande's generalized genetic distance (LGGD), which relate the ratio of between to within variations to population-genetic parameters (Lande, 1979; Lemos *et al.*, 2001, 2005; Lynch & Crease, 1990; Porto *et al.*, 2015; Weaver *et al.*, 2007). Specifically, by leveraging estimates of effective population size (N_e) and number of generations between species, or alternatively by assuming their constancy, these methods can test for departures from the null model of neutral evolution for a single trait. Such methods have been successful in identifying specific instances of diversifying

selection (Machado *et al.*, 2022; Schroeder & von Cramon-Taubadel, 2017) and near-drift (Machado *et al.*, 2023). However, N_e and the number of generations are complex parameters to correctly infer, and is usually done for a pair or only a few species, and ultimately requires large genomic datasets and heavy statistical methods (Wilder *et al.*, 2023). Instead, our diversity index opens new avenues to revisit these studies testing for the selective regime affecting the quantitative traits, by formally incorporating nucleotide divergence and polymorphism, bypassing estimation of N_e, generation time and calibration of ancestral node ages (Machado *et al.*, 2023).

As such, the main novelty of our study was to use the nucleotide divergence and polymorphism to normalize trait variation between and within species. In this context, our test bears many similarities to Q_{ST}–F_{ST} tests (and their derivatives) that have been developed to test for selection of a trait across several populations while also leveraging sequence variation (Leinonen *et al.*, 2013; Martin *et al.*, 2008) or co-ancestry between individuals (Ovaskainen *et al.*, 2011). Our method can be seen as an analog at the phylogenetic scale, where although the sequences used should be neutrally evolving, they can be obtained from different sampled individuals than for the trait. Importantly, by normalizing with sequence variation, we also showed using simulated data that our test was not sensitive to the assumption that N_e and mutation rates were constant across the phylogenetic tree, an unmet assumption empirically (Bergeron *et al.*, 2023; Wilder *et al.*, 2023). Indeed, under the neutral case of evolution, the normalization by nucleotide divergence and polymorphism automatically absorbed long-term and short-term changes in N_e, generation time and mutation rates, which canceled out in the neutrality index ρ.

In the context of phylogenetic comparative methods, modeling mean trait evolution as a function of nucleotide divergence (d) instead of time has more general consequences. As an example, trait variation is often modeled as a Brownian process running on a time-calibrated tree, which can produce biases (Litsios & Salamin, 2012). Indeed, for a neutrally evolving trait, trait variation depends directly on the number of generations, which in turn correlates with time. But, since species generation time might vary along the phylogenetic tree, d-scaled trees absorbing changes in generation time should be used instead of time-scaled trees. Using nucleotide divergence would also remove the potential effect of model assumptions required to calibrate ancestral node ages (e.g., molecular clocks). We argue, that the soundness of studying trait evolution on d-scaled trees can be evaluated by the absolute fit of a model to the data (Pennell *et al.*, 2015). More generally, genomic information could potentially be seen as a way to disentangle congruence models (Louca & Pennell, 2020), or as prior for methods that detect shifts in adaptive regimes (Ingram & Mahler, 2013; Khabbazian *et al.*, 2016; Mitov *et al.*, 2020; Uyeda & Harmon, 2014).

Even though our test was developed for a quantitative trait, analogies with other tests of selection developed for molecular sequences also provided insight into its behavior. First, we acknowledge that our test took inspiration from the McDonald and Kreitman (1991) test devised for protein-coding DNA sequences in a pair of species, except that the non-synonymous versus synonymous distinction is replaced by the comparison between quantitative trait and neutral genomic sequence. Second, at the phylogenetic scale, when comparison is done across several species, our test also bears analogy to codon-based test of selection, where the ratio of non-synonymous to synonymous substitutions (ω) is compared to 1 (Goldman & Yang, 1994; Muse & Gaut, 1994). As $\omega < 1$ is interpreted as purifying selection acting on the protein, $\rho < 1$ is interpreted as stabilizing selection acting on the trait. Similarly, the interpretation of adaptation for $\omega > 1$ is analogous to diversifying selection for $\rho > 1$. With this analogy in mind, we could leverage the vast literature discussing and interpreting the results of these tests and their pitfalls (Anisimova & Kosiol, 2009; Jensen *et al.*, 2019; Nielsen, 2005). First, not rejecting the neutral null model of $\rho = 1$ did not necessarily imply that the trait was effectively neutral, since diversifying and stabilizing selection could compensate each other resulting in $\rho = 1$, analogously to $\omega = 1$ under a mix of adaptation and purifying selection (Nielsen, 2005). Second, empirical evidence for $\rho < 1$ did not rule out diversifying selection, but rather that this diversifying selection was not strong enough to overcome the stabilizing selection, similarly to strong purifying selection resulting $\omega < 1$ even though those genes and sites are under adaptation (Latrille *et al.*, 2023). By explicitly modeling stabilizing selection as a moving optimum, it would theoretically be possible to tease apart the effect of diversifying and stabilizing selection in the context of quantitative traits to obtain a statistically more powerful test.

In the context of detecting diversifying selection on a trait, we argue that the main drawback of our method is that the additive genetic variance of the trait is required instead of the phenotypic variance. If phenotypic variance was used instead of additive genetic variance to estimate ρ, meaning that we assumed complete heritability, the neutrality index ρ was ultimately underestimated. Similarly, using broad-sense heritability instead of narrow-sense heritability would result in underestimated ρ. In such context, the test of stabilizing selection ($\rho < 1$]) would be statistically invalid. However, the test of diversifying selection ($\rho > 1$) was underpowered although not invalided, meaning that absence of evidence would not be evidence of absence. As an example, even though we assumed complete heritability for brain mass, we uncovered diversifying selection in mammals since $\rho > 1$. If available, any prior knowledge on heritability can be leveraged instead of assuming complete heritability to increase the statistical power to detect diversifying selection (Hansen & Pélabon, 2021; Hansen *et al.*, 2011). Additionally, phenotypic plasticity also affects the genotype-phenotype relationship with intricate consequences for our test of selection. First, at the level of within species variation, individuals might occupy different patches with different environments. Responding to these individual environmental conditions, phenotypic plasticity would then result in increased trait variation within species. In this scenario, as hypothesized in Rohlfs & Nielsen (2015), phenotypic plasticity then leads to a reduced ratio of between to within species variations, thus ultimately leading to our tests of diversifying selection being underpowered although not invalid. Alternatively, it is also possible that different species are experiencing different macro-environments, for example with species spread along a latitudinal or elevation gradient, with different temperatures or precipitation. These species could thus have different mean phenotypes solely because of phenotypic plasticity, while such changes are not encoded in their genome (Schraiber & Edge, 2023; Stamp & Hadfield, 2020). Such an effect can lead to $\rho > 1$ erroneously interpreting diversifying selection. The test of $\rho > 1$

would however be correct that the changes in mean phenotypes across species is due to change in environment, albeit in such a case not encoded by the genotype of individuals but due to phenotypic plasticity.

The development of our neutrality index was also based on several assumptions that could be relaxed in future studies. First, we cannot predict the behavior of our test in the context of population structures, gene flow and introgression. These factors should be thoroughly investigated using simulations. Second, loci were assumed to contribute additively to the phenotype. Although the effects of dominance and epistasis is typically weak compared to the additive effects on the quantitative traits, their influence should be assessed (Crow, 2010; Hill *et al.*, 2008). Third, the genetic architecture of the trait was assumed to be constant across the phylogenetic tree, whereas it might actually be variable among individuals and species (Huber *et al.*, 2015; Tung *et al.*, 2015). Such an assumption can theoretically be relaxed and changes in genetic architecture along the phylogenetic tree could jointly be estimated (Arnold *et al.*, 2008; Gaboriau *et al.*, 2020; Hohenlohe & Arnold, 2008; Kostikova *et al.*, 2016). Finally, from a statistical perspective, our Bayesian estimation could integrate uncertainty from the estimation of genetic variation, using sequences as input instead of estimated values of nucleotide diversity and divergence.

From an empirical point of view, our method required integrating genomic and trait variation, which could reduce the possible datasets to be used. However, such datasets will become more and more accessible and we showed the applicability of our method by applying it to the illustrative example of mammals' brain and body mass, both showing signals of diversifying selection. As such, this result corroborates studies relying solely on changes in mean trait values across mammals, showing strong statistical support for several distinct evolutionary regimes for body- and brain mass (Mitov *et al.*, 2019). Interestingly, our strongest signal is for brain mass, corroborating studies in hominids where skull size (related to brain mass) is the only trait that exceeded the expected rate of phenotypic evolution under a neutral model (Lynch, 1990). Hence, one first interpretation here is that brain mass might be an exceptional case among many phenotypic traits (e.g., dental and skeletal measures). Second, from a macro-evolutionary perspective, the consensus is that empirical rates of evolution calculated on phylogenetic trees and the fossil record are far inferior to the expected under drift (Lynch & Crease, 1990; Uyeda *et al.*, 2011), where such methods assume constancy of N_e, generation time and mutation rates. Our finding of diversifying selection on body and brain mass could be seen as an argument against that interpretation. In fact, rates of nucleotide evolution also show a tendency for slowing down on a longer timescale (Rolland *et al.*, 2023). One possible interpretation is that normalization by nucleotide divergence could absorb this observed slowing rate of evolution. Altogether, further empirical and theoretical studies are required to disentangle this discrepancy between these different results and interpretations Because our test was also based on several assumptions that might not hold on empirical data, we also provided a table containing the main assumptions and their consequences on the neutrality index and the test that can be performed (Table 2). For example, at the primate scale, the evidence for $\rho > 1$ does not necessarily imply that the brain mass was evolving under diversifying selection since the markers used for nucleotide divergences were not neutral, which can lead to a spurious $\rho > 1$. In conclusion, our study provided a statistical framework to test for diversifying selection acting on a quantitative trait while integrating the trove of genomic data available both within and between species, and we believe that our new approach is a promising tool to investigate the evolution of quantitative traits.

Funding

The work has been funded by Université de Lausanne and the Swiss National Science Foundation (315230_219757).

Acknowledgments

We gratefully acknowledge the help of Nicolas Lartillot, Philippe Veber, Isabela Jeronimo do Ó, Anna Marcionetti, Julien Clavel, and Daniele Silvestro for their insightful discussions and Julien Joseph for his advice and reviews concerning this manuscript.

Conflicts of interest

None declared.

Data availability

The data and code that support the findings of this study are openly available in Zenodo at https://doi.org/10.5281/zenodo.12666506. Snakemake pipeline, analysis scripts and documentation are available in the repository to replicate the study.

References

Adams, D. C. & Collyer, M. L. 2018. Multivariate phylogenetic comparative methods: Evaluations, comparisons, and recommendations. *Systematic Biology, 67*(1): 14–31.

Anisimova, M. & Kosiol, C. 2009. Investigating protein-coding sequence evolution with probabilistic codon substitution models. *Molecular Biology and Evolution, 26*(2): 255–271.

Arnold, S. J., Bürger, R., Hohenlohe, P. A., . . ., Jones, A. G. 2008. Understanding the evolution and stability of the G-matrix. *Evolution, 62*(10): 2451–2461.

Barton, N. H., Etheridge, A. M., & Véber, A. 2017. The infinitesimal model: Definition, derivation, and implications. *Theoretical Population Biology, 118*: 50–73.

Bergeron, L. A., Besenbacher, S., Zheng, J., . . ., Zhang, G. 2023. Evolution of the germline mutation rate across vertebrates. *Nature, 615*(7951): 285–291.

Blanquart, F. & Bataillon, T. 2016. Epistasis and the structure of fitness landscapes: Are experimental fitness landscapes compatible with fisher's geometric model? *Genetics, 203*(2): 847–862.

Catalán, A., Briscoe, A. D., & Höhna, S. 2019. Drift and directional selection are the evolutionary forces driving gene expression divergence in eye and brain tissue of Heliconius butterflies. *Genetics, 213*(2): 581–594.

Cooper, N., Thomas, G. H., Venditti, C., . . ., Freckleton, R. P. 2016. A cautionary note on the use of Ornstein Uhlenbeck models in macroevolutionary studies. *Biological Journal of the Linnean Society, 118*(1): 64–77.

Crow, J. F. 2010. On epistasis: Why it is unimportant in polygenic directional selection. *Philosophical Transactions of the Royal Society B: Biological Sciences, 365*(1544): 1241–1244.

Edelaar, P., Burraco, P., & Gomez-Mestre, I. 2011. Comparisons between QST and FST—how wrong have we been? *Molecular Ecology, 20*(23): 4830–4839.

Felsenstein, J. 1985. Phylogenies and the Comparative Method. *The American Naturalist, 125*(1): 1–15.
Felsenstein, J. 1988. Phylogenies and quantitative characters. *Annual Review of Ecology and Systematics, 19*(1): 445–471.
Felsenstein, J. 2008. Comparative methods with sampling error and within-species variation: Contrasts revisited and revised. *The American Naturalist, 171*(6): 713–725.
Foley, N. M., Mason, V. C., Harris, A. J., . . ., Murphy, W. J. 2023. A genomic timescale for placental mammal evolution. *Science, 380*(6643): eabl8189.
Fraser, H. B. 2020. Detecting selection with a genetic cross. *Proceedings of the National Academy of Sciences United States of America, 117*(36): 22323–22330.
Gaboriau, T., Mendes, F. K., Joly, S., . . ., Salamin, N. 2020. A multiplatform package for the analysis of intra- and interspecific trait evolution. *Methods in Ecology and Evolution*, 11(11): 1439–1447.
Gaboriau, T., Tobias, J. A., Silvestro, D., & Salamin, N. 2023. Exploring the Macroevolutionary Signature of Asymmetric Inheritance at Speciation.
Genereux, D. P., Serres, A., Armstrong, J., . . ., Zoonomia Consortium 2020. A comparative genomics multitool for scientific discovery and conservation. *Nature, 587*(7833): 240–245.
Gillard, G. B., Grønvold, L., Røsæg, L. L., . . ., Hvidsten, T. R. 2021. Comparative regulomics supports pervasive selection on gene dosage following whole genome duplication. *Genome Biology, 22*(1): 103.
Goldman, N. & Yang, Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution, 11*(5): 725–736.
Grabowski, M., Pienaar, J., Voje, K. L., . . ., Hansen, T. F. 2023. A Cautionary Note on "A Cautionary Note on the Use of Ornstein Uhlenbeck Models in Macroevolutionary Studies". *Systematic Biology, 72*(4): 955–963.
Hansen, T. F. 1997. Stabilizing selection and the comparative analysis of adaptation. *Evolution, 51*(5): 1341–1351.
Hansen, T. F. & Bartoszek, K. 2012. Interpreting the evolutionary regression: The interplay between observational and biological errors in phylogenetic comparative studies. *Systematic Biology, 61*(3): 413–425.
Hansen, T. F. & Martins, E. P. 1996. Translating between microevolutionary process and macroevolutionary patterns: The correlation structure of interspecific data. *Evolution, 50*(4): 1404–1417.
Hansen, T. F. & Pélabon, C. 2021. Evolvability: A Quantitative-Genetics Perspective. *Annual Review of Ecology, Evolution, and Systematics, 52*(1): 153–175.
Hansen, T. F., Pélabon, C., & Houle, D. 2011. Heritability is not Evolvability. *Evolutionary Biology, 38*(3): 258–277.
Harmon, L. 2018. Phylogenetic comparative methods: Learning from trees.
Hill, W. G., Goddard, M. E., & Visscher, P. M. 2008. Data and Theory Point to Mainly Additive Genetic Variance for Complex Traits. *PLOS Genetics, 4*(2): e1000008.
Hohenlohe, P. A. & Arnold, S. J. 2008. MIPoD: A hypothesis-testing framework for microevolutionary inference from patterns of divergence. *The American Naturalist, 171*(3): 366–385.
Hu, Z.-L., Park, C. A., & Reecy, J. M. 2022. Bringing the Animal QTLdb and CorrDB into the future: Meeting new challenges and providing updated services. *Nucleic Acids Research, 50*(D1): D956–D961.
Huber, B., Whibley, A., Poul, Y. L., . . ., Joron, M. 2015. Conservatism and novelty in the genetic architecture of adaptation in Heliconius butterflies. *Heredity*, 114(5): 515–524.
Huelsenbeck, J. P. & Rannala, B. 2003. Detecting correlation between characters in a comparative analysis with uncertain phylogeny. *Evolution, 57*(6): 1237–1247.
Ingram, T. & Mahler, D. 2013. SURFACE: Detecting convergent evolution from comparative data by fitting Ornstein-Uhlenbeck models with stepwise Akaike Information Criterion. *Methods in Ecology and Evolution, 4*(5): 416–425.
Jensen, J. D., Payseur, B. A., Stephan, W., . . ., Charlesworth, B. 2019. The importance of the Neutral Theory in 1968 and 50 years on: A response to Kern and Hahn 2018. *Evolution, 73*(1): 111–114.
Khabbazian, M., Kriebel, R., Rohe, K., & Ané, C. 2016. Fast and accurate detection of evolutionary shifts in Ornstein–Uhlenbeck models. *Methods in Ecology and Evolution, 7*(7): 811–824.
Khaitovich, P., Weiss, G., Lachmann, M., . . ., Pääbo, S. 2004. A neutral model of transcriptome evolution. *PLOS Biology, 2*(5): e132.
Kimura, M. 1962. On the probability of fixation of mutant genes in a population. *Genetics, 47*(6): 713–719.
Kimura, M. 1968. Evolutionary rate at the molecular level. *Nature, 217*(5129): 624–626.
Kostikova, A., Silvestro, D., Pearman, P. B., & Salamin, N. 2016. Bridging inter- and intraspecific trait evolution with a hierarchical bayesian approach. *Systematic Biology, 65*(3): 417–431.
Kuderna, L. F. K., Gao, H., Janiak, M. C., . . ., Marques Bonet, T. 2023. A global catalog of whole-genome diversity from 233 primate species. *Science*, 380(6648): 906–913.
Lamy, J.-B., Plomion, C., Kremer, A., & Delzon, S. 2012. QST < FST As a signature of canalization. *Molecular Ecology*, 21(23): 5646–5655.
Lande, R. 1979. Quantitative genetic analysis of multivariate evolution, applied to brain: Body size allometry. *Evolution, 33*(1): 402–416.
Lande, R. 1980a. Genetic variation and phenotypic evolution during allopatric speciation. *The American Naturalist, 116*(4): 463–479.
Lande, R. 1980b. Sexual dimorphism, sexual selection, and adaptation in polygenic characters. *Evolution, 34*(2): 292–305.
Lartillot, N. & Delsuc, F. 2012. Joint reconstruction of divergence times and life-history evolution in placental mammals using a phylogenetic covariance model. *Evolution, 66*(6): 1773–1787.
Lartillot, N. & Poujol, R. 2011. A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters. *Molecular Biology and Evolution, 28*(1): 729–744.
Latrille, T., Lanore, V., & Lartillot, N. 2021. Inferring long-term effective population size with mutation–selection models. *Molecular Biology and Evolution, 38*(10): 4573–4587.
Latrille, T., Rodrigue, N., & Lartillot, N. 2023. Genes and sites under adaptation at the phylogenetic scale also exhibit adaptation at the population-genetic scale. *Proceedings of the National Academy of Sciences of the United States of America, 120*(11): e2214977120.
Leinonen, T., O'hara, R. B., Cano, J. M., & Merilä, J. 2008. Comparative studies of quantitative trait and neutral marker divergence: A meta-analysis. *Journal of Evolutionary Biology, 21*(1): 1–17.
Leinonen, T., McCairns, R. J. S., O'Hara, R. B., & Merilä, J. 2013. QST–FST comparisons: Evolutionary and ecological insights from genomic heterogeneity. *Nature Reviews Genetics, 14*(3): 179–190.
Lemos, B., Marroig, G., & Cerqueira, R. 2001. Evolutionary rates and stabilizing selection in large-bodied opossum skulls (Didelphimorphia: Didelphidae). *Journal of Zoology, 255*(2): 181–189.
Lemos, B., Meiklejohn, C. D., Cáceres, M., & Hartl, D. L. 2005. Rates of Divergence in Gene Expression Profiles of Primates, Mice, and Flies: Stabilizing Selection and Variability Among Functional Categories. *Evolution, 59*(1): 126–137.
Litsios, G. & Salamin, N. 2012. Effects of Phylogenetic Signal on Ancestral State Reconstruction. *Systematic Biology, 61*(3): 533–538.
Louca, S. & Pennell, M. W. 2020. Extant timetrees are consistent with a myriad of diversification histories. *Nature, 580*(7804): 502–505.
Lynch, M. 1990. The Rate of Morphological Evolution in Mammals from the Standpoint of the Neutral Expectation. *The American Naturalist, 136*(6): 727–741.
Lynch, M. 1991. Methods for the analysis of comparative data in evolutionary biology. *Evolution, 45*(5): 1065–1080.
Lynch, M. & Crease, T. J. 1990. The analysis of population survey data on DNA sequence variation. *Molecular Biology and Evolution*, 7(4): 377–394.

Lynch, M. & Walsh, B. 1998. *Genetics and Analysis of Quantitative Traits*, volume 1. Sinauer Sunderland, MA.

Lynch, M., Latta, L., Hicks, J., and Giorgianni, M. 1998. Mutation, selection, and the maintenance of life-history variation in a natural population. *Evolution*, 52(3): 727–733.

Machado, F. A., Marroig, G., & Hubbe, A. 2022. The pre-eminent role of directional selection in generating extreme morphological change in glyptodonts (Cingulata; Xenarthra). *Proceedings of the Royal Society B: Biological Sciences*, 289(1967): 20212521.

Machado, F. A., Mongle, C. S., Slater, G., . . ., Uyeda, J. C. 2023. Using developmental rules to align microevolution with macroevolution.

Martin, G., Chapuis, E., & Goudet, J. 2008. Multivariate Qst–Fst comparisons: A neutrality test for the evolution of the G matrix in structured populations. *Genetics*, 180(4): 2135–2149.

McCandlish, D. M. & Stoltzfus, A. 2014. Modeling evolution using the probability of fixation: History and implications. *Quarterly Review of Biology, 89*(3): 225–252.

McDonald, J. H. & Kreitman, M. 1991. Adaptative protein evolution at Adh locus in Drosophila. *Nature, 351*(6328): 652–654.

Merilä, J. & Crnokrak, P. 2001. Comparison of genetic differentiation at marker loci and quantitative traits. *Journal of Evolutionary Biology, 14*(6): 892–903.

Mitov, V., Bartoszek, K., & Stadler, T. 2019. Automatic generation of evolutionary hypotheses using mixed Gaussian phylogenetic models. *Proceedings of the National Academy of Sciences, 116*(34): 16921–16926.

Mitov, V., Bartoszek, K., Asimomitis, G., & Stadler, T. 2020. Fast likelihood calculation for multivariate Gaussian phylogenetic models with shifts. *Theoretical Population Biology, 131*: 66–78.

Muse, S. V. & Gaut, B. S. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular Biology and Evolution, 1*(5): 715–724.

Nielsen, R. 2005. Molecular signatures of natural selection. *Annual Review of Genetics, 39*(1): 197–218.

O'Meara, B. C., Ané, C., Sanderson, M. J., & Wainwright, P. C. 2006. Testing for different rates of continuous trait evolution using likelihood. *Evolution, 60*(5): 922–933.

Ovaskainen, O., Karhunen, M., Zheng, C., . . ., Merilä, J. 2011. New method to uncover signatures of divergent and stabilizing selection in quantitative traits. *Genetics, 189*(2): 621–632.

Pennell, M. W., FitzJohn, R. G., Cornwell, W. K., & Harmon, L. J. 2015. Model Adequacy and the macroevolution of angiosperm functional traits. *The American Naturalist, 186*(2): E33–E50.

Porto, A., Sebastião, H., Pavan, S. E., . . ., Cheverud, J. M. 2015. Rate of evolutionary change in cranial morphology of the marsupial genus Monodelphis is constrained by the availability of additive genetic variation. *Journal of Evolutionary Biology, 28*(4): 973–985.

Price, P. D., Palmer Droguett, D. H., Taylor, J. A., …, Wright, A. E. 2022. Detecting signatures of selection on gene expression. *Nature Ecology & Evolution, 6*(7): 1035–1045.

Pujol, B., Wilson, A. J., Ross, R. I. C., & Pannell, J. R. 2008. Are QST–FST comparisons for natural populations meaningful? *Molecular Ecology, 17*(22): 4782–4785.

Rohlfs, R. V. & Nielsen, R. 2015. Phylogenetic ANOVA: The expression variance and evolution model for quantitative trait evolution. *Systematic Biology, 64*(5): 695–708.

Rohlfs, R. V., Harrigan, P., & Nielsen, R. 2014. Modeling gene expression evolution with an extended Ornstein–Uhlenbeck process accounting for within-species variation. *Molecular Biology and Evolution, 31*(1): 201–211.

Rolland, J., Henao-Diaz, L. F., Doebeli, M., …, Schluter, D. 2023. Conceptual and empirical bridges between micro- and macroevolution. *Nature Ecology & Evolution, 7*(8): 1181–1193.

Schraiber, J. G. & Edge, M. D. 2023. Heritability within groups is uninformative about differences among groups: Cases from behavioral, evolutionary, and statistical genetics. *Proceedings of the National Academy of Sciences, 121*(12): e2319496121.

Schroeder, L. & von Cramon-Taubadel, N. 2017. The evolution of hominoid cranial diversity: A quantitative genetic approach. *Evolution, 71*(11): 2634–2649.

Sella, G. & Barton, N. H. 2019. Thinking about the evolution of complex traits in the era of genome-wide association studies. *Annual Review of Genomics and Human Genetics, 20*(1): 461–493.

Silvestro, D., Kostikova, A., Litsios, G., …, Salamin, N. 2015. Measurement errors should always be incorporated in phylogenetic comparative analysis. *Methods in Ecology and Evolution, 6*(3): 340–346.

Silvestro, D., Tejedor, M. F., Serrano-Serrano, M. L., …, Salamin, N. 2019. Early arrival and climatically-linked geographic expansion of new world monkeys from tiny African ancestors. *Systematic Biology, 68*(1): 78–92.

Simons, Y. B., Bullaughey, K., Hudson, R. R., & Sella, G. 2018. A population genetic interpretation of GWAS findings for human quantitative traits. *PLOS Biology, 16*(3): e2002985.

Stamp, M. A. & Hadfield, J. D. 2020. The relative importance of plasticity versus genetic differentiation in explaining between population differences; a meta-analysis. *Ecology Letters, 23*(10): 1432–1441.

Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics, 123*(3): 585–595.

Tenaillon, O. 2014. The utility of fisher's geometric model in evolutionary genetics. *Annual Review of Ecology, Evolution, and Systematics, 45*(1): 179–201.

Tsuboi, M., van der Bijl, W., Kopperud, B. T., …, Kolm, N. 2018. Breakdown of brain–body allometry and the encephalization of birds and mammals. *Nature Ecology & Evolution, 2*(9): 1492–1500.

Tung, J., Zhou, X., Alberts, …, Gilad, Y. 2015. The genetic architecture of gene expression levels in wild baboons. *eLife, 4*: e04729.

Turelli, M. 1984. Heritable genetic variation via mutation-selection balance: Lerch's zeta meets the abdominal bristle. *Theoretical Population Biology, 25*(2): 138–193.

Turelli, M. 2017. Commentary: Fisher's infinitesimal model: A story for the ages. *Theoretical Population Biology, 118*: 46–49.

Uyeda, J. C. & Harmon, L. J. 2014. A novel Bayesian method for inferring and interpreting the dynamics of adaptive landscapes from phylogenetic comparative data. *Systematic Biology, 63*(6): 902–918.

Uyeda, J. C., Hansen, T. F., Arnold, S. J., & Pienaar, J. 2011. The million-year wait for macroevolutionary bursts. *Proceedings of the National Academy of Sciences, 108*(38): 15908–15913.

Walsh, B. & Lynch, M. 2018. *Evolution and selection of quantitative traits*. Oxford University Press.

Weaver, T. D., Roseman, C. C., & Stringer, C. B. 2007. Were neandertal and modern human cranial differences produced by natural selection or genetic drift? *Journal of Human Evolution, 53*(2): 135–145.

Wilder, A. P., Supple, M. A., Subramanian, A.,…Shapiro, B. 2023. The contribution of historical processes to contemporary extinction risk in placental mammals. *Science, 380*(6643): eabn5856.

9.4 Annexe autre

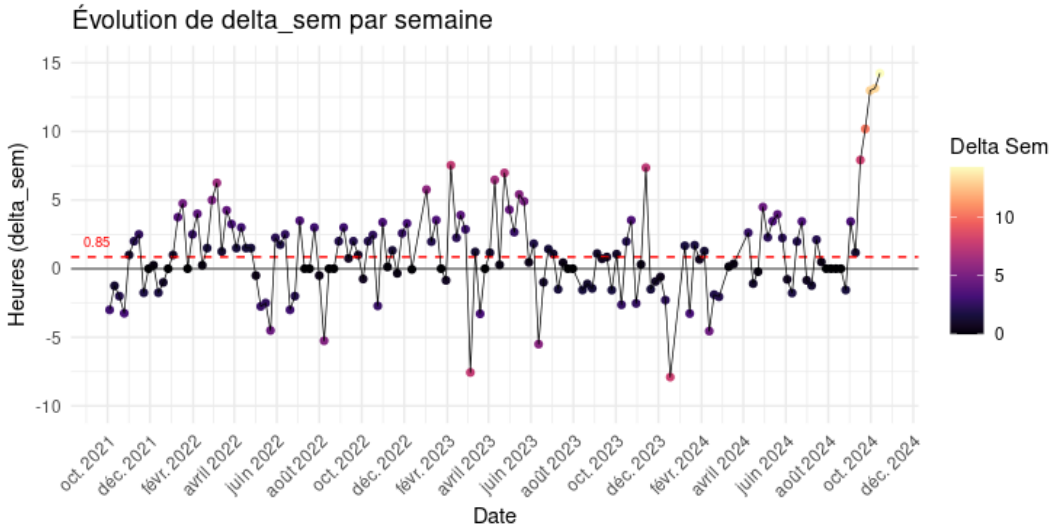


Figure 9.1 – *Représentation du temps de travail effectué pendant les 3 ans de thèse par rapport à une journée type de 7h (nommé delta_sem). Une valeur positive (ou négative) indique un temps de travail hebdomadaire supérieur (ou inférieur) à une moyenne de 7h par jours (soit 35h pour les semaines à 5 jours, 28h pour les semaines à 4 jours, etc.). La ligne rouge représente la mediane des delta_sem sur l'ensemble du temps de thèse. À noter que 33 jours de vacances sur les 135 prévus en 3 ans n'ont pas été pris.*

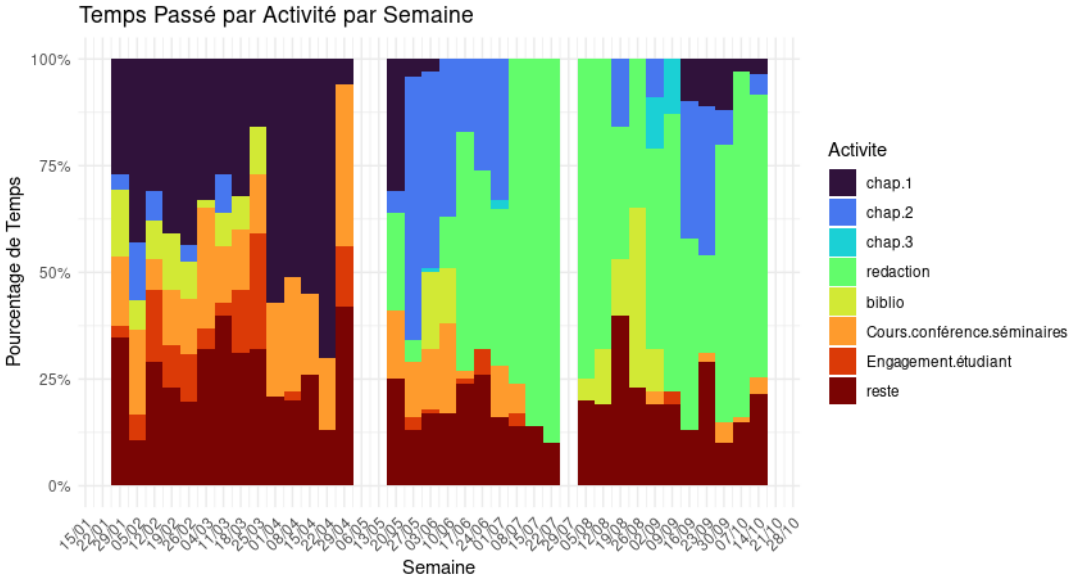


Figure 9.2 – *Répartition du temps de travail par catégorie de tâches, du 22/01/24 au 21/10/24. « Chap.1 » et « Chap.2 » correspondent respectivement aux manuscrits des chapitre 5 et chapitre 6, et « Chap.3 » au travail sur le modèle FastCoevolNe présenté en section 7.1. « Rédaction » inclut la rédaction de ce manuscrit ainsi que les recherches bibliographiques associées. « Biblio » se réfère aux recherches bibliographiques hors rédaction. « Cours/conférences/séminaires » englobe l'enseignement, les séminaires de laboratoire et externes auxquels j'ai assisté. « Engagement étudiant » désigne le temps consacré à l'association des Pinsons migRateurs et à la représentation des étudiant.e.s*

Enseignements :

Contrat ACE de 64h par ans sur 3 ans :

— TP de bio-informatique aux M2 de l'ue OSBAD et L3 de l'ue ASBIV : parcours de base de données, identifier une séquence avec Augustus et Blast, reconstruire et interpréter une phylogénie, utiliser Galaxy

— CM sur la reconstruction phylogénétique aux M2 de l'ue OSBAD

— TP Évolution aux L3 de l'ue Évolution : reconnaitre l'action de différents mécanismes évolutifs sur des séquences moléculaires via l'analyse du taux de GC et des longueurs de branches d'une phylogénie. A l'échelle intra-gène, entre gène et entre espèces.

— TP informatique aux L1 : recherche bibliographique et vérification des sources d'information

Vulgarisation :

— *Déclic* 2021 et 2022 : rencontre en speed dating avec des lycéen.ne.s

— Intervenante scientifique pour le projet *Mon collège sur Mars* en 2022. Suivis d'une classe de CM1-CM2 sur 5-6 demi-journées. Apprentissage du raisonnement scientifique, de la recherche d'informations et de la présentation des résultats dans un séminaire pour enfants.

— *Fête de la Science* 2022 et 2023 avec préparation d'ateliers sur le thème de la phylogénie et des séquences ADN.

— Conférence pour le festival *Pint of Science* 2024 sur la part de hasard dans l'évolution des séquences ADN.

— Intervention type *Ma thèse en 180 secondes*, en 2024, pour la journée des femmes en science organisée par l'université Lyon 1.

— Conférence pour l'association *Girls Can Code* lors d'un stage de programmation gratuit destiné à des jeunes filles de tout âge.

Formations

— MOOC « Se former pour enseigner dans le supérieur », 24h

— « Premiers secours en santé mentale standard », 14h

— MOOC « Éthique de la recherche », 15h

— « Assemblage et annotation de génome avec Galaxy », 20h

— « Histoire des sciences et épistémologie », 22h