Datations dans les arbres de gènes : spéciations, duplications, pertes

Guillaume Louvel

PSL
UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à l'École normale supérieure, Paris

Datations dans les arbres de gènes: spéciations, duplications, pertes

Soutenue par
Guillaume LOUVEL
Le 7 septembre 2020

Ecole doctorale n° 515
Complexité du vivant

Spécialité
Génomique comparative

Composition du jury :

Guillaume ACHAZ Pr, Collège de France	*Président*
Bastien BOUSSAU Dr, LBBE (Université Lyon 1)	*Rapporteur*
Julien DUTHEIL Dr, Max Planck Institute for Evolutionary Biology	*Rapporteur*
Maria ANISIMOVA Dr, ZHAW Life Sciences and Facility Management	*Examinateur*
Ingrid LAFONTAINE MC, IBPC (Sorbonne Université)	*Examinateur*
Hugues ROEST CROLLIUS DR2, IBENS	*Directeur de thèse*

ENS | PSL

Datations dans les arbres de gènes : spéciations, duplications, pertes

Guillaume LOUVEL

Thèse de doctorat
Septembre 2016 – Septembre 2020
Supervisée par : Hugues ROEST CROLLIUS

École normale supérieure, Paris Sciences et Lettres

Président du jury : Guillaume ACHAZ
Rapporteurs : Bastien BOUSSAU, Julien Y. DUTHEIL
Membres du jury : Maria ANISIMOVA, Ingrid LAFONTAINE

Table des matières

Introduction

Les échelles de temps sont indispensables à la compréhension de l'évolution. Avant d'avoir pu supposer que l'ensemble des êtres vivants est issu, par modification graduelle, d'un unique ancêtre commun, il faut avoir admis que l'âge de la Terre est suffisamment grand pour avoir laissé le temps à ces modifications, pour la plupart imperceptibles en une vie humaine. DARWIN (1859) a ainsi clairement souligné les échelles de temps considérables que révèlent l'accumulation de sédiments et l'érosion, et rapporte une mise en perspective appréhendable à notre échelle :

> Peu d'entre nous savent ce que signifie réellement un million d'années ; M. Croll propose l'illustration suivante : prenez une étroite bande de papier, longue de 83 pieds et 4 pouces [25,4 m. *Ndt.*], et déroulez-là contre le mur d'une vaste salle ; délimitez ensuite à l'une des extrémités un dixième de pouce [2,54 mm. *Ndt.*]. Ce dixième de pouce représentera cent ans, et la bande entière, un million d'années.

mais plus loin, il regrette :

> Malheureusement, nous n'avons aucun moyen de déterminer, en la rapportant à l'échelle en années, la longueur du temps qu'il faut pour modifier une espèce ;

(chapitre 11, *Sur le temps écoulé, inféré de la vitesse des dépôts et de l'étendue de la dénudation*).

Aujourd'hui, nous avons pu utiliser les datations radiométriques, absolues, pour déterminer l'âge de la Terre, 4,54 milliards d'années, connaître l'âge des couches géologiques, et trouver des indices de vie peut-être antérieure à 3,8 milliards d'années. Le tableau de cette odyssée, par le travail des scientifiques, nous force à la modestie, et en effet, « Our civilization would be pitifully immature without the intellectual revolution led by Darwin, even if we were equally well off economically without it. » (KIMURA 1983). Ce tableau est loin d'être achevé. Peut-être 86 % de la biodiversité reste à découvrir (MORA et al. 2011), et surtout celle-ci n'est pas figée. Comment elle réagira à la crise environnementale actuelle reste difficile à estimer.

Les dynamiques évolutives peuvent être comprises grâce au génome, un héritage commun à tous les organismes, au cœur de leurs cellules. Le travail d'archéologie génomique nécessite lui aussi des datations, et des mécanismes explicatifs. Les

génomes sont comparables à des palimpsestes, continûment effacés et réécrits, de sorte que les informations sur le passé deviennent fragmentaires, générant de nombreuses incertitudes.

Ces incertitudes font partie du travail scientifique, et ainsi, les datations absolues radiométriques, obtenues en mesurant dans les minéraux les proportions de certains isotopes à dégradation lente, ont une résolution limitée, quantifiée par un intervalle. D'ailleurs, les datations radiométriques furent accueillies suspicieusement, et des paléontologues eurent recours à la "géochronométrie" comme vérification indépendante de la radiochronométrie (HUSS 2017). Il s'agit d'analyser les cernes de croissance dans les fossiles, qui arborent des cernes annuelles, mais aussi journalières. Or, compter le nombre de jours par an informe sur la date, puisque la rotation terrestre ralentit à cause de la friction des marées (2 secondes tous les 100 000 ans). D'après les géophysiciens, il y a 390 millions d'années (Ma), une année contenait 393 à 396 jours. En mesurant 385 à 410 cernes par an dans des coraux du Dévonien moyen, WELLS (1963) a ainsi corroboré l'âge radiométrique de cette époque (sans toutefois parvenir à une meilleure résolution…).

En biologie moléculaire, ces datations absolues servent à calibrer les horloges moléculaires. Ce modèle a suscité de nombreux débats entre "molécularistes" et paléontologues, les âges des divergences entre espèces ne concordant pas toujours. Caractériser les incertitudes de chaque méthode, et combiner les différentes sources d'information, est donc un chantier nécessaire. De plus, l'horloge moléculaire peut s'avérer utile à la génomique en tant que tel, pour comprendre les dynamiques du génome, comme les duplications de gènes.

1 Génomes en évolution

1.1 Quelques bases pour non-biologistes

Le vivant, animaux, champignons, plantes, algues, plancton, bactéries, archées, etc., partage une organisation unique. Tous sont constitués de cellules, délimitant le milieu extérieur des composants métaboliques internes, et les macro-organismes que nous observons facilement sont pluri-cellulaires, chaque individu se développant à partir d'une cellule-œuf. L'unité remarquable du vivant se caractérise également par l'emploi des mêmes composants biochimiques assurant leur fonctionnement et leur reproduction. Les protéines et les acides ribonucléiques (ARN) assurent un rôle fonctionnel, ce sont les briques ou les moteurs des organismes. Étant formés d'une succession linéaire de résidus, respectivement les acides aminés (une vingtaine), et les ribonucléotides (quatre de base), c'est la séquence, puis son repliement tri-dimensionnel, qui détermine la fonction. Les protéines sont produites par traduction d'un ARN, dit messager, par correspondance entre triplets de nucléotides ou codons, et acides aminés, nommée code génétique. La séquence des ARNs est elle-même déterminée par une très longue molécule arrangée en double-hélice, l'acide désoxyribonucléique, ou ADN. L'ADN, structure très stable grâce à l'association complémentaire en paires de bases (A

vs T, C vs G), est transmis par réplication aux cellules filles lors de la division cellulaire, et transmis à la progéniture lors de la reproduction sexuée, constituant donc le support de l'hérédité. Dans la cellule, l'ADN est compacté avec des protéines sous forme de chromosome, et chez les organismes *eucaryotes* (étymologiquement, "à vrai noyau"), il se trouve majoritairement dans un compartiment interne, le noyau, mais une petite fraction se trouve dans les mitochondries, organelles intra-cellulaires assurant la production d'énergie par respiration, ou les chloroplastes des plantes effectuant la photosynthèse.

1.2 Contenu

Qu'est-ce donc que le "génome", sur lequel fonder toute une science comparative des organismes ? L'ADN en est le support matériel, mais le génome doit principalement être compris comme une "information", c'est-à-dire non pas comme le livre fait de papier, mais comme le texte inscrit dans le livre. L'ADN permet son stockage comme séquence de bases ("lettres" A, T, C, G). Ainsi, dans une cellule humaine, les vingt-trois paires de chromosomes du noyau contiennent le génome nucléaire, long de 3.10^9 paires de bases (pb), tandis que le génome mitochondrial est long de 16 569 pb. Du fait de la transmission avec erreurs de l'ADN aux cellules/mitochondries filles, le génome n'est donc jamais strictement identique entre cellules d'un organisme, encore moins entre individus d'une population, ou entre les chromosomes maternel et paternel des organismes diploïdes. Un génome étudié est donc toujours pris comme référence pour un ensemble de génomes jugés suffisamment similaires. Dans le cadre de comparaisons inter-espèces, nous nous rapportons donc à un génome de référence par espèce.

L'"information" du génome n'est pas synonyme de "signification", puisqu'en fait, tout le génome ne remplit pas de fonction pour l'organisme. Les gènes et les séquences qui régulent leur expression représentent une fraction variable du génome, propre à chaque taxon. La fraction qui encode les protéines est particulièrement faible chez les vertébrés, et chez les humains environ 1 % du génome est supposée codante, une proportion étonnamment infime à première vue. Pour illustrer la composition d'un génome, prenons le génome humain (chiffres d'après LYNCH 2007). On y trouve environ 20 000 gènes, tandis qu'au total les séquences intergéniques y occupent 68,3 %. Les gènes sont divisés en blocs codants discontinus, les exons, séparés par des introns, et les introns constituent 30 % du génome. Le génome non-codant contient des pseudogènes (gènes dégradés ayant perdu leur fonction), des séquences régulant l'expression des gènes (5 %), et en majorité des séquences répétées et des éléments transposables. Ces éléments transposables occupent 45 % du génome humain. Il s'agit de séquences capables de se copier dans un nouvel endroit du génome. Ce mécanisme est égoïste, dans le sens où il se fait au détriment du génome hôte, de sorte que les éléments transposables peuvent être vus comme des parasites du génome. La plupart de ces éléments est réprimée ou a perdu la capacité à se transposer. Leur grande proportion dans les génomes eucaryotes suggère un impact négatif faible rendant la sélection naturelle inefficace à les purger. Ils appartiennent à plusieurs classes disposant de mécanismes différents pour se copier, les éléments à LTR (Long-Terminal Re-

peats), qui sont des descendants de rétrovirus à ARN intégrés (8,3 %), les LINEs (Long Interspersed Non-coding Elements, 20,4 %), SINEs (Short Interspersed Non-coding Elements, 13,1 %) et les transposons (2,8 %). Les retrotransposons à LTR, LINEs et SINEs sont copiés-collés via leur transcription en ARN, tandis que les transposons produisent une protéine responsable de leur excision et insertion directement sur l'ADN. Les SINEs ne sont pas capable d'exprimer eux-mêmes les outils de leur transposition mais profitent de ceux des LINEs. Chez les primates, 70 % des SINEs sont de la famille *Alu*.

À cause de cette grande hétérogénéité fonctionnelle du génome, toutes les régions n'évoluent pas de la même façon. Ainsi, les pressions de sélection sont fortes sur les régions codantes et régulatrices, lesquelles sont plus conservées entre espèces que les régions répétées ou les éléments transposables évoluant essentiellement neutralement. La génomique comparative cherche à reconstruire l'histoire des organismes, et comprendre ses mécanismes, en interprétant les différences entre génomes de plusieurs organismes à la lumière de l'évolution. Pour remonter loin dans le temps, elle s'est donc principalement intéressée aux régions conservées, parmi lesquelles les séquences codantes des gènes sont les plus faciles à identifier et interpréter, grâce aux séquences promotrices reconnaissables, au code génétique, et à l'homologie des gènes entre espèces.

1.3 Les molécules comme archives de l'histoire évolutive

Si nous devons à LAMARCK l'invention du mot biologie, c'est à HAECKEL (1834–1919) que l'on doit celui de *phylogénie*, ou "origine des races/tribus/lignées". L'objectif de cette discipline est d'établir le degré d'apparentement entre organismes, ce que l'on représente sous la forme d'un arbre phylogénétique.

Même si les classifications hiérarchiques des êtres vivants remontent à l'Antiquité, et le système actuel à LINNÉ (1707–1778), l'idée de justifier une classification naturelle par l'*histoire évolutive* découle d'une théorie transformiste de l'origine des espèces, et à ce titre fait l'objet du chapitre 14 de l'*Origine des espèces* de DARWIN (1859). Le paradigme "qui ressemble à qui ?" évolue en "qui est le plus proche cousin de qui ?". À l'époque de Darwin, les descriptions de fossiles abondent, et les principes de sédimentation sont bien connus des géologues. Ils se révèlent alors être une manne d'information sur les apparentements entre espèces actuelles. À peine deux ans après la publication de l'Origine, un squelette fossile d'*Urvogel* remarquable fait sensation : dans des strates du Mésozoïque, avec ses plumes et ses dents, ses griffes aux ailes et une longue queue de lézard, l'archéoptéryx est bientôt présenté comme un "maillon" entre les dinosaures théropodes et les oiseaux, notamment par Thomas Henry HUXLEY (le "bulldog de Darwin"). Les outils méthodologiques à disposition pour construire cette classification sont alors essentiellement le concept que les ressemblances sont de deux natures : les *homologies* sont héritées d'un ancêtre commun, et sont donc pertinentes pour la classification, tandis que les *analogies* sont des similarités apparues indépendamment, en particulier à cause d'adaptations *convergentes*. Une formalisation rigoureuse des principes de systématique (classification) phylogénétique prend

corps avec la *cladistique* d'HENNIG (1966) [1], introduisant de nombreux concepts cruciaux comme la nécessité de n'utiliser que les caractères dérivés (par opposition à ancestraux) comme critère de regroupement, et de ne définir que des groupes *monophylétiques*, c'est-à-dire partageant un ancêtre *commun et exclusif*.

En parallèle de ces travaux exploitant les caractères morphologiques, la biologie moléculaire émerge, et s'intéresse de près aux protéines, puis à l'ADN, ces "cristaux apériodiques" selon Schrödinger, dont les séquences de résidus, à l'aune des mutations, se prêtent à une quantification de la similarité entre organismes —une approche dite *phénétique*, basée sur la similarité globale.

En effet, le lien entre ces molécules et l'hérédité a été progressivement déchiffré à l'aube du XX^{e} siècle, avec la synthèse néo-darwinienne (créditée à DE VRIES, JOHANNSEN et FISHER) qui apporte l'élément qui manquait à Darwin, un mécanisme générateur de la descendance avec modification. La découverte des lois de MENDEL, bientôt comprises sous l'angle de la ségrégation des chromosomes lors de la méiose, puis de la théorie chromosomique du gène initiée par MORGAN (DEUTSCH 2012), permet d'expliquer l'hérédité, et sa variation.

De fil en aiguille, la découverte de la structure en double-hélice de l'ADN et du code génétique passant par là, les mutations dans l'ADN, réflétées dans les protéines, fournissent une profusion de nouveaux caractères pertinents en phylogénie, qui ont l'avantage d'obéir à des "lois" physico-chimiques et mathématiques. En 1958, alors que sont connues seulement les séquences de l'insuline de cinq espèces, Francis CRICK présage (COBB 2017) :

> Biologists should realise that before long we shall have a subject which might be called "protein taxonomy" – the study of amino acid sequences of proteins of an organism and the comparison of them between species. It can be argued that these sequences are the most delicate expression possible of the phenotype of an organism and that vast amounts of evolutionary information may be hidden away within them.

L'image des "molécules comme archives de l'histoire évolutive" est formulée par ZUCKERKANDL et PAULING (1965) et amorce alors un nouveau pan de la biologie, l'évolution moléculaire, tout en introduisant la notion d'horloge moléculaire.

Pendant la première moitié du XX^{e} siècle, la génétique évolutive avait été une discipline essentiellement théorique. Avec des modèles mathématiques, WRIGHT, FISHER, HALDANE et consorts avaient posé le socle de ce qu'on appelle la génétique des populations, c'est-à-dire l'étude de la modification des fréquences alléliques dans des populations d'individus interféconds. Avec les progrès en biochimie, les prédictions des modèles allaient pouvoir se confronter aux données naturelles. À partir des toutes nouvelles informations sur le polymorphisme dans la nature, l'idée que les substitutions dans les protéines sont seulement le résultat de l'adaptation est remise en cause. Ce pan-adaptationisme moléculaire prend

1. dans sa parution anglaise, l'originale en allemand datant de 1950, et élaborant à partir de travaux du botaniste Walter ZIMMERMAN de 1931.

une sévère entaille avec la théorie neutre de l'évolution moléculaire (KIMURA 1968, 1983), qui, en plus de changer le paradigme, apporte une solution débloquant certains problèmes analytiques en génétique des population, grâce à l'approximation de la diffusion appliquée aux fréquences alléliques. Le message principal de cette théorie est sur l'impact quantitatif de l'évolution neutre : pour KIMURA, des contraintes sélectives agissent bien sûr sur les acides aminés, mais quantitativement, la majorité des sites évolueraient de façon neutre, par dérive génétique dans des populations de taille finie. Malgré les remous provoqués par cette proposition iconoclaste, cette théorie a fini par s'imposer, notamment comme une sorte de "modèle nul" pour tester l'adaptation. Avec les résultats expérimentaux, des modifications ont été apportées, comme avec la théorie presque neutre (OHTA 1992), postulant que la majorité des substitutions est légèrement délétère, mais pas suffisamment pour échapper complètement aux fluctuations aléatoires de la dérive.

Cette progression conceptuelle est accompagnée des tournants méthodologiques majeurs que sont le passage du séquençage protéique au séquençage de l'ADN (SANGER et al. 1977), et l'arrivée de l'amplification par Polymerase Chain Reaction, PCR (SAIKI et al. 1985) et de la technique "shotgun" aboutissant au génome humain vers l'an 2000. La phylogénie du vivant fut produite sans délai, grâce au séquençage des gènes d'ARN ribosomaux, très conservés entre organismes, par WOESE et FOX (1977) : l'arbre de la vie a trois branches primordiales, les eubactéries, les archées, (auparavant indistinctement regroupés sous le terme procaryote) et les eucaryotes.

Il est toutefois important de rappeler qu'un arbre phylogénétique est un modèle simplifié de l'histoire des organismes. En effet, un arbre représente uniquement les descendances par subdivisions successives. Bien que généralement adéquat, tout du moins pour les eucaryotes et les animaux, ce modèle ignore que l'évolution d'un organisme peut incorporer des réticulations, par les hybridations et les transferts horizontaux de gènes. L'un des meilleurs exemples en est l'origine des eucaryotes, résultant de l'endosymbiose d'une eubactérie (donnant les mitochondries) par une archée, et fusion de leurs génomes, ce qui a conduit certains à préférer la métaphore d'"anneau de la vie" plutôt qu'"arbre de la vie" (RIVERA et LAKE 2004).

2 Principes de la phylogénie

Pour établir une classification à partir de séquences protéiques ou nucléiques (figure 1), les principes suivants sont nécessaires : une séquence génétique est hérité par la descendance, avec la possibilité rare qu'une différence, c'est-à-dire une mutation apparaisse. Initialement dans un seul individu mutant d'une population, cette différence peut disparaître ou bien augmenter en fréquence, par le jeu de la dérive et de la sélection. Il y a donc une phase polymorphique, pendant laquelle deux "versions", ou alleles d'un même locus génétique coexistent dans la population, parfois suivie d'une *fixation* si l'allèle mutant finit par envahir toute la population. En comparant la même séquence homologue entre espèces, on observe donc ces différences. En supposant qu'elles se sont accumulées gra-

duellement, elles offrent une mesure de similarité entre espèces. Les mutations génétiques sont de nature variée, et celles qui sont pertinentes pour comparer des séquences sont de petite échelle : les mutations dites ponctuelles affectent un petit nombre de résidus de la séquence. Il s'agit soit d'une *substitution*, c'est-à-dire du remplacement d'un résidu par un autre, par exemple d'un nucléotide A (adénosine) par un nucléotide G (guanosine), ou bien de l'insertion ou délétion (indel) de quelques résidus.

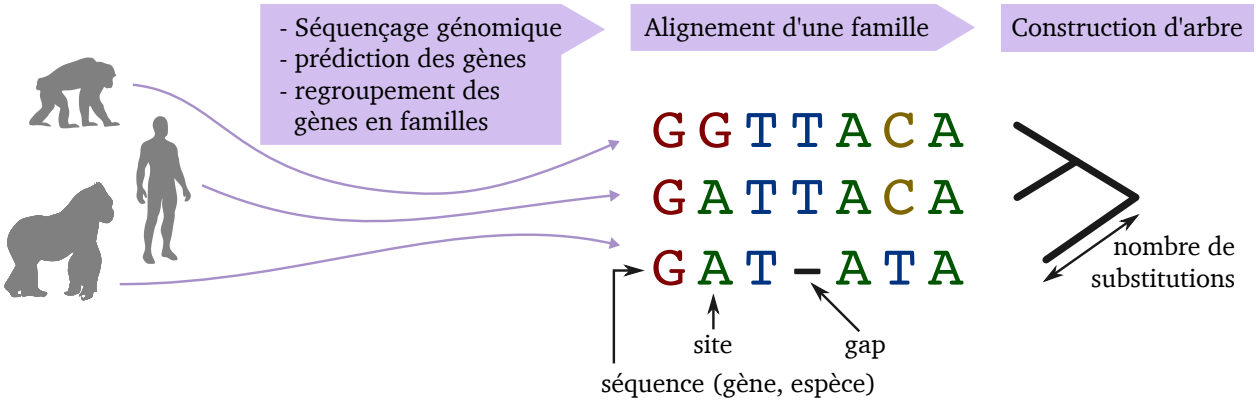


FIGURE 1 – Construire une phylogénie moléculaire

2.1 Aligner pour retrouver les positions homologues

Une substitution représente donc le changement d'un état de caractère homologue. La difficulté est de trouver quels sites sont homologues entre deux espèces. Lorsqu'il y a peu de différences, un humain peut aisément superposer les deux séquences et pointer les différences, éventuellement en introduisant les insertions-délétions sous forme de "gap". Lorsqu'il y en a davantage, ou que l'on veut comparer plusieurs espèces, la tâche est déléguée aux algorithmes d'*alignement*. Ces algorithmes superposent les séquences, produisant une matrice, avec les séquences en lignes, et les sites en colonne. Pour ce faire, des gaps sont introduits, et l'alignement recherché est celui qui maximise les similarités. On suppose alors que l'alignement a identifié les positions homologues entre séquences, et on peut passer à l'étape suivante de construction d'arbre, ou dendrogramme.

2.2 Le dendrogramme : des similarités au scénario évolutif

Cette tâche revient à trouver le meilleur arbre possible expliquant les similarités entre séquences, c'est-à-dire à formuler le scénario de descendance le plus probable. Un arbre phylogénétique raciné et étiqueté est constitué des composantes suivantes : les *feuilles* sont les nœuds terminaux, représentant les séquences actuelles ; la correspondance de chaque feuille avec une séquence particulière est un *étiquetage* ; la racine représente l'ancêtre commun le plus récent de toutes les séquences ; la *topologie* caractérise l'arrangement des branches, quelles que soient leur longueur et la rotation entre paires de nœuds frères par rapport au nœud parent.

Il faut cependant identifier les arbres compatibles parmi un nombre astronomique. Le nombre de possibilités augmente hyper-exponentiellement avec le nombre d'espèces : pour n espèces, les arbres possibles racinés, dichotomiques et étiquetés aux feuilles sont au nombre de :

$$3 \times 5 \times 7 \times \ldots \times (2n-3) = \frac{(2n-3)!}{2^{n-1}(n-1)!}$$

soit pour ne serait-ce que 20 séquences, 8.10^{21} (FELSENSTEIN 2004, chapitre 3).

Les méthodes de distances proposent de procéder par agglomération successive des groupes les plus similaires.

D'autres méthodes requièrent néanmoins d'explorer l'espace des arbres possibles afin de les évaluer, pour trouver celui qui optimise un critère d'adéquation aux données. L'exploration de l'espace des arbres n'étant la plupart du temps pas faisable exhaustivement, il faut donc procéder par heuristique, c'est-à-dire en sélectionnant avec des règles a priori les solutions que l'on suppose optimales. Dans les faits, on initialise souvent l'arbre par une méthode de distance, puis on procède à des petits changements de topologie jusqu'à ne plus être en mesure d'améliorer le critère d'évaluation de l'arbre.

L'un de ces critères est la parcimonie ou "quantité nette minimale d'évolution" (CAVALLI-SFORZA et EDWARDS 1963) : on suppose que l'hypothèse la plus plausible doit faire intervenir le moins de justifications non nécessaires. Un arbre autorisant un seul changement de caractère sera donc plus parcimonieux qu'un arbre nécessitant deux changements indépendants.

Alternativement, on peut inférer l'arbre par maximum de vraisemblance. Cela nécessite tout d'abord de construire un modèle d'évolution de séquence : le mécanisme évolutif est représenté comme résultant d'un phénomène aléatoire, mais obéissant à certaines règles, telles des lois de probabilités. Ce type de modèle laissant une part au hasard est nommé processus stochastique.

Cette formalisation fournit alors les formules nécessaires au calcul de la probabilité des données observées. Considérons par exemple des données de séquences $\mathscr{D}$, et un modèle d'évolution de séquence $\mathscr{M}$, alors la vraisemblance $\mathscr{V}$ est définie comme la probabilité que le modèle génère les données :

$$\mathscr{V}(\mathscr{M};\mathscr{D}) = P(\mathscr{D}|\mathscr{M})$$

Ainsi, avec un modèle de substitution où chaque nucléotide peut être remplacé équiprobablement par un autre, après un temps d'attente suivant une loi de probabilité exponentielle, on peut calculer la probabilité qu'un certain nombre de substitutions se produisent en un temps donné —autrement dit, le long d'une branche. Pour passer de la probabilité d'une branche à celle de l'arbre, comme les états ancestraux sont inconnus, il faut intégrer sur tous les états possibles, ce qui est faisable en un temps raisonnable grâce à l'algorithme d'élaguage ("pruning") de FELSENSTEIN (1981).

Une fois le modèle construit, et la vraisemblance calculable, on cherche alors à la maximiser en faisant varier les paramètres du modèle, par exemple l'arbre et les taux de substitution.

En résumé, pour construire un arbre à partir de données de séquences, dans un cadre probabiliste (maximum de vraisemblance ou bayésien), il faut proposer un modèle d'évolution de séquence, c'est-à-dire formaliser le processus stochastique responsable de l'accumulation de substitutions, puis inférer ses paramètres au moyen de la vraisemblance.

2.3 modèles d'évolution des séquences

Lorsqu'on compare deux séquences, le nombre de différences observées est une sous-estimation du nombre de substitution s'étant réellement produites. En effet, de multiple substitutions peuvent affecter un même site. Ce constat a mené JUKES et CANTOR (1969) à proposer la première modélisation du nombre de substitutions dans l'ADN, supposant que chaque nucléotide a une probabilité égale de remplacer le nucléotide d'origine, soit $1/3$. Ils en dérivent la formule reliant le nombre attendu de substitutions d au nombre de différences observées, p :

$$d = -\frac{3}{4}\ln(1-\frac{4}{3}p)$$

Les probabilités de chaque remplacement n'étant pas équiprobables, des modèles de substitutions plus complexes ont été introduits par la suite. Généralement, tous ont en commun une supposition : les sites d'une séquences évoluent indépendamment. Cela permet de calculer la vraisemblance de la séquence par le produit de la vraisemblance des sites.

L'idée sous-jacente est de représenter l'évolution d'un site comme une marche sautant d'un état à un autre. Ne sont observés que les états de départ et d'arrivée, mais entre les deux une infinité de chemins est possible. L'objectif est de calculer la probabilité de passage du nucléotide de départ par celui d'arrivée, en ayant pris en compte tous les chemins possibles.

Pour cela, des outils mathématiques bien établis sont à disposition, en commençant par remarquer que la probabilité de remplacement dépend seulement de l'état actuel du site, et pas des états précédents. Ce processus est donc sans mémoire. Pour intégrer sur le temps écoulé, une chaîne de Markov en temps continu est utilisée, qui nécessite d'expliciter uniquement les taux de transition instantanés. Pour des raisons pratiques, la substitution d'un nucléotide vers lui-même est considérée aussi. Dans le modèle de JUKES-CANTOR, les taux valent $r/4$, où r représente le taux global de substitution par unité de temps. Dans le modèle de FELSENSTEIN (1981), les proportions de chaque nucléotide sont représentées par des paramètres distincts $\pi_{A,C,G,T}$. Une matrice permet de représenter ces probabilités de transition, avec les états de départ en ligne et ceux d'arrivée en colonne :

$$Q = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{bmatrix} - & \pi_C & \pi_G & \pi_T \\ \pi_A & - & \pi_G & \pi_T \\ \pi_A & \pi_C & - & \pi_T \\ \pi_A & \pi_C & \pi_G & - \end{bmatrix} \end{array} \quad \text{(modèle F81)}$$

Pour simplifier les calculs, la diagonale est remplie de façon à ce que les sommes des lignes valent 0.

Les taux de transitions s'expriment alors comme le produit Qr de Q avec le taux global r. Pendant une durée de temps infinitésimale $\mathrm{d}t$, les probabilités de transition instantanées sont $(I + Qr\mathrm{d}t)$. On considère qu'un seul remplacement à la fois est possible pendant $\mathrm{d}t$.

En utilisant la propriété sans mémoire de Markov, on peut intégrer les probabilités de transition sur un intervalle de temps $t + \mathrm{d}t$. Si l'on connaît l'état au temps t :

$$P(t + \mathrm{d}t) = P(t)P(\mathrm{d}t) = P(t)(I + Qr\mathrm{d}t)$$

ce qui revient à

$$P'(t) = P(t)Qr\mathrm{d}t$$

et se résout par

$$P(t) = P(0)\exp(Qrt)$$

où l'exponentielle de matrice est définie comme $\sum_{k=0}^{\infty} Q^k/k!$. Heureusement, l'exponentielle de matrice se calcule beaucoup plus simplement après diagonalisation de Q, rendant les opérations réalisables numériquement.

Changer la paramétrisation de la matrice Q permet de formuler des modèles plus réalistes. Ainsi dans le modèle de KIMURA (1980), les taux de *transversion* sont distinct des taux de *transition* (au sens biochimique). Les transversions sont des substitutions d'une base de type purique (à deux cycles azotés, A et G) vers une base pyrimidique (un cycle, C et T) et vice-versa, les transitions conservent le type de base. D'autre part, les fréquences $\pi_{A,C,G,T}$ peuvent ne plus être égales. Le modèle le plus fréquemment utilisé est le modèle GTR, "General Time Reversible", dans lequel on attribue un taux par remplacement de nucléotide et son remplacement réciproque. C'est le plus général conservant la réversibilité temporelle, qui impose que $\pi_i Q_{ij} = Q_{ji}\pi_j$ (i et j étant des nucléotides), propriété importante pour simplifier l'optimisation de la vraisemblance.

Ces aspects mathématiques sont expliqués en profondeur dans (FELSENSTEIN 2004, chapitre 13) et BRYANT et al. (2005).

Avec l'algorithme d'élaguage de FELSENSTEIN (1981), l'inférence par maximum de vraisemblance permet ensuite d'optimiser tous les paramètres : la topologie de l'arbre, les longueurs de branches en substitutions ($r \times t$), et les ratios d'échanges de nucléotides de la matrice Q, faisant intervenir le cas échéant de multiples paramètres.

Ces modèles sont en réalité applicables à l'évolution de tout caractère discret. Pour les séquences d'acides aminés, on utilisera une matrice Q de taille 20×20, et pour un modèle de codon, 61×61 (GOLDMAN et YANG 1994).

Les modèles de codons apportent plusieurs avantages, notamment celui de prendre en compte la non-indépendance des sites d'un même codon, et, d'intérêt dans ma thèse, de mesurer le ratio $\omega = dN/dS$ de substitution non-synonymes sur synonymes. YANG et NIELSEN (1998) définissent la matrice de transition pour prendre en compte ω et κ (ratio transition/transversion) :

$$q_{ij} = \begin{cases} 0, & >2 \text{ différences,} \\ \pi_j, & \text{transversion synonyme,} \\ \kappa\pi_j, & \text{transition synonyme,} \\ \omega\pi_j, & \text{transversion non-synonyme,} \\ \omega\kappa\pi_j, & \text{transition non-synonyme.} \end{cases}$$

ce qui permet *in fine* d'estimer la valeur d'ω, puis le taux de substitution synonyme par codon, dS. Il faut donc réaliser que contrairement aux méthodes d'estimation du ratio dN/dS par comptage, comme NEI et GOJOBORI (1986), ici ω est un paramètre de taux instantané, obtenu en maximisant la vraisemblance. Pour obtenir dS, il faut utiliser le fait que la longueur de branche vaut $rt = S \times dS + N \times dN$, S et N représentant les nombres théoriques de sites synonymes et non-synonymes, respectivement. S et N sont en pratique calculés à partir des produits $Q_{ij}\pi_j$, i et j étant des codons, et π_j la fréquence à l'équilibre du codon, généralement obtenue à partir des fréquences empiriques dans l'alignement.

2.4 Un arbre et après ? La méthode phylogénétique comparative

Si reconstruire les relations de parenté est l'objectif premier de la phylogénie, c'est qu'il s'agit d'une étape préliminaire nécessaire. L'arbre phylogénétique est avant tout une description menant à de plus amples interprétations. D'ailleurs, le coproduit de l'inférence d'arbre par maximum de vraisemblance est peut-être au moins aussi important : les taux de substitution de nucléotides permettent en effet de caractériser le mécanisme d'évolution.

Dans ma thèse, je m'intéresse aux mécanismes de duplications de gène, et projette des perspectives sur leur connexion avec d'autres phénomènes, comme la diversification. C'est une préoccupation toute naturaliste que de vouloir tester les liens de causes à effets entre certains caractères, mais l'expérimentation directe étant la plupart du temps impossible pour des raisons pratiques (échelles de temps

longs, grandes tailles de population, paramètres environnementaux impossibles à contrôler, etc.), la déduction par méthode comparative est l'outil principal.

Malheureusement, les statistiques habituellement utilisées (tests de corrélation, test de Fisher exact, régression linéaire, etc.) partent d'une supposition forte : l'indépendance des points de mesure. Cette indépendance n'est pas validée pour un ensemble d'espèces, puisque leur existence est issue d'un processus de subdivision successives. Certaines paires d'espèces partagent donc des ancêtres communs plus récents. L'évolution procédant avec inertie, des caractères peuvent être partagés entre espèces à cause de leur ancêtre commun, et c'est d'ailleurs un des fondements permettant d'établir une phylogénie. Ainsi, considérer que les traits spécifiques sont indépendants entre espèces fausse totalement les tests d'hypothèse, en surestimant le nombre de degrés de libertés. Cela provoque des conclusions erronées, la plus problématique étant de rejeter à tort de l'hypothèse nulle.

De nombreuses justifications biologiques expliquent l'inertie phylogénétique (HARVEY et PAGEL 1991) :

1. Tout d'abord, la lenteur de l'évolution de certains caractères, même sous contrainte sélective. Cela peut être causé par un faible impact de la sélection naturelle, ou par des contraintes développementales, par la pléiotropie des gènes causant le maintien collatéral d'une fonction, etc.
2. Par ailleurs, l'écologie détermine la diversification des espèces : ainsi le conservatisme de niche écologique favorise les espèces apparentées à partager le même type d'environnement, donc les mêmes adaptations. En effet, imaginons la conquête d'un nouveau milieu, une forêt. Des espèces d'oiseaux déjà adaptées au milieu forestier auront une meilleure chance de s'établir avec succès que les espèces d'oiseaux de la prairie avoisinante.
3. Dans des taxons très éloignés, il y a de plus fortes chance que les mêmes adaptations soient résolues par des moyens physiologiques/développementaux différents, ou que des caractères différents répondent à une même force sélective.

Il est donc nécessaire de prendre en compte la non-indépendance entre espèces. Pour les traits discrets, la prise en compte des caractères sur les branches de l'arbre plutôt que sur les espèces est proposée par les méthodes de RIDLEY (1983) et MADDISON (1990), nécessitant une reconstruction préalable des caractères ancestraux par parcimonie. HARVEY et PAGEL (1991) proposent l'approche probabiliste modélisant le trait en temps continu, qui est parfaitement analogue aux modèles d'évolution de séquences. Elle permet de tester la corrélation par test de ratio de vraisemblance du modèle indépendant contre le modèle interdépendant (PAGEL 1994).

Du côté des traits continus, FELSENSTEIN (1985) apporte la méthode des contrastes indépendants phylogénétiques (PIC en anglais). Les différences entre paires de nœuds frères sont calculées, pour deux variables d'intérêt dont on teste la corrélation. Pour obtenir les contrastes des nœuds internes, une supposition sur le mode d'évolution du trait doit être effectuée, comme le mouvement

brownien. Le mouvement brownien est une marche aléatoire en temps continu, caractérisée par l'augmentation de la variance du changement du trait de façon proportionnelle au temps. Les PICs sont équivalents à la régression par moindres carrés phylogénétiques de GRAFEN (1989). La méthode de Grafen suppose aussi un mouvement brownien, qui se traduit par une matrice de covariances entre espèces, dont les éléments sont proportionnels à la longueur de branche partagée depuis la racine. L'avantage est de disposer du cadre flexible des moindres carrés généralisés, et de pouvoir reparamétrer facilement la covariance, par exemple en supposant un mode d'évolution plus contraint comme le processus d'Ornstein-Uhlenbeck.

Au cours de la progression de la discipline, il y a donc eu un passage d'une vision où l'arbre représente un facteur confondant à contourner à une vision où l'arbre devient l'outil de l'inférence et permet d'analyser le processus évolutif. Méthodologiquement, les modèles principalement utilisés proviennent des mêmes fondements mathématiques que pour l'évolution des séquences, ce cadre unificateur autorisant néanmoins une grande versatilité de modèles, incluant également la diversification (O'MEARA 2012). Les fondamentaux et les développements actuels de la méthode phylogénétique comparative sont vus en profondeur dans GARAMSZEGI (2014), et une introduction francophone a vu le jour récemment (ZAHARIAS et SANDERS 2018).

3 L'horloge moléculaire : hétérotachie universelle

L'horloge moléculaire est l'idée que les substitutions s'accumulent approximativement à taux constant dans l'évolution des espèces, ce qui offre l'opportunité de déterminer les âges de divergences par comparaison entre espèces actuelles. Le concept s'est rapidement élargi à des modélisations autorisant le taux à varier au cours du temps et des lignées, sous le nom d'horloges relâchées.

3.1 Historique d'une approche physicienne de l'évolution

La métaphore est formulée explicitement par ZUCKERKANDL et PAULING (1965), mais on peut en trouver un embryon dans les travaux de NUTTALL (1904), sur l'hémoglobine. Les prémices de la phylogénie moléculaire font alors appel à des techniques sérologiques : NUTTALL estime la dissimilarité entre hémoglobines de différentes espèces par l'intensité de la réaction sérum avec l'antigène.

À la fin des années 30, Linus PAULING, initialement un chimiste quantique, commence à s'intéresser à la biochimie, d'abord sous l'angle structural, puis aux enjeux biomédicaux. Travaillant par électrophorèse sur la structure de l'hémoglobine, il identifie la drépanocytose (anémie falciforme) comme une "maladie moléculaire" (PAULING et al. 1949). L'horloge moléculaire n'est en rien atomique, et pour cause : on doit peut-être sa conception à l'engagement de PAULING contre les essais nucléaires, un combat mené avec sa femme Ava Helen qui lui vaudra son second prix Nobel, pour la paix (MORGAN 1998). En effet, si les effets mutagènes des radiations sont connus à l'époque, leurs conséquences sont parfois négligées.

En 1958 un débat contre Edward Teller (père de la Bombe H, fervent partisan des essais nucléaires) qui affirme qu'un faible surplus de mutations peut même être bénéfique, oblige Pauling à trouver des arguments solides et il approfondit la génétique des populations.

Ayant rencontré Émile ZUCKERKANDL en 1957, il l'embauche à Caltech sur une bourse post-doctorale en 1959. Leurs premiers travaux examinent la similarité entre hémoglobines allant des grands singes aux vers polychètes (*Urechis*), d'après les empreintes bidimensionnelles générées par combinaison d'une électrophorèse et d'une chromatographie. Pour quantifier plus précisément les différences, ils utilisent le séquençage peptidique, alors naissant.

Cela donne lieu à leur article (ZUCKERKANDL et PAULING 1962) invité et non évalué par les pairs, volontairement provocateur, où ils utilisent l'idée d'horloge, sans la nommer, en calculant la date de divergence gorille-humain. Les taux de substitutions d'acides aminés sont considérés proportionnels non par génération, mais, quelque peu contre-intuitivement, par million d'années. Un taux constant est alors également observé sur le cytochrome c par MARGOLIASH (1963) et sur des fibrinopeptides par DOOLITTLE et BLOMBÄCK (1964). Les confrontations avec d'éminents évolutionnistes comme SIMPSON (1964) et MAYR (à "Evolving Genes and Proteins" 1965) ne se font pas attendre, car il s'agit pour eux d'une modélisation simpliste que les mécanismes d'évolution peinent à expliquer : en effet, si la sélection contrôle le taux de fixation des substitutions, comment se peut-il que celles-ci s'accumulent de manière constante ? Quid des temps de génération inégaux entre espèces ?

Toujours est-il que l'idée a une influence majeure sur la biologie évolutive, Zuckerkandl fondant en 1971 le "Journal of Molecular Evolution", et l'unité de mesure "pauling" (10^{-9} substitutions/site/Ma) ayant même été utilisée (par exemple dans OHTA et KIMURA 1971). Elle a également constitué un argument en faveur de la théorie neutre de l'évolution moléculaire, bien que pour Zuckerkandl et Pauling l'horloge était compatible avec la sélection.

3.2 Complications horologiques

Ces résultats très spécifiques (une protéine, peu d'espèces) ont rapidement été suivis de contre-exemples. Ainsi en 1964, Zuckerkandl observe que le cytochrome c évolue 1,5 fois plus vite que l'hémoglobine, et précise que l'horloge est une approximation. SARICH et WILSON (1966) proposent un test pour vérifier la constance des taux, et de nombreuses améliorations de ces "relative rate tests" sont proposées par la suite (FITCH et LANGLEY 1976 ; GU et LI 1992 ; TAJIMA 1993 ; cf NEI et KUMAR 2000, chapitre 10). Il a alors pu être démontré de nombreuses situations d'inconstance des taux. Tout d'abord le taux de substitutions par million d'années varie entre taxons (WU et LI 1985 ; BRITTEN 1986 ; PAGEL et al. 2006). De nombreuses causes à cette variation sont envisageables, comme les temps de générations, la non-neutralité des substitutions, la taille de population, l'efficacité du système de réparation de l'ADN, le niveau métabolique, etc. Pour un site donné, la variation de taux entre branches a été nommée *hétérotachie* (LOPEZ

et al. 2002). De plus, les taux varient également au sein d'un génome (WOLFE et al. 1989), sous les effets possibles de différences de pression de sélection, d'accessibilité de la chromatine, etc.

Bien qu'il soit argumenté que l'approximation de constance est valide en sélectionnant de multiples gènes et des espèces peu éloignées (KUMAR et HEDGES 1998), la plupart des modèles actuels prennent en compte des taux variables. Cependant, le mode de relaxation des taux fait débat : certains modèles supposent des taux autocorrélés d'une branche à l'autre, tandis que des modèles décorrélés supposent seulement une distribution globale comme la loi Gamma ou log-normale. Des études empiriques soutiennent soit des taux décorrélés (DRUMMOND et al. 2006 ; RANNALA et YANG 2007 ; HEATH et al. 2012), soit des taux autocorrélés (THORNE et al. 1998 ; SMITH et al. 2018 ; DOS REIS et al. 2018), et un modèle mixte mélangeant les deux suggère qu'une autocorrélation brownienne s'applique sur les grandes échelles temporelles, tandis que de brusques déviations locales sont possibles (LARTILLOT et al. 2016). En terme d'algorithme, l'implémentation probabiliste par maximum de vraisemblance d'une horloge à taux constant remonte à LANGLEY et FITCH (1974). En revanche, incorporer des taux relâchés entraîne des difficultés d'identifiabilité des taux avec les longueurs de branches : en cause, les branches d'un arbre phylogénétique représentent des nombres de substitution par site, c'est-à-dire le produit du taux et du temps $r \times t$. Il est donc possible d'expliquer une longueur de branche donnée par une infinité de taux et de durées, tant que leur produit reste le même. Ce problème d'identifiabilité est la raison pour laquelle les taux ne peuvent pas être relâchés indépendamment par branches, mais uniquement de façon interdépendante par les modes autocorrélés ou décorrélés présentés ci-dessus. SANDERSON (2002) résout ce problème dans un cadre semi-paramétrique supposant l'autocorrélation des taux : sa méthode par vraisemblance pénalisée autorise chaque branche à avoir un taux différent, mais une pénalité est appliquée sur la variance des taux entre branches consécutives. Le poids de la pénalité, relativement à la vraisemblance, doit être choisi en amont. Dans un cadre entièrement paramétrique CROTTY et al. (2020) ont récemment développé une méthode par maximum de vraisemblance dans le programme `IQ-Tree`. Alternativement, les algorithmes bayésiens sont un outil puissant pour modéliser les horloges relâchées, car ils offrent la possibilité d'ajuster des modèles plus riches en paramètres que les méthodes par maximum de vraisemblance. Parmi les programmes de datation bayésienne les plus utilisés, on trouve `Beast2` (BOUCKAERT et al. 2019) et `MCMCTree` (YANG 2007), fonctionnant par méthode MCMC (Monte-Carlo par chaînes de Markov). Les méthodes par MCMC lourdes en temps de calcul sont concurrencées par des algorithmes plus rapides, non paramétriques, fonctionnant par moindres carrés, comme LSD (TO et al. 2016) et TreeDater (VOLZ et FROST 2017), ce dernier autorisant des taux décorrélés. À la catégorie non-paramétrique s'ajoute également le Mean-Path-Length (MPL, BRITTON et al. 2002), qui obtient les profondeurs des nœuds en moyennant les longueurs de chemins descendants.

3.3 Jusqu'à quelle résolution ?

Les applications de l'horloge moléculaire consistent notamment à dater des spéciations, c'est-à-dire les divergences entre deux lignées de taxons, au moyen d'un grand nombre de gènes (des centaines) soigneusement sélectionnés et concaténés. Même dans cette situation optimale, les incertitudes de datation peuvent être conséquentes, et un corpus croissant de littérature aborde directement la précision, sujette à débat, des datations moléculaires (Graur et Martin 2004 ; Pulquério et Nichols 2007 ; Burbrink et Pyron 2008 ; Reis et Yang 2013 ; dos Reis et al. 2015 ; Zhu et al. 2015 ; Kumar et Hedges 2016 ; Warnock et al. 2017 ; Tao et al. 2019). De nombreuses limites ont été pointées, parmi lesquelles :

— l'artefact de densité de nœuds, résultant de la topologie de l'arbre. Lorsqu'il y a davantage de branchements, davantage de substitutions vont être inférées (Hugall et Lee 2007) ;
— la difficulté à caractériser le mode de relaxation des taux ;
— l'incertitude des points de calibrations eux-mêmes, résultant des âges et du placement des fossiles (dos Reis et al. 2015). Les incertitudes liées aux fossiles ont amené à développer des modes flexibles de rattachement des calibrations, les incorporant via un processus de naissance-mort.

Les datations moléculaires sont également d'intérêt crucial pour les divergences de lignées virales à l'échelle épidémiologique (To et al. 2016 ; Volz et Frost 2017 ; Bletsa et al. 2019), où l'on applique communément des modèles adaptés à des échantillonnages échelonnés dans le temps : contrairement aux arbres d'espèces où toutes les feuilles sont de même âge comparé aux profondeurs de nœuds en millions d'années, au cours d'une épidémie les dates de prélèvements doivent être prises en compte. Les méthodes adaptées sont dénommées "tip-dating" par opposition au "node-dating", car les informations de calibrations sont alors contenues dans les feuilles. De plus, la question des datations se pose aussi pour les duplications de gènes, pour l'évolution des microARNs (Sempere et al. 2006 ; Wheeler et al. 2009) et des éléments transposables (Boissinot et al. 2000 ; Ovchinnikov et al. 2002 ; Khan et al. 2006) ou dans la phylogénie appliquée au cancer (Lecca et al. 2015 ; Schwartz et Schäffer 2017). Il a été remarqué à maintes reprises que la longueur de séquence et le nombre de substitutions observables sont capitaux pour la précision de datation (Bromham et al. 2000 ; Lanfear et al. 2010 ; Smith et al. 2018 ; Bromham 2019 ; Bletsa et al. 2019), ce qui signifie qu'inférer des dates pour des virus, des éléments transposables ou des événements géniques sera plus difficile que de travailler avec des concaténations de centaines de gènes. En effet, la longueur limitée de telles séquences risque de compromettre toute conclusion, en accentuant les effets des incertitudes : 1) les erreurs d'origine méthodologique, comme lors de la reconstruction phylogénétique (topologie d'arbres et substitutions), 2) la variation d'origine biologique des taux et 3) la stochasticité inhérente au processus de substitution, bruitant le signal. Le premier chapitre de ma thèse s'attache donc à étudier la faisabilité de datations appliquées à l'échelle de l'arbre d'une seule famille de gènes, dont la longueur des séquences codantes varie d'environ mille à la centaine de milliers de nucléotides.

4 Les duplications de gènes

L'hémoglobine, point de départ des études historiques présentées ci-dessus, est un complexe formé de chaînes α et β chez l'adulte, dont la similarité n'a pas échappé à ZUCKERKANDL qui supposa une origine commune. Effectivement, les globines constituent une *famille* de protéines, des chaînes δ, ε existant également. En réalité, les gènes du génomes sont regroupés en familles multigéniques, c'est-à-dire que différents loci partagent une même séquence ancestrale. Un mécanisme mutationnel de type "copier-coller" en est à l'origine, la *duplication*. Deux gènes dont l'ancêtre commun est une duplication sont qualifiés de *paralogues*, comme les globines α et β, tandis que des gènes issus d'une spéciation (divergence de deux sous-populations devenant espèces distinctes) sont qualifiés d'*orthologues*.

4.1 Origine cellulaire des duplications de gènes

Quantitativement, les régions dupliquées représentent une source de variation entre individus plus importantes que les polymorphismes d'un seul nucléotide (SNP). Ainsi les sites variables entre humains dus à des régions dupliquées sont cinq fois plus nombreux que les SNP, soit 5 à 9 % (CONRAD et al. 2010 ; SUDMANT et al. 2015). Dans la lignée menant de notre ancêtre commun avec le chimpanzée, daté vers 6 Ma, à l'*Homo sapiens* actuel, environ 400 gènes auraient été dupliqués.

Les causes moléculaires des duplications sont variées : tout d'abord, les duplications complètes de génomes engendrent des organismes polyploïdes, possédant non pas deux paires de chaque chromosome, mais quatre ou plus. Cette modification drastique du caryotype peut en réalité être bien tolérée par l'organisme, car les proportions de chaque protéines demeurent inchangées. De nombreuses plantes, notamment d'intérêt agronomiques sont polyploïdes, et chez les métazoaires le phénomène n'est pas si infréquent : à la base des vertébrés, dont les génomes sont les mieux connus, il est admis qu'une double duplication complète de génome a eu lieu, et chez les poissons téléostéens, groupe de vertébrés le plus diversifié, une troisième duplication complète s'est produite. Des duplications complètes plus récentes se sont également produites, chez les salmonidés et chez la grenouille *Xenopus laevis*. Les mollusques et les arthropodes sont moins bien séquencés mais on commence à pouvoir y détecter des duplications complètes, comme chez les limules (double duplication) (KENNY et al. 2016), les arachnides (SCHWAGER et al. 2017), mais celle à la base des insectes reste débattue (ROELOFS et al. 2020). Les duplications *segmentales* affectent quant à elles une portion chromosomique. Leur existence, supposée notamment par STURTEVANT (1925), a été confirmée grâce aux colorations de chromosomes chez la drosophile (BRIDGES 1936). Ce sont les événements qui nous intéressent dans cette thèse car ils se produisent de façon continue. Cette mutation structurale peut résulter de plusieurs mécanismes cellulaires (revus par HASTINGS et al. 2009) :

— par rétro-transposition. Les ARNs des rétro-transposons peuvent parfois embarquer des séquences avoisinantes. La copie insérée du gène aura alors la particularité d'avoir été excisée, *i.e.* de ne plus contenir tous les introns ;

— les glissements de l'ADN-polymérase lors de la réplication peuvent répliquer de courts segments ;
— des crossing-over inégaux lors de la recombinaison méiotique peuvent accaparer l'un des chromosomes avec les deux allèles (tandis que l'autre n'en a plus aucun) ;
— les réparations des cassures double-brins, mécanisme de maintenance de l'ADN lors du cycle cellulaire, peuvent causer les duplications de plus grande taille. À l'endroit des cassures, la réparation utilise un patron de séquence similaire présent à un autre endroit du génome. La réparation peut se faire par recombinaison homologue (Homology Directed Repair), dirigée par une séquence homologue sur l'autre chromosome ou à un autre locus, ou alors elle peut exploiter les micro-homologies des extrémités simple-brins si présentes, et fait alors intervenir la NHEJ (Non-homologous End Joining). Des imprécisions dans ces deux systèmes peuvent créer des duplications.

4.2 Devenirs des duplicats

Les gènes dupliqués, ou paralogues, sont une source de redondance pouvant servir de matériel à l'évolution de nouvelles fonctions. En effet, l'apparition d'une copie d'un gène allège la sélection négative, puisqu'une "sauvegarde" peut maintenir la fonction originale. Bien que la majorité des duplications géniques aboutissent à une dégradation du gène le rendant non-fonctionnel (pseudogénéisation), il est possible que certaines mutations créent une nouvelle fonction, introduisant une innovation sélectionnée positivement (OHNO 1970). Cette idée de néofonctionnalisation s'oppose à une autre hypothèse sur le devenir des duplicats, la "duplication-dégénérescence-complémentation" (FORCE et al. 1999). Selon la DDC, la neutralité joue un rôle plus important dans le maintien des duplicats : les gènes ayant fréquemment plusieurs fonctions (pléiotropie), la dégénérescence des séquences dupliquées peut affecter les différentes fonctions aléatoirement dans l'une ou l'autre des copies. En revanche, une fois qu'une fonction est inactivée dans une copie, l'autre sera soumise à une sélection négative la conservant. Il y a alors complémentation des fonctions entre les deux paralogues. Enfin, un gène dupliqué entraîne une double production de la protéine ou de l'ARN. Si ce dosage a un avantage adaptatif, la sélection maintiendra le duplicat.

La prépondérance de chaque mécanisme de rétention des duplicats est débattue, mais les études soutiennent une combinaison de plusieurs mécanismes, dépendant probablement du gène dupliqué (INNAN et KONDRASHOV 2010).

À ces devenirs intra-spécifiques s'ajoutent des conséquences macro-évolutives. En effet, une duplication segmentale représente une variation dans la structure chromosomique, pouvant créer des infertilités, donc des barrières de reproduction menant à la spéciation. D'autre part, les incompatibilités génétiques peuvent résulter des fonctions des gènes dupliqués. Par exemple, l'incompatibilité de Bateson-Dobzhansky-Muller peut résulter d'une duplication (PRESGRAVES 2010) : un modèle proposé implique que deux populations isolées perdent indépendamment l'un des paralogues, de sorte qu'entre ces deux populations, l'unique version fonctionnelle restante soit localisée sur des chromosomes différents. Il

est alors possible que la seconde génération hybride produisent des génotypes sans aucune version fonctionnelle. Ainsi, en terme de biologie des espèces, les duplications pourraient non seulement avoir un impact adaptatif par l'apport de nouveau matériel génétique, mais aussi un impact sur la spéciation par la création d'incompatibilités génétiques.

4.3 Paralogie et phylogénie

Du point de vue de la phylogénie moléculaire, les duplications compliquent la reconstruction des arbres d'espèces : en effet, dans l'arbre de gènes, un événement de duplication engendre deux sous-arbres évoluant en parallèle dans les espèces descendantes, si bien que l'arbre de gène ne réprésente pas seulement les relations entre espèces, ou *orthologies*, mais aussi les relations entre loci, ou *paralogies* (figure 2).

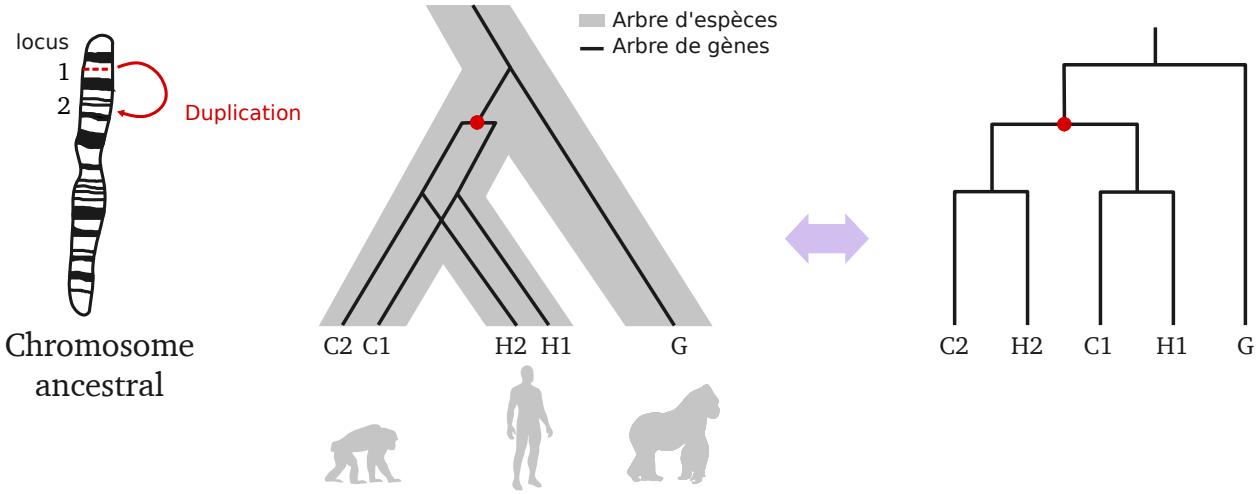


FIGURE 2 – Exemple d'arbre de gènes contenant une duplication, ayant lieu chez l'ancêtre humain-chimpanzée. Au centre, sur l'arbre d'espèces, l'évolution des loci 1 et 2 a lieu en parallèle, résultant en deux gènes paralogues H1 et H2 chez l'humain, et deux autres C1 et C2 chez le chimpanzée. À droite, l'arbre de gène correspondant ne montre donc pas la phylogénie des espèces.

Lorsqu'on cherche à reconstruire uniquement l'arbre d'espèces, on choisira donc préférentiellement des gènes présents en un seul exemplaire dans chaque espèce considérée. Cependant, si c'est l'histoire du gène qui est recherchée, il est nécessaire de déterminer pour chaque nœud s'il s'agit d'une duplication ou d'une spéciation, et à quel taxon l'associer. Comme les pertes de gènes compensent les duplications, un arbre de gène peut correspondre à de nombreux scénarios possibles. Une cartographie de l'arbre de gène par rapport à l'arbre d'espèces est nommée *réconciliation*, et des algorithmes permettent d'inférer la plus probable. Dans la base de données Ensembl utilisée dans ma thèse, la méthode `TreeBeST` combine la vraisemblance de l'arbre de gène avec un critère de parcimonie sur les duplications-pertes. Alternativement on peut modéliser les duplications-pertes par un processus naissance-mort prenant place dans chaque lignée d'espèce, avec taux de naissance (*i.e.* duplication) par gène constant (OTA et NEI 1994), et implémenter une réconciliation entièrement probabiliste (ARVESTAD et al. 2003 ;

COMTE et al. 2019). ALE (SZÖLLŐSI et al. 2013), et plus récemment GeneRax (MOREL et al. 2020), permettent d'incorporer les transferts horizontaux (modèle DLT: Duplication-Loss-Transfer), autre source de discordance entre arbre de gènes et arbre d'espèces. Ces deux programmes ne modélisent cependant pas l'horloge moléculaire, autrement dit les longueurs de branches en millions d'années. Introduire une horloge à taux variable a été réalisé dans un cadre bayésien (incluant les transferts, TOFIGH et al. 2009) et est applicable avec le programme `prime-DLRS` ("delirious", SJÖSTRAND et al. 2012). Ces modèles sont très riches en paramètres et nécessitent donc une grande quantité de données.

Les sources d'incongruences entre arbres de gènes et arbres d'espèces ne s'arrêtent cependant pas aux duplications-pertes-transferts, mais incluent également la conversion génique et le tri incomplet de lignées (Incomplete Lineage Sorting, ILS). La conversion génique est la modification d'une séquence d'après une homologue proche, pouvant donc affecter les paralogues, en particulier en tandem. En conséquence de ces échanges de fragments de séquences, la divergence entre deux gènes apparaîtra plus récente qu'elle ne l'est par pure descendance verticale. L'ILS est causé par la ségrégation partielle des allèles dans des populations divergeant rapidement et peut survenir ainsi : si une population présente un locus polymorphique, il est possible que les sous-populations descendantes, devenant des espèces filles, héritent également du polymorphisme, donc de plusieurs allèles. Si cette spéciation est rapidement suivie d'autre spéciations et que les allèles ancestraux ne se fixent qu'après, les espèces filles vont arborer des versions de gènes dont la divergence est antérieure à leur ancêtre commun. Prendre en compte l'ILS est donc essentiel en particulier pour dater les spéciations rapprochées. Des méthodes permettent de reconstruire les arbres de gènes avec un modèle Duplication-Perte-Transfert-ILS (DLTI) (RASMUSSEN et KELLIS 2012).

À l'échelle macroévolutive, les phylogénies de gènes présentent le potentiel pour reconstruire les dynamiques passées de duplications. Des variations entre lignées sont à attendre, et il a par exemple été observé une augmentation globale du taux de duplication et perte chez les primates (HAHN et al. 2007).

En perspective, étudier les duplications sur une échelle macroévolutive représente un défi technique, car cela présuppose des données déjà difficiles à construire : pour un arbre phylogénétique réconcilié, des incertitudes interviennent au moment du séquençage et de l'assemblage du génome, de la prédiction des gènes, du clustering des séquences homologues, du choix du transcrit, de l'alignement, de la reconstruction de l'arbre et de sa réconciliation (SCORNAVACCA et al. 2020, partie 2 : "Data quality, model adequacy").

5 Objectifs de la thèse

Le champ de la génomique comparative s'est penché de façon approfondie sur l'évolution des gènes, et sur l'impact des duplications de gènes. De nombreuses interrogations persistent néanmoins sur le rôle adaptatif de ces dernières. Pour tester le lien entre duplication et adaptation, il nous faut déterminer de façon

fiable la dynamique des duplications. C'est pourquoi l'objectif central de cette thèse est d'établir des dates et des taux de duplication à l'échelle génomique.

Dans une première partie, nous avons focalisé notre étude sur une vingtaine de primates, en quantifiant la précision des méthodes à notre disposition pour dater des événements géniques. En choisissant d'abord les événements de spéciation comme référence car leurs dates sont connues, nous avons pu déterminer les facteurs corrélés à la précision des dates obtenues, à partir des caractéristiques des arbres de gènes. Il s'agit du développement méthodologique principal de la thèse.

Dans une seconde partie nous utilisons le modèle de la première partie pour prédire l'erreur de datation sur les nœuds de duplication. Nous dressons ensuite un profil des taux de duplication : les familles de gènes arborent visiblement des taux de duplication variables et nous étudions donc leur distribution. Enfin nous présentons des analyses préliminaires sur la corrélation phylogénétique entre taux de duplication et taux de spéciation.

La troisième et dernière partie porte sur un autre type d'événement génique : la perte de gène par un taxon. Cet événement peut également être daté, au sens phylostratigraphique : grâce aux arbres réconciliés, la comparaison inter-espèces permet d'assigner la perte à un ancêtre donné. Dans le cadre d'une collaboration avec Chris GORDON et Jeanne AMIEL de l'Institut Imagine à Paris, nous nous intéressons aux pertes de gènes dans le cadre du système de latéralisation de l'embryon. L'un des gènes impliqué, MMP21, présente un profil de pertes très particulier qui reflète l'évolution de cette voie développementale. La recherche de profils similaires pour d'autres gènes et séquences régulatrices nous amène à proposer de nouveaux candidats qui pourraient également y jouer un rôle chez les mammifères.

Chapitre I

Précision empirique des datations géniques

> The clock in Lord Vetinari's anteroom didn't tick right. Sometimes the tick was just a fraction late, sometimes the tock was early. Occasionally, one or the other didn't happen at all. This wasn't really noticeable until you'd been in there for five minutes, by which time small but significant parts of the brain were going crazy.
>
> *Going Postal*, Terry PRATCHETT

L'objectif de la thèse étant de fournir des dates estimées à partir d'arbres de gènes individuels, nous quantifions d'abord la précision à attendre d'une telle méthodologie. Pour cela, nous mesurons la dispersion empirique des âges de spéciation estimés chez les primates, dans 5 235 arbres sans duplication ni perte de gène. Ces résultats font l'objet de l'article (soumission août 2020) : « Factors influencing the accuracy in dating single gene trees », LOUVEL et ROEST CROLLIUS (p. d.), retranscrit pour l'essentiel dans ce chapitre.

6 Procédure de datation et gamme de précision

6.1 Procédure de datation

Nous avons mis en place sept procédures de datation, se décomposant en deux phases principales (cf. Méthodes, figure I.8, C et D) :

1) calculer ou filtrer un alignement multiple,
2) dater les nœuds en ajustant une horloge moléculaire.

Les arbres réconciliés sont prérequis et obtenus d'Ensembl, de même que les alignements originaux.

6.1.1 Obtention de l'alignement

L'alignement est une source importante de variation dans les estimations de taux de substitution. De plus, ayant sous-échantillonné les alignements d'Ensembl, nous avons la possibilité de les recalculer plus finement. Dans cette première phase de la procédure de datation, nous avons testé quatre stratégies : (i) *original*, lorsque l'alignement d'Ensembl est retenu tel quel ; (ii) *cleaned*, lorsque l'alignement est filtré pour supprimer les segments mal alignés avec le programme `HmmCleaner` (Di Franco et al. 2019) ; (iii) *fsa*, lorsque les séquences ont été réalignées avec le programme `FSA`, un aligneur probabiliste et conservatif (Bradley et al. 2009) ; (iv) *fsa+cleaned*, combinant (iii) suivi du filtre (ii) sur l'alignement `FSA`.

6.1.2 Datation

Dans cette seconde phase, nous avons appliqué deux méthodes de datation, soit l'algorithme *Mean-Path-Length* (MPL) de Britton et al. (2002), soit un modèle d'horloge relâchée avec `Beast2` (Bouckaert et al. 2019). Le MPL requiert d'abord de précalculer les longueurs de branches synonymes, que nous produisons avec `Codeml`, du package `PAML` 4 (Yang 2007), selon deux modèles alternatifs : 1) le modèle par branche "free-ratio" (ici appelé "branch model"), ou 2) un modèle constant par branche mais avec un taux ω variable par site suivant une loi Gamma (ici appelé "site model"). Le MPL transforme ensuite ces longueurs de branche en âge absolu aux nœuds, au moyen d'un lissage non paramétrique : les longueurs de chemin d'un nœud à toutes ses feuilles descendantes sont moyennées, produisant un arbre ultramétrique. Cette distance moyenne au présent est ensuite étalonnée en âge absolu au niveau des points de calibration (cf. Méthodes, 12.6.1). Par contraste, `Beast2` est entièrement paramétrique et prend en compte l'incertitude dans la topologie de l'arbre, l'incertitude dans les calibrations, et un modèle de variation du taux de substitution par année (cf. Méthodes, 12.6.2.2).

6.1.3 Sept procédures alternatives

En combinant ces deux phases, nous avons testé les sept procédures suivantes :

(1) *original+siteMPL*,
(2) *original+branchMPL*,
(3) *cleaned+branchMPL*,
(4) *fsa+branchMPL*,
(5) *fsa+cleaned+branchMPL*,
(6) *fsa+Beast*,
(7) *fsa+cleaned+Beast*.

En entrée, les arbres de gènes et les séquences codantes d'Ensembl Compara ont été élagués pour ne retenir que les séquences *Simiiformes* et deux séquences extragroupe. *Simiiformes*, ou singes anthropoïdes, est le groupe primate frère des *Tarsiidae*, subdivisé en *Catarrhini* (singes de l'Ancien Monde et grands singes) et *Platyrhini* (singes du Nouveau Monde) (figure I.4). Dans le but que tous les

arbres de ce contrôle présentent la même topologie, nous avons éliminé les arbres avec des duplications ou des pertes, obtenant 5 235 arbres. Le nœud *Simiiformes*, ancêtre commun daté à 43,15 Ma (d'après TimeTree) a servi de calibration pour estimer conjointement l'âge des douze nœuds de spéciation restants, avec chaque procédure de datation.

Après application de ces procédures, je regroupe ensemble les âges obtenus d'une même spéciation et d'une même procédure, constituant 12×7 distributions d'âges constituées de 5 235 observations (arbres de gènes). La figure I.1 montre les deux procédures dont les distributions d'âge sont les moins étalées sur l'arbre d'espèces. En rappelant que l'attendu théorique est de retrouver chaque datation à proximité de l'âge de spéciation, il en ressort un étalement considérable en millions d'années, suggérant une incertitude de datation dépassant aisément 5 voire 10 Ma. Ainsi, selon ces deux procédures les moins étalées, certains âges d'ancêtres consécutifs semblent indistinguables : *Hominoidea* par exemple recouvre quasiment tout *Hominidae*.

6.1.4 Mesures d'erreur de datation

Afin de quantifier les erreurs de chaque procédure, dans cette section, nous rapportons la dispersion des dates dans chacune de ces 12×7 distributions, et deux mesures de leur magnitude :

1) l'intervalle entre les quantiles à 2,5 % et 97,5 %, que je nomme intervalle interquantile à 95 %. Cet intervalle contient les 95 % centraux des dates d'une spéciation, et donne une idée de l'étalement général.
2) la moyenne de la déviation absolue à la médiane (mAD) informe quant à elle de la tendance moyenne à l'écart, et constitue un bon compromis entre sensibilité (car elle prend en compte tous les points), et robustesse (en prenant la médiane comme centre).

Additionnellement, l'asymétrie des distributions (skew) et la kurtosis (caractérisant le poids des valeurs non-centrales) ont été calculées pour caractériser la forme de la dispersion des âges

Ici, pour simplifier le vocabulaire, nous assimilons la dispersion des dates à l'erreur de datation, mais la nuance entre les deux est détaillée en 6.3.

6.2 Intervalles interquantiles à 95 %

Retenons l'étendue de cet intervalle (figure I.2), en fraction de l'âge de l'ancêtre calibré (43,15 Ma) : l'intervalle d'*Hominoidea*, grand de 17 Ma, recouvre donc $17/43 = 40\,\%$ du temps jusqu'à *Simiiformes*.

Les intervalles en millions d'années contenant 95 % des dates pour chaque spéciation et chaque procédure sont représentés en figure I.3.

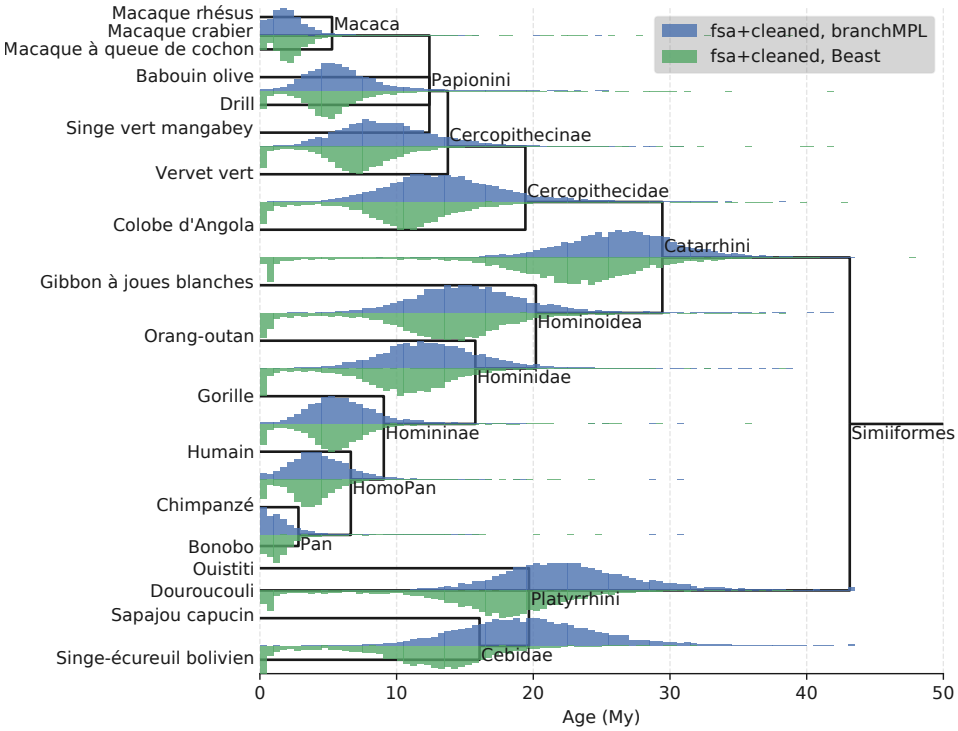


FIGURE I.1 – Distributions des âges des spéciations des 5 235 arbres selon les deux procédures de datation les moins étalées. L'arbre d'espèces sous-jacent est daté d'après TimeTree.

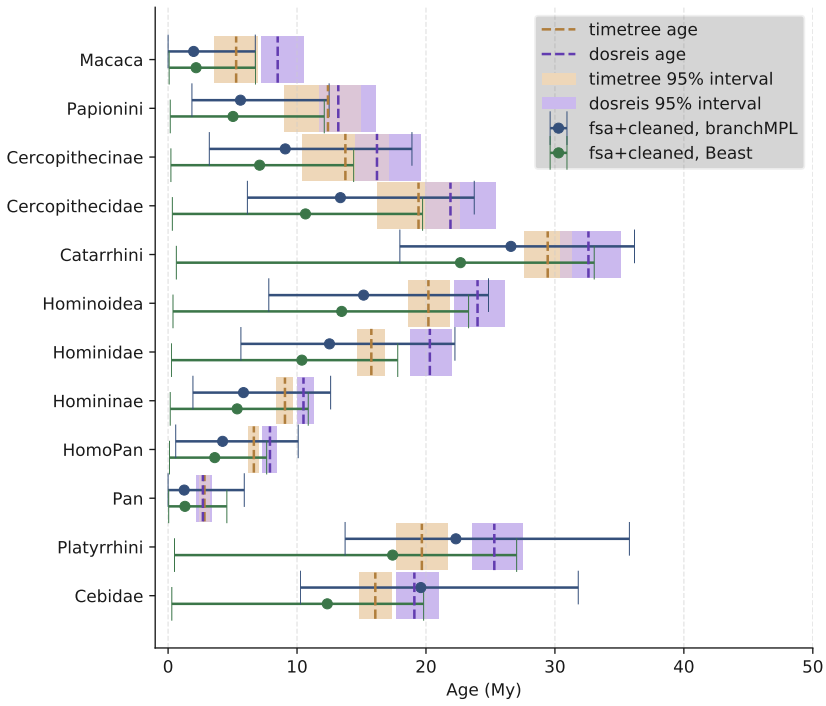


FIGURE I.2 – Intervalles interquantiles à 95 % des datations des deux procédures les moins étalées. Les âges de références (DOS REIS et al. 2018), et TimeTree (KUMAR et al. 2017) sont en arrière-plan.

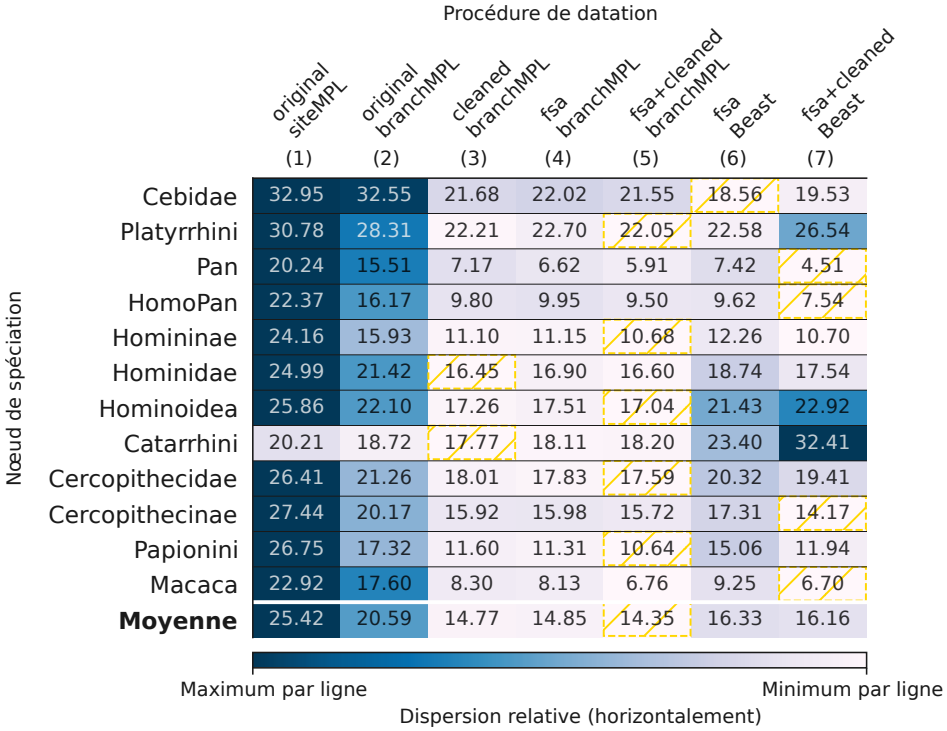


FIGURE I.3 – Tailles d'intervalles interquantiles à 95 % des datations (Ma), par spéciation et par procédure. Le gradient de couleur est calibré par ligne, et les minimums sont hachurés en jaune.

6.3 Comparaison avec des âges de référence extérieurs

Les mesures de dispersion exposées plus haut quantifient l'écart par rapport au centre de la distribution, ici la médiane. Il ne s'agit donc pas *stricto sensu* de mesures d'*erreur*. En effet, on peut imaginer des résultats très peu dispersés mais complètement décalés de la "vraie" valeur recherchée.

Nous avons comparé nos âges estimés avec deux références : la première source est TimeTree (KUMAR et al. 2017), une base de données d'âges consensus obtenus en rassemblant les datations moléculaires publiées, tandis qu'une seconde source provient d'une estimation entièrement paramétrique, effectuée chez les primates (DOS REIS et al. 2018), et fournissant chacune des intervalles de confiance/crédibilité à 95 % (figure I.2).

Une tendance forte en ressort : nos âges médians sont quasi systématiquement plus jeunes que ceux de nos deux références, plus particulièrement par rapport à l'estimation paramétrique de DOS REIS et al. (2018). Remarquablement, ils ne sont pas toujours contenus dans les intervalles à 95 % des références. Les décalages par rapport aux références sont plus marqués dans les descendants de *Catarrhini*.

Plusieurs raisons pourraient l'expliquer, notamment un modèle trop simple dans notre procédure ne prenant pas en compte des taux par espèces distincts ou les particularités propres à l'évolution d'un seul gène (hors tri incomplet de lignées, qui devrait produire des âges plus anciens que la spéciation).

Alternativement, TimeTree dépend d'études aux méthodologies variables et utilisant différentes calibrations fossiles, source d'incertitudes. Ce décalage de nos estimations vers le présent concerne aussi bien les résultats des méthodes MPL que Beast, malgré l'emploi par cette dernière d'une horloge relâchée et de distributions continues d'âges de calibrations.

La raison la plus probable me semble être le faible nombre de substitutions utilisables dans un seul arbre de gène, qui créerait un décalage ciblé sur les nœuds les plus récents. Dans des alignements de séquences pauvres en différences, il est plausible que les différences proviennent préférentiellement des branches internes, lesquelles séparent des groupes de séquences plus éloignés et nombreux, tandis qu'il y aurait trop peu de substitutions après les nœuds récents pour les placer suffisamment profondément. Enfin les sous-arbres descendants des nœuds *Catarrhini*, et surtout *Hominoidea* sont déséquilibrés, à l'instar d'*Hominoidea* qui engendre d'un côté une seule espèce, le gibbon, et de l'autre un groupe de cinq hominidés (figure I.1).

Pour finir, les deux références elles-mêmes ne sont étonnamment pas concordantes, certains de leurs intervalles à 95 % étant même disjoints. Nous contournons donc cette absence de consensus en nous basant sur nos médianes comme valeurs de référence, afin de disséquer plus en détail les facteurs influençant la précision de datation. De plus nous écartons ainsi les biais dûs au taux de substitutions spécifiques par lignée, ce qui permet de considérer seulement la variabilité gène à gène.

7 Évaluation de la procédure de datation

Dans un premier temps, nous montrons quelle stratégie est la plus à même de réduire l'erreur de datation, d'après la mAD. Les résultats présentés ici montrent le rôle crucial de la qualité de l'alignement, mais également que le modèle de substitution avec variation par branche est plus performant qu'avec variation par site.

7.1 Sélection pour une mAD minimale

7.1.1 *fsa+cleaned+branchMPL* produit les plus faibles erreurs, de façon stable

La mAD de chaque nœud de spéciation est calculée et représentée dans la figure I.4. En s'appuyant sur la moyenne par procédure (dernière ligne de la figure) comme critère principal de choix, la meilleure procédure est *fsa+cleaned+branchMPL*, avec une mAD moyenne de 2,7 Ma. Selon les spéciations, les mAD minimales sont atteintes majoritairement par *fsa+cleaned+branchMPL* et *fsa+cleaned+Beast*, qui ont toutes deux généré cinq minimums sur douze spéciations. Cependant, avec cette dernière, certaines mAD sont notablement plus élevées (*Catarrhini* et *Hominoidea*), pénalisant la moyenne. Ces deux procédures se distinguent seulement par la méthode de datation, suggérant que Beast produit des résultats plus

hétérogènes que le MPL en fonction des conditions. Ce résultat se produit également avec l'intervalle interquantile à 95 %. Il est surprenant de constater qu'une méthode sophistiquée comme Beast donne finalement des erreurs similaires au MPL, comparativement plus basique.

Du côté des statistiques d'asymétrie et kurtosis (annexe A4), la procédure *fsa+beast* (sans nettoyage des alignements) présente des valeurs extrêmes, signe de distributions très étirées. Cela indique qu'avec cette procédure, certains arbres produisent des dates extrêmement éloignées de la médiane. Notons que ce phénomène peut prendre moins d'ampleur dans le MPL, car cette méthode contraint strictement la date entre le présent et l'âge de la calibration, tandis que la calibration dans Beast est flexible. La variabilité de l'erreur produite par Beast en fonction de l'ancêtre, ainsi que ses asymétries et kurtosis, résultent probablement d'une sensibilité plus forte au faible nombre de sites dans l'alignement. Étonnamment, le filtrage des alignements inverse le sens de l'asymétrie avec Beast, résultant peut-être d'un nettoyage trop drastique par `HmmCleaner`. Ainsi Beast représente dans un sens plus fidèlement les données extrêmes, propageant l'incertitude de l'entrée au résultat final.

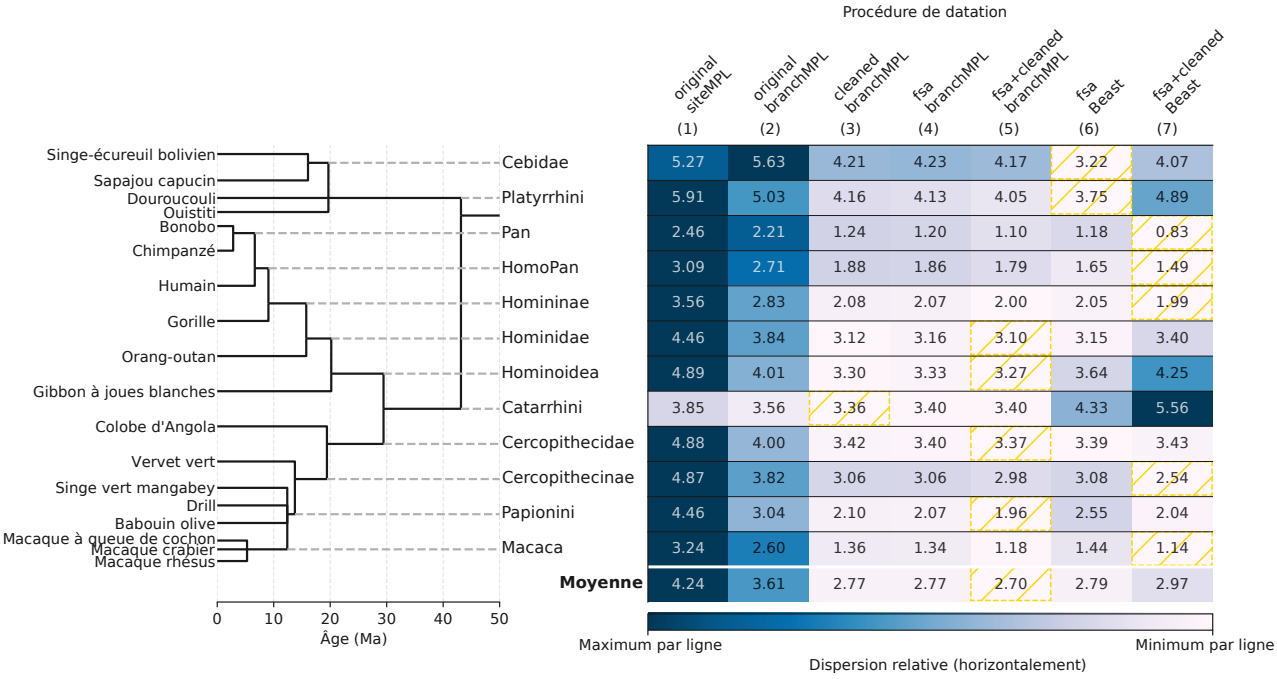


FIGURE I.4 – Déviations absolues moyennes à la médiane (mAD) de l'erreur de datation en Ma, par spéciation et par procédure. Le gradient de couleur est calibré par ligne, et les minimums sont hachurés en jaune.

7.1.2 La qualité de l'alignement et le modèle par branche réduisent drastiquement l'erreur

La comparaison quantitative des mAD entre les sept procédures (figure I.4) révèle principalement l'influence de l'alignement. Plus précisément, en comparant uniquement les procédures par *branchMPL* (2 à 5), l'alignement *original* produit quasi-systématiquement la plus grande erreur, tandis que les traitements *cleaned*,

fsa et *fsa+cleaned* produisent des erreurs similaires. Curieusement, parmi les douze spéciations, *Catarrhini* et *Hominoidea* réagissent différemment à l'alignement : *Catarrhini* est modérément affecté par les alignements originaux, mais bien plus par le distinguo MPL *vs* Beast.

Autre conclusion remarquable, notre évaluation statue sans ambiguité sur le choix de modèle de substitution (avant le MPL) : à partir des mêmes alignements (*originaux*), le modèle par branche (2) est systématiquement et substantiellement meilleur que le modèle par site (1), un résultat qui nous surprend au vu de la littérature sur le sujet (cf. Discussion I.11).

En conclusion, d'après les valeurs de dispersion (mAD) ci-dessus, nous décidons que la procédure la plus précise est *fsa+cleaned+branchMPL*, et nous l'utilisons dans la suite. De façon pratique, elle autorise également à fixer autant de points de calibrations que voulu, ce qui permet d'estimer les âges de spéciation une à une.

8 Caractéristiques propices à une datation précise

Nous déterminons maintenant quelles caractéristiques des arbres corrèlent avec la précision des dates estimées fournies par la procédure *fsa+cleaned+branchMPL*.

Afin d'obtenir une valeur d'erreur *par arbre*, nous moyennons les déviations d'âge des nœuds d'un même arbre, obtenant 5 235 mesures de déviation absolue moyenne.

Nous régressons ensuite cette erreur par arbre contre plus de 60 caractéristiques liées à la qualité de l'alignement, les taux de substitutions, ainsi que les mesures générées par les programmes de la procédure, tels que la vraisemblance du modèle de `Codeml`, ou la proportion de nucléotides supprimés par `HmmCleaner`. Parmi ces caractéristiques, plusieurs mesures utilisent le dS calculé par `Codeml`, c'est-à-dire le nombre de substitutions synonymes par site, ainsi que le dN, ou nombre de substitutions non-synonymes par site. Pour éviter toute confusion sur les termes, "taux de dS" est employé ici en tant que nombre de substitutions par site *et par million d'années*, par opposition au dS simple, qui est seulement le nombre de substitutions par site.

En premier lieu nous éliminons les arbres possédant des branches de longueur nulle (en dS ou en dN), ou contenant des séquences non chevauchantes (non alignables), donc impropres à la datation. Environ 3 000 arbres sont ainsi supprimés, résultant en un jeu final de 1 436 arbres à régresser. Deux vérifications justifient cette suppression en amont : d'abord, lorsque ces arbres sont gardés dans la régression, les caractéristiques considérées impropres obtiennent une forte corrélation avec l'erreur ; de plus en testant leur impact une par une, toutes ces caractéristiques sont associées à une plus grande erreur (p-valeur $< 0{,}01$, 12 tests de Student unilatéraux avec correction de Benjamini-Hochberg), sauf une (le non-chevauchement de certaines séquences). Ainsi, nous écartons en amont ces facteurs confondants que nous savons fortement associés à l'erreur.

Avec une régression Lasso de l'erreur par arbre, nous sélectionnons les variables avec un coefficient absolu supérieur à 0,01, et réajustons une régression par moindres carrés ordinaires. D'après les valeurs de coefficients de la régression multiple (figure I.5), la caractéristique principale augmentant l'erreur est l'hétérogénéité du taux de dS par millions d'années, ou hétérotachie du dS (p-valeur = 0), suivie par la longueur d'alignement (p-valeur = 0) et le pourcentage de GC (p-valeur = 0), et le ratio $\omega = dN/dS$ médian (p-valeur = 0,036). Parmi les variables éliminées par Lasso se trouve notamment le taux de dS moyen par arbre (proportionnel à la longueur totale des branches en nombre de substitutions). Le R^2 ajusté obtenu par les moindres carrés ordinaires est de 0,48.

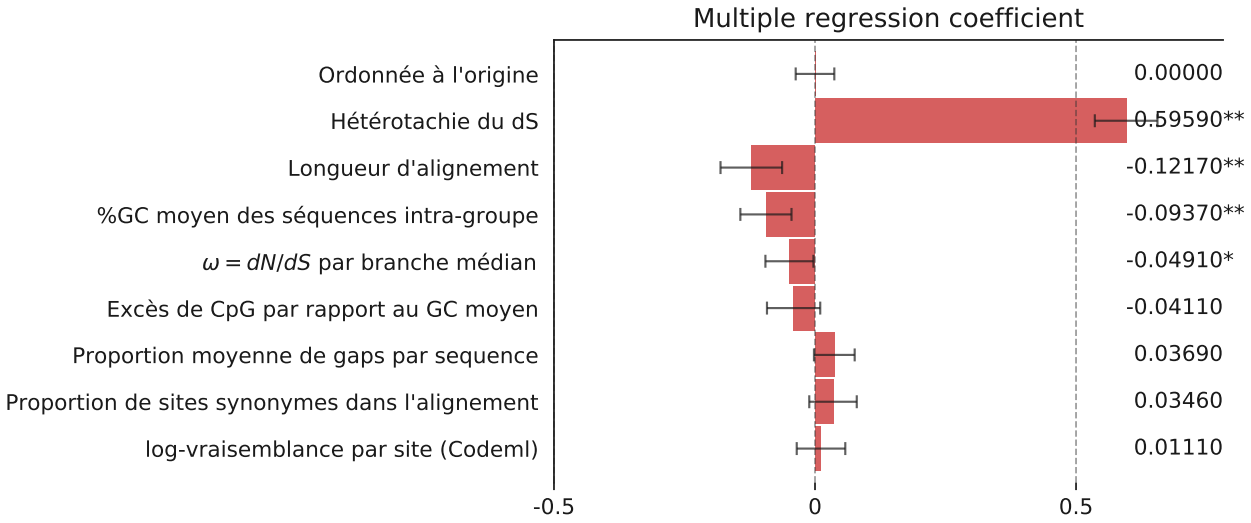


FIGURE I.5 – Coefficients de régression des caractéristiques normalisées contre la variable réponse "déviation d'âge moyenne par arbre". Les coefficients sont triés par valeur absolue (sauf l'ordonnée à l'origine), indiquant leur importance relative dans la déviation d'âge. Les barres d'erreur représentent les intervalles de confiance à 95 %. Cet ajustement des coefficients est obtenu par moindre carrés ordinaires après sélection par Lasso des variables à coefficient absolu supérieur à 0,01. (**: p-valeur < 0,01; *: p-valeur < 0,05)

Suite à cette comparaison relative, il est intéressant de ramener les caractéristiques à leurs unités d'origine, pour percevoir le contraste réel entre les caractéristiques des arbres les moins précisément datés et ceux les mieux datés. Pour cette comparaison, l'erreur *prédite* (i.e. l'ordonnée du plan de régression) sert à départager les arbres, en retenant le 1er et le dernier décile. Du décile inférieur de précision au décile supérieur, l'erreur *observée* passe de $3{,}12 \pm 1{,}13$ Ma, à $1{,}17 \pm 0{,}38$ Ma. Si ces erreurs observées paraissent toutes deux avantageusement faibles, notons cependant qu'elles résultent d'une moyenne entre nœuds, moyenne influencée principalement par les nœuds récents plus nombreux, aux erreurs plus faibles. La différence entre ces deux groupes extrêmes d'arbres, du moins précis au plus précis, correspond à une diminution d'un facteur 3,3 en hétérotachie, et à une augmentation d'un facteur 5,4 en longueur d'alignement, tandis que le pourcentage de GC et l'ω par branche varient plus modestement.

TABLE I.1 – Caractéristiques moyennes (± écart-type) des arbres aux deux extrémités d'erreur prédite.

	Erreur (Ma)	Hétérotachie du dS (10^{-3} subst/codon/Ma)	Longueur d'alignement (nucléotides)	GC moyen (%)	ω médian
10% de plus faible précision prédite	3,12 ± 1,13	1,90 ± 1,16	1324 ± 734	52,0 ± 8,7	0,20 ± 0,14
10% de plus haute précision prédite	1,17 ± 0,38	0,57 ± 0,27	7159 ± 3664	56,1 ± 7,4	0,31 ± 0,26

En conclusion, d'après les corrélations de ces caractéristiques avec l'erreur par arbre, nous concluons que la précision de datation est principalement liée à l'existence de suffisamment de substitutions, dépendant en premier lieu de la longueur d'alignement. L'autre influence majeure sur la précision provient de taux apparemment inconstants, causés soit par la reconstruction phylogénétique, soit par une hétérogénéité biologique.

9 Facteurs de précision localisés à un nœud inconnu

Après avoir établi les principaux facteurs influençant la précision de datation moyennée sur un arbre, considérons que chaque nœud occupe une position distincte. La précision de datation pourrait très bien être influencée par des facteurs spécifiques de certains nœuds. Une distinction évidente de chaque nœud réside dans leur âge de référence, allant de *Catarrhini* daté à 29,4 ± 1,9 Ma dans TimeTree, aux genres très récents *Pan* et *Macaca*. En outre, notre phylogénie d'espèces est constituée de paires de clades frères avec un nombre variable de descendants, pouvant être très équilibrés comme les deux sous-arbres descendants de *Catarrhini*, ou très déséquilibrés comme les sous-arbres descendant de *Hominoidea* (figure I.4), créant des artefacts de densité de nœuds (HUGALL et LEE 2007). D'autre part, dater les nœuds séparément permet de considérer leurs erreurs indépendamment des autres parties de l'arbre, c'est-à-dire sans être affecté par les branches situées au-delà des spéciations immédiatement voisines, puisque ces dernières sont maintenant calibrées. La procédure de datation *fsa+cleaned+branchMPL* a été exécutée pour chaque spéciation, en adaptant le schéma de calibration : tous les nœuds excepté celui d'intérêt sont calibrés, toujours d'après les âges de référence de TimeTree. Nous avons ensuite appliqué la même procédure de régression que précédemment, mais les variables de taux et d'hétérotachie ont été mesurées séparément pour les branches directement connectées au nœud d'intérêt, et pour les branches situées au-delà dans l'arbre.

D'importance, les R^2 des régressions par spéciation sont nettement plus faibles (allant de 0,103 pour *Papionini* à 0,317 pour *Pan*, annexe A6) que pour la régression de l'erreur des arbres entiers ($R^2 = 0{,}47$, figure I.5), vraisemblablement

parce que l'erreur locale est moins explicable par des caractéristiques moyennées sur l'ensemble de l'arbre.

Les douze régressions produisent des classements légèrement différents des caractéristiques ayant le plus d'influence sur l'erreur de datation, et nous avons sélectionné les cinq plus influentes de chaque régression (par union, 8 au total) pour comparaison (figure I.6). En remarque sur les méthodes, étant donné la corrélation entre le taux de dS local et le taux de dS non-local, nous avons retenu seulement les résidus du premier, interprétables comme une différence entre l'accélération locale de cet arbre et l'accélération locale moyenne (un résidu positif signifie que le taux local est supérieur à l'attendu d'après le taux non-local, attendu obtenu par régression simple sur tous les arbres). Les résultats de ces régressions par spéciation confirment que la constance des taux autour du nœud daté est importante pour réduire l'erreur (a), davantage que l'hétérotachie non-locale (b). De plus, on détecte ici le rôle majeur des taux de dS dans l'amélioration de la précision, avec une contribution plus forte du taux moyen non-local (c), et une contribution moindre de son accélération locale (d). En résumé, l'hétérotachie locale est un bon prédicteur de la précision, de même que le taux de dS de l'arbre entier (non-local), et dans une moindre mesure l'accélération locale du dS. Nos résultats sont cohérents avec l'attendu qu'un dS élevé fournit davantage de puissance statistique, et réduit donc l'erreur. En accord avec la régression globale, la longueur d'alignement (e) diminue l'erreur dans les datations par spéciation.

Enfin, l'objectif étant d'observer l'impact du nœud de spéciation sur l'erreur, nous les avons trié par âge dans la figure I.6, mais aucun effet d'âge évident n'en ressort. C'est plutôt inattendu car les nœuds les plus anciens dépendent de comparaisons de séquences plus distantes, entre lesquelles il paraît plus difficile d'estimer correctement les substitutions.

10 Âges médians de spéciation

Les âges médians sont présentés ici pour référence (tableau I.2). On remarque que la procédure (4) *fsa+branchMPL* sous-estime moins les âges que (5) *fsa+cleaned+branchMPL*, confortant l'hypothèse d'un nettoyage trop drastique par cette dernière.

Les taux médians par espèces pour procédure (5) sont présentés en figure I.7. Les branches les plus courtes (Macaque crabier, *Macaca*, *HomoPan*) affichent les taux les plus extrêmes, mais les valeurs ne diffèrent pas de plus d'un facteur 2. Les taux sont plus rapides sur la branche *Cercopithecidae* que sur *Hominoidea*, en accord avec ce qu'on connaît de la biologie de ces espèces, mais contrairement à des résultats antérieurs (Moorjani et al. 2016), la branche *Catarrhini* est aussi lente qu'*Hominoidea*, et le taux humain n'est pas inférieur aux taux chimpanzé-bonobo.

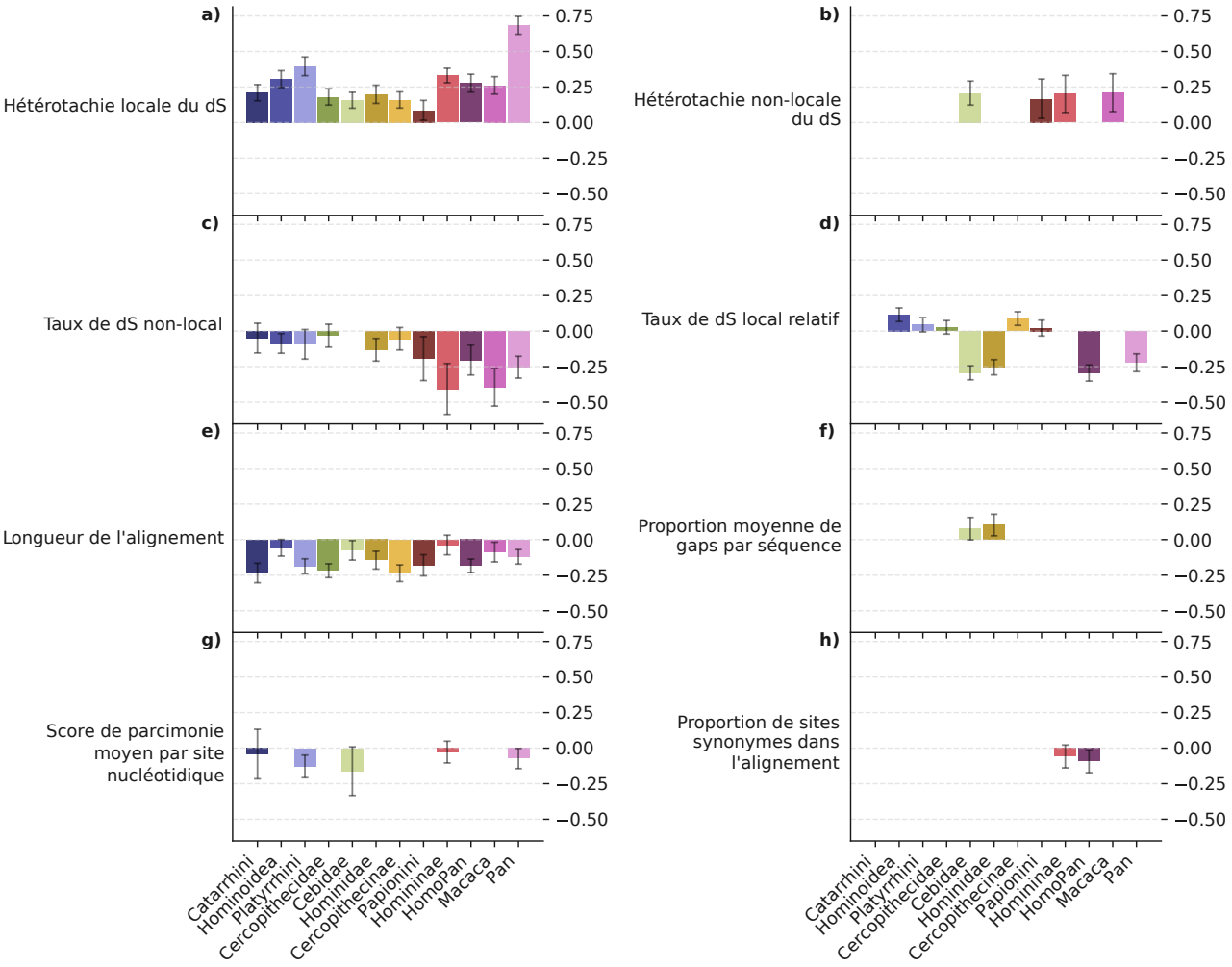


FIGURE I.6 – Coefficients de régression pour huit caractéristiques ayant le plus d'influence sur l'erreur de datation, pour chaque nœud ancestral (trié par âge). Les barres d'erreur représentent l'intervalle de confiance à 95 %

	TimeTree	dos Reis 2018	(1) original siteMPL	(2) original branchMPL	(3) cleaned branchMPL	(4) fsa branchMPL	(5) fsa+cleaned branchMPL	(6) fsa Beast	(7) fsa+cleaned Beast
Cebidae	16,07	19,10	18,60	19,97	19,61	19,60	18,11	13,84	12,35
Platyrrhini	19,68	25,30	22,67	22,69	22,41	22,39	20,71	18,79	17,41
Pan	2,82	2,70	1,41	1,37	1,29	1,31	1,17	1,55	1,31
HomoPan	6,65	7,90	4,51	4,50	4,28	4,32	3,83	4,12	3,62
Homininae	9,06	10,50	6,24	6,15	5,96	5,96	5,52	6,02	5,36
Hominidae	15,76	20,30	13,06	13,00	12,69	12,66	11,79	11,74	10,37
Hominoidea	20,19	24,00	16,37	15,78	15,32	15,26	14,63	14,93	13,46
Catarrhini	29,44	32,60	27,60	27,27	26,69	26,71	24,71	24,47	22,67
Cercopithecidae	19,42	21,90	14,71	14,39	13,56	13,54	12,21	11,79	10,66
Cercopithecinae	13,75	16,20	10,40	10,15	9,33	9,34	8,15	7,94	7,10
Papionini	12,40	13,20	6,58	6,42	5,82	5,78	4,79	5,67	5,04
Macaca	5,28	8,50	2,16	2,23	2,04	2,07	1,84	2,49	2,17

TABLE I.2 – Âges de spéciation médians en Ma, par procédure.

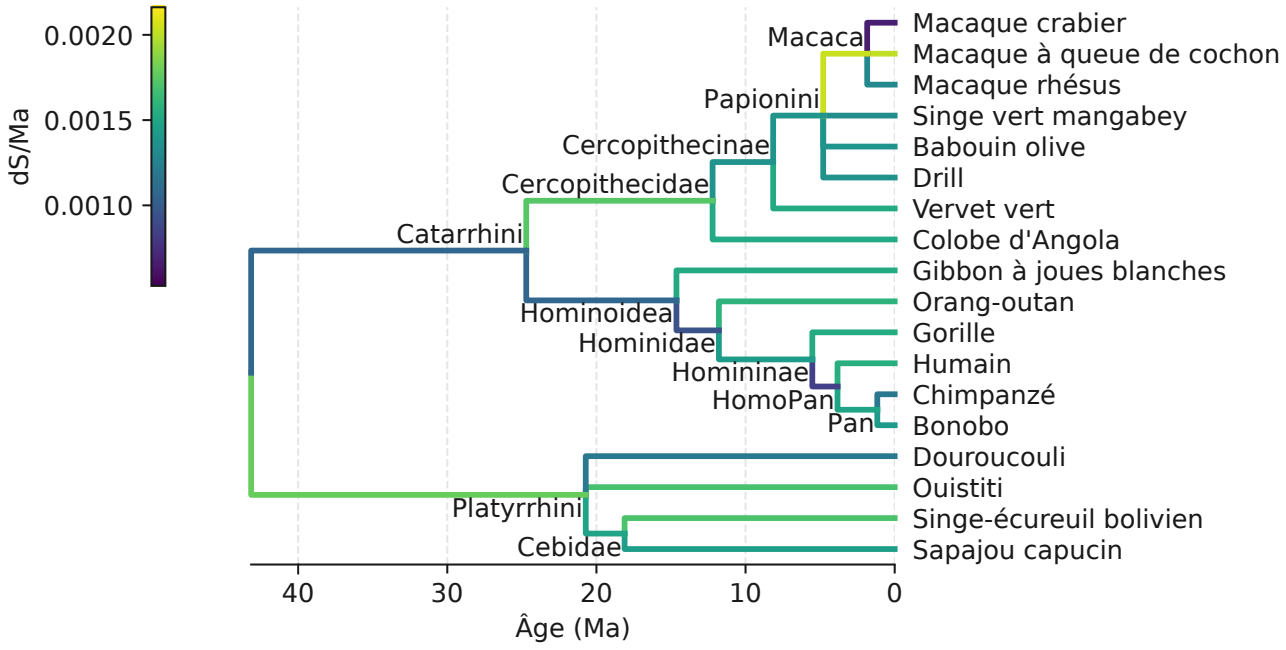


FIGURE I.7 – Taux de substitution médians par lignée, procédure (5) *fsa+cleaned+branchMPL*, en substitutions synonymes par codon par Ma. Les longueurs de branches en Ma proviennent de la même procédure (5). Tous les taux ont été calculés avec deux séquences extragroupe.

11 Discussion

Les datations moléculaires sont une étape préliminaire à de nombreuses analyses évolutives, comme relier l'évolution des espèces avec les climats et environnements passés, comparer l'évolution de plusieurs clades, mesurer la dynamique de diversification, trouver l'origine d'une épidémie, etc. Cependant, la puissance statistique des méthodes de datation dépend de la quantité de donnée disponible (LANFEAR et al. 2010 ; SMITH et al. 2018), et c'est la raison pour laquelle les âges de spéciation sont estimés à partir de concaténations d'alignements. Ici, nous

étudions en détail le comportement des datations moléculaires appliquées à des familles de gènes individuelles, où le nombre de sites est donc bien plus faible que dans les concaténations d'alignements typiquement utilisées. Toutefois, les histoires d'arbres de gènes individuels, notamment les duplications de gènes, intéressent les évolutionnistes. Comme illustration, l'exemple classique de la famille des globines a été revisité par AGUILETA et al. (2006) qui ont réestimé les taux de substitution et les âges de divergence des paralogues. Avec pour objectif ce type d'analyse à l'échelle du gène, nous avons donc estimé empiriquement la confiance dans les datations d'un seul gène, à partir de toutes les familles de gènes strictement orthologues des primates. Nous avons déterminé l'intervalle d'erreur obtenu pour les arbres de gènes individuels, statistiquement limités en termes de quantité d'information. D'abord, sur le jeu original de 5 234 arbres de gènes sans duplication ni perte, a priori de meilleure qualité, 3 558 d'entre eux révèlent des nombres de substitutions trop faibles, caractérisés par des branches de longueur nulle. Nous trouvons également que les autres 15 061 arbres incluant les duplications ou pertes sont encore moins appropriés pour une datation, car seulement 2 774 ont pu être retenus selon le critère des longueurs de branches non nulles. De plus, ces arbres-là ont également des alignements plus courts, ajoutant à la difficulté. Notre approche est empirique, puisque nous avons utilisé une base de donnée génomique d'arbres de gènes pour estimer l'erreur de datation dans cet échantillonnage. En tant que telle, notre approche est distincte et complémentaire d'estimations d'intervalles de confiance basées sur du bootstrap ou des distributions théoriques.

11.1 Magnitude et cause de la déviation moyenne des âges.

Premièrement, nos intervalles empiriques à 95 % atteignent jusqu'à 18,2 Ma, pour *Catarrhini*, ce qui représente une durée importante comparée aux 43 Ma de la calibration *Simiiformes*, bien que cette durée soit plus faible pour les nœuds plus récents. En complément, la déviation moyenne des âges (mAD) semble suffisamment basse pour autoriser des analyses ultérieures, malgré des données aberrantes, dont nous avons pu identifier les caractéristiques par régression.

Les 5 234 arbres sélectionnés pour ce contrôle représentent une fraction biaisée des 22 000 arbres de gènes enracinés à *Simiiformes*. En effet ils constituent vraisemblablement un sous-jeu optimal pour la datation, car l'absence de duplication ou perte pourrait corréler à une évolution plus stable en termes de taux et de pression de sélection. Dans cet échantillon optimal, nous avons détecté des variations de certaines caractéristiques influençant la précision de datation : comme attendu, davantage de puissance statistique est fournie par un taux de dS élevé et des alignements plus longs, tandis qu'une forte inconstance des taux entre branches augmente l'erreur. Cela confirme des observations précédentes pointant le nombre de sites et de substitutions comme facteur limitant la puissance statistique. Cela confirme également que les taux moléculaires ne sont généralement pas constants entre branches, même après avoir pris en compte l'incertitude due aux alignements courts. De plus, le taux moléculaire n'est pas non plus constant entre gènes d'un même génome. Cette variation affecte les

substitutions synonymes, ouvrant de nombreuses questions sur ses causes : les substitutions synonymes sont-elles suffisamment neutres pour une datation ? À quel point la stochasticité du processus de substitution affecte-elle l'hétérotachie, et comment l'incertitude s'amplifie-t-elle lors de l'inférence ? Ce manque de fiabilité des taux est connu dans le domaine, et des solutions pragmatiques ont été proposées pour gérer leur hétérogénéité, comme le "gene-shopping" de SMITH et al. (2018), où les arbres aux taux apparemment hétérogènes sont supprimés en amont des analyses. Notre régression des erreurs de datation a permis d'identifier de telles caractéristiques indésirables, et pourrait être utilisée pour chercher les explications biologiques.

Dans notre sélection de procédure, nous avons rapidement éliminé la stratégie *siteMPL*, clairement moins précise que la *branchMPL*. Ce résultat contraste avec les observations précédentes qu'un taux d'horloge ajusté par site serait plus précis qu'un taux par branche (SMITH et al. 2018 ; WANG et al. 2019). Il y a cependant une nuance importante entre les articles cité et notre résultat : ces études mentionnent le mode de variation du taux *de l'horloge* directement, tandis que dans nos calculs, il s'agit de la variation du $\omega = dN/dS$, lequel influe indirectement sur la datation. Naturellement, on peut s'attendre à une amélioration en ajustant simultanément les deux modes de variation, ce qui pourrait être fait avec les modèles de codons implémentés dans `HyPhy` (KOSAKOVSKY POND et al. 2019).

Notre analyse est une première tentative de quantification des facteurs prévalents dans la précision de datation par arbre de gènes. Elle révèle également quelles innovations récentes pourraient apporter une amélioration :

— une réconciliation des arbres de gènes avec l'arbre d'espèces entièrement probabiliste, telle qu'implémentée dans `GeneRax` (MOREL et al. 2020) ;
— pour les époques récentes, prendre en compte le tri incomplet de lignées, quantitativement important (SCALLY et al. 2012), et plus généralement les trajectoires individuelles des gènes. En effet, à l'échelle génomique, la spéciation semble survenir par "îlots de différenciation", c'est-à-dire que certaines régions génomiques se différencient précocément, tandis que d'autres restent sujettes aux flux de gènes (NOSIL et al. 2009 ; revu par WOLF et ELLEGREN 2017) ;
— des modèles de variation de taux de substitution à la fois par branche et par site, et traitant correctement les sites neutres. Ici nous avons utilisé un modèle de codon pour distinguer le dS et le dN, mais distinguer la 3^{e} position des codons pourrait être une approximation suffisante. C'est ce que nous avons appliqué avec Beast, et cela pourrait également être plus robuste à la stochasticité, car le modèle de codons utilisé ici, avec un $\omega = dN/dS$ par branche, requiert davantage de sites, un facteur limitant dans le cas des gènes individuels. Au sujet de cette question de la neutralité des sites, notons que la distribution des effets sur la valeur sélective des substitutions non-synonymes reste activement étudié (MOUTINHO et al. 2019 ; ROUSSELLE et al. 2019), et pourrait avoir des implications pour l'horloge moléculaire ;
— des modèles de mélange d'horloges relâchées comme dans LARTILLOT et al. (2016) et BLETSA et al. (2019) ;

— prendre en compte l'incertitude des calibrations fossiles, idéalement en employant plusieurs fossiles et un a priori de spéciation par naissance-mort, comme décrit par YANG et RANNALA (2006) et DOS REIS (2016), et mis en œuvre dans le programme `MCMCTree` (YANG 2007). En effet, ces auteurs ont observé que l'incertitude dans les calibrations fossibles, inévitable, cause des incertitudes proportionnelles à l'âge des nœuds.

11.2 Une inférence sophistiquée ne peut compenser les limites dans les données

Dans toutes nos régressions, les R^2 ajustés étaient faibles (0,47 pour la régression globale, et au minimum 0,104 pour la régression *Cercopithecinae*), traduisant une faible variance expliquée. Pour explorer la variance non expliquée, il faudrait incorporer des caractéristiques biologiques supplémentaires, mais il semble probable que la stochasticité intrinsèque des processus mutationnels et évolutifs soit la source principale de cette incertitude.

Ces limites inhérentes aux données persistent même avec une méthode statistiquement élaborée comme `Beast`. En effet nous n'avons pas réussi à produire des estimations moins dispersées, même si nous préconisons ce genre de programme pour ses nombreux atouts, notamment l'obtention d'intervalles de crédibilité à partir de la distribution postérieure. `Beast` a clairement produit des âges plus récents, mais compte tenu du désaccord de nos sources d'âges de référence, il est difficile de statuer sur leur justesse. Cependant, en prenant les données fossiles brutes, on observe que nos dates sont sous-estimées. Elle ne devraient en effet pas être plus jeunes que les fossiles attribués sans ambiguité aux descendants, hors flux de gènes minoritaire. Ainsi *HomoPan* devrait être plus ancien que *Sahelanthropus* (7,5 Ma), *Hominidae* que *Sivapithecus* (11,2 Ma), *Catarrhini* que *Kamoyapithecus* (25 Ma). De plus, il nous a paru en pratique numériquement impossible de dater les nœuds de spéciation individuellement, peut-être à cause du troncage des distributions a priori lorsque les points de calibrations sont trop nombreux.

Malgré l'attente qu'un modèle plus complet soit plus correct, non satisfaite par nos résultats avec `Beast`, il faut souligner que ce genre de méthode présente l'atout crucial de pouvoir correctement rendre compte de l'incertitude causée par des données insuffisantes. Le cadre bayésien se prête directement à cette analyse, en calculant la distribution a priori jointe —n'utilisant pas les données—, et en comparant la différence avec la distribution postérieure, qui si les données sont informatives, doit être plus précise que l'a priori. Un critère supplémentaire pour une bonne méthode de datation est de pouvoir transmettre les incertitudes entre chaque étape de l'analyse. Y parvenir est plus simple lorsque l'inférence est effectuée sur toutes les étapes simultanément, comme en estimant conjointement la topologie de l'arbre, les taux de substitution et de l'horloge, dans les limites de complexité imposées par l'identifiabilité des paramètres. D'après notre expérience, il semble difficile d'exécuter de tels calculs à l'échelle de la génomique comparative : premièrement, le MCMC nécessite de longs temps de calcul, et

des modèles trop complexes peuvent s'avérer insolubles en pratique ; deuxièmement, interpréter et valider les résultats d'un MCMC demande une inspection humaine approfondie et expérimentée. Face à cela, une base de donnée comme Ensembl Compara Vertebrata contient 197 espèces et 51 267 familles de gènes et devrait encore croître significativement dans un futur proche. Des algorithmes non paramétriques plus rapides sont donc un acteur pertinent dans ce domaine.

11.3 Guide de datation

Globalement, nos résultats fournissent des orientations pour appliquer une datation moleculaire précise et fiable par arbre de gènes. Ils soulignent la nécessité d'un nombre élevé de sites et de substitutions, essentiel pour disposer de suffisamment de puissance statistique. Cette limite est peut-être seulement contournable en incorporant de l'information complémentaire des séquences géniques, telles que le contexte génomique et les corrélations avec d'autres arbres de gènes, comme de récents développements le suggèrent (DUCHÊNE et al. 2019). Nos résultats peuvent aider à identifier les écueils d'une analyse par datation, à différentes étapes. De plus, notre contrôle empirique apporte des valeurs d'incertitude facilement comparables : un nouvel algorithme de datation pourrait être ajouté et comparé avec nos résultats actuels. Nous proposons donc quelques directives. Tout d'abord, l'alignement doit être produit de façon à minimiser les colonnes non-homologues, quitte à introduire trop de gaps. Une fois aligné, il est cependant possible que les séquences soient trop similaires, de sorte que trop peu de substitutions sont disponibles. Ces séquences évoluant trop lentement ne sont pas appropriées pour une datation. À l'inverse, la saturation des séquences rend les taux de substitution peu fiables et doit être évitée. Comme les familles multigéniques présentent différents taux de divergence, il faut donc une stratégie pour identifier les familles de taux intermédiaires, idéalement automatiquement. Avec des jeux de données à grande échelle, il paraît néanmoins encore nécessaire d'effectuer de nombreuses vérifications manuelles, pour repérer les alignements problématiques ou d'autres artefacts comme des *gene splits* non annotés et des erreurs de réconciliations. Pour analyser de nouveaux jeux de données, nos résultats sur les arbres primates peuvent servir de référence du moment que les caractéristiques des alignements et des taux de substitutions sont dans des intervalles similaires.

L'horloge moléculaire entre dans une époque où les données abondent et où les méthodes sont de plus en plus sophistiquées pour disséquer les mécanismes cachés de l'évolution des séquences. Les causes détaillées de variation du taux d'évolution moléculaire sont encore mystérieuses, tant à l'échelle des lignées qu'à l'échelle du génome. Les développements futurs permettant de répondre à cette question seront un défi technique mais devraient éclaircir d'importants processus évolutifs.

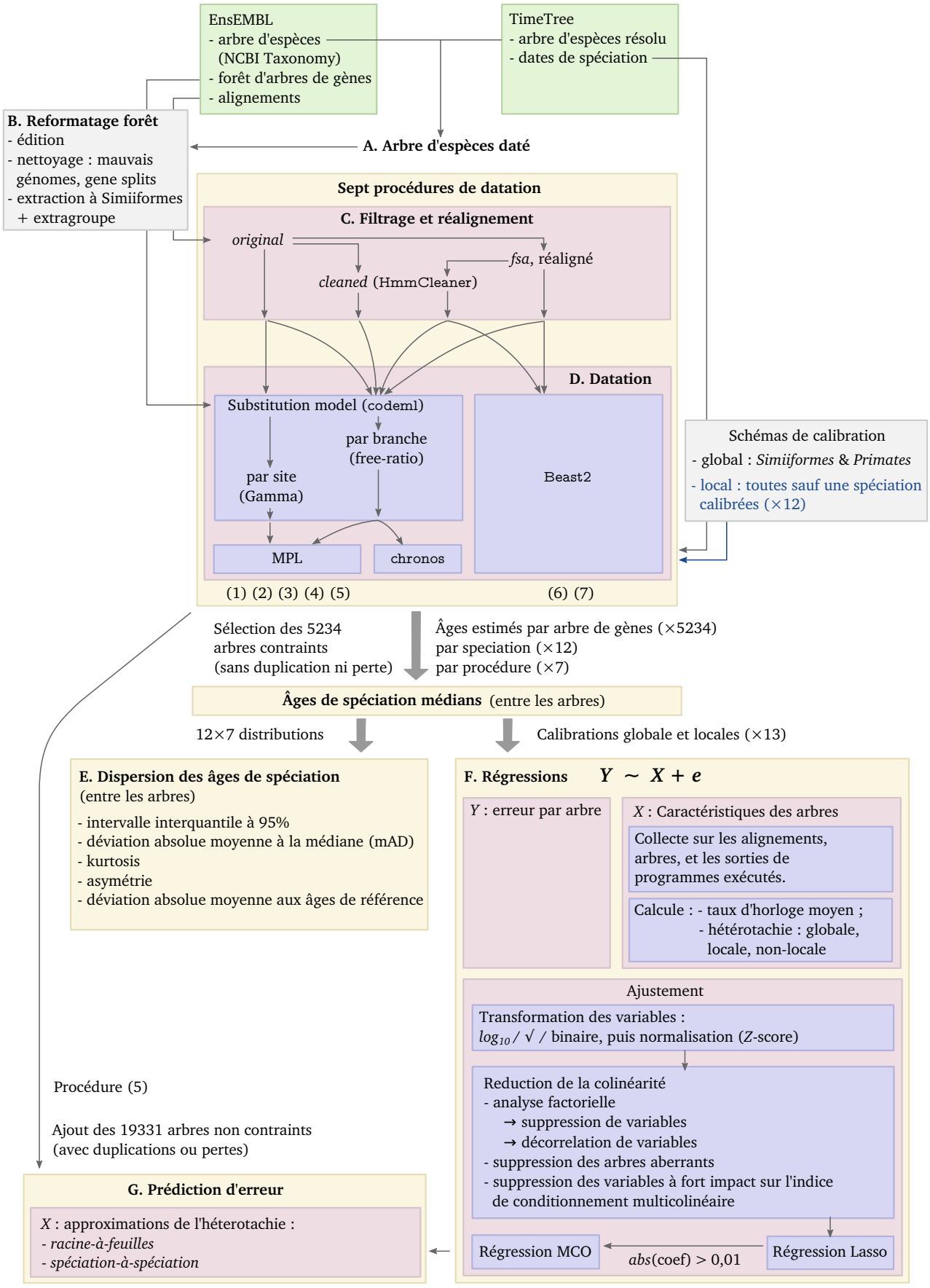


FIGURE I.8 – Analyse de la précision de datation, pipeline complet

12 Méthodes et données

12.1 Codes sources

Les développements méthodologiques de cette thèse ne font pas l'objet d'un package à la fonction spécifique. J'ai en revanche rassemblé tous mes outils à l'adresse `https://github.com/DyogenIBENS/Phylorgs` , avec la structure suivante :

— le module `dendro` pour les manipulations d'arbres (avec les packages `Ete3`, ainsi que `LibsDyogen`) ;
— `seqtools` pour manipuler les séquences (utilisant Biopython) ;
— `genchron` pour le projet de datation.
— `datasci` pour l'analyse de données et les graphes.
— `ensembltools` et `genomicustools` pour les utilitaires spécifiques à mes sources de données.

Ces scripts ont été écrits en Python 3.5 majoritairement, ainsi qu'en R (3.4.4) et en Bash (4.3).

12.2 Arbre d'espèces daté

L'arbre des 99 espèces d'Ensembl Compara Vertebrata version 93 (juillet 2018) (ZERBINO et al. 2018) est construit à partir de multiple sources (figure I.8, A).

La topologie provenant de la NCBI taxonomy est téléchargée sur Ensembl.

Une phylogénie datée est téléchargée depuis TimeTree (janvier 2019, KUMAR et al. 2017) [1]. Y sont insérées les espèces manquantes *Astyanax mexicanus*, *Ciona savignyi*, *Oreochromis niloticus* et *Poecilia formosa*, en reprenant les dates fournies par Ensembl.

L'arbre NCBI taxonomy contient des polytomies fréquentes, et lorsqu'il existe une résolution consensuelle, celles-ci sont résolues en se basant sur TimeTree ou des sources alternatives :

Polytomie	Nombre équivalent de dichotomies	Nouveaux nœuds	Source
Afrotheria	2	Paenungulata	
Caniformia	2	Arctoidea	
Cebidae	2	Cebidae_A	
Cricetidae	2	Cricetidae_A	
Euarchontoglire	2	laissé en polytomie (Toupayes incorrectement placés comme frère des *glires* dans TimeTree).	

1. TimeTree n'est malheureusement pas versionné, aux dépens de la reproducibilité. Les dates et la topologie obtenues en 2018 étaient différentes en 2019. On peut également leur reprocher de ne pas avoir d'API, de base de donnée ouverte, ou simplement une interface web "scrappable". Sans surprise, le logiciel MEGA, des mêmes auteurs, est sous licence fermée.

Polytomie	Nombre équivalent de dichotomies	Nouveaux nœuds	Source
Eutheria	2	Atlantogenata (Xenarthra+Afrotheria)	Upham et al 2019, Inferring the mammal tree, fig 2.
Homininae	2	HomoPan	
Macaca	2	laissé en polytomie	
Marsupialia	2	Australidelphia	
Muroidea	2	Eumuroida	
Cetartiodactyla	3	Artiofabula et Cetruminantia	
Hystricomorpha	3	Caviida + 1 polytomie	
Mus	3	Mus_B + 1 polytomie	
Rodentia	3	1 polytomie + MyodontaCastorimorpha	
Cercopithecinae	4	Papionini + tétratomie	Perelman et al 2011, A molecular phylogeny of living primates, ou Mittermeier et al, eds. 2013. Handbook of the Mammals of the World. Volume 3. Primates. p.951
Laurasiatheria	4	1 polytomie + Scrotifera + Fereuungulata	

Les sous-espèces, souches ou branches mâle/femelle sont fusionnées pour ne conserver que le rang espèce.

Cet arbre (annexe A1) sert ensuite à reformater la forêt d'arbres de gènes d'Ensembl.

12.2.1 Qualité des assemblages génomiques

Deux métriques permettent de mesurer la qualité des assemblages des génomes, en terme de continuité des scaffolds/chromosomes. Elles évaluent les tailles de scaffolds, en nombre de gènes :

— le *N50* est la taille maximale de scaffold tel que 50 % des gènes sont contenus dans des scaffolds plus grands. Formulé autrement, en considérant les gènes un à un, c'est la taille de scaffold médiane.
— le *K70* est une métrique approximativement inversement proportionnelle : c'est le *nombre* de scaffolds les plus longs contenant 70 % des gènes.

J'ai retiré les espèces au $N50 < 30$ ou $K70 > 400$. Pour les primates, cela retire 3 espèces parmi les 24 : *Rhinopithecus roxellana/bieti* et *Carlito syrichta* (tarsier) extragroupe immédiat de *Simiiformes*, ainsi que l'extragroupe immédiat *Tupaia belangeri* (toupaye).

12.2.2 Source alternative d'âges primates

En plus des âges de TimeTree, les estimations de DOS REIS et al. (2018) (stratégie de calibration A, horloge auto-corrélée) sont obtenues. Le nœud *Simiiformes* correspond à leur nœud *Anthropoidea*. L'arbre des primates d'Ensembl 93 avec les branchements des taxons non échantillonnés est en annexe A2.

12.3 Alignements multiples et arbres de familles de gènes

Les alignements multiples (format fasta) et les arbres phylogénétiques réconciliés (format NHX) sont téléchargés depuis Ensembl Compara Vertebrata 93 (juillet 2018).

Nous nous procurons uniquement les familles dont les gènes sont de type `protein_coding`, au total 23 904 arbres pouvant remonter jusqu'à *Opisthokonta* sont téléchargés via l'API d'Ensembl.

12.3.1 Filtrage et réalignement

Pour les procédures de datation (4) à (7), les alignements multiples ont été recalculés avec `FSA` (BRADLEY et al. 2009). Les paramètres par défaut ont été utilisés sur les alignements protéiques, ensuite rétrotraduits (avec le script personnel `seqtools/backtransX.py` prenant en compte les résidus ambigus [2]).

Au moyen d'`HmmCleaner` (DI FRANCO et al. 2019), les segments de séquences erronnés des alignements protéiques (résultant par exemple d'erreurs de séquençage) sont remplacés par des gaps. Cette transformation concerne les procédures de datation *cleaned* (3 et 5). Voir figure I.8, C. 31 alignements parmi les 5 235 produisent une erreur avec `HmmCleaner` et sont donc exclus dans la suite des procédures *cleaned*.

12.3.2 Construction des arbres par Ensembl

En bref, je résume ici les étapes qu'Ensembl applique pour construire les arbres réconciliés de familles de gènes :

1. un blast "tous contre tous" est exécuté sur les plus longues séquences codantes de chaque gène ;
2. à partir du graphe de proximité entre les gènes, basé sur les e-values, un clustering (`hcluster_sg2`) forme des groupes homologues ;
3. les séquences protéiques de ces groupes sont alignées par `M-coffee3` (produisant un consensus de plusieurs méthodes), ou `Mafft` lorsqu'ils sont trop grands ;
4. à partir des alignements rétrotraduits en ADN, `TreeBeST5` construit les arbres réconciliés avec la phylogénie d'espèces NCBI. En interne, `TreeBeST` utilise `PhyML` (GUINDON et al. 2010) pour calculer les vraisemblances et optimiser les topologies.

12.3.3 Reformatage de la forêt d'Ensembl

La forêt d'arbres de gènes doit subir les opérations suivantes (figure I.8, B).

12.3.3.1 Fusion des sous-espèces Les sous-espèces sont fusionnées (*Heterocephalus glaber* et *Cricetulus griseus*), en conservant en priorité l'une des sous-espèces si plusieurs sont présentes (*Heterocephalus glaber* femelle et *Cricetulus*

2. car l'outil `treebest backtrans` omet complètement les séquences dans ce cas.

griseus souche “CriGri”). Pour *Mus musculus*, seules la souche de référence est conservée.

12.3.3.2 Édition des réconciliations L'*édition* consiste à modifier a posteriori la topologie des arbres réconciliés, par exemple pour améliorer les prédictions d'orthologie ou de paralogie.

J'utilise ici la méthode d'édition de Genomicus (PERES et ROEST CROLLIUS 2015), basée sur un score de soutien des nœuds de duplications par les espèces descendantes, plus précisément la proportion d'espèces descendantes possédant la paralogie (*i.e.*, l'index de Jaccard des ensembles d'espèces de chaque côté de la duplication).

Comme je m'intéresse aux datations, l'information de longueur de branche m'est cependant cruciale, et l'édition pourrait avoir tendance à rapprocher en topologie des séquences pourtant très divergentes. J'applique donc ici une édition minimale (seuil de -1), dans le simple but de faire coincider l'arbre de gènes avec l'arbre d'espèces. Des incongruences surviennent en effet au niveau des polytomies de la NCBI taxonomy, où les arbres de gènes ont été résolus librement par TreeBest. Ils peuvent alors être incongruents avec notre arbre d'espèces plus résolu que la NCBI taxonomy, auquel cas le script réorganise les branches. 9 118 arbres sur 23 904 ont ainsi été reconstruits (nombres de nœuds par espèce en annexe A7).

12.3.3.3 Élaguage des espèces de basse qualité génomique Les espèces éliminées selon les critères de qualité d'assemblage (section 12.2.1) sont également retirées de la forêt de gènes, y compris les nœuds de spéciation rendus caduques. Les duplications dans les taxons éliminés sont réassignées à la branche descendante le cas échéant.

12.3.3.4 Nettoyage des branches aberrantes et des *gene splits* Un *gene split* est un nœud apparent de duplication, dont les descendants sont en fait les fragments d'un seul et même gène. Une partie est annotée dans la base de données d'Ensembl, et j'ai donc fusionné les descendants de ces nœuds en un seul. Dans les alignements correspondants, j'ai remplacé les *split genes* par la séquence fusionnée.

De plus, la distribution des longueurs des 3 418 322 branches de la forêt permet d'identifier un seuil de longueur aberrante : au-delà de 10 000 substitutions/site, on considère qu'il s'agit d'une erreur de `TreeBeST`, qui code en dur ces très grands nombres pour représenter l'infini. On détache alors le sous-arbre entier (annexe A3).

Au total, 6 962 *gene splits* sont fusionnés et 729 branches aberrantes (soutenant 6 359 feuilles) sont détachées, affectant 3 343 arbres.

12.3.4 Sous-arbres *Simiiformes*

Pour ce chapitre, les sous-arbres de gènes enracinés à *Simiiformes* sont extraits, en incluant obligatoirement deux séquences extragroupe. Pour cela, il est au préalable nécessaire de réinsérer les nœuds de spéciations implicites dans la forêt d'arbres. Ensuite, pour rajouter l'extragroupe, exactement deux séquences sont conservées, les deux dont les longueurs de branche sont les plus courtes si le clade externe en contient davantage. Notons qu'il est possible que l'extragroupe comprenne des paralogues *Simiiformes*, dans le cas où le nœud parent du gène *Simiiformes* est une duplication *Primates* ou antérieure. En résultent 24 562 sous-arbres.

12.4 Taux de substitution synonymes et non-synonymes

Nos datations (figure I.8, D) nécessitent des longueurs de branches en nombre de substitutions synonymes par site (dS). Pour les estimer, j'utilise `codeml` de la suite `PAML 4` (YANG 2007). `codeml` permet de mesurer le ratio de substitutions non-synonymes sur synonymes, $\omega = dN/dS$, optimisé par maximum de vraisemblance. En particulier, j'utilise les modèles suivants :

— "free-ratio" (paramètre `model=1`) : le ratio ω est indépendant entre les branches (et constant entre sites de l'alignements) ;
— ω suivant une loi Gamma entre sites (paramètres `model=0`, `NSsites=5`, `ncatG` laissé au défaut (10)).

Nous traitons les gaps comme des caractères ambigus (`cleandata = 0`). Un fichier de contrôle typique pour le modèle "free-ratio" est inclus en annexe A8.

12.5 Pipeline de préparation des données pour codeml

Les 5 235 calculs de `codeml` sont soumis à un cluster de calcul administré par `HTCondor`. Pour cela il est nécessaire de préparer les arbres, les alignements, et les fichiers de contrôle. J'automatise ces tâches via un pipeline `snakemake`, qui extrait les séquences de l'alignement total, reformate correctement les fichiers, écrit les fichiers de contrôle et le fichier de soumission sur le cluster. Ce pipeline fait appel aux outils que j'ai développés ainsi qu'aux Newick Utilities (JUNIER et ZDOBNOV 2010).

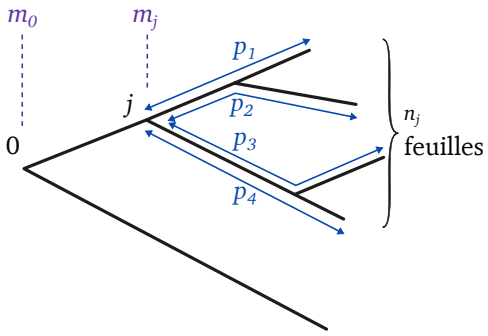
12.6 Datation

12.6.1 Mean-Path-Length

L'algorithme de datation est appliqué sur l'arbre de gène, dont les longueurs de branches sont en nombre de substitutions par site, c'est-à-dire pas forcément proportionnelles au temps.

Le Mean-Path-Length, ou MPL (BRITTON et al. 2002), permet de produire un arbre ultramétrique, c'est-à-dire dont toutes les feuilles sont à égale distance de la racine.

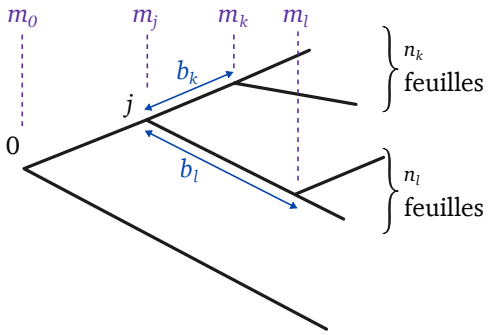
L'algorithme assigne à chaque nœud j rencontré la valeur m_j, la moyenne des longueurs chemins jusqu'aux n feuilles :



$$m_j = \frac{1}{n_j} \sum_{i=1}^{n_j} p_i$$

où p_i représente la longueur de chemin du nœud jusqu'à la feuille i, et n_j le nombre de feuilles descendant du nœud j.

En parcourant l'arbre des feuilles vers la racine, on peut profiter de la relation de récurrence suivante pour réutiliser les valeurs des nœuds enfants, par programmation dynamique. Dans le cas de deux enfants k et l du nœud j, situés sur des branches de longueur b_k et b_l respectivement :



$$m_j = \frac{1}{n_j}(n_k(b_k + m_k) + n_l(b_l + m_l))$$

Cette moyenne des distances aux nœuds enfants est conceptuellement équivalente au regroupement par UPGMA (Unweighted Pair Grouping Method with Arithmetic mean).

Pour convertir ces longueurs en âge, des points de calibration sont utilisés. Par exemple, si l'âge de la racine a_r est connu, on peut en déduire l'âge de n'importe quel nœud j :

$$a_j = a_r \times \frac{m_j}{m_r}$$

Comme cet algorithme peut éventuellement produire des incohérences d'âges (p.ex. un nœud parent plus jeune que son enfant), ces impossibilités sont rectifiées en forçant l'âge parent à valoir au moins celui de l'enfant (produisant alors une branche nulle).

Ce calcul est appliqué dans les procédures de datation (1) à (5), à partir des longueurs de branches en dS.

Comme décrit dans l'article original, il est possible de calculer la probabilité des longueurs de branches, suivant une loi Poisson, et d'obtenir une p-valeur pour l'hypothèse que les taux sont égaux entre sous-arbres enfants. Ce test fait cependant une supposition de normalité nécessitant au moins une trentaine de substitutions. Le MPL, non-paramétrique, peut être vu comme une façon de lisser les taux de proche en proche, c'est-à-dire d'ajuster une horloge à taux auto-corrélés (SANDERSON 2002).

12.6.1.1 Adaptation du MPL avec des points de calibration supplémentaires Dans une seconde publication, BRITTON (2007, appendix 1) décrivent comment incorporer des points de calibration supplémentaires, ce qui est implémenté dans leur programme `PATHd8`. Là encore, il s'agit d'un lissage non paramétrique. J'ai également introduit cette modification dans ma version, car elle permet de dater la situation suivante (figure I.9), occurant en section I.9 : lorsqu'une duplication est suivie de deux spéciations différentes, c'est-à-dire une duplication suivie d'une perte d'un des paralogues, faisant disparaître le nœud de spéciation attendu.

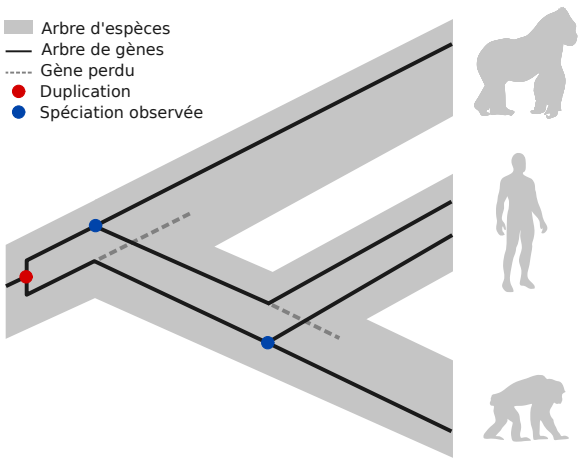
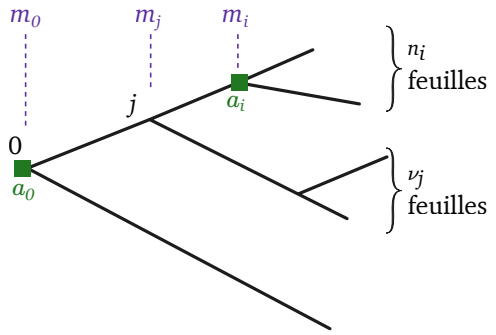


FIGURE I.9 – Duplication suivie de deux spéciations différentes (points bleus). Seuls ces nœuds de spéciation observés peuvent servir de point de calibration, or il faut dans ce cas adapter le MPL qui est initialement conçu pour calculer les âges depuis des feuilles du même âge (âge 0).

Je présente d'abord la formule originale de BRITTON et al. J'ai cependant préféré ne pas l'utiliser car elle pose un problème d'inhomogénéité des unités, et j'ai implémenté une variante homogène.

Considérons le nœud j dont nous estimons l'âge a_j. Le nœud ancêtre calibré le plus proche est noté 0, d'âge connu a_0. Notons $a_1, \ldots, a_k$ les âges des k nœuds calibrés descendant de j, et $n_1, \ldots, n_k$ les nombres de feuilles respectifs qu'ils supportent. Notons également ν_j le nombre de feuille descendant de j sans passer par aucun nœud calibré (de sorte que le nombre total de feuilles descendant de j vaut $\nu_j + n_1 + \ldots + n_k$).

Dans BRITTON et al, l'âge de j est calculé ainsi :



$$a_j = (\nu_j a_0 \frac{m_j}{m_0} + s_1 + \ldots + s_k)/(\nu_j + n_1 + \ldots + n_k)$$

où

$$s_i = n_i[a_i + (a_0 - a_i)(m_j - m_i)]/(m_0 - m_i)$$

pour tout i entier de 1 à k.

Autrement dit, chaque ensemble de chemins passant par un point de calibration est considéré simultanément, puis la différence d'âge entre les calibrations avant et après, $a_0 - a_i$ doit être ajustée pour représenter la fraction de temps s'étant écoulée uniquement entre le nœud j et le nœud i. Selon cette interprétation de l'algorithme, on voit clairement qu'au lieu de la formule ci-dessus, $(a_0 - a_i)$ aurait du être pondérée par la fraction de longueurs de chemin $(m_j - m_i)/(m_0 - m_i)$. Ça n'est pas le cas, et cela introduit une inhomogénéité : en effet, a_i, en Ma, est additionné à $(a_0 - a_i)(m_j - m_i)$, en Ma×subst, et c'est seulement ensuite que la division par $(m_0 - m_i)$ est appliquée.

Je concède que cette formule n'est de toute façon qu'une heuristique, sans modèle sous-jacent, ayant pour but de produire une approximation, et qu'en tant que telle son inhomogénéité pourrait être tolérée. Cependant ayant dû l'implémenter, j'en ai profité pour l'ajuster selon ma conception, redéfinissant le terme s_i :

$$s'_i = n_i[a_i + (a_0 - a_i)\frac{(m_j - m_0)}{m_0 - m_i}]$$

12.6.2 Horloges relâchées

Alternativement, le calcul des âges et des taux peut se faire par maximum de vraisemblance ou par méthode bayésienne, en autorisant un taux non constant.

12.6.2.1 Vraisemblance pénalisée J'ai calculé les âges avec la méthode de vraisemblance pénalisée (SANDERSON 2002) : les taux peuvent varier d'une branche à l'autre, mais une pénalité est ajoutée sur la variance des taux entre branches consécutives. De cette façon, une horloge auto-corrélée est ajustée.

Comme le MPL, cette méthode requiert en entrée un arbre de gènes avec longueurs de branches en substitutions. Ici nous l'appliquons sur les valeurs de dS obtenues par `codeml`.

J'ai utilisé la fonction `chronos` du package `APE` version 5.3 (PARADIS 2013), avec différentes paramétrisations :

— horloge constante : modèle "discrete", `nb.rate.cat=1` ;
— modèle "correlated" (défaut), paramètre lambda (poids de la pénalisation) à 1, puis 100, puis 100000 ;

— modèle “relaxed” (décorrélé suivant loi Gamma), lambda de 1, 100 et 100000.

La version de `chronos` utilisée a été modifiée de façon à autoriser des points de calibration plus nombreux (l'initialisation d'origine produisait en effet des situations impossibles).

Malheureusement un artefact d'optimisation produit des résultats dont l'évaluation par la mAD serait inappropriée, puisque les nœuds très incertains finissent positionnés sur le milieu exact entre les spéciations avoisinantes, créant un pic précis au milieu, donc une faible mAD malgré une totale incertitude (annexe A5).

Ce pic résulte apparemment d'une absence d'optimisation des âges initialisés, à cause d'une vraisemblance à la surface très plate : trop peu d'information dans les taux rendent la vraisemblance optimale quel que soit le point de départ choisi. On peut imaginer des contournements à ce problème : il est possible d'augmenter le poids de la pénalisation, λ, ou alternativement d'augmenter la précision du critère de différence de vraisemblance qui marque l'arrêt de l'optimisation. La première solution est peu pertinente car le modèle devient alors équivalent à un modèle à taux constant. La seconde solution n'a pas permis de supprimer le pic. Une troisième alternative serait de choisir un poids λ particulier à chaque arbre, choisi par validation croisée comme suggéré par SANDERSON (2002), mais il semble qu'une pénalité par arbre conduirait à un sur-ajustement non informatif.

La modification de l'initialisation, et d'autres d'ordre diagnostique liées à ces difficultés d'optimisation, sont maintenant incorporées dans `APE` version 5.4.

12.6.2.2 MCMC `Beast 2` (BOUCKAERT et al. 2019) permet d'inférer des taux décorrélés de façon entièrement paramétrique, par MCMC. En entrée, le programme prend les alignements, de sorte que les substitutions sont estimées simultanément, et réestime les arbres. Le modèle suivant a été appliqué sur chaque alignement :

— matrice de substitution GTR-Γ;
— partitionnement des taux de substitution par site entre les positions 1,2 et 3 des codons ;
— un prior de Yule sur le processus de branchement générant l'arbre ;
— une horloge décorrélée entre branches suivant une loi log-normale, dont la moyenne et l'écart-type sont estimés ;
— les calibrations *Primates* et *Simiiformes* avec les probabilités a priori suivantes, de façon a refléter la moyenne et l'intervalle fournis par TimeTree :
  – *Primates*: loi Gamma de paramètres $\alpha = 4{,}6$, $\beta = 0{,}656$ et localisation 70,8 Ma ;
  – *Simiiformes* : loi Gamma de paramètres $\alpha = 4$, $\beta = 0{,}575$ et localisation 40,9 Ma ;
— la monophylie du clade *Simiiformes* est imposée.

J'ai exécuté les MCMC sur 50 000 000 itérations, afin d'obtenir des ESS (Effective Sample Sizes) supérieures à 200. Les dates résultantes sont finalement annotées

sur la topologie d'arbre imposée (celle d'Ensembl) avec l'outil `TreeAnnotator` fourni par `Beast` 2.

12.7 Mesures de dispersion

Les estimations d'âge de chaque procédure sont comparées entre elles par plusieurs mesures de dispersion (figure I.8, E), dont l'intervalle entre les percentiles 2,5 % et 97,5 %, nommé intervalle interquantile à 95 % et également la moyenne de déviation absolue à la médiane (*mean absolute deviation from the median*, ou mAD). La mAD est une mesure robuste, car elle est moins influencée par l'asymétrie (skew) de la distribution que les mesures prenant la moyenne pour centre, et moins influencée par les valeurs extrêmes que l'écart-type qui fait intervenir les déviations au carré. Elle est cependant plus sensible que la *médiane* de la déviation absolue à la médiane et que l'écart interquartile. De plus, il s'agit de la statistique utilisée par le test Brown-Forsythe d'égalité de variance, ce qui nous autorise à tester la significativité de la différence de dispersion entre les procédures.

12.8 Caractéristiques des arbres de gènes

Pour tester les facteurs corrélés avec l'erreur, de nombreuses caractéristiques des arbres sont mesurées, puis régressées contre l'erreur des 5 235 arbres (figure I.8, F).

12.8.1 Caractéristiques de l'alignement

22 caractéristiques sont mesurées, sur les séquences intragroupe des alignements, parmi lesquelles :

- statistiques globales :
  - la longueur de l'alignement (i.e., nombre de sites alignés) ;
  - la proportion de paires de séquences non alignables ;
- statistiques par séquences :
  - fréquence moyenne de gaps, et son écart-type ;
  - le pourcentage de nucléotides ambigus moyen, et son écart-type ;
  - le pourcentage de GC moyen des séquences, et son écart-type ;
  - le pourcentage de CpG moyen, et son écart-type ;
- statistiques moyennées entre sites :
  - entropie moyenne de composition nucléotidique, et son écart-type ;
  - entropie moyenne de composition en codon, et son écart-type ;
  - en utilisant l'arbre, le score de parcimonie moyen et son écart-type sont calculés, pour les nucléotides et pour les codons.

12.8.2 Caractéristiques de l'arbre : arbres "contraints" comme contrôle

Dans ce chapitre, les arbres choisis servent de donnée contrôle pour mesurer l'erreur de datation. Afin de sélectionner des arbres de bonne qualité, et afin de

s'assurer que toutes les branches sont comparables, nous conservons seulement les arbres sans duplication ni perte, soit 5 235 sur les 24 565.

Les caractéristiques suivantes sont également mesurées :

— le nombre de nœuds reconstruits pour correspondre à l'arbre d'espèces (cf. 12.3.3.2) ;
— la valeur de bootstrap minimale et moyenne des nœuds, annotée par Ensembl ;

12.8.3 Caractéristiques des taux de substitution

En plus des scores de parcimonie calculés plus haut, `codeml` permet de rassembler les mesures suivantes sur les taux de substitution :

— proportions de *sites* synonymes et non-synonymes théoriques dans l'alignement ;
— $kappa$, le ratio transition/transversion ;
— longueur de l'arbre en dS, en dN et en t (taux de substitution total par codon) ;
— ω moyen, médian, écart-type et asymétrie par branche ;
— écart-type des longueurs de chemin *racine-à-feuilles*, en dS, dN et t.
— fraction de nœuds connectant des branches nulles consécutives, sœurs, ou en triplet, pour le dS, dN et t.

12.8.4 Caractéristiques des calculs effectués

— `codeml` :
  – score de maximum de parcimonie ;
  – warning de convergence ;
  – log-vraisemblance ;
  – temps de calcul ;
— `beast` (non régressées ici):
  – probabilités postérieures, vraisemblance, vraisemblance des positions de codon {1,2}, et 3, taux de substitutions à ces positions, taux de l'horloge moyen, forme de la loi Gamma du taux de l'horloge.

12.8.5 Caractéristiques de nettoyage des alignements

— par `Gblocks` [3] (Castresana 2000) :
  – nombre de blocs retenus ;
  – pourcentage de sites retenus ;
— par `HmmCleaner` :
  – proportion de séquences nettoyées ;
  – pourcentage moyen de séquence nettoyé ;
  – pourcentage maximal de séquence nettoyé.

3. Les résultats de la procédure utilisant Gblocks ne sont pas présentés ici, mais Gblocks améliore moins la précision qu'HmmCleaner.

12.8.6 Caractéristiques des taux d'horloge d'après les âges de référence

Dans ce chapitre, l'utilisation d'arbres avec uniquement des nœuds de spéciation datés par des références externes permet de mesurer les taux de l'horloge moléculaire (en subst/site/Ma). Pour les substitutions en dS, nous mesurons deux caractéristiques qui s'avèrent cruciales dans les résultats de la régression :

— le taux moyen $\bar{r}_{arbre}$: c'est le ratio de la somme de la longueur totale de l'arbre en dS (subst/site) sur sa longueur totale en Ma :

$$\bar{r}_{arbre} = \frac{\sum d_i}{\sum b_i}$$

où d_i représente le nombre de substitutions par site, et b_i les millions d'années, de la branche i. Cette estimation correspond au taux du maximum de vraisemblance d'après un modèle de Poisson à taux constant.

— l'hétérogénéité du taux, ou hétérotachie $\bar{\sigma}^2_{arbre}$: c'est l'écart des taux par branche, pondéré par la longueur de branche en Ma :

$$\bar{\sigma}^2_{arbre} = \frac{\sum b_i (\frac{d_i}{b_i} - \bar{r}_{arbre})^2}{\sum_i b_i}$$

Cette pondération a été choisie pour produire des estimations moins variables, ce qui est attendu des branches longues, et parce que les régressions obtenues avec cette mesure avait un R^2 légèrement meilleur.

12.9 Régression linéaire multiple

La variable dépendante, ou expliquée, est ici l'erreur par arbre. Les variables indépendantes, ou explicatives, sont les caractéristiques des arbres énoncées ci-dessus.

L'ensemble des analyses ci-après ont été effectuées dans des notebooks Jupyter en Python 3.5 et des paquets de calculs scientifiques, sous les versions suivantes :

— Python 3.5.2 (default, Nov 23 2017, 16:37:01)
— GCC 5.4.0 20160609
— `IPython 7.4.0`
— `jupyter 1.0.0`
— `numpy 1.17.3`
— `scipy 1.2.1`
— `pandas 0.24.2`
— `statsmodels 0.10.1`
— `sklearn 0.21.3`
— `matplotlib 3.0.3`
— `seaborn 0.9.0`
— `Biopython 1.73`

— `Ete3` `3.1.1` (HUERTA-CEPAS et al. 2016).

12.9.1 Transformation et normalisation des variables

Dans un premier lieu, l'application d'une régression linéaire nécessite des variables distribuées sans biais extrême. Pour cela, une procédure semi-automatisée a été utilisée, où les variables sont transformées par une fonction réduisant l'asymétrie (skew) le plus possible. Les transformations appliquées sont la racine carrée et le logarithme base 10, éventuellement avec un incrément ou une inversion de signe si nécessaire. Par ailleurs, certaines variables sont encodées binairement, manuellement, dans les cas où une grande proportion de valeurs est constante ou lorsque la distribution est visiblement bimodale.

12.9.2 Réduction de la dimensionalité

La multicolinéarité entre variables explicatives est un écueil connu à éviter lorsqu'on utilise les moindres carrés généralisés. Pour la réduire en amont de la régression, nous appliquons deux stratégies :

1. D'abord nous calculons la covariance par analyse factorielle (analogue à l'analyse en composantes principales ou ACP, mais pour les variables discrètes ou ordinales). Sur cette base, nous supprimons les caractéristiques appartenant à des groupes de variables très corrélées. Additionnellement, nous appliquons des "décorrélations" de paires de variables, par exemple en divisant l'une par l'autre, ou en obtenant les résidus de la régression simple de ces deux variables. Nous avons ainsi décorrélé principalement des mesures d'écarts-types étant fortement corrélés à leur moyenne.
2. En deuxième temps, nous vérifions le *conditionnement* indiquant la multicolinéarité, calculé comme la racine carrée de la plus haute valeur propre de X^tX, où X est la matrice des variables explicatives. Nous supprimons itérativement les variables augmentant le plus ce conditionnement, jusqu'à atteindre une valeur inférieur à 20.

12.9.3 Suppression d'arbres à caractéristiques éliminatoires

Les arbres possédant des paires de branches consécutives ou sœurs de longueur nulle ont été supprimés en amont de la régression, ainsi que les alignements présentant des paires de séquences non-chevauchantes, i.e., non alignables.

Leur impact sur l'erreur est testé individuellement par des tests de Students unilatéraux, corrigés par méthode de Benjamini-Hochberg.

Ce filtrage a priori réduit le nombre d'arbres de 5 235 à 1 436.

12.9.4 Régression Lasso

La régression Lasso (TIBSHIRANI 1996) est une technique d'optimisation de régression linéaire permettant de simultanément sélectionner les variables à partir d'un jeu colinéaire. Nous l'appliquons en première passe, afin de supprimer les

variables produisant un coefficient absolu inférieur à 0,01. Le paramètre de pénalité appliqué (`alpha`, dont l'effet dépend des données) est de 0,01. Comme les coefficients Lasso sont par construction biaisés, il est cependant compliqué d'obtenir des p-valeurs et des R^2. Pour cette raison, nous ajustons ensuite la régression par moindre carrés ordinaires, uniquement sur les variables sélectionnées par Lasso. La covariance est paramétrée selon la technique de robustesse à l'hétéroscédasticité de MacKinnon et White (1985).

12.10 Indices d'équilibre des sous-arbres

Pour mesurer le déséquilibre des sous-arbres dans la phylogénie *Simiiformes*, l'indice de Sackin (1972) a été calculé pour chacune des douze spéciations. Brièvement, cet indice peut s'exprimer comme la somme du nombre de nœuds ancêtres de chaque feuille.

D'autres indices de déséquilibre (Colless, Mir, voir Blum et François 2005 ; Mir et al. 2013) se prêtent moins à notre phylogénie à cause des polytomies.

Chapitre II

Datation et dynamique des duplications de gènes

> "Mistakes" is the word you're too embarrassed to use. You ought not to be. You're a product of a trillion of them.
>
> *Westworld*, Saison 1 épisode 1.

Les duplications de gènes figurent parmi les erreurs de transmission du matériel génétique les plus importantes en quantité de sites affectés. L'erreur initiale, une mutation portée par un individu, soumise à la dérive et la sélection, peut se fixer dans la population. Il s'agit d'un processus stochastique dont nous allons estimer les caractéristiques chez quelques espèces primates. Dans ce chapitre, j'utilise les résultats précédents pour circonscrire les datations aux arbres *Simiiformes* les plus fiables. J'analyse ensuite les taux sous-jacents, et les modalités de leur variation entre familles de gènes et entre lignées.

13 Prédiction de l'erreur dans les arbres avec duplications et pertes

Premièrement, comme dans le précédent chapitre, j'estime les facteurs corrélés à l'erreur de datation des 5 204 arbres contraints (sans duplication ni perte). Ces arbres contrôle servent ici à établir la relation linéaire entre erreurs et caractéristique des arbres, par régression. L'estimation diffère cependant du précédent chapitre car les valeurs d'hétérotachie sont approximées (cf. Méthodes, 18.1), afin d'être comparables entre arbres contraints et arbres avec duplications. Deuxièmement, grâce à la relation linéaire ainsi ajustée, je prédis l'erreur dans les arbres test, ceux qui contiennent des duplications ou des pertes.

13.1 Précision dans la forêt complète des 24 565 arbres

L'hétérotachie doit d'abord être mesurée de façon comparables entre les arbres contrôle (contraints) et les arbres de test (avec duplications/pertes). Or la mesure effectuée au chapitre I nécessite une information disponible uniquement pour les arbres contrôles —les âges de référence des nœuds, qui sont tous des spéciations—, et n'est donc pas utilisable telle quelle lorsqu'il y a des duplications. Pour que la régression soit ajustée avec la même variable d'hétérotachie que celle qui servira à la prédiction, il faut donc la mesurer de façon "approximée" dans les deux jeux de données (contrôle et test). J'ai donc approximé de deux façons cette hétérotachie, premièrement en lissant les taux de spéciation à spéciation, deuxièmement en lissant les taux de la racine aux feuilles (voir 18.1.1).

Dans les arbres contraints, l'hétérotachie calculée avec l'approximation *spéciation-spéciation* est identique à celle de la section I.8, mais la régression est ici calculée sur tous les arbres contraints, y compris ceux qui présentent des propriétés indésirables (séquences alignées non chevauchantes, absence de substitution sur certaines branches, cf. Méthodes 12.9.3). Dans cette régression ($R^2 = 0{,}54$), l'hétérotachie est le plus fort prédicteur, suivie de la longueur d'alignement et du taux de dS moyen. Relevons l'étendue de l'erreur ainsi modélisée, afin de pouvoir comparer avec la prédiction : le 9^e décile d'erreur modélisée est à 3,20 Ma/nœud, autrement dit la droite de régression dépasse ce seuil d'ordonnée pour 10 % des arbres. Qu'en est-il de l'erreur prédite sur les arbres test ?

On constate que l'erreur prédite sur les nouveaux arbres est globalement plus élevée (figure II.1a, en vert, échelle log). Cette augmentation est dûe comme attendu aux deux caractéristiques les plus influentes (figure II.1a, à droite), pour lesquelles les arbres test ont des hétérogénéités plus élevées, et des alignements plus courts. En revanche, le taux de dS moyen, prédisant une baisse de l'erreur, semble être souvent plus élevé dans les arbres test, à l'exception de quelques valeurs extrêmement faibles, qui pèsent elles pour une augmentation de l'erreur. Toutes variables combinées, le résultat de la prédiction est donc une augmentation moyenne de l'erreur des arbres avec duplications et pertes par rapport aux arbres contraints, tout en restant dans la même gamme d'erreur. D'après cette prédiction, les arbres non contraints auront pour les deux tiers une erreur inférieure à 3,20 Ma/nœud, ce qui confère une confiance raisonnable dans leur datation. En revanche, du fait de la sous-estimation de l'hétérotachie dans les arbres test, mais pas dans les arbres contrôle, il s'agit d'une prédiction d'erreur *minimale*.

En conséquence, nous souhaitons alors prédire l'erreur *maximale*, c'est-à-dire de façon plus conservative. Pour cela, l'approximation *racine-feuilles* assure que l'hétérotachie sera sous-estimée aussi bien dans le jeu contrôle que dans le jeu test. Cette seconde régression sur les 5 204 arbres contrôle produit cette fois un R^2 ajusté nettement plus faible (0,43). En outre, les caractéristiques les plus prédictives de l'erreur de datation deviennent 1) la longueur d'alignement, 2) le taux de dS moyen et 3) l'hétérotachie du dS. Pour résumer l'effet de cette nouvelle paramétrisation, cela revient à donner moins de pouvoir prédictif à l'hétérotachie en lissant ses valeurs ; c'est conçu pour pallier au fait que l'hétérotachie

mesurée dans les arbres non-contraints sera nécessairement sous-estimée, donc l'erreur également. Avec pour but de moins sous-estimer l'erreur prédite due à l'approximation sur l'hétérotachie, on s'attend à trouver une erreur d'autant plus accrue dans les arbres tests. En résultat, on obtient comme prévu davantage d'arbres prédits au-dessus du 9e décile d'erreur (figure II.1b). Cependant, cette augmentation est limitée en proportion d'arbres : en figure II.1b, graphique de gauche, sur 14 250 arbres test, 5 016 dépassent le 9e décile, soit 35 %, alors que pour l'approximation spéciation-spéciation (figure II.1a) 4 117 sur 14 155 dépassent, soit 29 %. Certes limitée en proportion d'arbres, cette augmentation d'erreur demeure importante en échelle, comme le révèle la compression de l'axe des abscisses entre la figure II.1a et II.1b : une fraction d'arbres dépasse les 20 Ma/nœud jusqu'à atteindre même 50 Ma/nœud. Ici, l'explication provient majoritairement de la longueur des alignements, et probablement des valeurs extrêmes inférieures du taux de dS, et supérieures de l'hétérotachie du dS.

Finalement, si l'estimation racine-feuille exclut davantage d'arbres au 9e décile, il s'agit en majeure partie (3 615 sur 5 536) d'arbres qui seront également exclus avec l'approximation spéciation-spéciation.

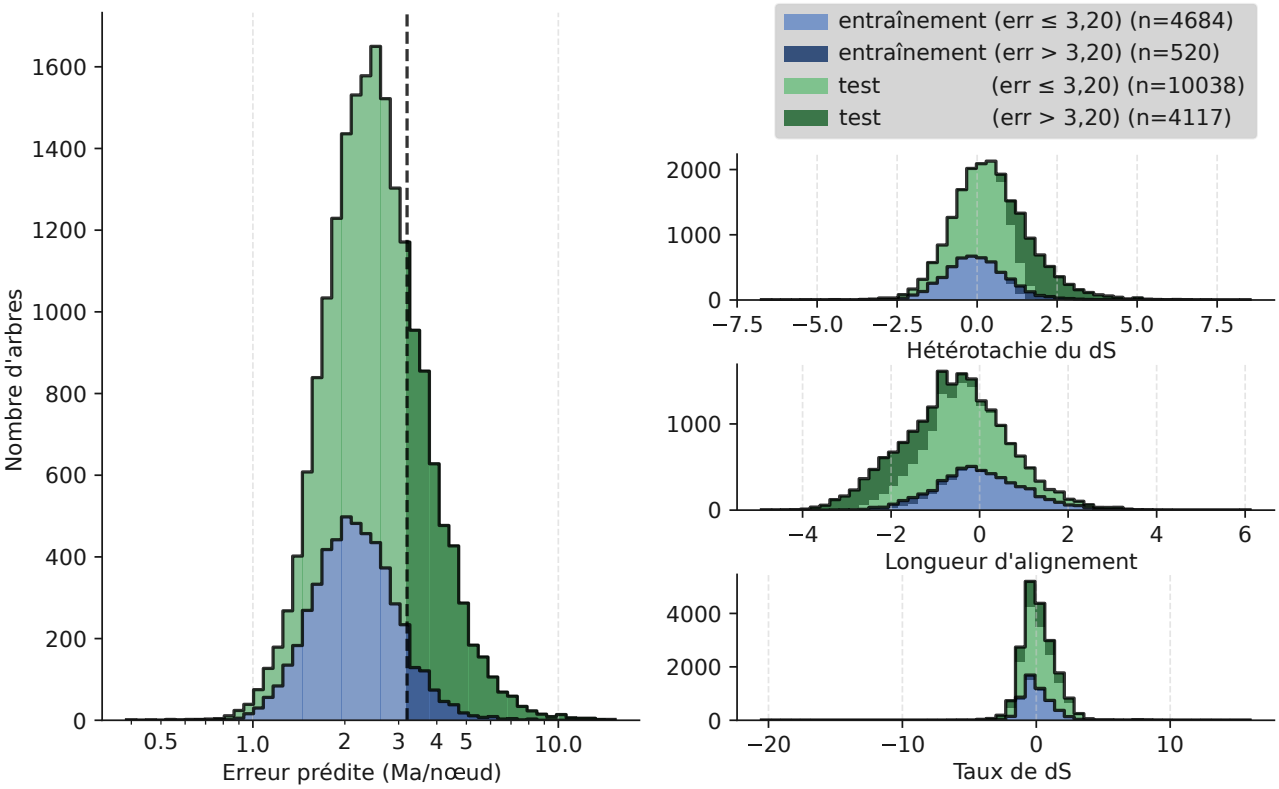
13.2 Suppression des arbres aux propriétés non-désirables

Dans le chapitre I, certains arbres sont filtrés *en amont* selon certaines caractéristiques binaires considérées comme augmentant l'erreur *a priori* (séquences alignées non chevauchantes, absence de substitution sur certaines branches). Elles entraînent une réduction importante du nombre d'arbres utilisés dans la régression contrôle, en passant de 5 204 à 1 436, mais on peut ainsi se concentrer sur d'autres caractéristiques moins "évidentes" et plus intéressantes évolutivement.

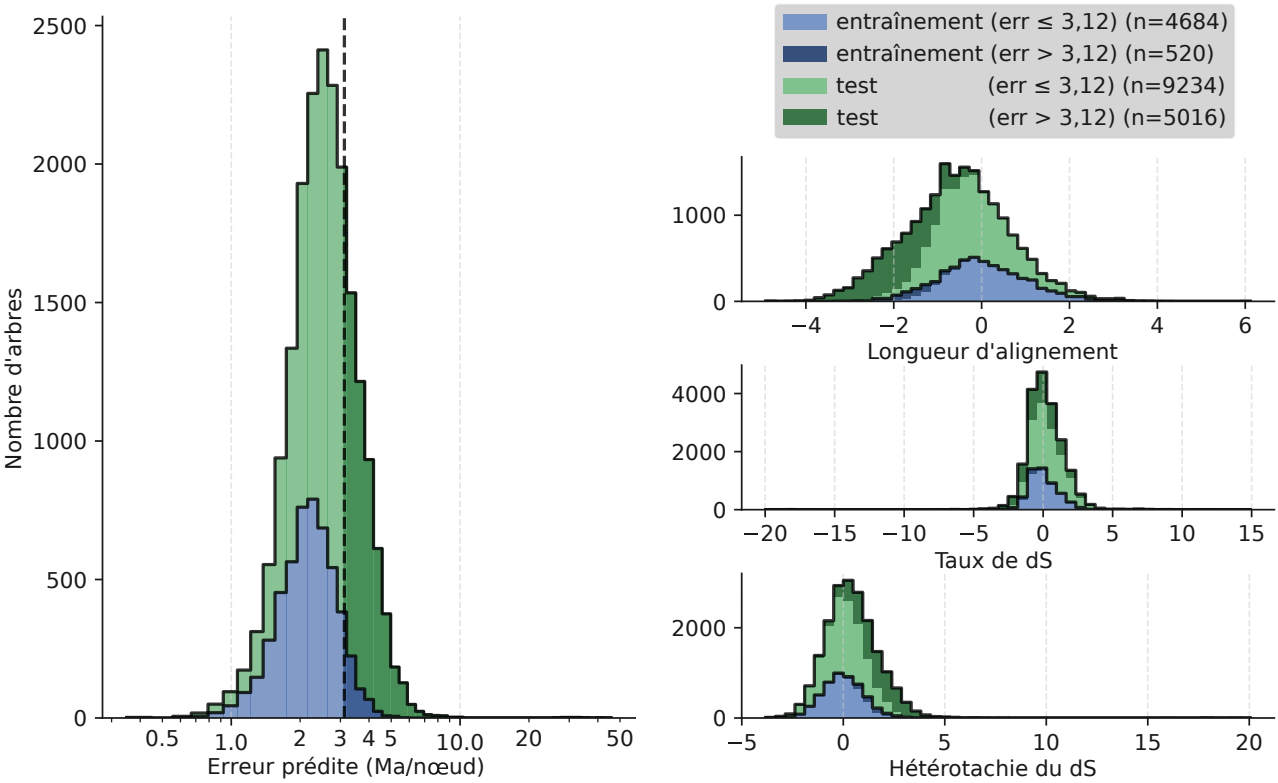
En conséquence, le jeu de test est réduit en amont de la même manière, passant de 14 250 à 2 774. Cette réduction encore plus drastique traduit le fait que les arbres avec duplications et pertes contiennent pour la grande majorité au moins une branche sans substitution, ce qui influencera négativement la datation.

Avec l'approximation *spéciation-spéciation*, la régression produit le même résultat qu'en section I.8, car l'hétérotachie ainsi calculée est identique dans les arbres contraints (R^2 ajusté de 0,47). Malgré un nombre d'arbres réduit par le filtrage en amont, avec l'approximation *spéciation-spéciation*, l'erreur prédite est toujours supérieure au jeu contrôle (figure II.2a), encore une fois à cause de l'hétérotachie et de la longueur d'alignement, tandis que le %GC moyen a une influence plus faible. Cette estimation d'erreur est une estimation minimale car l'hétérotachie spéciation-spéciation est sous-estimée seulement dans le jeu test.

Avec l'approximation *racine-feuilles* (estimation d'erreur plus conservative), les caractéristiques les plus prédictives deviennent la longueur d'alignement, le taux de dS moyen, et le score de parcimonie moyen par nucléotide (une estimation du taux de substitution moyen). L'erreur prédite a cette fois une allure plus étalée vers les hautes valeurs (figure II.2b), suggérant deux populations d'arbres avec duplication/perte : une population dans un interval d'erreur similaire aux arbres contraints —1 341 arbres en dessous de 2,46 Ma/nœud—, et une population



a) Approximation speciation-spéciation



b) Approximation racine-feuilles

FIGURE II.1 – Prédiction de l'erreur des 19 331 arbres test, comparés aux 5 204 arbres contrôle *Simiiformes*. Les histogrammes bleus (arbres contraints, jeu d'entraînement) et verts (arbres non contraints, jeu test) sont empilés, leur hauteur additionnée représentant la distribution complète. Les arbres au dessus du 9e décile de l'erreur entraînée (bleu) sont colorés en plus foncé.

potentiellement sujette à des erreurs allant de 2,46 à 10 Ma/nœuds. On apprend surtout ici qu'une partie des arbres non contraints présente en même temps des caractéristiques dans l'alignement qui les rendent moins propices à une datation précise.

Cette dernière prédiction est celle incluse dans l'article « Factors influencing the accuracy in dating single gene trees » (p. d.).

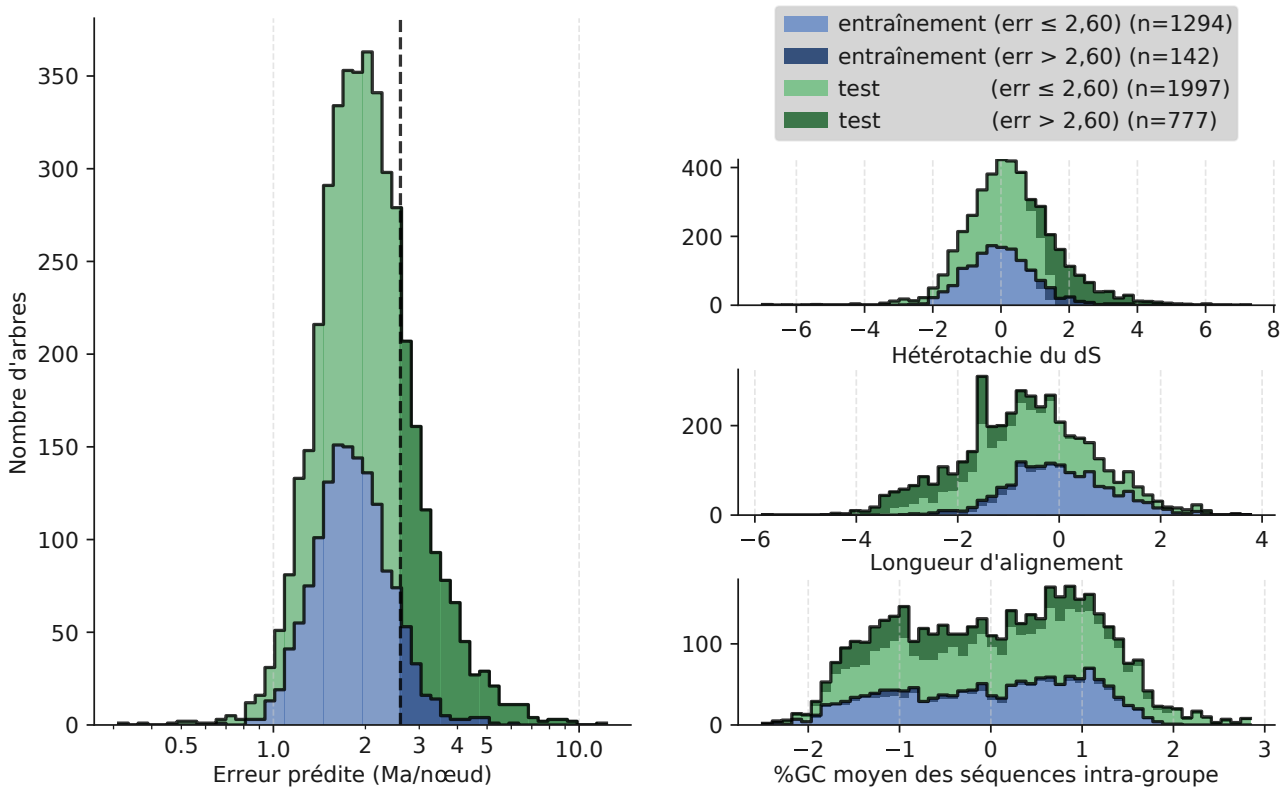
14 Distributions des âges de duplication chez les primates

L'un des objectifs de cette thèse est de dater les duplications dans le temps, le long des lignées, ici les lignées *Simiiformes*. Les choix méthodologiques concernant le calcul découlent directement du chapitre I : la procédure de génération des alignements est donc *fsa+cleaned*, et l'algorithme de datation est le MPL basé sur les longueurs en dS de codeml. De plus, la prédiction qualitative (section précédente) nous permet de sélectionner les meilleurs arbres : ici la moitié des arbres de meilleure qualité prédite (approximation spéciation-spéciation, section 13.1) a été utilisée. En conséquence, un total de 4 819 duplications est réparti le long de la phylogénie (figure II.3). Chaque nœud de spéciation est une calibration dure : impossible qu'un nœud de duplication sur une branche descendante soit plus ancien que la spéciation parente (et inversement sur une branche ancestrale).

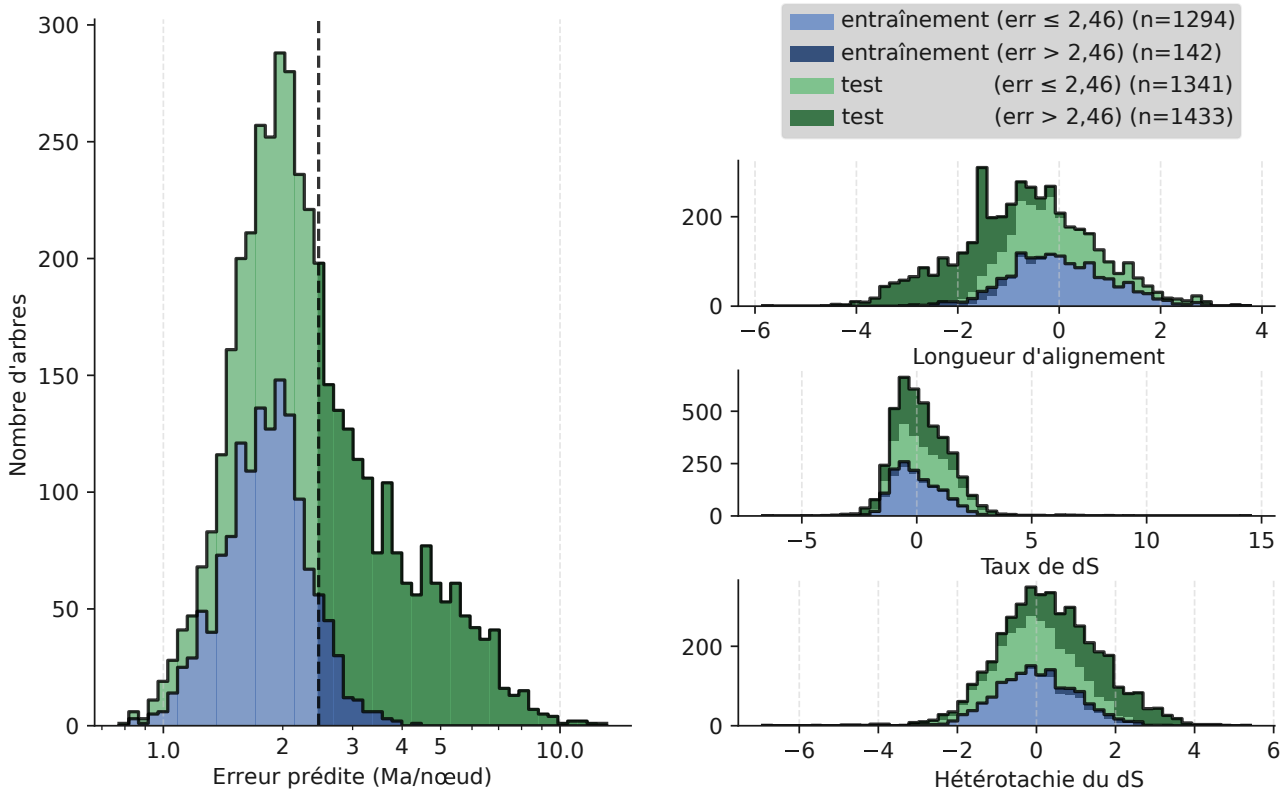
La distribution obtenue présente un signal assez faible : peu de duplications, et des pics éparpillés sur l'ensemble de la branche. Une tendance apparaît cependant : pour les branches internes (*Platyrrhini*, *Catarrhini*, *Hominoidea*, *Cercopithecidae*), les duplications sont toutes amassées au niveau de la spéciation ; similairement, pour beaucoup d'espèces (par exemple *H. sapiens* et *Chlorocebus sabaeus*), un pic d'âge zéro termine la branche : il s'agit de paralogues indistinguables par la séquence. Quelques espèces présentent néanmoins des duplications plus réparties sur la branche, en particuler *Pongo abelii* (Orang-outan) et l'ensemble des *Platyrrhini* (singes du nouveau monde).

Parmi ces derniers, les branches *Saimiri boliviensis* et *Cebus capucinus* laissent même suggérer une légère décrue des duplications depuis l'ancêtre commun. Cette décroissance corrèle avec les temps de générations des cébidés : leurs temps de générations sont en moyenne deux fois supérieurs à ceux des aotidés et callitrichidés (5 718 jours contre 3 048 et 2 371 respectivement; cf. annexe A10 d'après les données de Pacifici et al., (2013)) [1]. Un tel ralentissement de génération entraîne certainement un ralentissement du taux d'évolution moléculaire, mais il est incertain qu'il puisse produire un tel profil d'excès de duplications anciennes, à cause de deux mécanismes antagonistes : un ralentissement des taux de duplication génère *de facto* ce profil d'ancienneté, tandis qu'en compensation, un ralentissement simultané des taux de substitution réduirait l'âge *apparent*

1. Dans l'arbre *Platyrrhini*, deux familles absentes d'Ensembl (*Pitheciidae* et *Atelidae*) constituent des extra-groupes aux espèces présentées ici, avec des temps de générations intermédiaires, suggérant un ralentissement chez *Cebidae* parallèlement à une accélération chez *Callitrichidae*.



a) Approximation speciation-spéciation



b) Approximation racine-feuilles

FIGURE II.2 – Prédiction de l'erreur des 2 774 arbres test passant le filtre.

des duplications, les faisant paraître plus récentes qu'en réalité. Si ce profil est confirmé par d'autres méthodes, on pourrait imaginer que chez les cébidés, un allongement des générations cause un ralentissement des duplications plus accentué que le ralentissement des substitutions. Cette analyse dépend de l'arbre d'espèces *Platyrrhini*, difficile à résoudre, qui a ici la même topologie que celle calculée par DOS REIS et al. (2018) (figure 4), et diffère de celle de TimeTree (2020/06) qui inclut les callitrichidés dans les cébidés.

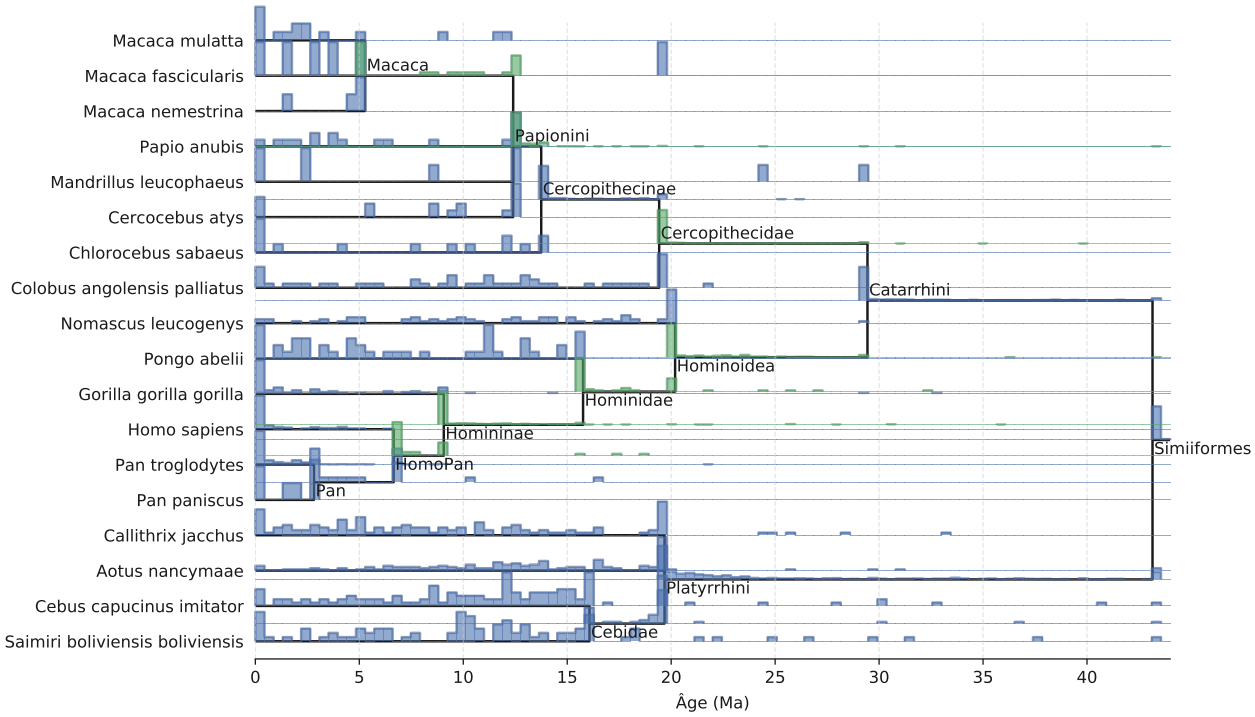


FIGURE II.3 – Âges de 4 819 duplications *Simiiformes*, procédure *fsa+cleaned+branchMPL*, moitié des arbres de plus haute précision prédite (9 680 arbres sélectionnés d'après la prédiction en section 13.1). La couleur bleu/vert sert uniquement à distinguer les histogrammes risquant de se superposer. L'échelle verticale des histogrammes représente le nombre d'événements, redimensionné pour remplir la hauteur entre chaque branche, et n'est donc pas comparable entre branches. Les événements datés hors branche (antérieurement à leur ancêtre immédiat) proviennent des arbres où certaines spéciations ne sont pas observées. Les nombres de duplications par branche sont fournis en annexe A9.

15 Distribution des taux de duplication entre familles de gènes

Dans les processus biologiques, l'hypothèse de constance d'un paramètre résulte plus souvent d'une simplification pratique que d'une réalité. Ainsi pour estimer les taux de substitution, il est plus simple et plus rapide de supposer le même taux pour chaque site d'un alignement ; aujourd'hui de nombreux modèles relâchent cette hypothèse de constance, une complexification qui est soluble en imposant une distribution paramétrée, souvent la loi Gamma pour les taux de substitution par site. Cette loi présente l'intérêt d'ordre méthodologique de bien représenter

un grand nombre de situations en faisant varier deux paramètres. Elle n'explique en revanche pas le mécanisme générateur de la distribution.

Parmi les outils concernant les duplications, la version 5 de CAFE (HAHN et al. 2005) permet d'ajuster une loi Gamma sur les taux par famille ; en revanche cela n'a pas été encore implémenté pour des modèles de calculs de taux faisant intervenir conjointement un modèle d'évolution de séquence et un modèle de réconciliation, comme `ALE` (SZÖLLŐSI et al. 2013). Cependant, le passage à l'échelle génomique est amorcé avec le développement de `GeneRax` (MOREL et al. 2020), qui applique le même modèle qu'`ALE` mais sur un ensemble de familles de gènes, permettant de coestimer les taux : il devient possible d'estimer un taux génomique global, ou bien un taux génomique variant selon les espèces, ou enfin un taux indépendant par famille. La prise en compte des taux par famille sous forme Gamma n'existe pas encore dans `GeneRax`, même si cela est possiblement en cours de développement.

Je me suis interrogé sur la possibilité et l'intérêt d'appliquer ce modèle de variation des taux suivant une loi Gamma. L'atout de ce type de regroupement statistique réside dans la possibilité de renforcer la confiance dans une estimation particulière, en s'appuyant sur plusieurs observations : si l'on connaît la forme de la distribution Gamma, on pourra exploiter l'ensemble des arbres pour attribuer un arbre particulier à une catégorie de taux. D'autre part cela permet d'estimer des taux moyens plus fiables. En construction d'arbres phylogénétiques, les topologies inférées dépendent fortement de l'utilisation d'un modèle de variation entre sites, signe que se fonder sur les taux moyens peut mener à des conclusions erronées. Il paraît donc intéressant d'intégrer la variation de taux en amont dans les modèles d'inférence de taux de duplication. Par contraste, l'analyse que nous avons effectuée est seulement exploratoire : elle étudie *a posteriori* la distribution des taux, c'est-à-dire après calcul par `GeneRax` de taux indépendants par famille.

Dans un premier temps, nous avons calculé les taux de duplication et de perte par famille, puis ajusté la loi Gamma (figure II.4). L'ajustement suggère des valeurs de forme $\alpha = 0{,}11$ et d'échelle inverse $\beta = 4{,}3$ pour l'ensemble des taux de duplication. Si l'on exclut les taux nuls, on obtient $\alpha = 1{,}7$ et $\beta = 26$. Comme attendu, ces deux valeurs indiquent une distribution très asymétrique avec une longue queue de valeurs extrêmes. La divergence de Kullback-Leibler de la densité ajustée par rapport à l'histogramme observé est de 0,186 pour la forêt complète, et 0,084 pour les familles à taux non nul, ce qui fait partie des valeurs les plus basses parmi les lois testées (figure II.5).

La distribution des taux de pertes est en revanche plus erratique, avec une tendance bi- ou tri-modale symptomatique d'une discrétisation des taux estimés. Nous nous intéresserons donc uniquement aux duplications.

Parmi les autres lois testées, la Gaussienne inverse (ou Wald) diverge de seulement 0,122, et la Chi (à degrés de liberté) de 0,188.

La forte proportion d'arbres n'arborant aucune duplication suggère cependant qu'un modèle avec "sites invariants" serait pertinent.

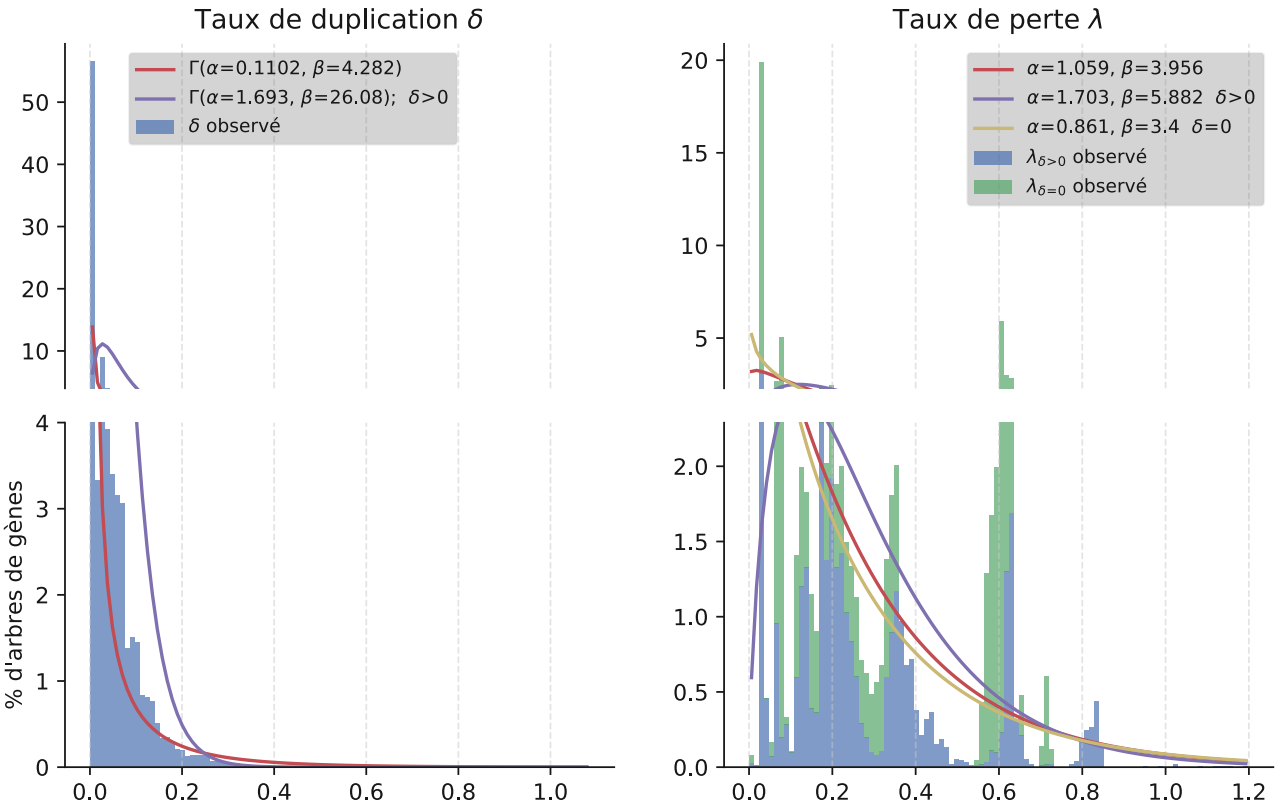


FIGURE II.4 – Distribution des taux de duplication/perte par famille et ajustement d'une loi Gamma. La courbe rouge représente la densité ajustée sur l'ensemble des familles, la violette sur les familles à taux non nul. L'échelle verticale est partitionnée pour mieux afficher les faibles fréquences.

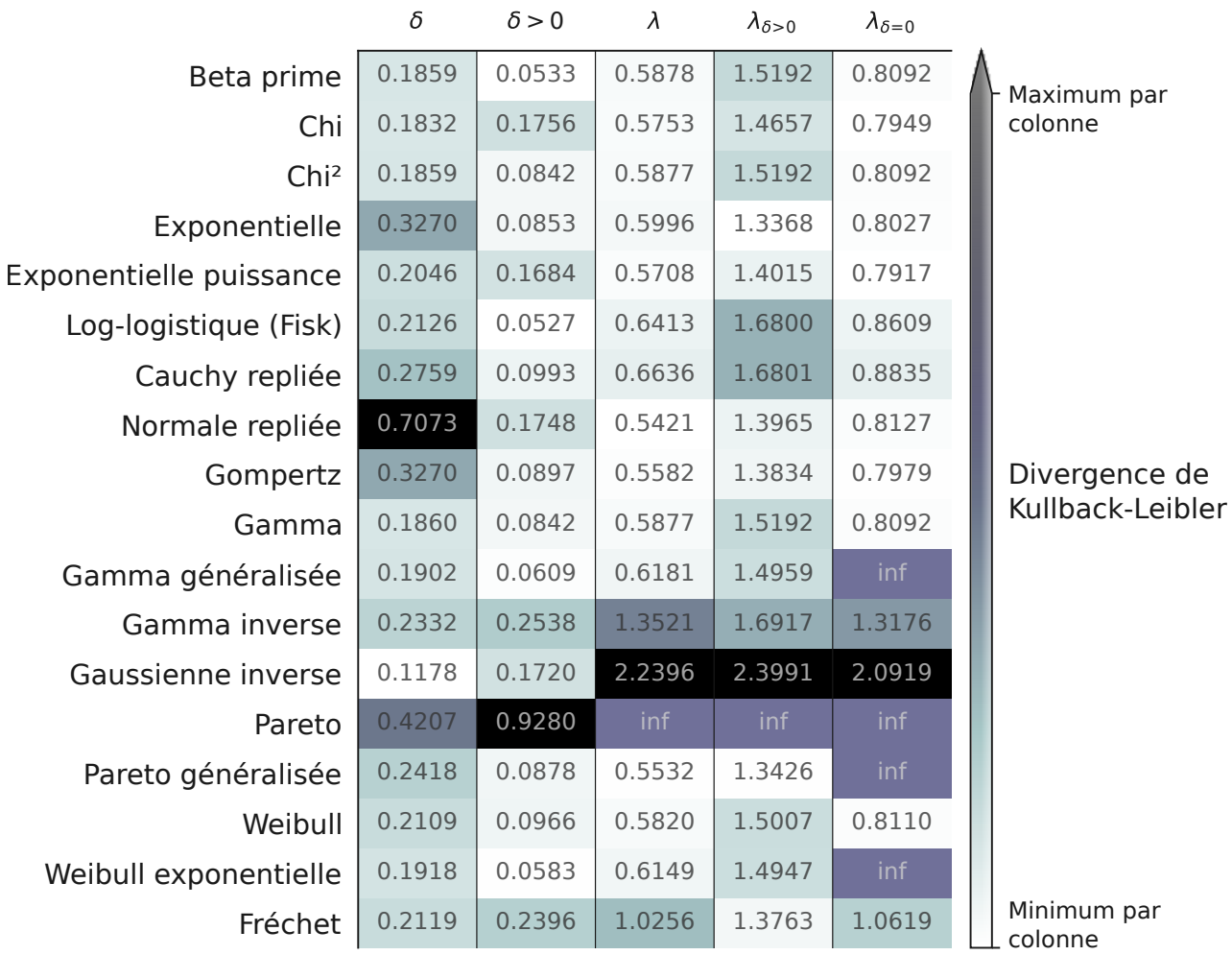


FIGURE II.5 – Divergences de Kullback-Leibler entre une loi théorique ajustée (lignes) et la distribution mesurée (colonnes). En colonnes, ont été mesurés les taux de duplications totaux (δ) et non nuls ($\delta > 0$), de pertes (λ), et de pertes selon la présence de duplications ($\lambda_{\delta=0}$ et $\lambda_{\delta>0}$)

Enfin, j'ai corrélé ces taux de duplication avec les caractéristiques des arbres *Simiiformes* rassemblés en chapitre I, qui comprennent les arbres avec duplications/pertes (total de 19 040). Une première régression est faite sur le jeu entier ($R^2 = 0{,}153$, figure II.6a), une seconde exclut les arbres au taux de duplication quasi-nul ($R^2 = 0{,}122$, figure II.6b).

La régression multiple (même procédure que chapitre I) a un faible pouvoir explicatif, ($R^2 = 0{,}153$). Les caractéristiques montrant une forte corrélation avec le nombre de branches ont été enlevées en amont. Cela concerne la log-vraisemblance de `codeml` par site, ainsi que l'écart-type du score de parcimonie entre sites, qui est positivement corrélé au nombre de branches.

Parmi les premières caractéristiques corrélées on trouve (figure II.6a) :

— la proportion moyenne de gaps dans l'alignement (positivement) ;
— la présence de triplets ou de paires de branches de longueur nulle (positivement) ;
— l'écart-type des longueurs de séquences dans un alignement (positivement);

suggérant que des difficultés de reconstruction sont responsables des taux de duplication élevés.

Les autres caractéristiques notables :

— l'écart-type des taux racine-feuilles (positivement) ;
— l'écart-type d'$\omega = dN/dS$ par branche (positivement) ;
— le score de parcimonie moyen par site (négativement) ;

vont dans ce sens, car elles montrent que les arbres aux taux de substitutions faibles, et hétérogènes entre branches correspondent aux taux de duplications élevés.

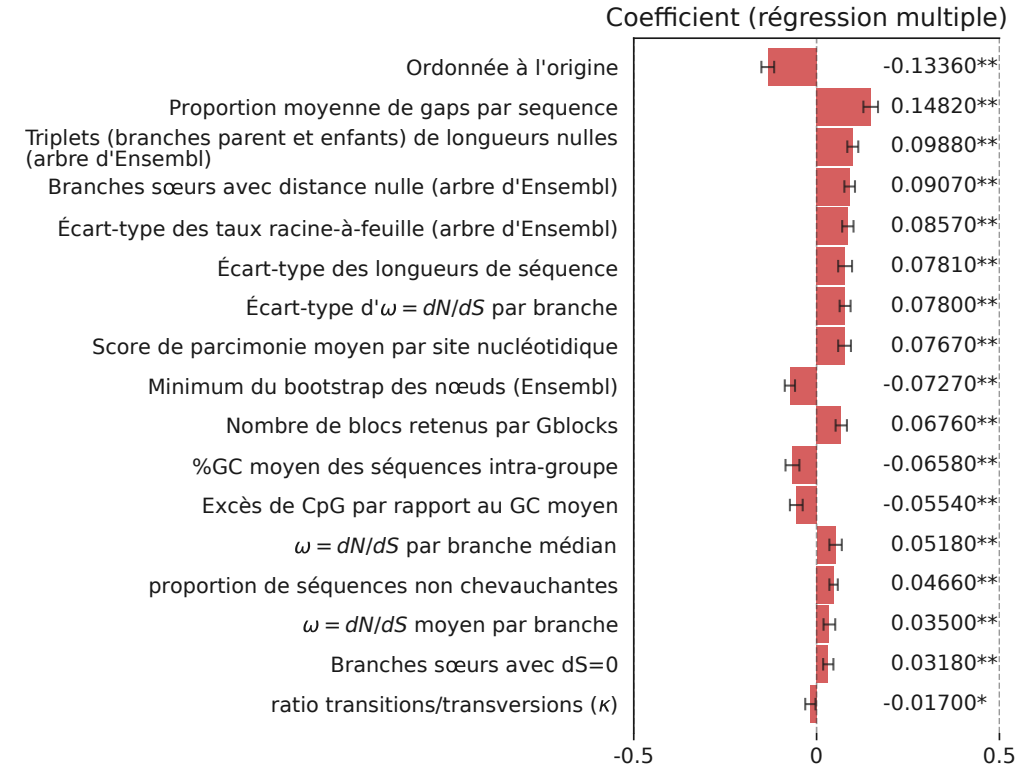
En revanche le pourcentage de G/C moyen (10^{e} coefficient) pourrait indiquer au contraire une influence de l'environnement génomique dans la duplicabilité des gènes.

La régression uniquement sur les arbres à taux de duplication non nul (figure II.6b) retient également comme coefficients principaux les variables sur la proportion de gaps, le score de parcimonie moyen, et les triplets de branches de longueurs nulles.

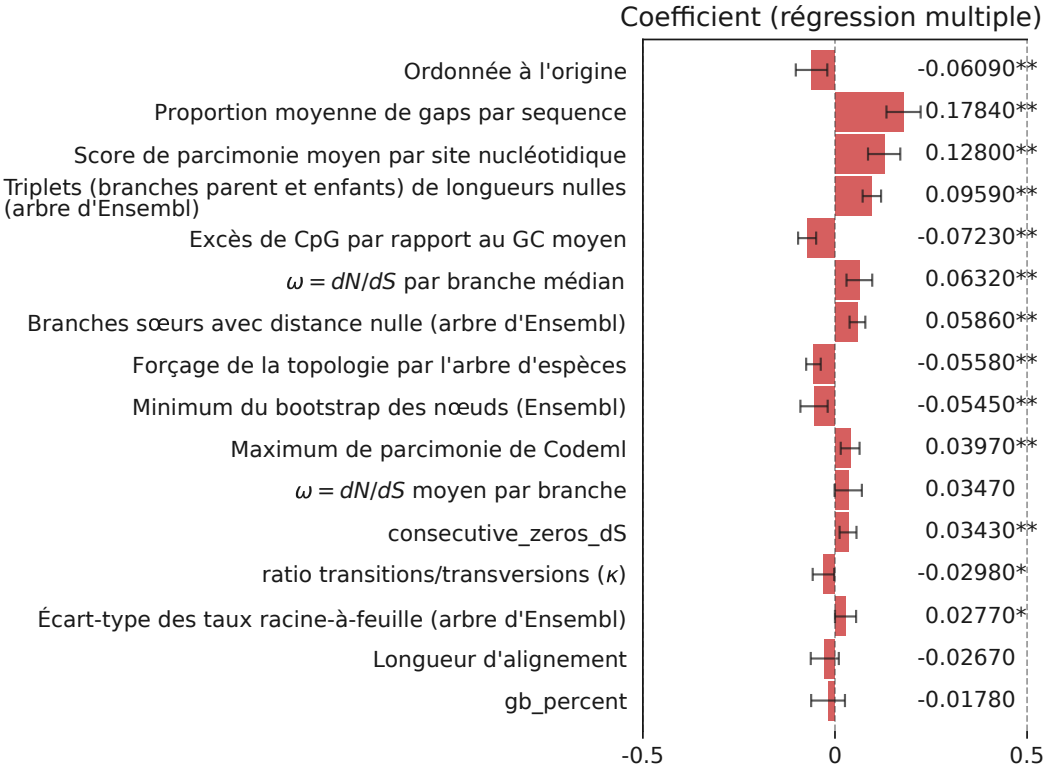
Il importe cependant de souligner que l'inspection visuelle des relations entre taux de duplication et des caractéristiques corrélées suggère des effets très faibles.

16 Évaluer le lien avec la dynamique de diversification en espèces

Dans cette partie, nous corrélons les taux d'évolution moléculaires avec les taux de diversification taxonomique (cf. Eo et DeWoody 2010).

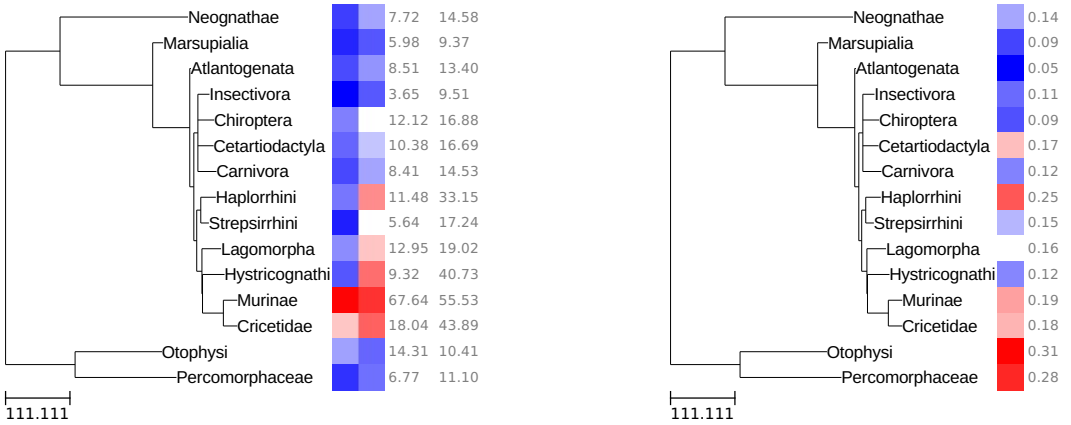


a) 19 040 arbres, incluant $\delta \sim 0$. $R^2 = 0{,}153$



b) 7 882 arbres à taux de duplication $\delta > 0$. $R^2 = 0{,}122$

FIGURE II.6 – Caractéristiques des arbres corrélées avec le taux de duplication (p-valeur $< 0{,}01$, p-valeur $< 0{,}05$)



a) Nombre de nœuds de duplications par Ma (en tandem à gauche, dispersés à droite)

b) Taux de spéciation, calculés selon un modèle de naissance-mort à taux constant avec fractions d'échantillonnage

FIGURE II.7 – Clades aggrégés d'âge inférieur à 150 Ma avec moins de 10 espèces dans Ensembl 90, et leurs caractéristiques. Échelle de couleur correspondant aux chiffres par colonne. Longueurs de branches en Ma.

L'approche la plus simple consiste à effectuer une corrélation des taux observés au niveau de clades terminaux. Cette approche a trois avantages pratiques : 1) il n'est pas nécessaire d'inférer les taux ancestraux (c'est seulement la régression phylogénétique qui fait les hypothèses concernant les valeurs ancestrales) ; 2) comme on aggrège les taux par clade, les données génomiques et les données taxonomiques deviennent comparables, bien que les échantillonnages d'espèces soient très différents ; 3) on peut estimer des taux de diversifications différents dans l'arbre, en ajustant simplement un modèle à taux constant par clade (au prix d'un découpage arbitraire). En particulier, nous considérons les clades selon leur échantillonnage dans Ensembl 90, d'âge inférieur à 150 Ma et rassemblant au plus 10 espèces, et traitons ensuite les taux de duplication et diversification comme des traits de ces clades (figure II.7).

Par régression phylogénétique permettant de prendre en compte l'inertie évolutive entre espèces apparentées (GRAFEN 1989), nous testons l'association entre taux de duplications retenues (en tandem ou dispersés) et taux de spéciation bruts. Avec cette méthode, on détecte une association avec les taux de duplications dispersées, mais qui ne résiste pas à une correction des tests multiples (seuil de p-valeur 0,05). Incorporer les temps de générations moyens des clades ne donne aucune association significative, car ceux-ci ne corrèlent pas avec les taux de diversification dans notre jeu de données.

Dans ce jeu de données, notons que les deux clades actinoptérygiens (poissons à nageoires rayonnées) *Otophysi* et *Percomorphaceae*, ont peu de duplications (figure II.7a) mais une extraordinaire diversification (II.7b), et donc, du fait de leur apparentement phylogénétique combiné au contraste de leurs traits avec le reste des clades, introduisent la phylogénie comme facteur principal expliquant la répartition des traits. En termes appropriés, le *signal phylogénétique* est fort, et ce principalement à cause de ces deux clades, ce qui serait plus adéquatement vérifié

avec une métrique existante (voir MÜNKEMÜLLER et al. 2012). En supprimant ces deux clades du jeu de données, on observe peu de différences entre les moindres carrés phylogénétiques et ordinaires, mais dans tous les cas aucune corrélation significative n'est obtenue, peut-être à cause du petit nombre de clades considérés.

Ainsi cette première approche approximative ne permet pas de détecter d'association évidente entre le nombre de duplications retenues et le taux de spéciation dans les clades étudiés. Afin de correctement exclure cette hypothèse, il faudra améliorer le calcul des taux de duplication, et utiliser un échantillonnage génomique mieux réparti taxonomiquement. Une méthodologie de test de corrélation phylogénétique plus puissante doit également être développée, qui soit adaptée aux processus de duplications et diversification, par forcément bien représentés par une évolution brownienne.

17 Discussion

Le travail fourni ici établit un portrait quantitatif de la dynamique de duplication de gènes chez les primates : en II.13 nous avons sélectionné les arbres de gènes dans lesquels la datation des nœuds est de plus grande précision, et comparé cette précision prédite à celle des arbres contraints. En II.14, à partir de ces arbres, nous avons daté les duplications et dressé leur profil temporel. En II.15, nous avons calculé les taux par arbre, en appliquant simultanément un modèle phylogénétique de réconciliation. Enfin, nous avons calculé les taux par lignée, et en **??** estimé sa relation avec les taux de diversification.

17.1 Fiabilité de l'hétérotachie comme prédicteur

Parmi les prédicteurs ayant une forte influence sur la précision des datations, de façon assez tautologique, on trouve l'hétérogénéité des taux de substitution entre branches, ou *hétérotachie*. Le tautologisme en question réside dans le fait que pour calculer l'hétérotachie, il faut connaître les âges des nœuds (pour ramener les nombres de substitution en taux par millions d'années), alors que pour connaître les âges des nœuds, il faut appliquer une horloge moléculaire qui estime simultanément taux et âges, en contraignant le mode de variation des taux et en calibrant l'âge de nœuds de référence. Il peut donc paraître prévisible que des taux inconstants influencent les âges calculés. Ce qui est en revanche moins évident, c'est que même les horloges à taux relachés décorrélés s'accomodent mal de cette hétérotachie, dans nos données. À cause de cette relation d'interdépendance entre âges et taux, ma prédiction d'erreur ne peut se faire sur des nœuds d'âge inconnu —par opposition au chapitre I, où des âges de référence sont disponibles.

Pour cette prédiction, puisque les nœuds de duplication sont d'âge inconnu, il faut d'abord estimer l'hétérotachie sans connaître les âges de tous les nœuds. Plusieurs méthodes permettent de s'affranchir de ces nœuds, au prix d'une approximation (figure II.8) : en utilisant uniquement les chemins joignant les nœuds connus, on considère implicitement que le taux le long du chemin est constant, puisque

les nœuds inconnus intermédiaires n'apportent aucune information. Cet effet de moyennage diminue donc la résolution du calcul, produisant une sous-estimation de l'hétérotachie. Cette sous-estimation est d'autant plus critique qu'elle provient justement de ces nœuds que l'on cherche à dater : l'approximation gomme donc l'hétérotachie de ces nœuds dont la précision de datation nous intéresse.

En complément, le taux moyen par arbre souffre lui aussi de l'approximation : certaines branches sont en effet comptabilisées plusieurs fois, puisque les branches ancêtres des nœuds inconnus sont réassociés à plusieurs chemins distincts. En revanche, contrairement à l'hétérotachie, le taux moyen n'est pas biaisé dans une direction particulière.

Il peut donc paraître impossible de prédire l'erreur de datation de nœuds avec cette approximation. Cependant, la régression de l'erreur *locale* effectuée en section I.9 nous indique que les facteurs les plus prédictifs localement comprennent le taux moyen local, ainsi que *non-local* du fait d'une certaine autocorrélation des taux. Cette autocorrélation laisse penser que le taux moyen global est tout de même un prédicteur influent de l'erreur locale de datation, pouvant compenser une sous-estimation de l'hétérotachie locale.

En gardant à l'esprit ces limitations, mes résultats montrent tous que l'erreur dans les arbres non contraints sera globalement plus élevée, si l'on considère que la relation linéaire utilisée pour prédire est valide, et l'approximation d'hétérotachie suffisamment sensible. Cependant, l'approximation la plus adéquate pour obtenir une borne supérieure d'erreur —racine-feuilles—, relègue à d'autres caractéristiques le pouvoir prédictif (à savoir essentiellement la longueur d'alignement et le taux moyen). En conséquence il serait souhaitable d'estimer l'hétérotachie de manière plus fine, mais comparable entre les arbres contraints et les arbres avec duplication/perte. Parmi les possibilités immédiates, nous aurions pu considérer l'hétérotachie des "cerises", c'est-à-dire des paires de branches sœurs terminales (défini par McKenzie et Steel 2000).

En résumé, la méthode de calcul d'hétérotachie que j'applique dans ce chapitre est une approximation non-paramétrique aux mêmes principes que le MPL. Elle correspond dans sa logique au test statistique de non-constance d'horloge dans la publication du MPL (Britton et al. 2002), basé sur l'écart entre longueurs de chemins, mais qui de plus associe l'écart à une distribution normale dans le cas où suffisamment de substitutions sont observées. Ce parallèle apporte deux perspectives : dans la mesure où l'hétérotachie est un prédicteur d'erreur majeur, le test de Britton pourrait être utilisé ici pour chiffrer la qualité de datation de nos arbres ; cependant, nous donnons en plus de l'importance à la taille des alignements et aux nombres de substitutions pour catégoriser nos arbres : cela permet d'explorer la zone de non-résolution du test de Britton, à savoir en dessous d'une trentaine de substitutions. Malheureusement, notre approche souffre de l'inconvénient majeur des techniques d'apprentissage, puisqu'elle dépend du jeu de données d'entraînement —les arbres contraints—, et extrapole aux nouveaux arbres en supposant qu'ils soient de même nature.

17.2 L'émergence de méthodes pour la diversification, applicables aux duplications ?

Dans un autre cadre statistique, les modèles phylogénétique d'horloge moléculaire peuvent se prêter à des estimations d'erreur, par exemple les intervalles crédibles résultant d'un calcul bayésien. Dans ce cadre, on ne se baserait plus sur des contrôles externes (nos âges de spéciation de référence), mais en autonomie grâce au produit du calcul bayésien : la distribution de probabilité postérieure. De cette distribution on peut immédiatement calculer l'intervalle où un paramètre (par exemple la date d'un nœud) se trouve avec 95 % de probabilité (postérieure). Ainsi un relief plat de probabilité postérieure se traduira par des intervalles crédibles larges. La distribution postérieure obtenue par échantillonnage en MCMC reste cependant une estimation, et utiliser un jeu contrôle reste une vérification indispensable pour détecter si un modèle est adéquat.

Dans le nouveau jeu de données analysé ici, la distribution des caractéristiques phylogénomiques des arbres avec duplication/perte comparativement aux arbres contraints suggère des difficultés de datations au delà d'un certain seuil : le tiers des arbres aux alignements plus courts et aux hétérotachies, même sous-estimées, supérieures, correspond à des erreurs de plus de 3 Ma/nœud (figure II.2) ; dans un arbre d'espèces où l'intervalle entre spéciations est parfois plus faible, cela suggère qu'appliquer une datation par horloge moléculaire n'apporte pas plus d'information que la réconciliation en elle-même. Il paraît donc plus judicieux à l'heure actuelle, où les séquences génomiques et transcriptomiques abondent, d'augmenter la couverture taxonomique du jeu de données. Suivant cette stratégie, il resterait uniquement les branches anciennes et les taxons peu diversifiés où un échantillonnage taxonomique plus large serait sans effet, dans lesquels une stratégie d'horloge pourrait alors permettre d'obtenir les âges de duplication, dans la limite statistique imposée par les tailles de séquence.

Néanmoins, à titre exploratoire, nous avons utilisé cette prédiction d'erreur pour restreindre aux données les plus fiables les duplications à considérer, puis déterminé leur distribution temporelle (section II.14). Ici, chaque spéciation était calibrée, dans un but de précision maximale. Cependant cette distribution d'âges assigne des duplications au niveau des spéciations. C'est le signe que le sous-jeu sélectionné contient encore des arbres difficilement datables, pour cause de trop peu de substitutions.

En résumé, l'approche par horloge manque de résolution pour dater des événements géniques. Il est possible, tout au plus, de détecter des tendances basées sur un grand nombre de gènes, comme c'est le cas lorsqu'on identifie des duplications complètes de génomes grâce aux distributions de dS, mais un unique gène contient le plus souvent trop peu d'information.

Pour cette raison, étudier les duplications en groupe, en se basant sur les taux le long des lignées est plus fiable. Nous avons utilisé des modèles de calcul de taux de duplication/perte permettant de décrire les différences entre familles et les différences entre lignées. Ils font appel à un processus de naissance-mort, un type de modèle fréquemment utilisé en diversification d'espèces (Nee et al.

1992), parfois joint à un modèle de réconciliation et d'évolution de séquences (`Generax`). Or lorsqu'on ne dispose d'aucune donnée sur les branches éteintes, ces modèles de naissance-mort sont surtout performants pour estimer le taux *net* de diversification (naissance moins mort) (LOUCA et PENNELL 2020). C'est ce que révèlent également mes valeurs de taux de perte discrétisés des familles primates. Il serait intéressant de vérifier ce problème d'identifiabilité dans le cadre des modèles de réconciliation, ou, à défaut, d'utiliser les taux nets. Autrement, on peut tirer parti du grand nombre de familles de gènes, et représenter leur variation de taux par une distribution, comme la Gamma que nous avons testée. Cela permettrait de combiner l'information de plusieurs arbres pour renforcer la confiance dans les résultats obtenus sur un seul arbre, notamment les taux et les topologies. Parmi les algorithmes de réconciliations, une procédure d'amalgamation de duplications segmentales est d'ailleurs développée par DONDI et al. (2019) : la méthode, optimisant un critère de parcimonie, consiste à tester les groupements possibles de duplications sur un même segment. Cette méthode est donc extrêmement attrayante pour reconstruire les réconciliations, mais pourrait également avoir un impact crucial sur les calculs de taux de duplication, et sur la compréhension du mécanisme à l'origine des duplications de gènes, qui bien souvent est modélisé en faisant abstraction du support chromosomique.

La notion d'aggrégation est également cruciale en datation, par l'augmentation de l'information de séquence. DUCHÊNE et al. (2019) ont montré que lier les longueurs de branches entre arbres améliorait la précision, mais surtout, que les arbres de gènes affichent essentiellement des proportions identiques malgré des taux moyens variables. En perspective, à travers ces deux études, amalgamer les duplications segmentales serait prometteur sur de nombreux plans.

Pour calculer les taux de duplications, beaucoup d'hypothèses restent encore à tester. Pour l'instant les modèles existants de naissance et mort testent peu de modes de variation des paramètres de naissance-mort. En plus de la variation entre familles et entre lignées, l'association à d'autres facteurs permettrait de résoudre de nombreuses questions, et c'est ce que nous ébauchons en étudiant le lien avec la diversification. De nombreuses complexifications sont envisageables : on pourrait notamment tester des modèles de duplication densité-dépendants, où le nombre total de copies aurait un rétro-effet négatif sur le taux de duplication. Ces modèles diversité-dépendants existent dans le champ de la diversification taxonomique (MORLON et al. 2010), inspirés de la théorie de la coalescence. À cela s'ajoutent, toujours en provenance des études de diversification, les modèles "State-dependent Speciation and Extinction" (SSE, MADDISON et al. 2007 ; FITZJOHN 2010 ; GOLDBERG et al. 2011) qui constituent une famille versatile permettant d'associer la dynamique de branchement avec d'autres variables observables… ou avec aucune : dans l'HiSSE (Hidden State Speciation and Extinction, BEAULIEU et al. 2013 ; BEAULIEU et O'MEARA 2016), des états cachés sont à l'origine des taux. Cela me semble pertinent d'ainsi imaginer que les branches d'arbres de gènes tombent dans un nombre fini de catégories de taux, sautant occasionnellement d'une catégorie à l'autre, par exemple en fonction du voisinage génomique, du nombre de gènes de fonctions similaires, etc). Une variable cachée présente l'intérêt de pouvoir révéler des associations sans a priori sur la variable

causative.

Ces complexifications, si séduisantes qu'elles soient, ne pourront cependant être concrètement utiles que lorsque les données seront suffisantes pour faire parler les modèles, par exemple lorsque des stratégies d'amalgamation seront mises en place. En effet, nous avons pu voir sur les familles dupliquées chez les primates que les événements de duplication sont globalement rares à l'échelle d'un gène, hormis dans quelques familles. D'autre part, les duplications de gènes ont leurs propres problématiques : par tri incomplet de lignées entre CNV, l'origine de paralogues éloignés risquerait d'être parcimonieusement positionnée trop récemment (Une prise en compte est proposée par RASMUSSEN et KELLIS 2012). Inversement, les paralogues, surtout en tandem, sont sujets à la conversion génique, pouvant les faire paraître plus récents qu'ils ne le sont réellement. Plus généralement, les exons d'un même gène peuvent générer des topologies conflictuelles, ce qui serait dû aux paralogies cachées, aux substitutions multiples non détectées ou aux erreurs d'alignement, davantage que l'ILS (SCORNAVACCA et GALTIER 2016). De plus, le mécanisme de maintien d'un nouveau duplicat est le sujet d'un débat perdurant depuis des décennies : en 1970, Ohno avait principalement envisagé la néofonctionnalisation ou le maintien à l'identique par sélection pour un dosage accru. FORCE et al. (1999) ont proposé un mécanisme neutraliste de Duplication-Dégénérescence-Complémentation. La prépondérance de l'un ou l'autre fait débat mais a des conséquences distinctes sur l'évolution de la séquence, de sorte que l'on peut vraisemblablement détecter le processus à l'œuvre dans les arbres de gènes (ZHAO et al. 2015). Enfin, pour faire le lien entre duplications et diversification, l'hypothèse d'effets d'incompatibilités génétiques causant la spéciation devrait être modélisée comme processus cladogénétique : en complément d'un processus anagénétique se produisant en continu le long de chaque lignée d'espèce, on modéliserait également les événements ponctuels survenant à l'origine des clades. Dans cet ordre d'idée, un modèle testant l'impact sur la spéciation des changements de caryotypes a été développé (`ChromoSSE`, FREYMAN et HÖHNA 2018). Pour l'adapter sur les duplications, il faudrait cependant les arbres de gènes réconciliés sur une très grande couverture taxonomique, ce qui est pour l'heure en attente de séquençages, et de méthodes de calcul.

En conclusion, l'inconstance fait loi : les arbres avec duplications ont des taux de substitution encore plus incohérents que les arbres contraints, les rendant plus difficiles à dater ; de plus, le rythme de duplication varie également d'un arbre de gène à l'autre. Ces sources de variabilité, au-delà d'une difficulté technique à résoudre, constituent un phénomène évolutif à défricher.

18 Méthodes

18.1 Prédiction de l'erreur de datation

Cette partie exploite les résultats de la précédente pour prédire des erreurs de datations sur l'ensemble des arbres de gènes, y compris ceux qui présentent des pertes et des duplications, soit 19 331 arbres supplémentaires pour le clade

Simiiformes (figure I.8, G).

Précisément, en section I.9, une régression multiple est effectuée entre l'*erreur moyenne par arbre* et les caractéristiques des arbres, ce qui constitue un modèle prédictif : la relation linéaire entre les caractéristiques et la variable dépendante (l'erreur moyenne) peut être appliquée à de nouvelles données.

Une difficulté survient cependant pour mettre en place cette prédiction à partir de la régression *contrôle* de la section I.9 : dans cette partie-là, j'appelle cette régression "*contrôle*" car on utilise des dates de spéciation de référence pour calculer l'une des caractéristiques : l'hétérogénéité du dS/Ma entre branches. Cependant cette information de contrôle est inconnue pour les duplications. Il faut donc dans un premier temps réexécuter une régression avec l'hétérogénéité des taux calculée autrement, approximée.

18.1.1 Mesures alternatives d'hétérogénéité des taux

Contrairement à la section I.9, il faut pouvoir attribuer l'hétérogénéité des taux par arbre lorsque certains âges de nœuds sont inconnus ; deux méthodes sont possibles (figure II.8) :

1. Approximation "racine-feuilles" (utilisée avec un objectif similaire dans SMITH et al. 2018) ;
2. Approximation racine-feuilles, dans les sous-arbres entre chaque spéciation : "spéciation-spéciation".

L'approximation "racine-feuilles" calcule les longueurs totales de chemins depuis la racine jusqu'à chaque feuille. L'hétérogénéité "racine-feuilles" est l'écart-type de ces longueurs de chemin. On comprend que du fait de la comptabilisation répétée des branches ancestrales, cette mesure est nettement moins puissante que la mesure exacte.

Pour l'approximation "spéciation-spéciation", c'est le même procédé, mais localement : en effet un arbre contenant des duplications peut être découpé en sous-arbres ultramétriques (en Ma) rejoignant les spéciations consécutives. On calcule les longueurs de chemin racine à feuilles pour chaque sous-arbre, puis l'on calcule l'écart-type de tous les taux spéciation à spéciation. Cela implique d'utiliser les âges de spéciation de référence, mais a l'avantage que pour un arbre sans duplication, le résultat est exact. Le revers de la médaille est que la régression sera faite sur les mesures exactes, tandis que la prédiction sur des mesures potentiellement sous-estimées.

18.1.2 Ré-application stricte des mêmes transformations de variables

Ma méthode de régression transforme au préalable les variables, afin de s'assurer i) d'un faible coefficient d'asymétrie et ii) d'une faible colinéarité des prédicteurs. Il faut donc traiter les variables explicatives de la même façon, avec les mêmes paramètres.

En particulier j'ai appliqué une décorrélation particulière entre certaines paires de

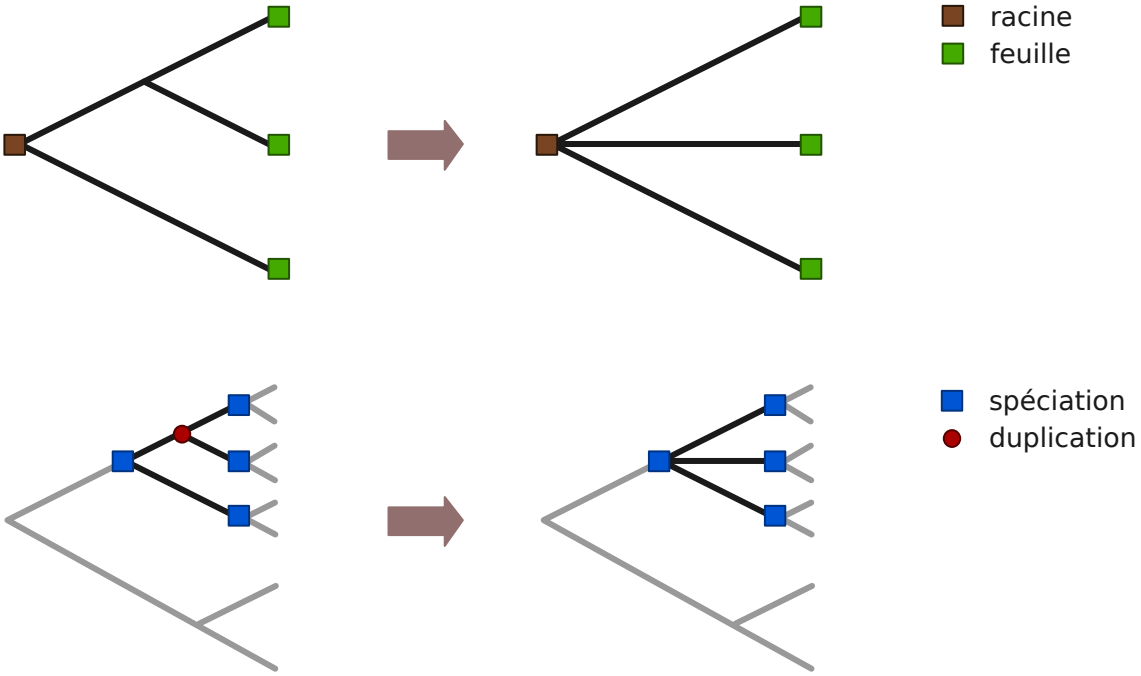


FIGURE II.8 – Transformation des branches pour calculer l'hétérogénéité des taux approximée

variables (difficiles à décorréler autrement), par exemple les paires "hétérogénéité de taux" *vs* "taux moyen", et "écart-type du score parcimonie par colonne" *vs* "score parcimonie moyen" : j'ai généré les résidus de la première variable régressée par la seconde (avec ordonnée à l'origine). Le jeu de données d'entraînement étant contenu dans l'ensemble des arbres, il faut donc réappliquer les mêmes coefficients de régression simple pour obtenir des "résidus" comparables dans l'ensemble des arbres, même si dans ce jeu élargi, les coefficients de décorrélation auraient été potentiellement différents (dans les faits, ils concordent approximativement).

18.2 Distribution des âges de duplication

Cette partie consiste à estimer l'âge de chaque duplication dans la forêt d'arbres *Simiiformes*, et à en dresser la distribution le long des lignées d'espèces. Pour chaque lignée d'espèce à espèce, j'obtiens donc une distribution temporelle des duplications. Comme en chapitre I, la méthode de datation utilisée est mon implémentation du Mean-Path-Length (MPL, voir section 12.6.1 pour le détail de l'implémentation). Les points de calibration sont chaque nœud de spéciation, datés d'après TimeTree, lorsqu'ils existent dans l'arbre.

J'ai également essayé sans succès le programme `DLRSOrthology` (ULLAH et al. 2015), censé produire les âges de duplication directement à partir de l'alignement et de l'arbre d'espèces. Ce programme dépend de PrimeDLRS, qui modélise conjointement l'évolution de séquence, l'horloge moléculaire et les taux de duplication/perte. Deux implémentations en sont disponibles (en C, et en java), mais toutes deux échouent lorsqu'on souhaite spécifier une matrice de substitution différente de Jukes-Cantor. La soumission d'une issue sur BitBucket n'a pas suffi à résoudre le problème, et la mailing-liste du programme n'est plus disponible

(lien mort).

18.3 Distribution des taux de duplication par famille de gènes

Par contraste avec la date des duplications observées, on s'intéresse ici au *taux*, c'est-à-dire, pour un arbre de gènes donné, combien de duplications se produisent et se fixent par million d'année, en prenant simultanément en compte les pertes. Si l'on calculait le taux simplement en divisant le nombre de duplications *observées* par le temps, on sous-estimerait le taux réel, et ce d'autant plus que les duplications sont anciennes. Un modèle de naissance et mort est donc plus approprié pour inférer le processus le long de l'arbre, en faisant ici la supposition que le taux est constant dans l'arbre.

Pour inférer les taux de duplication et perte j'ai utilisé `GeneRax 1.0.0` (MOREL et al. 2020). Un taux a été ajusté indépendamment pour chacun des 24 562 arbres *Simiiformes*. Les scripts développés pour ces calculs sont contenus dans le dossier `duprates` du dépôt `Phylorgs`.

18.3.1 Ajustement d'une distribution de probabilité

Les lois de probabilités continues positives implémentées dans `Scipy 1.4.1` ont été ajustées par maximum de vraisemblance. J'ai concentré mon interprétation sur les lois de types puissance et avec de longues queues de distribution, comme la Gamma, Pareto, ou Chi, au total 18 lois. Leur paramètre de localisation a été fixé à zéro. La densité de la loi Gamma se paramétrise en forme (α) et en échelle ($1/\beta$) :

$$f(x) = x^{\alpha-1}\frac{\beta^{\alpha}e^{-\beta x}}{\Gamma(\alpha)} \qquad (x \geq 0)$$

avec Γ la fonction gamma, analogue à la factorielle pour les valeurs continues. Les lois du χ^2 et exponentielle sont notamment des cas particuliers de la Gamma.

Étant donné le grand nombre de familles avec un taux de duplication estimé à 10^{-7}, c'est-à-dire virtuellement nul, deux ajustements de paramètres ont été effectués : le premier avec l'ensemble des familles, le deuxième sans les familles à taux (quasi) nul.

18.3.2 Évaluation de la qualité d'ajustement des lois de probabilité

1. Le critère d'information d'Akaike (AIC) permet d'estimer la qualité d'un modèle en pénalisant la vraisemblance par le nombre de paramètres :

   $$AIC = 2k - 2\ln\mathscr{L}$$

   avec k le nombre de paramètres à estimer et $\mathscr{L}$ la vraisemblance. Dans les lois sélectionnées, le nombre de paramètres va de 1 à 3.

2. La divergence de Kullback-Leibler (KL) est une mesure quantifiant la différence relative entre 2 distributions de probabilités, sa formule provenant comme l'AIC de la théorie de l'information. J'ai mesuré la divergence KL entre la densité de probabilité théorique ajustée et la distribution réelle (découpée en 100 intervalles). Celle-ci a l'avantage d'être comparable entre tous mes ajustements.

18.3.3 Régression du taux de duplication par arbre

Une régression multiple des caractéristiques des arbres contre le taux de duplication a été effectué, selon la même méthode qu'en chapitre I.

18.4 Contrastes indépendants phylogénétiques et régression phylogénétique

18.4.1 Corrélation chez des grands groupes de vertébrés

Cette approche beaucoup plus rapide est appliquée sur 15 clades de vertébrés, par régression phylogénétique.

18.4.1.1 Sélection de clades L'échantillonnage taxonomique dans Ensembl 90 est fortement biaisée vers les mammifères, notamment primates et rongeurs. De manière à être aussi équilibré que possible en terme d'âge et d'espèces échantillonnées, les premiers nœuds d'âges inférieur à 150 Ma et contenant moins de 10 espèces ont été sélectionnés, en partant de l'ancêtre vertébré :

— *Otophysi* (poissons actinoptérygiens, environ 8000 espèces) ;
— *Percomorphaceae* (poissons actinoptérygiens, environ 7000 espèces) ;
— *Neognathae* (oiseaux sauf ratites et tinamous, environ 9000 espèces) ;
— *Atlantogenata* (clade proposé de mammifères regroupant Xénarthres et *Afrotheria*) ;
— *Eulipotyphla* (ou *Insectivora* : taupes, musaraignes, hérissons, environ 500 espèces)
— *Chiroptera* (chauve-souris, 1200 espèces) ;
— *Cetartiodactyla* (ongulés et cétacés, 380 espèces) ;
— *Carnivora* (mammifères carnivores, 280 espèces) ;
— *Haplorrhini* (tarsiers et *Simiiformes*, 297 espèces) ;
— *Strepsirrhini* (lémurs, galagos, 133 espèces) ;
— *Lagomorpha* (lapins, pikas, 91 espèces) ;
— *Hystricognathi* (rongeurs dont cobaye et rats-taupes, 250 espèces) ;
— *Murinae* (rongeurs dont rat et souris, $\sim$ 500 espèces) ;
— *Cricetidae* (rongeurs dont hamster et lemming, $\sim$ 600 espèces).

18.4.1.2 Calculs des traits des clades Dans chaque clade, j'ai comptabilisé les nœuds de duplications, séparés en deux catégories —en tandem, ou dispersée—, puis converti ces nombres en taux par million d'années. Il s'agit

donc de taux de duplications *retenues* jusqu'aux espèces actuelles. Ces taux sont transformés en échelle logarithmique base 10 pour la régression.

Pour les taux de diversification, un modèle de naissance et mort a été appliqué sur les arbres d'espèces de TimeTree correspondant à chaque clade (avec `RPANDA`). Les fractions d'échantillonnage de TimeTree ont été estimées, en faisant le ratio d'espèces des mêmes clades dans la taxonomie NCBI. Le taux de naissance (spéciation) a été retenu.

Les temps de générations proviennent de PACIFICI et al. (2013), et ont été moyennés par clade.

18.4.1.3 Régression phylogénétique

Les taux de différents types de duplication ont été regressés par moindres carrés généralisés (fonction `gls` du package `R nlme`) contre les taux de spéciation en imposant une matrice de covariance phylogénétique (fonction `corBrownian` du package `R APE`). Les p-valeurs ont été ajustées par correction de Benjamini-Hochberg.

18.4.2 Corrélation chez les *Simiiformes*

Alternativement, la manipulation suivante est proposée mais n'a pas pu être terminée : une régression entre i) les taux de duplication et ii) les taux de diversification le long des lignées de la phylogénie *Simiiformes*.

i. Pour calculer les taux de duplication par lignée (i), `GeneRax` version 1.2.0 a été exécuté sur les alignements protéiques (options `--per-species-rates --rec-model UndatedDL`, modèle `LG+G`). Cette version de `GeneRax` échoue cependant en début de calcul sur mes données. Comme alternative, `CAFE version 5` (HAHN et al. 2005) a pu être utilisée.
ii. Pour calculer les taux de diversification par lignée (ii), l'arbre phylogénétique des *Simiiformes* de TimeTree a été extrait ; l'ajustement des taux a été exécuté avec le modèle ClaDS (MALIET et al. 2019) implémenté dans `RPANDA` (version 1.6) (MORLON et al. 2016). Le calcul n'a pas atteint des ESS suffisantes. Le report serait ensuite fait sur les branches de l'arbre d'Ensembl.
iii. La régression a été effectuée en prenant chaque branche d'espèce de l'arbre d'Ensembl comme une observation indépendante.

Chapitre III

Pertes corrélées de gènes et d'enhancers impliqués dans la latéralisation de l'embryon

> "Improbability Drive," she said patiently. "You explained it to me yourself. We pass through every point in the Universe, you know that."
> "Yeah, but that's one wild coincidence, isn't it?"
> "Yes."
>
> *The Hitchhiker's Guide to the Galaxy*, Douglas ADAMS

Les animaux à symétrie bilatérale peuvent posséder des organes disposés asymétriquement entre la gauche et la droite, comme le cœur et les viscères chez l'humain, ce qui implique la mise en place d'une polarité gauche-droite lors du développement embryonnaire.

L'expression différentielle du gène NODAL est responsable de l'asymétrie gauche-droite lors de l'embryogénèse des vertébrés, et de certains non-vertébrés (NAKAMURA et HAMADA 2012). Cette polarisation prend place à la fin du stade blastula, au moment de la gastrulation. À ce stade, les axes antéro-postérieur et dorso-ventral définissent déjà une *symétrie bilatérale* de l'embryon. La symétrie parfaite est ensuite rompue par la création d'un axe gauche-droite qui entraînera un positionnement asymétrique des organes internes dont le cœur. Cette rupture de symétrie gauche-droite est coordonnée par une région organisatrice sur la lèvre dorsale du blastopore (LRO, Left-Right Organiser) appelé nœud ventral chez la souris, nœud de Hensen chez les oiseaux, organisateur de Spemann-Mangold chez les amphibiens, vésicule de Kuppfer chez le poisson-zèbre, sous l'action du gène NODAL (appartenant à la famille fonctionnelle des Transforming Growth Factors TGF-β). La face interne de cette monocouche de cellules donne sur le blastocœle, cavité emplie de liquide. Chez les mammifères et le poisson-zèbre, ce liquide joue un rôle clé dans la polarisation gauche-droite : des cils motiles en

rotation horaire, situés sur la face interne de l'organisateur génèrent un courant causant la dégradation de l'ARN messager d'un inhibiteur de NODAL, DAND5 [1], dans le mésoderme latéral gauche. Chez les oiseaux et les reptiles cependant, on ne trouve pas de cils motiles dans l'embryon, et c'est un mouvement cellulaire qui cause l'expression différentielle de NODAL [2] (KAJIKAWA et al. 2020). De plus, chez les mammifères, un enhancer reconnu par FoxH1 activerait l'expression de NODAL dans le mésoderme latéral gauche, et a été qualifié d'enhancer asymétrique (KAJIKAWA et al. 2020).

En aval de ce flux de liquide dirigé, c'est la protéine PKD1L1, en se fixant sur le canal ionique PKD2 des cils immotiles latéraux qui entraîne la dégradation des ARNm de DAND5. Suite à l'activation latéralisée de NODAL, une cascade d'activation de gènes fait intervenir entre autres LEFTY , un inhibiteur de NODAL (BLUM et al. 2014).

Cette phase du développement fait intervenir des structures homologues conservées entre les vertébrés, le blastopore et le nœud organisateur gauche-droite. Cependant le mécanisme précis de latéralisation diffère chez les sauropsides, n'utilisant pas de cils motiles. À cela s'ajoute la découverte que les gènes DAND5 et MMP21 (Matrix Metalloproteinase-21, impliqué dans la voie, GUIMIER et al. 2015 ; PERLES et al. 2015) sont non seulement absents des sauropsides, mais aussi des cétartiodactyles (ongulés à nombre de doigts pair et cétacés). Ces deux altérations développementales similaires indépendantes vont à l'encontre de la supposition d'une ontogénèse strictement décalquée sur la phylogénèse.

Chez l'humain, des mutations inactivatrices dans ces deux gènes entraînent des malformation cardiaques et des anomalies de positionnement des organes, que l'on appelle *hétérotaxies*. Une hétérotaxie peut se traduire par une forme parfaitement inversée, donc non pathologique, le *situs inversus totalis*, ou par tout un spectre d'anomalies. Parmi les patients étudiés dans le monde, il n'a pas été possible de toujours déterminer la mutation causative. L'objectif de ce travail est de trouver des gènes ou des séquences régulatrices pouvant causer une hétérotaxie. Pour cela, nous partons de l'hypothèse que deux altérations indépendantes (chez les sauropsides et les cétartiodactyles) d'un mécanisme développemental devraient être corrélée à l'évolution d'autres gènes fonctionnellement impliqués, spécifiquement identifiables car la coïncidence d'une telle double perte semble improbable (voir l'arbre de MMP21, figure III.1).

19 Sélection de gènes candidats par corrélation phylogénétique

Suite à l'observation, sur un petit nombre d'espèces, de l'absence de MMP21 chez les oiseaux et les cétartiodactyles par Chris GORDON et collègues (GUIMIER

1. ou *Cerberus 2*, *Cer2*.

2. les gènes NODAL mammifère et sauropside sont en fait des *paralogues* retenus différentiellement, d'après KAJIKAWA et al, et KURAKU et KURATANI (2011), OPAZO et al. (2019). On trouve les deux paralogues chez les amphibiens, les actinoptérygiens, les chondrichtyens.

et al. 2015), une collaboration s'est mise en place pour effectuer un criblage plus systématique. L'objectif est de prendre en compte davantage d'espèces, et de détecter davantage de candidats en tolérant certains écarts par rapport à la répartition de référence (figure III.1). En effet, il est possible d'observer des séquences présentes dans les clades où l'on cherche une absence, si celles-ci sont en voie de pseudogénéisation. Inversement, s'il y a rétention d'un unique homologue et que sa présence est concomitante d'une perte assez basale dans le clade, on pourra considérer qu'il y a une bonne corrélation. Enfin, nous disposons des données de présence-absence d'enhancers conservés chez les vertébrés (CLÉMENT et al. 2020), permettant d'élargir notre recherche au-delà des séquences codantes.

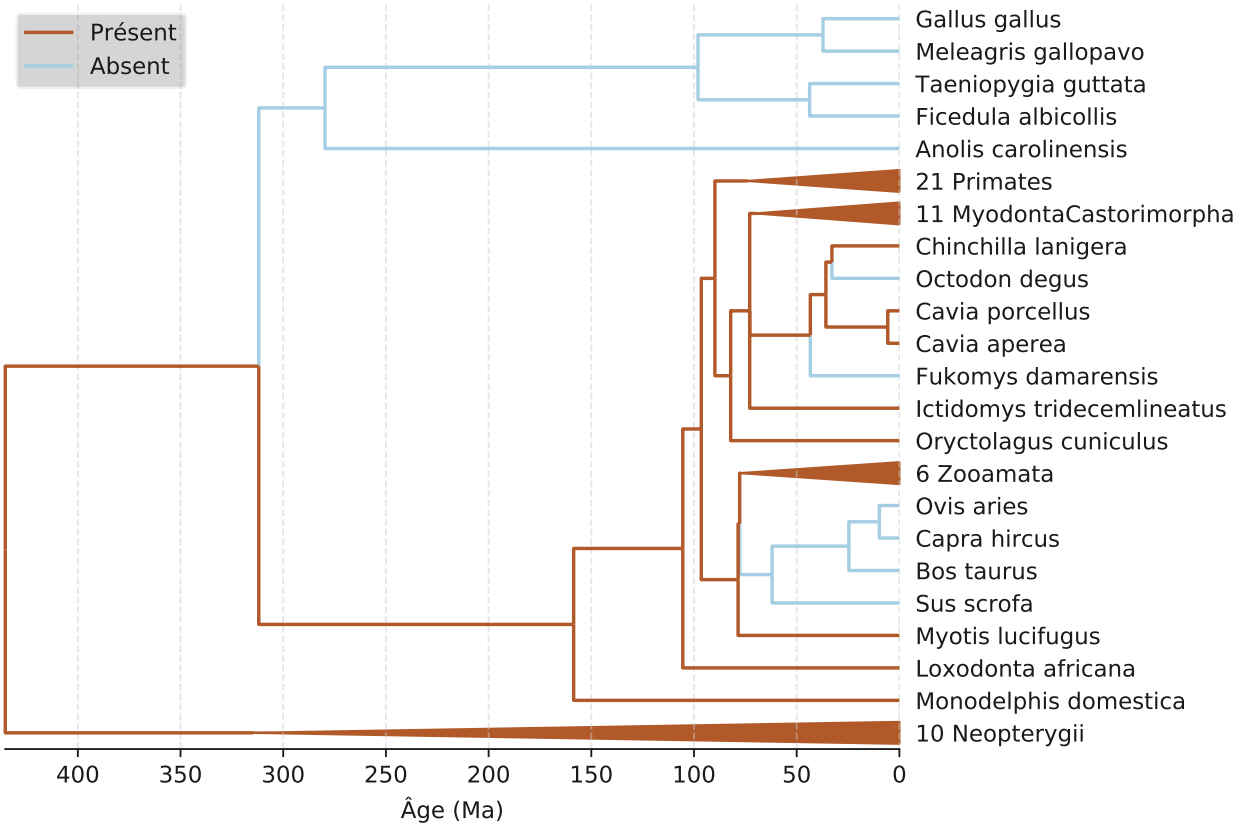


FIGURE III.1 – Présence-absence d'un orthologue de MMP21 chez les *Euteleostomi*, réconciliation d'après Ensembl. Cette répartition sert de référence à comparer aux autres gènes, à l'exception des pertes apparentes chez les rongeurs *Octodon degus* et *Fukomys damarensis*. *Neopterygii* (en majorité des poissons à nageoires rayonnées) sert d'extragroupe chez qui un orthologue doit obligatoirement être présent.

19.1 Classement des arbres de gènes selon quatre clades de référence

Dans cette première approche, nous calculons un score d'après la répartition du gène dans quatre clades :

— *Euarchontoglires* et *Carnivora* chez qui les absences sont pénalisées (23.4.1) ;
— *Sauria* (oiseaux et reptiles) et *Artiofabula* (Cétartiodactyles excepté les camélidés), chez qui les présences sont pénalisées. Comme ces deux clades

définissent la zone de recherche d'un taux de perte accru, je les appelle "clades focaux" dans la suite.

Le classement subséquent permet d'obtenir la tête de liste en tableau III.1.

TABLE III.1 – Arbres de gènes les mieux classés. Le score maximum théorique est de 2, signifiant adéquation parfaite avec le motif de présence-absence de référence, et le minimum est -3, signifiant opposition parfaite.

Gène amniote ancestral	Score	Gène humain	description
ENSGT00910000144034.A	1,948	MMP21	matrix metallopeptidase 21
ENSGT00390000010375.a.a	1,897	CDC42SE2	CDC42 small effector 2
ENSGT00530000062926.a.c.a	1,807	ADIPOR1	adiponectin receptor 1
ENSGT00530000063926.a	1,756	DAND5	DAN domain BMP antagonist family member 5
ENSGT00530000062944.a.a.e.b	1,748	SLC25A18	solute carrier family 25 member 18
ENSGT00760000119204.B.b.b	1,671	TNFRSF14	TNF receptor superfamily member 14
ENSGT00760000118893.b.a.a.b	1,628	TRIM60;TRIM75P	tripartite motif containing 60; 75, pseudogene
ENSGT00700000104221.B.a.b.b	1,410	PKD1L1	polycystin 1 like 1, transient receptor potential channel interacting
ENSGT00910000144387	1,346	RPL41	ribosomal protein L41
ENSGT00390000005707.a.a	1,307	ARL6IP1	ADP ribosylation factor like GTPase 6 interacting protein 1
ENSGT00730000110654.f.b	1.282	MS4A4A;MS4A4E	membrane spanning 4-domains A4A; A4E
ENSGT00760000119023.C.b.a.b	1,243	HS3ST3A1	heparan sulfate-glucosamine 3-sulfotransferase 3A1
ENSGT00390000003592	1,230	C1orf127	chromosome 1 open reading frame 127
ENSGT00770000120884	1,230	SMIM22	small integral membrane protein 22
ENSGT00760000119024.B.a.a.a	1,205	L3MBTL4	L3MBTL4, histone methyl-lysine binding protein
ENSGT00390000006311.a.a.b	1,2	PPP2R2D	protein phosphatase 2 regulatory subunit Bdelta
ENSGT00930000150821.A.a.a.b	1,182	ADGRG4	adhesion G protein-coupled receptor G4
ENSGT00760000118944.a.a.b.a.b	1,141	PSKH2	protein serine kinase H2

Gène amniote ancestral	Score	Gène humain	description
ENSGT00930000150984	1.141	AC067968.1;ZNF284;ZNF155; ZNF230;ZNF222;ZNF223; ZNF224;ZNF225;ZNF221;ZNF234	Zinc fingers proteins
ENSGT00390000008796.b	1.128	LMLN2	leishmanolysin like peptidase 2
ENSGT00760000119005.a.a.b.e.b	1.102		
ENSGT00540000071542.a	1.102	AC022167.5	lipopolysaccharide-induced tumor necrosis factor-alpha factor-like
ENSGT00550000074323.a.a	1.082	CHIT1	chitinase 1
ENSGT00390000014814.a.b	1.076	RPL39	ribosomal protein L39
ENSGT00920000149076.D.b	1.076	KIAA1586	KIAA1586
ENSGT00550000074473.B.b.b.b.b.a.a	1.051	MYL9	myosin light chain 9
ENSGT00760000119034.D.a.b	1.023	S100P	S100 calcium binding protein P
ENSGT00390000016398	1.012	NAT16	N-acetyltransferase 16 (putative)
ENSGT00760000118977.a.a.a.b	1.0	OPN3	opsin 3

Ce classement est cohérent pour les candidats déjà connus (MMP21, DAND5, PKD1L1) qui se retrouvent en haut. En revanche le score décroît rapidement (figure III.2), indiquant une faible similarité des candidats suivants. Sur un tel nombre d'espèces (notamment 39 *Euarchontoglires*), il est en effet attendu d'observer des déviations par rapport à la répartition recherchée. Sur une échelle de -3 à 2, seuls une vingtaine d'arbres sur les 19 096 dépasse 1. Cette décroissance rapide est due à une augmentation des pertes dans les clades où on ne les attend pas, en particulier chez les rongeurs.

Lorsqu'on restreint la recherche aux protéines possédant un peptide signal — indiquant une voie d'export extra-cellulaire—, comme suggéré par nos collaborateurs, la liste se retreint à MMP21, TNFRSF14, C1orf127, LMLN2 et CHIT1. Parmi ceux-ci TNFRSF14 et CHIT1 sont présents chez le lézard et surtout présentent des dizaines de paralogues dans chaque espèce ce qui en fait de faux positifs possibles.

C1orf127 et LMLN2 sont connus de nos collaborateurs qui ont pu démontrer leur implication dans la latéralisation de l'embryon (en cours de publication).

Ceux-ci n'ont pas exactement la même répartition (voir annexe A11) phylogénétique : C1orf127 est également absent des carnivores (il est pseudogénéisé chez le cheval et le chien), à la différence de MMP21. Quant à LMLN2, il présente une similarité peu convaincante avec MMP21 : en effet il est absent de tous les rongeurs sauf rat et souris ; on aurait donc pu penser que les pertes chez *Sauria* étaient dues au simple hasard, ce gène ayant une certaine propension à être perdu dans l'ensemble de la phylogénie. À l'opposé, et puisque le rôle de LMLN2 a été confirmé expérimentalement par E. SZENKER, B. RÉVERSADE et collègues, je suppose qu'il n'a pas été correctement annoté ou séquencé dans les rongeurs d'Ensembl mais que la séquence existe bel et bien.

L'article soumis par nos collaborateurs SZENKER-RAVI et al. (p. d.) contient les

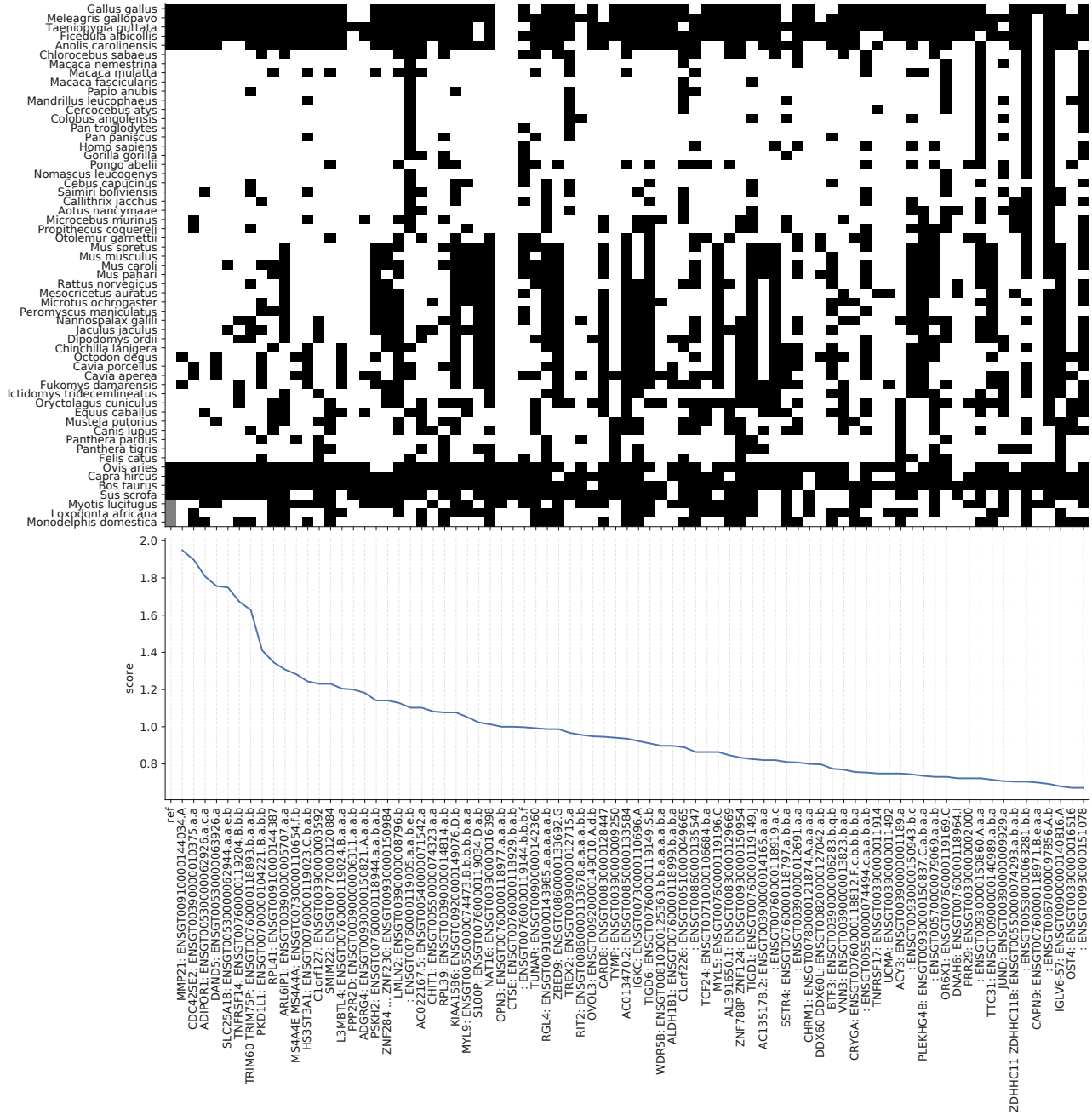


FIGURE III.2 – Top 80 des arbres triés : en haut, motif de présence/absence (blanc/noir respectivement); en bas le score des arbres. La 1re colonne contient la répartition de référence. Les 3 espèces en gris ne sont pas prises en compte dans le score.

résultats ci-dessus comme prémices à l'identification de gènes impliqués dans la voie de latéralisation, et rapporte la caractérisation expérimentale chez le poisson-zèbre du rôle de C1orf127 et LMLN2, rebaptisés ALED (ALENDROIT) et TDT (TOUT-DE-TRAVERS) ; ils ont ainsi pu mettre en évidence que TDT s'exprime précisément dans l'Organiseur Gauche-Droite des embryons poisson-zèbre, xénope et souris, et que son Knock-Out chez le poisson-zèbre cause des altérations de la latéralisation. Ils montrent également le positionnement relatif de ces acteurs dans la voie de régulation, en rapportant que ces altérations peuvent être partiellement compensées par une sous-régulation de DAND5, et que TDT agit en amont de MMP21, indépendamment de la ciliogénèse.

19.2 Classement alternatif : corrélations phylogénétiques

Le score précédent permet de trier les gènes selon un critère arbitraire, et sans validation statistique. J'ai donc ici mis en place deux méthodes comparatives phylogénétiques existantes, adaptées aux traits binaires.

19.2.1 Test de répartition des événements de pertes

J'ai implémenté la méthode de MADDISON (1990) qui teste si les gains et pertes d'un caractère binaire sont plus concentrés dans une région de l'arbre. Il s'agit d'un test non-paramétrique qui calcule le nombre de combinaisons possibles de répartition des événements entre les branches. Dans notre cas il est simplifié puisqu'on ne s'intéresse qu'aux pertes, et que celles-ci sont irréversibles. Par rapport au score précédent, cette méthode présente l'avantage de considérer l'arbre phylogénétique.

Avec cette méthode, MMP21, DAND5, C1orf127, PKD1L1 sont classés en positions 2158, 4649, 8041, 9822, 11067 respectivement, ce qui ne correspond pas à l'attendu. La raison en est que par construction, cette méthode va privilégier les grands nombres d'événements dans les clades *Sauria* et *Artiofabula*. Comme ici les pertes sont irréversibles, les arbres les mieux classés sont ceux pour lesquels une séquence par clade à été conservée. Si cette répartition n'est pas l'attendue, les candidats ainsi obtenus peuvent cependant se révéler intéressants s'il s'agit de véritables pertes multiples, ou si au contraire l'unique orthologue détecté est en fait un pseudogène. Il est possible d'adapter la méthode pour éviter ce problème, en limitant la zone focale à quelques branches autour d'un nœud, sans descendre jusqu'aux feuilles. Dans ce cas, le calcul de probabilité correspondrait davantage à ce que l'on cherche, mais la définition de la zone focale resterait quand même arbitraire.

19.2.2 Test de corrélation des taux de pertes en temps continu

La méthode actuelle canonique concernant ce type de problème a été formalisée par PAGEL (1994), et modélise l'évolution du trait binaire grâce à une chaîne de Markov en temps continu. Pour l'étude de deux caractères binaires dépendants, la chaîne de Markov est paramétrée par huit taux de transitions possibles, que l'on réduit à quatre car les gains sont impossibles (cf. Méthodes 23.4.3).

En triant par ratio de vraisemblance, c'est-à-dire degré de significativité, on obtient ainsi MMP21, DAND5, PKD1L1, C1orf127 et LMLN2 en positions 1, 3, 12, 61 et 1344 respectivement (tableau III.2). Encore une fois, LMLN2 pâtit ici de son absence dans la plupart des rongeurs. La méthode de Pagel détecte à raison un taux élevé en dehors des clades focaux, ce qui rend le différentiel de taux de perte trop faible, et disqualifie donc les pertes de LMLN2 chez les oiseaux et cétartiodactyles. Cependant, cette stringence sur le différentiel de perte a pour effet de favoriser des gènes très peu perdus hors des clades focaux, comme WDR55 qui est classé deuxième, malgré sa présence dans l'ensemble des cétartiodactyles et chez le poulet.

Gène	LRV	$q_1 - q_4$	$q_2 - q_3$	p-valeur	p-valeur corrigée
MMP21	7,11	99,99745	0,030	0,0008	0,50
DAND5	5,71	99,99682	0,017	0,0033	0,78
PKD1L1	4,90	99,98672	0,013	0,0075	0,96
C1orf127	3,98	99,99516	0,009	0,0186	0,96
LMLN2	1,48	99,98470	0,004	0,2281	1,00

TABLE III.2 – Paramètres estimés et statistiques du test `corDISC`. LRV : Log-ratio de vraisemblances ; $q_1 - q_4$: augmentation du taux de perte dans la zone focale ; $q_2 - q_3$: augmentation du taux de perte de la référence lorsque le gène comparé est perdu. Les ratios q_1/q_4 sont de l'ordre de 10^4 et les q_3 estimés à zéro. Un test à un seul degré de liberté (en imposant $q_2 = q_3$) donne un classement en LRV similaire, si ce n'est que PKD1L1 est relégué derrière C1orf127.

20 Sélection d'enhancers conservés

Les données de *Pegasus* générées par l'équipe établissent une liste d'environ deux millions *Conserved Non-coding Elements* (CNEs) à travers les vertébrés, sur la base de conservation synténique des gènes cibles, et leur présence-absence chez 35 espèces dont 33 amniotes (CLÉMENT et al. 2020).

20.1 Classement par clades de références

J'applique le même système de score que précédemment (19.1) sur ce nouveau jeu d'espèces, permettant d'obtenir la tête de liste en tableau III.3.

TABLE III.3 – CNEs les mieux classés, avec leurs cibles conservées

CNE	Score	Gènes cibles	chromosome	début	fin
chr13_23611	2.0		chr13	64190485	64190495
chr13_38667	2.0		chr13	90365669	90365696
chr13_38668	2.0		chr13	90365820	90365836
chr13_40295	2.0	GPC5	chr13	93211043	93211064
chr1_101679	2.0	RGS7	chr1	240872540	240872738
chr8_44042	1.923	TMEM55A OSGIN2 NBN DECR1 OTUD6B TMEM64 RIPK2 NECAB1 CALB1	chr8	91472993	91473002
chr6_66364	1.846	VNN3 VNN1 VNN2 STX7 CTGF RPS12 MOXD1 SLC18B1	chr6	132967948	132968077

CNE	Score	Gènes cibles	chromosome	début	fin
chr17_12565	1.846	DDX52 DHRS11 AATF MYO19 HNF1B TADA2A LHX1 DUSP14 ACACA PSMC3 KBTBD4 SLC39A13 PACSIN3 FNBP4 HARBI1 C11orf49 CREB3L1 NDUFS3 ATG13 CKAP5 DGKZ AMBRA1 F2 ARHGAP1 LRP4 RAPSN ARFGAP2 ACP2 NR1H3 MADD DDB2 MYBPC3	chr17	35639584	35639601
chr4_41348	1.846	HERC6 HERC5 PYURF PPM1K ABCG2 FAM13A	chr4	89307003	89307021
chr5_38296	1.846	ATP6AP1L SSBP2 ZCCHC9 ACOT12 RPS23	chr5	80906274	80906294
chr9_19623	1.846	RUSC2 RECK TMEM8B HINT2 FANCG RGP1 KIAA1045 VCP FAM214B GLIPR2 MSMP CD72 NPR2 CREB3 TLN1 GBA2 TESK1 PIGO UNC13B	chr9	35451769	35451779
chr5_69686	1.846	SLC4A9 ARAP3 DND1 FCHSD1 PCDHB1-8, 10,11,13-16 HDAC3 APBB3 ZMAT2 HARS2 HARS TMCO6 WDR55 PFDN1 RELL2 HBEGF	chr5	140478255	140478265
chr4_45796	1.846	STPG2	chr4	97711059	97711098
chr7_10253	1.833	ITGB8 ABCB5 MACC1	chr7	20582557	20582580
chr20_18088	1.833	ZHX3 PTPRT CHD6 PLCG1 EMILIN3 LPIN3	chr20	40175561	40175585
chr2_79934	1.833	ERMN ACVR1C GALNT5 ACVR1 GPD2 CYTIP	chr2	158002414	158002431
chr17_6332	1.833	MYO15A ATPAF2 SHMT1 RAI1 LRRC48 MED9 PEMT LLGL1 MIEF2 TOP3A DRG2 FLII TOM1L2 NT5M	chr17	17800640	17800669
chr22_7222	1.769	HMGXB4 MCM5 ISX HMOX1 RASD2 TOM1	chr22	35804004	35804014
chr12_28323	1.769	LRIG3	chr12	59257614	59257632
chr11_24398	1.769	OSBP PATL1	chr11	59277674	59277693
chr4_86345	1.769	GALNTL6	chr4	172778495	172778649
chr6_74883	1.769	PPIL4 IYD ZC3H12D UST LATS1 GINM1 LRP11 NUP43 TAB2 KATNA1	chr6	150175014	150175033

CNE	Score	Gènes cibles	chromosome	début	fin
chr16_27968	1.769	MON1B NUDT7 ADAMTS18 VAT1L CLEC3A	chr16	77532819	77532840
chr12_18257	1.769	CPNE8 ALG10B			
chrX_79152	1.769	ZNF449 SLC9A6 DDX26B SAGE1 CT45A6 CT45A4 CT45A3 CT45A1 CT45A2 CT45A5 ZNF75D	chrX	134419082	134419096
chr1_70583	1.756	PAPPA2 BRINP2 ASTN1 RFWD2	chr1	176818964	176818979
chr4_91350	1.756		chr4	181117723	181117747
chr18_6753	1.756	MC2R	chr18	13239472	13239538
chr18_6752	1.756	MC2R	chr18	13239369	13239419
chr9_29607	1.756	TRPM6	chr9	77234065	77234080
chr9_29607	1.756	RORB			
chr1_65871	1.756	MGST3 ALDH9A1 TMCO1	chr1	166330137	166330149
chr8_44043	1.756	TMEM55A OSGIN2 NBN DECR1 OTUD6B TMEM64 RIPK2 NECAB1	chr8	91473101	91473116

Étant donné le nombre initial de CNEs, il est probable d'obtenir avec cette méthode une majorité de CNEs n'ayant aucun lien fonctionnel avec la latéralisation de l'embryon. Les cinq premiers CNEs ont la répartition exacte recherchée, ce qui est explorable manuellement. Les trois premières n'ont cependant pas de cible conservée à travers les vertébrés, reste donc la cible GPC5 (Glypican 5, qui pourrait jouer un rôle dans la division cellulaire et la régulation de la croissance) et RGS7 (un régulateur de protéine G).

20.2 Restriction aux CNEs dans le voisinage de gènes candidats connus

Pour extraire une information moins bruitée de ce classement, nous l'avons restreint aux CNEs dans le voisinage (moins de 1,2 Mb) de 29 gènes connus ou soupçonnés d'avoir un rôle dans la latéralisation, d'après les sources bibliographiques fournies par nos collaborateurs Jeanne AMIEL et Chris GORDON. Cette intersection capture un CNE à 30 kb du gène SESN1 avec un score de 1,6, tandis que les suivants ont un score en dessous de 1,45 : ces CNEs voisins ont donc une faible corrélation phylogénétique avec la répartition de référence.

21 Restriction des candidats par similarité fonctionnelle

Les sélections précédentes de gènes et CNEs candidats indique un faible niveau d'association phylogénétique, de surcroît confondue par la possibilité de nom-

breux "faux positifs", c'est-à-dire des éléments bien corrélés mais n'ayant aucun lien fonctionnel.

Afin de détecter des candidats à la fonction pertinente, nous mesurons donc leur similarité avec les termes GO des gènes "contrôle" MMP21, DAND5 et PKD1L1, et trions selon ce nouveau critère tous les gènes/CNEs de score phylogénétique supérieur ou égal à 1. Notons que les gènes impliqués dans la latéralisation mais non identifiés comme tels ne bénéficient pas encore de l'annotation de fonction correspondante. La meilleure illustration en est "Polycystic Kidney Disease Protein 1-Like 1", alias PKD1L1, dont le premier rôle connu a donc été en lien avec une néphropathie. Malgré l'existence de cette pléiotropie trompeuse, notre démarche se base donc sur l'hypothèse que les gènes non découverts possèdent tout de même certaines autres fonctions en commun avec les contrôles.

En résultat, les gènes candidats ont des similarités GO inférieures à 0,7 (sur une échelle de 0 à 1) et le plus similaire (avec DAND5) est CDC42SE2, qui était classé deuxième par score sur la présence du gène (section III.19). Les cibles de CNEs retenues sont quant à elles plus nombreuses à être convaincantes, en particulier pour leur similarité fonctionnelle avec DAND5 (figure III.3) : les quatre premières cibles dans ce tri, TESK1, APBB3, FAM13A et HBEGF, avec un score phylogénétique supérieur à 1,8, mériteraient une vérification manuelle des raisons de leur similarité de fonction avec DAND5.

22 Discussion

Dans ce chapitre, j'ai pu expérimenter l'application de méthodes phylogénétiques pour l'étude d'un système particulier, la voie développementale menant à la symétrie gauche-droite des vertébrés. À partir d'un gène déjà connu ayant été perdu deux fois indépendamment, j'ai cherché des gènes ou des CNEs répartis de la même manière, qui pourraient être ensuite vérifiées expérimentalement, ou médicalement grâce aux génomes de patients atteints d'hétérotaxie.

22.1 Choisir une méthode évolutive pour représenter l'interdépendance entre branches

Tout d'abord, un simple score sur les présences-absences par espèce a révélé de nombreux gènes et CNEs similaires, incluant les gènes déjà validés antérieurement PKD1L1 et DAND5. Dans un cadre exploratoire, cette méthodologie devient vite limitante : en effet, ce score ajuste mal le mécanisme évolutif à l'origine des pertes, et ne fournit donc pas d'arguments pour juger de la pertinence des gènes les mieux classés. Il ne permet par exemple pas de mesurer la probabilité que le scénario de référence se produise, et ne permet pas non plus de proprement quantifier l'écart au scénario de référence. En fait, ce score revient à considérer indépendantes les espèces au sein d'un clade, une stratégie que les évolutionnistes évitent depuis Felsenstein (1985). Un ajustement *ad hoc* que nous avons utilisé, de pondérer les clades selon les longueurs de branches, répond en partie à ce problème. Dans la même ligne d'idées, il aurait été possible de "peaufiner" encore ce score,

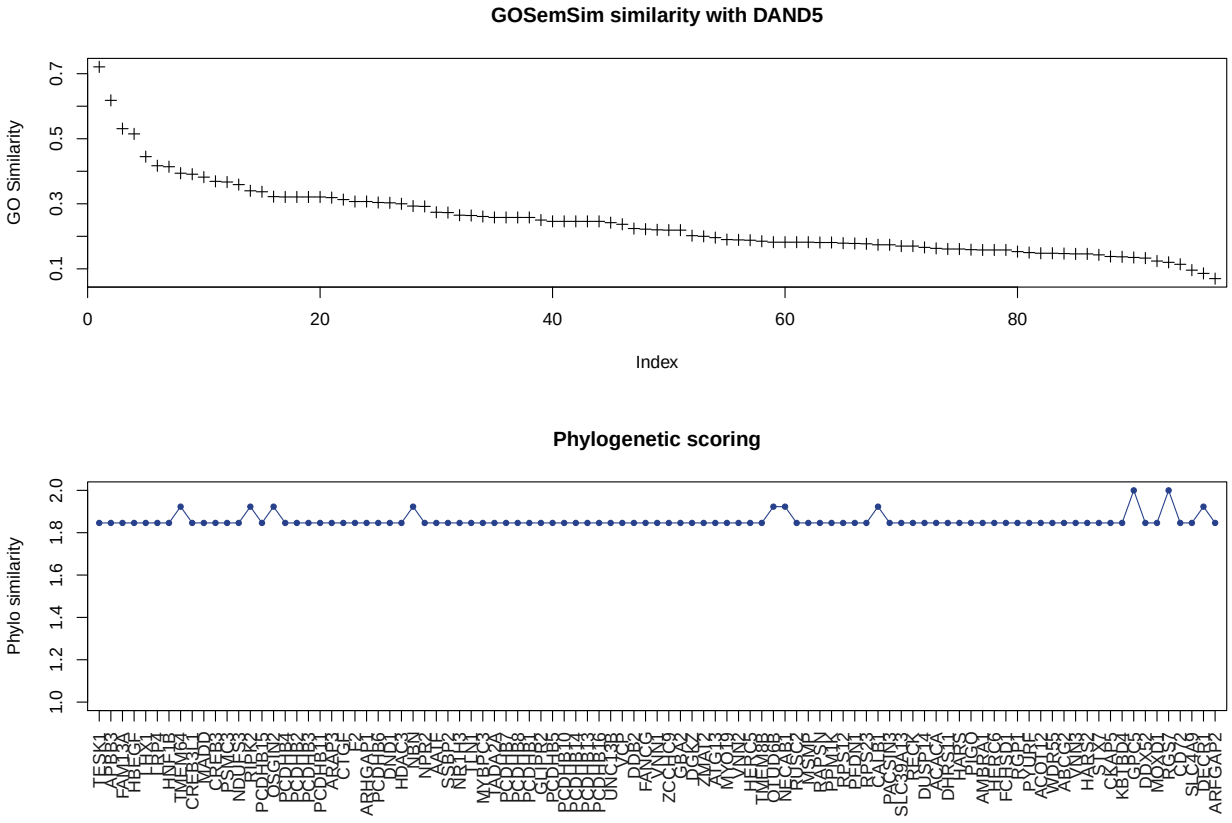


FIGURE III.3 – Similarité fonctionnelle avec DAND5 des cibles de CNEs les plus corrélés phylogénétiquement (en haut). Le graphe du bas partage la même abscisse et rapporte le score phylogénétique de présence-absence du CNE correspondant.

en mesurant les proportions moyennes de perte (ou absence d'annotation) par espèce, tous gènes confondus, ou en prenant en compte pour les CNEs le fait que la séquence humaine a servi de référence lors de l'alignement multiple. Ces stratégies requièrent une supervision manuelle et des hypothèses peu explicites.

Pour cette raison, j'ai imité mes précurseurs phylogénéticiens qui un jour sont sortis des eaux de la statistique par espèce et ont conquis la terre ferme des dépendances évolutives. Il faut cependant noter qu'appliquer la *méthode comparative phylogénétique* à ce problème particulier présente encore des difficultés : ici, les traits sont binaires et les changements irréversibles. Cela exclut les méthodes adaptées aux traits continus comme les moindres carrés généralisés phylogénétiques (PGLS). Maddison lui-même a publié "The unsolved challenge to phylogenetic correlation tests for categorical characters" (2015), visant notamment sa propre méthode (1990) et celle de PAGEL (1994), qui alerte sur nombre de difficultés toujours d'actualité dans ce cadre précis, que nous pouvons illustrer ici.

Premièrement, certaines de ces difficultés sont évitables en choisissant le bon jeu de données. Ainsi, si l'on choisit un trait qui change une seule fois dans tout l'arbre, tester sa corrélation avec d'autres traits n'a pas de sens car une multitude d'autres facteurs a pu apparaître de façon contingente dans le même clade : il devient impossible de dissocier quel caractère spécifique au clade est fonctionnellement lié au trait testé [3]. Or, par construction, la méthode de Pagel détecte ce genre de corrélation. Notre cas est limite, avec seulement deux transitions indépendantes dans l'arbre, mettant sous caution les associations découvertes.

Deuxièmement, la méthode phylogénétique de MADDISON (1990) a montré ici des limites spécifiques au problème en question : considérant des événements irréversibles, les pertes, comptabiliser leur nombre n'est pas forcément approprié. Il faudrait *a minima* ajuster la zone focale, par exemple en ne retenant que les branches autour des nœuds de pertes, mais cela entraîne les questions suivantes : 1) qui est le plus corrélé, entre *une* perte simultanée et *deux* pertes en décalé ? 2) Sur quelle distance peut-on considérer les branches "corrélées", au sens de partageant des mêmes facteurs écologiques et développementaux causant les transitions recherchées ? Il semblerait pertinent de prendre ici en compte les covariances *entre branches*. Cette notion de "pseudo-réplication des facteurs spécifiques à la lignée" est décrite par Maddison dans son article critique et à ma connaissance n'a pas fait l'objet de développements méthodologiques, en tout cas sous ce nom [4]. Peut-être quelques connaissances sur les méthodes d'autocorrélations spatiales (appliquées aux graphes) apporteraient des solutions. Il faudrait alors résoudre la difficulté imposée par le lien de causalité : si le gène 1 est nécessaire pour le gène 2, alors les pertes doivent toujours être observées dans le sens 1 puis 2. Notons que nous n'avons pas eu le temps dans cette étude

3. Maddison prend l'exemple de l'apparition "corrélée" du poil et des trois os de l'oreille moyenne. Ils n'apparaissent qu'une fois, chez les mammifères, pourtant nous ne concluons pas qu'il y a un lien fonctionnel.

4. Maddison cite néanmoins la méthode de la variable latente brownienne (FELSENSTEIN 2012), qui modélise donc cette covariance temporelle. Cependant elle me semble difficilement s'appliquer ici, les transitions étant irréversibles.

de mettre en place un modèle de markov avec état caché (avec `corHMM`), mais cela me semble approprié pour mieux représenter la phase de transition des deux gènes, celle-ci étant justement incertaine (chez *Artiofabula* pour MMP21, *Fereungulata* pour C1orf127).

22.2 L'improbabilité de deux pertes similaires dans des données hautement dimensionnelles

Le troisième point souligné par Maddison est la difficulté de tester des corrélations avec autant de caractères. Nous passons au crible l'intégralité des arbres amniotes (19 096) et des CNEs conservées (2 millions), posant un évident problème de de tests multiples, et plus concrètement ne prenant pas en compte la probabilité de détecter des gènes similaires sans aucun lien fonctionnel. Ici les données devraient donc idéalement être traitées comme multidimensionnelles. La méthode comparative phylogénétique s'est attaquée à ce problème, par exemple avec l'analyse en composantes principales phylogénétique (Revell 2009) et plus récemment les régressions linéaires multivariées lorsque le nombre de traits est supérieur au nombre d'espèces (Clavel et al. 2015 ; Clavel et Morlon 2020). Malheureusement, alors qu'en statistiques classiques l'analyse factorielle remplace l'ACP pour des variables binaires ou catégorielles, en phylogénie ces méthodes n'ont pas d'équivalent catégoriel. À quand un modèle de seuil multidimensionnel, ou une analyse factorielle phylogénétique ?

À cela s'ajoutent quelques observations qui découlent des gènes "contrôles" sur lesquels j'ai pu tester les corrélations phylogénétiques, en particulier LMLN2. Ce gène n'a d'homologues *glires* que chez la souris et le rat, un parcours qui implique 10 pertes (annexe A11). C'est évidemment un contre-argument fort pour dire que ses pertes sont significatives dans les clades focaux, et en toute cohérence aucune méthode phylogénétique utilisée ici ne le classe bien. Dans Ensembl 100 (avril 2020), ces lacunes sont partiellement comblées, avec 8/28 homologues glires, et 12/17 homologues carnivores. En conséquence de ces incertitudes, il paraît donc nécessaire, quelle que soit la méthode, de croiser les sources de preuves, comme nous l'avons fait en considérant les annotations fonctionnelles.

Par ailleurs, il convient de rappeler que les jeux de données génomiques ne sont pas conçus pour inférer les pertes. Dans Ensembl par exemple, l'utilisateur recherche avant tout des relations d'orthologie-paralogie, des liens avec des séquences régulatrices, etc., mais la présence ou l'absence d'un homologue est une information sujette à caution : il faut en effet vérifier que les absences ne résultent pas d'un problème de séquençage/assemblage/annotation/clustering, et inversement que les présences ne cachent pas un pseudogène. Nous devrions donc envisager de repenser notre procédure en amont de l'obtention des arbres, en effectuant des Blasts systématiques, ainsi qu'en vérifiant la qualité de l'information de synténie. Autrement, une méthode comme Pagel (1994), que l'on a vu sensible au nombre des événements, inférera des taux de pertes incorrects.

En conclusion, au-delà d'un cas d'étude sur la latéralisation des vertébrés, les données en mains se sont révélées complexes au fur et à mesure de l'analyse :

au premier abord, reconstruire des histoires évolutives de gènes individuels, représentés de façon binaire, avec des pertes irréversibles, paraît simple, et en effet, l'algorithme de parcimonie de Dollo adapté à cette situation est intuitif. Pourtant, trois caractéristiques combinées rendent l'analyse de corrélation non triviale : 1) l'irréversibilité, ou tout du moins une certaine unidirectionalité (voir GOLDBERG et IGIĆ 2008, pour une discussion plus large) ; 2) le caractère binaire ; 3) la multidimensionnalité. J'en viens donc à me ranger à l'avis que ce problème est encore irrésolu [5].

L'évolution est un constant remodelage, impliquant l'acquisition de caractères mais également leur perte ; le cas de la latéralisation des vertébrés montre comment des gènes impliqués tôt dans l'embryogénèse peuvent néanmoins devenir accessoires et dégénérer, en de multiples occasions, ce qui amène à reconsidérer la flexibilité développementale dans l'évolution (CHURCH et EXTAVOUR 2020). Plus généralement, les données génomiques de gènes et d'enhancers révèlent que les pertes sont fréquentes. De plus, elles sont hétérogènes entre lignées animales, ainsi les ecdysozoaires ont subi davantage de pertes que les lophotrochozaires, et les protostomes environ 20 fois plus que les deutérostomes (SIMAKOV et al. 2013 ; revu dans ALBALAT et CAÑESTRO 2016). Ces variations sont également affectées par les duplications, en particulier les duplications complètes de génome. Considérant la façon dont les réseaux de régulation de gènes sont cablés, par de multiples voies impliquant des inhibiteurs, eux-mêmes inhibés, la perte d'un tel nœud peut en fait réactiver certaines voies, conférant un "gain" de fonction par perte. Par ailleurs l'évolution par réduction est aussi communément observée dans de nombreux organismes —tels les myxozoaires, ces métazoaires redevenus unicellulaires, et dans la plupart des cas de parasitismes—, suggérant des avantages adaptatifs dus aux pertes de gènes. La question se pose donc de la mesure avec laquelle des contraintes sélectives peuvent se relâcher, voire s'inverser, similairement dans des taxons distincts, et du nombre de gènes affectés par l'obsolescence d'un module fonctionnel.

23 Méthodes

23.1 Présence-absence des gènes dans les arbres Ensembl

La forêt d'arbres de gènes réconciliés d'Ensembl 93 a été utilisée, après suppression des espèces aux génomes mal assemblés (voir section 12.3). La position des duplications peu soutenues a été corrigée pour atteindre le score minimal de 0,35 —comme dans Genomicus 93—, car cela permet de ne conserver que les pertes les plus certaines, celles où aucun homologue, même divergent en séquence, n'existe dans les espèces focales (figure III.4).

Pour convertir la forêt d'arbre réconciliées en présences-absences depuis l'ancêtre *Amniota*, il faut :

5. Même si Joseph FELSENSTEIN, répondant sur la liste "`R-sig-phylo`", déclare « though I'd say that the problem is not as "unsolved" as they think. ».

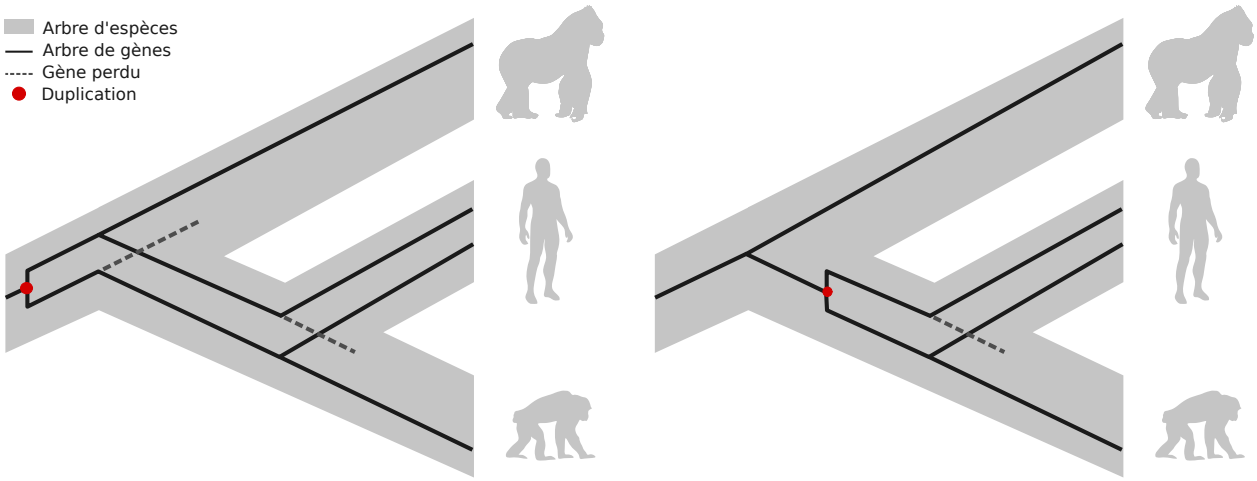


FIGURE III.4 – Correction d'une duplication avec seuil de soutien de 0,35. À gauche, la duplication est parente d'une paralogie chez une espèce descendante sur 3 (soutien de 0,33). Ce genre d'arbre survient lorsque les paralogues semblent très divergents en séquence. À droite, après correction, la duplication a été "poussée" vers le présent, et elle est désormais soutenue par une espèce sur 2 (soutien de 0,5). En conséquence, si l'on regarde le devenir des gènes depuis l'ancêtre commun, l'arbre corrigé n'implique pas de perte chez le gorille.

1. réinsérer les nœuds de spéciations non-observées (par exemple si un arbre a des descendants amphibiens mais pas *Sauria*, le nœud Amniote n'y sera pas observé) ;
2. Extraire tous les arbres passant par *Amniota*. Par souci d'exhaustivité, j'ai extrait à la fois les nœuds de duplication et les nœuds de spéciation, générant ainsi des sous-arbres redondants, mais racinés en aval);
3. comptabiliser les gènes descendants par espèce.

J'ai donc ainsi extrait 19 096 arbres enracinés chez *Amniota*, dont 2 497 n'ont pas de branche humaine.

23.2 Visualisation d'arbres de gènes réconciliés

Les figures en annexe A11 ont été générées avec le script `genetree_drawer.py`, utilisable en ligne de commande et dépendant de `matplotlib`, `Ete3`, et `LibsDyogen`. À partir d'arbres en newick simple ou en NHX avec les champs TreeBest, il permet d'afficher le parcours des lignées géniques par-dessus l'arbre d'espèces. Les nœuds implicites doivent au préalable être reinsérés avec `genchron/prune2family.py`.

D'autres outils de visualisation de réconciliations existent aussi, comme SylVX (CHEVENET et al. 2015), et recphylovisu.

23.3 Présence-absence des CNEs (Pegasus)

Les *Conserved Non-coding Elements* (CNEs) proviennent du jeu de données *Pegasus* généré dans l'équipe par Yves CLÉMENT (2020), à partir d'Ensembl 75. Cette méthode utilise des alignements génomiques multiples produits par LastZ, et interprète la conservation de synténie entre gènes et régions non-codantes pour

prédire 1) l'existance d'un enhancer conservé, et 2) l'association conservée de cet enhancer avec des gènes cibles. Les espèces utilisées ici rassemblent 33 amniotes, la grenouille *Xenopus tropicalis* et le cœlacanthe. Seuls les CNEs présents chez le xénope ou le cœlacanthe sont retenues. Les CNEs sans gène cible humain sont inclus également.

23.4 Métriques de similarité phylogénétique

On cherche ici les arbres les plus similaires à la répartition de MMP21 (figure III.1) :

Arbres de gènes			Arbres de CNEs			Présence
Clade	Espèces	Poids	Clade	Espèces	Poids	
Extragroupe *Neopterygii*	10	-	cœlacanthe/xénope	2	-	oui
Euarchontoglires	39	1		13	1	oui
Zooamata	6	1		5	1	oui
Sauria	5	-1		6	-1	**non**
Artiofabula	4	-2		3	-1	**non**

TABLE III.4 – Choix des clades sur lesquels pénaliser la présence : les présences chez *Sauria* et *Artiofabula* sont pénalisées par un poids négatif.

23.4.1 Score linéaire entre 4 clades

Pour trier les arbres/CNEs selon leur similarité avec la répartition de référence, le score suivant est calculé pour chaque arbre de gènes :

$$\text{Score(arbre)} = p_{Eu} + p_{Zoo} - p_{Sau} - 2p_{Art}$$

où p_{Eu}, p_{Zoo}, p_{Sau}, p_{Art} sont les proportions d'espèces possédant un homologue dans les clades *Euarchontoglires*, *Zooamata*, *Sauria* et *Artiofabula* respectivement. Le choix d'une pénalisation deux fois plus forte pour *Artiofabula* se justifie par le fait que ce clade est beaucoup plus récent que *Sauria*. En effet, les longues branches menant à *Sauria* puis aux oiseaux rendent la probabilité de perte plus élevée, tous gènes confondus.

Pour les CNEs, le calcul a été fait avec une version antérieure du score, où p_{Sau} et p_{Art} ont le même poids de -1.

23.4.2 Probabilité de répartition des événements de Maddison 1990

L'algorithme de MADDISON (1990) compare le nombre de gains/pertes dans la zone focale de l'arbre avec le nombre en dehors.

Pour cela, l'arbre est pris en compte, et les transitions de l'état du caractères doivent être préalablement positionnée sur les branches, par exemple par parcimonie. Dans notre cas, l'état dérivé du caractère est la perte d'un gène, et la réversion vers un état fonctionnel est considérée impossible, ce qui simplifie l'algorithme.

L'algorithme est combinatoire : il comptabilise le nombre de façons de répartir les événements sur les branches de l'arbre, chaque branche étant considérée équiprobablement, avec la contrainte supplémentaire qu'une perte ne peut être positionnée qu'une fois le long d'une lignée entière.

Par programmation dynamique dans le sens feuilles-racine (post-order), on calcule c, le nombre de répartitions possibles de n pertes dans l'arbre entier. De même, on calcule c_a, le nombre de répartitions possibles de a pertes dans les sous-arbres constitués des clades focaux *Artiofabula* et *Sauria*, et $n - a$ pertes en dehors.

On estime alors que la probabilité de a pertes dans la zone focale sachant n pertes au total est c_a/c, et l'on en déduit la p-valeur en sommant cette probabilité sur toutes les répartitions de plus de a pertes dans la zone focale.

Mon implémentation en Python et Numpy (`https://www.github.com/DyogenIBENS/Phylorgs/dendro/dollocorr.py`) gère les polytomies et les nœuds à un seul descendant (ce qui permet de discrétiser les longues branches en plusieurs segments). Le calcul a pris environ une demi journée sur un seul processeur pour les 19 096 arbres, ce qui pour donner une idée approximative, est deux ou trois fois plus rapide que la fonction `corDISC` du package R `corHMM` (voir partie suivante). Il n'est donc pas nécessaire de recourir à l'approximation également développée dans l'article original, du moins avec un faible nombre d'espèces comme le nôtre (57 amniotes).

23.4.3 Modèle d'évolution de deux traits binaires en temps continu

Le modèle général décrit par PAGEL (1994) est implémenté en R par maximum de vraisemblance avec la fonction `corDISC` du package `corHMM` (BEAULIEU et al. 2013). Deux calculs doivent être effectués : 1) sous le modèle *dépendant*, où les deux traits évoluent de façon corrélée ; 2) sous le modèle *indépendant*, ou modèle 0, cas particulier du premier.

Les matrices de transitions doivent être spécifiées en entrées. Dans notre cas avec deux traits binaires, il y a quatre états possibles : (0,0), (0,1), (1,0) et (1,1), et comme les gains sont impossibles, les taux de gain sont nuls. Comme habituellement dans les modèles en temps continu, les doubles transitions simultanées ont des taux nuls, ce qui produit les matrices suivantes, les lignes représentant les états de départ :

Matrice de transition avec dépendance

	(0,0)	(0,1)	(1,0)	(1,1)
(0,0)				
(0,1)	q_1			
(1,0)	q_2			
(1,1)		q_3	q_4	

Matrice de transition avec indépendance

	(0,0)	(0,1)	(1,0)	(1,1)
(0,0)				
(0,1)	q'_1			
(1,0)	q'_2			
(1,1)		q'_2	q'_1	

Dans la matrice à taux indépendants, seuls deux paramètres sont donc estimés, tandis que quatre doivent l'être dans le cas dépendant. Un test de ratio de vraisemblance à deux degrés de liberté est effectué pour obtenir les p-valeurs,

corrigées par méthode de Benjamini-Hochberg, et uniquement les arbres pour lesquels $q_1 > q_4$ ou $q_2 > q_3$ sont sélectionnés, puis classés par leur log-ratio de vraisemblance.

J'ai appliqué ces calculs entre la répartition de référence et chaque gène. Le modèle recalcule les états ancestraux, y compris pour la répartition de référence, que nous connaissons déjà, mais ce n'est pas un problème car les états ancestraux sont non ambigus du fait du taux de gain nul.

Cette modélisation présente l'intérêt que la dépendance est testée dans les deux sens : soit l'évolution du gène dépend de l'état du gène de référence, soit c'est la référence qui dépend du gène.

`corDISC` a été exécutée avec les paramètres supplémentaires `ntraits=2`, `root.p=c(0,0,0,1)` (tous les gènes présents à la racine), `model="ARD"` ("all rates differ") et `node.states="marginal"`, avec `corHMM` version 1.22, `APE` 5.3 sous R 3.4.4 (2018-03-15).

Conclusion

Le point de départ de ma thèse a été la question du lien entre génome et diversification. Comment l'évolution du génome influence-t-elle la diversification, et vice-versa ? Les échantillonnages génomiques étant encore bien plus parcellaires que les échantillonnages taxonomiques, trouver l'ancrage entre ces deux sources de données demande de nouvelles méthodes.

Les duplications et pertes peuvent avoir des conséquences sur l'évolution organismale, en autorisant des adaptations, ou en générant des incompatibilités génétiques. À grande échelle cependant, il s'agit d'un processus continu, au rythme soutenu, qui pourrait ne révéler aucun lien avec l'adaptation. Pour en savoir plus, caractériser leur dynamique temporelle est une première étape. Nous avons ici montré que l'approche par horloge moléculaire s'y prête mal, en l'absence d'intégration plus poussée des données génomiques. Il semble plus réalisable en pratique de proposer des hypothèses sur le taux de duplication, que l'on pourrait tester avec les méthodes phylogénétiques comparatives adaptées.

24 Avenues de développement des méthodes phylogénétiques comparatives

Malgré l'accroissement des données séquences, combler la discordance entre données génomiques et taxonomique est encore à l'ordre du jour. Il est donc nécessaire d'imaginer comment traiter les espèces et branches manquantes en phylogénie, au moyen d'imputation par exemple.

Dans le cadre des traits binaires irréversibles (chapitre III), je fais écho au besoin de nouvelles méthodes comparatives prenant en compte la non-indépendance des facteurs spécifiques à la lignée. Cela pourrait s'appliquer plus généralement, par exemple avec l'évolution des taux de duplications en lien avec les taux de diversification (chapitre II, section II.16).

Enfin, les modes de variation des taux de duplication doivent encore être élucidés, notamment afin de mesurer la part relative de neutralité et de sélection, ce qui nécessite d'intégrer de multiples informations, parfois difficiles à manipuler : fonction des gènes, structure chromosomique, tailles de populations, etc.

25 Enjeux : utiliser plus complètement l'information présente dans les génomes

Cette thèse a soulevé des limites de résolution des données utilisées : faible nombre de sites, peu de substitutions entre espèces proches, erreurs d'alignements, erreurs de construction d'arbres. Concernant l'évolution génique, il est pourtant possible de prendre un point de vue plus global et faire appel à l'intégration d'information. Si les simplifications initiales sont utiles, pour répondre à certaines questions il est nécessaire de faire usage de l'ensemble du contenu informationnel génomique. La recherche en génomique et en spéciation prend naturellement cette direction, et en tirera très certainement des informations passionnantes. En effet, les génomes restent encore des univers majoritairement inexplorés. Malgré les séquençages d'une diversité croissante d'organismes, les données ne proviennent que d'une fraction infinitésimale des organismes existants. Pourtant cette fraction ridicule représente déjà une quantité d'information si grande que nous n'en avons que gratté la surface. Dans les génomes bien connus, la proportion de séquence "fonctionnelle" fait débat. Parmi les fonctions annotées, on maîtrise cependant mal les interactions entre composants, sans parler de la relation génotype-phénotype. Or, les génomes eucaryotes sont très partiellement annotés, souvent par extrapolation grâce aux homologies de séquence. Surtout, la plupart des scénarios évolutifs n'ont été obtenus que par et pour les séquences géniques orthologues, et un petit nombre d'entre elles. L'évolution des zones intergéniques, et de la structure même des génomes, recèle pourtant de nombreuses informations, pour peu que l'on puisse la reconstruire. Il y a donc de vastes abysses génomiques à explorer.

Bibliographie

Articles de la thèse

LOUVEL, Guillaume et Hugues ROEST CROLLIUS (p. d.). « Factors influencing the accuracy in dating single gene trees ». biorxiv : 10.1101/2020.08.24.264671v1. Soumis (cf. p. 26, 62).

SZENKER-RAVI, Emmanuelle et al. (p. d.). « A functional operon delineates an extracellular pathway that controls Left-Right patterning only in animals with a ciliated organizer ». Soumis (cf. p. 85).

Références

AGUILETA, Gabriela, Joseph P. BIELAWSKI et Ziheng YANG (2006). « Evolutionary rate variation among vertebrate β globin genes : Implications for dating gene family duplication events ». In : *Gene* 380.1, p. 21-29. DOI : 10.1016/j.gene.2006.04.019 (cf. p. 39).

ALBALAT, Ricard et Cristian CAÑESTRO (2016). « Evolution by gene loss ». In : *Nature Reviews Genetics* 17.7, p. 379-391. DOI : 10.1038/nrg.2016.39 (cf. p. 95).

ARVESTAD, L. et al. (2003). « Bayesian gene/species tree reconciliation and orthology analysis using MCMC ». In : *Bioinformatics* 19 (Suppl 1), p. i7-i15. DOI : 10.1093/bioinformatics/btg1000 (cf. p. 23).

BEAULIEU, Jeremy M., Brian C. O'MEARA et Michael J. DONOGHUE (2013). « Identifying Hidden Rate Changes in the Evolution of a Binary Morphological Character : The Evolution of Plant Habit in Campanulid Angiosperms ». In : *Systematic Biology* 62.5, p. 725-737. DOI : 10.1093/sysbio/syt034 (cf. p. 74, 98).

BEAULIEU, Jeremy M. et Brian C. O'MEARA (2016). « Detecting Hidden Diversification Shifts in Models of Trait-Dependent Speciation and Extinction ». In : *Systematic Biology* 65.4, p. 583-601. DOI : 10.1093/sysbio/syw022 (cf. p. 74).

BLETSA, Magda et al. (2019). « Divergence dating using mixed effects clock modelling : An application to HIV-1 ». In : *Virus Evolution* 5.2. DOI : 10.1093/ve/vez036 (cf. p. 20, 40).

BLUM, Martin et al. (2014). « The evolution and conservation of left-right patterning mechanisms ». In : *Development* 141.8, p. 1603-1613. DOI : 10.1242/dev.100560 (cf. p. 82).

BLUM, Michael G. B. et Olivier FRANÇOIS (2005). « On statistical tests of phylogenetic tree imbalance : the Sackin and other indices revisited ». In : *Mathematical Biosciences* 195.2, p. 141-153. DOI : 10.1016/j.mbs.2005.03.003 (cf. p. 57).

BOISSINOT, Stéphane, Pascale CHEVRET et Anthony V. FURANO (2000). « L1 (LINE-1) Retrotransposon Evolution and Amplification in Recent Human History ». In : *Molecular Biology and Evolution* 17.6, p. 915-928. DOI : 10.1093/oxfordjournals.molbev.a026372 (cf. p. 20).

BOUCKAERT, Remco et al. (2019). « BEAST 2.5 : An advanced software platform for Bayesian evolutionary analysis ». In : *PLOS Computational Biology* 15.4, e1006650. DOI : 10.1371/journal.pcbi.1006650 (cf. p. 19, 27, 52).

BRADLEY, Robert K. et al. (2009). « Fast Statistical Alignment ». In : *PLOS Computational Biology* 5.5, e1000392. DOI : 10.1371/journal.pcbi.1000392 (cf. p. 27, 46).

BRIDGES, Calvin B (1936). « The BAR "gene" a duplication ». In : *Science* 83.2148, p. 210-211. DOI : 10.1126/science.83.2148.210 (cf. p. 21).

BRITTEN, R. J. (1986). « Rates of DNA sequence evolution differ between taxonomic groups ». In : *Science* 231.4744, p. 1393-1398. DOI : 10.1126/science.3082006 (cf. p. 18).

BRITTON, Tom et al. (2002). « Phylogenetic dating with confidence intervals using mean path lengths ». In : *Molecular Phylogenetics and Evolution* 24.1, p. 58-65. DOI : 10.1016/S1055-7903(02)00268-3 (cf. p. 19, 27, 48, 72).

BRITTON, Tom et al. (2007). « Estimating Divergence Times in Large Phylogenetic Trees ». In : *Systematic Biology* 56.5. Sous la dir. de Frank ANDERSON, p. 741-752. DOI : 10.1080/10635150701613783 (cf. p. 50).

BROMHAM, Lindell (2019). « Six Impossible Things before Breakfast : Assumptions, Models, and Belief in Molecular Dating ». In : *Trends in Ecology & Evolution* 34.5, p. 474-486. DOI : 10.1016/j.tree.2019.01.017 (cf. p. 20).

BROMHAM, Lindell et al. (2000). « The Power of Relative Rates Tests Depends on the Data ». In : *Journal of Molecular Evolution* 50.3, p. 296-301. DOI : 10.1007/s002399910034 (cf. p. 20).

BRYANT, David, Nicolas GALTIER et Marie-Anne POURSAT (2005). « Likelihood Calculation In Molecular Phylogenetics ». In : *Mathematics of evolution and phylogeny*. Oxford University Press, p. 33-62. ISBN : 0-19-856610-7 (cf. p. 14).

BRYSON, Vernon et Henry J. VOGEL (1965). « Evolving Genes and Proteins ». In : *Science* 147.3653. Discussion de E. MAYR sur l'horloge moléculaire, p. 68-71. ISSN : 0036-8075 (cf. p. 18).

BURBRINK, Frank T. et R. Alexander PYRON (2008). « The Taming of the Skew : Estimating Proper Confidence Intervals for Divergence Dates ». In : *Systematic Biology* 57.2, p. 317-328. DOI : 10.1080/10635150802040605 (cf. p. 20).

CASTRESANA, J. (2000). « Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis ». In : *Molecular Biology and Evolution* 17.4, p. 540-552. DOI : 10.1093/oxfordjournals.molbev.a026334 (cf. p. 54).

CAVALLI-SFORZA, LL et AWF EDWARDS (1963). « The reconstruction of evolution ». In : *Ann. Hum. Genet* 27, p. 105-106 (cf. p. 12).

CHEVENET, François et al. (2015). « SylvX : A viewer for phylogenetic tree reconciliations ». In : *Bioinformatics*. DOI : 10.1093/bioinformatics/btv625 (cf. p. 96).

CHURCH, Samuel H. et Cassandra G. EXTAVOUR (2020). « Null hypotheses for developmental evolution ». In : *Development* 147.8. DOI : 10.1242/dev.178004 (cf. p. 95).

CLAVEL, Julien, Gilles ESCARGUEL et Gildas MERCERON (2015). « mvMORPH : an R package for fitting multivariate evolutionary models to morphometric data ». In : *Methods in Ecology and Evolution* 6.11. Sous la dir. de Timothée POISOT, p. 1311-1319. DOI : 10.1111/2041-210X.12420 (cf. p. 94).

CLAVEL, Julien et Hélène MORLON (2020). « Reliable Phylogenetic Regressions for Multivariate Comparative Data : Illustration with the MANOVA and Application to the Effect of Diet on Mandible Morphology in Phyllostomid Bats ». In : *Systematic Biology*. corrected proof. DOI : 10.1093/sysbio/syaa010. Sous presse (cf. p. 94).

CLÉMENT, Yves et al. (2020). « Enhancer–gene maps in the human and zebrafish genomes using evolutionary linkage conservation ». In : *Nucleic Acids Research* 48.5, p. 2357-2371. DOI : 10.1093/nar/gkz1199 (cf. p. 83, 88, 96).

COBB, Matthew (2017). « 60 years ago, Francis Crick changed the logic of biology ». In : *PLOS Biology* 15.9, e2003243. DOI : 10.1371/journal.pbio.2003243 (cf. p. 9).

COMTE, Nicolas et al. (2019). « Treerecs : an integrated phylogenetic tool, from sequences to reconciliations ». In : *bioRxiv*, p. 782946. DOI : 10.1101/782946 (cf. p. 23).

CONRAD, Donald F. et al. (2010). « Origins and functional impact of copy number variation in the human genome ». In : *Nature* 464.7289, p. 704-712. DOI : 10.1038/nature08516 (cf. p. 21).

CROTTY, Stephen M. et al. (2020). « GHOST : Recovering Historical Signal from Heterotachously Evolved Sequence Alignments ». In : *Systematic Biology* 69.2, p. 249-264. DOI : 10.1093/sysbio/syz051 (cf. p. 19).

DARWIN, Charles (1859). *L'Origine des espèces*. par le moyen de la sélection naturelle, ou la préservation des races favorisées dans a lutte pour la vie. Éd. établie et préf. par Patrick TORT. Trad. de l'anglais par Aurélien BERRA. 6e éd. Révision de 1876. Paris : Honoré Champion Éditeur. ISBN : 978-2-7453-1862-6 (cf. p. 5, 8).

DEUTSCH, Jean (2012). *Le Gène. Un concept en évolution*. Science Ouverte. Le Seuil. 210 p. ISBN : 978-2-02-103701-2 (cf. p. 9).

DI FRANCO, Arnaud et al. (2019). « Evaluating the usefulness of alignment filtering methods to reduce the impact of errors on evolutionary inferences ». In : *BMC Evolutionary Biology* 19.1, p. 21. DOI : 10.1186/s12862-019-1350-2 (cf. p. 27, 46).

DONDI, Riccardo, Manuel LAFOND et Celine SCORNAVACCA (2019). « Reconciling multiple genes trees via segmental duplications and losses ». In : *Algorithms for molecular biology : AMB* 14, p. 7. DOI : 10.1186/s13015-019-0139-6 (cf. p. 74).

DOOLITTLE, Russell F. et Birger BLOMBÄCK (1964). « Amino-Acid Sequence Investigations of Fibrinopeptides from Various Mammals : Evolutionary Implications ». In : *Nature* 202.4928, p. 147-152. DOI : 10.1038/202147a0 (cf. p. 18).

DOS REIS, Mario (2016). « Notes on the birth–death prior with fossil calibrations for Bayesian estimation of species divergence times ». In : *Philosophical Transactions of the Royal Society B : Biological Sciences* 371.1699, p. 20150128. DOI : 10.1098/rstb.2015.0128 (cf. p. 41).

DOS REIS, Mario et al. (2015). « Uncertainty in the Timing of Origin of Animals and the Limits of Precision in Molecular Timescales ». In : *Current Biology* 25.22, p. 2939-2950. DOI : 10.1016/j.cub.2015.09.066 (cf. p. 20).

DOS REIS, Mario et al. (2018). « Using Phylogenomic Data to Explore the Effects of Relaxed Clocks and Calibration Strategies on Divergence Time Estimation : Primates as a Test Case ». en. In : *Systematic Biology* 67.4, p. 594-615. DOI : 10.1093/sysbio/syy001 (cf. p. 19, 29, 30, 45, 64).

DRUMMOND, Alexei J et al. (2006). « Relaxed phylogenetics and dating with confidence. » In : *PLoS biology* 4.5. Publisher : Public Library of Science, e88. DOI : 10.1371/journal.pbio.0040088 (cf. p. 19).

DUCHÊNE, David A. et al. (2019). « Linking Branch Lengths Across Sets of Loci Provides the Highest Statistical Support for Phylogenetic Inference ». In : *Molecular Biology and Evolution*. DOI : 10.1093/molbev/msz291 (cf. p. 42, 74).

EO, Soo Hyung et J. Andrew DEWOODY (2010). « Evolutionary rates of mitochondrial genomes correspond to diversification rates and to contemporary species richness in birds and reptiles ». In : *Proceedings of the Royal Society of London B : Biological Sciences* 277.1700. DOI : 10.1098/rspb.2010.0965 (cf. p. 68).

FELSENSTEIN, Joseph (1981). « Evolutionary trees from DNA sequences : A maximum likelihood approach ». In : *Journal of Molecular Evolution* 17.6, p. 368-376. DOI : 10.1007/BF01734359 (cf. p. 12, 13, 15).

– (1985). « Phylogenies and the Comparative Method ». In : *The American Naturalist* 125.1, p. 1-15. DOI : 10.2307/2461605 (cf. p. 16, 91).

– (2004). *Inferring Phylogenies*. Sinauer. ISBN : 0-87893-177-5 (cf. p. 12, 14).

– (2012). « A comparative method for both discrete and continuous characters using the threshold model. » In : *The American naturalist* 179.2, p. 145-56. DOI : 10.1086/663681 (cf. p. 93).

FITCH, W. M. et C. H. LANGLEY (1976). « Protein evolution and the molecular clock ». In : *Federation Proceedings* 35.10, p. 2092-2097. ISSN : 0014-9446 (cf. p. 18).

FITZJOHN, Richard G. (2010). « Quantitative Traits and Diversification ». In : *Systematic Biology* 59.6, p. 619-633. DOI : 10.1093/sysbio/syq053 (cf. p. 74).

FORCE, A et al. (1999). « Preservation of duplicate genes by complementary, degenerative mutations. » In : *Genetics* 151.4, p. 1531-45. ISSN : 0016-6731 (cf. p. 22, 75).

FREYMAN, William A. et Sebastian HÖHNA (2018). « Cladogenetic and Anagenetic Models of Chromosome Number Evolution : A Bayesian Model Averaging Approach ». In : *Systematic Biology* 67.2, p. 195-215. DOI : 10.1093/sysbio/syx065 (cf. p. 75).

GARAMSZEGI, László Zsolt, éd. (2014). *Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology*. Berlin, Heidelberg : Springer Berlin Heidelberg. DOI : 10.1007/978-3-662-43550-2 (cf. p. 17).

GOLDBERG, Emma E. et Boris IGIĆ (2008). « On Phylogenetic Tests of Irreversible Evolution ». In : *Evolution* 62.11, p. 2727-2741. DOI : 10.1111/j.1558-5646.2008.00505.x (cf. p. 95).

GOLDBERG, Emma E., Lesley T. LANCASTER et Richard H. REE (2011). « Phylogenetic Inference of Reciprocal Effects between Geographic Range Evolution and Diversification ». In : *Systematic Biology* 60.4, p. 451-465. DOI : 10.1093/sysbio/syr046 (cf. p. 74).

GOLDMAN, N et Z YANG (1994). « A codon-based model of nucleotide substitution for protein-coding DNA sequences. » In : *Molecular biology and evolution* 11.5, p. 725-36. ISSN : 0737-4038 (cf. p. 15).

GRAFEN, A. (1989). « The Phylogenetic Regression ». In : *Philosophical Transactions of the Royal Society of London B : Biological Sciences* 326.1233 (cf. p. 17, 70).

GRAUR, Dan et William MARTIN (2004). « Reading the entrails of chickens : molecular timescales of evolution and the illusion of precision ». In : *Trends in Genetics* 20.2, p. 80-86. DOI : 10.1016/j.tig.2003.12.003 (cf. p. 20).

GU, Xun et Wen-Hsiung LI (1992). « Higher rates of amino acid substitution in rodents than in humans ». In : *Molecular Phylogenetics and Evolution* 1.3, p. 211-214. DOI : 10.1016/1055-7903(92)90017-B (cf. p. 18).

GUIMIER, Anne et al. (2015). « MMP21 is mutated in human heterotaxy and is required for normal left-right asymmetry in vertebrates ». In : *Nature Genetics* 47.11, p. 1260-1263. DOI : 10.1038/ng.3376 (cf. p. 82).

GUINDON, Stéphane et al. (2010). « New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies : Assessing the Performance of PhyML 3.0 ». In : *Systematic Biology* 59.3, p. 307-321. DOI : 10.1093/sysbio/syq010 (cf. p. 46).

HAHN, Matthew W et al. (2005). « Estimating the tempo and mode of gene family evolution from comparative genomic data ». In : *Genome research* 15.530, p. 1153-1160. DOI : 10.1101/gr.3567505 (cf. p. 65, 80).

HAHN, Matthew W., Jeffery P. DEMUTH et Sang-Gook HAN (2007). « Accelerated rate of gene gain and loss in primates ». In : *Genetics* 177.3, p. 1941-1949. DOI : 10.1534/genetics.107.080077 (cf. p. 24).

HARVEY, Paul H et Mark D PAGEL (1991). *The comparative method in evolutionary biology*. Oxford University Press, Oxford. ISBN : 978-0-19-854640-5 (cf. p. 16).

HASTINGS, P J et al. (2009). « Mechanisms of change in gene copy number. » In : *Nature reviews. Genetics* 10.8, p. 551-64. DOI : 10.1038/nrg2593 (cf. p. 21).

HEATH, Tracy A., Mark T. HOLDER et John P. HUELSENBECK (2012). « A Dirichlet Process Prior for Estimating Lineage-Specific Substitution Rates ». In : *Molecular Biology and Evolution* 29.3, p. 939-955. DOI : 10.1093/molbev/msr255 (cf. p. 19).

HENNIG, Willi (1966). *Phylogenetic systematics* (cf. p. 9).

HUERTA-CEPAS, Jaime, François SERRA et Peer BORK (2016). « ETE 3 : Reconstruction, Analysis, and Visualization of Phylogenomic Data ». In : *Molecular Biology and Evolution* 33.6, p. 1635-1638. DOI : 10.1093/molbev/msw046 (cf. p. 56).

HUGALL, Andrew F. et Michael S. Y. LEE (2007). « The Likelihood Node Density Effect and Consequences for Evolutionary Studies of Molecular Rates ». In : *Evolution* 61.10, p. 2293-2307. DOI : 10.1111/j.1558-5646.2007.00188.x (cf. p. 20, 35).

HUSS, John (2017). « Paleontology : Outrunning Time ». In : *Time of Nature and the Nature of Time*. Sous la dir. de Christophe BOUTON et Philippe HUNEMAN. T. 326. Springer International Publishing, p. 211-235. ISBN : 978-3-319-53725-2 (cf. p. 6).

INNAN, Hideki et Fyodor KONDRASHOV (2010). « The evolution of gene duplications : classifying and distinguishing between models. » en. In : *Nat. Rev. Genet.* 11.2, p. 97-108. DOI : 10.1038/nrg2689 (cf. p. 22).

JUKES, Thomas H et Charles R CANTOR (1969). « Evolution of protein molecules ». In : H N MUNRO. *Mammalian protein metabolism*. T. 3, p. 21-132 (cf. p. 13).

JUNIER, Thomas et Evgeny M. ZDOBNOV (2010). « The Newick utilities : high-throughput phylogenetic tree processing in the Unix shell ». In : *Bioinformatics* 26.13, p. 1669-1670. DOI : 10.1093/bioinformatics/btq243 (cf. p. 48).

KAJIKAWA, Eriko et al. (2020). « Nodal paralogues underlie distinct mechanisms for visceral left–right asymmetry in reptiles and mammals ». In : *Nature Ecology & Evolution* 4.2, p. 261-269. DOI : 10.1038/s41559-019-1072-2 (cf. p. 82).

KENNY, N. J. et al. (2016). « Ancestral whole-genome duplication in the marine chelicerate horseshoe crabs ». In : *Heredity* 116.2, p. 190-199. DOI : 10.1038/hdy.2015.89 (cf. p. 21).

KHAN, Hameed, Arian SMIT et Stéphane BOISSINOT (2006). « Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates ». In : *Genome Research* 16.1, p. 78-87. DOI : 10.1101/gr.4001406 (cf. p. 20).

KIMURA, Motoo (1968). « Evolutionary Rate at the Molecular Level ». en. In : *Nature* 217.5129. Number : 5129 Publisher : Nature Publishing Group, p. 624-626. DOI : 10.1038/217624a0 (cf. p. 10).

– (1980). « A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences ». In : *Journal of Molecular Evolution* 16.2, p. 111-120. DOI : 10.1007/BF01731581 (cf. p. 14).

– (1983). *The neutral theory of molecular evolution*. Cambridge University Press (cf. p. 5, 10).

KOSAKOVSKY POND, Sergei L. et al. (2019). « HyPhy 2.5 - a customizable platform for evolutionary hypothesis testing using phylogenies ». In : *Molecular Biology and Evolution*. DOI : 10.1093/molbev/msz197 (cf. p. 40).

KUMAR, Sudhir et S. Blair HEDGES (1998). « A molecular timescale for vertebrate evolution ». In : *Nature* 392.6679, p. 917-920. DOI : 10.1038/31927 (cf. p. 19).

– (2016). « Advances in Time Estimation Methods for Molecular Data ». In : *Molecular Biology and Evolution* 33.4, p. 863-869. DOI : 10.1093/molbev/msw026 (cf. p. 20).

KUMAR, Sudhir et al. (2017). « TimeTree : A Resource for Timelines, Timetrees, and Divergence Times ». In : *Molecular Biology and Evolution* 34.7, p. 1812-1819. DOI : 10.1093/molbev/msx116 (cf. p. 29, 30, 44).

KURAKU, Shigehiro et Shigeru KURATANI (2011). « Genome-Wide Detection of Gene Extinction in Early Mammalian Evolution ». In : *Genome Biology and Evolution* 3, p. 1449-1462. DOI : 10.1093/gbe/evr120 (cf. p. 82).

LANFEAR, Robert, John J. WELCH et Lindell BROMHAM (2010). « Watching the clock : Studying variation in rates of molecular evolution between species ». In : *Trends in Ecology & Evolution* 25.9, p. 495-503. DOI : 10.1016/j.tree.2010.06.007 (cf. p. 20, 38).

LANGLEY, Charles H. et Walter M. FITCH (1974). « An examination of the constancy of the rate of molecular evolution ». In : *Journal of Molecular Evolution* 3.3, p. 161-177. DOI : 10.1007/BF01797451 (cf. p. 19).

LARTILLOT, Nicolas, Matthew J. PHILLIPS et Fredrik RONQUIST (2016). « A mixed relaxed clock model ». In : *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 371.1699. DOI : 10.1098/rstb.2015.0132 (cf. p. 19, 40).

LECCA, Paola, Nicola CASIRAGHI et Francesca DEMICHELIS (2015). « Defining order and timing of mutations during cancer progression : the TO-DAG probabilistic graphical model ». In : *Frontiers in Genetics* 6, p. 309. DOI : 10.3389/fgene.2015.00309 (cf. p. 20).

LOPEZ, P., D. CASANE et H. PHILIPPE (2002). « Heterotachy, an Important Process of Protein Evolution ». In : *Molecular Biology and Evolution* 19.1, p. 1-7. DOI : 10.1093/oxfordjournals.molbev.a003973 (cf. p. 18).

LOUCA, Stilianos et Matthew W. PENNELL (2020). « Extant timetrees are consistent with a myriad of diversification histories ». In : *Nature*, p. 1-4. DOI : 10.1038/s41586-020-2176-1 (cf. p. 74).

LYNCH, Michael (2007). *The Origins of Genome Architecture*. Sinauer Associates, Inc. ISBN : 978-0-87893-484-3 (cf. p. 7).

MACKINNON James, G et Halbert WHITE (1985). « Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties ». In : *Journal of Econometrics* 29.3, p. 305-325. DOI : 10.1016/0304-4076(85)90158-7 (cf. p. 57).

MADDISON, Wayne P. (1990). « A Method for Testing the Correlated Evolution of Two Binary Characters : Are Gains or Losses Concentrated on Certain Branches of a Phylogenetic Tree ? » In : *Evolution* 44.3, p. 539. DOI : 10.2307/2409434 (cf. p. 16, 87, 93, 97).

MADDISON, Wayne P. et Richard G. FITZJOHN (2015). « The Unsolved Challenge to Phylogenetic Correlation Tests for Categorical Characters ». In : *Systematic Biology* 64.1, p. 127-136. DOI : 10.1093/sysbio/syu070 (cf. p. 93).

MADDISON, Wayne P., Peter E. MIDFORD et Sarah P. OTTO (2007). « Estimating a binary character's effect on speciation and extinction ». In : *Systematic Biology* 56.5, p. 701-710. DOI : 10.1080/10635150701607033 (cf. p. 74).

MALIET, Odile, Florian HARTIG et Hélène MORLON (2019). « A model with many small shifts for estimating species-specific diversification rates ». In : *Nature Ecology & Evolution* 3.7, p. 1086. DOI : 10.1038/s41559-019-0908-0 (cf. p. 80).

MARGOLIASH, E. (1963). « PRIMARY STRUCTURE AND EVOLUTION OF CYTOCHROME C ». In : *Proceedings of the National Academy of Sciences of the United States of America* 50.4, p. 672-679. ISSN : 0027-8424 (cf. p. 18).

McKenzie, Andy et Mike Steel (2000). « Distributions of cherries for two models of trees ». In : *Mathematical Biosciences* 164.1, p. 81-92. DOI : 10.1016/S0025-5564(99)00060-7 (cf. p. 72).

Mir, Arnau, Francesc Rosselló et Luci A. Rotger (2013). « A new balance index for phylogenetic trees ». In : *Mathematical Biosciences* 241.1, p. 125-136. DOI : 10.1016/j.mbs.2012.10.005 (cf. p. 57).

Moorjani, Priya et al. (2016). « Variation in the molecular clock of primates ». In : *Proceedings of the National Academy of Sciences* 113.38, p. 10607-10612. DOI : 10.1073/pnas.1600374113 (cf. p. 36).

Mora, Camilo et al. (2011). « How Many Species Are There on Earth and in the Ocean ? » In : *PLoS Biology* 9.8. DOI : 10.1371/journal.pbio.1001127 (cf. p. 5).

Morel, Benoit et al. (2020). « GeneRax : A tool for species tree-aware maximum likelihood based gene family tree inference under gene duplication, transfer, and loss ». In : *bioRxiv*, p. 779066. DOI : 10.1101/779066 (cf. p. 24, 40, 65, 78).

Morgan, Gregory J. (1998). « Emile Zuckerkandl, Linus Pauling, and the Molecular Evolutionary Clock, 1959-1965 ». In : *Journal of the History of Biology* 31.2. Publisher : Springer, p. 155-178. ISSN : 0022-5010 (cf. p. 17).

Morlon, Hélène, Matthew D. Potts et Joshua B. Plotkin (2010). « Inferring the Dynamics of Diversification : A Coalescent Approach ». In : *PLoS Biology* 8.9. Sous la dir. de Paul H. Harvey, e1000493. DOI : 10.1371/journal.pbio.1000493 (cf. p. 74).

Morlon, Hélène et al. (2016). « RPANDA : an R package for macroevolutionary analyses on phylogenetic trees ». In : *Methods in Ecology and Evolution* 7.5. Sous la dir. de Richard Fitzjohn, p. 589-597. DOI : 10.1111/2041-210X.12526 (cf. p. 80).

Moutinho, Ana Filipa, Thomas Bataillon et Julien Y. Dutheil (2019). « Variation of the adaptive substitution rate between species and within genomes ». In : *Evolutionary Ecology*. DOI : 10.1007/s10682-019-10026-z (cf. p. 40).

Münkemüller, Tamara et al. (2012). « How to measure and test phylogenetic signal ». In : *Methods in Ecology and Evolution* 3.4, p. 743-756. DOI : 10.1111/j.2041-210X.2012.00196.x (cf. p. 71).

Nakamura, Tetsuya et Hiroshi Hamada (2012). « Left-right patterning : conserved and divergent mechanisms ». In : *Development* 139.18, p. 3257-3262. DOI : 10.1242/dev.061606 (cf. p. 81).

Nee, S., A. O. Mooers et P. H. Harvey (1992). « Tempo and mode of evolution revealed from molecular phylogenies ». In : *Proceedings of the National Academy of Sciences of the United States of America* 89.17, p. 8322-8326. DOI : 10.1073/pnas.89.17.8322 (cf. p. 73).

Nei, Masatoshi et Takashi Gojobori (1986). « Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. » In : *Molecular biology and evolution* 3.5, p. 418-26. ISSN : 0737-4038 (cf. p. 15).

Nei, Masatoshi et Sudhir Kumar (2000). *Molecular Evolution and Phylogenetics*. Oxford University Press. 348 p. ISBN : 978-0-19-535051-7 (cf. p. 18).

NOSIL, Patrik, J FUNK Daniel et Daniel ORTIZ-BARRIENTOS (2009). « Divergent selection and heterogeneous genomic divergence ». In : *Molecular Ecology* 18.3, p. 375-402. DOI : 10.1111/j.1365-294X.2008.03946.x (cf. p. 40).

NUTTALL, George Henry Falkiner (1904). *Blood immunity and blood relationship : a demonstration of certain blood-relationships amongst animals by means of the precipitin test for blood*. University Press (cf. p. 17).

O'MEARA, Brian C. (2012). « Evolutionary Inferences from Phylogenies : A Review of Methods ». In : *Annual Review of Ecology, Evolution, and Systematics* 43.1, p. 267-285. DOI : 10.1146/annurev-ecolsys-110411-160331 (cf. p. 17).

OHNO, S (1970). *Evolution by Gene Duplication*. Springer-Verlag. ISBN : 0-04-575015-7 (cf. p. 22).

OHTA, T (1992). « The Nearly Neutral Theory of Molecular Evolution ». In : *Annual Review of Ecology and Systematics* 23.1. Publisher : Annual Reviews 4139 El Camino Way, P.O. Box 10139, Palo Alto, CA 94303-0139, USA, p. 263-286. DOI : 10.1146/annurev.es.23.110192.001403 (cf. p. 10).

OHTA, Tomoko et Motoo KIMURA (1971). « On the constancy of the evolutionary rate of cistrons ». en. In : *Journal of Molecular Evolution* 1.1, p. 18-25. DOI : 10.1007/BF01659391 (cf. p. 18).

OPAZO, Juan C. et al. (2019). « Evolution of nodal and nodal-related genes and the putative composition of the heterodimers that trigger the nodal pathway in vertebrates ». In : *Evolution & Development* 21.4, e12292. DOI : 10.1111/ede.12292 (cf. p. 82).

OTA, Tatsuya et Masatoshi NEI (1994). « Divergent evolution and evolution by the birth-and-death process in the immunoglobulin VH gene family. » In : *Molecular Biology and Evolution* 11.3, p. 469-482. DOI : 10.1093/oxfordjournals.molbev.a040127 (cf. p. 23).

OVCHINNIKOV, Igor, Adrienne RUBIN et Gary D. SWERGOLD (2002). « Tracing the LINEs of human evolution ». In : *Proceedings of the National Academy of Sciences* 99.16, p. 10522-10527. DOI : 10.1073/pnas.152346799 (cf. p. 20).

PACIFICI, Michela et al. (2013). « Generation length for mammals ». In : *Nature Conservation* 5, p. 89-94. DOI : 10.3897/natureconservation.5.5734 (cf. p. 62, 80).

PAGEL, Mark (1994). « Detecting Correlated Evolution on Phylogenies : A General Method for the Comparative Analysis of Discrete Characters ». In : *Proceedings : Biological Sciences* 255, p. 37-45. DOI : 10.2307/49836 (cf. p. 16, 87, 93, 94, 98).

PAGEL, Mark, Chris VENDITTI et Andrew MEADE (2006). « Large Punctuational Contribution of Speciation to Evolutionary Divergence at the Molecular Level ». In : *Science* 314.5796. Publisher : American Association for the Advancement of Science Section : Report, p. 119-121. DOI : 10.1126/science.1129647 (cf. p. 18).

PARADIS, Emmanuel (2013). « Molecular dating of phylogenies by likelihood methods : A comparison of models and a new information criterion ». In : *Molecular Phylogenetics and Evolution* 67.2, p. 436-444. DOI : 10.1016/j.ympev.2013.02.008 (cf. p. 51).

PAULING, Linus et al. (1949). « Sickle Cell Anemia, a Molecular Disease ». In : *Science* 110.2865, p. 543-548. ISSN : 0036-8075 (cf. p. 17).

PERES, Amélie et Hugues ROEST CROLLIUS (2015). « Improving duplicated nodes position in vertebrate gene trees ». In : *BMC Bioinformatics* 16 (Suppl 3), A9. DOI : 10.1186/1471-2105-16-S3-A9 (cf. p. 47).

PERLES, Zeev et al. (2015). « A human laterality disorder caused by a homozygous deleterious mutation in MMP21 ». In : *Journal of Medical Genetics* 52.12, p. 840-847. DOI : 10.1136/jmedgenet-2015-103336 (cf. p. 82).

PRESGRAVES, Daven C. (2010). « The molecular evolutionary basis of species formation ». In : *Nature Reviews Genetics* 11.3, p. 175-180. DOI : 10.1038/nrg2718 (cf. p. 22).

PULQUÉRIO, Mário J. F. et Richard A. NICHOLS (2007). « Dates from the molecular clock : how wrong can we be ? » In : *Trends in Ecology & Evolution* 22.4, p. 180-184. DOI : 10.1016/j.tree.2006.11.013 (cf. p. 20).

RANNALA, Bruce et Ziheng YANG (2007). « Inferring Speciation Times under an Episodic Molecular Clock ». In : *Systematic Biology* 56.3, p. 453-466. DOI : 10.1080/10635150701420643 (cf. p. 19).

RASMUSSEN, Matthew D. et Manolis KELLIS (2012). « Unified modeling of gene duplication, loss, and coalescence using a locus tree ». In : *Genome Research* 22.4, p. 755-765. DOI : 10.1101/gr.123901.111 (cf. p. 24, 75).

REIS, Mario dos et Ziheng YANG (2013). « The unbearable uncertainty of Bayesian divergence time estimation ». In : *Journal of Systematics and Evolution* 51.1, p. 30-43. DOI : 10.1111/j.1759-6831.2012.00236.x (cf. p. 20).

REVELL, Liam J. (2009). « Size-Correction and Principal Components for Interspecific Comparative Studies ». In : *Evolution* 63.12, p. 3258-3268. DOI : 10.1111/j.1558-5646.2009.00804.x (cf. p. 94).

RIDLEY, Mark (1983). *The explanation of organic diversity : the comparative method and adaptations for mating*. Oxford University Press, USA (cf. p. 16).

RIVERA, Maria C. et James A. LAKE (2004). « The ring of life provides evidence for a genome fusion origin of eukaryotes ». In : *Nature* 431.7005, p. 152-155. DOI : 10.1038/nature02848 (cf. p. 10).

ROELOFS, Dick et al. (2020). « Multi-faceted analysis provides little evidence for recurrent whole-genome duplications during hexapod evolution ». In : *BMC Biology* 18.1, p. 57. DOI : 10.1186/s12915-020-00789-1 (cf. p. 21).

ROUSSELLE, M. et al. (2019). « Is adaptation limited by mutation ? A timescale-dependent effect of genetic diversity on the adaptive substitution rate in animals ». In : *bioRxiv*. ver. 4, évalué par les pairs et recommandé par Peer Community in Evolutionary Biology, p. 643619. DOI : 10.1101/643619 (cf. p. 40).

SACKIN, M. J. (1972). « “Good” and “Bad” Phenograms ». In : *Systematic Biology* 21.2, p. 225-226. DOI : 10.1093/sysbio/21.2.225 (cf. p. 57).

SAIKI, R. K. et al. (1985). « Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia ». In : *Science* 230.4732, p. 1350-1354. DOI : 10.1126/science.2999980 (cf. p. 10).

SANDERSON, M. J. (2002). « Estimating Absolute Rates of Molecular Evolution and Divergence Times : A Penalized Likelihood Approach ». In : *Molecular Biology and Evolution* 19.1, p. 101-109. DOI : 10.1093/oxfordjournals.molbev.a003974 (cf. p. 19, 50-52).

SANGER, F., S. NICKLEN et A. R. COULSON (1977). « DNA sequencing with chain-terminating inhibitors ». In : *Proceedings of the National Academy of Sciences of the United States of America* 74.12, p. 5463-5467. ISSN : 0027-8424 (cf. p. 10).

SARICH, Vincent M. et Allan C. WILSON (1966). « Quantitative Immunochemistry and the Evolution of Primate Albumins : Micro-Complement Fixation ». In : *Science* 154.3756, p. 1563-1566. DOI : 10.1126/science.154.3756.1563 (cf. p. 18).

SCALLY, Aylwyn et al. (2012). « Insights into hominid evolution from the gorilla genome sequence ». In : *Nature* 483.7388, p. 169-175. DOI : 10.1038/nature10842 (cf. p. 40).

SCHWAGER, Evelyn E. et al. (2017). « The house spider genome reveals an ancient whole-genome duplication during arachnid evolution ». In : *BMC Biology* 15.1, p. 62. DOI : 10.1186/s12915-017-0399-x (cf. p. 21).

SCHWARTZ, Russell et Alejandro A. SCHÄFFER (2017). « The evolution of tumour phylogenetics : principles and practice ». In : *Nature Reviews. Genetics* 18.4, p. 213-229. DOI : 10.1038/nrg.2016.170 (cf. p. 20).

SCORNAVACCA, Celine, Frédéric DELSUC et Nicolas GALTIER (2020). *Phylogenetics in the Genomic Era*. No commercial publisher | Authors open access book (cf. p. 24).

SCORNAVACCA, Celine et Nicolas GALTIER (2016). « Incomplete Lineage Sorting in Mammalian Phylogenomics ». In : *Systematic Biology* 18.1, syw082. DOI : 10.1093/sysbio/syw082 (cf. p. 75).

SEMPERE, Lorenzo F. et al. (2006). « The phylogenetic distribution of metazoan microRNAs : insights into evolutionary complexity and constraint ». In : *Journal of Experimental Zoology Part B : Molecular and Developmental Evolution* 306B.6, p. 575-588. DOI : 10.1002/jez.b.21118 (cf. p. 20).

SIMAKOV, Oleg et al. (2013). « Insights into bilaterian evolution from three spiralian genomes ». In : *Nature* 493.7433, p. 526-531. DOI : 10.1038/nature11696 (cf. p. 95).

SIMPSON, George Gaylord (1964). « Organisms and Molecules in Evolution ». In : *Science* 146.3651, p. 1535-1538. ISSN : 0036-8075 (cf. p. 18).

SJÖSTRAND, Joel et al. (2012). « DLRS : gene tree evolution in light of a species tree ». In : *Bioinformatics* 28.22, p. 2994-2995. DOI : 10.1093/bioinformatics/bts548 (cf. p. 24).

SMITH, Stephen A., Joseph W. BROWN et Joseph F. WALKER (2018). « So many genes, so little time : A practical approach to divergence-time estimation in the genomic era ». In : *PLOS ONE* 13.5. Sous la dir. d'Hector ESCRIVA. Publisher : Public Library of Science, e0197433. DOI : 10.1371/journal.pone.0197433 (cf. p. 19, 20, 38, 40, 76).

STURTEVANT, A. H. (1925). « The Effects of Unequal Crossing over at the Bar Locus in Drosophila ». In : *Genetics* 10.2, p. 117-147. ISSN : 0016-6731 (cf. p. 21).

SUDMANT, Peter H. et al. (2015). « An integrated map of structural variation in 2,504 human genomes ». In : *Nature* 526.7571, p. 75-81. DOI : 10.1038/nature15394 (cf. p. 21).

SZÖLLŐSI, Gergely J. et al. (2013). « Lateral Gene Transfer from the Dead ». In : *Systematic Biology* 62.3, p. 386-397. DOI : 10.1093/sysbio/syt003 (cf. p. 24, 65).

TAJIMA, F. (1993). « Simple methods for testing the molecular evolutionary clock hypothesis. » In : *Genetics* 135.2, p. 599-607. ISSN : 0016-6731, 1943-2631 (cf. p. 18).

TAO, Qiqing et al. (2019). « Reliable Confidence Intervals for RelTime Estimates of Evolutionary Divergence Times ». In : *Molecular Biology and Evolution*. DOI : 10.1093/molbev/msz236 (cf. p. 20).

THORNE, J. L., H. KISHINO et I. S. PAINTER (1998). « Estimating the rate of evolution of the rate of molecular evolution ». In : *Molecular Biology and Evolution* 15.12, p. 1647-1657. DOI : 10.1093/oxfordjournals.molbev.a025892 (cf. p. 19).

TIBSHIRANI, Robert (1996). « Regression shrinkage and selection via the Lasso ». In : *Journal of the Royal Statistical Society : Series B (Statistical Methodology)* 58.1, p. 267-288. DOI : 10.1111/j.2517-6161.1996.tb02080.x (cf. p. 56).

TO, Thu-Hien et al. (2016). « Fast Dating Using Least-Squares Criteria and Algorithms ». en. In : *Systematic Biology* 65.1, p. 82-97. DOI : 10.1093/sysbio/syv068 (cf. p. 19, 20).

TOFIGH, Ali et al. (2009). « Detecting LGTs using a novel probabilistic model integrating duplications, LGTs, losses, rate variation, and sequence evolution ». In : (cf. p. 24).

ULLAH, Ikram et al. (2015). « Integrating Sequence Evolution into Probabilistic Orthology Analysis. » In : *Systematic biology* 64.6, p. 969-82. DOI : 10.1093/sysbio/syv044 (cf. p. 77).

VOLZ, E. M. et S. D. W. FROST (2017). « Scalable relaxed clock phylogenetic dating ». In : *Virus Evolution* 3.2. Publisher : Oxford University Press. DOI : 10.1093/ve/vex025 (cf. p. 19, 20).

WANG, Huai-Chun, Edward SUSKO et Andrew J. ROGER (2019). « The Relative Importance of Modeling Site Pattern Heterogeneity Versus Partition-Wise Heterotachy in Phylogenomic Inference ». In : *Systematic Biology* 68.6, p. 1003-1019. DOI : 10.1093/sysbio/syz021 (cf. p. 40).

WARNOCK, Rachel C. M., Ziheng YANG et Philip C. J. DONOGHUE (2017). « Testing the molecular clock using mechanistic models of fossil preservation and molecular evolution ». In : *Proceedings of the Royal Society B : Biological Sciences* 284.1857, p. 20170227. DOI : 10.1098/rspb.2017.0227 (cf. p. 20).

WELLS, John W. (1963). « Coral Growth and Geochronometry ». In : *Nature* 197.4871, p. 948-950. DOI : 10.1038/197948a0 (cf. p. 6).

WHEELER, Benjamin M. et al. (2009). « The deep evolution of metazoan microRNAs ». In : *Evolution & Development* 11.1, p. 50-68. DOI : 10.1111/j.1525-142X.2008.00302.x (cf. p. 20).

WOESE, C R et G E FOX (1977). « Phylogenetic structure of the prokaryotic domain : the primary kingdoms. » In : *Proceedings of the National Academy of Sciences of the United States of America* 74.11, p. 5088-5090. ISSN : 0027-8424 (cf. p. 10).

WOLF, Jochen B. W. et Hans ELLEGREN (2017). « Making sense of genomic islands of differentiation in light of speciation ». In : *Nature Reviews. Genetics* 18.2, p. 87-100. DOI : 10.1038/nrg.2016.133 (cf. p. 40).

WOLFE, Kenneth H., Paul M. SHARP et Wen-Hsiung LI (1989). « Mutation rates differ among regions of the mammalian genome ». In : *Nature* 337.6204, p. 283-285. DOI : 10.1038/337283a0 (cf. p. 19).

WU, C. I. et W. H. LI (1985). « Evidence for higher rates of nucleotide substitution in rodents than in man ». In : *Proceedings of the National Academy of Sciences* 82.6, p. 1741-1745. DOI : 10.1073/pnas.82.6.1741 (cf. p. 18).

YANG, Ziheng (2007). « PAML 4 : phylogenetic analysis by maximum likelihood. » In : *Molecular biology and evolution* 24.8, p. 1586-91. DOI : 10.1093/molbev/msm088 (cf. p. 19, 27, 41, 48).

YANG, Ziheng et Rasmus NIELSEN (1998). « Synonymous and nonsynonymous rate variation in nuclear genes of mammals ». In : *Journal of Molecular Evolution* 46.4, p. 409-418. DOI : 10.1007/PL00006320 (cf. p. 15).

YANG, Ziheng et Bruce RANNALA (2006). « Bayesian Estimation of Species Divergence Times Under a Molecular Clock Using Multiple Fossil Calibrations with Soft Bounds ». In : *Molecular Biology and Evolution* 23.1, p. 212-226. DOI : 10.1093/molbev/msj024 (cf. p. 41).

ZAHARIAS, Paul et T SANDERS Malcolm (2018). *Introduction aux méthodes comparatives phylogénétiques*. Éditions Matériologiques. Biosystema 31. ISBN : 978-2-37361-188-5 (cf. p. 17).

ZERBINO, Daniel R. et al. (2018). « Ensembl 2018 ». In : *Nucleic Acids Research* 46 (D1), p. D754-D761. DOI : 10.1093/nar/gkx1098 (cf. p. 44).

ZHAO, Jing et al. (2015). « A generalized birth and death process for modeling the fates of gene duplication. » In : *BMC evolutionary biology* 15, p. 275. DOI : 10.1186/s12862-015-0539-2 (cf. p. 75).

ZHU, Tianqi, Mario DOS REIS et Ziheng YANG (2015). « Characterization of the Uncertainty of Divergence Time Estimation under Relaxed Molecular Clock Models Using Multiple Loci ». In : *Systematic Biology* 64.2, p. 267-280. DOI : 10.1093/sysbio/syu109 (cf. p. 20).

ZUCKERKANDL, Emile et Linus PAULING (1962). « Molecular disease, evolution, and genetic heterogeneity ». In : *Horizons in biochemistry*, p. 189-225 (cf. p. 18).

– (1965). « Molecules as documents of evolutionary history ». In : *Journal of theoretical biology* 8.2, p. 357-366 (cf. p. 9, 17).

Table des figures

Liste des tableaux

Liste des abréviations

ACP Analyse en Composantes Principales
GO Gene Ontology
ILS Incomplete Lineage Sorting (Tri incomplet de lignée)
LRV Log-Ratio de Vraisemblances
Ma Million d'années
mAD mean Absolute Deviation from the median
MCO Moindres Carrés Ordinaires
MPL Mean-Path-Length
PIC Phylogenetic Independent Contrasts
PGLS Phylogenetic Generalized Least Squares

Remerciements

Ces quatre années de thèse ont été un cheminement très exigeant et enrichissant, et je suis extrêmement reconnaissant envers mon superviseur Hugues d'avoir rendu ce parcours possible. Tu m'as très souvent accordé ta confiance, ton soutien et une grande liberté, me permettant de me former dans des domaines nouveaux qui n'avaient pas été envisagés, ainsi que l'opportunité d'interagir avec la communauté, même lorsque les résultats n'étaient pas tout à fait prêts à être présentés. Tu as également su me forcer à me remettre constamment en question scientifiquement, en ne laissant jamais de côté les zones d'ombres. J'espère un jour être à la hauteur de ce niveau d'exigence scientifique.

Je remercie chaleureusement mes rapporteurs Julien Yann Dutheil et Bastien Boussau, d'avoir accepté d'évaluer ma thèse —pendant l'été—, et mon jury Ingrid Lafontaine, Maria Anisimova et Guillaume Achaz. Cela représente un réel honneur d'être évalué par des scientifiques que j'admire et qui m'inspirent à continuer dans cette voie.

Je suis également endetté auprès d'Hélène Morlon, qui m'a accordé beaucoup de temps et de précieux conseils, et je regrette seulement de ne pas avoir réussi à concrétiser cette aide scientifique en un travail plus collaboratif. Mon comité de suivi de thèse, Gilles Fisher et Nicolas Lartillot m'a également été d'un soutien inestimable, grâce à ses encouragements et ses conseils, que je m'excuse de n'avoir pas correctement rémunéré au moyen de chouquettes lors de ces longues réunions.

L'environnement de travail est une composante essentielle dans ce marathon, et j'ai été particulièrement chanceux d'avoir l'équipe Dyogen comme accueil : merci Alex pour ta disponibilité sur tous les sujets ; merci Camille, pour tes corrections très exigeantes, tes questions perçantes, tes rappels de deadlines. Merci Yves, toujours apte à faire dériver nombre de discussions en dehors de la science, avec une mention spéciale pour les pauses babyfoot d'une violence verbale inénarrable. Merci Lambert, aîné de thèse, qui m'a fourni un modèle à suivre, et pour avoir initié les pauses cookies et les soirées molky. Merci enfin à Nga, Élise et François sur qui j'ai aussi pu compter en cas de difficultés. Merci aussi aux Dyogens que j'ai croisé plus brièvement.

Il serait impossible de faire de la recherche sans le personnel de l'IBENS et de la section Génomique Fonctionnelle : d'abord un immense merci au service informatique, qui fournit un travail titanesque en apportant à l'institut entier

toute l'infrastructure nécessaire, mais de façon personnalisée et flexible. Grâce à Pierre Vincens, n'importe quel programme écrit par un bioinformaticien et ne compilant que dans un contexte particulier de phases de la lune peut être disponible. Merci également à Bilel, Phi Phong, Catherine et Nolwenn, pour leurs interventions patientes et efficaces. Et puis, l'expertise en mécanique vélo de Phi Phong m'a été très instructive. Un grand merci également à l'équipe gestion : Brigitte et Abdoul, toujours d'humeur taquine. Le pôle doctorant de l'ENS aura également été d'une disponibilité précieuse dans les démarches administratives.

L'équipe-dans-les-bureaux-d'en-face-du-couloir, ou CSB, a également contribué à la convivialité et la bienveillance de notre coin des bioinformaticiens. Merci en particulier à Morgane, qui m'a accompagné pour l'enseignement. Merci à Aurélien, Laura, Swann et Denis pour les agréables temps de pause et les pots. Dans l'institut, j'ai également eu la chance de rencontrer d'autres étudiants, partageant avec moi des préoccupations fondamentales de thésard comme l'heure de la cantine et la pause babyfoot. Merci Nikita, qui a un don pour rassembler les gens, merci à Toni, Tiphaine, Marco, Guillaume "Dédé", Caroline pour cette équipée et tous ceux que j'ai rencontrés de façon plus ou moins brève.

Hors du labo, des personnes incroyables m'ont permis de toujours rester enthousiaste, à commencer par mes colocs : Guilhem et Adam, puis Lambert, trois thésards dans un même bâteau. Guilhem, il est impossible d'énumérer tout ce que j'ai appris ou retiré de nos discussions. Merci à l'équipe Laure, Oriane, Adam pour votre grain de folie ressourçant ; au clan "Zen" Benjamin, François, Maxime, Hugo, détenteurs de la sagesse ultime ; à la tribu "concerts" David, Bertrand, Bilgé ; aux camarades de promo dont Alexis et Romain ; aux agros de Ginette, envers qui je dois pourtant m'excuser d'avoir si souvent posé des lapins ; aux Phoenix de Montrouge, en particulier coachs, bureau, coéquipiers, parce que courir après un frisbee est une défoulade qui justifie de se lever le samedi matin comme en semaine ; aux amitiés du collège et lycée. Parmi les personnes ci-dessus, je suis spécialement redevable à tous ceux qui m'ont généreusement aidé lors de mes déménagements annuels…

Si ma passion pour l'évolution a pu aboutir à cette thèse, c'est aussi grâce aux professeurs formidables que j'ai pu avoir, ainsi qu'à ma famille. Merci à mon père et à mon frère. Merci à ma mère. Parmi ce que je te dois, il y a sans doute la cultivation de ma curiosité, un nourrissement constant de ma passion, ce choix de prépa, et tout le travail que tu as fourni pour que ça soit possible.

Et puis pour Thérèse, ta réactivité face à mes questionnements parfois rébarbatifs, ta tolérance à mes horaires trop nocturnes, et surtout ton obstination à toujours m'alléger de contraintes : ce n'est qu'une faible illustration de ton soutien inconditionnel.

Annexes

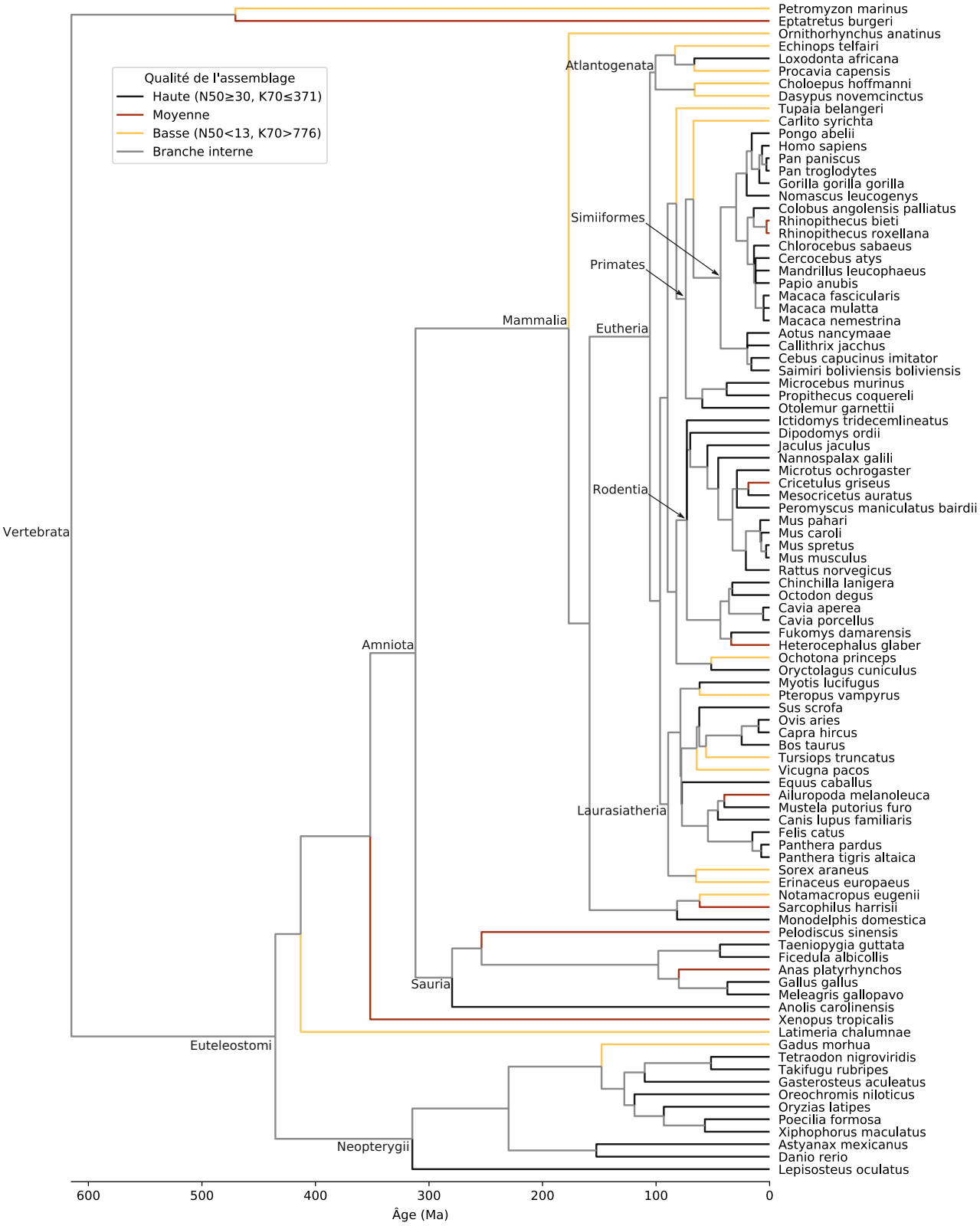


FIGURE A1 – Phylogénie adaptée des vertébrés d'Ensembl

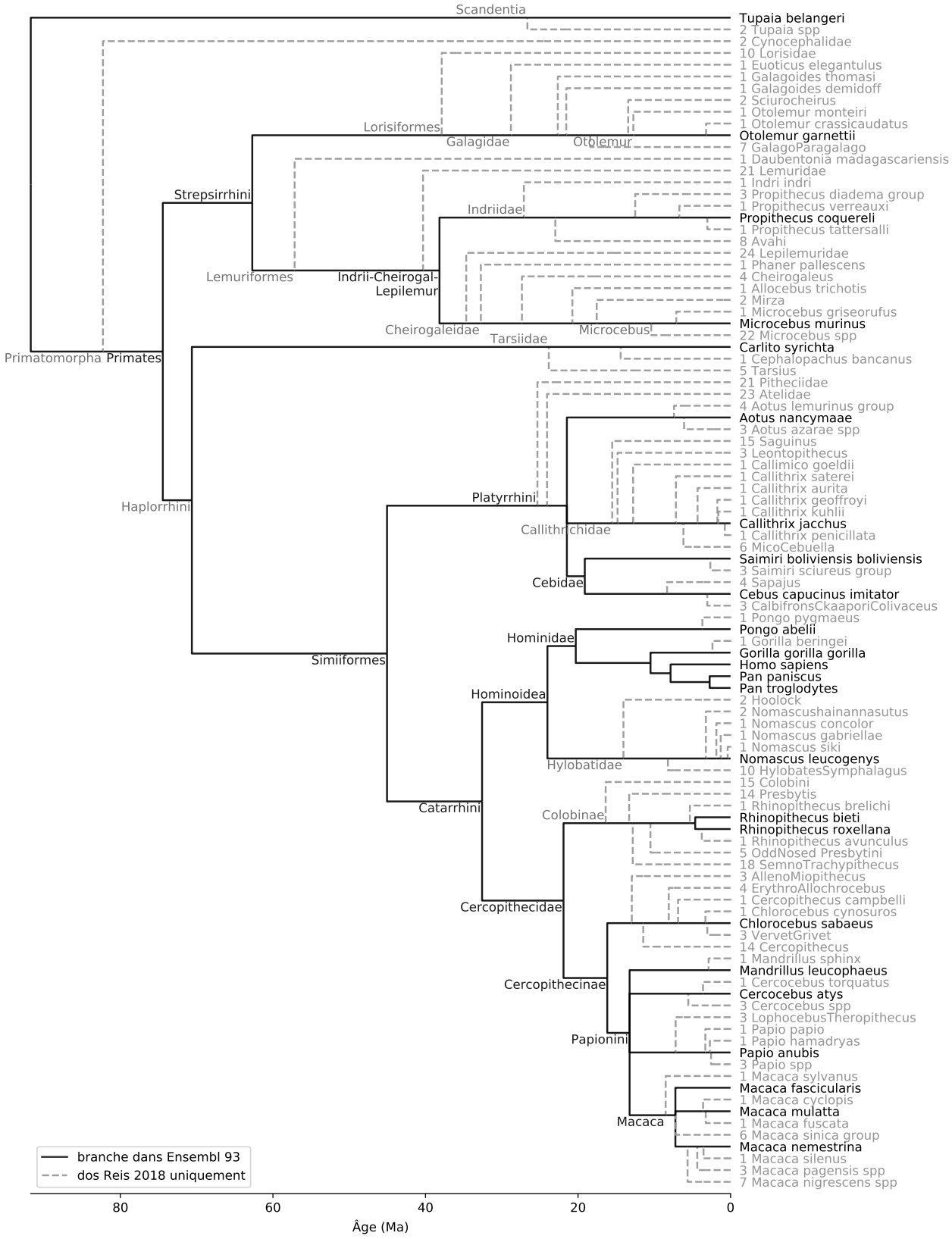


FIGURE A2 – Phylogénie Primates utilisée, incluant les espèces à mauvais assemblage. Les taxons non échantillonnés dans Ensembl 93 sont en gris pointillé pour complétude.

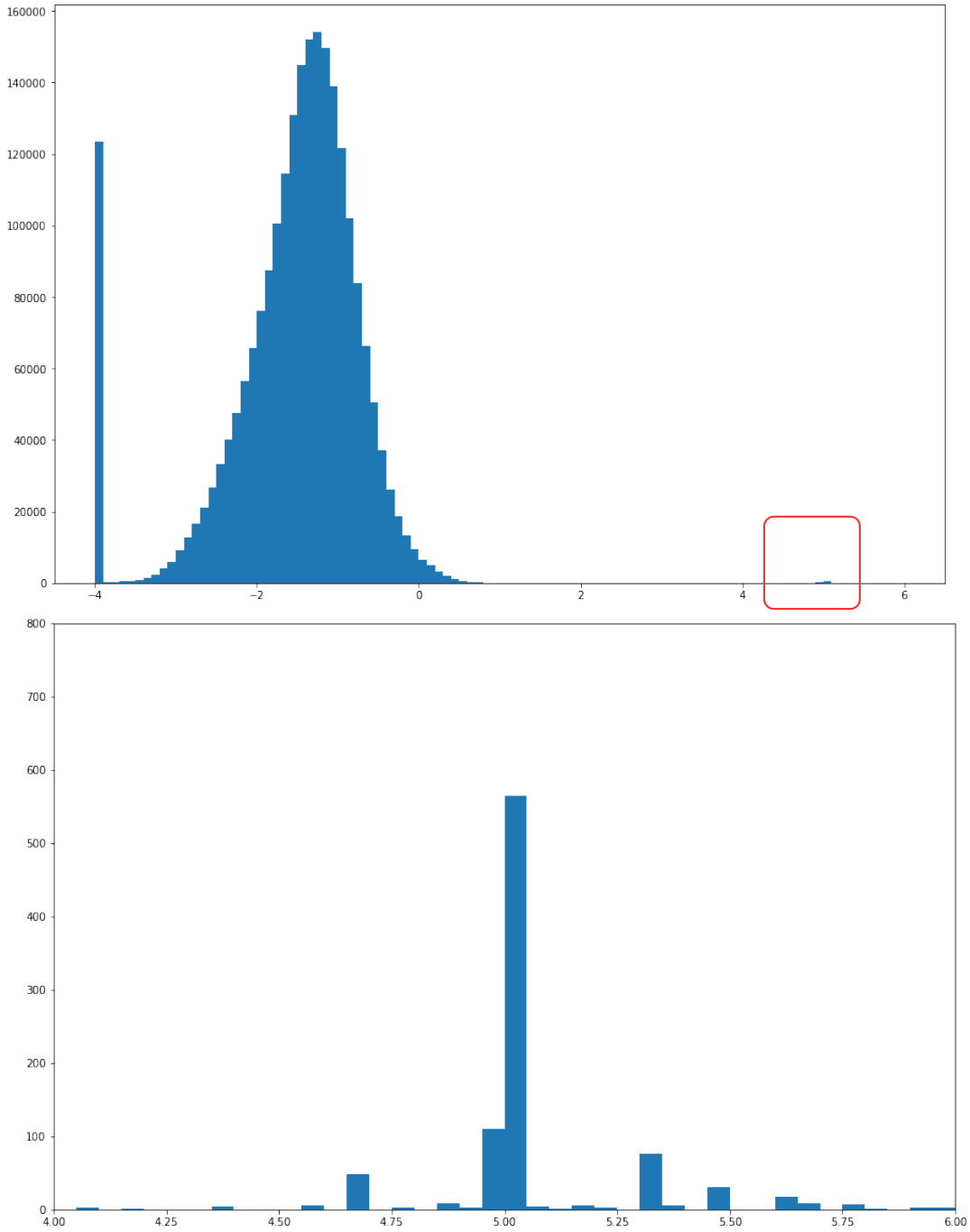
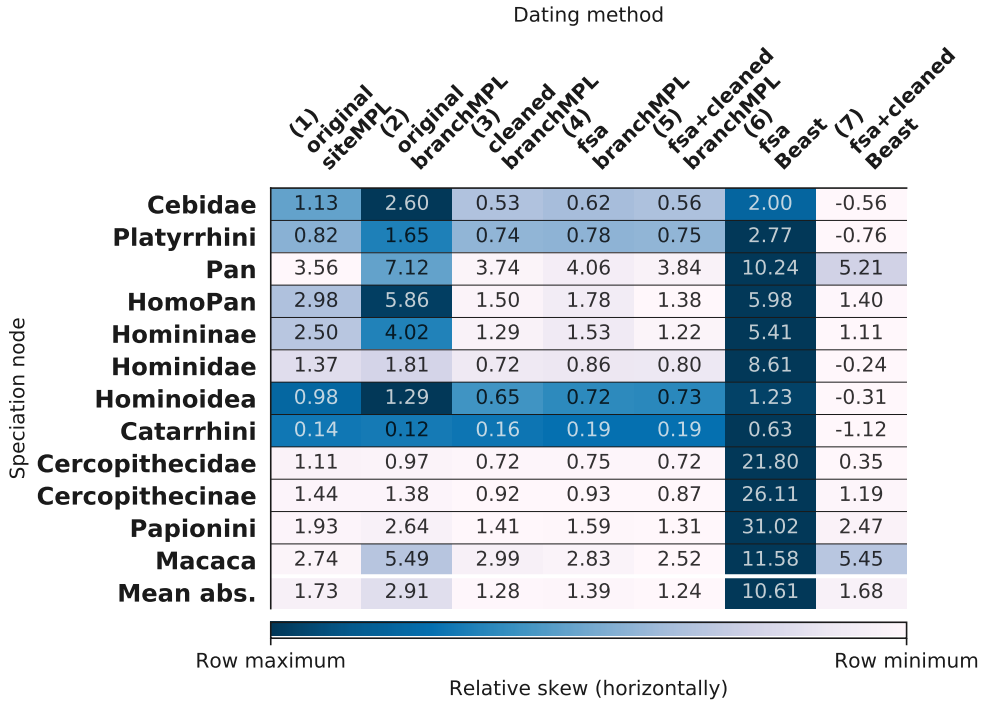
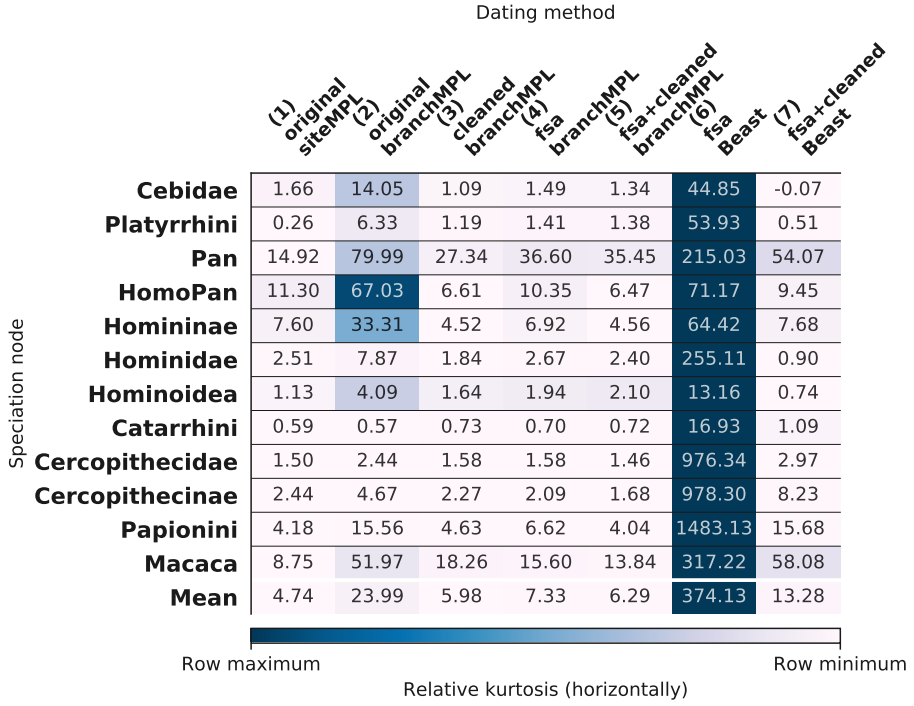


FIGURE A3 – Distribution des longueurs de branche en logarithme (base 10) dans la forêt entière d'Ensembl. Ces longueurs sont initialement en substitutions par nucléotide, calculées par PhyML via TreeBest (modèle GTR-Γ). En bas, zoom sur le carré rouge.



a) Asymétrie (skew)



b) Kurtosis (poids des valeurs non centrales)

FIGURE A4 – Statistiques de dispersion par procédure

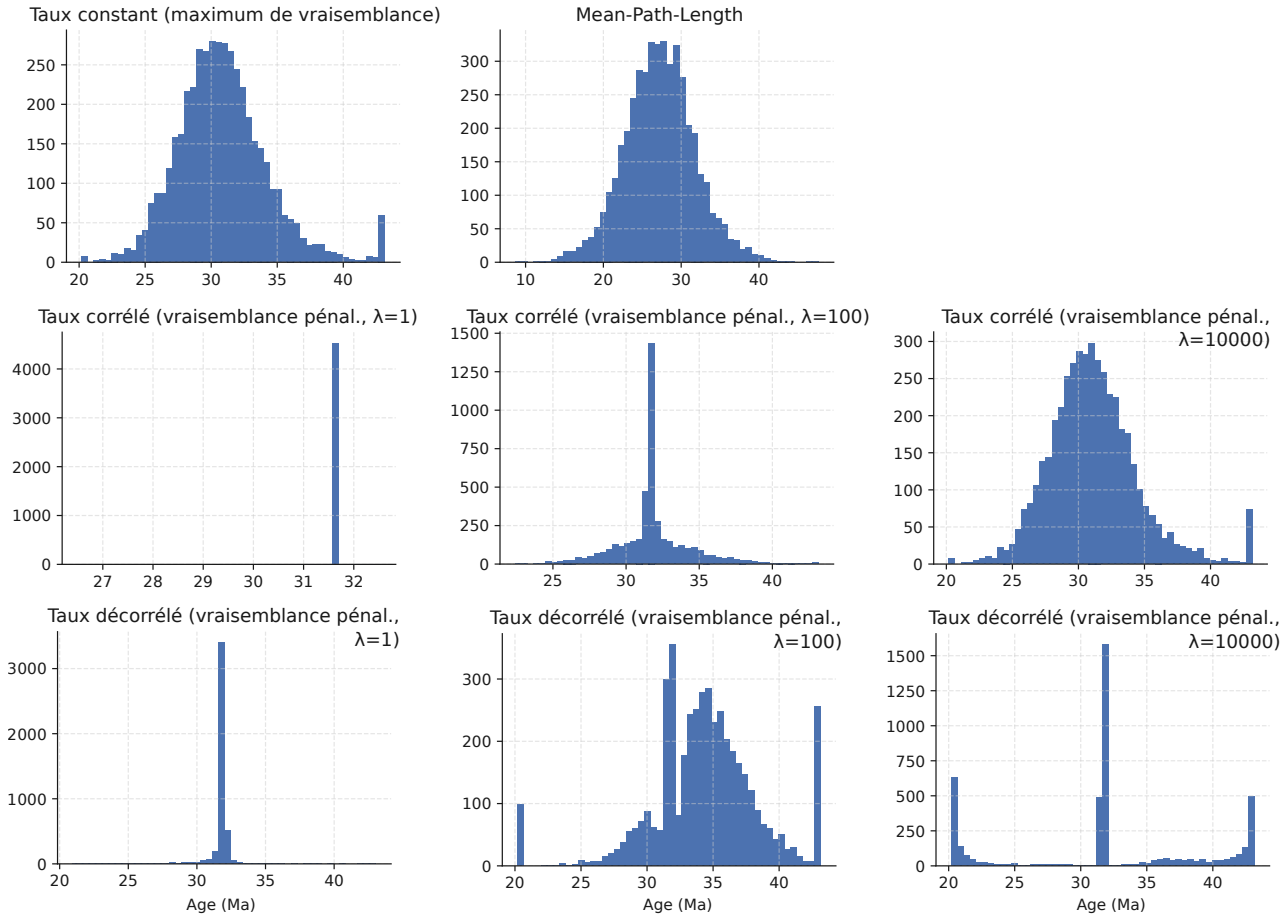


FIGURE A5 – Datations avec APE chronos. Âges de *Catarrhini* selon différentes paramétrisations de `chronos`

TABLE A6 – R^2 ajustés des régressions des erreurs locales par spéciation

Régression	Adjusted R^2
Catarrhini	0.1234
Hominoidea	0.1392
Platyrrhini	0.1796
Cercopithecidae	0.1063
Cebidae	0.1376
Hominidae	0.1433
Cercopithecinae	0.1054
Papionini	0.1056
Homininae	0.1456
HomoPan	0.1484
Macaca	0.1370
Pan	0.3121
globale	0.47

TABLE A7 – Nombre de nœuds reconstruits pour l'arbre d'espèces. L'édition (section 12.3.3.2) reconstruit la topologie de la forêt d'arbres produisant les nouveaux nœuds suivants.

Fereuungulata	13045
Zooamata	10664
Atlantogenata	7039
HomoPan	6506
Scrotifera	6134
Papionini	5977
ChinchillaOctodon	5557
Euarchonta	5484
Cebidae	5153
Australidelphia	4882
MyodontaCastorimorpha	4611
Caviomorpha	3913
Cetruminantia	3766
Artiofabula	3726
Arctoidea	2712
Mus_B	1678
Paenungulata	1361
Eumuroida	337
Rodentia	53
Platyrrhini	29
total	92627

A8 Fichier de contrôle de `codeml`, modèle "free-ratio"

```
seqfile  = SimiiformesENSGT00390000000002.a.a.a.a.a.b.a_genes.phy
treefile = SimiiformesENSGT00390000000002.a.a.a.a.a.b.a_codeml.nwk
outfile  = SimiiformesENSGT00390000000002.a.a.a.a.a.b.a_m1w04.mlc

        noisy = 1   * 0,1,2,3,9: how much rubbish on the screen
      verbose = 2   * 1: detailed output, 0: concise output
      runmode = 0   * 0: user tree;  1: semi-automatic;  2: automatic
                    * 3: StepwiseAddition; (4,5):PerturbationNNI; -2: pairwise

      seqtype = 1   * 1:codons; 2:AAs; 3:codons-->AAs
    CodonFreq = 2   * 0:1/61 each, 1:F1X4, 2:F3X4, 3:codon table
        ndata = 1   * specifies the number of separate data sets in the file
        clock = 0   * 0: no clock, unrooted tree, 1: clock, rooted tree

       aaDist = 0   * 0:equal, +:geometric; -:linear, 1-6:G1974,Miyata,c,p,v,a
                    * 7:AAClasses

        model = 1   * models for codons:
                        * 0:one, 1:b, 2:2 or more dN/dS ratios for branches
                    * models for AAs or codon-translated AAs:
                        * 0:poisson, 1:proportional, 2:Empirical, 3:Empirical+F
                        * 6:FromCodon, 8:REVaa_0, 9:REVaa(nr=189)

      NSsites = 0   * 0:one w; 1:neutral; 2:positive selection; 3:discrete; 4:freqs;
                    * 5:gamma; 6:2gamma; 7:beta; 8:beta&w; 9:beta&gamma;
                    * 10:beta&gamma+1; 11:beta&normal>1; 12:0&2normal>1;
                    * 13:3normal>0

        icode = 0   * 0:universal code; 1:mammalian mt; 2-11:see below
        Mgene = 0   * 0:rates, 1:separate;

    fix_kappa = 0   * 1: kappa fixed, 0: kappa to be estimated
        kappa = 2.05154 * initial or fixed kappa
```

```
   fix_omega = 0   * 1: omega or omega_1 fixed, 0: estimate
       omega = 0.4

       getSE = 0   * 0: don't want them, 1: want S.E.s of estimates
RateAncestor = 0   * (0,1,2): rates (alpha>0) or ancestral states (1 or 2)
  Small_Diff = .5e-6 * small value used in the difference approximation
                     * of derivatives

   cleandata = 0   * remove sites with ambiguity data (1:yes, 0:no)?
 fix_blength = 1   * 0: ignore, -1: random, 1: initial, 2: fixed
      method = 0   * 0: simultaneous; 1: one branch at a time
```

TABLE A9 – Nombre de duplications par branche *Simiiformes*

taxon	Meilleure moitié	Arbres sous l'erreur de 3,20 Ma	Tous
Aotus nancymaae	147	314	521
Callithrix jacchus	123	277	568
Catarrhini	890	1650	2796
Cebidae	38	72	112
Cebus capucinus	120	271	538
Cercocebus atys	16	23	70
Cercopithecidae	598	1024	1734
Cercopithecinae	229	333	575
Chlorocebus sabaeus	17	38	112
Colobus angolensis	62	121	270
Gorilla gorilla	90	140	266
Hominidae	75	122	218
Homininae	265	374	636
Hominoidea	308	458	697
Homo sapiens	175	252	460
HomoPan	41	62	111
Macaca	31	56	185
Macaca fascicularis	11	20	67
Macaca mulatta	19	46	143
Macaca nemestrina	7	17	54
Mandrillus leucophaeus	15	29	68
Nomascus leucogenys	83	154	304
Pan	30	46	112
Pan paniscus	10	19	64
Pan troglodytes	397	574	975
Papio anubis	23	61	168
Papionini	616	876	1456
Platyrrhini	557	1028	1485
Pongo abelii	74	114	229
Saimiri boliviensis	88	199	374
Simiiformes	100	298	739

TABLE A10 – Temps de génération moyens chez les primates, en jours, d'après @Pacifici2013.

Famille	Moyenne	Écart-type
Aotidae	3048	181
Atelidae	4825	1066
Callitrichidae	2371	347
Cebidae	5102	1214
Cercopithecidae	4035	578
Cheirogaleidae	2088	141
Daubentoniidae	2920	NaN
Galagidae	1877	285
Hominidae	8030	894
Hylobatidae	5374	247
Indriidae	3644	127
Lemuridae	3222	336
Lepilemuridae	1638	33
Lorisidae	2437	505
Pitheciidae	3174	295
Tarsiidae	2190	0

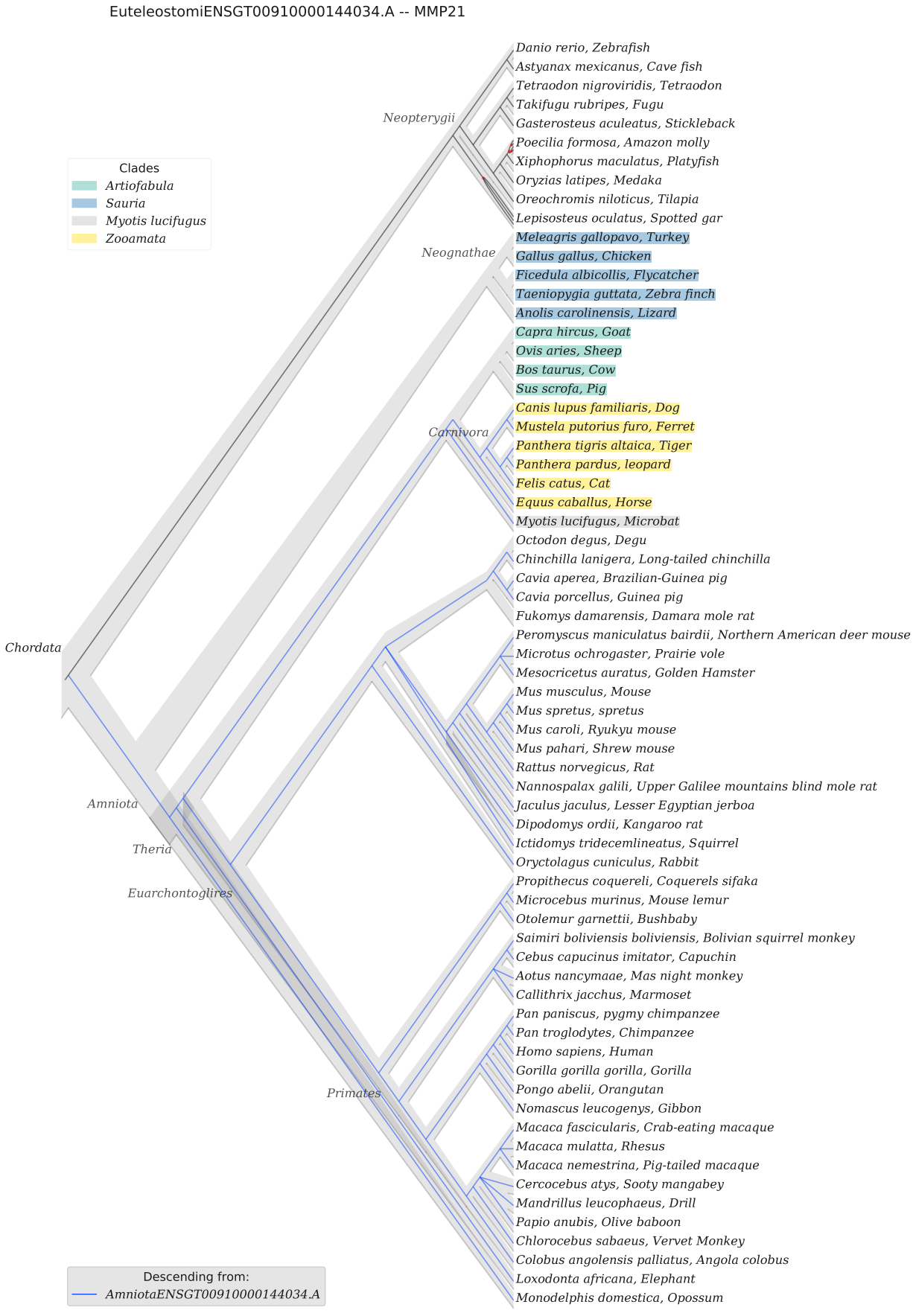


FIGURE A11a – Arbre réconcilié de MMP21, d'après Ensembl 93 avec édition à 0,35

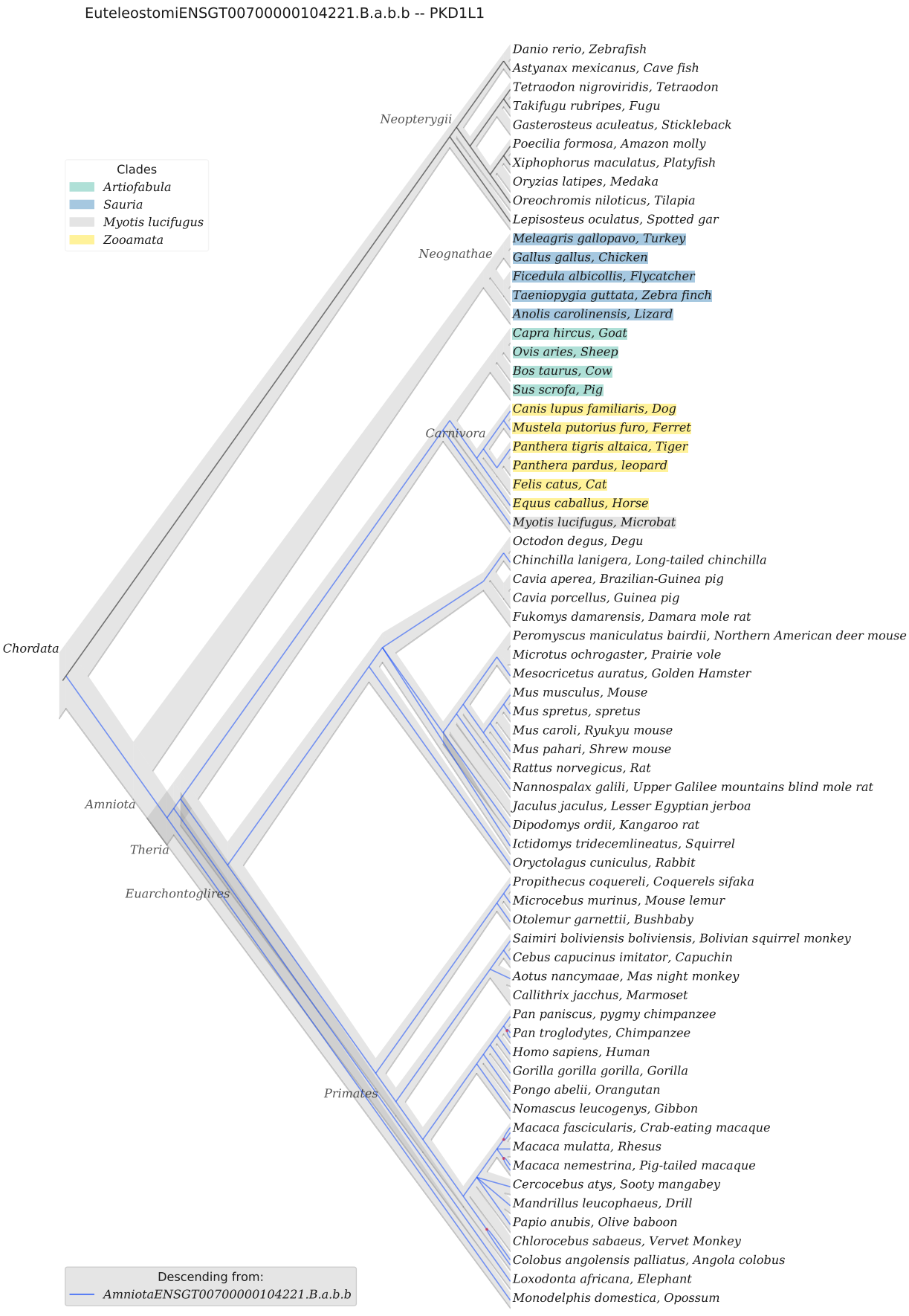


FIGURE A11b – Arbre réconcilié de PKD1L1, d'après Ensembl 93 avec édition à 0,35

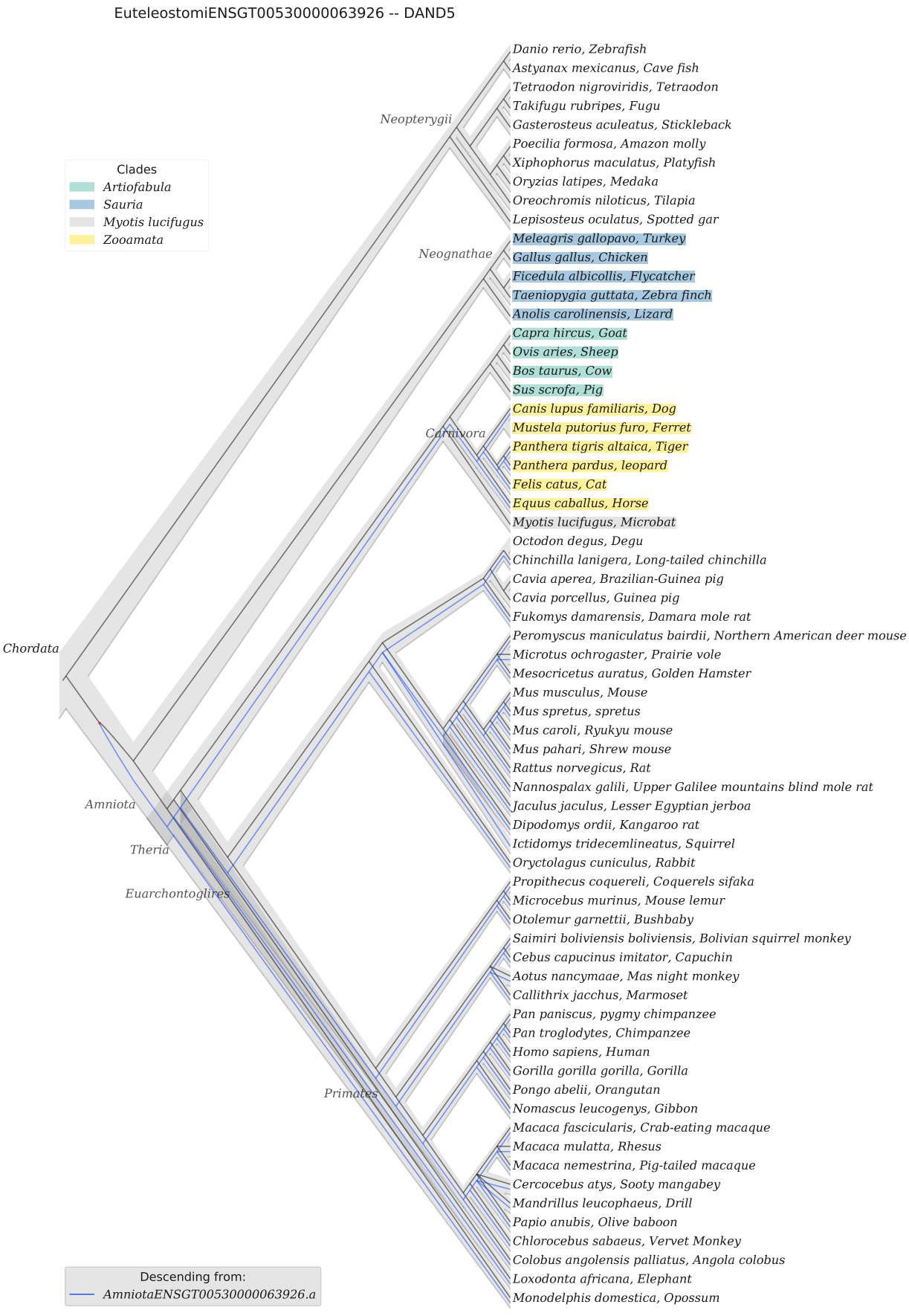


FIGURE A11c – Arbre réconcilié de DAND5, d'après Ensembl 93 avec édition à 0,35

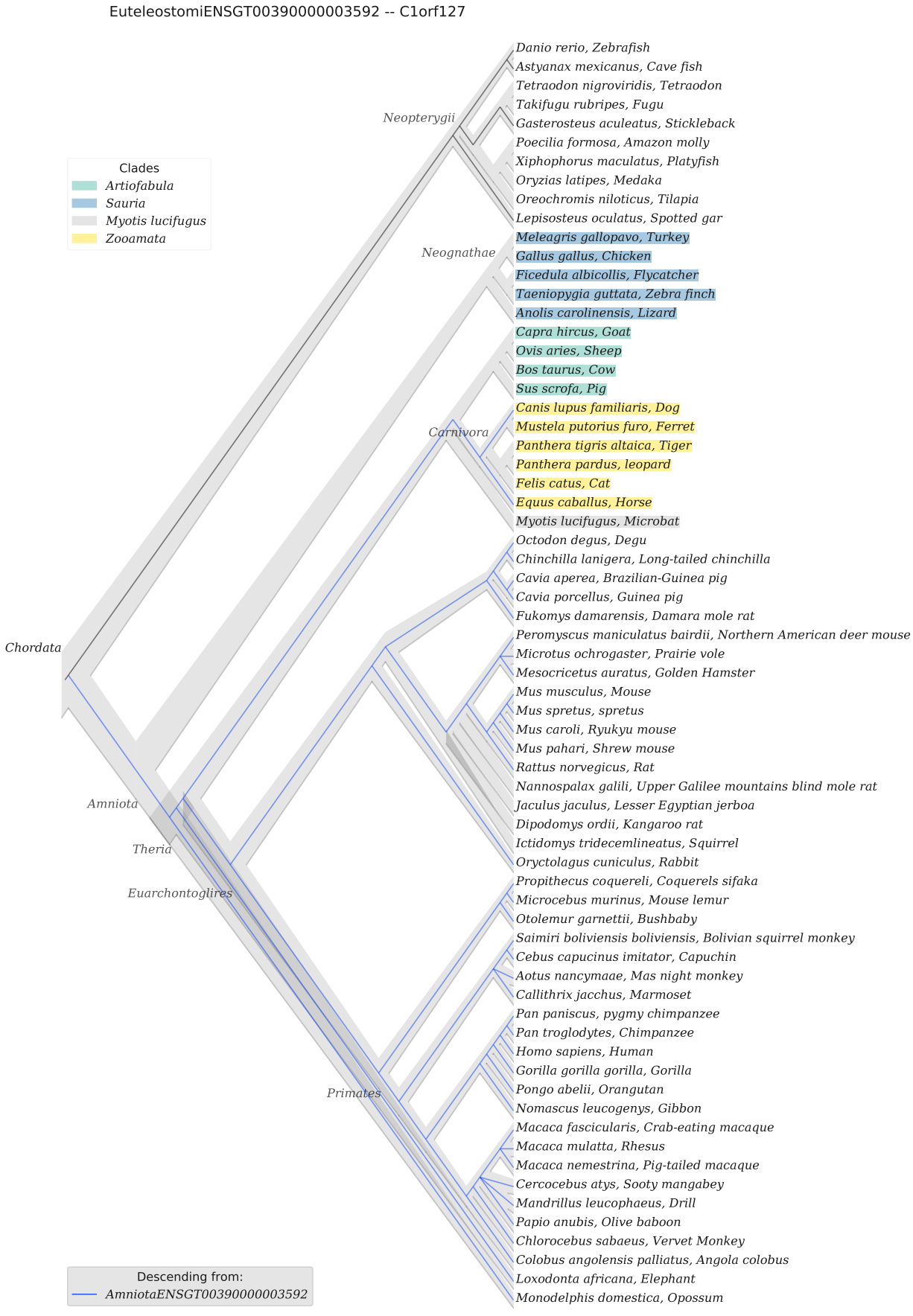


FIGURE A11d – Arbre réconcilié de C1orf127, d'après Ensembl 93 avec édition à 0,35

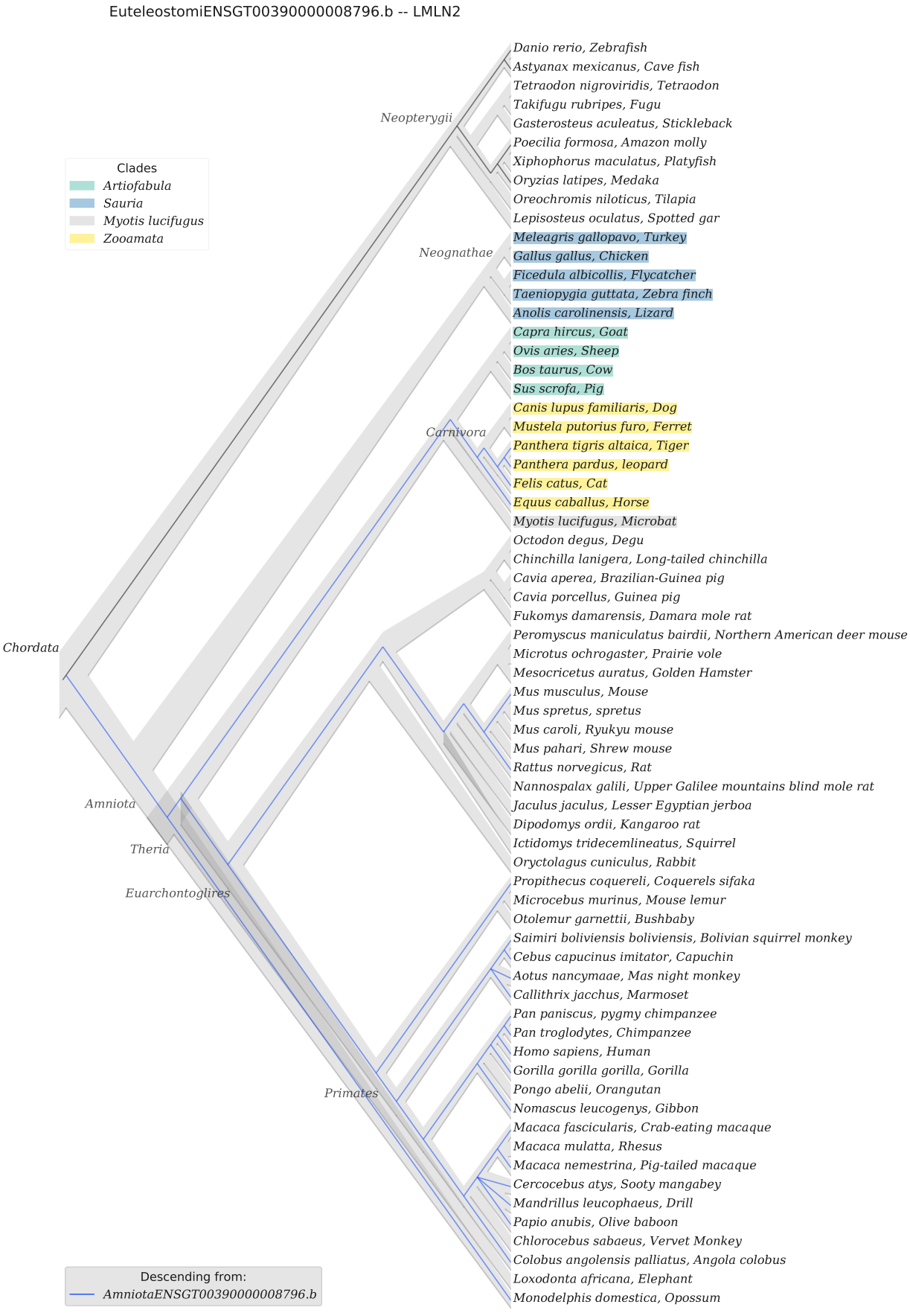


FIGURE A11e – Arbre réconcilié de LMLN2, d'après Ensembl 93 avec édition à 0,35

RÉSUMÉ

Ma thèse applique le concept d'horloge moléculaire à l'échelle de l'arbre de gène. L'arbre de gène ne se décalque pas exactement sur celui des espèces, car le nombre de copies fonctionnelles dans un même génome varie, soit par duplication sur un nouveau locus, soit par perte (pseudogénéisation ou délétion). Ces événements fréquents de duplication et perte sont cruciaux pour l'adaptation des organismes, en leur fournissant une grande plasticité génétique. Pour cette raison, j'ai travaillé sur l'ensemble des arbres de gènes d'une vingtaine d'espèces de primates, avec pour objectif de dater les duplications.

Par contraste avec la concaténation de plusieurs alignements, l'utilisation d'une seule famille génique impose une limite sur la puissance statistique. C'est ce que nous quantifions dans un premier temps, en effectuant un contrôle comparant les datations des spéciations dans les arbres de gènes avec les âges de référence des taxons. Avec ce contrôle nous sélectionnons une procédure de datation précise, qui optimise la qualité de l'alignement en amont. Nous déterminons la distribution de la précision de datation et l'associons ensuite à diverses caractéristiques mesurables sur les arbres de gènes et les alignements. Notre analyse confirme l'impact de la longueur d'alignement dans la précision, mais aussi de l'hétérogénéité des taux de substitution entre branches, qui est compliquée à accomoder par les modèles d'horloge moléculaire. Concrètement, notre stratégie permet de prédire un niveau de précision sur de nouvelles données, et nous l'appliquons aux datations de duplications.

À partir de cette prédiction de confiance sur les datations d'arbres avec duplications, nous sélectionnons le sous-jeu de meilleure qualité pour établir la distribution temporelle des duplications le long des lignées. Outre les dates, nous calculons les taux de duplications et caractérisons leur variation : ils sont en effet inégaux entre arbres de gènes, avec de nombreux arbres sans duplication et une faible proportion d'arbres se dupliquant beaucoup, ce qui peut se modéliser par une loi Gamma. De plus, le taux de duplication varie entre lignées d'organismes. Nous testons la corrélation phylogénétique entre taux de duplication génomique moyen par lignée, et diversification de cette lignée.

Enfin, des pertes de gènes impliqués dans la latéralisation de l'embryon caractérisent certains taxons de vertébrés. Nous établissons donc par corrélation de nouvelles séquences potentiellement fonctionnelles chez l'humain en criblant les gènes et enhancers montrant des pertes similaires.

Ainsi, après avoir évalué les méthodes appropriées pour une inférence fiable, nous avons caractérisé les dynamiques de renouvellement des gènes. Cette étape ouvre la voie pour comprendre l'association entre ces dynamiques génomiques et les dynamiques macroévolutives et l'adaptation ou la diversification des organismes.

MOTS-CLÉS

Génomique comparative; Horloge moléculaire; Datation; Réconciliation; Puissance statistique; Primates.

ABSTRACT

My PhD work applies the molecular clock concept at the scale of the gene tree. A gene tree does not match exactly a species tree, because the number of functional copies in a genome varies, either by duplication to a new locus, or by loss (pseudogenisation or deletion). These events of gain and loss are frequent and crucial to organismal adaptation, by providing genetic plasticity. Hence I worked on the whole set of gene trees of twenty primate species, and aimed at dating duplications.

By contrast with alignment concatenation, the use of a single gene family enforces a limit on statistical power. This is what we first quantify in performing a control comparing the speciation dates in gene trees with reference ages of the taxa. With this control we select an accurate dating procedure, which optimizes upstream the quality of the alignment. We determine the distribution of the dating accuracy and then associate it with various measurable characteristics on the gene trees and alignments. Our analysis confirms the impact of the alignment length in the accuracy, but also of the heterogeneity of the substitution rates between branches, which is complicated to accommodate by molecular clock models. In concrete terms, our strategy allows us to predict a level of accuracy on new data, and we apply it to the duplication dates.

From this confidence prediction on dating trees with duplications, we select the best quality subset to establish the temporal distribution of duplications along lineages. In addition to the dates we calculate the duplication rates and characterise their variation: indeed it differs substantially between gene trees, with many trees without duplications and a low proportion of trees that duplicate a lot, which can be modeled by a Gamma law. Moreover, the duplication rate varies between organism lineages. We test the phylogenetic correlation between average genomic duplication rate per lineage, and diversification of this lineage.

Finally, the loss of genes involved in the lateralization of the embryo is characteristic of certain vertebrate taxa. We therefore determine by correlation new sequences that are potentially functional in humans, by screening for genes and enhancers showing similar losses.

Thus, after evaluating the appropriate methods for reliable inference, we have characterised the dynamics of gene turnover. This paves the way to understanding the association between these genomic dynamics and the adaptation and diversification of organisms.

KEYWORDS

Comparative genomics; Molecular clock; Dating; Reconciliation; Statistical power; Primates.