Université Montpellier II
Sciences et Techniques du Languedoc

THÈSE

pour obtenir le grade de

Docteur de l'Université Montpellier II

Discipline : Biologie des Populations et Écologie
Formation Doctorale : Biologie de l'Évolution et Écologie
École Doctorale : Biologie des Systèmes Intégrés, Agronomie - Environnement

présentée et soutenue publiquement

par

Jonathan Romiguier

le 22 novembre 2012

Titre :

Phylogénomique et stratégies d'histoire de vie des mammifères placentaires : apports de la théorie de la conversion génique biaisée.

JURY

M. Jacques David, Professeur, SupAgro, Montpellier	Président
M. Henrik Kaessmann, Professeur, Université de Lausanne	Rapporteur
M. Toni Gabaldón, Directeur de recherche, CGR, Barcelone	Rapporteur
M. Gabriel Marais, Directeur de recherche, CNRS, Lyon	Examinateur
M. Emmanuel Douzery, Professeur, CNRS, Montpellier	Invité
M. Nicolas Galtier, Directeur de recherche, CNRS, Montpellier	Directeur
M. Vincent Ranwez, Professeur, SupAgro, Montpellier	Directeur

Phylogenomic and life-history strategies of placental mammals : insights from the biased gene conversion theory.

From mice to whales through humans, placental mammals present a stunning diversity. Despite being one of the most studied group ever, mysteries persist about their origin. Indeed, their most basal relationships still remain uncertain, and nothing is really known about the lifestyle of our cretaceous ancestors, these placental mammals which lived side by side with non-avian dinosaurs during 30 My.

To answer these evolutionnary questions, comparative genomic studies of placental mammals have been conducted. One of its originalities is to take into account biased gene conversion. Rigging the genetic lottery, this recombination-associated mechanism involves a reparation bias favouring the G and C nucleotides over the A and T ones, which mark the mammalian genomic landscapes by inducing localized peaks of GC-content.

This phenomenon has been so far studied in few model species. The exploration of biased gene conversion in more than 30 mammal genomes led to several key results. In particular, GC content evolution has proved to be correlated to the longevity and the body mass of species. By linking together molecular evolution and life history traits, the reconstruction of ancestral sequences allowed us to estimate a life-span above 25 years for early placental mammals. This value is markedly different from that of mice or shrews, although our mammalian ancestors have often been represented as such.

In addition to these results, GC-rich genes were found to be prone to produce false phylogenies. Less affected by recombination associated artifacts, AT-rich genes are shown to be more reliable, and to support species of African origin as the sister group of all other placental mammals - perhaps resolving one of the most controversial nodes of the mammalian tree.

From nucleotide to the birth of a 4,000 species infraclass, this work reveals how molecular evolution can shed new light on our deepest origins.

Keywords : Placental mammals, Comparative genomic, Isochore and nucleotidic landscapes, Ancestral life history traits, Phylogenomic, Biased gene conversion.

Phylogénomique et stratégies d'histoire de vie des mammifères placentaires : apports de la théorie de la conversion génique biaisée

Des souris aux baleines en passant par les humains, la diversité écologique des mammifères placentaires est des plus fascinantes. Bien qu'il s'agisse là d'un des groupes les plus étudiés, leur origine fait pourtant l'objet de bien des mystères. Leurs relations de parenté les plus basales restent en effet incertaines, et l'on ignore encore beaucoup du mode de vie qu'avaient nos ancêtres du Crétacé, ces mammifères placentaires qui auraient côtoyé les dinosaures pendant plus de 30 millions d'années.

Afin d'aborder ces questions, cette thèse a utilisé l'outil de la génomique comparative. L'une de ses principales originalités est la prise en compte d'un distorteur majeur de notre évolution moléculaire : la conversion génique biaisée. Truquant la loterie génétique, ce mécanisme associé à la recombinaison méiotique avantage les nucléotides G et C au détriment des nucléotides A et T. Façonnés par son influence, nos paysages nucléotidiques présentent ainsi ponctuellement des taux de GC anormalement élevés.

Jusque là, ce phénomène n'avait été étudié que chez une poignée d'organismes modèles. Son analyse chez plus d'une trentaine de génomes mammaliens a mis en évidence une série de résultats clés. En particulier, l'évolution du contenu en GC des gènes s'est avéré dépendre de la masse corporelle et la longévité des espèces. En reliant ainsi évolution moléculaire et traits d'histoire de vie, des reconstructions de séquences ancestrales ont permis d'estimer la durée de vie des premiers mammifères placentaires à plus de 25 ans. Cette longévité va bien au delà de ce que peuvent espérer atteindre les souris ou musaraignes actuelles, des animaux au mode de vie pourtant jusqu'ici supposé proche de celui de nos ancêtres.

Parallèlement à ces résultats, une tendance à produire des phylogénies inexactes a été detectée chez les gènes les plus GC-riches. Moins soumis à la conversion génique biaisée, les gènes AT-riches se sont montrés plus fiables, tout en soutenant que les espèces originaires d'Afrique sont situés à la base de l'arbre des placentaires. Ce résultat suggère ainsi la possible résolution d'un des noeuds les plus controversés de notre histoire évolutive.

Du simple nucléotide à la naissance d'une infraclasse de plus de 4000 espèces, ce travail révèle comment l'évolution moléculaire peut porter un nouveau regard sur nos origines les plus profondes.

Mots-clés : Mammifères placentaires, Génomique comparative, Isochores et paysages nucléotidiques, Traits d'histoire de vie ancestraux, Phylogénomique, Conversion génique biaisée.

Remerciements

Merci aux membres du jury de m'avoir fait l'honneur d'accepter d'évaluer ce travail.

Evidemment, je ne remercierai jamais assez Nicolas, Vincent et Emmanuel pour m'avoir encadré durant ces trois ans de doctorat. Travailler ensemble m'a toujours paru facile et plaisant, ce qui a largement contribué à faire de cette thèse un moment agréable de ma vie.

Merci à mes co-autheurs, Julien Dutheil, Bastien Boussau et Frédéric Delsuc. Un merci tout particulier à Julien pour son aide inestimable lors de ma première année. Personne n'a jamais mieux su résoudre mes problèmes existentiels d'apprenti bio-informaticien. Merci également à Khalid Belkhir pour l'administration du cluster de calcul sans lequel rien n'aurait été possible.

Merci à tous ceux qui m'ont apporté leur aide et orienté au grès de comités de thèses (Laurent Duret, Nicolas Lartillot, Renaud Vitalis), stage de master (Johan Michaux, Marie Pagès) ou autre (Rumsais Blatrix, Doyle McKey).

Merci à Emeric Figuet que j'ai eu la chance de co-encadrer lors de deux stages, et qui a largement contribué à une partie de ce travail.

Merci à Nicolas Faivre et Marion Ballenghien pour m'avoir assisté lors de mes rares escapades au labo. Merci également au reste de l'équipe (Vincent, Aurélien, Etienne, Phillipe, Sylvain, Benoît, Camille, Fidel, Céline, Laurana...) pour ces nombreux repas et débats partagés.

Merci à mes camarades et amis doctorants, Laure, Arnaud, Anthony, Laurana, Anusha et tout le reste des membres de l'association de vulgarisation scientifique que j'ai eu le plaisir de fonder avec eux. Merci également à Rolando et Vivi pour les pauses et soirées nécessaires à l'équilibre du thésard.

Je tiens également à remercier l'Université Montpellier 2 dans laquelle j'ai réalisé l'intégralité de mes études supérieures. Merci aux enseignants qui ont jalonné mon parcours (E. Douzery, J.B. Ferdy, B. Godelle, I. Olivieri, C. Moulia, M. Raymond...), et ainsi influencé sans le savoir ce parcours académique qui est le mien.

Enfin, merci à ma famille, mes proches, moins proches et tous ceux que j'aurai pu oublier.

Table des matières

Première partie

Introduction

For english readers :

Just like a lot of PhD students, one of the most common question I heard during the last three years was "What are you working on ?". This simple question was a pleasure to address when asked by a specialist, but could turned into a nightmare otherwise. Whether it be for colleagues from other fields, undergraduate students during my teaching activities or with relatives and friends, explaining my work in a nutshell has been a tedious task. For all these people who did not understand why studying genomic GC-content is exciting in itself and how it can help us to better understand the evolution of mammals, I wrote the first two chapters of this thesis report. For a specialist, this introductive part, which was written in my mother tongue, is not essential to understand the research I conducted. It contains a review about mammalian evolution and nucleotidic composition evolution, structured as follows :

The first part is focused on mammals :

-**Part 1.1** : A brief overview of evolution, phylogeny and the place of human and mammals in the living world.

-**Part 1.2** : Generalities about mammals and the origin of Placentalia species, traditionnaly related to the famous Cretaceous-Tertiary mass extinction, 65 My ago.

-**Part 1.3** : A small introduction to molecular phylogeny and its advantages.

-**Part 1.4** : A comparison of the morphological and molecular phylogenies of mammals, the molecular time estimations which predates the Cretaceous-Tertiary boundary for the origin of modern placental species, and the conflicting hypothesis about the root position of their tree.

The second part is focused on molecular evolution and nucleotidic composition landscapes :

-**Part 2.1** : A general introduction about the concepts of neutralism, selfish DNA and the units and levels of selection.

-**Part 2.2** : A presentation of the nucleotidic landscapes of mammals, the so-called GC-rich isochores and GC-poor isochores.

-**Part 2.3** : Adaptative and neutral hypothesis for the origin of isochores.

-**Part 2.4** : A presentation of the biased gene conversion hypothesis, the evidence for such a non-adaptative mechanism, and its evolutionnary consequences on molecular maladaptations and methodological biases in selection detection methods.

-**Part 2.5** : The link among GC-content, biased gene conversion, recombination, chromosome number and genome size, including a so far unpublished work of mine regarding the isochore origin an maintenance in non-mammal Vertebrates.

-**Part 2.6** : Published links between molecular evolution and life history traits in mammals.

When relevant, most of this content is presented in the introduction of the corresponding reseach articles that constitutes the core of this PhD report.

Chapitre 1

Origine et diversification des mammifères placentaires

1.1 Pourquoi s'intéresser à nos origines ? Evolution de la place de l'homme dans le vivant

Quelque soit sa culture, l'homme a toujours tenté d'élucider l'énigme de sa propre origine. Sésame indispensable pour esquisser un sens à sa vie et au monde qui l'entoure, il fit émerger de cette préoccupation majeure les premières religions, matrices des plus grandes œuvres, guerres et merveilles architecturales de l'humanité. Chaque mythologie possède sa propre cosmogonie, ensemble de mythes fondateurs mettant en scène divers bestiaires fantastiques et une ou plusieurs divinités, êtres par essence supérieurs. Cette notion générale d'êtres supérieurs et inférieurs influencera longtemps la pensée occidentale. Dès l'Antiquité, Aristote propose en ce sens d'établir une gradation des être vivants. Il relègue ainsi les espèces les plus simples aux échelons inférieurs, tandis qu'il fait trôner l'espèce humaine près du sommet de cette échelle des êtres. Cette vision perdurera par ailleurs des siècles, en témoigne cette fameuse représentation d'une *scala natu-*

rae adaptée au folklore judéo-chrétien (Figure 1.1). Ici, toute les espèces sont vues comme ayant toujours existé, fixées depuis leur création telle qu'elle est relatée dans le Livre de la Genèse. On parle alors de "fixisme". Les frontières entre les divers échelons sont immuables, sans aucune notion de changement ou d'évolution.

Cette limite bien nette entre humanité et le reste des êtres vivants s'effritera à partir du siècle des Lumières. Plusieurs philosophes et scientifiques français contribuèrent à mettre en place la notion d'évolution. Citons en particulier Pierre Maupertuis (1698-1759), Denis Diderot (1713-1784), Georges-Louis Buffon (1707-1788), ou Jean-Baptiste Lamarck (1744-1829). Bien que ces derniers aient été les précurseurs de cette révolution idéologique, ce sont les britanniques Charles Darwin (1809-1892) et Alfred Wallace (1823-1913) qui donnèrent à la théorie de l'évolution sa forme quasi-finale. Depuis, les espèces ne sont plus vues comme étant indépendantes les unes des autres, mais reliées entre elles par des relations de parenté. Les espèces évoluent et se transforment en d'autres, on parle de "transformisme". Ernst Haeckel est le premier à avoir popularisé la représentation de ce phénomène sous la forme d'un arbre (Figure 1.2). On peut y voir l'homme désormais apparenté à d'autres animaux, les mammifères, définis en 1758 par Carl Von Linné comme le groupe des animaux qui allaitent leur progéniture. Héritage de siècles d'anthropocentrisme, ce groupe est situé tout en haut de cet arbre du vivant. L'évolution y est faussement dépeinte comme dirigée vers une cime représentée par l'homme, un idéal subjectif. Figées à un stade plus primitif, les autres espèces sont elles symbolisées par de simples branches latérales.

La science qui étudie la généalogie des espèces s'appelle aujourd'hui la phylogénie, et considère que toutes les espèces actuelles sont au même niveau d'évolution. Contrairement à l'adage populaire, le chimpanzé n'est pas l'ancêtre de l'homme, mais bien son plus proche cousin. Ils ne sont pas reliés par un lien de descendance : leur ancêtre commun a disparu, et chacune des deux espèces a suivi sa propre histoire. Cette évolution, buissonnante et non pas dirigée, est

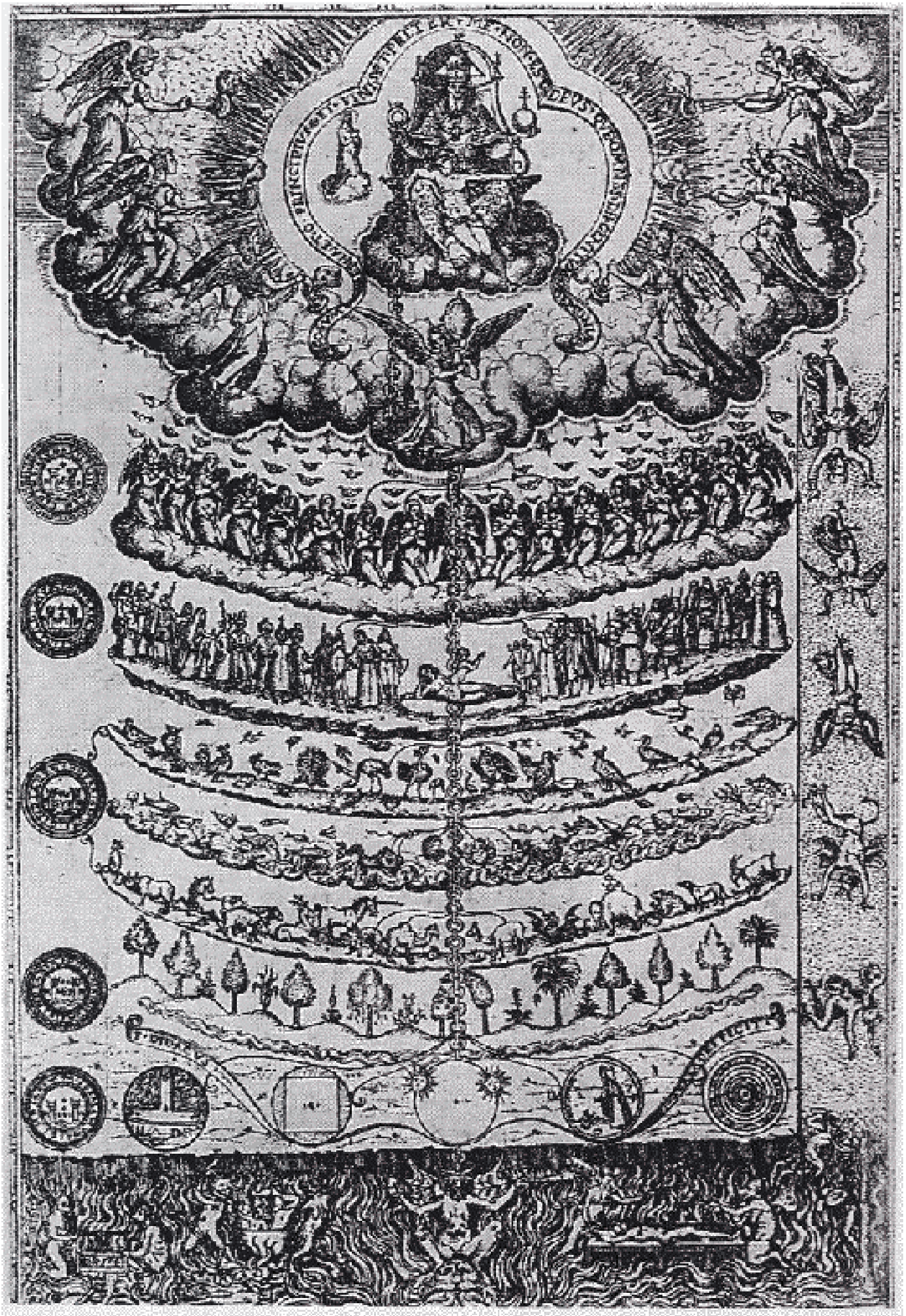


FIGURE 1.1 – L'échelle des êtres, ou *scala naturae*, telle que figurée dans *Rhetorica Christiana* de Diego Valadés (1579). Le monde minéral est en bas de l'échelle, suivi par les végétaux, les animaux, les hommes, les anges et Dieu. Image dans le domaine public.

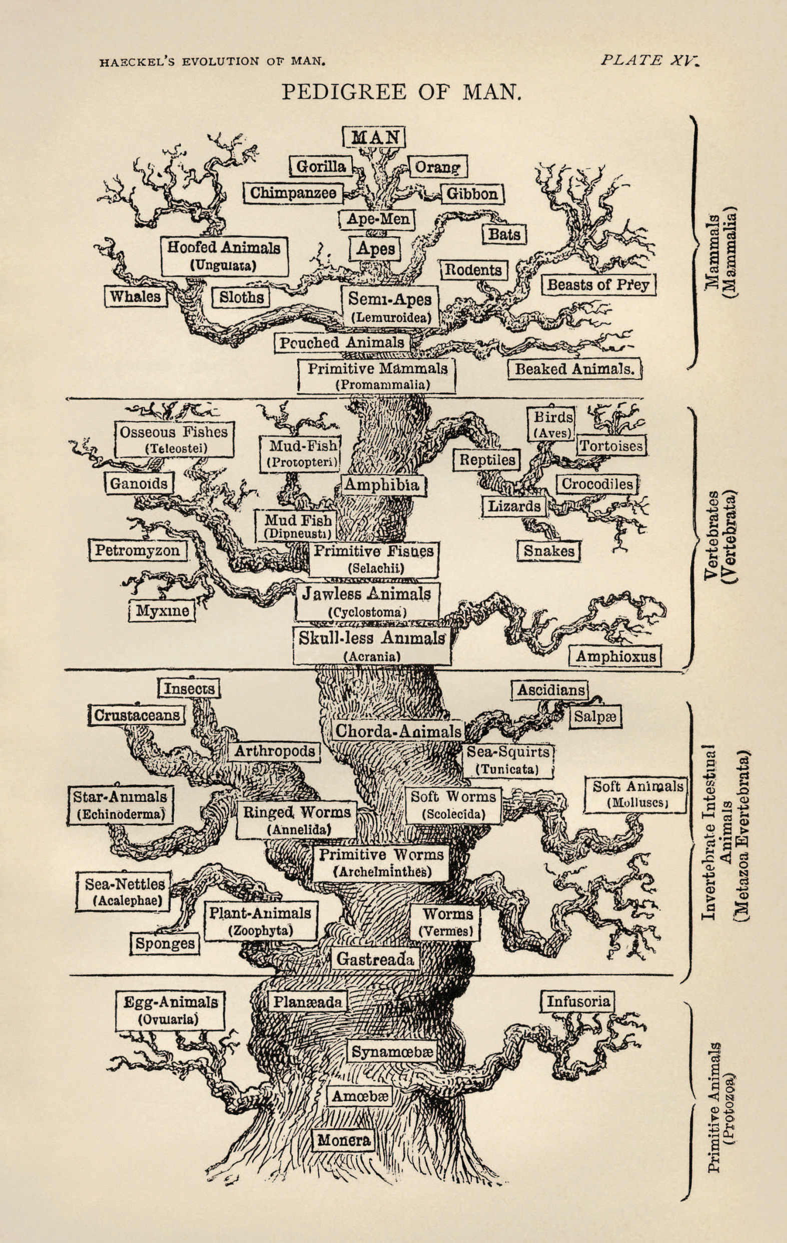


FIGURE 1.2 – La lignée humaine de Ernst Haeckl (1874).

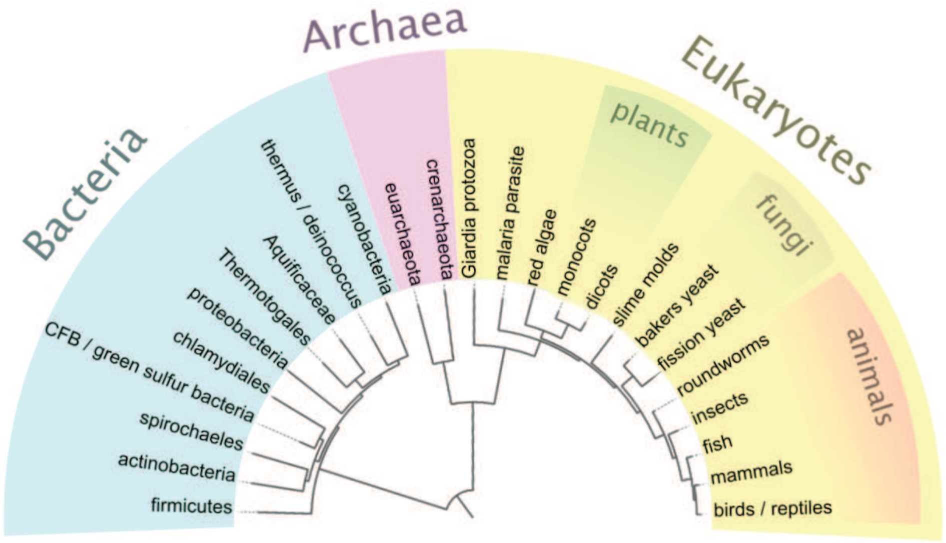


FIGURE 1.3 – Arbre phylogénétique du vivant. Représentation simplifiée obtenue à partir des informations disponibles dans "The interactive Tree of Life" [Ciccarelli et al., 2006].

parfaitement bien illustrée par les arbres du vivant les plus modernes (Figure 1.3). Dans cet arbre phylogénétique, chaque branche représente la persistance d'une lignée à travers le temps, les nœuds la naissance d'une nouvelle espèce (un événement de spéciation). L'arbre se sépare à la base en trois grands groupes majeurs, les eubactéries, les archées et les eucaryotes. Les mammifères, et par conséquent l'homme, se situent au milieu des eucaryotes, sans positionnement particulier pouvant les distinguer des autres êtres vivants.

En comparant les figures 1 à 3, on peut ainsi observer la descente progressive de l'homme de son piédestal initial. D'abord supposé supérieur à toute autre forme de vie, son histoire est ensuite confondue avec celle de ses proches parents, les mammifères. Ces derniers sont d'abord considérés comme un modèle de complexité vers lequel tendrait l'évolution des espèces, pour finir comme une simple voie parmi d'autres qui aurait été prise par l'évolution.

Cette dernière représentation est cependant peu connue du grand public.

Cette incompréhension est le terreau de diverses théories, dites créationnistes, rejetant catégoriquement l'idée que l'évolution puisse ne pas être dirigée par un dessein divin aboutissant nécessairement à l'espèce humaine. Bien qu'inexacte, la vision d'Haeckel (Figure 1.2) est donc toujours bel et bien présente dans l'inconscient collectif. Elle dépeint ainsi les mammifères comme un groupe qui trouverait son aboutissement dans les grands singes, puisant ses racines vers un tronc d'où émergent des branches inférieures telle que celle des rongeurs. Mais que sait-on au juste sur l'origine des mammifères ? A quel point cette intuition confortable de nos précurseurs est elle éloignée de la réalité scientifique actuelle ? Parce que s'intéresser à l'origine des mammifères, c'est s'intéresser à nos origines, le sujet a toujours suscité beaucoup d'attention. Objet principal de cette thèse, leur histoire évolutive sera passée en revue au cours des prochaines parties.

1.2 Origine des mammifères placentaires : généralités

Les mammifères regroupent environ 5400 espèces [Wilson and Reeder, 2005]. Définis pour la première fois par Carl Von Linné en 1958, ils sont notamment caractérisés par la possession de poils et de glandes mammaires. Bien qu'ayant hérité de ces caractéristiques d'un seul et même ancêtre commun, les mammifères disposent d'une des plus grandes diversités de formes et d'écologies du monde animal. D'une masse allant de 2g (chauve-souris bourdon) à près de 200 tonnes (baleine bleue), ils peuvent être de type aquatique, terrestre, fouisseur, arboricole ou volant.

Presque toute l'étendue de cette diversité est contenue dans une seule sous-classe, celle des placentaires (aussi appelés euthériens[1]). Ils rassemblent plus de

[1] Si le terme euthérien désigne bien les même espèces actuelles, il prend un sens différent en paléontologie. Placentaire inclut tous les descendants de l'ancêtre commun des placentaires actuels. Eutheria rassemble placentaires + les lignées éteintes externes qui restent plus apparentés aux placentaires qu'aux Marsupiaux.

5000 espèces (groupées en divers ordres, tels que les primates, rongeurs ou carnivores), et tiennent leur nom du placenta, un organe qui permet le développement complet de l'embryon dans le corps de la mère. Ce développement est en revanche incomplet chez les marsupiaux, dont le nom provient de leur poche marsupiale. L'embryon peut y terminer son développement, qui contrairement aux placentaires, est incomplet à la naissance. Bien que moins diversifiés que les placentaires (près de 300 espèces), ils présentent une grande variété (kangourou, koala, souris marsupiale, taupe marsupiale, opossum, diable de tasmanie...). La troisième sous-classe, les monotrèmes, ne comportent que 5 espèces (4 echidnés et l'ornithorynque). Ils présentent la caractéristique étrange (pour des mammifères) d'être ovipares, bien qu'allaitant leurs petits. Ils seraient ainsi les premiers à avoir divergé du tronc commun des mammifères, faisant des marsupiaux et des placentaires, tous deux vivipares, les deux groupes les plus apparentés.

De ces trois sous-classes, ce sont les placentaires qui ont toujours suscité le plus d'attrait. Parce que leur origine est directement reliée à celle de l'homme, celle-ci a cristallisé l'une des histoire les plus racontées en biologie évolutive. Tous le monde a déjà entendu parler de cette extinction brutale des dinosaures, qui il y a 65.5 Ma aurait permis à des mammifères, jusque là restreints à de petites tailles, de se diversifier rapidement pour aboutir aux espèces actuelles. Cette extinction massive est appelée crise K-Pg (pour Crétacé-Paléogène). Elle est reconnue comme l'une des plus dévastatrices de l'histoire du vivant, et marque la fin du règne sans partage des dinosaures non-aviens (dinosaures à l'exception des oiseaux). Celle-ci est souvent attribuée à l'impact d'un astéroïde dans la péninsule mexicaine du Yuccatan [Schulte et al., 2010]. Les modalités exactes de cette extinction font cependant toujours débat. Certains privilégient une cause multifactorielle (changements climatiques, retrait des eaux et activité volcanique) accompagnée d'une extinction graduelle [Macleod et al., 1997, Keller et al., 2004, Archibald et al., 2010, Keller et al., 2010]. D'autres privilégient plutôt la thèse d'un événement brutal, cataclysmique, qui aurait décimé en quelques heures toutes les espèces incapables de survivre à la déflagration en s'enterrant ou s'immergeant sous l'eau [Robertson et al., 2004].

Quelles qu'en soient les causes, le registre fossile supporte l'idée selon laquelle les placentaires actuels doivent tous leur origine à cet événement [Archibald et al., 2001]. Pour les paléontologues, les niches écologiques libérées par les dinosaures non-aviens auraient de plus permis d'accompagner cette diversification explosive des placentaires par une augmentation de leur taille moyenne [Smith et al., 2010]. Largement vulgarisée dans d'innombrables manuels scolaires, documentaires ou académiques [Dawkins, 2004, Feldhamer et al., 2007], cette épopée évolutive des placentaires a cependant été partiellement remise en cause. Ce revirement est dû à l'émergence d'une nouvelle discipline, la phylogénie moléculaire, qui a également largement bouleversé la systématique de ce groupe si étudié.

1.3 L'outil moléculaire : la machine à remonter le temps

La phylogénie est la science qui étudie les relations de parenté entre espèces. Historiquement, elle s'y atelle en comparant les différences morphologiques, anatomiques ou histologiques des espèces qu'elle étudie. Cette manière de procéder est cependant tombée en désuétude depuis l'avènement de la phylogénie moléculaire. Pour établir la généalogie des espèces, cette branche de la phylogénie analyse directement les différences contenues dans les molécules de l'hérédité, l'ADN. Nos tests de paternité modernes reposent par ailleurs sur ce même type d'information.

C'est en 1962 que Watson et Crick reçoivent le prix Nobel de médecine pour avoir mis au point le modèle moléculaire de l'acide désoxyribo-nucléique (ADN), support de l'hérédité et de l'information génétique [Watson and Crick, 1953]. Formée de deux brins complémentaires qui s'enroulent en double-hélice, elle est composée de 2 paires de nucléotides : l'Adénine (A) toujours appariée à la Thymine (T), et la Guanine (G) appariée à la Cytosine (C) (Figure 1.4). L'information génétique est ainsi résumée par l'enchaînement particulier de ces 4 "bases" de l'ADN : A,C,G et T. On parle alors de séquence ADN (ex : ATGC-CGTAG).

Codant la composition et la structure de toutes les protéines du vivant, ces séquences ont mémorisé l'histoire évolutive des organismes. Dérivant toutes d'un seul ancêtre commun, chaque espèce a ainsi accumulé son propre lot de mutations[2]. Ainsi, plus deux espèces sont proches (plus leur dernier ancêtre commun est récent), plus elles auront tendance à disposer de séquences génétiques similaires. A l'aide d'un alignement de séquences (Figure 1.5), on peut comparer la séquence d'un même gène chez différentes espèces, et ainsi établir leurs relations de parenté en inférant un arbre phylogénétique. Ce raisonnement est le fondement de la phylogénie moléculaire. L'ADN étant une caractéristique commune à l'ensemble des êtres vivants cellulaires connus à ce jour, c'est cette discipline qui a permis d'établir l'arbre du vivant moderne (Figure 1.3). Si l'on suppose un taux constant de mutations à travers le temps (hypothèse d'horloge moléculaire), on peut même dater les divers événements de spéciation à l'aide d'un ou plusieurs points de calibration fossiles.

L'inférence d'un arbre à partir de données moléculaires peut se faire à l'aide de plusieurs méthodes [Yang and Rannala, 2012]. La première méthode utilisée est issue des analyses morphologiques. Elle consiste à reconstruire l'arbre phylogénétique qui minimise le nombre de changements de caractères (ici, les mutations), autrement dit, le plus parcimonieux. Cette méthode dite de parcimonie est cependant aujourd'hui supplantée par les méthodes dites probabilistes, dont l'utilisation s'est rapidement généralisée avec l'augmentation rapide des données moléculaires. Ces méthodes sont particulièrement bien adaptées aux caractéristiques des séquences moléculaires qui disposent de nombreux caractères (sites) pouvant être considérés comme la réalisation de variables aléatoires. Ainsi, la reconstruction phylogénétique peut être vue comme un problème d'es-

[2] Modification quelconque de la séquence, souvent due à une erreur de copie lors de sa réplication. Les plus simples, dites ponctuelles, sont les substitutions (changement de lettres), les délétions (oubli de lettres) et les insertions (rajouts de lettres).

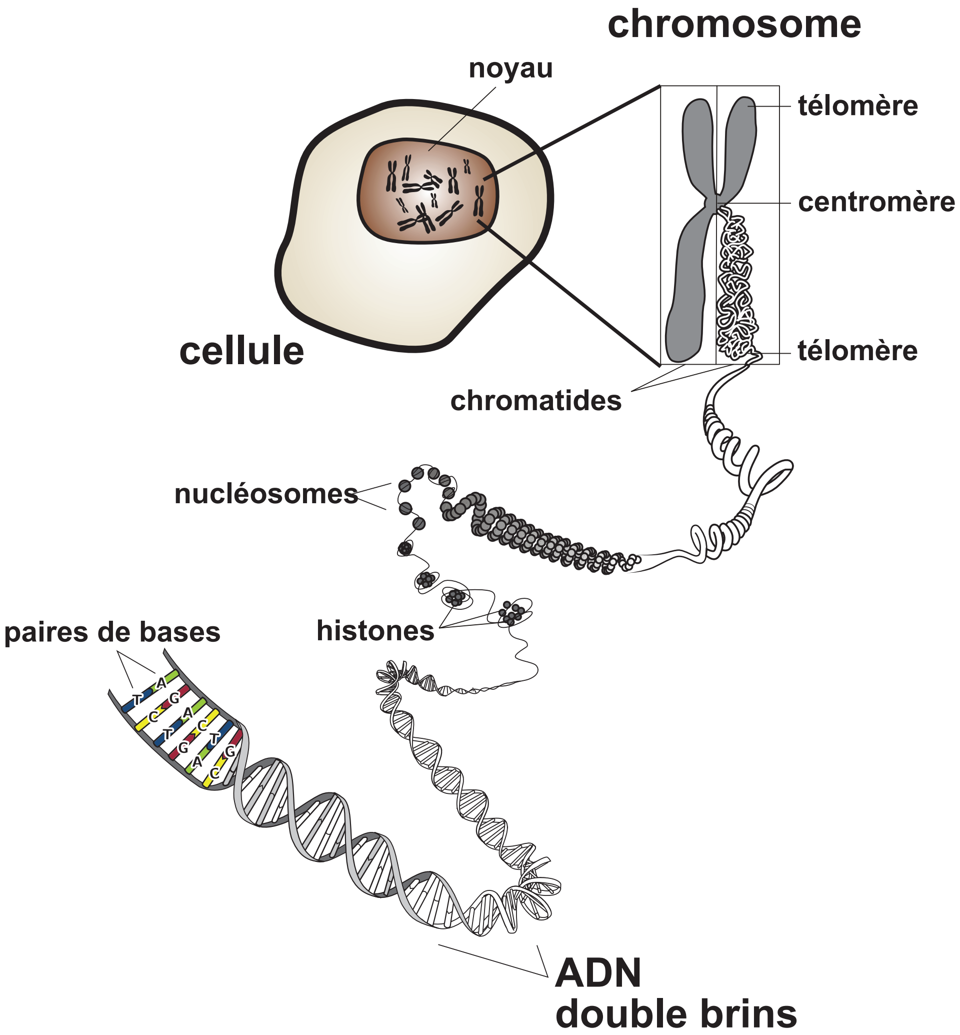
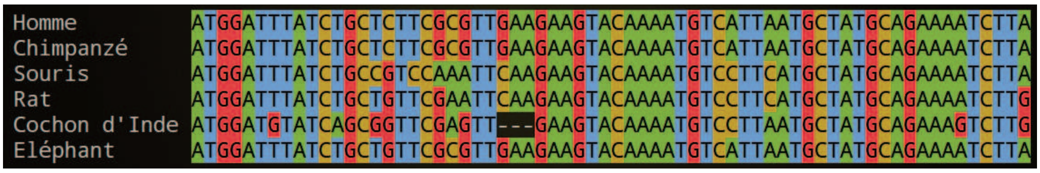


FIGURE 1.4 – Organisation de l'ADN au sein des chromosomes. Image du domaine public.



```
Homme          ATGGATTTATCTGCTCTTCGCGTTGAAGAAGTACAAAATGTCATTAATGCTATGCAGAAAATCTTA
Chimpanzé      ATGGATTTATCTGCTCTTCGCGTTGAAGAAGTACAAAATGTCATTAATGCTATGCAGAAAATCTTA
Souris         ATGGATTTATCTGCCGTCCAAATTCAAGAAGTACAAAATGTCCTTCATGCTATGCAGAAAATCTTA
Rat            ATGGATTTATCTGCTGTTCGAATTCAAGAAGTACAAAATGTCCTTCATGCTATGCAGAAAATCTTG
Cochon d'Inde  ATGGATGTATCAGCGGTTCGAGTT---GAAGTACAAAATGTCCTTAATGCTATGCAGAAAGTCTTG
Eléphant       ATGGATTTATCTGCTGTTCGCGTTGAAGAAGTACAAAATGTCATTAATGCTATGCAGAAAATCTTA
```

FIGURE 1.5 – Exemple d'alignement : les 66 premières positions du gène BRCA1. On remarque la forte similitude des séquences de l'homme et du chimpanzé, très proches parents. Sur la séquence du cochon d'inde, on peut noter un exemple de délétion de 3 positions nucléotidiques : un codon.

timation statistique, ce qui requiert des modèles d'évolution des séquences [Felsenstein, 1988]. Un modèle regroupe plusieurs paramètres qui vont résumer au mieux le comportement d'une séquence au cours de son histoire évolutive. Les modèles de substitution nucléotidique sont stochastiques et "sans mémoire", on les dit Markoviens d'ordre 1. Les taux pour chaque type de substitution (A vers C, A vers G...) sont donnés comme paramètres du modèle. Le modèle le plus simple, Jukes-Cantor 69, ne prend en compte qu'un seul taux global de substitution [Jukes and Cantor, 1969] ; les plus complexes ont des taux de substitution différents pour chaque type de substitutions [Tavaré, 1986]. Une fois le modèle choisi, les valeurs de ses divers paramètres sont estimées à l'aide de la fonction de vraisemblance. De manière générale, cette vraisemblance est définie comme étant la probabilité d'observer des données sachant des hypothèses. Dans un cadre de phylogénie moléculaire, les données sont un alignement de séquences associé à un modèle d'évolution. Les hypothèses rassemblent les valeurs de paramètres du modèle et l'arbre phylogénétique que l'on cherche à inférer. Ainsi, si l'on procède par maximum de vraisemblance, on retiendra l'arbre qui maximisera la probabilité d'observer notre alignement de séquences. Ce choix de l'arbre retenu peut aussi se faire via une méthode alternative popularisée ces dernières années, l'inférence bayésienne [Yang and Rannala, 2012], elle aussi basée sur la modélisation Markovienne de l'évolution des séquences.

Plus encore que le raffinement de ces méthodes dédiées aux séquences, c'est la nature même des données moléculaires qui les rend si précieuses en phylogénie. Un seul gène peut contenir plusieurs milliers de paires de bases, un génome mammifère jusqu'à plusieurs milliards (3.2 pour l'homme). Comparé aux données morphologiques, cette abondance de caractères est sans commune mesure. Parce qu'elles ne nécessitent pas d'intervention humaine directe pour être inventoriées, les données moléculaires sont de surcroît souvent considérées comme plus objectives. Pour toutes ces raisons, la phylogénie moléculaire est à présent la discipline reine pour traquer les relations de parenté entre espèces actuelles. On utilise cependant encore les données morphologiques en paléontologie, puisque l'obtention d'ADN sur un fossile plus ancien de quelques dizaines de milliers d'année est impossible à l'heure actuelle.

1.4 Origine et systématique des Placentaires à l'ère moléculaire

Les évolutionnistes n'ont pas attendu l'avènement de la phylogénie moléculaire pour se pencher sur l'arbre des mammifères. Sommes nous plus proches du dauphin ou du rat ? Au delà de la simple curiosité, élucider notre propre généalogie à travers celle des mammifères fournit un cadre évolutif indispensable pour interpréter nos spécificités morphologiques, physiologiques, comportementales ou génomiques. Parce que l'on connaît les mammifères mieux que tout autre groupe, décrypter entièrement leur histoire a pour vocation naturelle de faire office de modèle pour la résolution du reste de l'arbre de la vie.

Ainsi, les morpho-anatomistes et paléontologues n'ont eu de cesse d'essayer d'établir au mieux la systématique de ce groupe. Ces efforts ont trouvé leur aboutissement dans un arbre qui faisait encore office de référence il y a une dizaine d'années (Figure 1.6a) [Novacek, 1992, Shoshani and McKenna, 1998, Liu et al., 2001]. Sa caractéristique majeure était de placer les Xenarthres (tatou, grand fourmilier, paresseux et autres animaux exclusivement d'Amérique du Sud) à la base des placentaires, regroupant la majorité des ordres restants en 3 grands

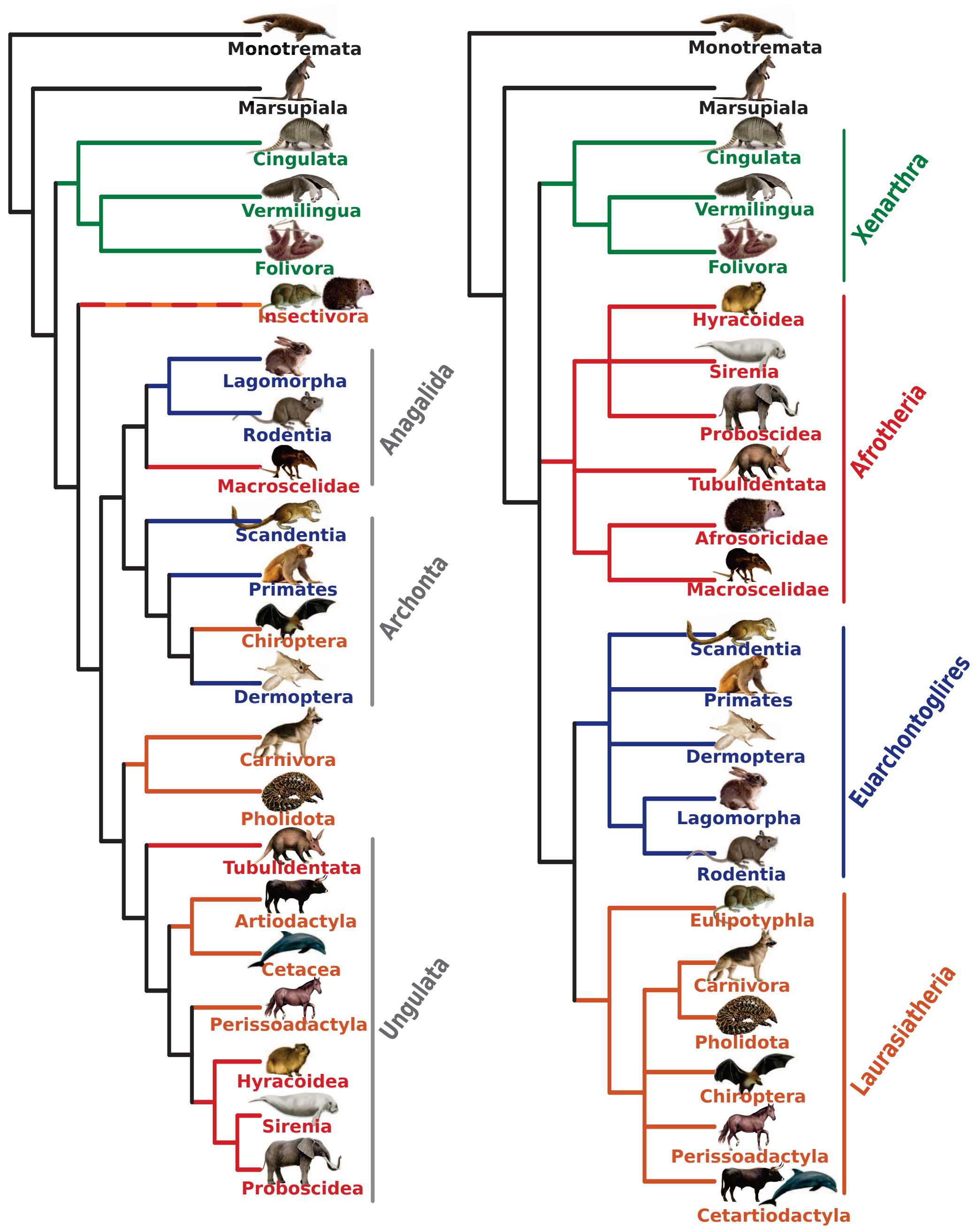


FIGURE 1.6 – Comparaison entre a) un arbre de référence basé sur des caractères morphologiques vs b) un arbre de référence basé sur des caractères moléculaires. Les couleurs sont attribuées en fonction de l'appartenance à l'un des 4 super-ordres modernes de Placentaires. Adapté d'après [Springer et al., 2004].

groupes : les ongulés (*Ungulata*, regroupant les animaux qui marchent sur le bout de leur doigts et tous les mammifères aquatiques), les archontes (*Archonta*, regroupant une majorité d'animaux arboricoles ou volants, dont les primates et les chauves-souris) et les anagalides (*Anagalida*, regroupant de petits animaux terrestres tels que les rongeurs, lapins ou rats à trompe). Enfin, relégués à part, les Insectivores sont aujourd'hui considérés comme un groupe fourre-tout, rassemblant tous types de musaraignes, hérissons ou taupes aux origines pourtant bien différentes.

Cette classification est complètement remise en cause au début des années 2000 [Madsen et al., 2001, Murphy et al., 2001]. Faisant aujourd'hui office de références largement acceptées, les classifications moléculaires dressent en effet un tout autre tableau. En lieu et place des ongulés, archontes et angalides, la classification actuelle considère 4 super-ordres : Xenarthres, Afrothériens, Euarchontoglires et Laurasiathériens (Figure 1.6b).

Parmis ces super-ordres, le groupe des Xenarthres est le seul à demeurer inchangé : il reste tel qu'il a été établi par la classification morphologique. En revanche, le groupe des Afrothériens est le plus inattendu. Comme son nom l'indique, il rassemble divers mammifères d'origine africaine, jusque là éparpillés dans la classification morphologique : éléphants (Proboscidea), lamantins (Sirenia), damans (Hyracoidea), tenrecs (Afrosoricida) ou rats à trompe (Macroscelidea). Autre bouleversement majeur, les chauve-souris (Chiroptera) ne sont plus apparentées aux écureuil volants (Dermoptera). Ces derniers restent donc seuls proches parents des Primates. Le groupe des Euarchontoglires réunit ainsi les anciens archontes (à l'exception des chauve-souris), et les glires (rongeurs et lagomorphes). Les Laurasiathériens rassemblent eux des animaux très variés, tels que les musaraignes (Eulipotyphla), dauphins (Cetartiodactyla), chevaux (Perrissodactyla), chiens (Carnivora) ou chauve-souris (Chiroptera). Laurasiathériens et Euarchontoglires forment de plus tous deux un groupe majeur supplémentaire. En référence à leur origine supposée nordique, on a appelé l'union de ces deux super-ordres Boreoeutheria, les euthériens (=placentaires) originaires du Nord.

Comment expliquer un tel écart entre données morphologiques et moléculaires ? Au vu de cette nouvelle phylogénie, il semble que des espèces éloignées aient parallèlement développé des morphologies très proches. La figure 1.7 illustre ce concept de convergence évolutive, où sont présentés de véritables sosies morphologiques pourtant très éloignés moléculairement. Cette notion de sosie est par ailleurs très utile pour comprendre l'échec des morphologistes. Tout le monde aura en effet constaté qu'en dépit d'une ressemblance frappante, certaines personnes ne partagent aucun lien de parenté. Or, s'il est possible de reconnaître des personnes d'une même famille via des similitudes dans leurs visages, le raisonnement atteint vite ses limites si l'on incorpore de parfaits sosies. Encore une fois, l'unique moyen de reconstituer les bonnes familles sera d'avoir recours à l'ADN. Ce raisonnement issu de la généalogie classique s'applique à la généalogie des espèces. Ainsi, et cela en dépit de plus d'un siècle d'efforts, les morphologistes n'ont jamais pu résoudre les liens de parentés entre les ordres actuels.

Mais les données moléculaires ne se sont pas contentées de bouleverser la systématique. Alimentant encore aujourd'hui la polémique avec les morphologistes, elles fournissent des dates de divergence en désaccord majeur avec le registre fossile. Comme évoqué lors de la partie 1.2, l'ancien dogme formulé par les paléontologues faisait coïncider la racine des placentaires avec la crise K-Pg, il y a de cela 65 Ma. D'après les estimations moléculaires, cette origine serait en réalité bien plus ancienne, et daterait d'il y a plus de 100 Ma [Springer et al., 2003]. Ce résultat impliquerait donc que les placentaires ne doivent pas leur origine au cataclysme qui a mis fin au règne des dinosaures non-aviens. Ainsi, les super-ordres actuels seraient apparus bien avant cette crise majeure de la biodiversité. Certaines études vont même plus loin et suggèrent que cet événement considéré si important n'aurait en fait pas eu le moindre impact sur la diversification des ordres [Bininda-Emonds et al., 2007]. Ce point de vue est toutefois nuancé par une étude plus récente [Meredith et al., 2011]. Quoiqu'il en soit, ces deux études (et d'autres) s'accordent sur une origine des placentaires vieille de plus de 100

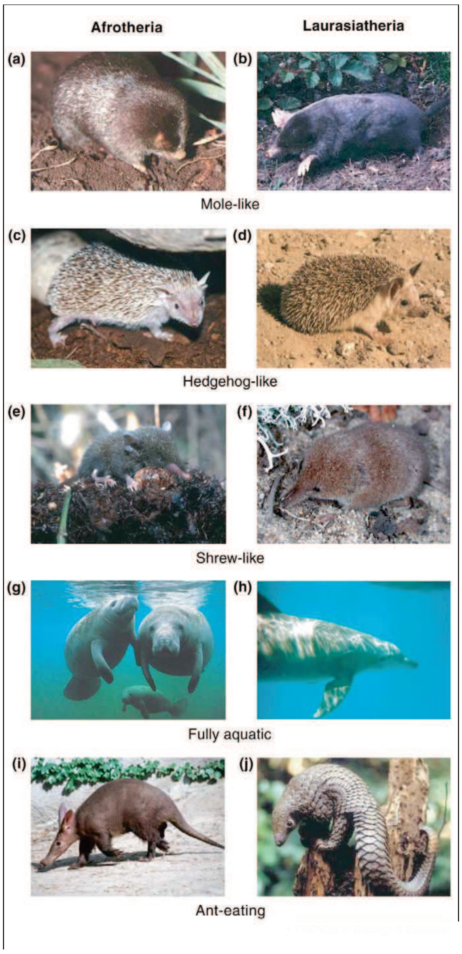


FIGURE 1.7 – Convergence morphologique chez Laurasiatheria et Afrotheria. a) Taupe dorée (Chrysochlorinae) ; b) Taupe commune (Talpinae) ; c) Tenrec hérisson (Tenrecinae) ; d) Hérisson commun (Erinaceinae) ; e) Tenrec musaraigne (Oryzorictinae) ; f) Musaraigne commune (Soricinae) ; g) Lamantin (Trichechidae) ; h) Dauphin (Delphininae) ; i) Oryctérope (Orycteropodidae) ; Pangolin (Maninae). D'après [Springer et al., 2004].

Ma [Hedges et al., 2006]. Le mythe d'un cimetière de dinosaures en guise de berceau des placentaires modernes est lézardé jusque dans ses fondements. Mais si la crise K-Pg n'est pas à l'origine des 4 super-ordres actuels, quel phénomène peut expliquer leur diversification ?

C'est ce nouveau regard sur la phylogénie des placentaires qui a révélé au grand jour l'influence majeure de la biogéographie au travers de la tectonique des plaques. Comme évoqué plus haut, chacun des 4 super-ordres moléculaires correspond à des espèces d'origine géographique commune. Les Boréoeuthériens (Euarchontoglires + Laurasiathériens) sont ainsi originaires de l'ancien super-continent nordique, la Laurasie, tandis qu'Afrothériens et Xénarthres proviennent respectivement d'Afrique et d'Amérique du Sud. C'est la dérive des continents qui expliquerait ainsi la phylogénie des espèces actuelles.

Cependant, le scénario exact n'est pas encore résolu. Comme le montre la figure 1.8b, il n'y a actuellement pas de consensus clair pour la racine des placentaires. Trois hypothèses sont envisageables : Epitheria (Boreoeutheria + Afrotheria), Exafrotheria (Boreoeutheria + Xenarthra) et Atlantogenata (Afrotheria + Xenarthra) (Figure 1.8a). Epitheria est l'hypothèse classique des morphologistes (Xénarthres à la racine des placentaires), mais est la moins soutenue par les données moléculaires, bien que supportée par certaines analyses récentes [Shoshani and McKenna, 1998, Waddell et al., 2001, Kriegs et al., 2006, Churakov et al., 2009]. Parmi les deux hypothèses restantes, Atlantogenata propose le scénario bio-géographique le plus élégant : les espèces originaires d'Afrique et d'Amérique du Sud sont rassemblées, tout comme l'étaient ces deux continents par le passé dans le super-continent du Gondwana (Figure 1.8b). Cette hypothèse est de fait très populaire, et est la plus soutenue par les données nucléaires les plus récentes [Huchon et al., 2002, Murphy et al., 2007, Wildman et al., 2007, Hallström et al., 2007, Kjer and Honeycutt, 2007, Hallström and Janke, 2008, Prasad and Allard, 2008, Meredith et al., 2011, Song et al., 2012]. Malgré une adéquation moins stricte avec la dérive des continents, l'hypothèse Exafrotheria (Afrotheria à la racine) est suggérée par un nombre non-négligeable d'études [Murphy et al., 2001, Delsuc et al., 2002, Amrine-Madsen et al., 2003, Waddell and Shelley, 2003, Nikolaev et al., 2007, McCormack et al., 2012].

Nishihara et al [Nishihara et al., 2009] montrent quant à eux que la séparation de la Pangée (l'unique continent primordial) en Gondwana (Afrique + Amérique du Sud) et Laurasie (Eurasie + Amérique du Nord) ne correspond pas forcément à un brutal isolement géographique. Plusieurs ponts ont pu relier Laurasie, Afrique et Amérique du Sud, rendant la résolution de la racine des placentaires quasi-impossible. Résoudre définitivement la racine des placentaires reste donc l'un des derniers et plus importants défis de la phylogénie moléculaire.

Comme nous avons pu le voir, l'outil moléculaire a donc largement redéfini les relations profondes entre ordres placentaires. Notons cependant que malgré ces bouleversements inter-ordinaux, les ordres eux même, bien qu'initialement établis morphologiquement, restent pertinents. La seule exception notable est celle des Insectivores, désormais éclatés en deux ordres bien distincts : les Afrosoricidae (situés au sein des Afrothériens) et les Eulipotyphles (situés au sein des Laurasiathériens). D'après Madsen et al. [Madsen et al., 2001], cette surprise pourrait suggérer que le morphe "insectivores" est la base à partir de laquelle tous les autres groupes ont divergé. Cette hypothèse expliquerait pourquoi les morphologistes n'ont jamais pu différencier les "insectivores" Laurasiathériens et Afrothériens. Un ancêtre à l'aspect de musaraigne est de plus suggéré depuis longtemps par les paléontologues [Novacek, 1986]. Cette idée d'un ancêtre de petite taille, nocturne, insectivore et fouisseur a en effet souvent été avancée pour expliquer la survie des euthériens au cataclysme brutal qui aurait marqué la fin du Crétacé [Robertson et al., 2004, Smith et al., 2010]. Bien que l'ADN conteste toute relation entre origine des placentaires et crise K-Pg, le portrait-robot de cet ancêtre commun n'a jamais vraiment été remis en cause. Toujours largement relayé par la plupart des ouvrages de vulgarisation scientifique [Dawkins, 2004, Feldhamer et al., 2007], il s'agit d'un des rares points où fossiles et molécules ne sont pas en désaccord frontal. Et pour cause, notons que ni l'un

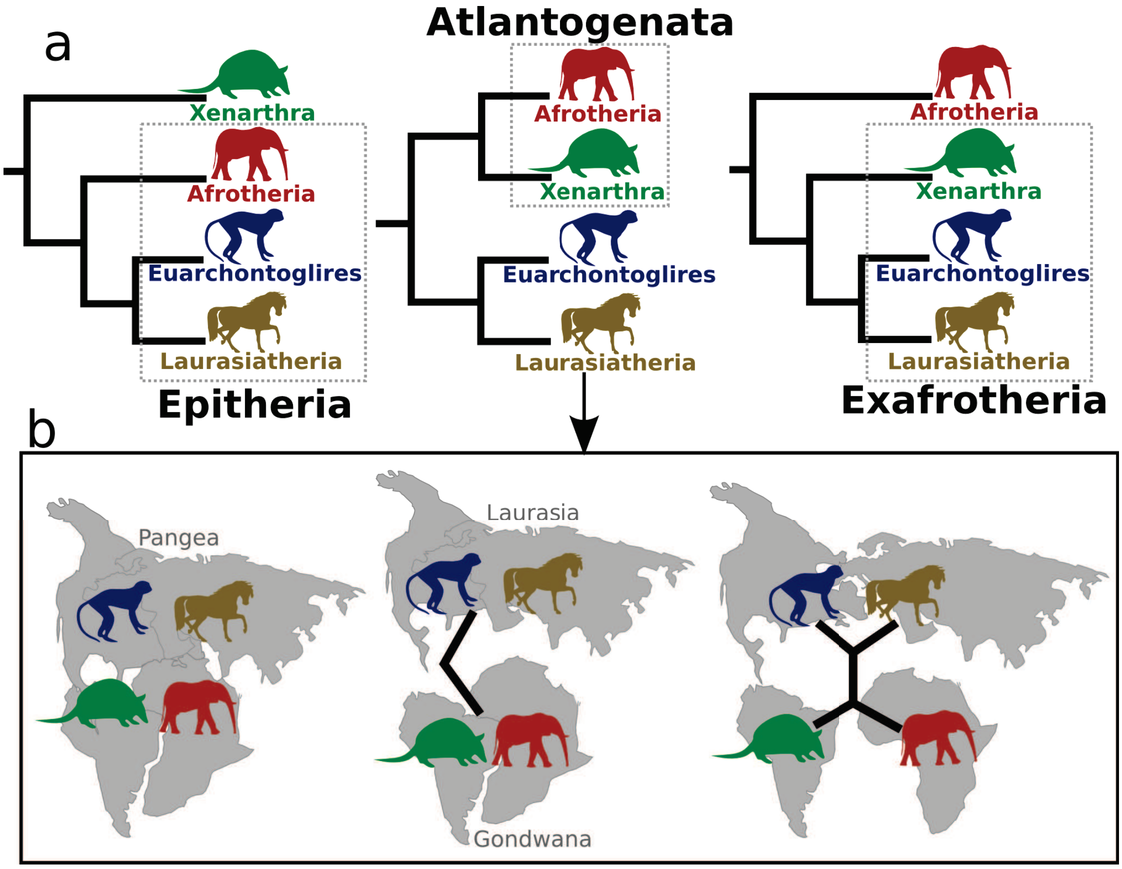


FIGURE 1.8 – 3 Hypothèses pour la racine des placentaires (a). L'hypothèse Atlantogenata est celle qui s'accorde le mieux avec un scénario de vicariance strict selon la dérive des continents (b).

ni l'autre n'a de réels arguments. Comme vu en figure 1.7, la convergence morphologique est répandue chez les mammifères, et rien ne prouve que cela ne soit pas aussi le cas du morphe musaraigne. Rappelons également que l'on n'a jamais découvert le moindre fossile placentaire datant du Crétacé, période présumée de leur origine. Cette déficience est probablement attribuable à l'effet Signor-Lipps [Signor and Lipps, 1982], principe qui stipule que le registre fossile est par nature incomplet, nécessairement orphelin du premier organisme du groupe que l'on étudie. Il est en effet raisonnable de penser que plus on remonte vers les origines d'un groupe, moins la diversité ancestrale découverte n'a de chances d'y appartenir. Les fossiles mammifères vieux de plus de 65 Ma n'ont ainsi jamais été attribués au groupe des placentaires. On ne sait donc rien de la forme et du mode de vie de nos ancêtre du Crétacé, ces placentaires qui auraient côtoyé les dinosaures pendant plus de 30 Ma. Caractériser ces fantômes du registre fossile et déterminer comment ils ont formé les 4 super-ordres actuels est donc l'une des dernières pièces du puzzle de nos origines.

Comment s'atteler à cette tâche ? Tout au long de cette première partie, nous avons évoqué l'ADN en tant qu'outil du systématicien pour traquer l'histoire des espèces. L'étude des séquences ADN peut toutefois être abordée avec un point de vue différent : celui de l'évolutionniste moléculaire. Celui-ci ne place pas nécessairement les espèces et leur généalogie au centre de son étude, mais essaye davantage de comprendre l'évolution et l'architecture de leurs génomes. A partir d'une même donnée moléculaire, on peut donc s'intéresser à deux histoires évolutives différentes : celle des espèces ou celle de leurs génomes, intimement liées, mais chacune régie par ses propres règles. Une partie de l'originalité de cette thèse réside dans cette dualité. C'est à partir d'une problématique d'évolution des génomes, celle de l'évolution de leurs paysages nucléotidiques, que seront révélés des indices sur l'origine et la diversification des placentaires. Dans la deuxième partie de cette introduction, nous aborderons donc le génome en tant qu'objet biologique à part entière, point de vue indispensable pour répondre aux questions évoquées lors de cette première partie.

Chapitre 2

Origine des paysages nucléotidiques mammaliens

2.1 Le génome : un objet biologique à part entière

Contrairement à ce que son nom laisse entendre, un génome n'est pas qu'un catalogue de gènes mis bout à bout. C'est pourtant l'idée qu'on s'en faisait jusque dans les années 70. Plan de montage ou vulgaire notice de fonctionnement des êtres vivants, l'évolution d'un génome était alors fondue avec celle de l'espèce. Comme nous allons le voir, l'évolution moléculaire est pourtant régie par des règles qui lui sont propres.

Centrée sur les individus, l'évolution darwinienne a longtemps supposé un rôle quasi-hégémonique de la sélection naturelle. L'essentiel des caractères d'un organisme, séquences génomiques y compris, devaient ainsi résulter d'une nécessaire adaptation des espèces à leur milieu. Ce dogme adaptationiste fut ébranlé une première fois par la théorie neutraliste de l'évolution moléculaire [Kimura, 1983]. Dans ce nouveau paradigme, Motoo Kimura soutient que l'évolution moléculaire s'effectue sans un rôle prépondérant de la sélection naturelle. Les attendus théoriques de cette dernière sont en effet très loin d'expliquer à eux seuls les taux de substitutions et niveaux de polymorphismes observés chez certains mammifères [Kimura, 1968]. Les mutations qui procurent un avantage aux individus qui les reçoivent sont en réalité rarissimes. La quasi-totalité des mutations qui parviennent à envahir une population (on parle de fixation) sont donc neutres, sans bénéfice direct pour les individus mutants (Figure 2.1). Ce constat sonne le glas d'un adaptationisme forcené, considérant qu'évolution du génome n'est synonyme que d'optimisation des organismes. Si elle l'amoindri, la théorie neutraliste ne néglige pas pour autant le rôle de la sélection naturelle. Autrefois considérée comme moteur de l'évolution, elle est vue aujourd'hui comme son indispensable frein moléculaire. Limitant l'évolution débridée de nos gènes soumis au joug de mutations inopinées, elle assure l'intégrité fonctionnelle des génomes et empêche la fixation d'innombrables mutations délétères [King and Jukes, 1969]. Cette notion clé de sélection purifiante explique pourquoi les variations dans nos génomes se détectent essentiellement sur des sites sans grand intérêt fonctionnel (telles que les 3$^{\text{ème}}$ position de codons[1] ou certaines régions non-codantes), les autres étant contraints à conserver leur état optimisé.

Ce dernier point est d'autant plus important qu'une écrasante majorité de notre génome est concerné. On estime en effet que le génome humain ne contient qu'1.5% de régions codantes [Lander et al., 2001]. Ce faible pourcentage est la part exprimée des gènes, ces fragments d'ADN qui codent la synthèse de toutes les protéines nécessaires à notre bon fonctionnement. Comment expliquer que nos génomes contiennent 98.5% d'ADN non-codant ? Est il nécessairement utile ?

Ces questions trouvent une partie de leur réponse dans la théorie dite du "gène égoïste" [Dawkins, 1976]. Qui est la cible de la sélection ? S'il semblait jusque là naturel de considérer qu'il s'agissait de l'individu, ce nouveau point de vue remet en cause son statut de sujet central de l'évolution. En déclarant que "les unités de sélection ne sont en aucun cas les individus biologiques, mais leurs gènes et leurs

[1] La plupart des mutations en 3$^{\text{ème}}$ position de codons sont dites silencieuses, l'acide aminé pour lequel code le codon restant identique.

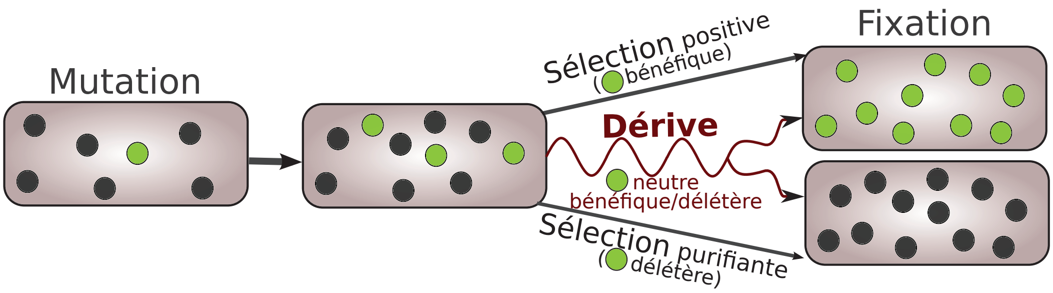


Figure 2.1 – Fixation d'un caractère.
Au niveau des individus : par mutation, un variant avantageux apparait dans une population (ex : un zèbre plus rapide que les autres). Plus apte à la survie, il a plus de chance d'avoir une descendance, descendance qui pourra à son tour propager le caractère avantageux. Par sélection positive, celui-ci deviendra de plus en plus fréquent jusqu'à envahir la population (fixation).
Au niveau moléculaire : une nouvelle mutation étant presque toujours neutre ou délétère, la sélection positive est un processus rare. Les mutations sont en réalité plus souvent éliminées par sélection purifiante, ou sont fixées par dérive génétique. Cette force majeure de l'évolution moléculaire est souvent qualifiée de stochastique, et réside dans la possibilité qu'a une mutation de se retrouver par chance dans le bon gamète (ovule ou spermatozoïde), et cela de générations en générations. Envahir une population de la sorte est un processus considéré comme lent, mais est d'autant plus facile que la taille de population est faible (probabilité de fixation 1/N avec N le nombre d'individus dans la population, 1/2N dans le cas d'individus diploïdes). En faible taille de population, la sélection (positive ou purifiante) peut être outrepassée par la dérive, et une mutation faiblement délétère peut parvenir à se maintenir ou se fixer dans une population

chromosomes", G. Ostergren [Östergren, 1945] fut le premier à jeter un pavé dans la mare. Plusieurs biologistes lui emboîtèrent le pas : J. Maynard-Smith, G. C. Williams [Williams, 1966], W. D. Hamilton ou R. Dawkins [Dawkins, 1976]. Ces derniers firent justement remarquer que seule l'information génétique perdure à travers les générations : puisque c'est elle qui se reproduit le plus fidèlement, c'est avant tout elle qu'il faut considérer. L'entité qui la transmet, l'individu, n'est ainsi envisagé que comme un véhicule éphémère, inventé et sans cesse reconstruit par les gènes pour perdurer à travers les générations.

"*Nous sommes des machines à survie - des robots programmés à l'aveugle pour préserver les molécules égoïstes connues sous le nom de gènes*" [Dawkins, 1976]. Un brin provocatrice et imagée, cette vision fut sévèrement critiquée, notamment par le naturaliste E. Mayr et le paléontologue S. J. Gould. Aujourd'hui assoupi, ce débat a donné naissance à l'une des notions les plus importantes de la biologie évolutive moderne : celle des niveaux de sélection. Il est en effet aujourd'hui entendu que la sélection n'est pas exclusive à l'individu, mais agit bel et bien sur d'autres niveaux d'intégration du vivant.

Toutes les entités qui présentent des variations et peuvent se reproduire en tant que telle sont potentiellement concernées : gènes, chromosomes, organismes, groupes, espèces... un nombre de candidats à la sélection plus grand qu'envisagé au départ. Si leurs intérêts respectifs sont à l'évidence imbriqués les uns aux autres, sont-ils pour autant toujours compatibles ? La réponse est non. Ainsi, la compétition entre individus se fait parfois au détriment de l'espèce : en témoigne la queue disproportionnée des paons mâles, utile individuellement pour s'attirer les faveurs d'une femelle, mais si lourde et si visible qu'elle n'est clairement pas favorable à la survie de l'espèce. Il en va de même pour tout type de comportement tricheur, avantageux pour son auteur bien que néfaste au groupe. De tels conflits d'intérêts sont en réalité monnaie courante, et ces derniers ne s'arrêtent pas à la frontière de l'individu. Autrefois indivisible, on distingue en effet des intérêts sélectifs variés au sein d'un même organisme : ceux de la cellule (figure

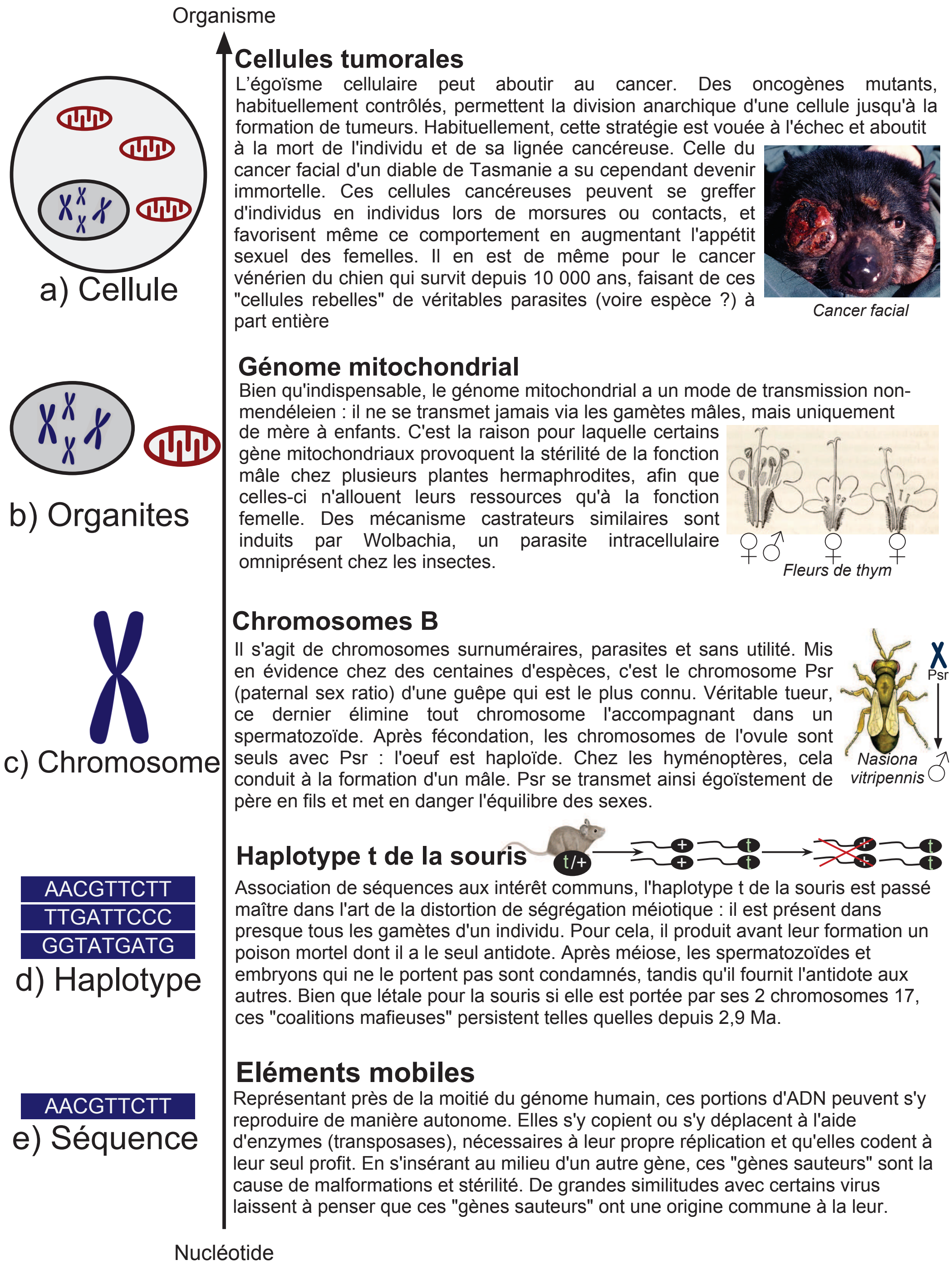
2.2a), de l'organite[2] (figure 2.2b), du chromosome (figure 2.2c), de l'haplotype[3] (figure 2.2d) ou de la séquence (figure 2.2e).

On parle alors de conflits intra-génomiques, conflits qui impliquent la plupart du temps des éléments génétiques dits "égoïstes" [Werren et al., 1988, Werren, 2011, Burt and Trivers, 2006]. Parasites de nos génomes, ces séquences peuvent augmenter leur nombre ou leur transmission, et cela même si elles n'apportent rien ou sont néfastes à l'ensemble. "Parasite ultime" des génomes eucaryotes, il a été proposé qu'une partie de notre ADN non-codant soit d'origine égoïste [Orgel and Crick, 1980, Doolittle and Sapienza, 1980]. Sélectionnées avant tout pour elles même, ces séquences perdurent par leur seule capacité à se multiplier et se transmettre d'un organisme à l'autre. Pour ce faire, elles disposent d'un arsenal des plus larges : parmi les mécanismes les plus connus, citons la transposition des éléments mobiles (figure 2.2e) et la distorsion de ségrégation méiotique[4] (exemple en figure 2.2d). Face à ces menaces bien réelles, le génome contre-attaque, et a su mettre en place de nombreux moyens de défense contre la multiplication anarchique de ces parasites internes [Johnson, 2007]. Une fois contrôlés ou inactivés, ces derniers peuvent même finir par s'avérer utiles [Doolittle and Sapienza, 1980, Hurst and Werren, 2001, Burt and Trivers, 2006, Werren, 2011], que cela soit au travers des divers exemples de "domestication" d'éléments transposables [Feschotte, 2008, Sinzelle et al., 2009] ou d'un quelconque rôle qu'on leur attribue dans l'adaptabilité des génomes [Hurst and Werren, 2001, Biémont and Vieira, 2006]. Mais comme pour les mutations, l'apparition d'éléments égoïstes est presque toujours momentanément

[2] Structures spécialisées des cellules. Parmi elles, la mitochondrie (centrale énergétique des eucaryotes) et le chloroplaste (responsable de la photosynthèse végétale) ont leur propre ADN et peuvent entrer en conflit avec le génome nucléaire.

[3] Un groupe d'allèles, versions de gènes situés sur un même chromosome et habituellement transmis ensemble

[4] La méiose est une suite de divisions cellulaires aboutissant à la formation des gamètes (chez les mammifères : ovule et spermatozoïdes). Habituellement, les deux versions d'un même gène (les allèles) sont répartis équitablement entre ces derniers. Les distorteurs de ségrégation méiotique "trichent", et parviennent à se retrouver dans plus de la moitié des gamètes. La loterie génétique ainsi truquée, ils ont plus de chance d'être transmis à la descendance.



Cellules tumorales

L'égoïsme cellulaire peut aboutir au cancer. Des oncogènes mutants, habituellement contrôlés, permettent la division anarchique d'une cellule jusqu'à la formation de tumeurs. Habituellement, cette stratégie est vouée à l'échec et aboutit à la mort de l'individu et de sa lignée cancéreuse. Celle du cancer facial d'un diable de Tasmanie a su cependant devenir immortelle. Ces cellules cancéreuses peuvent se greffer d'individus en individus lors de morsures ou contacts, et favorisent même ce comportement en augmentant l'appétit sexuel des femelles. Il en est de même pour le cancer vénérien du chien qui survit depuis 10 000 ans, faisant de ces "cellules rebelles" de véritables parasites (voire espèce ?) à part entière

Génome mitochondrial

Bien qu'indispensable, le génome mitochondrial a un mode de transmission non-mendéleien : il ne se transmet jamais via les gamètes mâles, mais uniquement de mère à enfants. C'est la raison pour laquelle certains gène mitochondriaux provoquent la stérilité de la fonction mâle chez plusieurs plantes hermaphrodites, afin que celles-ci n'allouent leurs ressources qu'à la fonction femelle. Des mécanisme castrateurs similaires sont induits par Wolbachia, un parasite intracellulaire omniprésent chez les insectes.

Chromosomes B

Il s'agit de chromosomes surnuméraires, parasites et sans utilité. Mis en évidence chez des centaines d'espèces, c'est le chromosome Psr (paternal sex ratio) d'une guêpe qui est le plus connu. Véritable tueur, ce dernier élimine tout chromosome l'accompagnant dans un spermatozoïde. Après fécondation, les chromosomes de l'ovule sont seuls avec Psr : l'oeuf est haploïde. Chez les hyménoptères, cela conduit à la formation d'un mâle. Psr se transmet ainsi égoïstement de père en fils et met en danger l'équilibre des sexes.

Haplotype t de la souris

Association de séquences aux intérêt communs, l'haplotype t de la souris est passé maître dans l'art de la distortion de ségrégation méiotique : il est présent dans presque tous les gamètes d'un individu. Pour cela, il produit avant leur formation un poison mortel dont il a le seul antidote. Après méiose, les spermatozoïdes et embryons qui ne le portent pas sont condamnés, tandis qu'il fournit l'antidote aux autres. Bien que létale pour la souris si elle est portée par ses 2 chromosomes 17, ces "coalitions mafieuses" persistent telles quelles depuis 2,9 Ma.

Eléments mobiles

Représentant près de la moitié du génome humain, ces portions d'ADN peuvent s'y reproduire de manière autonome. Elles s'y copient ou s'y déplacent à l'aide d'enzymes (transposases), nécessaires à leur propre réplication et qu'elles codent à leur seul profit. En s'insérant au milieu d'un autre gène, ces "gènes sauteurs" sont la cause de malformations et stérilité. De grandes similitudes avec certains virus laissent à penser que ces "gènes sauteurs" ont une origine commune à la leur.

Figure 2.2 – Niveaux de sélection et conflits intra-génomiques. Werren et al. 2011, Burt et Trivers 2006 pour revue. Au delà du niveau de l'organisme, s'ajoutent ceux du super-organisme (insectes sociaux), de la population, de l'espèce ou de l'écosystème.

neutre ou délétère pour l'individu. Penser que leur naissance et leur maintien initial résulte avant tout d'une sélection pour un potentiel bienfait est souvent révélateur d'un excès d'adaptationisme, vision dont l'histoire de l'évolution moléculaire a appris à se méfier.

Si l'évolution darwinienne a fait tomber l'espèce humaine de son piédestal, l'évolution moléculaire fait subir un sort comparable à l'individu : jusque là sujet de l'évolution, il cède sa place à l'information génétique. De ce basculement de perspective, il nous reste du génome l'image d'un microcosme à part entière, peuplé de parasites omniprésents, domestiqués, réduits au silence ou toujours actifs face à des gènes qui œuvrent pour la reproduction de l'individu. Au fur et à mesure, l'accumulation de ces divers champs de bataille a façonné des paysages génomiques variés. Parmi eux, cette thèse s'intéresse à ceux que l'on peut trouver chez les mammifères. Enigmatique, la structuration de leurs paysages nucléotidiques suggérerait presque que nos génomes sont le théâtre d'une lutte bien singulière : celle qui oppose deux à deux les lettres de l'alphabet du vivant, AT et GC. Comme nous allons le voir, cette guerre fratricide a laissé des traces durables dans nos génomes. Inutiles pour l'organisme, ces traces s'avéront en revanche bel et bien utiles pour nous permettre d'un peu mieux comprendre l'origine et l'évolution des mammifères placentaires.

2.2 Les isochores : une énigmatique structuration de nos génomes

Indispensables au code génétique, tout pourrait porter à croire que les bases A,T,G et C sont également distribuées au sein du vivant. On sait pourtant depuis longtemps que la composition nucléotidique n'est presque jamais aussi homogène qu'escompté [SUEOKA, 1962]. La composition en GC peut en effet s'étendre de 15 à 75%, tous groupes vivants confondus [Lynch and Walsh, 2007]. Pire, au sein d'un même génome, on observe parfois des variations à première vue inexplicable : c'est le cas chez les mammifères. Paysages nucléotidiques alternant les pics de

GC aux vallées d'AT (Figure 2.3), les génomes mammaliens n'ont eu de cesse d'intriguer les chercheurs depuis plus de 30 ans.

Cette forte hétérogénéité en composition de base a d'abord été mise en évidence par Bernardi et son équipe [Bernardi, 1985], bien avant l'ère du séquençage. Par des techniques d'ultracentrifugation en gradient de densité, il est en effet possible de séparer des fragments d'ADN en fonction de leur composition en bases G+C. Menés sur quelques génomes de mammifère et d'oiseaux, ces premiers travaux ont permis d'identifier l'existence de grands fragments chromosomiques (>100kb) aux taux de GC étonnamment contrastés (de 35 à 55%). Chacune caractérisée par son propre taux de GC, ces régions furent baptisées isochores[5] : ainsi, qualifia-t-on les génomes de vertébrés à sang chaud de "mosaïques d'isochores". Ce modèle fut cependant critiqué avec l'avènement des séquençages complets de génomes. En règle générale, il n'y a pas de frontière nette aux isochores, mais plutôt une variation continue de la composition en bases [Lander et al., 2001]. L'analyse des séquences génomiques confirme cependant bel et bien l'existence de variations significatives du contenu en GC le long des chromosomes (Figure 2.3). En accord avec Eyre-Walker et Hurst [Eyre-Walker and Hurst, 2001], on considère ainsi que le terme d'isochore, bien qu'imparfait, est utile pour décrire l'organisation particulière de ces génomes. Affectant aussi bien les régions codantes que non-codantes, ces variations du taux de GC ne peuvent qu'être difficilement le fruit du hasard.

Si leur origine suscite à elle seule la curiosité, leur étude revêt d'autant plus d'intérêt lorsqu'on les sait associées à d'autres caractéristiques génomiques. En effet, les isochores les plus riches en GC sont aussi les plus denses en gènes. Dotés d'introns plus courts [Duret et al., 1995], ces derniers sont répliqués plus tôt [Costantini and Bernardi, 2008] et présentent des taux de méthylation [Kazanskaya et al., 1997], d'expression [Konu and Li, 2002] et de recombinaison [Fullerton, 2001, Duret and Arndt, 2008] plus élevés. La structuration en isochore est par ailleurs associée à la distribution en éléments transposables

[5] Du grec isos (égal) et choros (région)

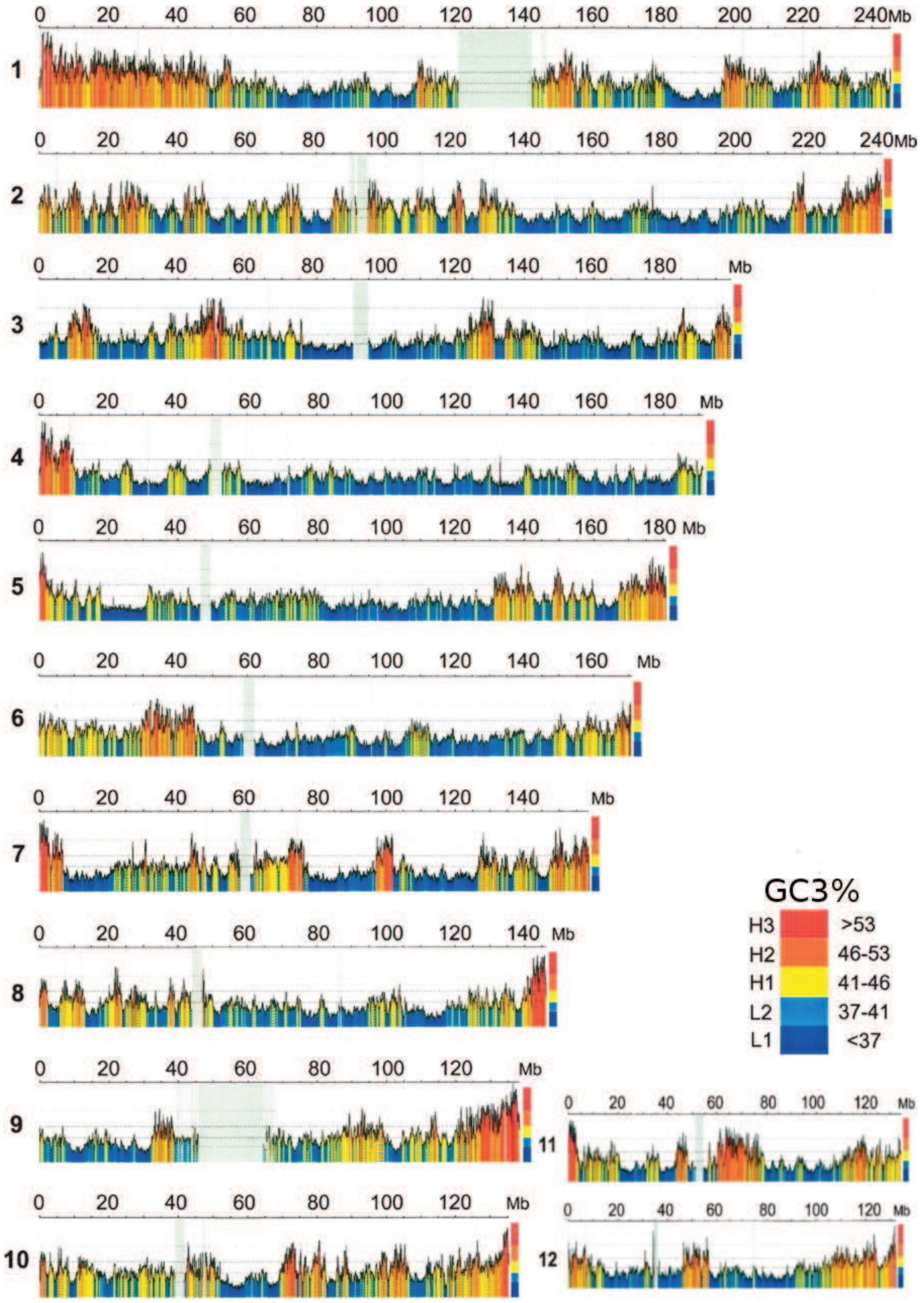


FIGURE 2.3 – Composition nucléotidique des chromosomes humains 1 à 12. Les couleurs sont attribuées selon le taux de GC de fenêtres de 100 kb, du bleu marine (GC-pauvre) au rouge (GC-riche). D'après [Costantini et al., 2006]

.

[Smit, 1999]. Ainsi, cette structuration semble clairement refléter un aspect fondamental de l'évolution de nos génomes.

Pourquoi seraient-ils le théâtre d'une lutte inégale entre AT et GC ? Qu'est ce qui explique la prévalence d'un couple sur l'autre selon la région du chromosome ? Et surtout, est-ce que cette structure confère un avantage sélectif aux espèces qui la possèdent, ou n'est-elle que le reflet d'un processus évolutif non-adaptatif ? Plusieurs hypothèses ont tenté de tirer ces questions au clair.

2.3 Origine des isochores : ni adaptation, ni mutation

2.3.1 Hypothèse sélectioniste : Adaptation à l'endothermie

L'hypothèse selon laquelle la structuration des génomes en isochore serait expliquée par un avantage adaptatif est la première à voir le jour. Bernardi n'observe en effet la présence d'isochores que chez les mammifères et les oiseaux, les seuls Vertébrés dont la température est produite et régulée de manière interne (raison pour laquelle on parle d'animaux endothermes[6]). Bernardi suggère ainsi que l'apparition des isochores serait une adaptation à cette endothermie [Bernardi, 1985, Bernardi, 2000], et appuie son raisonnement par leur absence chez des vertébrés à sang froid, tels que les lissamphibiens[7] ou les téléostéens[8] [Bernardi, 1990].

Pourquoi l'apparition d'isochores GC-riches procurerait-elle un avantage aux endothermes ? L'appariement entre G et C (3 liaisons hydrogène) est chimiquement plus stable que celui entre A et T (2 liaisons hydrogène). D'après Bernardi, les nombreuses liaisons GC des régions les plus riches en gènes offriraient une

[6] Ou plus communément : "à sang chaud"

[7] Plus communément : amphibiens.

[8] Poissons osseux.

meilleure thermostabilité de l'ADN, protection nécessaire après l'apparition de l'endothermie [Bernardi, 1985, Bernardi, 2000].

Bien que séduisante, cette hypothèse adaptationiste est contredite par de nombreux résultats. D'après des études préliminaires n'exploitant pas les génomes complets, tortues, crocodiliens et serpents (des organismes ectothermes) posséderaient également des isochores [Hughes et al., 1999, Kuraku et al., 2006, Hamada et al., 2003]. Il n'y a de plus aucune corrélation entre le taux de GC des Vertébrés et leur température corporelle [Belle et al., 2002, Ream et al., 2003], pas plus qu'il n'y en a avec la température optimale de croissance de divers organismes unicellulaires [Galtier and Lobry, 1997, Hurst and Merchant, 2001]. Outre ces divers résultats, cette hypothèse se heurte à d'importants problèmes théoriques : les mutations surviennent généralement une par une. Pour que chacune d'elle puisse se fixer, il faudrait imaginer qu'un seul et unique changement de base vers G ou C puisse aboutir à un réel avantage sélectif. Celui-ci est rendu d'autant plus improbable par la moindre efficacité de la sélection en cas de faible taille de population, ce qui est tout particulièrement le cas des mammifères et des oiseaux. Bien que plus abondants et parfois soumis à des températures plus élevées (sélection plus efficace associée à une pression de sélection plus forte), même les unicellulaires ne semblent pas optimiser de la sorte leurs taux de GC génomiques [Galtier and Lobry, 1997, Hurst and Merchant, 2001].

Tout ceci rend particulièrement improbable l'hypothèse d'un avantage sélectif aux isochores. Bien que la théorie neutraliste de l'évolution moléculaire [Kimura, 1983] nous ait appris à se méfier de ce genre de réflexe adaptationiste, cette théorie séduit toujours et est encore défendue par son créateur [Bernardi, 2007]. Celui-ci bénéficie par ailleurs du récent doute suggéré par l'absence d'isochores dans le premier génome complet de reptile, l'anole vert (Anolis carolinensis)[9]. Plusieurs théories alternatives ont cependant émergé, parmi lesquelles un corpus d'hypothèses neutralistes regroupées sous le terme d'hypothèses de biais mutationnels.

[9] Ce point sera plus longuement discuté en partie 2.5.3

2.3.2 Hypothèse neutraliste : Biais mutationnels

L'hypothèse sélectionniste semblant peu plausible, plusieurs auteurs ont proposé l'existence de biais mutationels variables à l'origine des isochores. Différents mécanismes ont ainsi été avancés. L'un d'eux suppute qu'il existe plus de nucléotides libres G ou C, et que les régions qui se répliquent en premier seraient ainsi plus promptes à piocher dans ce réservoir et subir des mutations vers GC [Wolfe et al., 1989, Eyre-Walker and Hurst, 2001]. Des effets mutagènes du GC local [Fryxell and Zuckerkandl, 2000] ou de la recombinaison ont également été proposés [Lercher and Hurst, 2002, Hellmann et al., 2003], de même que des variations de l'efficacité de la machinerie cellulaire [Filipski, 1987].

Selon toutes ces théories, on observerait ainsi un excès de mutations vers GC dans les isochores GC-riches. Cette supposition est cependant invalidée par l'ensemble des données de polymorphismes qui suggèrent plutôt un biais mutationnel vers AT [Cargill et al., 1999, Smith and Eyre-Walker, 2001, Webster and Smith, 2004, Capra and Pollard, 2011]. Tout tend à faire penser que les mutations vers G ou C sont plus rares, mais qu'elles bénéficient d'une plus grande probabilité de fixation. Ceci est non seulement démontré par des données de polymorphisme chez la souris [Eyre-Walker, 1999], mais également confirmé à partir de données de génomes entiers [Duret et al., 2002, Spencer et al., 2006]. Il existerait donc un avantage sélectif aux allèles résultant d'une mutation vers GC. Mais si ce n'est au travers d'une meilleure adaptation de l'individu, qu'est ce qui peut en expliquer la cause ? Comme nous l'avons vu en partie 2.1, la sélection n'agit pas qu'au niveau de l'individu, mais sur tous les niveaux d'intégration de l'information génétique. Se pourrait-il que cette règle s'applique jusqu'au nucléotide, unité même du code génétique ? C'est ce que peut laisser entendre l'hypothèse la plus en vogue du moment : celle de la conversion génique biaisée vers GC.

2.4 La conversion génique biaisée vers GC

2.4.1 Une distorsion de ségrégation méiotique généralisée

Chez les mammifères, comme chez la plupart des eucaryotes, il y a deux exemplaires de chaque chromosome : on parle de paires de chromosomes homologues. L'un étant d'origine maternelle, l'autre d'origine paternelle, seul l'un d'entre eux est transmis à la descendance. Une séquence présente sur l'un et l'autre (homozygotie) est donc assurée de passer à la génération suivante. Lorsque ce n'est pas le cas (hétérozygtie), les deux allèles concurrents sont répartis équitablement entre les gamètes. Cette équité est assurée par la méiose, une étape clé de la gamétogénèse au cours de laquelle les chromosomes maternels et paternels sont séparés et redistribués. Les distorteurs de ségrégation méiotique "trichent", et parviennent à se retrouver dans plus de la moitié des gamètes. La loterie génétique ainsi truquée, ils ont plus de chance d'être transmis à la descendance. Il s'agit notamment du cas de l'haplotype t de la souris (figure 2.2d).

La conversion génique peut aboutir à un effet similaire. Il s'agit d'un transfert non-réciproque d'information génétique : une séquence peut ainsi être littéralement "copiée-collée" d'un chromosome à l'autre, augmentant ses chances d'être transmise à la descendance. C'est par le biais d'un tel mécanisme que les nucléotides G et C parviendraient à dominer certaines régions de nos génomes.

La conversion génique est un processus associé à la recombinaison. La recombinaison est indispensable à la méiose, et permet des échanges d'informations génétiques entre chromosomes homologues. Elle débute par la cassure double brin de l'un d'entre eux, et est résolue en crossing over (échange bilatéral entre les deux chromosomes) ou non crossing-over (échange unilatéral) ([Whitby, 2005] pour revue). Qu'il y ait échange bilatéral ou pas, les événements de recombinaison aboutissent systématiquement à la formation d'hétéroduplex, une association de segments d'ADN d'origine différente.

Ces derniers engendrent de fréquents mésappariements (un T en face d'un G par exemple, Figure 2.4). Illégitimes, les mésappariements sont généralement

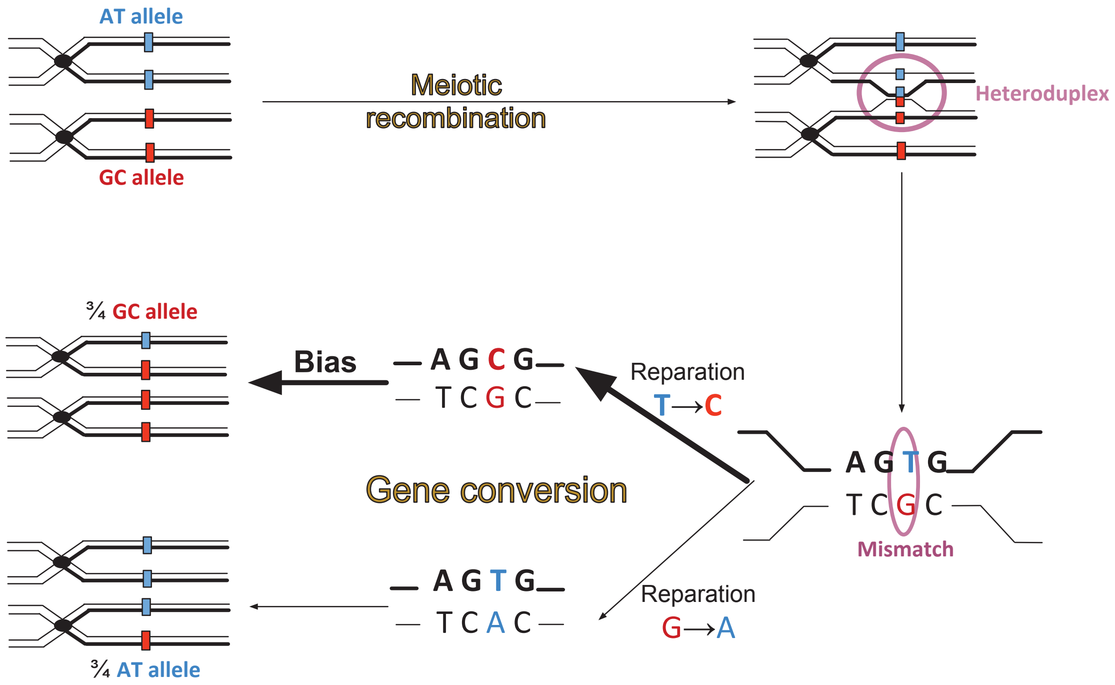


FIGURE 2.4 – Conversion génique biaisée vers GC. D'après [Eyre-Walker, 1993]
.

reconnus par les systèmes de réparation, et sont réparés par la conversion d'une séquence par l'autre : c'est la conversion génique ([Chen et al., 2007] pour revue). Celle-ci est dite biaisée lorsqu'une des conversions possibles est favorisée par les systèmes de réparations. Plusieurs auteurs ont ainsi proposé que les isochores seraient la conséquence d'un biais de conversion favorisant les nucléotides G et C [Brown and Jiricny, 1988, Eyre-Walker, 1993, Galtier et al., 2001, Duret, 2009] (Figure 2.4). Ainsi sur-représentés dans le réservoir de gamètes d'un individu, ils bénéficieraient d'un avantage reproductif suffisant pour se fixer plus facilement dans une population [Eyre-Walker, 1999, Duret et al., 2002, Spencer et al., 2006]. Ce biais de fixation augmentant avec le taux de recombinaison, l'apparition des isochores GC-riches correspondrait ainsi à des régions chromosomiques particulièrement recombinantes.

Associés à ce biais des systèmes de réparation, cette théorie ferait donc des nucléotides G et C les plus petits et plus omniprésents distorteurs de ségrégation méiotique de nos génomes. Mais quelles sont les preuves d'une fraude à si grande échelle de la loterie génétique ? Qu'ils soient expérimentaux ou théoriques, les arguments soutenant l'existence de la conversion génique biaisée sont légions.

2.4.2 Liens entre recombinaison et taux de GC

Plusieurs preuves empiriques suggèrent que la conversion génique biaisée est à l'origine des isochores ([Duret, 2009] pour revue). Parmi elles, le biais de fixation des allèles résultant d'une mutation GC [Eyre-Walker, 1999, Webster and Smith, 2004] qui serait plus fort encore dans les points chaud de recombinaison [Spencer et al., 2006]. En parallèle, de nombreuses corrélations entre taux de GC et recombinaison ont été reportées, que ce soit chez les oiseaux [Webster et al., 2006] ou chez les mammifères [Montoya-Burgos et al., 2003, Meunier and Duret, 2004], y compris lors de conversions géniques ectopiques[10] [Galtier, 2003, Kudla et al., 2004]. Ces relations entre GC et recombinaison n'ex-

[10] Survenant lors d'événements de recombinaisons entre séquences localisées sur des positions génomiques différentes

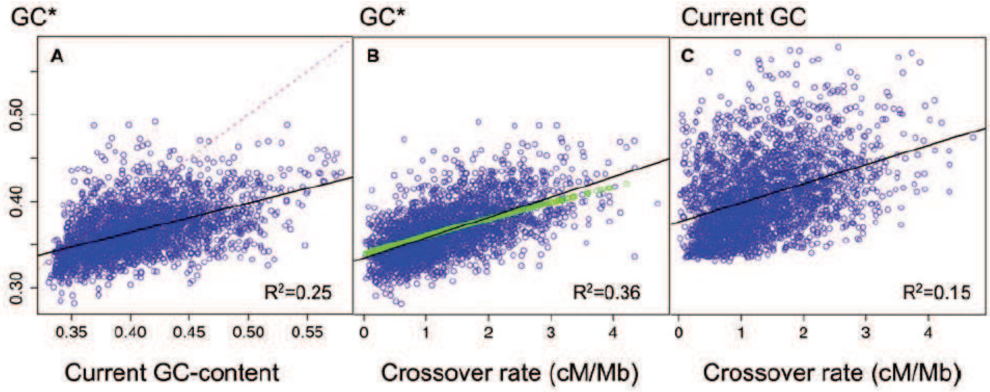


FIGURE 2.5 – Corrélation entre le GC d'équilibre (GC*), le GC actuel et le taux de crossover chez les chromosomes non-sexuels de l'homme. Le GC d'équilibre (GC*) correspond au taux de GC qu'atteindrait une séquence si la dynamique évolutive en cours se poursuivait suffisamment longtemps. Tiré de [Duret and Arndt, 2008].

cluent bien évidemment pas le génome humain [Fullerton, 2001, Li et al., 2008, Duret and Arndt, 2008], où les taux de GC d'équilibre[11] prédisent mieux la recombinaison que les taux de GC actuels (Figure 2.5). En accord avec le modèle de conversion génique biaisée, ceci démontre clairement une influence majeure de la recombinaison sur le taux de GC, et pas seulement un effet recombinant des motifs GC-riches (comme suggéré par certains auteurs [Surtees et al., 2004]).

Le gène de la souris Fxy fournit un autre exemple spectaculaire. On sait que ce gène a subi une translocation récente dans la région pseudo-autosomale (homologue entre chromosomes X et Y), où le taux de recombinaison est extrêmement élevé. Cette translocation a impliqué une forte augmentation du taux de substitution [Perry and Ashworth, 1999] et une très forte augmentation du contenu en GC des sites codants et non-codants [Galtier and Duret, 2007]. En moins de 3 Ma, le pourcentage de GC à la troisième position des codons est ainsi passé de 56% à 87%. On observe de plus de nombreuses délétions au sein des

[11] Taux de GC prédits en supposant que les taux de substitutions restent constants

introns, ce qui peut expliquer la forte densité en gènes dans les isochores GC-riches [Montoya-Burgos et al., 2003]. Ces délétions peuvent s'expliquer par le fait que des séquences disposant d'une composition en base G+C trop extrême (ADN non-codant dans un hotspot de recombinaison) provoqueraient des sauts de réplication [Lunter et al., 2006]. Les régions codantes seraient prémunies de ces compositions extrêmes car seules les troisièmes positions de codon (sites synonymes) seraient massivement touchées par la conversion génique biaisée.

2.4.3 Evidence et origine d'un biais de réparation vers GC

Chez la levure et chez les mammifères, la réparation des mésappariements est connue pour être biaisée vers GC [Brown and Jiricny, 1988, Birdsell, 2002] (voir Table 2.1). Ce biais de réparation avéré dans les cellules mitotiques pourrait expliquer la conversion génique biaisée vers GC lors des méioses. Cette explication est d'autant plus probable que ce sont les même structures qui sont responsables des réparations et de la recombinaison. Corroborant expérimentalement cette hypothèse, un biais de conversion favorisant les allèles GC a été démontré par typage des produits de la méiose de la levure [Mancera et al., 2008].

Pourquoi un biais de réparation vers GC ? Le biais de réparation extrême du mésappariement G:T vers G:C (92% des cas, Table 2.1) suggère que cela serait un moyen de contrer l'hypermutabilité des nucléotides C méthylés (voir Figure 2.6). En effet, les C suivis d'un G (dinucléotide noté CpG) sont souvent associés à un groupement méthyl. Chez les mammifères, la méthylation des CpG permet d'éteindre l'expression d'un gène. Ce mécanisme qualifié d'épigénétique est indispensable. Il a probablement été inventé au cours de l'évolution des vertébrés pour répondre à la nécessité de réprimer un très grand nombre de gènes tissus-spécifiques, ou encore stopper l'activité des éléments transposables [Suzuki et al., 2007] (ce point pourrait par ailleurs avoir été une pré-adaptation indispensable à l'évolution du placenta [Sekita et al., 2008]). Regroupés en "ilôts" en amont d'un gène, les CpG font de plus office de régions régulatrices. Il ne serait donc pas étonnant que l'évolution ait favorisé des pro-

Mismatch	**Repaired to G:C**	Repaired to A:T	Unrepaired
Monkey			
G:T	**92**	4	4
A:C	**41**	37	22
C:T	**60**	12	28
A:G	**27**	12	61
Yeast			
G:T	**53**	37	10
A:C	**44**	34	21
C:T	**48**	33	18
A:G	**48**	36	16

Table 2.1 – Biais de réparation dans les cellules de singe et de levure (pourcentages). Des plasmides avec des mésappariement du type indiqué ont été introduits dans des cellules pour tester d'éventuels biais dans la direction de la réparation. D'après [Brown and Jiricny, 1988] et [Birdsell, 2002].

cessus de réparation restaurant efficacement les fréquentes mutations de ces cytosines dont la méthylation joue un rôle si crucial. En cas de mésappariement T:G, les mécanismes de réparations corrigent ainsi plus vers C:G (restauration de l'appariement originel) que T:A. Les T sont ainsi considérées comme étant, à juste titre, plus promptes à résulter d'un événement de mutation. Consistant avec cette idée, le biais de réparation du mésappariement T:G chez la levure (qui n'a pas de méthylation) est moins prononcé (voir Table 2.1).

S'ajoute à cela l'existence d'un biais généralisé des mutations vers AT [Lynch and Walsh, 2007]. Basé sur de nombreuses sources, M. Lynch calcule un biais s'étendant de 1.59 à 2.77[12] chez des organismes aussi divers que le bacille de la lèpre, le colibacille, la levure, la drosophile, le hamster ou l'homme. Pour contrer un tel biais mutationnel vers AT, un biais de réparation vers GC semble donc logique et nécessaire.

Cet à priori négatif vis à vis des nucléotides AT profite ainsi aux nucléotides GC lorsque les systèmes de réparation doivent arbitrer un mésappariement

[12] ratio du nombre de mutation vers AT sur le nombre de mutations vers GC

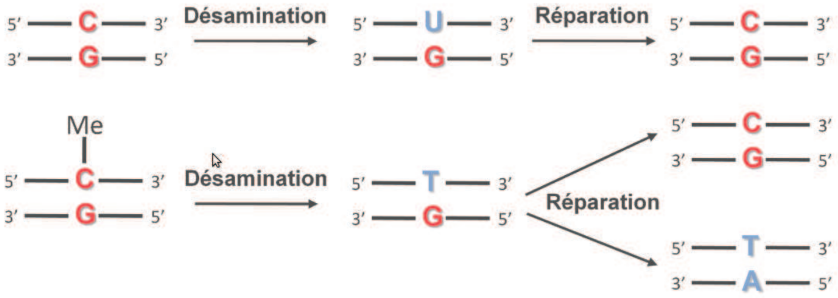


FIGURE 2.6 – Hypermutabilité des Cytosines (C) méthylées. En temps normal, la fréquente désamination (élimination d'un groupe amine) d'une cytosine produit un uracile (U). L'uracile (U) remplace la thymine (T), mais n'est utilisé que lorsque l'information génétique est sous forme d'ARN. Dans l'ADN, le mésappariement U:G est ainsi facilement réparée. En revanche, la désamination d'une cytosine (C) méthylée aboutit à une thymine (T). Le mésappariement T:G peut ainsi être réparé en C:G (état initial) ou en T:A (mutation du C originel vers T).

en méiose. Mais les nucléotides GC ne sont pas les seuls à profiter des imperfections des systèmes de réparation de l'ADN. Certains éléments égoïstes tels que les endonucléase de homing les exploitent même de manière directe [Burt and Trivers, 2006]. Ces introns, optionnels et sans fonction connue, codent une enzyme qui reconnaît et coupe des chromosomes qui ne contiennent pas l'une de leur copie. Ainsi, ils obligent les systèmes de réparation à les copier en lieu et place du chromosome coupé pour réparer les dégâts. Bien que leur nature égoïste soit exprimée de manière plus active, ils ne causent que des dommages négligeables au génome. Loin d'être aussi anecdotique, la conversion génique biaisée vers GC concerne la moitié de tout l'alphabet du vivant. Cette omniprésence lui confère la possibilité d'être la cause de maladaptation moléculaire, tout particulièrement autour des points chauds de recombinaison. Sous couvert de bonne réputation auprès des systèmes de réparation, les mutations GC délétères y bénéficient de passe-droits pour envahir nos génomes.

2.4.4 Conversion génique biaisée et sélection naturelle

Le talon d'achille de nos génomes

Au niveau de l'individu, la sélection naturelle n'avantage que la fixation des mutations positives (augmentation de la valeur sélective de l'organisme). La conversion génique biaisée ne favorise elle que celles qui sont vers GC, sans se soucier qu'elles soient ou non délétères. Ces deux niveaux de sélection peuvent ainsi entrer en conflit lorsque survient une mutation délétère vers GC : bien que contre-sélectionnée au niveau de l'individu, les systèmes de réparation la font bénéficier d'un avantage sélectif au niveau du nucléotide. Ceci est également valable pour les mutations positives vers AT : bien que bénéfiques pour l'individu, elles sont contre-sélectionnées au niveau du nucléotide.

Ce constat est d'autant plus important qu'il s'agit souvent du niveau de sélection le plus bas qui a la priorité sur les autres : un allèle n'aura jamais l'occasion d'être sélectionné au niveau de l'individu s'il est contre-sélectionné au niveau du gène[13]. Parce que la conversion génique biaisée est une forme de sélection qui agit au niveau le plus bas, selon la nature même du nucléotide, son influence est théoriquement colossale.

Peut-elle jouer un rôle positif et favoriser la fixation des mutations avantageuses vers GC ? Ou bien encore empêcher celle des délétères vers AT ? Probablement, mais une analyse théorique indique que le faible bénéfice qu'elle pourrait procurer n'équivaut pas au fardeau de délétère qu'elle crée [Galtier et al., 2009]. Chez les primates, elle aurait deux conséquences majeures : une augmentation du taux de substitution total, et une proportion accrue de substitutions délétères (de 4 à 16%). Cela suggère bien qu'elle favorise la fixation de mutations "non-désirées", des mutations qui seraient normalement éliminées en son absence.

Cet exemple théorique peut être illustré par la translocation récente du gène Fxy dans la région pseudoautosomale du chromosome X (point chaud de re-

[13] C'est ce raisonnement qui a poussé certains évolutionniste à considérer que la sélection n'agissait quasiment jamais au niveau du groupe, niant la possibilité évolutive d'un altruisme désintéressé [Dawkins, 1976, Williams, 1966].

combinaison). Comme nous l'avons déjà vu, cette translocation a été suivie d'une augmentation du taux de GC non-codant (sites synonymes et introns), mais aussi et surtout de 28 substitutions non-synonymes, toutes de AT vers GC [Montoya-Burgos et al., 2003]. Ces 28 substitutions se sont déroulées en moins de 3 millions d'années (date estimée de la translocation du gène Fxy). A titre de comparaison, on ne compte que 4 différences entre le rat et l'homme, des animaux qui ont pourtant divergé depuis plus de 80 millions d'années. Comme nous l'avons évoqué en partie 2.1, la sélection purifiante limite habituellement les substitutions non-synonymes, majoritairement neutres ou faiblement délétères. Si elle semble avoir rempli son office lors des 80 Ma d'évolution qui séparent le rat et l'homme, seuls 3 Ma de recombinaison intensive chez la souris auront suffit à la conversion génique biaisée pour l'outrepasser et promouvoir la fixation de 28 mutations non-synonymes vers GC. Sous forte recombinaison, ces dernières ont pu passer entre les mailles du filet de la sélection purifiante, et cela malgré leur nature vraisemblablement délétère [Galtier and Duret, 2007].

Véritables "Talon d'Achille" de nos génome, nos points chauds de recombinaison maintiennent les mutations potentiellement délétères [Galtier and Duret, 2007, Galtier et al., 2009]. Ainsi, Necsulea et al. démontrent que leur propagation dans la population humaine est favorisée par la conversion génique biaisée [Necșulea et al., 2011]. Le spectre des mutations non-synonymes observées porte en effet son empreinte, y compris lorsqu'on ne considère que celles qui ont un effet néfaste sur la fonction de la protéine ou qui sont responsables de maladies génétiques humaines.

Un niveau de sélection à part entière

Lors d'un mésappariement, nous avons vus que les nucléotides G et C sont sélectionnés par les système de réparation. Si l'on peut bel et bien parler de sélection "naturelle", celle-ci opère à un niveau tout autre que celui de l'individu. Sous son influence, les allèles GC ont un avantage populationnel sur les allèles AT : leur dynamique de population est donc très similaire à celle d'un allèle qui procure un avantage adaptatif [Nagylaki, 1983].

Déceler les épisodes adaptatifs qui font de l'homme un primate différent du chimpanzé est un objectif des plus attrayant. C'est avec cette idée que des chercheurs ont scruté le génome humain à la recherche d'éléments non-codant conservés entre les Vertébrés, mais très divergents chez l'homme [Prabhakar et al., 2006]. Ont ainsi été identifiées plusieurs régions accélérées chez l'homme (notées HARs, pour human accelerated regions). L'accélération brutale et spécifique des plus rapides d'entre elles sont typiquement interprétées comme autant de signatures de la sélection naturelle. Ces signatures ont cependant été systématiquement interprétées en terme adaptatifs, allant jusqu'à expliquer les différences cérébrales entre l'homme et le chimpanzé [Pollard et al., 2006, Prabhakar et al., 2008].

Croire qu'accélération de l'évolution moléculaire rime toujours avec adaptation est pourtant erroné. Tels que la théorie neutraliste de l'évolution moléculaire et les différents niveaux de sélection nous l'ont appris, les patrons de substitutions ne sont que très rarement le signe d'une adaptation dirigée vers l'individu. Une analyse plus fine a ainsi permis de démontrer que ces HARs se trouvent non-seulement dans des régions fortement recombinantes, mais surtout qu'elles subissent une majorité de substitutions AT→GC, ce biais s'étendant même au-delà des séquences supposées fonctionnelles [Galtier and Duret, 2007]. Ces trois critères réunis indiquent bien qu'il s'agit là de l'œuvre de la conversion génique biaisée. Cette brusque fixation de substitutions vers GC n'a ainsi probablement rien à voir avec une séduisante spécificité adaptative du cerveau humain, mais résulte d'une sélection au niveau du nucléotide, inutile voire délétère pour l'organisme. Ainsi, il a été récemment démontré qu'une des HARs avait en réalité perdu sa fonction originelle chez l'homme (chez la souris, remplacer cette séquence de 81 paires de base par la séquence de l'homme aboutit au même résultat que de la supprimer) [Sumiyama and Saitou, 2011]. Avant de considérer que l'évolution d'une séquence est adaptative, il convient donc de considérer en premier lieu l'hypothèse d'une évolution neutre ou d'un mécanisme non-adaptatif. Dans nos génomes, la sélection est non-seulement rarissime, mais elle n'agit de plus que très peu au niveau de l'individu : nucléotides, séquences, haplotypes et chromosomes sont autant de cibles moins attendues, mais bel et bien prioritaires.

2.5 Génèse, maintien et évolution des isochores

2.5.1 La conversion génique biaisée : un processus universel ?

Potentiellement maladaptative, la conversion génique biaisée ne concerne pas que l'homme et ses apparentés, mais pourrait bien être largement répandue chez l'ensemble des eucaryotes. En effet, par des preuves plus ou moins directes, des liens entre taux de recombinaison et taux de GC sont trouvés chez divers organismes : la levure [Gerton et al., 2000], le nématode [Marais et al., 2001], la drosophile [Marais et al., 2003], la paramécie [Duret et al., 2008] et les angiospermes[14] [Glémin et al., 2006, Muyle et al., 2011, Serres-Giardi et al., 2012]. Plusieurs analyses comparatives récentes aboutissent ainsi à cette conclusion : la conversion génique biaisée vers GC semble bel et bien active chez la plupart des eucaryotes [Escobar et al., 2011, Capra and Pollard, 2011, Pessia et al., 2012], bien que son influence puisse varier en intensité.

Malgré une telle omniprésence, les génomes complets actuellement disponibles sont loin de tous présenter une structuration en isochore aussi extrême que celle des mammifères et des oiseaux. Une telle hétérogénéité le long des chromosomes, affectant à la fois codant et non-codant, n'est en effet jusque-là relevée que chez l'abeille [Jø rgensen et al., 2006]. Dans un article faisant suite à la publication du premier génome complet de reptile [Alföldi et al., 2011], Fujita et al. concluent à l'absence quasi-totale d'isochores dans le génome du lézard *Anolis carolinensis* [Fujita et al., 2011]. Faut-il en conclure pour autant que la conver-

[14] Plantes à fleurs

sion génique biaisée n'y a jamais sévi ? Bien qu'indispensable à la genèse des isochores, elle n'est pourtant pas le seul élément nécessaire à leur maintien. Contrainte majeure des patrons de recombinaison, l'évolution de la taille du génome et du caryotype ne doit pas être négligée.

2.5.2 L'influence de la taille du génome et du caryotype

Sous l'hypothèse de conversion génique biaisée, la variabilité du taux de GC est intrinsèquement liée à la variabilité spatiale des taux de recombinaisons. Outre un biais de réparation vers GC, la condition *sine qua none* pour l'apparition d'un isochore GC-riche est donc le maintien prolongé d'un point chaud de recombinaison. Identifier les paramètres qui influencent leur distribution apparaît donc comme crucial pour comprendre la distribution des taux de GC.

A l'échelle caryotypique, on sait que les taux de recombinaisons sont fortement contraints par la taille des bras chromosomiques : chez l'homme et le poulet, les chromosomes les plus petits ont des taux de recombinaison plus élevés [Lander et al., 2001, Chicken and Sequencing, 2004] (Figure 2.7a). Ce phénomène est dû à la nécessité d'avoir au moins un événement de recombinaison sur chaque bras chromosomique lors de chaque méiose [de Villena and Sapienza, 2001]. Conformément aux attendus de la conversion génique biaisée, les taux de GC sont ainsi corrélés négativement à la taille d'un chromosome (Figure 2.7b) : un grand chromosome subit moins de recombinaison méiotique par base, ce qui procure moins d'occasions à la conversion génique biaisée d'affecter son GC moyen. En revanche, le fort taux de recombinaison par base d'un chromosome de petite taille permet une inexorable augmentation de son taux de GC.

Ainsi, l'évolution des caryotypes (par translocation, fission ou fusion chromosomique) est un déterminant majeur du taux de recombinaison. Parmi les mammifères, celui-ci est d'ailleurs presque parfaitement bien prédit par le nombre de bras chromosomiques (R2>0.8) [de Villena and Sapienza, 2001]. Parmi les exemples les plus extrêmes, le génome de l'opossum (*Monodelphis domes-*

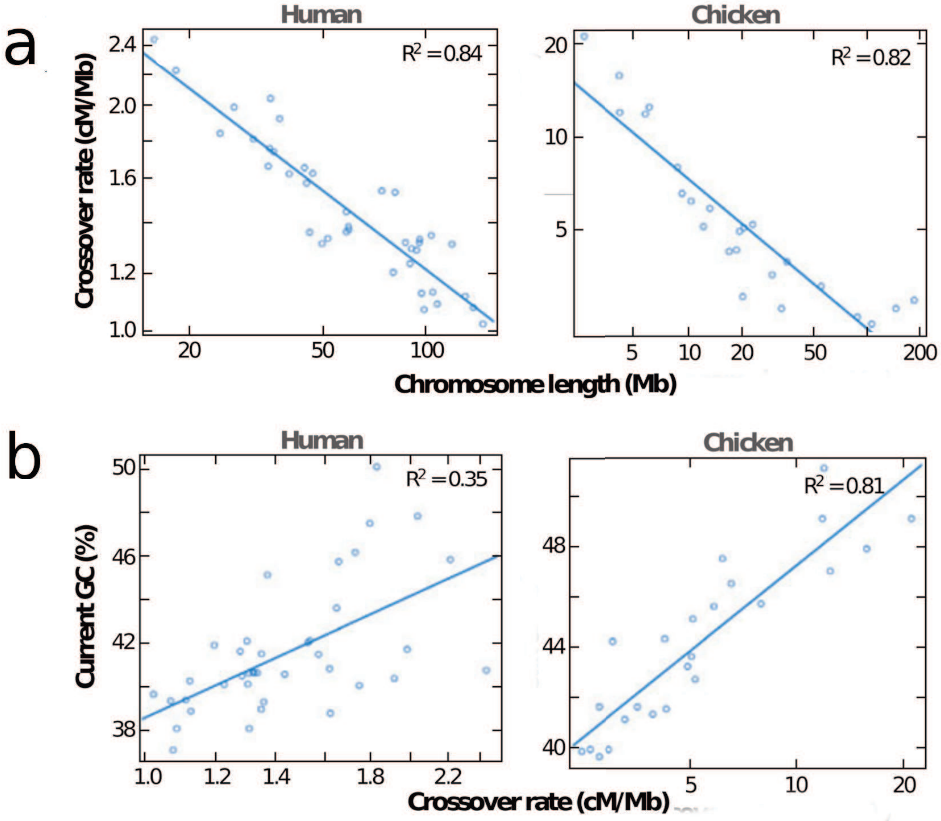


FIGURE 2.7 – Corrélation entre tailles de chromosomes, taux de recombinaison et contenu en GC de l'homme et du poulet. Chez l'homme, le taux de recombinaison corrèle encore plus avec le taux de GC équilibre ($R2 = 0.66$, non montré) qu'avec le taux de GC actuel ($R2 = 0.35$, graph b de gauche). D'après [Duret, 2009].

tica, un marsupial) est composé de 8 chromosomes géants accompagnés d'un petit chromosome X. Comme attendu, les 8 autosomes ont un taux de recombinaison faible et sont AT-riches, tandis que le petit chromosomes X a un taux de recombinaison élevé et est GC-riche [**?**]. L'exemple opposé est fourni par l'ornithorynque (*Ornithorhynchus anatinus*), dont le génome est morcelé en 52 chromosomes et affiche un taux de GC record (45.5%, à comparer au 37.7% de l'opossum [Warren et al., 2008]).

A l'échelle des Vertébrés, de tels liens entre caryotype et distribution du GC ont été montrés au cours de cette thèse. Conformément à l'hypothèse de conversion génique biaisée, on peut observer qu'un génome aux tailles de chromosomes hétérogènes fait preuve de taux de GC tout aussi hétérogènes (Figure 2.8). La taille moyenne des chromosomes d'un génome dépend en grande partie de la taille du génome lui même. Doté en moyenne de chromosomes plus courts, un petit génome reçoit en moyenne plus de recombinaisons par positions nucléotidiques. De telles corrélations entre taille de génome et taux de recombinaisons ont ainsi été rapportées à l'échelle de tous les eucaryotes, mais aussi au sein même des unicellulaires, des plantes terrestres, des invertébrés et des vertébrés [Lynch et al., 2006]. Ici encore, nos résultats sont conformes aux prédictions de la conversion génique biaisée, et montrent une corrélation entre les tailles des génomes de vertébrés et leur taux de GC en troisième position des codons (noté GC3%, Figure 2.9). L'augmentation spectaculaire du génome des dipneustes (plus gros génome animal connu à ce jour, jusqu'à 37 fois la taille du génome humain) a ainsi probablement noyé l'apparition ou le maintien d'isochores. Sans surprise, son taux de recombinaison théorique est si infime qu'il présente le taux de GC3 génique le plus bas et le plus homogène du jeu de données.

En réduisant à néant les taux de recombinaison, les augmentations de taille du génome peuvent ainsi empêcher la conversion génique biaisée de façonner durablement leur composition nucléotidique. De la même manière, des réarrangements caryotypiques trop fréquents devraient empêcher le main-

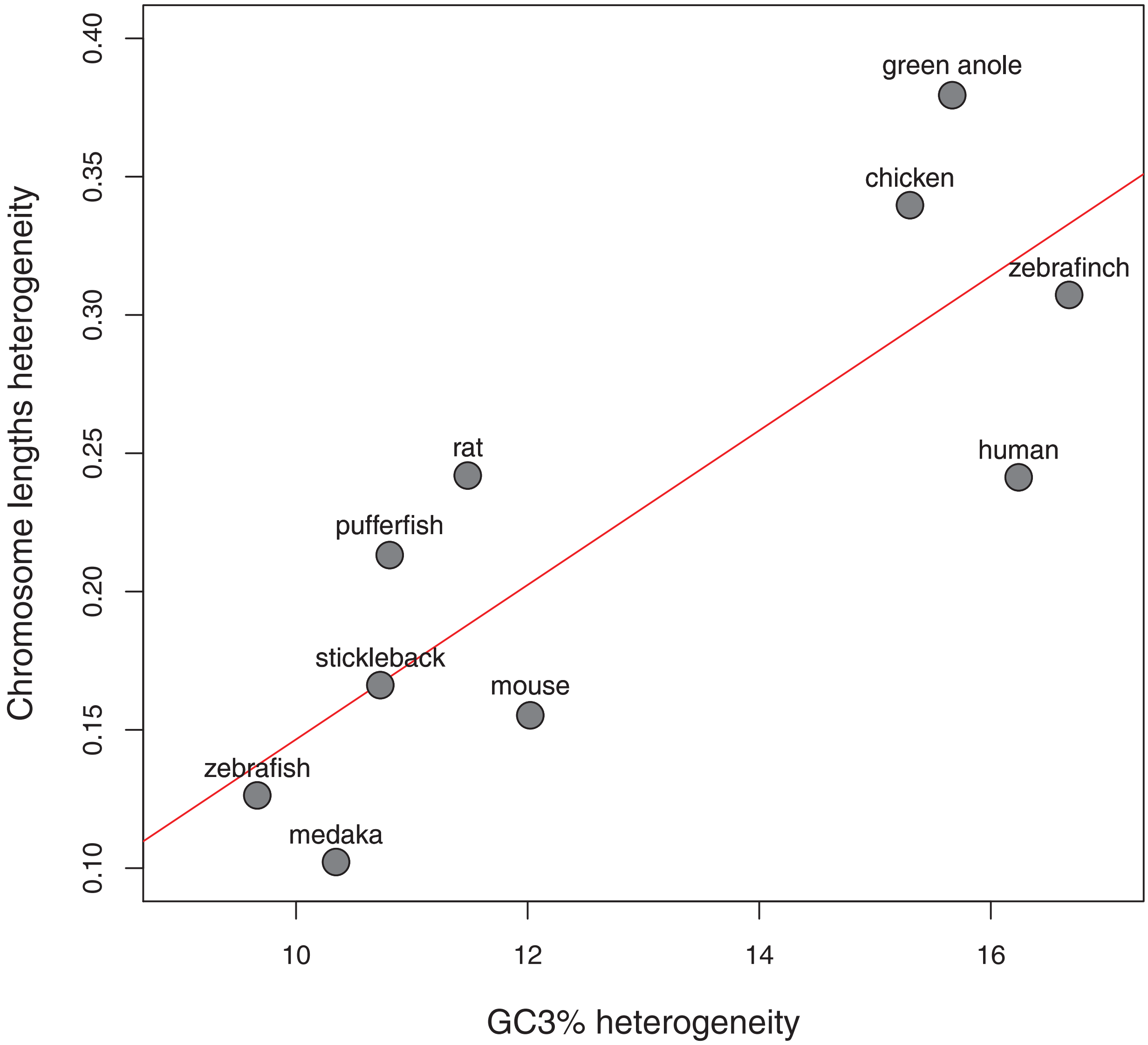


FIGURE 2.8 – Corrélation entre hétérogéneité du GC3 de 10 Vertébrés et hétérogénéité de leur taille de chromosomes ($R^2 = 0.67$). L'hétérogéneité du GC3 est mesurée via l'écart-type des GC3 de tous les gènes d'une espèce disponibles sur Ensembl [Birney et al., 2004]. Cette analyse a été réalisée dans le cadre du stage de Master 2 recherche de Emeric Figuet, encadré au cours de cette thèse.

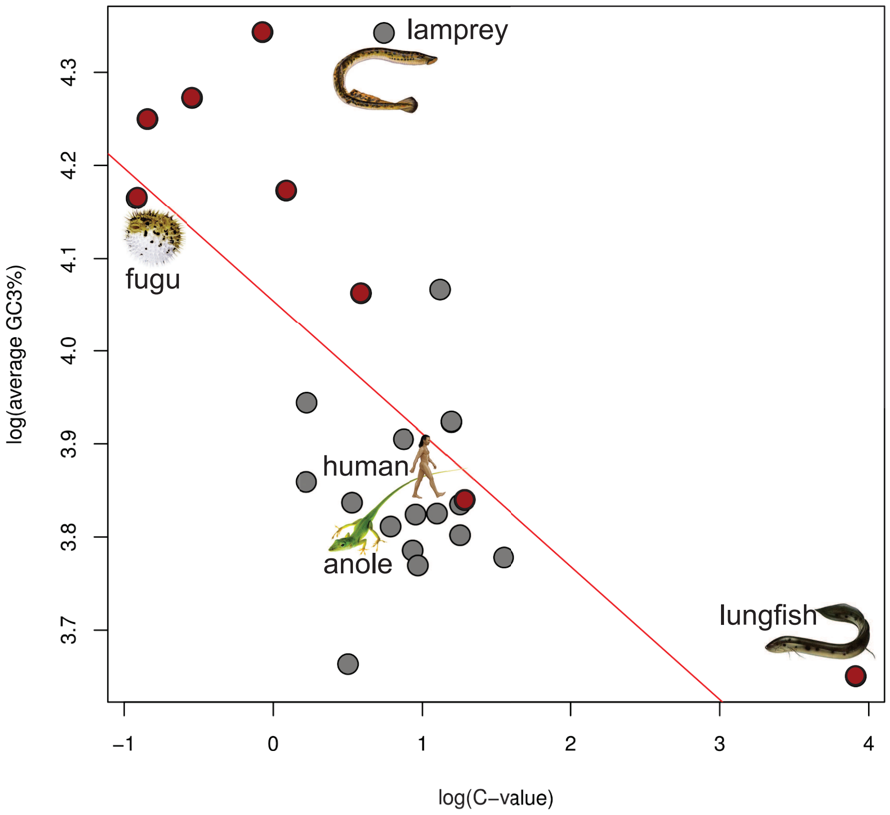


FIGURE 2.9 – Corrélation entre le GC3 moyen de 23 Vertébrés et leur taille de génomes. Résultats obtenus à l'aide de 248 alignements de gènes orthologues (complétés à partir d'un article de Chiari et al. [Chiari et al., 2012]), tailles de génomes obtenues à partir de The Animal Genome Size Database [Gregory et al., 2007]. On note que la Lamproie *Petromyzon marinus* est l'espèce qui s'éloigne le plus de la tendance générale. Son taux de GC trop élevé pour la taille de son génome s'explique cependant par l'extrême morcellement de son caryotype (2n=168). Notons que si l'on essaye de gommer les effets du caryotype en ne conservant que les sarcoptérygiens aquatiques (8 "poissons" aux caryotypes très homogènes et relativement conservés, points rouge sur le graph), le R^2 passe de 0.5 à 0.83.Cette analyse a été réalisée dans le cadre du stage de Master 2 recherche de Emeric Figuet, encadré au cours de cette thèse.

tien de points chauds de recombinaison stables, conduisant inexorablement à une homogénéisation des taux de GC. Conformément à cette prédiction, on observe des taux de réarrangements chromosomiques exceptionnellement élevé chez le rat et la souris [Gibbs et al., 2004], ce qui peut expliquer les nombreux éléments qui indiquent que leur structuration en isochore est en pleine érosion [Mouchiroud et al., 1988, Gibbs et al., 2004].

2.5.3 Pourquoi le génome d'Anolis carolinensis n'a-t-il pas d'isochores ?

Comme évoqué précédemment, le seul génome complet de reptile actuellement disponible révèle avec surprise une absence d'isochores [Fujita et al., 2011] (Figure 2.10). Les lézards descendent du même ancêtre que celui des mammifères et des oiseaux. Or, la similitude entre les paysages nucléotidiques de l'homme et du poulet suggère qu'ils l'ont hérité de leur ancêtre commun le plus récent. Comment se fait-il que cette structuration ait été conservée chez les oiseaux et mammifères, mais pas chez l'anole vert ? Fujita et al. suggèrent deux explications : un arrêt ou une inversion de la conversion génique biaisée chez *Anolis carolinensis* (perte secondaire de la structuration commune des mammifères et oiseaux) ou une double apparition indépendante de cette structuration chez les animaux à sang chaud (évoquant l'hypothèse historique d'adaptation à l'homéothermie proposé par Bernardi).

Comme nous l'avons vu lors de la partie précédente, l'absence d'isochore n'implique cependant pas l'absence de conversion génique biaisée. Celle-ci semblant de plus un processus général répandu à l'échelle des eukaryotes [Escobar et al., 2011, Capra and Pollard, 2011, Pessia et al., 2012], des vertébrés (Figure 2.8 et Figure 2.9) et a priori à plus forte raison encore aux amniotes (origine commune des isochores oiseaux et mammifère, Figure 2.10), tout porte à croire qu'elle pourrait être encore active chez *Anolis carolinensis*. Pour le vérifier, nous avons comparé les compositions nucléotidiques des gènes portés par ses chromosomes. Divisés

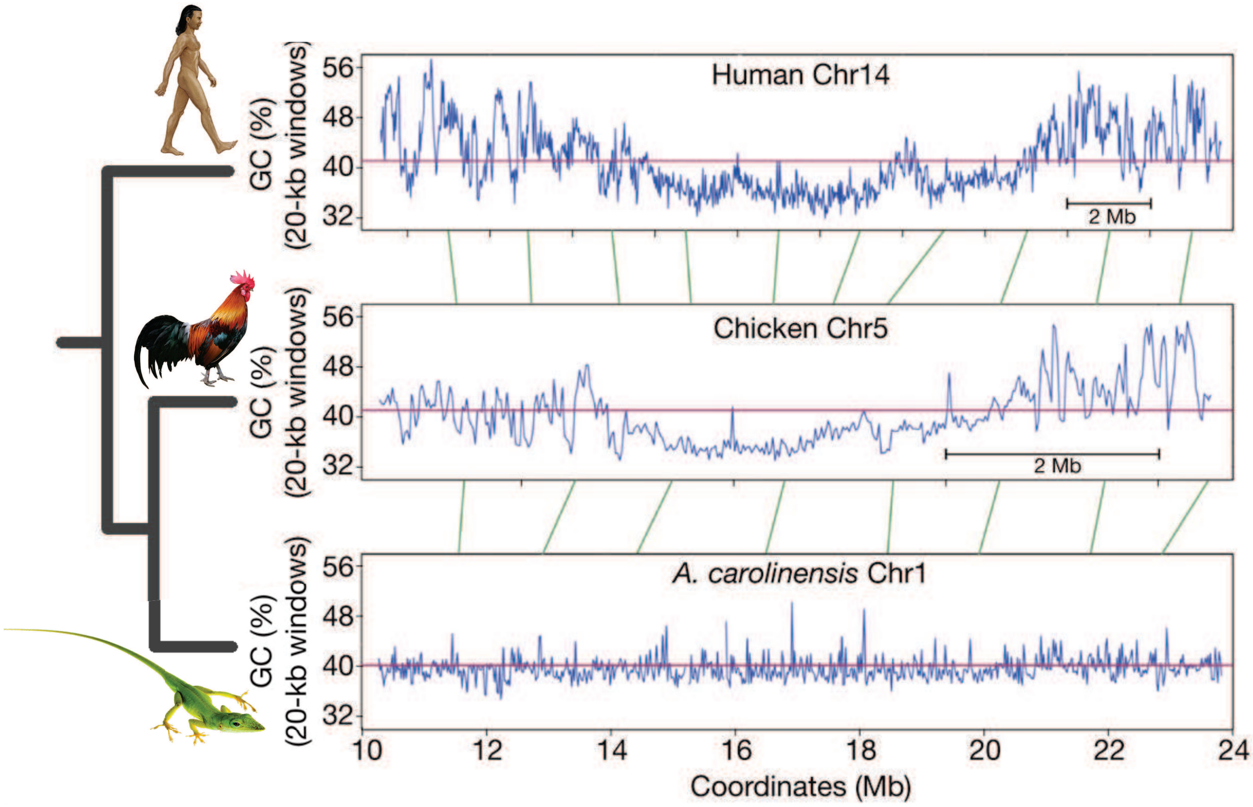


FIGURE 2.10 – Absence d'isochores chez le lézard *Anolis carolinensis*. Ceux des mammifères et des oiseaux semblent pourtant bien être d'origine commune. Adapté d'après [Alföldi et al., 2011].
.

en deux classes de taille très marquées, on s'attend à ce que les microchromosomes d'Anolis aient un taux de GC plus élevé que ses macrochromosomes. Nos résultats sont bien conformes à cet attendu (55.6% pour les microchromosomes, 42.9% pour les macrochromosomes). De plus, le taux de GC3 stationnaire, c'est à dire l'équilibre vers lequel tend la composition en bases, prévoit une poursuite de l'enrichissement du GC des microchromosomes (GC3* = 59,6) et un appauvrissement de celui des macrochromosomes (GC3* = 34.4).

Cet écart qui continue à se creuser suggère ainsi bel et bien que la conversion génique biaisée est encore active chez *Anolis carolinensis*. Comment expliquer qu'elle ne parvienne pas à maintenir une structuration en isochore telle qu'on l'observe chez les mammifères ou les oiseaux ? Des points chauds de recombinaisons particulièrement instables ou trop épars peuvent l'expliquer. Pourtant, le caryotype de ce lézard ne semble pas particulièrement instable (la structuration en macro et micro-chromosomes est partagée avec le poulet), et la taille de son génome n'a rien d'exceptionnelle. Ce n'est pas à l'échelle du génome ou des chromosomes qu'il faut ici se placer, mais à l'échelle de la séquence. Il a en effet été rapporté que le génome d'*Anolis carolinensis* était parsemé d'éléments transposables, provenant de familles inédites, diversifiées et extrêmement jeunes [Novick et al., 2011]. Se répandant de manière rapide dans le génome, il est aisé de comprendre que de telles séquences puissent profondément modifier une structuration nucléotidique lentement acquise, substitution après substitution. Leur insertion anarchique a de plus toutes les chances d'affecter la distribution des points chauds de recombinaisons, en facilitant notamment les réarrangements chromosomiques.

L'absence d'isochore rend ainsi difficile l'observation d'une quelconque trace de la conversion génique biaisée à large échelle, codant et non-codant y compris. En revanche, les 3$^{\text{èmes}}$ positions de codons des gènes en conservent la trace. Peu soumises à sélection (code génétique dégénéré) mais nécessairement conservées malgré les invasions récentes d'éléments transposables, elles sont idéales pour rendre compte des patrons de substitutions de long terme. Bien que ne reflétant pas forcément une structuration en isochores (du moins, en dehors des mammifères et des oiseaux), elles sont donc les plus pertinentes pour scruter l'activité de la conversion génique biaisée.

Ainsi, c'est uniquement via l'analyse des 3$^{\text{ème}}$ positions de codons d'*Anolis carolinensis* que nous avons pu mettre en évidence une forte hétérogénéité des patrons de substitutions de cette espèce, notamment expliquée par son caryotype. Un tel constat s'est avéré impossible à déduire de l'analyse du génome complet, probablement trop marqué par les effets confondants d'une récente invasion d'éléments transposables [Fujita et al., 2011, Novick et al., 2011].

2.6 Conversion génique biaisée et biologie des espèces : des liens potentiels ?

Comme nous l'avons vu tout au long de cette seconde partie d'introduction, les génomes disposent bel et bien de leurs propres règles d'évolution. Bien que spécifiques, ces dernières n'en demeurent pas moins connectées à celles des organismes. Tout comme un individu est contraint par son environnement, un génome est contraint par les attributs de son espèce. Taille, longévité, abondance ou mode de reproduction sont ainsi autant de caractéristiques qui peuvent influencer l'évolution moléculaire.

Chez les mammifères, ces traits d'histoire de vie sont particulièrement diversifiés. De manière générale, on peut organiser cette diversité tout au long d'un continuum. Située à l'une de ses extrémités, la stratégie *r* est celle choisie par la majorité des espèces de petite taille, à longévité faible mais à fécondité et abondance maximale (exemple : la souris). A l'autre extrême, on trouve la stratégie dite *K*, qui rassemble les animaux de grande taille, longévifs mais peu abondants et à faible fécondité (exemple : l'éléphant).

Couplée au nombre abondant de données génomiques, cette diversité biologique des mammifères a permis de mettre en évidence plusieurs liens entre évolution moléculaire et traits d'histoire de vie. Ainsi, les espèces à stratégie *r* présentent des taux de substitution plus élevés [Nabholz et al., 2008, Bromham, 2011], tandis que les espèces à stratégie *K* accumulent une plus grande part de substitutions non-synonymes, la faute à la moins grande efficacité de la sélection purifiante dans de si petites tailles de populations [Popadin et al., 2007, Nikolaev et al., 2007].

Se pourrait-il que de tels liens existent avec la conversion génique biaisée ? Si son action est clairement contrainte par les caryotypes et les tailles de génome, pourquoi la biologie des espèces n'aurait-elle pas son mot à dire ?

A partir de plus d'une trentaine de génomes complets de mammifères, cette thèse se propose d'évaluer l'évolution de leurs taux de GC. Du Crétacé à nos jours, cette guerre nucléotidique révélera des liens insoupçonnés avec la biologie des espèces, nous fournissant de précieux indices sur la nature et la diversification de nos mystérieux ancêtres contemporains des dinosaures non-aviens.

Deuxième partie

Articles

This brief foreword explains how the following research articles are organized.

Chapter 3, "Contrasting GC-content dynamics across 33 mammalian genomes: Relationship with life-history traits and chromosome sizes", is the first article published during this PhD thesis (Genome Research 2010). Its initial aim was to depict the nucleotidic landscapes, the so-called isochore structure[15] of a representative number of mammal species. Indeed, this necessary work filled a gap in the previous litterature, which was focused on a handful of model species, like apes, mouse or rat. Actually, as is shown in this article, the isochore evolution of these species did not reflect the placental main trends. This analysis revealed the existence of a relationship between GC-content evolution and species life history traits, in agreement with the biased gene conversion hypothesis[16]. This result was the starting point of the next article.

Chapter 4, entitled "Genomic Evidence for Large, Long-Lived Ancestors to Placental mammals" (Molecular Biology and Evolution 2012), it attempts to shed light on early mammalian evolution. Around ten years ago, molecular phylogeny shaked the old morphological tree[17]. In addition to this taxonomical upheaval, molecular clock estimations revealed that the origin of current placentals probably predates the famous KT crisis, 65 Mya. However, none of the cretaceous mammalian fossils known so far are considered as placentals. Nothing is therefore known about the form, size or lifestyle of our early ancestors, which lived side by side with non-avian dinosaurs. In order to caracterize them, we used two molecular markers of life history traits, namely GC conservation level and dN/dS ratio. Unexpectedly, our results exclud small shrew-like placental ancestors, contradicting one of the most frequently told story in evolution.

The next biggest mystery about our placental ancestors is probably the way they diversified, particularly regarding the continental drift which occured during the cretaceous. Indeed, the placental root position is one of the most controversial

[15] see Figure 2.3 in Introduction for a quick view of isochore structure in human chromosomes.

[16] see Figure 2.4 in Introduction for a quick view of the mecanism.

[17] see Figure 1.6 in Introduction

node of the mammal tree. During the last decade, a number of contradicting studies have supported one out of three different hypothesis[18]. The most popular one is the Atlantogenata hypothesis, which joins species originating from South America (Xenarthra) and Africa (Afrotheria). In the work presented in chapter 5, we show that this hypothesis is best supported by GC-rich genes, which are prone to produce topological errors because of incomplete lineage sorting or homoplasy induced by biased gene conversion. AT-rich, which are more reliable phylogenetic markers, suggest that Afrotheria is the most basal clade within Placentalia.

The last chapter joins two methodological articles. Based on probabilistic substitution mapping, this method allows to characterize variations in the substitution process across lineages. Without the need to estimate one parameter per branch as in non-homogenous models, this approach is fast and well adapted to the increasing number of genomic datasets. Available with the first article, the software MapNH was used in the chapter 4 article to estimate a dN/dS ratio in each branch.

[18] see Figure 1.8 in Introduction

Chapitre 3

Contrasting GC-content dynamics across 33 mammalian genomes: Relationship with life-history traits and chromosomes sizes

Contrasting GC-content dynamics across 33 mammalian genomes: Relationship with life-history traits and chromosome sizes

Jonathan Romiguier, Vincent Ranwez, Emmanuel J.P. Douzery, et al.

Genome Res. 2010 20: 1001-1009 originally published online June 7, 2010
Access the most recent version at doi:10.1101/gr.104372.109

Contrasting GC-content dynamics across 33 mammalian genomes: Relationship with life-history traits and chromosome sizes

Jonathan Romiguier, Vincent Ranwez, Emmanuel J.P. Douzery, and Nicolas Galtier[1]

Université Montpellier 2, CNRS UMR 5554—Institut des Sciences de l'Evolution, 34095 Montpellier, France

The origin, evolution, and functional relevance of genomic variations in GC content are a long-debated topic, especially in mammals. Most of the existing literature, however, has focused on a small number of model species and/or limited sequence data sets. We analyzed more than 1000 orthologous genes in 33 fully sequenced mammalian genomes, reconstructed their ancestral isochore organization in the maximum likelihood framework, and explored the evolution of third-codon position GC content in representatives of 16 orders and 27 families. We showed that the previously reported erosion of GC-rich isochores is not a general trend. Several species (e.g., shrew, microbat, tenrec, rabbit) have independently undergone a marked increase in GC content, with a widening gap between the GC-poorest and GC-richest classes of genes. The intensively studied apes and (especially) murids do not reflect the general placental pattern. We correlated GC-content evolution with species life-history traits and cytology. Significant effects of body mass and genome size were detected, with each being consistent with the GC-biased gene conversion model.

[Supplemental material is available online at http://www.genome.org.]

The mammalian genome is characterized by its high spatial heterogeneity in base composition. The average GC content of a 100-kb fragment of the human genome can be as low as 35% or as high as 60%, a range that is twice as wide as that typically observed in teleostean fishes, for instance (International Human Genome Sequencing Consortium 2001). This property of the human genome, identified in the pre-genomic era (Bernardi et al. 1985), was called the "isochore structure." The discovery of isochores immediately raised questions as to the reasons for their existence. How did isochores originate? How have they been maintained? Are they adaptive or merely the consequence of neutral evolutionary processes? These questions are important because GC content in mammals is correlated with a number of genomic features that are potentially relevant from a functional viewpoint, for example, gene density, transposable element distribution, methylation rate, recombination rate, and expression levels (Eyre-Walker and Hurst 2001; Kudla et al. 2006).

To further investigate these issues, a comparative approach was undertaken to characterize the evolutionary dynamics of isochores. At the vertebrate level, human-like isochores were reported in sauropsids (birds and "reptiles") (Hughes et al. 1999; International Chicken Genome Sequencing Consortium 2004; Kuraku et al. 2006), but not in teleostean fishes and lissamphibians (Bernardi and Bernardi 1990), suggesting that compositional heterogeneity evolved in the amniote ancestor. Within mammals, it was established early that mouse and rat genomes have a substantially more homogeneous GC content than the human genome (Mouchiroud et al. 1988). Phylogenetic analyses suggested that the ancestral placental genome structure was probably close to the human pattern, with the homogeneous mouse and rat pattern being the derived state (Galtier and Mouchiroud 1998). Duret et al. (2002) and Belle et al. (2004) confirmed this result and proposed that the GC-richest components of the mammalian genome have been eroded in rodents, but also to various extents in primates, artiodactyls, and marsupials. This view is debatable. Alvarez-Valin et al. (2004) suggested that the detected erosion of isochores in Duret et al. (2002) was due to methodological bias. Li et al. (2008) reproduced the results of Belle et al. (2004), but reported that there apparently has been no erosive trend in lagomorphs.

The above-reviewed studies used the third-codon positions of aligned coding sequences to characterize genomic GC-content evolution. GC3 (third-codon position GC content) is strongly correlated with the flanking GC content in humans (Mouchiroud et al. 1988). This is a convenient measure because orthologous genes are reasonably easy to identify, align, and compare across mammals. Sequence analysis of repeated elements—an independent source of data—provided a similar picture of GC-content evolution in placental mammals (Arndt et al. 2003; Webster et al. 2005), namely, a decline in GC content in the GC-richest regions of the genome, thus strengthening the erosion hypothesis. All of this literature, however, was based on just a handful of model species, with the main focus being on the human versus mouse comparison.

In addition to this exploration of GC-content dynamics of genomes, progress has been made in our understanding of the underlying evolutionary forces. Bernardi et al. (1985) first claimed that isochores were an adaptation to endothermy, but the discovery of a similar structure in cold-blooded amniotes (Hughes et al. 1999) disqualified this proposal (see also Belle et al. 2002; Ream et al. 2003). Subsequent selective scenarios invoking a higher stability of RNA and proteins in the GC-rich context (Bernardi 2007) have received no empirical support and have failed to explain the spatial heterogeneity in GC content. Neutral processes therefore attracted attention. Wolfe et al. (1989) proposed that the point mutation process could be GC-biased in some regions of the genome, and AT-biased in others in relation with replication timing. Although replication origins were locally associated with spatial shifts in base composition (Watanabe et al. 2002; Schmegner et al. 2007), this effect would probably not account for the genome-wide distribution in GC content (Eyre-Walker and Hurst 2001).

[1] **Corresponding author.**
E-mail nicolas.galtier@univ-montp2.fr; fax 33-467-14-36-10.

Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.104372.109.

Over the last decade, an alternative hypothesis potentially explaining GC-content evolution in mammals has been examined: GC-biased gene conversion (gBGC) (Galtier et al. 2001; Webster et al. 2005; Duret and Galtier 2009). According to this model, a bias in the DNA repair machinery would result in meiotic distortion favoring G and C over A and T alleles in highly recombining regions (Eyre-Walker 1993; Galtier et al. 2001). Indirect empirical evidence supporting the gBGC hypothesis has accumulated: GC-biased pattern of allele segregation at polymorphic sites in humans and mice (Eyre-Walker 1999; Webster and Smith 2004), especially in high-recombining regions (Spencer 2006); recombination-driven increase in GC content in primates (Meunier and Duret 2004), mice (Montoya-Burgos et al. 2003), and birds (Webster et al. 2006); and GC-content increase in genes undergoing ectopic gene conversion (Galtier 2003; Kudla et al. 2004). Duret and Arndt (2008) showed that the gBGC model correctly predicts genome-wide patterns of nucleotide substitution in humans, given the available information on recombination rates in this species. Although the evidence is compelling (Duret and Galtier 2009), note again that most of these studies were conducted in hominid primates or murid rodents, i.e., just two mammalian families. Hence, interpreting the isochore dynamics in terms of spatiotemporal variations in gBGC strength would currently be highly speculative.

Insufficient taxonomic sampling is therefore an obvious limitation of current studies on isochores. The findings of previous studies, in which just six to eight genomes were typically analyzed, could not link GC-content evolution to species biology, ecology, or cytology, so there is still no convincing explanation for the observed diversity. We do not know why isochores are being eroded and whether this pattern applies to all mammals, especially since the single potential GC-driving force for which we have arguments, i.e., gBGC, has been documented in only two groups. This study was designed to clarifying the evolutionary dynamics of GC content in placental mammals through an analysis of 1138 orthologous genes and their flanking regions from 33 fully sequenced genomes. We reconstructed—in a time-heterogeneous maximum-likelihood framework—the distribution of GC3 in the placental ancestral genome, characterized its evolution in representatives of 16 orders and 27 families, and correlated GC-content dynamics with species life-history traits and karyotypes. Our results substantially modify the current view of isochore dynamics in mammals. We show that erosion is not a general rule, while revealing highly diverse trends across lineages and investigating evolutionary processes that could potentially explain these variations.

Results

GC3 dynamics in placental mammals

Table 1 provides the main characteristics of the GC3 distribution across 1138 genes for each of the 33 analyzed species (Fig. 1), plus the reconstructed GC3 of the most recent common ancestor of

Table 1. Characteristics of the distribution of GC3 across 33 species and 1138 genes

Species	Common name	Abbreviation[a]	Mean GC3%	SD GC3 (%)	r^2[b]	$D_{i,\mathrm{anc}}$
Ornithorynchus anatinus	Platypus	Orn	57.89	14.11	0.55	—
Monodelphis domesticus	Opossum	Mon	43.89	10.86	0.48	—
Choloepus hoffmanni	Sloth	Cho	46.57	10.36	0.27	148.24
Dasypus novemcinctus	Armadillo	Das	47.23	11.25	0.26	194.41
Echinops telfairi	Tenrec	Ech	53.55	11.11	0.35	339.65
Loxodonta africana	Elephant	Lox	47.66	9.26	0.24	133.56
Procavia capensis	Hyrax	Pro	48.96	9.50	0.31	240.08
Tupaia belangeri	Tree shrew	Tup	49.18	10.94	0.32	243.66
Homo sapiens	Human	Hom	46.1	9.70	0.10	95.64
Pan troglodytes	Chimp	Pan	46.09	9.69	0.12	97.28
Gorilla gorilla	Gorilla	Gor	46.07	9.71	0.14	96.22
Pongo pygmaeus	Orangutan	Pon	45.97	9.61	0.12	95.79
Macaca mulatta	Macaque	Mac	46	9.64	0.12	102.17
Tarsius syrichta	Tarsier	Tar	47.34	9.97	0.23	178.97
Microcebus murinus	Mouse lemur	Mic	47.91	11.17	0.14	169.33
Otolemur garnettii	Bushbaby	Oto	47.63	9.65	0.10	177.44
Oryctolagus cuniculus	Rabbit	Ory	51.87	12.00	0.29	308.63
Ochotona princeps	Pika	Och	52.52	10.18	0.24	326.49
Spermophilus tridecemlineatus	Squirrell	Spe	46.16	10.44	0.25	166.96
Cavia porcellus	Guinea pig	Cav	49.97	11.19	0.32	304.10
Dipodomys ordii	Kangaroo rat	Dip	48.04	10.60	0.19	244.08
Rattus norvegicus	Rat	Rat	51.46	7.50	0.14	307.35
Mus musculus	Mouse	Mus	51.24	7.80	0.11	299.93
Erinaceus europeaus	Hedgehog	Eri	48.06	11.15	0.37	289.97
Sorex araneus	Shrew	Sor	53.40	14.20	0.41	438.96
Bos taurus	Cow	Bos	49.94	10.88	0.22	228.33
Tursiops truncatus	Dolphin	Tur	48.82	10.59	0.18	169.75
Vicugna pacos	Alpaca	Vic	50.47	10.85	0.27	238.13
Myotis lucifugus	Microbat	Myo	51.70	12.85	0.30	345.27
Pteropus vampyrus	Megabat	Pte	47.65	10.76	0.31	212.26
Equus caballus	Horse	Equ	48.74	9.93	0.18	153.12
Canis familiaris	Dog	Can	47.67	10.08	0.15	157.55
Felis catus	Cat	Fel	48.62	10.13	0.23	174.89
Placental ancestor	—	—	46.16	10.65	—	—

[a] Species name abbreviations used in all figures.
[b] The squared correlation coefficient between GC3 and flanking GC content.

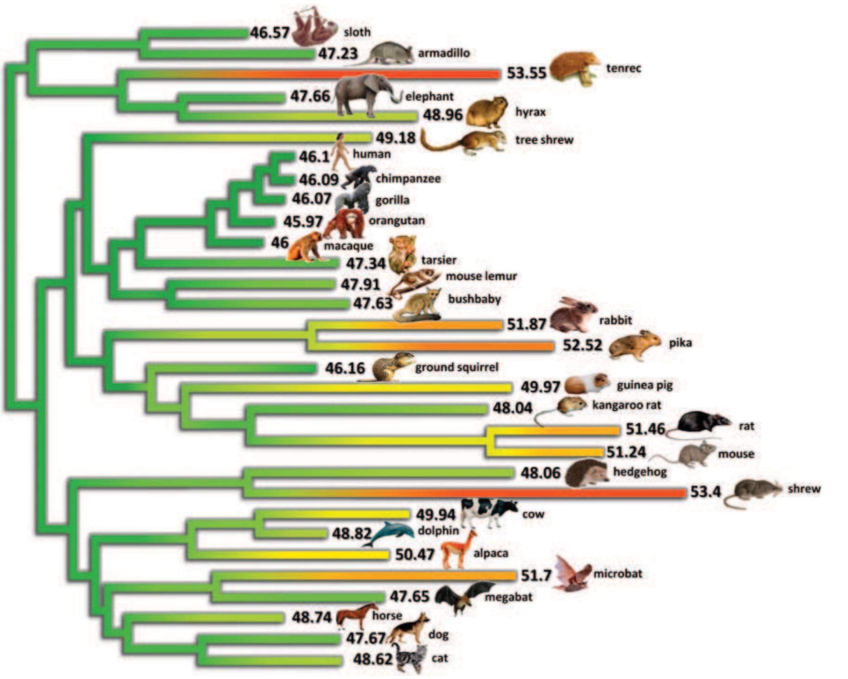


Figure 1. Genomic third-codon position GC-content (GC3) evolution in placental mammals. Colors reflect current or estimated average GC3 on 1138 orthologous genes (green, low GC; yellow, medium GC; red, high GC). current or estimated average GC3 on 1138 orthologous genes. Branch lengths quantify the amount of GC3 divergence: the branch connecting nodes *i* and *j* has a length proportional to $D_{i,j}$ (Equation 2). The estimated ancestral GC content of the placental ancestor is 46.2%.

placental mammals. The mean and standard error of the distribution varied substantially among species. The GC3 distribution in the ancestral placental appeared to be quite similar to that of humans, and much more heterogeneous than in mice, thus confirming previous study findings. Table 1, however, reveals that the newly sampled placental groups had distinctive patterns. The tenrec and shrew genomes, for instance, appeared to be substantially more GC-rich than the ancestral one, and more heterogeneous. The non-placental platypus (Monotremata) and opossum (Marsupialia) were the GC-richest (57.89%) and GC-poorest (43.89%) (on average), respectively, of all the analyzed genomes. The pattern diversity is illustrated in Figure 2, which displays the observed distribution of GC3 in three representative species (human, mouse, tenrec) compared to the estimated ancestral distribution.

It is noteworthy that the strength of the isochore structure, as measured by the standard deviation of GC3 across genes, was positively correlated with the species average GC3 (Fig. 3; $\rho = 0.42$, P-value < 0.05). With the exception of mouse and rat, which behaved very differently from other placentals in this respect, GC-rich species tended to show high variance in genomic GC3. Consistent with previous findings, in humans and apes there was a moderate decline in the standard deviation of GC3 across genes (9.7%), whereas this decline was substantial in mouse (7.8%) and rat (7.5%) as compared to the estimated ancestral value (10.6).

The various levels of GC3 divergence since the placental ancestor are represented in Figure 1. Colors in this figure reflect current or estimated average GC3, and branch lengths quantify the amount of GC3 divergence—the length of the branch connecting nodes *i* and *j* is proportional to $D_{i,j}$ (see Methods, Equation 2). Figure 1 shows the existence of lineages in which the transcriptome-wide GC3 had evolved slowly (e.g., apes, horse, armadillo) since the placental ancestor, and of lineages showing elevated amounts of GC3 divergence. Among the latter, muroid rodents (mouse, rat), as well as kangaroo rat, guinea pig, and hedgehog, showed a limited change in average GC3. In these species, gene-specific GC3 tended to diverge quickly, but the average was moderately affected, with the decrease in some genes somehow compensating for the increase in some others. Several species, finally, showed both a gene-by-gene ($D_{i,j}$) and collective increase in GC3—this pattern has yet to be documented. This was observed in shrew, microbat, tenrec, and lagomorphs, i.e., four phylogenetically distant lineages of placental mammals. Notably, no marked decrease in average GC3 was noted among the 31 placental species examined.

The contrasted dynamics across species are represented in Figure 4. In this figure, genes were divided into five bins according to their ancestral GC3. For each placental species *i* and each category of genes, the average GC3 was calculated and plotted against $D_{i,\text{anc}}$, i.e., the amount of gene GC3 divergence since the ancestor. The average ancestral GC3 of each bin is represented as a vertical dotted line. Figure 4 shows that an increase in GC3 was the most common evolutionary trend in placentals. This is especially true of ancestrally GC3-poorest genes, but also involved the GC3-medium and, in some species, GC3-rich categories. The previously documented erosion of GC-rich isochores was only detected from the 20% GC3-richest genes (rightmost graph), and this erosive process, which we confirmed in hominid primates

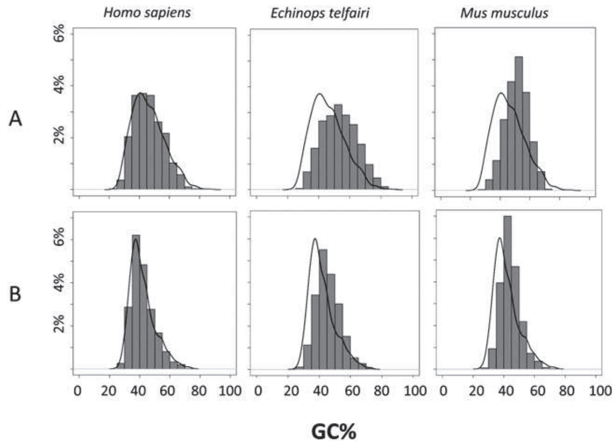


Figure 2. Gene GC-content distribution in three representative placental species. (*A*) Third-codon position GC content (GC3); (*B*) 5′- and 3′-flanking GC content. (Curved line) Estimated GC distribution of the common placental ancestor. (Gray histogram) Observed distributions of extant species. As compared to the estimated ancestral state, humans show a conservative pattern, tenrecs (*Echinops telfairi*) a global enrichment in GC, and mice (*Mus musculus*) a decreased variance in GC across genes.

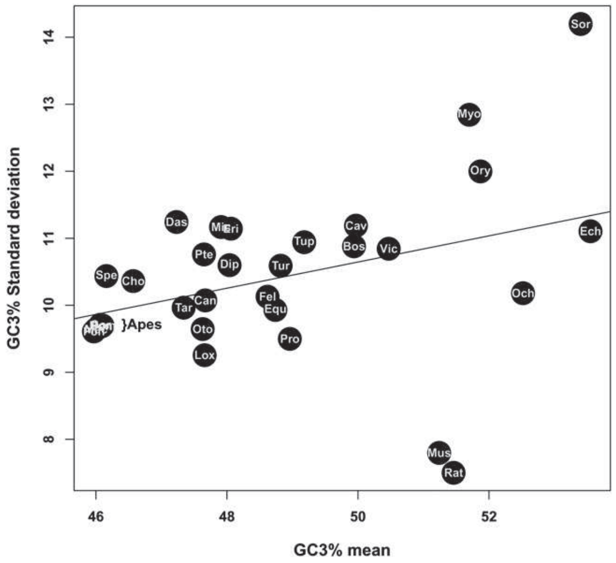


Figure 3. Relationship between genomic average and standard deviation of GC3.

and muroid rodents, was not a general trend. In several species (tenrec, shrew, microbat, rabbit), even the 20% ancestrally GC3-richest genes had undergone an increase in average GC3 since the placental ancestor. Figure 4 also revealed a highly specific pattern in mouse and rat: Their GC3-poorest genes were highly GC3-enriched, and their GC3-richest genes very GC3-depleted. This was reflected by the remarkably low standard deviation of GC3 distribution in these two species (Table 1).

It should be noted that the method used here to estimate ancestral GC3 did not account for the hypermutability of CpG doublets, which is known to significantly impact the nucleotide substitution process in mammals. To account for this potential bias, we replicated the analysis after removing from the data set the third-codon positions immediately flanked by a 5′ C or a 3′ G in >50% of the species. The remaining sites, representing 67.2% of the original data set, were presumably only weakly affected by the CpG effect. Despite a general decline in all GC3 values (by ~2%), highly similar trends were observed, i.e., a global GC3 enrichment since the last common placental ancestor. Similarly, Belle et al. (2004) showed, by simulations, that NHML estimates of ancestral GC content were only slightly affected by CpG hypermutability.

Causes of GC3 enrichment

We investigated evolutionary forces that could potentially account for this diversity of patterns by correlating the GC3 dynamics with several life-history and cytologic variables across the 33 species of the data set. Within placentals, we found a significant negative correlation between average GC3 and (log-transformed) body mass ($\rho = -0.44$, P-value = 0.013), and between $D_{i,\mathrm{anc}}$ and body mass ($\rho = -0.69$, P-value $< 10^{-4}$; Fig. 5). Similar trends were noted when we correlated GC3 and $D_{i,\mathrm{anc}}$ with species longevity ($\rho = -0.58$, P-value $< 10^{-4}$; $\rho = -0.73$, P-value $< 10^{-4}$, respectively), age of sexual maturity, and gestation time (data not shown). The latter trends reflect high positive correlations between body mass and these variables. Small-sized placentals tended to evolve faster with respect to GC3, and the main trend was toward an increase in average GC3 in fast-evolving genomes.

All of these correlations remained significant when phylogenetic inertia was regressed out of the analysis. Spearman correlation coefficients were −0.40 (P-value = 0.028), −0.52 (P-value = 0.0037), and −0.51 (P-value = 0.0067) for the GC3/body mass, $D_{i,\mathrm{anc}}$/body mass, and GC3/genome size relationships, respectively, after phylogenetic correction. It can be noted from Figure 1 that the body size effect was detectable even within orders and superorders. Within afrotherians, for instance, tenrec evolved faster and was GC3-richer than elephant, with hyrax being intermediate. A similar trend was found within primates ("monkeys" faster than apes), cetartiodactyls (bovids and camelids faster than cetaceans), Chiroptera (microbats faster than megabats), and Eulipotyphla (shrews faster than hedgehogs). The pattern was only less clearcut in rodents, but body mass was also less contrasted in this group.

Karyotypes, and especially chromosome length, have been connected to GC-content evolution in various vertebrate taxa (International Human Genome Sequencing Consortium 2001; International Chicken Genome Sequencing Consortium 2004; Kuraku et al. 2006), with chromosome length being inversely related to the recombination rate (Li and Freudenberg 2009), hence to gBGC. Chromosome length data, however, were only available for a limited number of taxa included in this study. We thus tried to indirectly approach the relationship by linking GC3 dynamics to the current number of chromosome arms in each species, under the assumption that more fragmented genomes should contain shorter chromosomes on average. We found that, consistent with the hypothesis of chromosome length-driven GC-content evolution, platypus ($2n = 54$, 104 chromosome arms) was the GC3-richest, and opossum ($2n = 18$, 20 chromosome arms) the GC3-poorest, on average, of the 33 analyzed mammals, as also previously documented (Goodstadt et al. 2007; Warren et al. 2008). Placentals showed an intermediate number of chromosome arms, and an

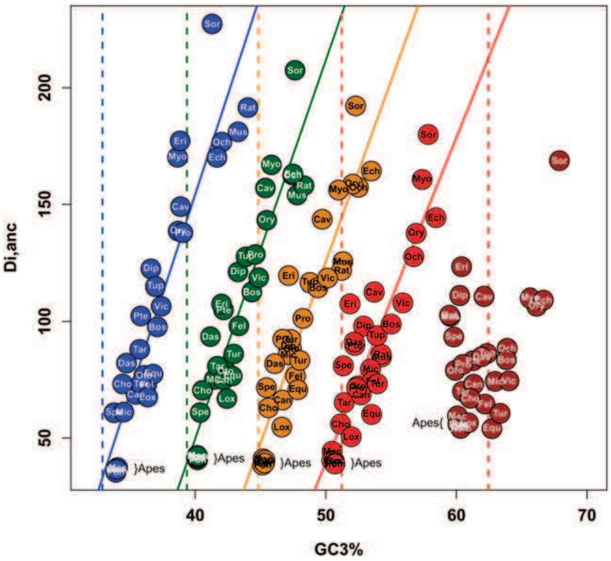


Figure 4. GC3 dynamics in GC3-rich versus GC3-poor genes. Genes are divided into five categories, depending on their ancestral GC3%. Blue, green, orange, red, and brown are used from the least GC-rich to the most GC-rich categories. Within each category, the average GC3 of each placental species is plotted against $D_{i,\mathrm{anc}}$ (i.e., the average amount of GC3 divergence among the 1138 genes since the placental ancestor). Dotted lines represent the average ancestral GC3 within each category.

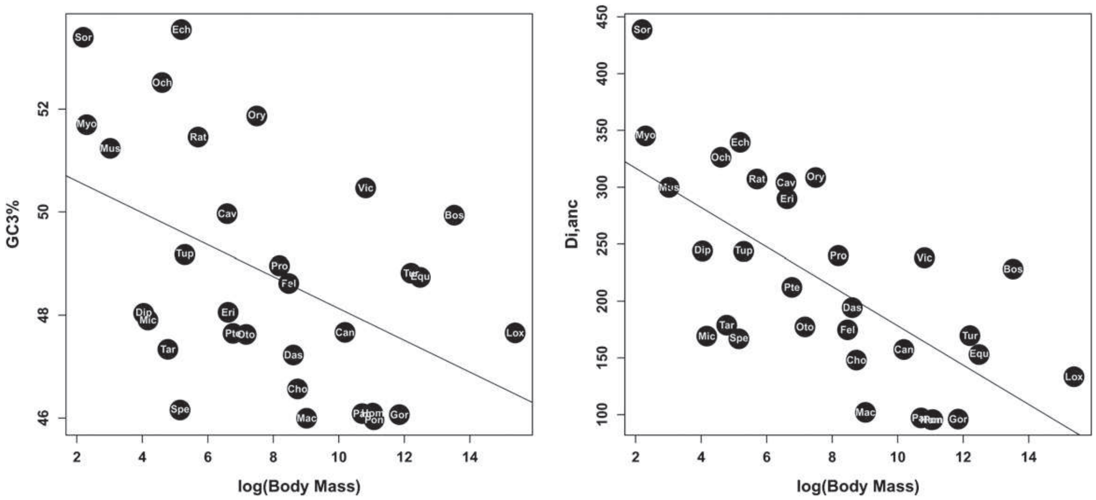


Figure 5. Relationship between adult body mass and GC3 (*left*) and $D_{i,\mathrm{anc}}$ (*right*).

intermediate average GC3. However, no significant relationship was detected between these two variables within placentals.

The above rationale is based on the implicit assumption of a constant genome size across species. Alternatively, one could suppose that small genomes tend to have short chromosomes, and large genomes long chromosomes, while assuming a relatively constant genome fragmentation level (i.e., diploid number) across lineages. We found a significant negative correlation between genome size (measured by the C-value) and average GC3 ($\rho = -0.48$, *P*-value = 0.01; Fig. 6), which was robust to phylogenetic control ($\rho = -0.51$, *P*-value < 0.01). The correlation remained marginally significant when we excluded platypus and opossum ($\rho = -0.37$, *P*-value = 0.06).

Flanking gene regions

To check whether the GC3 analyses reflected genome-wide patterns, we analyzed noncoding sequences flanking the 1138 genes of this study. We first calculated, for each species, the correlation, across genes, between GC3 and the GC content of 100-kb-long flanking regions (50 kb each side). For all species, highly significant correlations were found (*P*-value < 0.001) (see r^2 values in Table 1). Then ancestral GC-content analyses were performed using 1460 alignments (each longer than 300 bp) of noncoding flanking regions of our genes. The observed and estimated GC-content distributions across flanking regions confirmed the main trends of the GC3 analysis (Fig. 2B; Supplemental Table 1), the positive correlation between the average and standard deviation in GC content ($\rho = 0.42$, *P*-value < 0.05), and the negative correlation between GC content and species longevity (average GC content: $\rho = -0.46$, *P*-value = 0.01; $D_{i,\mathrm{anc}}$: $\rho = -0.8$, *P*-value < 0.0001) and genome size ($\rho = -0.46$, *P*-value = 0.016).

Discussion

In this study, we analyzed GC-content evolution in 1138 genes across 33 mammalian species, reconstructed the ancestral distribution of GC3, and characterized isochore evolution in distinct lineages. In line with previous reports, we found that the estimated ancestral placental GC3 distribution was close to the current one in humans. Apes have evolved relatively slowly since the common placental ancestor as far as GC3 is concerned, and this trend also applies to other large-sized species (e.g., elephant, xenarthrans, horse, cow, dolphin). GC3 evolved more rapidly in smaller mammals (e.g., rabbit, tenrec, rodents, shrew, microbat), and the main trend was an increased average GC3, and more structured isochores, in these fast-evolving lineages.

Genomic and taxonomic sampling

The so-called erosion of GC-rich isochores (GC3-decrease of GC3-rich genes) was documented as a major process in previous studies of GC-content dynamics in mammals (Duret et al. 2002; Arndt et al. 2003; Belle et al. 2004). Our findings, in contrast, suggested that the erosion of GC-rich isochores is not a general process: It apparently affected a limited set of placental mammals, and only

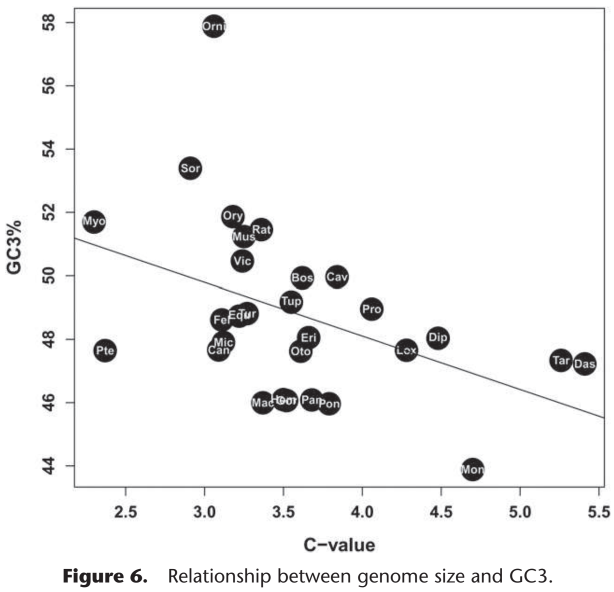


Figure 6. Relationship between genome size and GC3.

the top 20% GC-richest genes of the genome. Our larger sample size, both at taxonomic (33 mammals) and genomic (more than 1000 genes) levels, could likely explain the difference between our findings and those of previous studies. Previous works typically analyzed a small number of taxa and focused on model organisms such as humans and mice, which turned out to undergo detectable erosion (Fig. 4). Belle et al. (2004), furthermore, used a set of 41 genes available at that time, which actually appeared to be biased toward GC-rich genes (mean GC3: 62.96), as compared to our 1138-gene data set (mean GC3: 48.08). This also applies to the data set of Gu and Li (2006), which was similarly enriched in GC-rich genes (178 genes, mean GC3: 62.95). This explains why the detected erosive trend appeared more prevalent to Belle et al. (2004) and Gu and Li (2006) than it did here—although we confirmed it in apes and murids. This work, therefore, emphasizes the importance of large numbers of species and genes in evolutionary genomics.

Along the same lines, we noted that the pattern observed in mouse and rat, often taken as reference genomes in comparative studies, appeared to be highly unusual among placental mammals. These species were the only ones in this data set for which decreased variance of GC3 across genes was observed despite the fast evolution rate. Ground squirrel (*Spermophilus tridecemlineatus*) also stands as an exception: Despite its small size (body mass: 0.17 kg), it behaved as a large organism with respect to GC3. It should be noted, however, that Sciuridae, and especially Marmotini, include relatively large species of rodents, like prairie dogs (1 kg) and marmots (up to 7 kg), which are close relatives to *S. tridecemlineatus* (Herron et al. 2004). This suggests that the terminal branch leading to ground squirrels in Figure 1 actually corresponds to medium-sized ancestors, thus perhaps explaining the slow evolutionary rate in this lineage.

Karyotypic correlates

Thanks to the substantial number of genomes analyzed in this study, for the first time we were able to correlate GC3 evolution with various species characteristics. We report a negative correlation between GC3 (or $D_{i,\mathrm{anc}}$) and body mass (or longevity, or age of sexual maturity), and a negative correlation between GC3 and genome size. We noted that GC3 and $D_{i,\mathrm{anc}}$ reflected long-term evolutionary processes, and had been influenced by the (unknown) properties of ancestral species, when we can only observe extant species. We conclude that the true biological relationships between GC-content evolution and species biology/cytology are probably even closer than measured here.

The reported GC3 increase in small genomes and the specific patterns found in marsupials and monotremes are consistent with the relationship between chromosome length and GC content observed in humans and chicken, and predicted by the gBGC model. In every meiosis, each chromosome arm undergoes at least one (and not many more than one) crossover, which is probably required for proper meiotic segregation. This results in a higher per-megabase recombination rate in short than in long chromosomes, with gBGC thus having a greater impact on short chromosomes (e.g., Montoya-Burgos et al. 2003). Observing a relationship between GC3 and genome size (our best estimate of long-term average chromosome length) across 33 mammals strongly suggests that gBGC is a general process in this group, not a unique feature of hominid primates and murine rodents. Greater insight into this process will be gained by correlating GC-content dynamics to karyotypic evolution in placental mammals—chromosomal maps are currently available for just a handful of fully assembled mammalian genomes.

The combination of chromosome arm number and genome size analyses suggested that chromosomes were rearranged at a relatively fast rate in placental mammals, consistent with the high amount of karyotypic variation observed among species. The diploid number, for instance, ranged from 22 to 80 across species sampled in this study, and was as low as six in muntjak (*Muntiacus muntjak*, Cervidae, Cetartiodactyla), and as high as 88 in the western pocket gopher (*Thomomys bottae*, Geomyidae, Rodentia). The diploid number could evolve very rapidly in placental mammals, as illustrated by the observation of very different numbers among species of the same genus, for example, $2n = 18$ in *Microtus oregoni* versus $2n = 62$ in *Microtus duodecimcostatus*, i.e., two vole species that diverged less than 5 million yr ago (Galewski et al. 2006). The current chromosome number is therefore probably not a reliable estimate of the long-term average genome fragmentation level. Marsupials and monotremes, in contrast, appear to have a relatively stable karyotypic structure (Wrigley and Graves 1988). The diploid number ranges from 12 to 22 across the 39 marsupial species available in the Genome Size Database, and is high (54 in platypus, 64 in echidna) in the two available monotremes. Karyotypic stability might explain why these two taxa fit predictions regarding GC content and chromosome number better than placentals, since it takes a long time for GC content to reach equilibrium following a change in chromosome size. In line with this hypothesis, the chromosome length/GC content relationship is especially close in chicken (International Chicken Genome Sequencing Consortium 2004), and a very high level of karyotypic conservation has been documented in birds (Griffin et al. 2007).

Another explanation for the relationship between genome size and GC content could be a potential effect of gBGC on genome size. Indeed, GC-rich isochores are characterized by a high gene density, and it was noted that the marked increase in GC content in the mouse *Fxy* gene after its translocation into the high-recombining pseudoautosomal region was accompanied by large deletions within its introns (Montoya-Burgos et al. 2003). GC-rich sequences could be generally prone to deletions, for example, because sequences of extreme base composition might trigger replication slippage. Under this (still speculative) hypothesis, gBGC could indirectly promote the deletion of noncoding DNA sequences.

Life-history correlates

The relationship between body mass and GC3 evolution is not as easy to interpret, since body mass may be correlated with a number of potentially relevant variables. First, body mass could affect molecular evolution through generation time. Small mammals tend to have more generations per time unit than larger ones, resulting in a higher per unit time mutation rate, and therefore a higher propensity for GC3 to diverge during evolution. This effect probably largely explains the $D_{i,\mathrm{anc}}$/body mass relationship—slow-evolving genomes cannot diverge faster than fast-evolving ones—but the mutation rate effect does not explain the trend of an increased average GC3 in small species. Generation time, however, also affects the per-year number of meioses in germline evolution. This could result in a higher per-year recombination rate, hence more effective gBGC, and an increased equilibrium GC content in short-lived species.

Body mass, finally, could affect molecular evolution through its relationship with population size. Analyses of nonsynonymous (d_N) versus synonymous (d_S) substitution rates revealed a higher

d_N/d_S ratio in large mammals (Nikolaev et al. 2007; Popadin et al. 2007), which was interpreted in terms of effective population size. Natural selection was found to be less efficient in small populations (large animals), in which a number of slightly deleterious nonsynonymous mutations could reach fixation through increased genetic drift. These studies suggested that body mass is a good indicator of the long-term effective population size in mammals. Just like directional selection, gBGC is supposed to be more efficient in large populations. What matters is the product of the effective population size by the recombination rate by the repair bias (Duret and Galtier 2009). So a higher equilibrium GC content and a faster increase in GC content (in nonequilibrium conditions) would be expected in large populations under the gBGC model (Duret and Arndt 2008). It should be noted that the two potential effects of body mass, through generation time and population size, are not mutually exclusive.

GC3 as an isochore marker

In this study, as in many previous ones, GC3 was taken as a proxy for genomic GC content, a strategy that was recently criticized by Elhaik et al. (2009). Analyzing the noncoding, flanking regions of the 1138 genes of this study, we found a significant correlation between GC3 and flanking GC content across genes, and similar evolutionary dynamics for the two data sets. Elhaik et al. (2009) compared GC3 with the GC content of a noncoding window gradually moving away from the focal gene. We suggest that the rapid decline in correlation coefficient they reported reflects the heterogeneous nature (especially) of GC-rich regions at the 5-kb scale (International Human Genome Sequencing Consortium 2001). Our analysis suggested that GC3 is a reasonable marker of local genomic GC content and one that overcomes the problem of whole genome alignment between distantly related species. Interestingly, the GC3/flanking-GC correlation coefficient was especially elevated in species for which a marked increase in GC3 was found (e.g., shrew, bat, tenrec), suggesting that the newly reported GC increase in these species affects the whole genome, not just third-codon positions.

Conclusions

The findings of this analysis modified our view of GC-content dynamics in placental mammals. We showed that the erosion of GC-rich isochores is not a general trend, and that several species, especially small-sized ones, have undergone a substantial increase in gene GC3 over the last 100 million yr. Our results are consistent with the hypothesis of chromosome-length-driven GC-content evolution, in agreement with the gBGC model. We noted, finally, that gBGC was apparently very strong in a number of non-model taxa, for example, Soricidae, Lagomorpha, Chiroptera, and Afrosoricida, which appear to be more suitable than the highly studied murids and hominids for the analysis of the gBGC process, and more generally of molecular evolutionary processes in mammals.

Methods

Sequences, alignments, trees

The 1138 orthologous genes were extracted from the OrthoMam database (release v5) (Ranwez et al. 2007) and corresponded to all CDS available for the 33 mammalian species documented by Ensembl v54. Sequence alignments provided by the database were cleaned with Gblocks (Castresana 2000) to exclude the least conserved regions and select third-codon positions. The well-accepted phylogenetic trees of Nishihara et al. (2006) and Prasad et al. (2008) were used to perform ancestral GC3 estimations (Fig. 1).

For each of the 1138 genes in the study, aligned 5′ and 3′ noncoding flanking regions (5000 bp each with reference to humans) were downloaded from the EPO section of Ensembl v54. These alignments included up to 31 species, but platypus and opossum were not available. Alignments were cleaned with Gblocks (Castresana 2000). The 1460 alignments, which were longer than 300 bp after this step, were kept for subsequent analysis (730 and 730 alignments on the 5′ and 3′ sides, respectively). Sequences including more than 50% missing nucleotides were removed from the alignments.

Estimating ancestral GC3

Alignments of third-codon positions were separately analyzed using the method introduced by Galtier and Gouy (1998), implemented in the NHML and bpp_ML programs (Dutheil and Boussau 2008). This method relies on a nonhomogeneous, nonstationary Markov model of nucleotide substitution to obtain an estimate of ancestral GC content at each internal node of the underlying phylogenetic tree in the maximum likelihood framework. Under this model, each branch of the underlying tree has its own specific equilibrium GC content. Gamma-distributed rates across sites were assumed. The method and program have been validated in a number of biological applications (e.g., Galtier and Mouchiroud 1998; Galtier et al. 1999; Rodríguez-Trelles et al. 2000; Belle et al. 2004; Herbeck et al. 2005; Boussau and Gouy 2006). We focused on the placental ancestral node, with platypus and opossum being used as outgroups, so our analysis was unaffected by the high uncertainty in the GC-content estimation at the mammalian root node (Galtier and Gouy 1998). Gaps and lacking exons were treated as missing data.

Measuring average gene GC3 divergence

For any extant species i, the average amount of GC3 divergence among n genes since the placental ancestor is denoted by $D_{i,\mathrm{anc}}$ and measured as:

$$D_{i,anc} = \sqrt{\sum_{k=1}^{n} (GC3_k^i - GC3_k^{anc})^2}, \qquad (1)$$

where $GC3_k^i$ is the GC3 observed for gene k, species i, while $GC3_k^{anc}$ is the estimated ancestral GC3 for gene k. $D_{i,\mathrm{anc}}$ is the Euclidean distance between the two n-dimension vectors of GC3. Similarly, GC3 divergence between any two nodes i and j of the tree was defined as:

$$D_{i,j} = \sqrt{\sum_{k=1}^{n} (GC3_k^i - GC3_k^j)^2}. \qquad (2)$$

Quantitative variables

For each of the 33 species included in the study, karyotypic information was collected from the *Atlas of Mammalian Chromosomes* (O'Brien et al. 2006) and from the appendix of the study by Pardo-Manuel de Villena and Sapienza (2001). Genome size values (C-value) for 29 of our 33 species were taken from the Animal Genome Size Database (Gregory et al. 2007). When more than one C-value was available for a given species, we took the most recent estimate. The C-value of *Myotis lucifugus* (microbat) is not documented. It was estimated by the mean C-value of 13 species from the genus *Myotis*. Data on life history traits (body mass, longevity, sexual maturity) were taken from the AnAge database (de Magalhaes and Costa 2009).

Correlation analyses

Nonparametric Spearman correlation tests between GC3%, life-history traits, and genome size values were performed using R. For each of these tests, a phylogenetic control was performed with the method of independent contrasts (PIC) implemented in the PHYLIP software package (Felsenstein 1995). Parametric Pearson correlation tests between GC3% of the 1138 genes and the GC content of 100-kb-long noncoding flanking regions (50 kb each side) were also performed with R.

Acknowledgments

This work was supported by the French Agence Nationale de la Recherche "Génomique Animale" (ANR-08-GENM-036-01 "CoGeBi"), European Research Council ("PopPhyl"), and by the computer grid of Université Montpellier 2. We thank Kerstin Lindblad-Toh, the Broad Institute, Washington University, and Baylor College of Medicine for giving us permission to analyze unpublished low-coverage mammalian genomes. This publication is contribution number 2010-041 of the Institut des Sciences de l'Evolution de Montpellier (UMR 5554–CNRS).

References

Alvarez-Valin F, Clay O, Cruveiller S, Bernardi G. 2004. Inaccurate reconstruction of ancestral GC levels creates a "vanishing isochores" effect. *Mol Phylogenet Evol* **31:** 788–793.

Arndt PF, Petrov DA, Hwa T. 2003. Distinct changes of genomic biases in nucleotide substitution at the time of mammalian radiation. *Mol Biol Evol* **20:** 1887–1896.

Belle EM, Smith N, Eyre-Walker A. 2002. Analysis of the phylogenetic distribution of isochores in vertebrates and a test of the thermal stability hypothesis. *J Mol Evol* **55:** 356–363.

Belle E, Duret L, Galtier N, Eyre-Walker A. 2004. The decline of isochores in mammals: An assessment of the GC content variation along the mammalian phylogeny. *J Mol Evol* **58:** 653–660.

Bernardi G, Bernardi G. 1990. Compositional transitions in the nuclear genomes of cold-blooded vertebrates. *J Mol Evol* **31:** 282–293.

Bernardi G. 2007. The neoselectionist theory of genome evolution. *Proc Natl Acad Sci* **104:** 8385–8390.

Bernardi G, Olofsson B, Filipski J, Zerial M, Salinas J. 1985. The mosaic genome of warm-blooded vertebrates. *Science* **228:** 953–958.

Boussau B, Gouy M. 2006. Efficient likelihood computations with nonreversible models of evolution. *Syst Biol* **55:** 756–768.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* **17:** 540.

De Magalhaes JP, Costa J. 2009. A database of vertebrate longevity records and their relation to other life-history traits. *J Evol Biol* **22:** 1770–1774.

Duret L, Arndt PF. 2008. The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genet* **4:** e1000071. doi: 10.1371/journal.pgen.1000071.

Duret L, Galtier N. 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu Rev Genomics Hum Genet* **10:** 285–311.

Duret L, Semon M, Piganeau G, Mouchiroud D, Galtier N. 2002. Vanishing GC-rich isochores in mammalian genomes. *Genetics* **162:** 1837–1847.

Dutheil J, Boussau B. 2008. Non-homogeneous models of sequence evolution in the Bio++ suite of libraries and programs. *BMC Evol Biol* **8:** 255. doi: 10.1186/1471-2148-8-255.

Elhaik E, Landan G, Graur D. 2009. Can GC content at third-codon positions be used as a proxy for isochore composition? *Mol Biol Evol* **26:** 1829–1833.

Eyre-Walker A. 1993. Recombination and mammalian genome evolution. *Proc Biol Sci* **252:** 237–243.

Eyre-Walker A. 1999. Evidence of selection on silent site base composition in mammals: Potential implications for the evolution of isochores and junk DNA. *Genetics* **152:** 675–683.

Eyre-Walker A, Hurst LD. 2001. The evolution of isochores. *Nat Rev Genet* **2:** 549–555.

Felsenstein J. 1995. *PHYLIP (phylogeny inference package), version 3.57 c*. Department of Genetics, University of Washington, Seattle, WA.

Galewski T, Tilak MK, Sanchez S, Chevret P, Paradis E, Douzery EJP. 2006. The evolutionary radiation of Arvicolinae rodents (voles and lemmings): Relative contribution of nuclear and mitochondrial DNA phylogenies. *BMC Evol Biol* **6:** 80. doi: 10.1186/1471-2148-6-80.

Galtier N. 2003. Gene conversion drives GC content evolution in mammalian histones. *Trends Genet* **19:** 65–68.

Galtier N, Gouy M. 1998. Inferring pattern and process: Maximum likelihood implementation of non-homogeneous model of DNA sequence evolution for phylogenetic analysis. *Mol Biol Evol* **15:** 871–879.

Galtier N, Mouchiroud D. 1998. Evolution of isochores in mammals: A human-like ancestral pattern. *Genetics* **150:** 1577–1584.

Galtier N, Tourasse NJ, Gouy M. 1999. A nonhyperthermophilic common ancestor to extant life forms. *Science* **283:** 220–221.

Galtier N, Piganeau G, Mouchiroud D, Duret L. 2001. GC-content evolution in mammalian genomes: The biased gene conversion hypothesis. *Genetics* **159:** 907–911.

Goodstadt L, Heger A, Webber C, Ponting CP. 2007. An analysis of the gene complement of a marsupial, *Monodelphis domestica*: Evolution of lineage-specific genes and giant chromosomes. *Genome Res* **17:** 969–981.

Gregory TR, Nicol JA, Tamm H, Kullman B, Kullman K, Leitch IJ, Murray BG, Kapraun DF, Greilhuber J, Bennett MD. 2007. Eukaryotic genome size databases. *Nucleic Acids Res* **35:** D332–D338.

Griffin DK, Robertson LB, Tempest HG, Skinner BM. 2007. The evolution of the avian genome as revealed by comparative molecular cytogenetics. *Cytogenet Genome Res* **117:** 64–77.

Gu J, Li WH. 2006. Are GC-rich isochores vanishing in mammals? *Gene* **385:** 50–56.

Herbeck JT, Degnan PH, Wernegreen JJ. 2005. Nonhomogeneous model of sequence evolution indicates independent origins of primary endosymbionts within the enterobacteriales (gamma-Proteobacteria). *Mol Biol Evol* **22:** 520–532.

Herron MD, Castoe TA, Parkinson CL. 2004. Sciurid phylogeny and the paraphyly of Holarctic ground squirrels (Spermophilus). *Mol Phylogenet Evol* **31:** 1015–1030.

Hughes S, Zelus D, Mouchiroud D. 1999. Warm-blooded isochore structure in Nile crocodile and turtle. *Mol Biol Evol* **16:** 1521–1527.

International Chicken Genome Sequencing Consortium. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432:** 695–716.

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.

Kudla G, Helwak A, Lipinski L. 2004. Gene conversion and GC-content evolution in mammalian Hsp70. *Mol Biol Evol* **21:** 1438–1444.

Kudla G, Lipinski L, Caffin F, Helwak A, Zylicz M. 2006. High guanine and cytosine content increases mRNA levels in mammalian cells. *PLoS Biol* **4:** e180. doi: 10.1371/journal.pbio.0040180.

Kuraku S, Ishijima J, Nishida-Umehara C, Agata K, Kuratani S, Matsuda Y. 2006. cDNA-based gene mapping and GC3 profiling in the soft-shelled turtle suggest a chromosomal size-dependent GC bias shared by sauropsids. *Chromosome Res* **14:** 187–202.

Li W, Freudenberg J. 2009. Two-parameter characterization of chromosome-scale recombination rate. *Genome Res* **19:** 2300–2307.

Li MK, Gu L, Chen SS, Dai JQ, Tao SH. 2008. Evolution of the isochore structure in the scale of chromosome: insight from the mutation bias and fixation bias. *J Evol Biol* **21:** 173–182.

Meunier J, Duret L. 2004. Recombination drives the evolution of GC-content in the human genome. *Mol Biol Evol* **21:** 984–990.

Montoya-Burgos JI, Boursot P, Galtier N. 2003. Recombination explains isochores in mammalian genomes. *Trends Genet* **19:** 128–130.

Mouchiroud D, Gautier C, Bernardi G. 1988. The compositional distribution of coding sequences and DNA molecules in humans and murids. *J Mol Evol* **27:** 311–320.

Nikolaev SI, Montoya-Burgos JI, Popadin K, Parand L, Margulies EH. 2007. Life-history traits drive the evolutionary rates of mammalian coding and noncoding genomic elements. *Proc Natl Acad Sci* **104:** 20443–20448.

Nishihara H, Hasegawa M, Okada N. 2006. Pegasoferae, an unexpected mammalian clade revealed by tracking ancient retroposon insertions. *Proc Natl Acad Sci* **103:** 9929–9934.

O'Brien SJ, Menninger JC, Nash WG. 2006. *Atlas of mammalian chromosomes*, 1st ed. Wiley-Liss, New York.

Pardo-Manuel de Villena F, Sapienza C. 2001. Female meiosis drives karyotypic evolution in mammals. *Genetics* **160:** 1263–1264.

Popadin K, Polishchuk LV, Mamirova L, Knorre D, Gunbin K. 2007. Accumulation of slightly deleterious mutations in mitochondrial protein-coding genes of large versus small mammals. *Proc Natl Acad Sci* **104:** 13390–13395.

Prasad AB, Allard MW, Green ED. 2008. Confirming the phylogeny of mammals by use of large comparative sequence data sets. *Mol Biol Evol* **25:** 1795–1808.

Ranwez V, Delsuc F, Ranwez S, Belkhir K, Tilak MK, Douzery EJ. 2007. OrthoMaM: A database of orthologous genomic markers for placental mammal phylogenetics. *BMC Evol Biol* **7:** 241. doi: 10.1186/1471-2148-7-241.

Ream RA, Johns GC, Somero GN. 2003. Base compositions of genes encoding alpha-actin and lactate dehydrogenase-A from differently adapted vertebrates show no temperature-adaptive variation in G + C content. *Mol Biol Evol* **20:** 105–110.

Rodríguez-Trelles F, Tarrío R, Ayala FJ. 2000. Evidence for a high ancestral GC content in *Drosophila*. *Mol Biol Evol* **17:** 1710–1717.

Schmegner C, Hameister H, Vogel W, Assum G. 2007. Isochores and replication time zones: A perfect match. *Cytogenet Genome Res* **116:** 167–172.

Spencer CC. 2006. Human polymorphism around recombination hotspots. *Biochem Soc Trans* **34:** 535–536.

Warren WC, Hillier LW, Marshall Graves JA, Birney E, Ponting CP. 2008. Genome analysis of the platypus reveals unique signatures of evolution. *Nature* **453:** 175–183.

Watanabe Y, Fujiyama A, Ichiba Y, Hattori M, Yada T, Sakaki Y, Ikemura T. 2002. Chromosome-wide assessment of replication timing for human chromosomes 11q and 21q: Disease-related genes in timing-switch regions. *Hum Mol Genet* **11:** 13–21.

Webster MT, Smith NG. 2004. Fixation biases affecting human SNPs. *Trends Genet* **20:** 122–126.

Webster MT, Smith NG, Hultin-Rosenberg L, Arndt PF, Ellegren H. 2005. Male-driven biased gene conversion governs the evolution of base composition in human Alu repeats. *Mol Biol Evol* **22:** 1468–1474.

Webster MT, Axelsson E, Ellegren H. 2006. Strong regional biases in nucleotide substitution in the chicken genome. *Mol Biol Evol* **23:** 1203–1216.

Wolfe KH, Sharp PM, Li WH. 1989. Mutation rates differ among regions of the mammalian genome. *Nature* **337:** 283–285.

Wrigley JM, Graves JA. 1988. Karyotypic conservation in the mammalian order Monotremata (subclass Prototheria). *Chromosoma* **96:** 231–247.

Received December 18, 2009; accepted in revised form May 19, 2010.

Chapitre 4

Genomic evidence for large, long-lived ancestors to placental mammals

Genomic Evidence for Large, Long-Lived Ancestors to Placental Mammals

J. Romiguier,[1] V. Ranwez,[1,2] E.J.P. Douzery,[1] and N. Galtier*,[1]

[1]CNRS, Université Montpellier 2, UMR 5554, ISEM, Montpellier, France

[2]Montpellier SupAgro, UMR 1334, AGAP, Montpellier, France

***Corresponding author:** E-mail: nicolas.galtier@univ-montp2.fr.

Associate editor: Naruya Saitou

Abstract

It is widely assumed that our mammalian ancestors, which lived in the Cretaceous era, were tiny animals that survived massive asteroid impacts in shelters and evolved into modern forms after dinosaurs went extinct, 65 Ma. The small size of most Mesozoic mammalian fossils essentially supports this view. Paleontology, however, is not conclusive regarding the ancestry of extant mammals, because Cretaceous and Paleocene fossils are not easily linked to modern lineages. Here, we use full-genome data to estimate the longevity and body mass of early placental mammals. Analyzing 36 fully sequenced mammalian genomes, we reconstruct two aspects of the ancestral genome dynamics, namely GC-content evolution and nonsynonymous over synonymous rate ratio. Linking these molecular evolutionary processes to life-history traits in modern species, we estimate that early placental mammals had a life span above 25 years and a body mass above 1 kg. This is similar to current primates, cetartiodactyls, or carnivores, but markedly different from mice or shrews, challenging the dominant view about mammalian origin and evolution. Our results imply that long-lived mammals existed in the Cretaceous era and were the most successful in evolution, opening new perspectives about the conditions for survival to the Cretaceous–Tertiary crisis.

Key words: phylogeny, GC-content, d*N*/d*S* ratio, GC-biased gene conversion, placentalia, fossils.

Introduction

It is commonly assumed that early mammals were small creatures that only evolved into a variety of forms and sizes after the massive extinction of large reptiles, at the Cretaceous/Tertiary (KT) boundary, 65 Ma (Dawkins 2004; Feldhamer et al. 2007). This scenario is consistent with theoretical considerations: Cope's rule (Alroy 1998) states that current living lineages generally descend from small ancestors, because large forms have a short-term advantage but tend to be more prone to extinction in the long run. The hypothesis of a small ancestral size is also largely supported by the fossil record: most of the Cretaceous mammals are smaller than a few inches, whereas post-KT deposits include numerous large mammals (Luo 2007; Smith et al. 2010).

Paleontology, however, is not conclusive regarding the ancestry of extant mammals due to the difficulty of linking Cretaceous and Paleocene fossils to modern lineages (Archibald et al. 2001; Asher et al. 2005). Genomic data provide an attractive opportunity to characterize the ancestral features of extant species: genome dynamics is imprinted by species life-history traits (Nikolaev et al. 2007; Nabholz et al. 2008; Romiguier et al. 2010), and ancestral genome characters can be reconstructed by phylogenetic methods (Galtier et al. 1999; Blanchette et al. 2004; Boussau et al. 2009; Lartillot and Poujol 2011). If molecular evolution can be traced back to the last common ancestor of extant placentals, then we could potentially learn about its macroscopic characteristics, even though this ancestor is not physically observable because missing from the fossil record.

In mammals, two genomic variables are known to correlate with species life-history traits. First, species longevity and body mass influence the ratio of nonsynonymous (= amino acid changing, d*N*) to synonymous (d*S*) nucleotide substitution rates. It has been shown that large and long-lived species display a higher d*N*/d*S* ratio, on average, than small and short-lived ones, presumably because of the smaller average population sizes, and hence the less effective purifying selection, in long-lived animals (Nikolaev et al. 2007). Second, large species tend to show a lower GC3 (percentage of G and C at the third position of codons) than small species. This effect is supposed to be caused by GC-biased gene conversion (Duret and Galtier 2009), a mechanism by which a biased DNA-repair process favors G and C alleles during meiotic recombination. Because short-lived species experience a higher rate of meiosis per time unit, their genome shows a faster divergence in gene GC3 and an increase in average GC3 (Romiguier et al. 2010).

Here, we quantify the influence of species longevity on gene coding sequence evolutionary dynamics (d*N*/d*S* ratio and GC3 divergence) in modern placental mammals. Then we reconstruct ancestral gene sequences using nonhomogeneous phylogenetic models and estimate the d*N*/d*S* and GC3 dynamics in the deepest branches of the mammalian tree. On the basis of the existing correlation between genomic processes and traits, we estimate the maximal life span of early placental mammals. Our analysis suggests that the ancestors of living placentals had a longevity and body mass similar to current primates, cetartiodactyls, or carnivores but differed markedly from mice or shrews, contradicting the prevailing view about mammalian origins and evolution.

Materials and Methods

Phylogeny and Divergence Dates

The phylogeny used in this study (fig. 1) was adapted from Meredith et al. (2011). Divergence dates were obtained from the TimeTree of Life database (http://www.timetree.org), which summarizes current paleontological and molecular knowledge. In Rodentia, the divergence dates proposed by this database are inconsistent with the tree topology assumed here. We used the following dates, consistent with recent literature on the subject (Springer et al. 2003; Huchon et al. 2007): *Mus*/*Rattus*: 16.4 My; *Mus*/*Dipodomys*: 70 My; *Mus*/*Cavia*: 73 My; and *Mus*/*Spermophilus*: 74 My.

Orthologous Genes and Alignments

Aligned orthologous gene coding sequences were obtained from the ORTHOMAM database (Ranwez et al. 2007) version 6, which is based on ENSEMBL orthology annotations. Alignments were cleaned using the Gblocks program with default parameters (Castresana 2000).

Life-History Traits

The maximum life span and body mass of placental mammals were retrieved from the AnAge database (de Magalhaes and Costa 2009), build 11. For each of the 33 placental species in the genomic data set, longevity was estimated by taking the mean longevity across all species of its family, for example, dog longevity was calculated as the mean of the longevities of all documented Canidae species. The family average was considered here as an estimate of the long-term average, thus avoiding potential problems due to recent changes in longevity (e.g., in domesticated species). Human was excluded when calculating the Hominidae average because the maximal record in human (120 years) is irrelevant from an evolutionary viewpoint.

Ancestral GC3 Estimation

For each of the 787 genes shared by all 33 species and the outgroups *Ornithorhynchus*, *Macropus*, and *Monodelphis*, we estimated ancestral GC3 at each node of the placental tree using Galtier and Gouy's nonhomogeneous model of sequence evolution (Galtier and Gouy 1998), as implemented in the Bio++ library (Dutheil et al. 2006; Dutheil and Boussau 2008), with branch lengths being separately estimated for each gene. This model is such that the GC content can

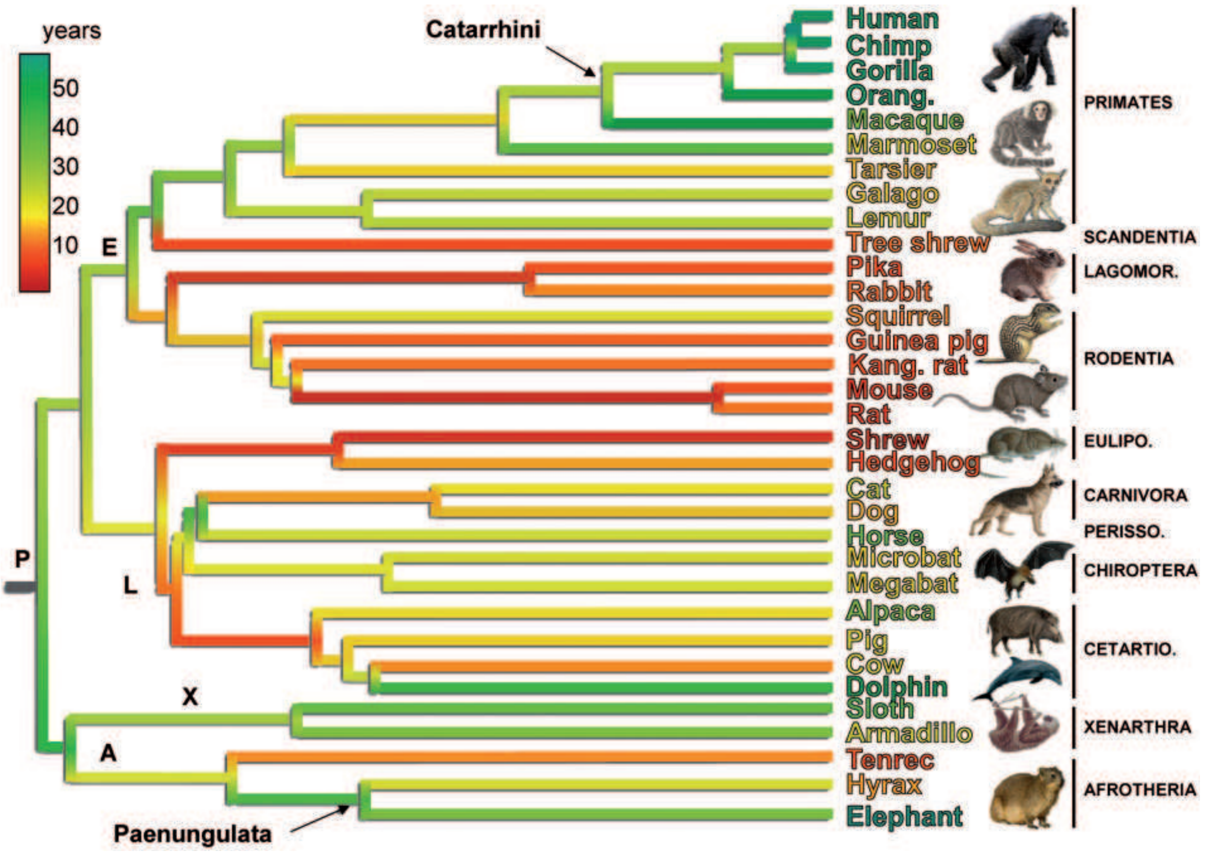


Fig. 1. Predicted evolution of longevity across the placental phylogeny. The maximal longevity of current species (family average) is indicated by the color of species names (red, short lived; green, long lived, see the top-left scale bar). Time-independent estimates of ancestral longevity are similarly indicated by the color of branches. P, most recent common ancestor (MRCA) of Placentalia; E, Euarchontoglires MRCA; L, Laurasiatheria MRCA; X, Xenarthra MRCA; A, Afrotheria MRCA.

fluctuate in time and across lineages, with each branch of the tree being assigned its own process, and its own equilibrium GC content. Outgroups are necessary for reliable estimation of GC3, especially, at the most basal nodes of the tree.

GC3 Plots

For each pair of species, plots of gene GC3 in species 1 versus species 2 were drawn using all orthologous genes available for the considered pair. This number varied across species pairs, from 3,682 (*Choloepus*/*Ornithorhynchus*) to 11,526 (Human/Chimpanzee). Kendall's coefficient was used to measure the level of correlation of GC3 across genes between any two species or ancestors. To ensure comparability across species pairs, this coefficient was only calculated using the subset of 787 genes available in all 33 species of the data set. Importantly, in these correlation analyses, the observed GC3 values at tip nodes (extant species) were replaced with values predicted by the model, that is, expected GC3 values assuming that the estimated branch lengths and parameters of the nucleotide substitution model are true. This was necessary because, for a given gene, the model tends to smooth estimated GC3 across nodes of the tree and slightly underestimate the heterogeneity of GC3 across nodes. The comparison between internal and terminal pairs of nodes is therefore only relevant when estimated values are used for all nodes. Using actual GC3 values at tip nodes would lead to (perhaps implausibly) higher estimates of ancestral longevity.

GC3-Based Ancestral Longevity Estimation

The time-corrected index of GC3 conservation that was used for a given pair of species was $\gamma = -t/\log(\tau)$, where t is the divergence time, and τ is Kendall's correlation coefficient. This is based on the assumption of an exponential decay in the coefficient correlation over time, at a rate that is inversely proportional to the species longevity, here considered as constant over time and between the two considered lineages. Under this assumption, we have $\tau = \exp(-\alpha t/l)$, where l is the species longevity, and α is a scaling factor. The γ index was calculated for the (Catarrhini ancestor and Paenungulata ancestor) pair and for 13 independent pairs of modern species of comparable divergence times. Pairs of modern species were chosen so as to maximize the number of within-order comparisons and avoid within-family comparisons. Among Afrotheria, the hyrax/tenrec pair was preferred to the hyrax/elephant pair to better match the assumption of equal longevity between species within a pair. The divergence time of the ancestral Catarrhini ancestor/Paenungulata ancestor pair was defined as ([Pl − Ca]+[Pl − Pa])/2, where Pl is the age of the placental ancestor (105 My), Ca is the age of the Catarrhini ancestor (30 My), and Pa is the age of the Paenungulata ancestor (61 My). Regression analyses were performed using R, similar to all the statistical analyses in this study.

d*N*/d*S* Analysis

The branch-specific nonsynonymous/synonymous substitution rate ratio was calculated through substitution mapping, in the spirit of Jobson et al. (2010). For each codon of each gene of the data set, synonymous and nonsynonymous changes were mapped onto the branches of the tree by probabilistic mapping (Dutheil et al. 2005), assuming Yang and Nielsen's (1998) model of coding sequence evolution. Then for each branch, the numbers of synonymous and nonsynonymous changes were summed across genes and their ratio calculated. This approach gives equal weight to codons, not to genes, and is computationally faster than maximum-likelihood approaches (Romiguier et al. 2012). Summing counts across genes avoids taking the ratio between small numbers, thus escaping the upward bias reported by Wolf et al. (2009) when estimating d*N*/d*S* from very short amounts of sequence divergence. Linear regression of species longevity on log-transformed branch-specific d*N*/d*S* ratio was performed using terminal branches only. Then, for each internal branch of the tree, ancestral longevity was predicted based on estimated d*N*/d*S* ratio.

Results

The phylogeny of the 33 placental mammalian species used in this study is shown in figure 1. These species, whose complete genomes have been sequenced, widely differ in their life-history traits. Here, we focused on maximum life span, represented by the color of species names in figure 1. Our aim is to estimate the maximum life span of the ancestors to these living species, here represented by the root and other basal nodes of the tree. We first introduce the main arguments of this study based on illustrative examples before presenting the whole statistical analyses.

GC3 Plots

We examined the rate of genomic divergence using GC3 plots, in which the GC content at codon third positions of orthologous genes is compared between two species. Figure 2 shows GC3 plots for four species pairs, including human–tarsier (two primates, diverged ca. 71 Ma, fig. 2*A*), elephant–hyrax (two afrotherians, diverged ca. 61 Ma, fig 2*B*), and human–elephant (fig. 2*C*), which diverged approximately 105 Ma, and whose common ancestor is that of all extant placentals. The first two plots illustrate the faster GC3 dynamics of short-lived species—many genes have undergone a substantial increase in GC3 in tarsier and hyrax lineages, when compared with human and elephant lineages, resulting in asymmetric plots. This is in agreement with the documented trend for GC3 enrichment in small mammals (Romiguier et al. 2010).

Notably, the human–elephant plot reveals a high level of GC3 conservation between these two species, such that the human/elephant correlation coefficient, here measured by Kendall's $\tau = 0.80$, is higher than those computed for the human/tarsier ($\tau = 0.72$) and elephant/hyrax ($\tau = 0.73$) pairs, despite the >1.5-fold longer divergence period. The 105 My of human/elephant divergence was more conservative regarding GC3 than the 64 My of cow/pig divergence ($\tau = 0.77$), and the 40 My of rabbit/pika divergence ($\tau = 0.78$), not to speak of

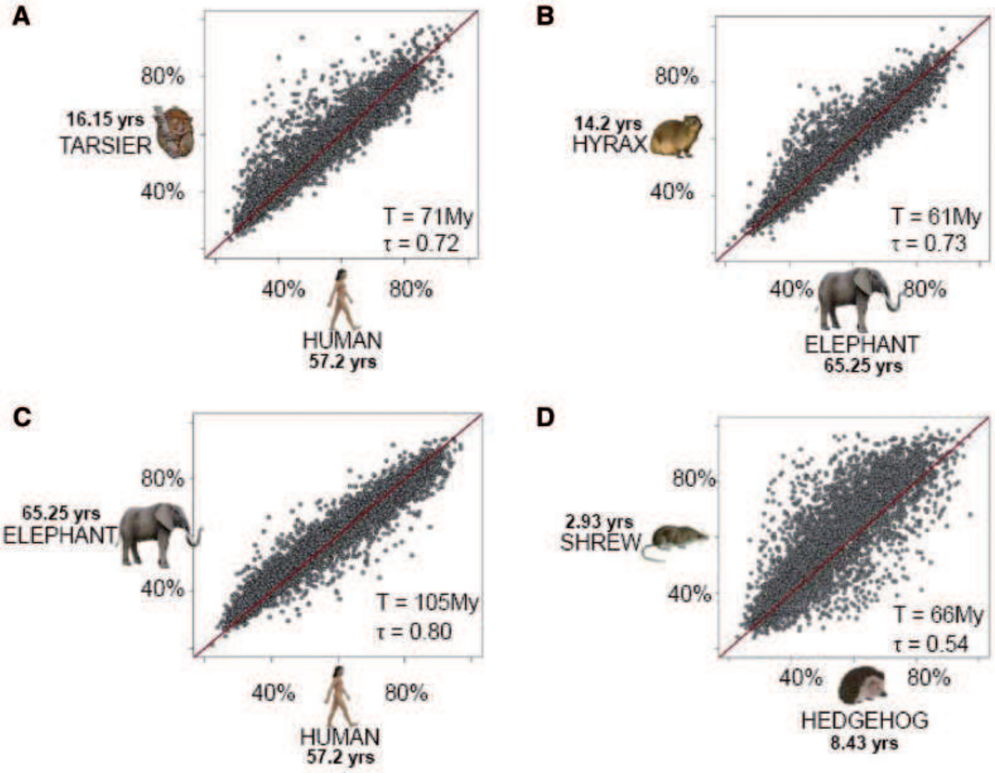


Fig. 2. Typical GC3 plots between pairs of long-lived and/or short-lived species. The family average in maximal longevity is given for each species.

that of mouse/kangaroo rat (70 My, $\tau = 0.57$) or shrew/hedgehog (66 My, $\tau = 0.54$, fig. 2D).

This result confirms that long-lived species evolve slowly regarding GC3 genes, while also strongly suggesting that all human and elephant ancestors were long-lived animals. If the first primates and afrotherians had had a short life span, then the GC3 of these ancestors should have diverged quickly during early placental evolution, and reached a state similar to figure 2D, thus indelibly marking the human/elephant plot. We suggest that an ancestral GC3 pattern typical of short-lived taxa (fig. 2D) cannot lead to a modern GC3 pattern typical of long-lived species (fig. 2C) during the course of evolution: even if GC-biased gene conversion had stopped, the random accumulation of AT→GC and GC→AT mutations through genetic drift is not expected to drive individual gene GC3 back to similar values in two independently evolving lineages (see fig. 3 for illustration). Note that high levels of GC3 conservation were also found between human and the long-lived sloth ($\tau = 0.78$), and between elephant and sloth ($\tau = 0.75$), extending the rationale to the xenarthran ancestral branch.

Ancestral Longevity Estimation

To quantify the early rate of GC3 divergence, we reconstructed ancestral gene GC3 using the maximum-likelihood method and a nonhomogeneous model of sequence evolution. From these reconstructions, we built GC3 plots between internal nodes of the tree and measured levels of GC3 conservation between these ancestors. We focused on divergence between the most recent common ancestor (MRCA) to extant Catarrhini (Old World monkeys and apes) and the MRCA to extant Paenungulata (elephants, sirenians, and hyraxes; fig. 1). These two MRCAs diverged over an average 59.5 My since their common placental ancestor (see Materials and Methods). Their level of GC3 correlation is $\tau = 0.84$. We compared this number to levels of GC3 correlation calculated between independent pairs of extant species with a comparable divergence time (59.5 ± 30 My). We found that among 13 such species pairs, only the long-lived chimpanzee/*Macaca* ($\tau = 0.93$, 44 years) and cow/dolphin ($\tau = 0.85$, 36 years) pairs showed levels of GC3 correlation similar to, or higher than, the Paenungulata/Catarrhini ancestral pair (supplementary table S1, Supplementary Material online). All 11 species pairs with an average longevity lower than 30 years diverged more rapidly than Paenungulata/Catarrhini regarding GC3.

With this data set, a linear regression of longevity on a time-corrected GC3 conservation level (see Materials and Methods) revealed a strong, positive correlation ($r^2 = 0.93$) and predicted that the average longevity during early Catarrhini/Paenungulata divergence was 33.3 ± 7.6 years (fig. 4). Among the 897 placental species with a documented maximal longevity, 174 (19%) are within this range, and 65 (7%) are more long lived. The list of modern species matching the 33.3 ± 7.6 prediction interval includes 76 primates, 38 cetartiodactyls, and 30 carnivores but only five rodents (mostly porcupines), whereas rodents represent ~40% of placental species overall. In this list of 144 species, 95% of the body mass distribution is within (0.65–18 kg) in arboreal

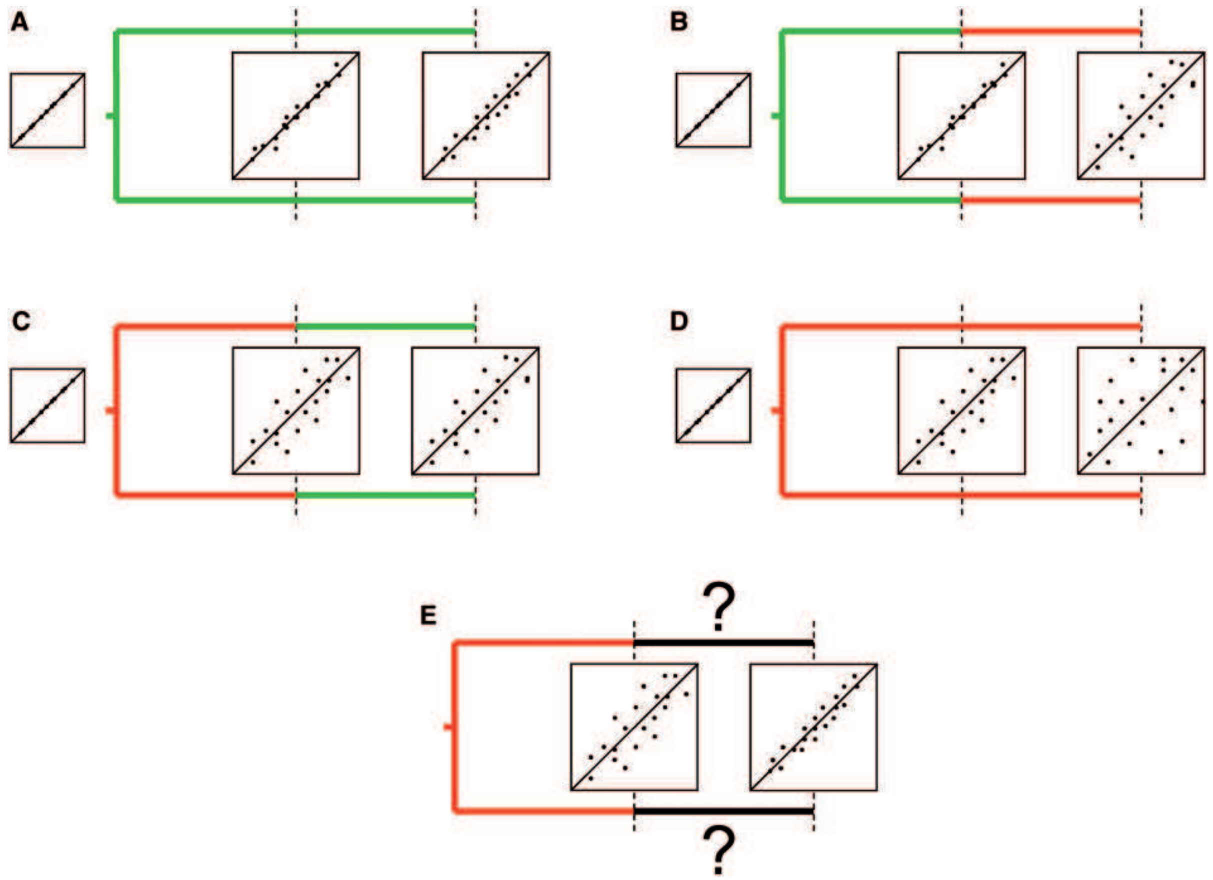


Fig. 3. Predicted behavior of GC3 plot under four scenarios of life span evolution. Two periods of time, early and late, are considered. Green branches are for long-lived lineages and red branches for short-lived ones. GC3 plots between the two diverging lineages are represented at three time points: initial stage (no divergence, correlation coefficient = 1), end of the early period, and end of the late period. In scenario *A*, an ancestral long life span is kept throughout the two periods. In this case, gene GC3 evolution is slow, and a strong level of correlation is to be expected from GC3 plots. Scenario *B* represents convergent evolution from a long ancestral life span to a short derived life span. Here, the GC3 plot is only little perturbed during the early period, like in (*A*), but a fast decay of correlation coefficient is expected during the late period. In scenarios *C* and *D*, the ancestral life span was short and was either kept throughout (*D*) or convergently evolved to a derived long life span (*C*). In both cases, a low correlation coefficient is expected for GC3 plots, if only because the intermediate GC3 plot is weakly correlated. Under scenario *C*, convergent evolution toward a long life span is not expected to result in an increased correlation coefficient but rather in a freeze of GC3 plot. Reverse evolution toward a higher correlation coefficient, represented in panel *E*, can only occur under an implausible scenario in which a large number of initially diverged genes would convergently reach a common equilibrium GC3 in the two lineages.

species and within (3.75–800 kg) in terrestrial species—arboreal mammals are known to be longer lived than terrestrial mammals, for a given body mass (Shattuck and Williams 2010). Notably, all members of the former Lipotyphla group ("insectivores": Eulipotyphla, Afrosoricida, Macroscelidae, Scandentia, and Dermoptera), often considered as having retained placental plesiomorphies (Madsen et al. 2001; Asher 2005), have a maximal longevity of less than 19 years (mean = 7 years), that is, well outside the prediction interval.

d*N*/d*S* Analysis

The above calculations rely on prior knowledge of the age of internal nodes in the placental tree, here taken as errorless data. Although somewhat consensual in recent molecular phylogenetic literature (Meredith et al. 2011), divergence dates between mammals are obviously uncertain (Murphy and Ezirik 2009). Even though our GC3-based analysis is only dependent on relative, not absolute, divergence dates, confirming these results in a time-independent manner would appear desirable. To achieve this aim, we focused on terminal branches of the tree and measured the branch-specific d*N*/d*S* ratio. This time-independent statistics is known to be correlated to species life-history traits: d*N*/d*S* is higher, on average, in long-lived species than in short-lived ones (see earlier). We here confirmed this relationship: a significant ($r^2 = 0.86$, *P* value < 0.0001), positive correlation was obtained between log-transformed terminal branch d*N*/d*S* and species longevity (fig. 5). Then, we measured the d*N*/d*S* ratio in internal branches of the placental tree. On the basis of the linear regression of figure 4, we predicted the average longevity in ancestral placental lineages, as shown by colors in figure 1.

Although the prediction interval for any specific branch was wide, this analysis essentially corroborated the GC3-based estimates. Among the 10 internal branches separating the

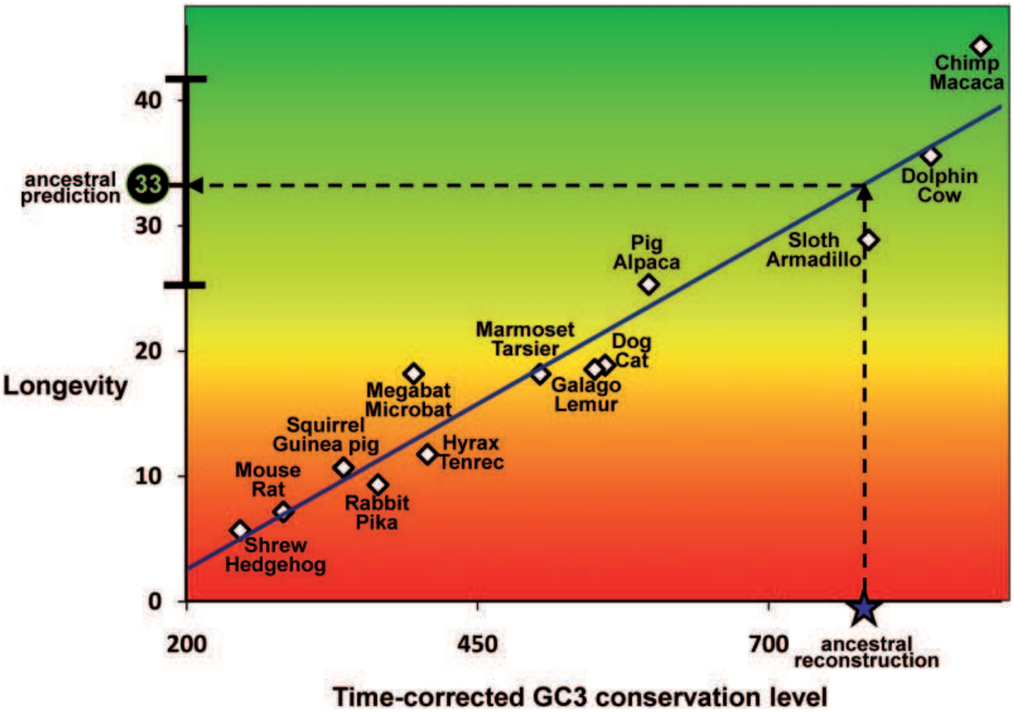


FIG. 4. Linear regression of longevity on time-corrected GC3 conservation. Each dot is for a pair of species. *x* axis is the time-corrected index of GC3 conservation $-t/\log(\tau)$, where t is the divergence time between the two considered species, and τ is Kendall's correlation coefficient calculated from GC3 plots. *y* axis is the average maximal longevity of the considered pair. The star marks the position of the Catarrhini MRCA/Paenungulata MRCA pair in *x* axis. The estimated longevity and prediction interval for this pair of ancestors is shown on *y* axis (Supplementary Material online).

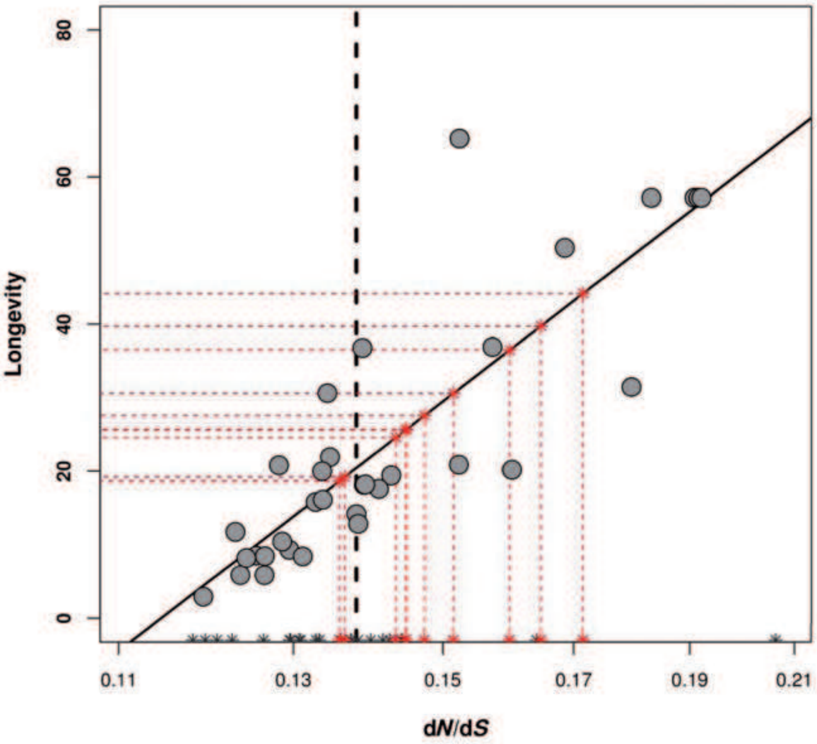


FIG. 5. Linear regression of longevity on d*N*/d*S* ratio. Each dot is for a terminal branch of the placental tree. *x* axis: d*N*/d*S* ratio (log scale). *y* axis: species maximal longevity. Vertical dotted line: median d*N*/d*S* ratio across terminal branches. Stars in *x* axis represent the estimated d*N*/d*S* ratio in internal branches of the tree. Red stars correspond to the 10 internal branches from the path connecting the Catarrhini MRCA to the Paenungulata MRCA. Estimated ancestral longevities for these 10 branches are shown by projection onto the *y* axis. Raw data available from supplementary table S2, Supplementary Material online.

Catarrhini MRCA from the Paenungulata one (red stars in fig. 5), the predicted longevity ranged from 19 to 45 years and averaged 29.0 years, confirming the results of figure 4. This analysis suggests that the MRCAs of Afrotheria, Xenarthra, Laurasiatheria, and Euarchontoglires—the four major placental superorders (indicated by their initial in fig. 1)—were long lived, whereas the MRCAs of Eulipotyphla, Lagomorpha, and, surprisingly, Cetartiodactyla, were short-lived animals. In this analysis, the average d*N*/d*S* ratio of internal branches (0.141) was almost identical to the average d*N*/d*S* ratio of terminal branches (0.140), and no correlation was observed between the estimated d*N*/d*S* ratio and the phylogenetic depth of an internal branch, here defined as the age of its bottom node ($r^2 = -0.09$, P value = 0.6). These results suggest that our inferences are not affected by a systematic upward bias of d*N*/d*S* estimates in ancient branches, as could be expected in case of substitutional saturation or alignment errors in deeply branching lineages.

Controls

Additional control analyses were achieved to check the robustness of ancestral longevity estimates. The impact of missing data was assessed by removing from the data set alignments that included at least one sequence containing a proportion of gaps above some threshold. The results were only slightly affected. When the threshold was set at 50%, only 205 genes were retained, and time-dependent predicted ancestral longevity for the Catarrhini/Paenungulata MRCA was 32.8 ± 7.8. The impact of CpG dinucleotides, which are hypermutable toward TpG or CpA in mammals, was assessed by removing all CpG-affected codon sites from the alignments. A codon site was considered CpaG affected as soon as any position in the codon was involved in a CpG, a TpG, or a CpA doublet in >50% of the sequences. The results were robust with respect to this removal (time-dependent predicted ancestral longevity for the Catarrhini/Paenungulata MRCA: 29.7 ± 9.8). This control is especially needed knowing that a shift in CpG substitution rate at the time of early placental divergence has been reported (Arndt et al. 2003). Finally, qualitatively similar results were obtained when we used the age of female sexual maturity, rather than maximal life span, as a predictor of GC3 dynamics.

Discussion

Here, we used a phylogenetic approach to reconstruct ancestral genome dynamics and estimate ancestral life-history traits, based on the evidence in modern taxa that species traits influence molecular evolution. This analysis, which benefits from the strength of nonhomogeneous models of sequence evolution, illustrates the power of genomic data to unravel species evolutionary history. The link between species characteristics and GC-content dynamics has been previously observed (Romiguier et al. 2010; Nabholz et al. 2011), but its implications regarding ancestral trait reconstruction have never been considered so far.

The main signal we extract from the data is related to the much higher evolutionary instability of gene GC3 in short-lived than in long-lived species. We suggest that the high level of conservation of gene GC3 during early placental divergence is only compatible with a long ancestral life span, because putative short-lived ancestors would have left indelible mark in modern GC3 plots (fig. 1). This was formalized in our method by the assumption of an exponential decay of correlation coefficient in time, which neglects possible instances of convergent evolution in gene GC3. The very strong correlation we observed between time-corrected GC3 divergence level and species longevity (fig. 4) suggests that this assumption is largely met by the data and that GC3 plots correctly capture the evolutionary dynamics of longevity.

This result is reinforced by the analysis of d*N*/d*S* variations between lineages, an independent source of information that corroborated the GC3-based inferences. Note that the agreement between GC3 plots and d*N*/d*S* is not limited to the deepest branches of the tree. The report of a low d*N*/d*S* ratio in the ancestral Cetartiodactyl branch, for instance, is consistent with the relatively low level of GC3 correlation between long-lived members of this group and other long-lived mammals. The GC3 correlation coefficient between alpaca and horse (0.72), for instance, is lower than between the more distantly related horse and *Macaca* (0.8) or horse and sloth (0.75)—to talk only about species of maximum life span ~30 years. So the short-lived predicted ancestral Cetartiodactyl has apparently left an indelible mark in the GC3 plots of this group, in agreement with the rationale of figure 3.

Our study of past genome dynamics predicts that the maximum life span of the ancestors to modern placentals was more than 20 years and probably ~30 years. According to our results, these early placental mammals were either large-sized terrestrial or medium-sized arboreal animals. Ancestral mouse-like or shrew-like life-history traits are excluded, questioning one of the most frequently told stories in evolutionary biology. Our analysis suggests that very small size in placental mammals is a derived state, which evolved several times independently (e.g., in Rodentia, Chiroptera, Eulipotyphla, and Macroscelidea), in contradiction with Cope's rule (Monroe and Bokma 2010). We note that the hypothesis of a long ancestral life span for placentals appears consistent with the ecological theory. Evolution of the placenta (and more generally viviparity) means that reproduction is delayed to the benefit of juvenile survival, that is, increased investment in parental care, which could appear unexpected in a short-lived, r-strategy species (Stearns 1992). This comment is only loosely relevant to our results, though, because the evolution of the placenta may have preceded to a non-negligible extent of the diversification of the MRCA of modern placental superorders.

From a paleontological standpoint, our results call for re-examination of large or arboreal mammalian taxa as potential stem groups to extant placental superorders. The Cenozoic fossil record includes relatively large mammals, such as the late Cretaceous, dinosaur-eating *Repenomamus giganteus* (15 kg) (Hu et al. 2005), which is indicative of the existence of potentially long-lived mammals before the KT crisis.

Chapitre 5

Less is more in mammalian phylogenomics : AT-rich genes minimize tree conflicts and support Afrotheria as sister group to all other placentals

Less is more in mammalian phylogenomics : AT-rich genes minimize tree conflicts and support Afrotheria as sister group to all other placentals.

Jonathan Romiguier*[1]and Vincent Ranwez[2] , Frederic Delsuc[1], Nicolas Galtier[1]and Emmanuel JP Douzery[1]

[1]CNRS, Université Montpellier 2, UMR 5554, ISEM, Montpellier, France
[2]Montpellier SupAgro, UMR 1334, AGAP, Montpellier France

Email: Jonathan Romiguier*- jonathan.romiguier@gmail.com; Vincent Ranwez - ranwez@supagro.inra.fr; Frederic Delsuc - frederic.delsuc@univ-montp2.fr; Nicolas Galtier - nicolas.galtier@univ-montp2.fr; Emmanuel JP Douzery - emmanuel.douzery@univ-montp2.fr;

*Corresponding author

Abstract

Background: Despite the rapid increase of phylogenomic data sets, a number of important nodes of the tree of life are still unresolved. Among these, the rooting of the placental mammals remains a matter of debate, even though the abundance of genomic data in placentals has increase at an impressive pace during the last decade. The difficulty lies in the pervasive phylogenetic conflicts existing among genes, each one telling its own story, which may be reliable or not. Here we identify a simple criterion, GC-content, which substantially helps in determining which genes most likely reflect the species history.

Results: We assessed the ability of 13,111 coding sequence alignments to correctly reconstruct the placental phylogeny. We found that GC-rich genes induce a higher amount of conflict among gene trees, and perform worse than AT-rich genes in retrieving well-supported, consensual nodes of the placental tree. We interpret this GC-effect as a consequence of genome wide variations in recombination rate - recombination is known to drive GC-content evolution, and is problematic for phylogenetic reconstruction in the context of incomplete lineage sorting. When we focused on the AT-richest fraction of the data set, the level of resolution of the placental phylogeny was greatly increased, and a strong support was obtained in favour of an Afrotheria versus Exafrotheria rooting for placental mammals.

Conclusions: We show that in mammals most of the conflict among genes, which has so far hampered the resolution of the placental tree, is concentrated in the GC-rich fraction of the genome. We suggest that, because it is a reliable marker of the long-term recombination rate, GC-content is an easy-to-calculate criterion likely to help separating the wheat from the chaff in various phylogenomic data sets.

Background

Most of evolutionary biology studies rely on a well-resolved phylogenetic tree. Over the last decade, this requirement has been mainly achieved through molecular data. Initially reduced to few genes, they are now growing more and more thanks to high-throughput sequencing. With whole genome datasets, resolving a species phylogeny is no longer a matter of quantity of unambiguously aligned sites. However, despite the rise of phylogenomics [1,2], important nodes of the Tree of Life remain unresolved.

One of the most iconic example is the phylogeny of placental mammals, among the first studied under the light of molecular data. Challenging classical morphological classification, first studies suggested three principal clades of placental mammals [3, 4]: Afrotheria (e.g. elephant and tenrec), Xenarthra (e.g. armadillo and sloth), and Boreoeutheria, which groups most of its species in two super-orders, namely Euarchontoglires (e.g. primates and rodents) and Laurasiatheria (e.g. ruminants, cetacean, bats, carnivores). These clades are now well-established and confirmed by several other studies [5–8]. Nevertheless, other nodes are more tricky.

Among the most difficult nodes to resolve, the root of placental mammals is one of the most debated. Several contradicting studies suggest competing hypotheses, namely (i) Epitheria (Afrotheria + Boreoeutheria) [9–12], (ii) Exafrotheria (Xenarthra + Boreoeutheria) [4, 13–15] or (iii) Atlantogenata (Afrotheria + Xenarthra) [7, 8, 16–22]. Although essential to understand the mammal diversification regarding to the continental drift, transposons insertions studies suggest that this node is probably impossible to resolve [23]. In addition, part of inter-order relationships remains unclear inside Laurasiatheria, as well as the place of Scandentia (tupaia) and some rodent clades inside Euarchontoglires. These discrepancies were first believed resolvable with the advent of whole genome sequencing. Nevertheless, despite 39 mammalian genomes now available (Ensembl, release 67 [24]), the uncertainties still persist, highlighting the current failure of phylogenomic to unravel the trickiest nodes of the Tree of Life. Why such a massive amount of characters could be kept in check ?

Actually, a growing corpus of coding sequences revealed conflicting evolutionary histories among genes [25, 26]. According to literature, these discordances could be mainly due to coalescent stochasticity [26, 27]. Coalescent theory describes the genealogy of gene copies since their most recent common ancestor. During and after speciation, each gene copies evolve distinctly within the two species. Yet, some copies may have preserved identical sequences when the next speciation took place, leading to the possibility that at least one of them first coalesces with a copy from a less closely species. Thus, this random, independent sorting of genes could lead to genealogies different from the surrounding species phylogeny. This phenomenon is called incomplete lineage sorting (ILS), and was detected in several different taxa [28–32], for example in hominids [27, 33, 34]. Actually, ILS is mainly observed during rapid bursts of speciations, a common phenomenon in placental mammals. This is particularly true for the trickiest nodes, e.g. the basal divergence and the explosive radiation in Laurasiatheria orders [8, 35].

For all these reasons, dealing with ILS seems one of the key to unravel the phylogeny of placental mammals. In order to fulfill this purpose, Liu et al. proposed several coalescent methods [36, 37]. Recently, these approaches give insights on the complete phylogeny of placentals, but still provide conflicting results for the placental root. Thus, flanking regions of ultraconserved elements strongly support the Exafrotheria hypothesis [15], whereas coding sequences support as strongly Atlantogenata [22]. Despite appropriate tree-building methods, the nature of data seems a primary concern (e.g. coding versus non-coding). Here, we suggest an alternative approach. Instead of trying to deal with ILS, we propose to cut it from the source, and work on curated datasets.

Commonly used, orthologous coding sequences are the biggest resource available for mammals. Given current computing limitations, handling the whole dataset in optimal conditions is nearly impossible. Furthermore, adding more and more genes seems useless in case of frequent ILS, and only leads to more and more gene discrepancies. To address these two issues, eliminating genes with histories differing from species genealogy would be ideal. Searching factors that increase ILS probability of a gene is a first step.

Through a deep analysis of hominid genomes, Hobolth et al [27] reported that recombination is associated to more frequent ILS. An explanation of this effect is that recombination maintain polymorphism against the effect of background selection [38], increasing the variants of a gene hence more difficult to sort. Beyond ILS, recombination is a common issue in phylogeny and can mix genes in neighboring segments with different histories [25, 39, 40]. For all these reasons, genes with low recombination rates are supposed to give better species phylogenies. However, local recombination rates are known to change frequently through the placentals history and is a scarce data in most of species. Fortunately, recombination rate is positively associated to GC content in mammals [41]. This effect is due to a neutral mechanism, the biased gene conversion, a GC reparation bias occuring during meiotic recombination and shaping mammalian genomes since the beginning of their history [42, 43]. GC content correlates even more strongly with ILS than current recombination rate [27]. Indeed, GC content is probably a better long-term predictor of recombination than current recombination rates themselves.

Easily available, GC content could be a proxy for the probability of a gene to experiment ILS. To test this hypothesis, we propose to analyse the accuracy of the phylogeny of a gene with respect to its GC content in mammals. A significant effect of nucleotidic composition would provide an easy way to reduce gene conflict in phylogenies and promising options to finally settle the mammal tree.

Results

Higher probability of topology errors in GC-rich gene trees

We compared 13 111 mammal gene trees to an assumed true species genealogy (controversial nodes kept unresolved, see topology of Figure 1). Gene trees were then divided in 131 subsets of 100 genes according to their GC3 content. Figure 2A shows the main evidence that GC-richer genes are markers more prone to inaccuracy for inferring mammal species phylogeny. Indeed, the mean of gene tree error correlates strongly with GC3 content. Regardless the metrics used to measure the distance between topologies, this correlation remains highly significant : Robinson-Foulds distance (P-value < 0.0001, $r^2 = 0.70$), quartet similarity (P-value < 0.0001, $r^2 = 0.77$) or triplet similarity (shown in figure 1A, gray dots, P-value < 0.0001, $r^2 = 0.85$). Thus, the error rate on GC-rich genes is up to 5 times higher than that of most AT-rich genes.

GC and AT cumulative subsets (red and blue dots) illustrate the same trend. Adding more and more GC-rich gene trees to a subset increase nearly always its level of discrepancy among genes (red dots). By contrast, adding more AT-rich gene trees always improve the global agreement (blue dots). Thus, the whole dataset (right part of the plot) could be curated by the removal of GC-rich genes, whereas removing AT-rich genes seems often a bad idea.

This fact alone suggests that using GC-rich markers should be avoided for phylogenomic studies. Nevertheless, the reason behind this strong effect is still unclear. Obviously, alignments with less characters (less sites and/or species) generally produce less accurate phylogeny. Such a bias in current genomic databases could be the simplest explanation. In agreement with this hypothesis, we found a negative correlation between the GC-content of an alignment and its number of sites (P-value $<$ 0.0001, r^2 = -0.03) or species (P-value < 0.0001, r^2 = -0.17). This is probably due to GC-biases reported in various genome sequencing methods based on PCR [44–46]. Thus, it clearly appears that the GC-rich genes of current database releases yield to alignments that are less informative. Yet, one can wonder if this elevated error rate in GC-rich gene trees is only due to missing characters. To answer this question, we repeated our analyses after having homogenized the number of sites and species along the GC gradient of our 13 111 alignments (see Material and Methods). Results displayed in Figure 2B shows the effect of the GC-content without any putative bias due to character quantity. The correlation is weaker, but remains highly significant for all metrics: Robinson-Foulds distance (P-value < 0.0001, $r^2 = 0.33$), quartet similarity (P-value < 0.0001, $r^2 = 0.49$) and triplet similarity (Figure 2B, gray dots, P-value < 0.0001, $r^2 = 0.59$).

Independently of their lower quality in current databases, GC-rich genes clearly present a rise of inferring erroneous topology.

GC-rich genes are prone to miss monophyly of the deepest placental nodes

Supertree approach

To characterize the errors induced by GC-content, we computed one supertree per non-cumulative subsets, extracted the support values of each well-established node (accordingly to our reference tree in Figure 1), and correlated them to GC3-content.

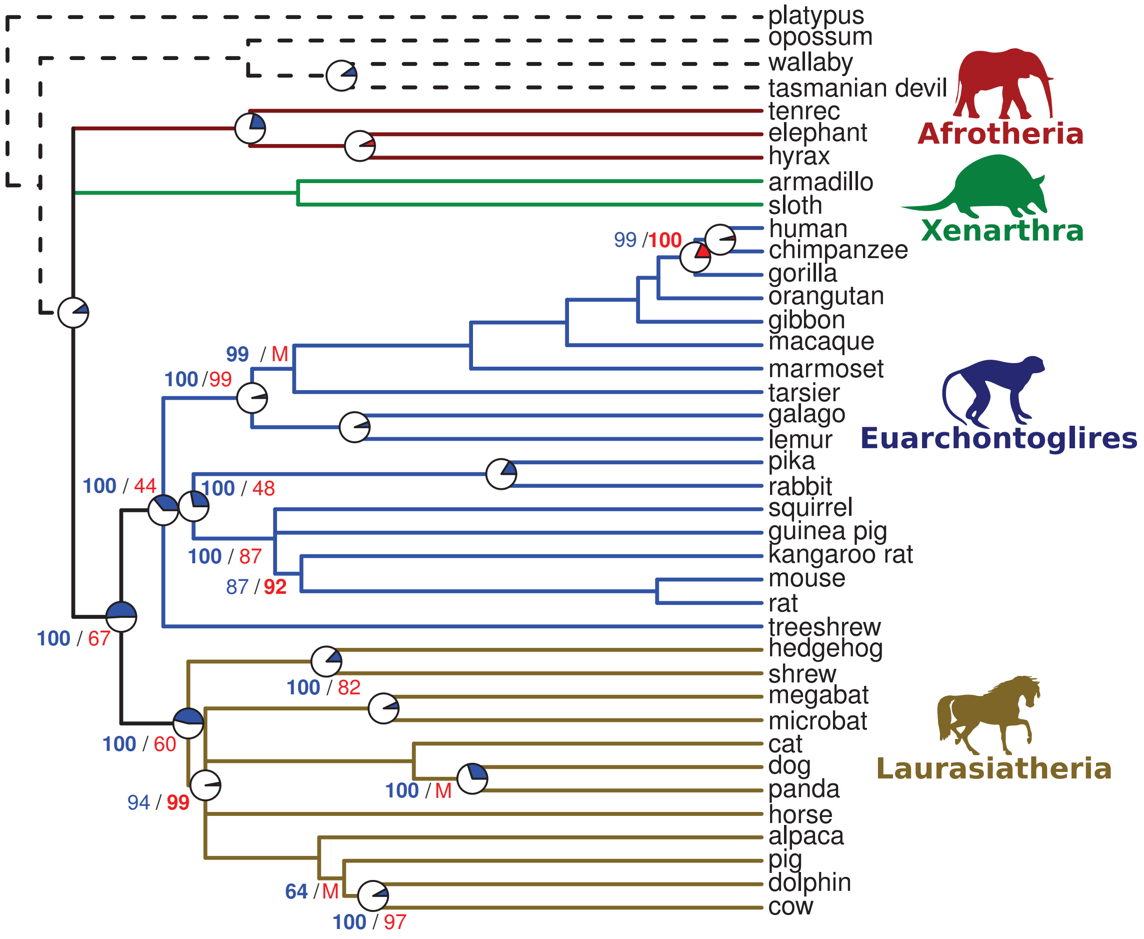


Figure 1 - GC effect on reference tree topology

This reference tree topology keeps unresolved the most debated nodes of the mammal phylogenetic tree (see Reference tree section in Methods). Branch lengths are proportional to the divergence dates (except for distant non-placental species with dotted branches).

Pie charts refer to section "supertree approach", and show the explained variance of node support by GC3 (blue for a negative effect of GC3, red for a positive one). We show only significant effect for the gap homogenized dataset (same alignment quality among GC-rich and AT-rich genes).

Blue and red numbers refer to a different analysis, "supermatrix approach"), and show bootstrap supports for the concatenation of the top 100 AT-rich genes (in blue) versus the concatenation of the top 100 GC-rich genes (in red). M is for Missing node.

The pie charts on nodes of Figure 1 show the percent of their support value variance explained by GC3 (i.e. r2). Blue pies are for a negative effect of GC-content on support values, red pies for a positive one. To be as conservative as possible, we only consider significant correlations of the gap homogenized dataset.

As expected, the support value of most nodes is negatively correlated with GC3%. This result is in agreement with Figure 2, where GC-rich genes are prone to produce topological errors, thus decreasing support values of consensual nodes. Obviously, this effect is more striking with raw data, which encompass the effect of alignment gappiness (not shown).

Interestingly, the 3 nodes the most affected by GC-content are deep in the placental tree. Their support values are all negatively correlated with GC3 (P-value < 0.0001): Euarchontoglires ($r^2 = 0.77$ for raw datas, $r^2 = 0.36$ after gap homogenization), Laurasiatheria ($r^2 = 0.79$ for raw datas, $r^2 = 0.46$ after gap homogenization) and Boreoeutheria ($r^2 = 0.82$ for raw datas, $r^2 = 0.51$ after gap homogenization). The other nodes with high explained variance by GC3 are Caniformia (dog+panda) and Afrotheria. Unexpectedly, the relationship among human, chimp and gorilla is positively affected by GC-content. Although the explained variance is high, the slope of the correlation is lower than for other nodes (+0.12 for hominids, compared to -0.27 for the shrew+hedgehog node and up to -0.38 for Boreoeutheria).

Supermatrix approach

We concatenated the top 100 GC-richest and the top 100 AT-richest genes in two supermatrices. For a fair comparison, we used alignments obtained after gap homogenization. Thus, the two supermatrices contain both 81 924 sites, with the same number of species (39) and the same proportion of missing data. We used each of these two supermatrices to infer a phylogeny of mammals thanks to RAxML [47], and compared the support values in Figure 1 (blue numbers for the AT-rich supermatrix, red numbers for the GC-rich). Equal support values are not shown.

These results are in strong agreement with the supertree approach. Once again, nodes related to the super-order scale are better supported by AT-rich genes. In the GC-rich super-matrix tree, Boreoeutheria, Euarchontoglires and Laurasiatheria support values respectively drop from 100 to 67, 100 to 44 and 100 to 60. Some support values decrease more dramatically, e.g Glires (drop from 100 to 48). Moreover, the dog+panda, pig+cow+dolphin and tarsier+anthropoidea clade are not recovered in the tree inferred from the GC-richest supermatrix.

Nowadays largely accepted, super-order nodes are among the most important revealed by molecular phylogeny. Interestingly, these nodes are deep and often result from rapid speciation events, a feature shared with the most unresolved nodes. This is particularly true for the root of Placentalia (Afrotheria, Xenarthra and Boreoeutheria relationship), leading to promising perspective to unravel this node.

AT-rich genes support the Exafrotheria hypothesis for the Placental root

Atlantogenata : the most common error in mammal phylogeny ?

AT-content and increased support for deep nodes are clearly linked. But is it true for the most debated one, i.e. the root of Placentalia ? To answer this question, we concatenated the alignments used in Figure 2a into 131 100-gene supermatrices and performed a maximum likelihood analysis on each of them. Corresponding bootstrap supports for the three hypotheses about placental root are displayed in Figure 3, according to GC3 content and tree quality. The resulting trees are in strong conflict, clearly divided between the Atlantogenata and Exafrotheria hypotheses. This is consistent with the last studies, which exclude an Epitheria root of Placentalia and support an Atlantogenata [22] or an Exafrotheria one [15]. Interestingly, support values are here clearly non-randomly distributed. Producing better trees, AT-rich genes (left part of Figure 3, 77% of the first quarter) mainly support Exafrotheria, whereas GC-rich genes (right part of Figure 3) prefer Atlantogenata. To understand this tendency, we computed the root-to-tip branch lengths for each species of Afrotheria and Xenarthra. The resulting mean value for each ML tree is reported on the top of the Figure 3. Clearly, GC-rich genes produced trees with longer branches, not only for these taxa (P-value < 0.0001, r^2=0.59) but also for the whole tree (P-value < 0.0001, r^2=0.66). Associated with the increase of topology error, it strongly suggests long branch attraction artefacts [48]. Indeed, Afrotheria and Xenarthra are poorly sampled clades and include the two longest branches of the tree. Longer and longer, the more the GC-content is high, the

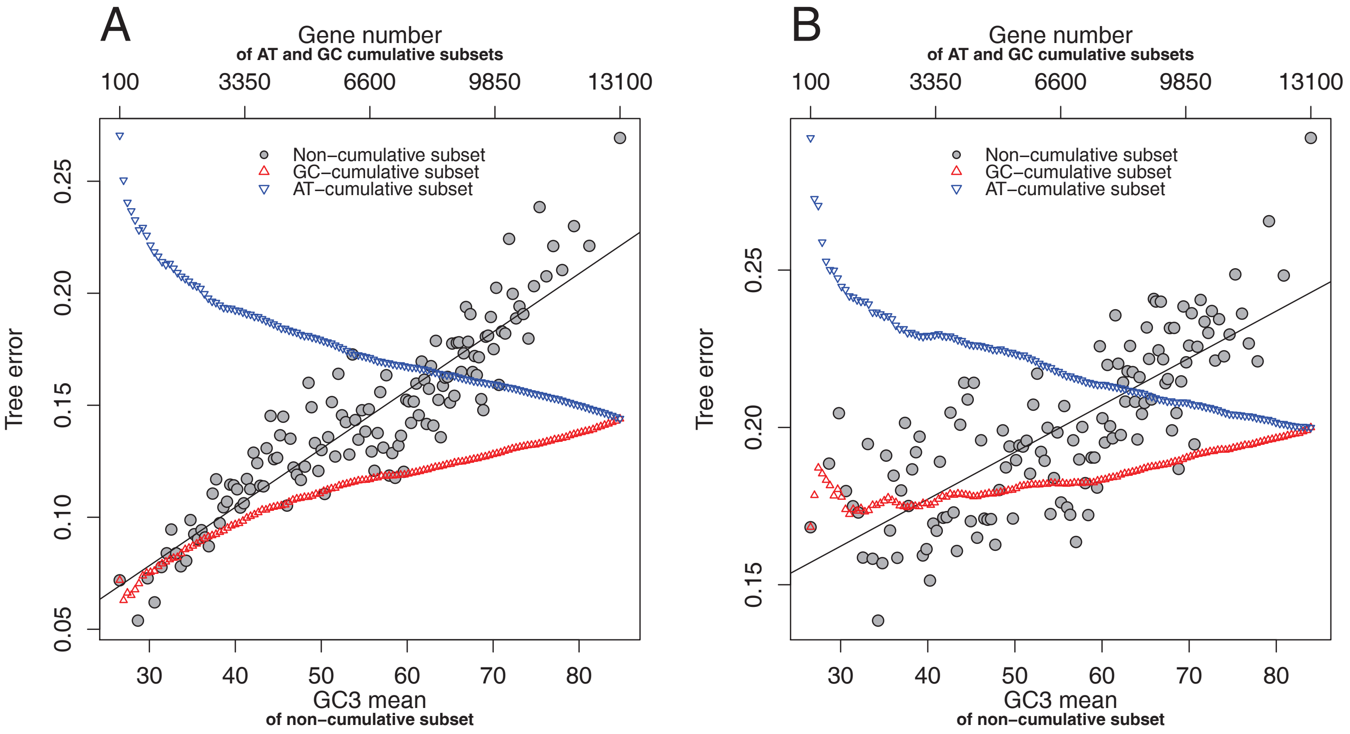


Figure 2 - GC effect on gene tree error

Each gray dot represent 100 alignments (for a total of 131), positionned according to its GC3 mean (X-axis) and its average tree errors (Y-axis). A gene tree error is here the proportion of false species triplet of a tree compared to the true species genealogy (reference tree topology of Figure 1). Red and blue dots are for cumulative subsets, where X-axis is the number of genes of the subset. GC cumulative subsets (red) contains more and more GC-rich genes, AT cummulative subsets (blue) more and more AT-rich genes. Figure 2A is for the raw dataset without any control. Figure 2B is for a modified dataset where alignment gappiness (site and taxon number) is homogenized along the GC-gradient.

more they attract each other to form an artefactual Atlantogenata association. In the light of these results, it suggests that Exafrotheria might be the root of the species tree, whereas Atlantogenata could be the most common topology error due to long-branch attraction. Such a relation between GC-content and homoplasy make sense with the biased gene conversion action mode. We shall return to this point in Discussion.

Exclusion of the unreliable GC-rich genes : a first attempt

Taking care of GC-rich genes seems then highly relevant to resolve the placental root. Is a curated dataset without such recombining genes able to resolve the whole tree ? We propose a first attempt to get rid of this evolutionary noise. To be sure to discard highly recombining locus, we excluded alignments with a GC3% superior to the average rate of the whole genome. In mammals, this average rate is roughly equal to 40% (e.g. 40.91 for the human), with extreme ranging from 37.82% (opossum) to 45.49% (platypus) [49]. However, although better, AT-rich genes can always give conflicting histories, even below 40% of GC3-content. Obviously, GC-content can not track some short and intense recombination events that could bring local discrepancies on a specific node. To keep only the best of these genes, we excluded those which produce a gene tree with more than 10% of triplet error (regarding to the reference tree of Figure 1). It gives a large dataset of 1640 genes, which we concatenated to perform a maximum likelihood analysis from RAxML [47]. To our knowledge, this alignment of 4,417,485 sites is the biggest ever analyzed in mammal phylogeny, pushing current computational capacities near their limits. Because such a large concatenated dataset could produce over-estimated bootstrap values [48], we analyzed another dataset of more modest size : the AT-rich alignments (GC3 below 40%) which contain the full species set (39 taxa, 175 genes). Bootstrap values of the large and small datasets are presented in Figure 4. As expected following Figure 3, these AT-rich alignments support the Exafrotheria hypothesis. The bootstrap values of the large dataset are all equals to 100, with the exception of the Cetartiodactyla + Chiroptera node. However, these values drop significantly with the small dataset : 90 to 63 for the Cetartiodactyla+Chiroptera node or 100 to 78 for the Perissodactyla+Carnivora node. The topology changes for the tree shrew, which is related to Glires or Primates depending on the dataset. Exafrotheria and the squirell sister-group of all other remaining rodents are however still very well supported.

Discussion

The GC syndrome

Here, we prove that GC-rich genes are generally less reliable to reconstruct the species phylogeny of placental mammals. Two distinct reasons may explain this result : i) the number of available characters (i.e. less sites and species in GC-rich alignments) and ii) the GC-content itself.

The first factor, i.e. the smaller number of character in GC-rich regions, could be explained by biological reasons, but there is no clear consensus on this subject. Oliver et al. [50] suggested that GC-rich coding sequence regions are longer, presumably because of the AT-bias in the stop-codon composition. On the other hand, Duret et al [51] reported that AT-rich genes code for longer proteins, and suggested biological links related to the isochore structure of mammals [52]. However, a bias in databases is not excluded, and GC-rich genes could be under-represented just because of methodological issues. Indeed, GC-content biases during standard and high-throughput sequencing are well-known in litterature, even though extreme AT-rich and GC-rich genes seem both affected by the phenomenon [44–46]. Either biological or methodological, we note that this effect was never taken into account in phylogenetic studies. Here and now, AT-rich genes are represented by less-gapped alignments in Ensembl, and probably in other genomic databases. From a practical point of view, this report is relevant for most phylogeny projects. Particularly in mammals, this GC syndrome should be kept in mind when selecting a subset of phylogenetic markers already present in databases or when sequencing a specific marker in several non-model species.

In addition to the trivial effect of alignment gappiness, we report an effect of the GC-content itself. GC-rich genes increase the discrepancies between the gene histories and the species phylogeny. This result is in agreement with our expectations, i.e. that GC-content is a good long-term recombination marker, which in turn is known to increase incomplete lineage sorting [27]. The latter is supposed to occur

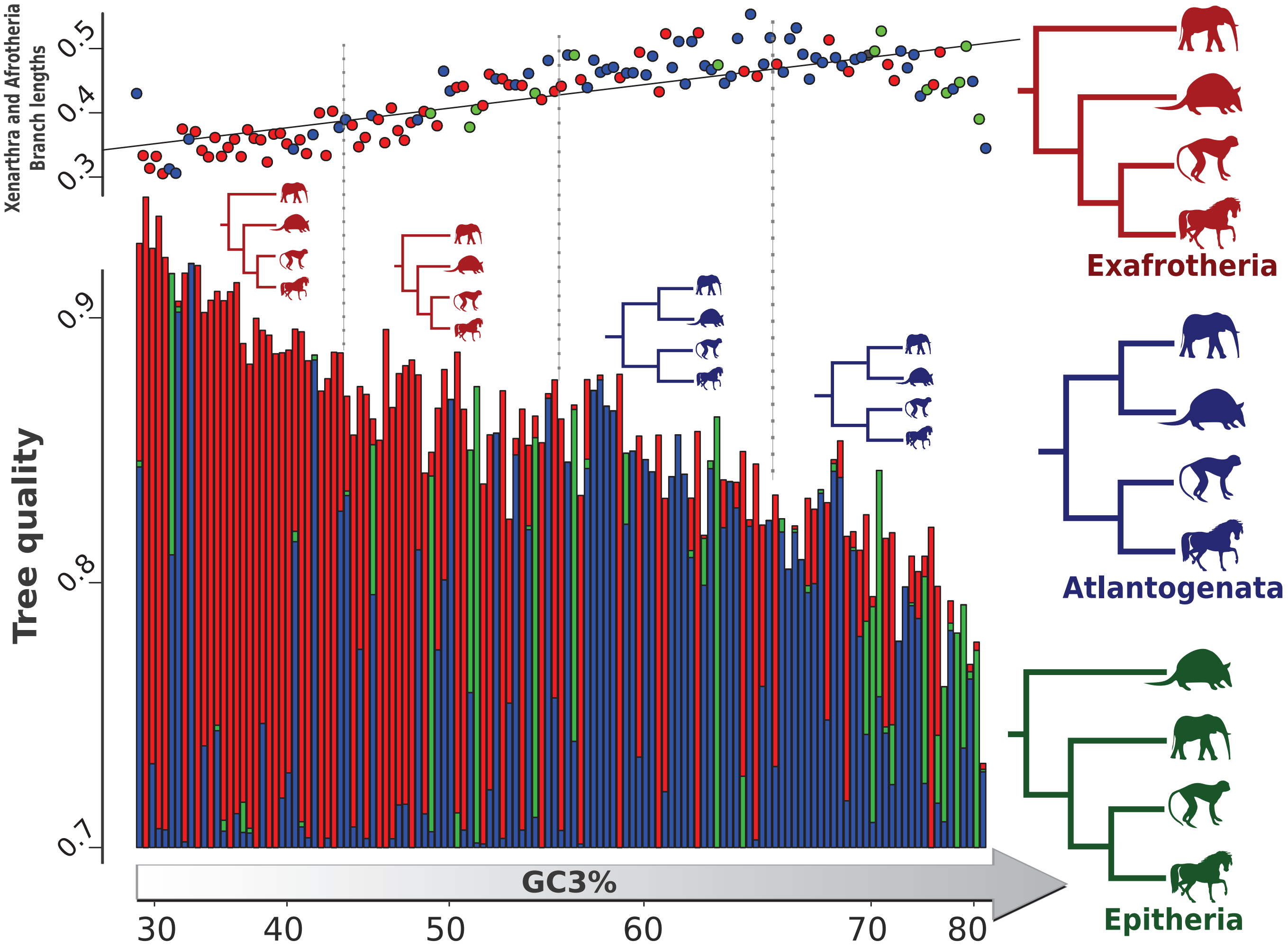


Figure 3 - Support values for the root of Placentalia according to GC content and tree quality.
Each bar represents a concatenate of 100 genes grouped accordingly to their GC3-content (as for Figure 2A). The heights of the bars are proportional to the average quality value of the 100 gene trees. Tree quality represent the proportion of congruent triplets between a gene tree and the reference species tree (Figure 1) (same value than 1 - gene tree error from Figure 2A). Red, blue and green area represent respectively the proportion of Exafrotheria, Atlantogenata and Epitheria bootstrap support provided by RAxML. On the top of the figure, the mean values of the branch length distances from the root to elephant, hyrax, tenrec, armadillo and sloth tips are displayed for the 131 subsets. Red, blue and green colors of the dots represent which color dominates their corresponding bar.

more they attract each other to form an artefactual Atlantogenata association. In the light of these results, it suggests that Exafrotheria might be the root of the species tree, whereas Atlantogenata could be the most common topology error due to long-branch attraction. Such a relation between GC-content and homoplasy make sense with the biased gene conversion action mode. We shall return to this point in Discussion.

Exclusion of the unreliable GC-rich genes : a first attempt

Taking care of GC-rich genes seems then highly relevant to resolve the placental root. Is a curated dataset without such recombining genes able to resolve the whole tree ? We propose a first attempt to get rid of this evolutionary noise. To be sure to discard highly recombining locus, we excluded alignments with a GC3% superior to the average rate of the whole genome. In mammals, this average rate is roughly equal to 40% (e.g. 40.91 for the human), with extreme ranging from 37.82% (opossum) to 45.49% (platypus) [49]. However, although better, AT-rich genes can always give conflicting histories, even below 40% of GC3-content. Obviously, GC-content can not track some short and intense recombination events that could bring local discrepancies on a specific node. To keep only the best of these genes, we excluded those which produce a gene tree with more than 10% of triplet error (regarding to the reference tree of Figure 1). It gives a large dataset of 1640 genes, which we concatenated to perform a maximum likelihood analysis from RAxML [47]. To our knowledge, this alignment of 4,417,485 sites is the biggest ever analyzed in mammal phylogeny, pushing current computational capacities near their limits. Because such a large concatenated dataset could produce over-estimated bootstrap values [48], we analyzed another dataset of more modest size : the AT-rich alignments (GC3 below 40%) which contain the full species set (39 taxa, 175 genes). Bootstrap values of the large and small datasets are presented in Figure 4. As expected following Figure 3, these AT-rich alignments support the Exafrotheria hypothesis. The bootstrap values of the large dataset are all equals to 100, with the exception of the Cetartiodactyla + Chiroptera node. However, these values drop significantly with the small dataset : 90 to 63 for the Cetartiodactyla+Chiroptera node or 100 to 78 for the Perissodactyla+Carnivora node. The topology changes for the tree shrew, which is related to Glires or Primates depending on the dataset. Exafrotheria and the squirell sister-group of all other remaining rodents are however still very well supported.

Discussion

The GC syndrome

Here, we prove that GC-rich genes are generally less reliable to reconstruct the species phylogeny of placental mammals. Two distinct reasons may explain this result : i) the number of available characters (i.e. less sites and species in GC-rich alignments) and ii) the GC-content itself.

The first factor, i.e. the smaller number of character in GC-rich regions, could be explained by biological reasons, but there is no clear consensus on this subject. Oliver et al. [50] suggested that GC-rich coding sequence regions are longer, presumably because of the AT-bias in the stop-codon composition. On the other hand, Duret et al [51] reported that AT-rich genes code for longer proteins, and suggested biological links related to the isochore structure of mammals [52]. However, a bias in databases is not excluded, and GC-rich genes could be under-represented just because of methodological issues. Indeed, GC-content biases during standard and high-throughput sequencing are well-known in litterature, even though extreme AT-rich and GC-rich genes seem both affected by the phenomenon [44–46]. Either biological or methodological, we note that this effect was never taken into account in phylogenetic studies. Here and now, AT-rich genes are represented by less-gapped alignments in Ensembl, and probably in other genomic databases. From a practical point of view, this report is relevant for most phylogeny projects. Particularly in mammals, this GC syndrome should be kept in mind when selecting a subset of phylogenetic markers already present in databases or when sequencing a specific marker in several non-model species.

In addition to the trivial effect of alignment gappiness, we report an effect of the GC-content itself. GC-rich genes increase the discrepancies between the gene histories and the species phylogeny. This result is in agreement with our expectations, i.e. that GC-content is a good long-term recombination marker, which in turn is known to increase incomplete lineage sorting [27]. The latter is supposed to occur

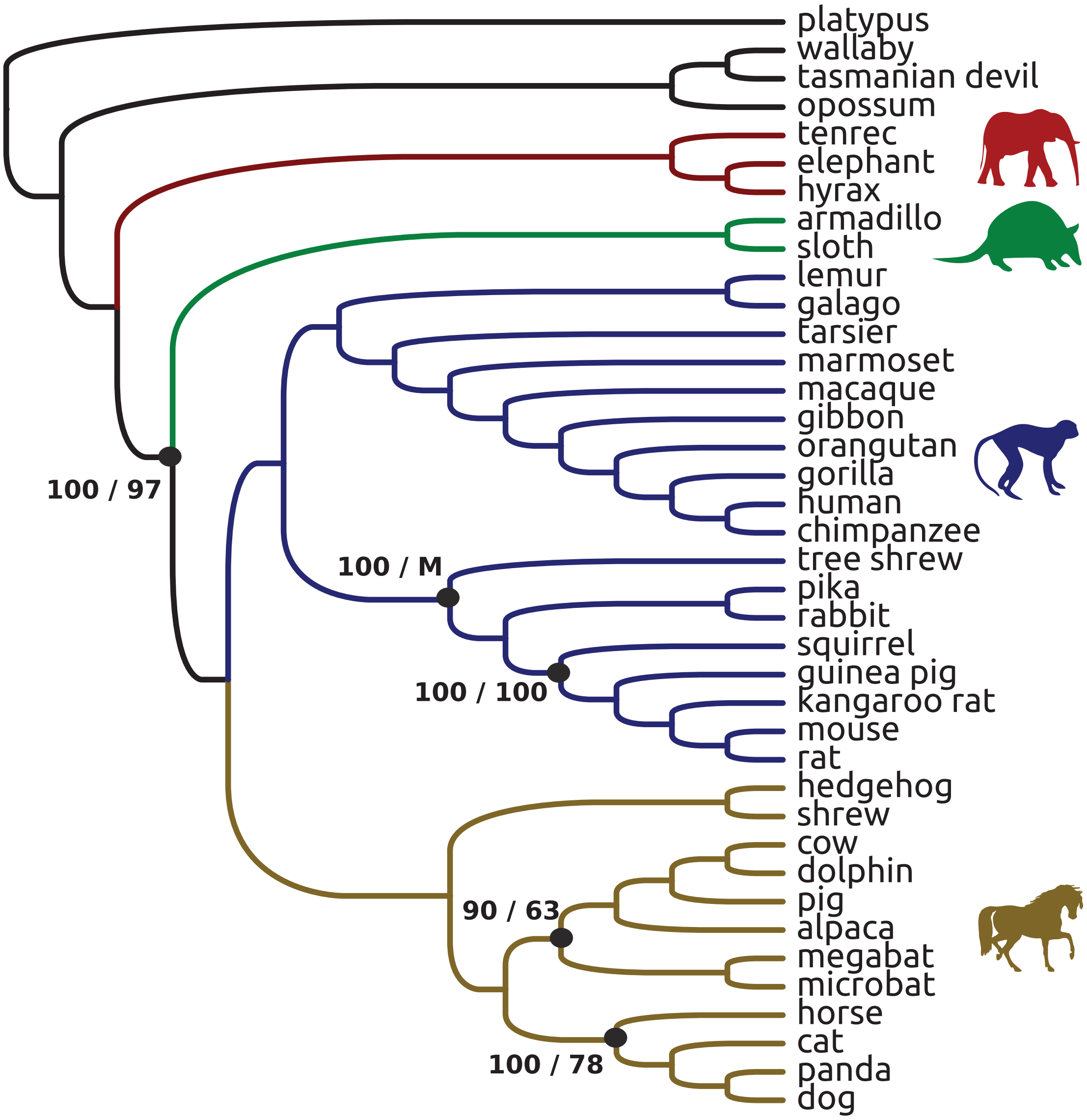


Figure 4 - Support values of AT-rich datasets for the debated nodes of placental phylogeny

The first bootstrap value is for the 1640-locus dataset, the second value for the 174-locus dataset. The boostraps values of all the other nodes are equal to 100.

more frequently during quick succession of speciation events [15, 53]. Agreeing with this statement, we report a stronger negative effect of GC-content on the Glires, Euarchontoglires, Laurasiatheria and Boreoeutheria nodes. Defining the higher taxa of placental mammals, these nodes are near their explosive radiation, making them prone to ILS.

More recent rapid speciation events, such as observed in hominids, seems poorly affected by GC-content. This contradicts partly a result of Hobolth et al. [27] which report positive correlation between incomplete lineage sorting and GC equilibrium. Note however that we found such a correlation in raw data, but the effect disappears after control for alignment gappiness. Furthermore, ILS involves only 1.6% of the human, chimp and orangutan genomes, and is believed to rarely affect their relationship in phylogeny [27]. This low proportion was found on the 3 whole genomes, including non-coding regions, while we use orthologous coding sequence regions for 39 species. Without non-coding regions, we certainly have a lack of data to detect a such low rate of ILS among hominoids.

Biased gene conversion leads to homoplasy

Incomplete lineage sorting does not seem sufficient to explain observed discrepancies among GC-rich gene trees. Here, we propose an extra hypothesis. An abundant litterature hypothesized that GC-content distribution is due to a neutral mechanism, the so-called biased gene conversion [42, 54–56]. According to this model, a bias in the DNA repair machinery would result in meiotic distortion favouring GC over AT alleles in highly recombination regions [57]. Increasing the GC-content of recombination hotspots, this mechanism would give birth to around 100kb regions of high average GC-content. Called GC-rich isochore [52], these regions present higher gene density, methylation rate and expression levels [58, 59]. Though important from a functional viewpoint, these regions could be penalized by biased gene conversion. Indeed, studies report that the neutral nature of this mechanism could counteract natural selection, and promote the fixation of deleterious mutations from A or T toward G or C [60–62]. However, recombination hotspots born and die quickly in these regions. This conducts GC-rich genes to undergo short biased gene conversion events followed by long periods where natural selection takes over again, and should promote compensatory mutations toward AT (partly repairing deleterious substitutions fixed after biased gene-conversion episodes). Under this dynamic pattern of substitution, sites are prone to alternate GC and AT states, depending whether biased gene conversion switch on or off. Such a turnover leads to multiple substitutions, which are known to blur phylogenetic signal. In our opinion, this homoplasy phenomenon could explain the higher topology error of GC-rich genes.

This point of view is in agreement with most of our results. Indeed, GC-content dynamic of mammals was described recently [43]. In this study, non-model species with very fast GC-content evolution were identified, among which a tenrec, lagomorphs (rabbit and pika) and a shrew. Interestingly, nodes implying these species are among the few recent nodes significantly affected by GC-content (Figure 2). This is particularly striking for the shrew, reported as the placental mammal the most influenced by biased gene conversion [43].

ILS and homoplasy induced by biased gene conversion are not exclusive hypotheses. In our opinion, long branches leading to recent nodes are particularly prone to homoplasy. On the other hand, short branches leading to deep nodes could be more strongly affected by ILS. Whatever the reason, the two explanations state that recombination is the main culprit.

In this regard, our real ability to estimate long-term recombination is crucial. As stated in introduction, GC-content induced by biased gene conversion was reported as the best predictor. However, biased gene conversion is supposed to need time to properly inprint the signature of a recombination hotspot. Consequently, GC-content better reflects recombination in case of higher hotspot stability. Interestingly, such a high hotspot stability was reported in Canidae, one of the nodes here the most affected by GC-content (Figure 2, dog + panda node). Indeed, dogs and their wild relatives exhibit a non-functional PRDM9, a protein involved in the short life cycle of recombination hotspots of all other mammals [63]. Because of this loss of function, GC-content reflect particularly well the recombination patterns of dog [64]. This fact probably explain the strong GC-effect found for the Caniformia node (dog + panda).

The root of Placentalia : Exafrotheria ?

As shown in Figure 3, GC-rich genes are under faster evolution. In phylogeny, it is well-known that such evolutionary rates can conduct to homoplasy and long-branch attraction [48]. Here, we suggest that these artefacts are by-products of biased gene conversion. Indeed, in addition to the theoretical proposition explained above, concrete example of such accelerating molecular evolution were reported. Inducing sudden burst of GC-increase (a striking example is provided by the *Fxy* gene of the mouse [60]), biased gene conversion is responsible of the fastest evolving regions of the human genome, misleading natural selection scans [54, 65]. Near the placental root, such episode increases the length of the already long Afrotheria and Xenarthra branches, previously reported prone to long-branch attraction [48]. Biased gene conversion increases this risk, and could lead to over-estimate confidence in their sister relationship, the so-called Atlantogenata clade. Thus, a strong Atlantogenata support was reported by most of phylogenomic studies [7, 17–21]. Our results show that, actually, this node is nearly only supported by the less-reliable phylogenetic markers of our genome : fast-evolving GC-rich genes, prone to recombination, ILS, homoplasy and topological errors on widely accepted nodes. On a genomic scale, averaging equally the signal of all genes leads to mixing support for the root of Placentalia. Taking into account the better reliability of AT-rich genes allow to distinguish their agreement to support the Exafrotheria hypothesis.

Interestingly, a recent study based on non-coding ultra-conserved elements gives similar conclusions [15]. In our opinion, this is mainly due to the fact that non-coding regions are concentrated in AT-rich regions, the recombination coldspot of mammalian genomes [41, 66]. Confirming this assumption, the GC content of the whole dataset of McCormack et al. is of 38% (917 loci), which is to be compared to the 55% of our 13 111 coding sequences, known to be often located near recombination hotspots [41, 66]. This difference could explain why similar coalescent methods [36] support Atlantogenata with coding sequences [22], whereas Exafrotheria is supported with non-coding sequences [15]. However, our AT-rich coding sequences are also nearly recombination free, and give the same support for Exafrotheria. They are probably the biggest source of reliable markers now available, sharing the abundancy of coding sequence alignments and the reliability of non coding ultra-conserved elements.

In order to unravel the trickiest nodes of the placental tree, we suggest to focus on these AT-rich genes for a larger taxon sampling (similar to [8]). Indeed, trying to cure existing dataset does not solve all nodes (see the tree shrew placement and the laurasiatherian orders), and relies on arbitrary threshold which are far from ideal. In addition to a taxon sampling increase, this issue could be addressed with improvement of current inference methods. It is well-known that gene trees could be regarded as contradicting testimonies of the past. Current coalescent methods [36, 37] try to resolve the conflicts, even when the stories differ strongly, whereas curated datasets ignore the claims of the least reliable witnesses. We think that the ideal solution is somewhere in-between : taking into account all the information available, but putting different weight on it. Who want to believe equally an obvious liar and a regular marker? Based on several criterion such as GC-content, coalescent methods should give more importance to more reliable gene trees. Reconstruction of ancestral substitutions along each branch of a tree [67, 68] can pinpoint local increase of GC-content, and could be used to infer more precisely which genes are unreliable in regard of which nodes. Such a combination of increased taxon sampling [8], improved coalescent methods [36, 37] and accurate sequence reliability measure is probably the key to produce a fully resolved placental tree, which does not change depending on loci, taxa or methods.

Conclusion

Through a genomic scale analysis, this study provides the first evidence of a strong base composition effect on the accuracy of a gene phylogeny with respect to the species phylogeny. Footsteps of recombination hotspots, GC-rich genes tend to give rise to erroneous species phylogenies, which is in agreement with well-studied phenomena such as ILS and biased gene conversion. This result is particularly relevant in the post-genomic era, where evolutionary biology is overwhelmed under massive and sometimes discordant datasets. We believe that GC-content is only the first from a series of clearly needed criterion to separate the wheat from the chaff, allowing phylogeny to make a new breakthrough in the resolution of the Tree of Life. Based on a wide genomic tendency, we suggest a possible resolution of the long-debated position of the placental root. Such a claim is based on a solid biological explanation for previous biaises : the combined effect of biased gene conversion and ILS. Allowing to explain recent disagreeing results [15, 22], the present study suggests that recombination hotspots, and then GC-rich sequences, were the plague that prevented us to identify Arotherians as the first offshot of the placental tree.

Methods

Dataset

Sequence alignments were provided by the OrthoMam v7 database [69], which contains curated 1-to-1 orthologous markers from the 39 mammalian genomes available on Ensembl, release 67 [24]. We used all the coding sequences available (13,111), which represent more than 24,000,000 sites. We also the maximum likelihood tree provided with each alignment in OrthoMam.

GC3 content

GC content is related to recombination because of biased gene conversion, a neutral mechanism [42, 54–56]. In order to avoid confounding effect of natural selection, we computed the GC percent on the third position of codons (GC3%). Mostly synonymous, substitutions on these positions are supposed to measure more accurately recombination. One GC3 content was computed for each alignment, thanks to home-made Python scripts. The GC3 content of an alignment is defined as the mean of the GC3 content of each taxon.

Control for gene length and number of species

The number of sites and species available for each gene could infuence the accuracy of its phylogeny. To distinguish the effect of GC content/recombination and an effect of the alignment gappiness, analyses were performed on two datasets : the original one made of alignments from OrthoMam, and the gap homogenized one. The latter dataset tries to homogenize the number of species and the number of sites along the GC3 gradient of our 13111 genes. First, we sort all genes accordingly to their GC3 content, in order to pair each gene with another : the least GC-rich gene is paired with the most GC-rich, the second least GC-rich gene is paired with the second most GC-rich, and so on. For each pair, when a site contains a missing character for a given species in one alignment, we replace the corresponding character in the other alignment by the same missing character. Thus, alignments of a pair have the same structure, with same species, same gaps and same missing data. Within this gap homogenized dataset, we control for the fact that a contrast between the AT and GC richest genes will not be due to a different amount of characters. As a counterpart, this homogenization introduces many gap in extreme AT-rich alignments to mimic the gappiness of GC-rich ones. Because it does not affect the contrast between AT and GC-rich genes, this does not bias our analysis toward an over-estimation of AT and GC-rich genes contrast. Actually, it even tends to underestimate the coefficient of the correlation between gene tree accuracies and GC-gradient.

A likelihood phylogenetic tree was re-computed for each gap homogenized alignments. Those trees have been inferred according to the OrthoMam pipeline [69]. Note that in 0.6% cases, alignments were too gappy and were rejected by the pipeline.

Gene tree subsets

We divide our 13 111 gene trees in 131 subsets acording to three different strategies.

In the first method, clusters of 100 gene trees have been built, according to the GC3 content of their alignments : the first cluster contains the 100 trees inferred from the 100 most AT-rich alignements, the last one contains the 100 trees inferred from the 100 most GC-rich. We called this the non-cumulative method, in opposition to methods 2 and 3.

In the second strategy, the first subset contains the 100 most AT-rich genes, the second subset the 200 most AT-rich and so on, until the 131th that contains all the gene trees. Hence those subsets contain more and more information, but also more and more trees from GC-rich genes. We called this the GC-cumulative method.

Similarly, in the AT-cumulative method (third method), subsets contain more and more information, but also more and more trees from AT-rich genes.

Reference tree

We compared all our gene trees with a reference species tree. This reference tree, based on litterature [8, 70], contains only nodes well-accepted by the scientific community (see topology of Figure 1). Debated nodes are the root of Placentalia (relationship among Afrotheria, Xenarthra and Boreoeutheria), relationships among Cetartiodactyla, Perissodactyla, Chiroptera and Carnivora, placement of Scandentia (Tupaia) in Euarchontoglires and placement of squirrel relative to murid and caviomorph rodents.

Measuring the gene tree error

We used several topological distances to measure the topological error of a gene tree : their percent of absent bipartitions, triplets and quartets compared to the resolved nodes of the reference species tree. Absent bipartitions were computed thanks to a part of the Robinson-foulds metrics (number of partitions implied by the reference tree but not in the gene tree) using a home-made program based on the Bio++ library [71]. Triplet and quartet distances were computed with the DQUAD program [72].

Supertree approach

To know which node is more affected by GC-content, we used a supertree approach. Thanks to the SuperTriplet program [72], we computed 131 supertrees for each subset of 100 trees. Our supertrees were constrained to the reference tree, in order to get a support value for each well-established nodes.

Supermatrix method

We compared topology and support values obtained from a supermatrix containing the 100 most AT-rich genes versus those obtained from a supermatrix containing the 100 most GC-rich genes. Concatenated alignments were used to infer the best maximum likelihood tree thanks to RAxML 7.2.8, GTR+CAT model [47, 73] and 100 bootstraps replicates. We used the same procedure for the 131 subsets of 100 genes (results of Figure 3), the 1640-locus dataset (all alignments below 40% of GC3-content with a triplet topology error proportion below 0.1), and the 175-locus dataset (all alignments below 40% of GC3 content containing the 39 species of this study).

References

1. Eisen JA, Fraser CM: **Phylogenomics: intersection of evolution and genomics.** *Science* 2003, **300**(5626):1706–1707, [http://www.ncbi.nlm.nih.gov/pubmed/12805538].

2. Delsuc F, Brinkmann H, Philippe H: **Phylogenomics and the reconstruction of the tree of life.** *Nature Reviews Genetics* 2005, **6**(5):361–375, [http://www.ncbi.nlm.nih.gov/pubmed/15861208].

3. Madsen O, Scally M, Douady CJ, Kao DJ, DeBry RW, Adkins R, Amrine HM, Stanhope MJ, De Jong WW, Springer MS: **Parallel adaptive radiations in two major clades of placental mammals.** *Nature* 2001, **409**(6820):610–614, [http://www.ncbi.nlm.nih.gov/pubmed/11214318].

4. Murphy WJ, Eizirik E, Johnson WE, Zhang YP, Ryder OA, O'Brien SJ: **Molecular phylogenetics and the origins of placental mammals.** *Nature* 2001, **409**(6820):614–618, [http://www.ncbi.nlm.nih.gov/pubmed/11214319].

5. Delsuc F, Scally M, Madsen O, Stanhope MJ, De Jong WW, Catzeflis FM, Springer MS, Douzery EJP: **Molecular phylogeny of living xenarthrans and the impact of character and taxon sampling on the placental tree rooting.** *Molecular Biology and Evolution* 2002, **19**(10):1656–71, [http://www.ncbi.nlm.nih.gov/pubmed/12270893].

6. Scally M, Madsen O, Douady CJ, Jong WWD, Stanhope MJ, Springer MS: **Molecular Evidence for the Major Clades of Placental Mammals**. *Journal of Mammalian Evolution* 2002, **8**(4):239–277, [http://www.springerlink.com/index/HV9FVWAVBJX4WCA0.pdf].

7. Prasad AB, Allard MW: **Confirming the phylogeny of mammals by use of large comparative sequence data sets**. *Molecular Biology and Evolution* 2008, **25**(9):1795.

8. Meredith RW, Janecka JE, Gatesy J, Ryder OA, Fisher CA, Teeling EC, Goodbla A, Eizirik E, Simão TLL, Stadler T, Rabosky DL, Honeycutt RL, Flynn JJ, Ingram CM, Steiner C, Williams TL, Robinson TJ, Burk-Herrick A, Westerman M, Ayoub NA, Springer MS, Murphy WJ: **Impacts of the Cretaceous Terrestrial Revolution and KPg Extinction on Mammal Diversification.** *Science* 2011, **334**(6055):521–4, [http://www.ncbi.nlm.nih.gov/pubmed/21940861].

9. Shoshani J, McKenna MC: **Higher taxonomic relationships among extant mammals based on morphology, with selected comparisons of results from molecular data.** *Molecular Phylogenetics and Evolution* 1998, **9**(3):572–584, [http://www.ncbi.nlm.nih.gov/pubmed/9668007].

10. Waddell PJ, Kishino H, Ota R: **A phylogenetic foundation for comparative mammalian genomics.** *Genome informatics International Conference on Genome Informatics* 2001, **12**(0919-9454 LA - eng PT - Journal Article SB - IM):141–154, [http://www.ncbi.nlm.nih.gov/pubmed/11791233].

11. Kriegs JO, Churakov G, Kiefmann M, Jordan U, Brosius J, Schmitz J: **Retroposed Elements as Archives for the Evolutionary History of Placental Mammals**. *PLoS Biology* 2006, **4**(4):e91, [http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1395351\&tool=pmcentrez\&rendertype=abstract].

12. Churakov G, Kriegs JO, Baertsch R, Zemann A, Brosius J, Schmitz J: **Mosaic retroposon insertion patterns in placental mammals**. *Genome Research* 2009, **19**(5):868–875, [http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2675975\&tool=pmcentrez\&rendertype=abstract].

13. Waddell PJ, Shelley S: **Evaluating placental interordinal phylogenies with novel sequences including RAG1, gamma-fibrinogen, ND6, and mt-tRNA, plus MCMC-driven nucleotide, amino acid, and codon models.** *Molecular Phylogenetics and Evolution* 2003, **28**(2):197–224, [http://www.sciencedirect.com/science/article/B6WNH-48S334P-1/2/5790b7b87439a5dd89442238f2945bef].

14. Nikolaev SI, Montoya-Burgos JI, Popadin K, Parand L, Margulies EH, Antonarakis SE: **Life-history traits drive the evolutionary rates of mammalian coding and noncoding genomic elements.** *Proceedings of the National Academy of Sciences of the United States of America* 2007, **104**(51):20443–8, [http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2154450\&tool=pmcentrez\&rendertype=abstract].

15. McCormack JE, Faircloth BC, Crawford NG, Gowaty PA, Brumfield RT, Glenn TC: **Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis.** *Genome research* 2012, [http://www.ncbi.nlm.nih.gov/pubmed/22207614].

16. Huchon D, Madsen O, Sibbald MJJB, Ament K, Stanhope MJ, Catzeflis F, De Jong WW, Douzery EJP: **Rodent phylogeny and a timescale for the evolution of Glires: evidence from an extensive taxon sampling using three nuclear genes.** *Molecular Biology and Evolution* 2002, **19**(7):1053–65, [http://www.ncbi.nlm.nih.gov/pubmed/12082125].

17. Murphy WJ, Pringle TH, Crider TA, Springer MS, Miller W: **Using genomic data to unravel the root of the placental mammal phylogeny**. *Genome Research* 2007, **17**(4):413–421, [http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1832088\&tool=pmcentrez\&rendertype=abstract].

18. Wildman DE, Uddin M, Opazo JC, Liu G, Lefort V, Guindon S, Gascuel O, Grossman LI, Romero R, Goodman M: **Genomics, biogeography, and the diversification of placental mammals.** *Proceedings of the National Academy of Sciences of the United States of America* 2007, **104**(36):14395–400, [http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1958817\&tool=pmcentrez\&rendertype=abstract].

19. Hallström BM, Kullberg M, Nilsson MA, Janke A: **Phylogenomic data analyses provide evidence that Xenarthra and Afrotheria are sister groups.** *Molecular Biology and Evolution* 2007, **24**(9):2059–2068, [http://www.ncbi.nlm.nih.gov/pubmed/17630282].

20. Kjer KM, Honeycutt RL: **Site specific rates of mitochondrial genomes and the phylogeny of eutheria**. *BMC Evolutionary Biology* 2007, **7**:8, [http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1796853\&tool=pmcentrez\&rendertype=abstract].

21. Hallström BM, Janke A: **Resolution among major placental mammal interordinal relationships with genome data imply that speciation influenced their earliest radiations**. *BMC Evolutionary Biology* 2008, **8**:162, [http://www.ncbi.nlm.nih.gov/pubmed/18505555].

22. Song S, Liu L, Edwards SV, Wu S: **Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model**. *Proceedings of the National Academy of Sciences* 2012, [http://www.pnas.org/cgi/doi/10.1073/pnas.1211733109].

23. Nishihara H, Maruyama S, Okada N: **Retroposon analysis and recent geological data suggest near-simultaneous divergence of the three superorders of mammals.** *Proceedings of the National Academy of Sciences of the United States of America* 2009, **106**(13):5235–40, [http://www.pnas.org/cgi/content/abstract/106/13/5235].

24. Birney E, Andrews TD, Bevan P, Caccamo M, Chen Y, Clarke L, Coates G, Cuff J, Curwen V, Cutts T, Down T, Eyras E, Fernandez-Suarez XM, Gane P, Gibbins B, Gilbert J, Hammond M, Hotz HR, Iyer V, Jekosch K, Kahari A, Kasprzyk A, Keefe D, Keenan S, Lehvaslaiho H, McVicker G, Melsopp C, Meidl P, Mongin E, Pettett R, Potter S, Proctor G, Rae M, Searle S, Slater G, Smedley D, Smith J, Spooner W, Stabenau A, Stalker J, Storey R, Ureta-Vidal A, Woodwark KC, Cameron G, Durbin R, Cox A, Hubbard T, Clamp M: **An Overview of Ensembl**. *Genome Research* 2004, **14**(5):925–928, [http://www.ncbi.nlm.nih.gov/pubmed/15078858].

25. Degnan JH, Rosenberg NA: **Discordance of Species Trees with Their Most Likely Gene Trees**. *PLoS Genetics* 2006, **2**(5):7, [http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1464820\&tool=pmcentrez\&rendertype=abstract].

26. Degnan JH, Rosenberg NA: **Gene tree discordance, phylogenetic inference and the multispecies coalescent.** *Trends in ecology & evolution* 2009, **24**(6):332–40, [http://dx.doi.org/10.1016/j.tree.2009.01.009].

27. Hobolth A, Dutheil J, Hawks J, Schierup M: **Incomplete lineage sorting patterns among human, chimpanzee, and orangutan suggest recent orangutan speciation and widespread selection**. *Genome Research* 2011, :1–8, [http://gb.cw.com.tw/m2m-0000/genome.cshlp.org/content/21/3/349.full].

28. Takahashi K, Terai Y, Nishida M, Okada N: **Phylogenetic relationships and ancient incomplete lineage sorting among cichlid fishes in Lake Tanganyika as revealed by analysis of the insertion of retroposons.** *Molecular Biology and Evolution* 2001, **18**(11):2057–2066, [http://www.ncbi.nlm.nih.gov/pubmed/11606702].

29. Jennings WB, Edwards SV: **Speciational history of Australian grass finches (Poephila) inferred from thirty gene trees.** *Evolution: International Journal of*

Organic Evolution 2005, **59**(9):2033–2047, [http://www.ncbi.nlm.nih.gov/pubmed/16261740].

30. Pollard DA, Iyer VN, Moses AM, Eisen MB: **Widespread discordance of gene trees with species tree in Drosophila: evidence for incomplete lineage sorting.** *PLoS genetics* 2006, **2**(10):e173, [http://dx.plos.org/10.1371/journal.pgen.0020173].

31. Carstens BC, Knowles LL: **Estimating species phylogeny from gene-tree probabilities despite incomplete lineage sorting: an example from Melanoplus grasshoppers.** *Systematic Biology* 2007, **56**(3):400–11, [http://www.ncbi.nlm.nih.gov/pubmed/17520504].

32. Matzke A, Churakov G, Berkes P, Arms EM, Kelsey D, Brosius J, Kriegs JO, Schmitz J: **Retroposon Insertion Patterns of Neoavian Birds: Strong Evidence for an Extensive Incomplete Lineage Sorting Era**. *Molecular Biology and Evolution* 2012, **29**(6):1–5, [http://www.ncbi.nlm.nih.gov/pubmed/22319163].

33. Chen FC, Li WH: **Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees.** *The American Journal of Human Genetics* 2001, **68**(2):444–56, [http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1235277\&tool=pmcentrez\&rendertype=abstract].

34. Patterson N, Richter DJ, Gnerre S, Lander ES, Reich D: **Genetic evidence for complex speciation of humans and chimpanzees.** *Nature* 2006, **441**(7097):1103–1108, [http://www.ncbi.nlm.nih.gov/pubmed/16710306].

35. Springer MS, Murphy WJ, Eizirik E, O'Brien SJ: **Placental mammal diversification and the Cretaceous-Tertiary boundary.** *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100**(3):1056–61, [http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=298725\&tool=pmcentrez\&rendertype=abstract].

36. Liu L, Yu L, Pearl DK, Edwards SV: **Estimating species phylogenies using coalescence times among sequences.** *Systematic Biology* 2009, **58**(5):468–477, [http://www.ncbi.nlm.nih.gov/pubmed/20525601].

37. Liu L, Yu L, Edwards SV: **A maximum pseudo-likelihood approach for estimating species trees under the coalescent model**. *BMC Evolutionary Biology* 2010, **10**:302, [http://www.biomedcentral.com/1471-2148/10/302].

38. Charlesworth B, Morgan MT, Charlesworth D: **The effect of deleterious mutations on neutral molecular variation.** *Genetics* 1993, **134**(4):1289–303, [http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1205596\&tool=pmcentrez\&rendertype=abstract].

39. Posada D, Crandall Ka: **The effect of recombination on the accuracy of phylogeny estimation.** *Journal of molecular evolution* 2002, **54**(3):396–402, [http://www.ncbi.nlm.nih.gov/pubmed/11847565].

40. Ruths D, Nakhleh L: **Recombination and phylogeny: effects and detection.** *International Journal of Bioinformatics Research and Applications* 2005, **1**(2):202–212.

41. Duret L, Arndt PF: **The impact of recombination on nucleotide substitutions in the human genome**. {*PLoS*} *genetics* 2008, **4**(5).

42. Galtier N, Piganeau G, Mouchiroud D, Duret L: **GC-content evolution in mammalian genomes: the biased gene conversion hypothesis.** 2001, [http://www.genetics.org/cgi/content/full/159/2/907?ijkey=1cd808d1057871dbd4b3a06943fe4da3da34398a\&keytype2=tf_ipsecsha].

43. Romiguier J, Ranwez V, Douzery EJP, Galtier N: **Contrasting GC-content dynamics across 33 mammalian genomes: Relationship with life-history traits and chromosome sizes.** *Genome research* 2010, :1001–1009, [http://www.ncbi.nlm.nih.gov/pubmed/20530252].

44. Aird D, Ross MG, Chen WS, Danielsson M, Fennell T, Russ C, Jaffe DB, Nusbaum C, Gnirke A: **Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries.** *Genome Biology* 2011, **12**(2):R18, [http://genomebiology.com/2011/12/2/R18].

45. Benjamini Y, Speed TP: **Summarizing and correcting the GC content bias in high-throughput sequencing.** *Nucleic Acids Research* 2012, **40**(10):1–14, [http://www.ncbi.nlm.nih.gov/pubmed/22323520].

46. Dabney J, Meyer M: **Length and GC-biases during sequencing library amplification: a comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries.** *BioTechniques* 2012, **52**(2):87–94, [http://www.ncbi.nlm.nih.gov/pubmed/22313406].

47. Stamatakis A: **RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models.** *Bioinformatics* 2006, **22**(21):2688–2690, [http://www.ncbi.nlm.nih.gov/pubmed/16928733].

48. Nishihara H, Okada N, Hasegawa M: **Rooting the eutherian tree: the power and pitfalls of phylogenomics**. *Genome Biology* 2007, **8**(9):R199, [http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2375037\&tool=pmcentrez\&rendertype=abstract].

49. Kryukov K, Sumiyama K, Ikeo K, Gojobori T, Saitou N: **A new database (GCD) on genome composition for eukaryote and prokaryote genome sequences and their initial analyses.** *Genome biology and evolution* 2012, **4**(4):501–12, [http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3342873\&tool=pmcentrez\&rendertype=abstract].

50. Oliver JL, Marín A: **A relationship between GC content and coding-sequence length.** *Journal of molecular evolution* 1996, **43**(3):216–23, [http://www.ncbi.nlm.nih.gov/pubmed/12794931].

51. Duret L, Mouchiroud D, Gautier C: **Statistical analysis of vertebrate sequences reveals that long genes are scarce in GC-rich isochores.** *Journal of Molecular Evolution* 1995, **40**(3):308–317, [http://www.springerlink.com/index/10.1007/BF00163235].

52. Bernardi G: **The Mosaic Genome of Warm-Blooded Vertebrates**. *Science* 1985, **228**(4702):953–958.

53. Maddison WP: **Gene Trees in Species Trees**. *Systematic Biology* 1997, **46**(3):523–536, [http://sysbio.oxfordjournals.org/cgi/doi/10.1093/sysbio/46.3.523].

54. Galtier N, Duret L: **Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution.** *Trends in genetics : TIG* 2007, **23**(6):273–7, [http://dx.doi.org/10.1016/j.tig.2007.03.011].

55. Lynch M, Walsh B: *The Origins of Genome Architecture*. Sinauer Associates Inc.,U.S. 2007, [http://www.amazon.fr/Origins-Genome-Architecture-Michael-Lynch/dp/0878934847].

56. Duret L: **Mutation patterns in the human genome: more variable than expected.** *PLoS biology* 2009, **7**(2):e1000028, [http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2634789\&tool=pmcentrez\&rendertype=abstract].

57. Eyre-Walker A: **Recombination and mammalian genome evolution.** *Proceedings of the Royal Society B: Biological Sciences* 1993, **252**(1335):237–243, [http://www.ncbi.nlm.nih.gov/pubmed/8394585].

58. Eyre-Walker A, Hurst LD: **The evolution of isochores**. *Nature Reviews Genetics* 2001, **2**(7):549–555, [http://dx.doi.org/10.1038/35080577].

59. Kudla G, Lipinski L, Caffin F, Helwak A, Zylicz M: **High Guanine and Cytosine Content Increases mRNA Levels in Mammalian Cells**. *PLoS Biology* 2006, **4**(6):e180, [http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1463026\&tool=pmcentrez\&rendertype=abstract].

60. Montoya-Burgos JI, Boursot P, Galtier N: **Recombination explains isochores in mammalian genomes**. *Trends in Genetics* 2003, **19**(3):128–130.

61. Galtier N, Duret L, Glémin S, Ranwez V: **GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates.** *Trends in Genetics* 2009, **25**:1–5, [http://www.ncbi.nlm.nih.gov/pubmed/19027980].

62. Necşulea A, Popa A, Cooper DN, Stenson PD, Mouchiroud D, Gautier C, Duret L: **Meiotic recombination favors the spreading of deleterious mutations in human populations.** *Human mutation* 2011, **32**(2):198–206, [http://www.ncbi.nlm.nih.gov/pubmed/21120948].

63. Muñoz Fuentes V, Di Rienzo A, Vilà C: **Prdm9, a Major Determinant of Meiotic Recombination Hotspots, Is Not Functional in Dogs and Their Wild Relatives, Wolves and Coyotes**. *PLoS ONE* 2011, **6**(11):e25498, [http://dx.plos.org/10.1371/journal.pone.0025498].

64. Axelsson E, Webster MT, Ratnakumar A, Ponting CP, Lindblad-Toh K: **Death of PRDM9 coincides with stabilization of the recombination landscape in the dog genome.** *Genome research* 2012, **22**:51–63, [http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3246206\&tool=pmcentrez\&rendertype=abstract].

65. Kostka D, Hubisz MJ, Siepel A, Pollard KS: **The role of GC-biased gene conversion in shaping the fastest evolving regions of the human genome.** *Molecular Biology and Evolution* 2011, **29**(3):1–35, [http://www.ncbi.nlm.nih.gov/pubmed/22075116].

66. Fullerton SM: **Local rates of recombination are positively correlated with GC content in the human genome**. *Mol. Biol. Evol.* 2001, :1139–1142, [http://www.ncbi.nlm.nih.gov/pubmed/22638613].

67. Romiguier J, Figuet E, Galtier N, Douzery EJP, Boussau B, Dutheil JY, Ranwez V: **Fast and Robust Characterization of Time-Heterogeneous Sequence Evolutionary Processes Using Substitution Mapping**. *PLoS ONE* 2012, **7**(3):e33852, [http://dx.plos.org/10.1371/journal.pone.0033852].

68. Dutheil JY, Galtier N, Romiguier J, Douzery EJ, Ranwez V, Boussau B: **Efficient selection of branch-specific models of sequence evolution.** *Molecular biology and evolution* 2012, **29**(7):1861–1874, [http://www.ncbi.nlm.nih.gov/pubmed/22319139].

69. Ranwez V, Delsuc F, Ranwez S, Belkhir K, Tilak MK, Douzery EJ: **OrthoMaM: A database of orthologous genomic markers for placental mammal phylogenetics**. *BMC Evolutionary Biology* 2007, **7**:241, [http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2249597\&tool=pmcentrez\&rendertype=abstract].

70. Springer MS, Stanhope MJ, Madsen O, de Jong WW: **Molecules consolidate the placental mammal tree.** *Trends in ecology & evolution* 2004, **19**(8):430–8, [http://www.ncbi.nlm.nih.gov/pubmed/16701301].

71. Dutheil J, Gaillard S, Bazin E, Glémin S, Ranwez V, Galtier N, Belkhir K: **Bio++: a set of C++ libraries for sequence analysis, phylogenetics, molecular evolution and population genetics**. *BMC Bioinformatics* 2006, **7**:188, [http://www.ncbi.nlm.nih.gov/pubmed/16594991].

72. Ranwez V, Criscuolo A, Douzery EJP: **SuperTriplets: a triplet-based supertree approach to phylogenomics**. *Bioinformatics* 2010, **26**(12):i115–i123, [http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2881381\&tool=pmcentrez\&rendertype=abstract].

73. Stamatakis A: **Phylogenetic models of rate heterogeneity: a high performance computing perspective**. *Proceedings 20th IEEE International Parallel Distributed Processing Symposium* 2006, **21**(4):8 pp., [http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1639535].

Chapitre 6

Methodological articles: Substitution mapping to characterize the heterogeneity of substitution processes

Fast and Robust Characterization of Time-Heterogeneous Sequence Evolutionary Processes Using Substitution Mapping

Jonathan Romiguier[1]*, Emeric Figuet[1], Nicolas Galtier[1], Emmanuel J. P. Douzery[1], Bastien Boussau[2,4], Julien Y. Dutheil[1]⁹, Vincent Ranwez[1,3]⁹

[1] Institut des Sciences de l'Evolution de Montpellier, CNRS-Université Montpellier 2, Montpellier, France, [2] Laboratoire de Biométrie et Biologie Evolutive, CNRS-Université Lyon 1, Villeurbanne, France, [3] Unité Mixte de Recherche Amélioration génétique et adaptation des plantes méditerranéennes et tropicales, Montpellier SupAgro, Montpellier, France, [4] Department of Integrative Biology, University of California, Berkeley, California, United States of America

Abstract

Genes and genomes do not evolve similarly in all branches of the tree of life. Detecting and characterizing the heterogeneity in time, and between lineages, of the nucleotide (or amino acid) substitution process is an important goal of current molecular evolutionary research. This task is typically achieved through the use of non-homogeneous models of sequence evolution, which being highly parametrized and computationally-demanding are not appropriate for large-scale analyses. Here we investigate an alternative methodological option based on probabilistic substitution mapping. The idea is to first reconstruct the substitutional history of each site of an alignment under a homogeneous model of sequence evolution, then to characterize variations in the substitution process across lineages based on substitution counts. Using simulated and published datasets, we demonstrate that probabilistic substitution mapping is robust in that it typically provides accurate reconstruction of sequence ancestry even when the true process is heterogeneous, but a homogeneous model is adopted. Consequently, we show that the new approach is essentially as efficient as and extremely faster than (up to 25 000 times) existing methods, thus paving the way for a systematic survey of substitution process heterogeneity across genes and lineages.

Citation: Romiguier J, Figuet E, Galtier N, Douzery EJP, Boussau B, et al. (2012) Fast and Robust Characterization of Time-Heterogeneous Sequence Evolutionary Processes Using Substitution Mapping. PLoS ONE 7(3): e33852. doi:10.1371/journal.pone.0033852

Editor: David Liberles, University of Wyoming, United States of America

Received December 16, 2011; **Accepted** February 22, 2012; **Published** March 27, 2012

Funding: This work was supported by an Agence Nationale de la Recherche grant (ANR-08-GENM-036-01 CoGeBi) and a European Research Council Advanced Grant (ERC 232971 PopPhyl) to NG. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: jonathan.romiguier@univ-montp2.fr

⁹ These authors contributed equally to this work.

Introduction

Mapping the history of nucleotide or amino-acid changes onto the evolutionary history of a gene, as depicted by a phylogenetic tree, is of central interest to researchers in molecular evolution. This procedure, called mutation or substitution mapping, is useful for characterizing the molecular evolutionary processes of DNA and protein sequences, and their variations across sites and lineages. Substitution mapping has been successfully applied to study various aspects of molecular evolution, including coevolution [1], [2], selective constraints in proteins [3], deviations from the molecular clock hypothesis [4], and changes in selective regimes [5]. Beyond this, substitution mapping has also enabled the implementation of a number of models that were otherwise intractable [6],[7].

Over the past 10 years, several inference methods have been developed to achieve substitution mapping. Formally, the problem is to identify, for every site in a sequence alignment, the kinds of character changes that occurred, and their location in the underlying phylogeny. So a substitution mapping method would take an alignment and a tree as input and return, as output, an estimate of the number/nature of substitutions that have occurred, for each site of the alignment and each branch of the tree. The "naive" substitution mapping procedure [8] involves first reconstructing all ancestral sequences at each node of the phylogenetic tree. Secondly, for each site, one substitution is mapped on a branch when two different states are observed for this site at the two extremities of the branch. The main drawback of such an approach is that it overlooks the uncertainty of the ancestral sequence inference.

Two improved mapping methods have been proposed: Bayesian Mutational Mapping (BMM, [9]) and Probabilistic Substitution Mapping (PSM, [1],[10]). They both use Markov chains to model the substitution process and account for the uncertainty in the ancestral states [11], [12]. BMM is a procedure that generates a substitution scenario compatible with the data, together with its associated likelihood. This procedure was not designed to produce human-readable substitution maps, but rather to integrate a statistic of interest over the set of possible substitution maps. Because it is a sampling procedure, BMM is fairly computer-expensive, although some more stable or efficient samplers have been proposed lately [13], [14], [15]. PSM is an analytical procedure, which computes the probability distribution of the number of substitutions that occurred at each site of the alignment and each branch of the phylogenetic tree. Dutheil et al

2005 [1] report how to compute the mean number of total substitutions per branch and site, but it is also possible to compute higher-order moments of the distribution, or distinguish between different types of substitutions ([14],[15] and the present study). PSM is a maximum likelihood solution of BMM for some particular statistics (the mean of the branch and site-specific distributions of the expected number of substitutions in the case of Dutheil et al 2005 [1]) and is therefore quite fast to compute for a given tree and substitution model, which is a significant advantage with respect to the increasing amount of molecular data provided by high-throughput sequencing. In addition, the relative simplicity and computer efficiency of substitution mapping procedures have promoted them for use in several analyses (e.g. [16]). They have been shown to facilitate parameter estimation of complex models when used within expectation-maximization procedures [17]. Adequate statistics based on substitution maps could therefore serve as straight-forward descriptors of molecular evolution that can be used as proxies for more complex ones.

One of the major advantages of substitution mapping is its power to detect and characterize time-heterogeneous processes, i.e. processes that vary across branches of the tree. Such variations, when identified, can be linked to variations in selective pressure (e.g. [18]) and mutation/fixation biases (e.g. [19]), or linked to macroscopic features of species such as effective population size (e.g. [20], [21]), ecological preferences [22] or life-history traits [23]. To detect heterogeneous processes, explicit models of non-homogeneous sequence evolution have been implemented in the maximum-likelihood or Bayesian frameworks [22], [24], [25], [26]. However, these parameter-rich models could lead to over-parametrization issues and are computationally demanding, so their usage is limited to relatively small subsets of the large amounts of currently available sequence data. Being fast and flexible, substitution mapping may potentially offer the opportunity to detect heterogeneous processes without fitting parameter-rich heterogeneous models. One possibility would be to map substitutions under a simple, fast, time-homogeneous model of sequence evolution, and then rely on the inferred changes to assess the heterogeneity of the evolutionary process, at low computational cost.

This, however, raises concern as to the ability of substitution mapping procedures to infer characteristics of the data which are not explicitly hard-coded in the model used for the mapping. This study is the first attempt to assess the extent to which substitution mapping is robust with respect to time-heterogeneous model choice. Using simulations under realistic non-homogeneous models of substitutions, both at the nucleotide and codon level, we demonstrate that probabilistic substitution mapping is robust to the *a priori* choice of substitution model. We show that even a homogeneous model with roughly approximate branch lengths captures most of the signal in the data, and allows to very efficiently infer complex aspects of the real process, including non-homogeneity. A dataset of 139 mammalian mitochondrial genomes and 243 ribosomal DNA sequences (18 S) from vertebrates is then used to illustrate the scalability of this method. Finally, we tested the method as a substitute to the famous *codeml* software from the PAML package [27] for large database analysis. Based on 993 vertebrate gene families from the *Selectome* database [28], we show that substitution mapping is a faster and better way to describe variations in the substitution process across each branch of a tree.

Results and Discussion

Analyses of simulations at the codon level

To test the robustness of substitution mapping, we propose to evaluate its ability to infer the dN/dS (non-synonymous/synonymous substitutions) ratio under various conditions. For this, we simulated 50 alignments of 1000 sites under a non-homogeneous YN98 codon model [25]. This model assumes a distinct omega value (dN/dS) for each branch. From a 33-leaf tree, we obtained 33 simulated sequences and 63 dN/dS ratios for each branch of the phylogeny. The tree topology was taken from reference [23]. Branch lengths were estimated from a real data set of 987 orthologous genes in 33 mammals obtained from the OrthoMam database [29] (Figure S1).

Substitution mapping was then used on these data. Directly inferred from sequences, this mapping provides synonymous and non-synonymous substitution counts, and allows deduction of a dN/dS ratio per branch. Since it is based on homogeneous models, this dN/dS estimation strategy does not require long optimization of multiple omega parameters. To test its model tolerance, substitution mapping was performed under different substitution models, i) a non-homogeneous YN98 model, the same used in simulations, ii) a homogeneous YN98 model, with a single omega value shared by all branches, iii) a Jukes Cantor model [30], where all substitution rates are fixed and equal.

Branch lengths and parameter values were re-estimated under these three models before substitution mapping computations. Because these re-estimations represent the largest percentage of the computation time, it is interesting to assess their real impact on substitution mapping. We thus added a fourth substitution mapping condition to test the robustness of substitution mapping with respect to branch length parameters. For this purpose, we performed substitution mapping under a homogeneous YN98 model with fixed branch lengths, randomly distorted values (in a range of+/−25% of their original value).

The simulation results are summarized in figure 1, and allow comparison of the dN/dS ratio deduced from substitution mapping with the real one resulting from the sequence simulation. Overall, the dN/dS ratio is well deduced by substitution mapping and estimated distributions are in the range of the corresponding simulated ones. Using a homogeneous or non-homogeneous model for the mapping seems to have very little effect. A single omega value shared by all branches already provides a correct estimation of the real molecular evolution process where the omega value varies across branches. Indeed, even the Jukes-Cantor model seems to provide reliable dN/dS estimation, despite the fact that it does not distinguish synonymous from non-synonymous substitution rates. Note, however, that it performs slightly worse than other models for very low omega values (<0.02). Since the Jukes-Cantor model assumes equal non-synonymous and synonymous substitution parameters, it tends to over-estimate dN when omega is $\ll 1$.

Another major result of this analysis is the very strong similarity between inferences, conducted with and without distorted branch lengths. For a homogeneous model, using correct branch length values does not improve the dN/dS inference by substitution mapping. This result promises very short computation time for substitution mapping, since a rough estimation of branch lengths is sufficient to obtain a reliable inference of molecular evolution processes.

Figure 2 shows the influence of the branch length (fig. 2A) and substitution rate (fig. 2B) on dN and dS estimation accuracy. Our substitution mapping estimations are evaluated according to their relative error in dN/dS, defined as follows:

$$\frac{\frac{dN_{map}}{dS_{map}} - \frac{dN_{sim}}{dS_{sim}}}{\frac{dN_{sim}}{dS_{sim}}}$$

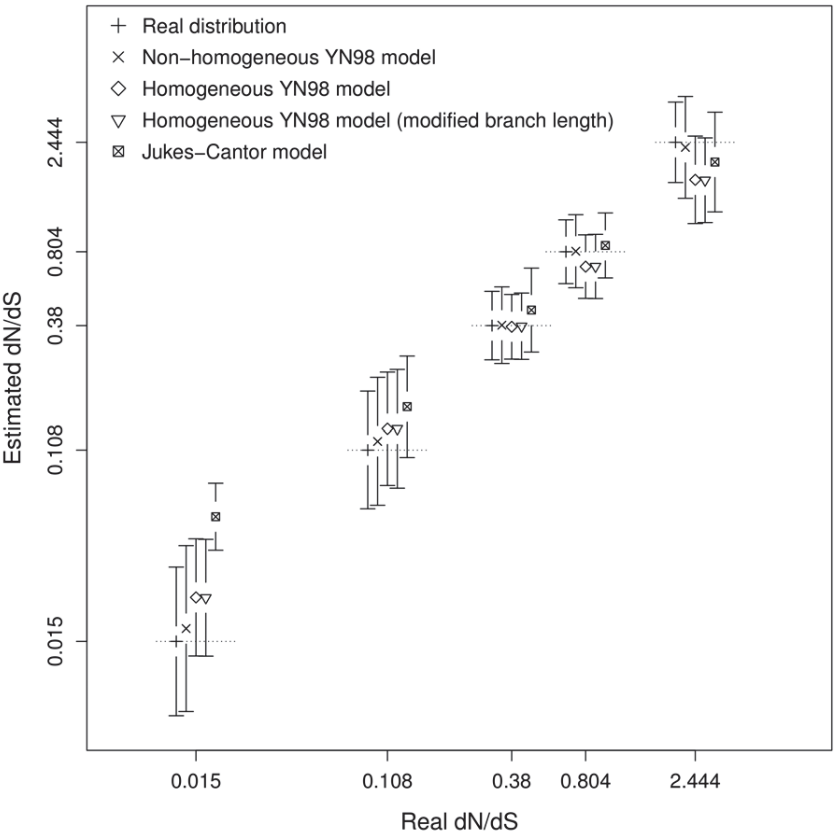


Figure 1. Estimated dN/dS under various substitution mapping conditions compared with real dN/dS from simulations. Axes are in log scale. Each point represents the median of one of the five classes. Bars represent the median absolute deviation of their corresponding class. The real distribution is plotted to give an idea of the initial variability of values to estimate.
doi:10.1371/journal.pone.0033852.g001

where $\frac{dN_{map}}{dS_{map}}$ is the dN/dS ratio obtained from mapping substitutions, and $\frac{dN_{sim}}{dS_{sim}}$ is the true dN/dS observed during simulations which generate sequences.

As we can see in figure 2A, branch lengths seem to have no major effect with YN98 models (non-homogeneous or homogeneous), but the Jukes-Cantor model tends to over-estimate dN/dS, and the relative error of this estimation increases with the branch length.

This effect could be explained by the occurrence of multiple substitutions. When a site presents a single substitution on a particular branch, the mapping method easily discriminates between a synonymous and a non-synonymous substitution, whatever the model used. On the other hand, in case of multiple substitutions, their distribution in non-synonymous or synonymous categories is mostly determined by the model, and forced to occur at an equal rate with the Jukes-Cantor model. As expected, this bias is stronger for long branches (figure 2A) and fast evolving sites (figure 2B). The Jukes-Cantor model clearly underestimates synonymous substitutions at sites where multiple substitutions take place. This effect also occurs with YN98 models, but affects sites much less. On the other hand, the Jukes-Cantor model does not seem to underestimate non-synonymous substitutions, even when there are 9 or more substitutions. This could be explained by the fact that the Jukes-Cantor model always overestimates dN because of its lack of omega value which implies equal synonymous and non-synonymous substitution probabilities. Note that a large majority of sites are subjected to fewer than 10 substitutions on the whole tree, a threshold where substitution mapping with a homogeneous YN98 model gives results similar to those obtained with a non-homogeneous model.

Note that there is no correlation between branch depth and relative error (Pearson's $R^2 = 0.0002$, p-value = 0.45 for homogeneous T92 model, similar results are obtained with other models).

In conclusion, the simulation results show very similar performances with respect to non-homogeneous and homogeneous substitution mapping for estimating the dN/dS ratio. With a single omega value, the homogeneous YN98 model captures the evolutionary history of sequences that evolved under distinct substitution processes. In spite of its simplicity, the Jukes-Cantor model shows comparable performances, except in the presence of long branches, where multiple substitutions can lead it to overestimate dN/dS ratio (by underestimating dS and overestimating dN). For highly divergent datasets, the homogeneous YN98 model therefore generates more accurate estimations, while still saving the time cost of a full-optimized non-homogeneous model.

Analysis of simulations at the nucleotide level

Simulation analyses were performed on GC-content patterns using a simulation protocol similar to that used for dN/dS. We simulated 50 000 sites under a non-homogeneous model [24], which assumes a different GC-equilibrium (theta value) per branch. We used the same topology (33 leaf tree) as that used for codon simulations.

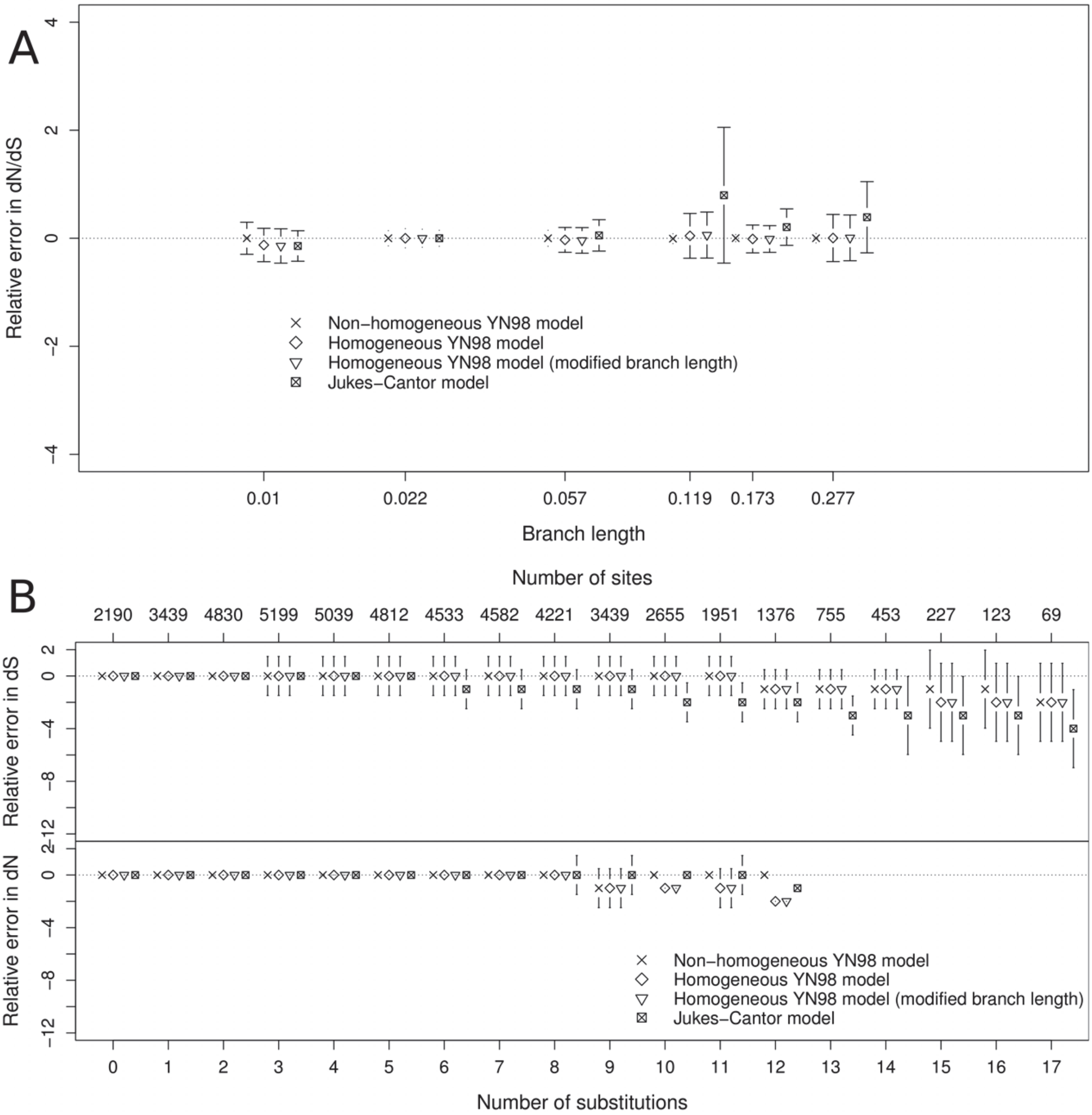


Figure 2. Relative error in dN/dS estimation by branch and by site. Panel A is the relative error variation in dN/dS estimation of a branch according to its branch length value. The branch length axis is in log scale. Each point represents the median of one of the six classes, bars represent the median absolute deviation of their corresponding class. Panel B is the relative error in dN and dS of a site according to its synonymous substitution number (top plot) and non-synonymous substitution number (bottom plot) in all the branches of the tree. Each point represents the median of one of the 17 classes (for readability, we do not show the 5 last classes that concern only 0.2% of our sites), bars represent the median absolute deviation of their corresponding class. The top axis gives the number of sites concerned by the corresponding number of synonymous substitutions.
doi:10.1371/journal.pone.0033852.g002

We then tested the ability of substitution mapping to infer the proportion of (A or T) to (G or C) substitution, under different conditions: i) a non-homogeneous model, ii) a homogeneous T92 model [31], iii) a homogeneous T92 model with a distorted branch length tree, iv) a homogeneous Jukes-Cantor model. These results (presented here as supplementary material, Figure S2 and Figure S3) led to conclusions similar to those of previous codon simulations: substitution mapping with a homogeneous model of sequence evolution is sufficient to recover variations in AT to GC and GC to AT substitutions throughout the phylogeny.

dN/dS estimation on data from mammalian mitochondria

The ability of substitution mapping to estimate heterogeneous substitution processes from a simple homogeneous model provides an opportunity to achieve a great speed boost for dN/dS inference.

For several reasons, this molecular evolutionary aspect has attracted considerable attention over the last years.

Because dN/dS is expected to be equal to 1 in case of neutral evolution, this ratio can give some clues about natural selection pressure on molecular evolution. Such traces of positive selection have been detected, revealing genetic differentiation to herbivory [27] or genes under selection in the human genome [32]. Furthermore, dN/dS provides a way to detect life history trait imprints on molecular sequences. According to the nearly neutral theory of molecular evolution [33], most non-synonymous substitutions are slightly deleterious. Because of stronger genetic drift, species with a small population size (*Ne*) are more prone to accumulate these slightly deleterious mutations. This less effective purifying selection leads to a higher dN/dS ratio in large mammal species which exhibit small *Ne* in comparison to small species [21]. Similar effects have been found between lineages with asexual or sexual reproductive modes [20], [34], [35].

These links between molecular evolution, natural selection and life history traits are used to an increasing extent in analyses with the dN/dS ratio, typically conducted with the highly popular *PAML* [27] or *HyPhy* [36] packages. Such programs infer a dN/dS ratio through non-homogeneous models, often with one-per-branch omega values, and can conduct likelihood comparison tests for statistical detection of positive selection. However, these procedures need to fit multiple parameters, which is not only computationally costly, but can also lead to poor estimations due to over-parameterization.

Our simulations show that substitution mapping using simple models with a single parameter is a relevant alternative. Substitution mapping provides one dN/dS ratio per branch without requiring complex multiple parameter optimizations, therefore providing the fastest and easiest way to perform dN/dS analyses. Furthermore, as substitution mapping is very robust to branch length errors, extra computation time can be saved using rough branch length optimization.

We used data similar to that used by Popadin et al [21] to further test this approach. From 139 mammalian mitochondrial genomes, they found positive correlations between dN/dS ratios (using *codeml*) and body mass, as expected through theoretical population size effects on purifying selection. Because estimating a dN/dS per branch was not feasible (due to computation time limitation), they conducted their analysis on 11 monophyletic subtrees. To assess the efficiency and the speed of our substitution mapping method, we used a similar dataset with more species (165), and estimated dN/dS ratios on the whole 165-tip tree.

Branch lengths and the single omega parameter value were estimated under an YN98 homogeneous model in BppML [37]. Since the simulations indicated a strong robustness to branch length errors, we estimated them with a low precision (optimization stopped when the improvements in log-likelihood were under 1 log likelihood point). We then used MapNH (software available at http://biopp.univ-montp2.fr/forge/testnh), to obtain dN and dS counts. The results of this analysis are shown in figure 3, where the estimated dN/dS of each terminal branch is compared to the body mass of the corresponding leaf. A strong significant correlation was found between these two variables ($p<0.0001$, Kendal's tau = 0.35) which is in agreement with the conclusions of Popadin et al and theoretical predictions. The total computation time of our substitution mapping procedure (including MapNH and prior estimation of parameters by BppML) was 1 h 34 min on

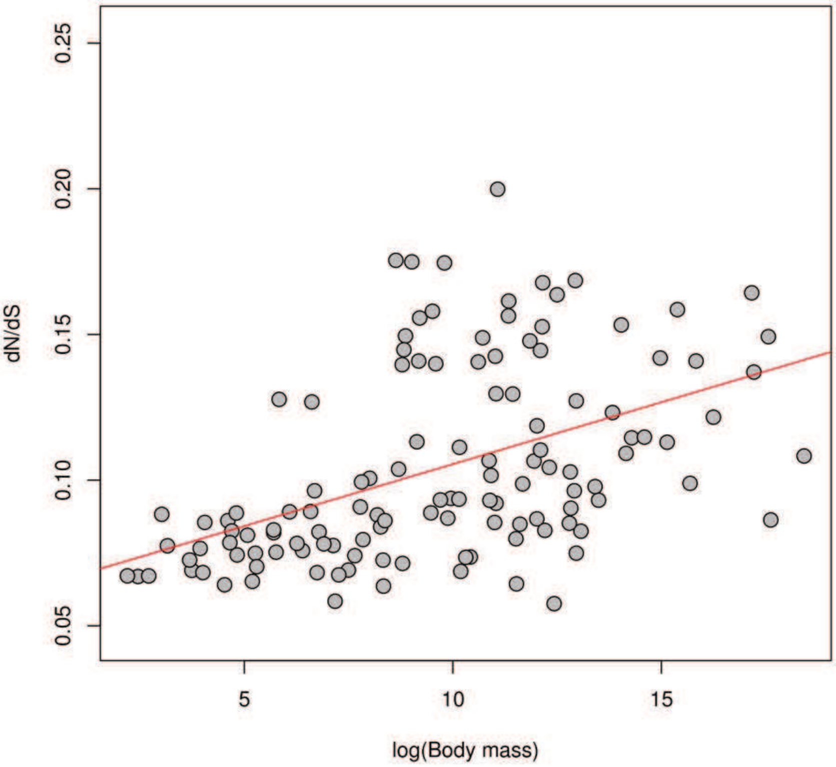


Figure 3. dN/dS ratio of terminal branches of 139 mammals according to their body mass.
doi:10.1371/journal.pone.0033852.g003

a desktop computer (Intel Xeon 2.27Ghz CPU) whereas, by comparison, fitting an equivalent non-homogeneous model with the PAML package required 38 days.

Substitution mapping thus seems to be a reliable tool to perform dN/dS ratio analyses, while being remarkably faster than classical approaches (more than 60 times). Moreover, as mentioned in the introduction, substitution mapping is theoretically not limited to dN/dS and can explore all types of substitution process heterogeneity, as illustrated in the following section where mapping is used to explore GC heterogeneity.

GC equilibrium estimation on data from vertebrate 18 S ribosomal RNA

Here, we tested the reliability of substitution mapping at the nucleotide level using a benchmark dataset with a documented variation in GC-content caused by biased gene conversion [38]. To evaluate the impact of biased gene conversion [39] on 18 S ribosomal RNA, Escobar et al [38] estimated the GC equilibrium of 243 vertebrate species through a non-homogeneous model. We used the alignments and tree provided with the article to conduct a mapping analysis (BppML parameter estimations followed by MapNH substitution mapping). We obtained (A or T) to (G or C) and (G or C) to (A or T) substitution counts for all branches of the tree, and computed a rough approximation of GC equilibrium defined by $\frac{AT\rightarrow GC}{AT\rightarrow GC+GC\rightarrow AT}$ where AT→GC and GC→AT are the substitution counts inferred by substitution mapping for each substitution type.

We compared these substitution mapping estimations to GC-equilibrium values of the internal branches shown by Escobar et al in their figure 2. Because of the lack of statistical power for small branches (too few substitution events to compute a reliable GC-equilibrium), we excluded all branches with fewer than 10 substitutions, and report a significant correlation (Pearson's $R^2 = 0.85$, p-value<0.0001) (figure 4). The total computation time was 44 s.

This comparison confirms that a simple homogeneous model used with substitution mapping can capture heterogeneous molecular evolution processes as well as massive time-consuming heterogeneous models. In this case, optimization through a non-homogeneous model takes more than 13 days, i.e. is more than 25 000 time slower than our substitution mapping approach. This huge gain of time, a lot better than for our codon case study, is probably due to various differences between the two datasets, particularly the number of species (here 243 compared to 165, i.e. more parameters per branch to estimate for the non-homogeneous model).

Substitution mapping for scalable genomic-scale selection analyses

As reviewed before, dN/dS computations and searches for traces of selection in substitution patterns are popular analyses used in recently published studies [20], [21], [32], [34], [35]. Consequently, several databases include such analyses on a large amount of data: Human PAML browser [40], The Adaptative Evolution Database [41], or Selectome [28]. All of these databases use *codeml* from the PAML package (Yang 1998). Because of the recent increase in available high throughput sequencing data, it will be increasingly difficult to quickly update such databases. Even with future increases in computer performance, it seems nearly impossible to imagine that one-per-branch non-homogeneous

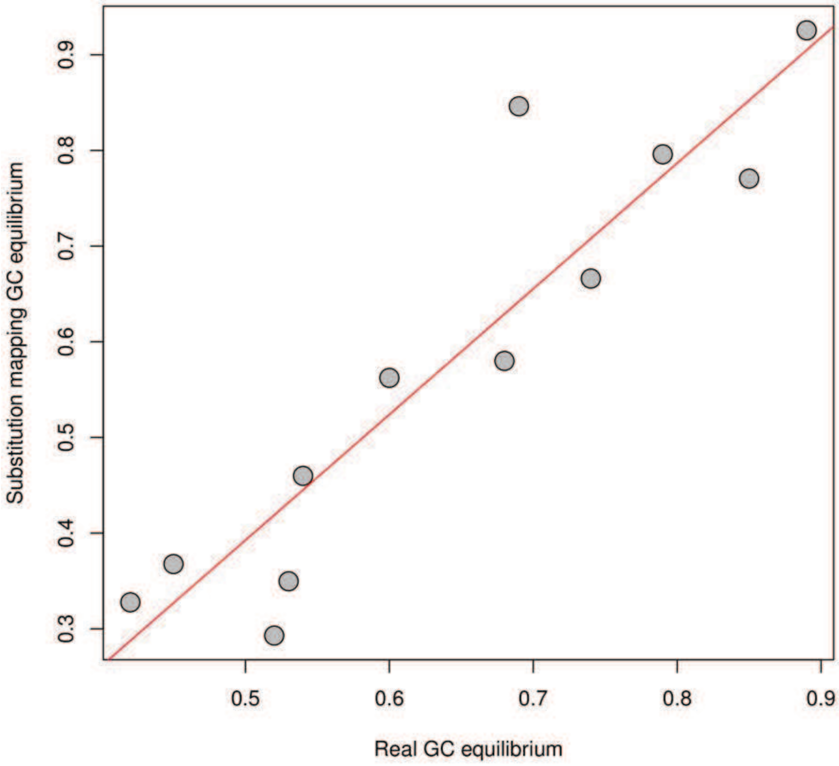


Figure 4. Estimation of equilibrium GC obtained by substitution mapping compared to equilibrium GC obtained with a non-homogeneous model (one GC equilibrium parameter per branch, values obtained from Escobar et al 2011).
doi:10.1371/journal.pone.0033852.g004

models could perform whole dN/dS analyses, e.g. in recent projects like 10 000 vertebrate genomes [42]. Despite the lack of likelihood comparison tests to detect accurately positive selection, fast and robust substitution mapping provides a scalable alternative to *codeml* for massive dN/dS analyses.

As a proof-of-concept, we used 993 gene alignments from the *Selectome* database, and compared dN/dS computations from *codeml* and substitution mapping. With a one-per-branch dN/dS parameter value, the CPU computation time of *codeml* was approximately 1 month (740 h). The same task took 10.72 h using substitution mapping (*BppML* and *MapNH*).

Figure 5 summarizes the PAML/mapping comparison for 30 643 dN/dS estimations (one per branch for 993 genes). dN/dS PAML estimations ranged from 0.003 to 3, and the heatmap highlights a high majority of correlated points. As seen before, this indicates that the two methods generated very similar results: 83.2% of our 30 643 dN/dS estimations are highly similar (less than 0.1 difference). Nevertheless, there is clearly a substantial number of outliers on both sides of this correlation line, revealing incongruent dN/dS values.

On the right of the plot, outliers are for a mainly due to extreme dN/dS values estimated by PAML, from 3 to over-represented aberrant dN/dS values near 1000 (2665 points). These erroneous values estimated by PAML are due to branches with very few substitution events, as illustrated by the two colored scales at the bottom, which indicate corresponding branch length values and the number of sites in the alignments. Under the same conditions, substitution mapping estimations were almost never above 3 (excepted for 310 infinite values, i.e. 0 synonymous substitutions detected), which makes more sense from a biological standpoint.

Outliers on the left of the plot correspond to very low dN/dS estimations by PAML (<0.002). As illustrated by the branch length color scale, these outliers are mainly due to aberrant branch length estimations obtained by PAML, with a median value ranging from 20 to 50. These extreme branch length values probably lead to overestimation of dS, and produce abnormally low dN/dS values for PAML. For those branches, substitution mapping probably benefits from its robustness to branch length estimation errors, and from the fact that those estimations are achieved through a more constrained homogeneous model. Whereas there are 693 branch length values above 10 with PAML, there are only 31 with BppML. This difference is probably caused by over-parametrization in the non-homogeneous PAML model, thus illustrating the technical difficulties that may arise when estimating numerous branch length values plus several substitution parameters simultaneously.

Outliers produced by substitution mapping are easily identifiable and correspond to 0 and infinite values, respectively, when there is 0 non-synonymous and 0 synonymous substitutions detected. They represent only 1.5% of points for plausible dN/dS values estimated by PAML (between 0.003 and 10), whereas PAML outliers (values lower than 0.003 and higher than 10 for plausible mapping values) represent 9% of points.

For our 30,643 estimations, 384 are greater than 1 with substitution mapping (74 without infinite values), and only 10 of these instances of positive selection are out of line with the PAML results. The extent of agreement between the two estimations (defined by the number of cases where PAML and substitution mapping agree on whether a dN/dS value is above or below 1, divided by the total number of estimations) is 0.92. PAML detects more positive selection events (3,029 versus 384 for mapping), but 88% of them are due to unreliable dN/dS values ranging from 3 to more than 1000. The substitution mapping is thus more stringent than PAML and seems to restrict the number of potential false detections of selection events, a key feature to help separate the wheat from the chaff in future studies of datasets with thousands of species.

Providing more reliable estimations, BppML+MapNH substitution mapping is able to perform a 1 month PAML task in less than half a-day. Fast, robust and stringent, substitution mapping appears to be tailored for dealing with future large dataset analyses.

Conclusion

Both simulations and real case studies clearly indicate that substitution mapping only requires a simple homogeneous model for branches and a phylogenetic tree with approximate branch lengths to achieve results similar to those generated by complex, non-homogeneous models.

Substitution mapping, which is more robust, not prone to over-parametrization problems and much faster (up to more than 25 000 times compared to non-homogeneous models), appears to be a good descriptor of molecular evolution. However, contrary to non-homogeneous models, it does not provide a framework for hypothesis testing though e.g. likelihood ratio tests.

Thanks to high throughput sequencing, biology has reached a new stage where commonly-used methods push computers to their limits. To overcome these limits, more efficient tools [16] are needed. Moreover, as large amounts of data go together with automated analysis pipelines, model misspecification and high false discovery rate become prominent concerns. In this respect, the application of robust statistics is particularly important. For these reasons, we consider that substitution mapping is one promising avenue for studies in molecular evolution, and could facilitate complex analyses such as the detection of natural selection in huge datasets that could not be dealt with using parameter-rich heterogeneous models.

Materials and Methods

Simulation

Simulations were performed with programs developed in C++ using the Bio++ libraries [43]. Initial branch lengths of the tree (same topology than [23]) were obtained from real data, i.e. 987 orthologous genes from 33 mammals obtained from the OrthoMam database [29]. We drew branch-specific Omega values from a gamma distribution of mean 0.2 (shape 0.5) and simulated 50 sequence alignments of 1000 sites each.

Third positions of the same orthologous genes were used for nucleotide simulations (GC analysis). Theta values (one per branch) were drawn from a uniform distribution of between 0 and 1.

Substitution mapping

The procedure is an extension of that described by Dutheil et al [1], which only provides the mean number of substitutions per branch and per site in the alignment. Here, we also estimated the detail counts for each type of substitution (e.g. synonymous or non-synonymous, AT→GC or GC→AT).

At each position i in the alignment, we computed the substitution vector $(v^s_{i,l},\ldots,v^s_{i,b},\ldots,v^s_{i,m})$ where $v^s_{i,b}$ is the posterior estimate of the number of substitutions of type s that occurred on branch b for m branches in the tree. $v^s_{i,b}$ was estimated by averaging all possible ancestral states at top x_p and bottom x_q nodes [1]:

$$v^s_{i,b} = \sum_{x_p}\sum_{x_q}\Pr(x_p,x_q|D_i,\theta)\times n^s_{x_p,x_q}(t)$$

Given the data and parameters, is the joint probability of having state x_p at the bottom node and state x_p at the top node. It is

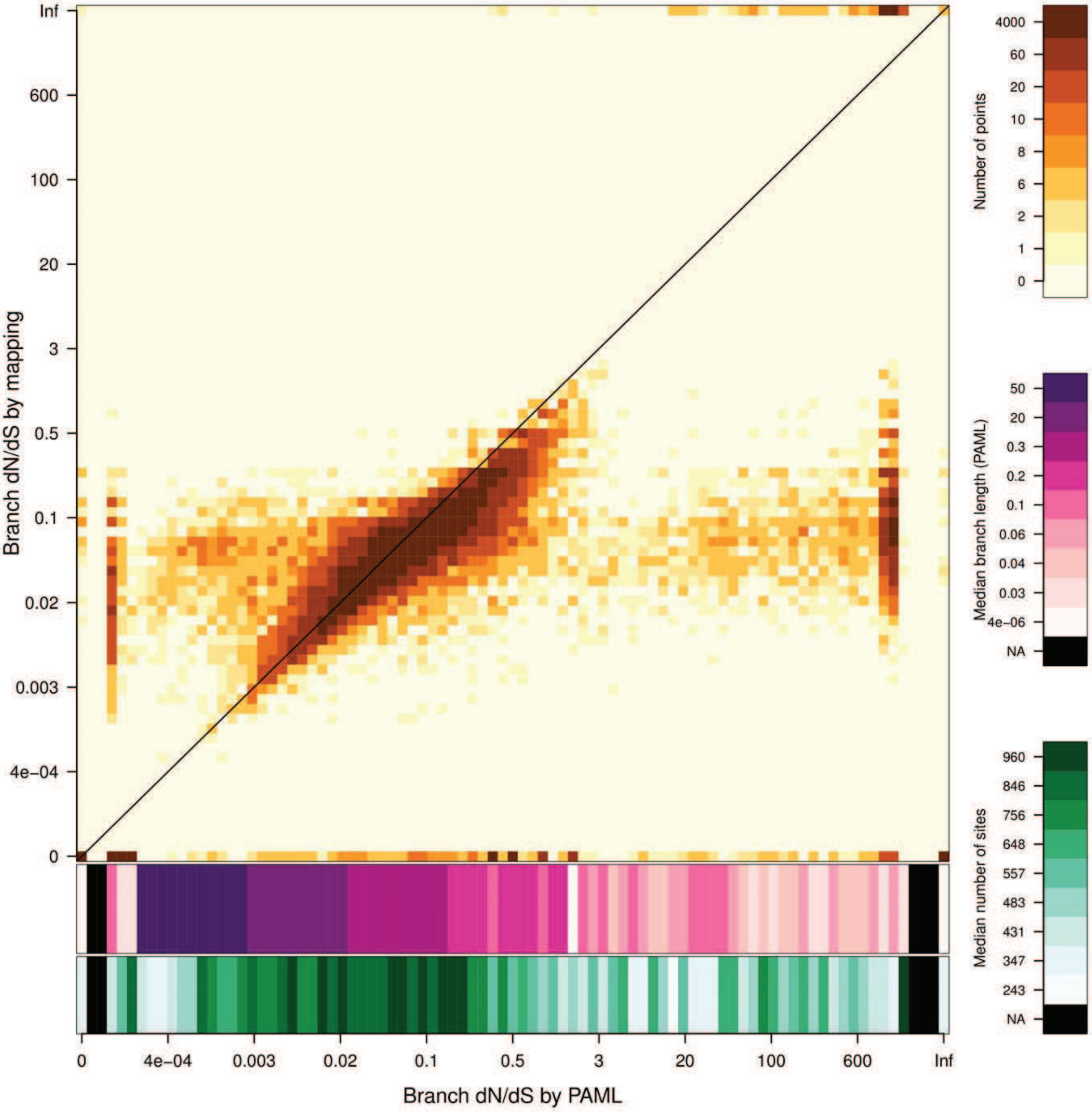


Figure 5. Estimation of dN/dS values by substitution mapping compared with dN/dS values estimated by PAML. Axes are in log scale. The orange color gradient represents the density of points. The two scales in purple and green show the median branch length (estimated by PAML) and the median number of sites for corresponding dN/dS values estimated by PAML. The 0 value includes estimations where dN is equal to 0 and estimations where dN and dS are both equal to 0 (0/0).
doi:10.1371/journal.pone.0033852.g005

computed as follows [1], [44], [45]:

$$\Pr(x_p,x_q|D_i,\theta)=\frac{\Pr(x_p,x_q|D_i,\theta)}{\Pr(D_i|\theta)}$$

The denominator is the likelihood for site i [46], while the numerator is obtained in a very similar way, but considering the ancestral states x_p and x_q as known in the Felsenstein recursion. Given the initial state x_p and final state x_q, term $n^s_{x_p,x_q}(t)$ is the mean number of substitutions of type s that occurs on a branch of length t. To compute this mean number, we used the uniformization method [14], [15] as it is exact, numerically more stable than the method proposed in Dutheil et al [1], and provides counts for each type of substitution.

Substitution counts obtained for each site were summed, then pooled depending on the type of data: AT→GC and GC→AT for nucleotides, and synonymous vs non-synonymous for codon sequences. On amino-acid sequences, we could have pooled substitutions in conservative vs non-conservative substitutions.

Substitution mapping was implemented in the MapNH program built using Bio++ libraries (Dutheil et al., 2006). The uniformization method for counting substitution was implemented from the R code kindly provided by Tataru and Hobolth [15], and made available to the libraries (version 2.0.2 and later). MapNH is available in the TestNH package at http://biopp.univ-montp2.fr/forge/testnh, as source code and binary versions. The program is written in standard C++ and it compiles in Linux, Windows and MacOS systems, and depends only on Bio++ libraries.

Mitochondrial genomes of mammals

Mitochondrial genomes for the 139 mammals were obtained from NCBI. Body mass values were obtained from the AnAge database [47].

Statistics

Statistical analyses were performed and graphs drawn up with the R software [48].

We used median absolute deviation (defined as $median(|X_i - median(X)|)$) instead of 95% intervals to measure dispersion in Figure 1 and 2. 95% intervals were highly dominated by outliers, and the resulting measure of dispersion reflected false distinction between model performances.

Supporting Information

Figure S1 Tree used in simulations. Initial branch lengths of the tree (same topology than [23]) were obtained from real data, i.e. 987 orthologous genes from 33 mammals obtained in the OrthoMam database [29].
(TIF)

Figure S2 Estimated GC equilibrium approximation under various substitution mapping conditions compared with real GC equilibrium approximation from simulations. GC equilibrium approximation is defined by, where AT→GC and GC→AT are number of substitutions of each type inferred by substitution mapping. Each point represents the median of one of the five classes, confidence intervals are the median absolute deviation of their corresponding class. The real distribution is plotted to give an idea of the initial variability in the values to estimate.
(TIF)

Figure S3 Relative error in GC equilibrium estimation by branch and by site. Panel A is the relative error variation in GC equilibrium approximation of a branch according to its length. Each point represents the median of one of the six classes, bars represent the median absolute deviation of its corresponding class. Panel B is relative error in GC equilibrium approximation of a site according to its AT to GC substitution number (top plot), and GC to AT substitution number (bottom plot). Each point represents the median of one of the 17 classes, bars represent the median absolute deviation of its corresponding class. The top axis gives the number of sites concerned by the corresponding number of substitutions.
(TIF)

Acknowledgments

The authors would like to thank Asger Hobolth and Paula Tataru for sharing code on the computation of the conditional number of jumps of a Markov chain.

Sebastien Moretti provided sequence alignments from the Selectome database.

Author Contributions

Conceived and designed the experiments: NG EJPD VR JR JD BB. Performed the experiments: JR EF. Analyzed the data: JR EF JD. Contributed reagents/materials/analysis tools: JD BB JR VR. Wrote the paper: JR JD VR NG EJPD.

References

1. Dutheil J, Pupko T, Jean-Marie A, Galtier N (2005) A model-based approach for detecting coevolving positions in a molecule. Molecular biology and evolution 22: 1919–1928.
2. Dimmic MW, Hubisz MJ, Bustamante CD, Nielsen R (2005) Detecting coevolving amino acid sites using Bayesian mutational mapping. Bioinformatics 21 Suppl 1: i126–i135.
3. Dutheil J (2008) Detecting site-specific biochemical constraints through substitution mapping. Journal of molecular evolution 67: 257–265. doi:10.1007/s00239-008-9139-8.
4. Mayrose I, Otto SP (2011) A likelihood method for detecting trait-dependent shifts in the rate of molecular evolution. Molecular biology and evolution 28: 759–770. doi:10.1093/molbev/msq263.
5. Zhai W, Slatkin M, Nielsen R (2007) Exploring Variation in the dN/dS Ratio Among Sites and Lineages Using Mutational Mappings: Applications to the Influenza Virus. Journal of Molecular Evolution 65: 340–348-348. doi:10.1007/s00239-007-9019-7.
6. Lartillot N (2006) Conjugate Gibbs sampling for Bayesian phylogenetic models. Journal of computational biology a journal of computational molecular cell biology 13: 1701–1722.
7. Robinson DM, Jones DT, Kishino H, Goldman N, Thorne JL (2003) Protein evolution with dependence among codons due to tertiary structure. Molecular Biology and Evolution 20: 1692–1704.
8. Shindyalov IN, Kolchanov NA, Sander C (1994) Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? Protein Engineering 7: 349–358.
9. Nielsen R (2002) Mapping mutations on phylogenies. Systematic biology 51: 729–739. doi:10.1080/10635150290102393.
10. Dutheil J, Galtier N (2007) Detecting groups of coevolving positions in a molecule: a clustering approach. BMC Evolutionary Biology 7: 242.
11. Felsenstein J (2004) Inferring Phylogenies. Sinauer As. Sinauer Associates Inc., U.S.
12. Yang Z (2006) Computational molecular evolution. Oxford University Press.
13. Rodrigue N, Philippe H, Lartillot N (2008) Uniformization for sampling realizations of Markov processes: applications to Bayesian implementations of codon substitution models. Bioinformatics 24: 56–62.
14. Hobolth A, Stone EA (2009) Simulation from endpoint-conditioned, continuous-time Markov chains on a finite state space, with applications to molecular evolution. The Annals of Applied Statistics.
15. Tataru P, Hobolth A (2011) Comparison of methods for calculating conditional expectations of sufficient statistics for continuous time Markov chains. BMC Bioinformatics 12: 465.
16. Minin VN, Suchard Ma (2008) Fast, accurate and simulation-free stochastic mapping. Philosophical transactions of the Royal Society of London Series B, Biological sciences 363: 3985–3995. doi:10.1098/rstb.2008.0176.
17. Hobolth A, Jensen J (2005) Applications of hidden Markov models for comparative gene structure prediction. J Comput Biology 12: 186–203.
18. Jobson RW, Nabholz B, Galtier N (2010) An evolutionary genome scan for longevity-related natural selection in mammals. Molecular Biology and Evolution 27: 840–847.
19. Duret L, Arndt PF (2008) The impact of recombination on nucleotide substitutions in the human genome. PLoS genetics 4.
20. Paland S, Lynch M (2006) Transitions to asexuality result in excess amino acid substitutions. Science (New York, N Y) 311: 990–992. doi:10.1126/science.1118152.
21. Popadin K, Polishchuk LV, Mamirova L, Knorre D, Gunbin K (2007) Accumulation of slightly deleterious mutations in mitochondrial protein-coding genes of large versus small mammals. Proceedings of the National Academy of Sciences of the United States of America 104: 13390–13395. doi:10.1073/pnas.0701256104.
22. Boussau B, Blanquart S, Necsulea A, Lartillot N, Gouy M (2008) Parallel adaptations to high temperatures in the Archaean eon. Nature 456: 942–945. doi:10.1038/nature07393.
23. Romiguier J, Ranwez V, Douzery EJP, Galtier N (2010) Contrasting GC-content dynamics across 33 mammalian genomes: Relationship with life-history traits and chromosome sizes. Genome research. pp 1001–1009. doi:10.1101/gr.104372.109.
24. Galtier N, Gouy M (1998) Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. Molecular biology and evolution 15: 871–879.

25. Nielsen R, Yang Z (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148: 929–936.
26. Lartillot N, Poujol R (2011) A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters. Molecular Biology and Evolution 28: 729–744.
27. Yang Z (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Molecular biology and evolution 15: 568–573.
28. Proux E, Studer RA, Moretti S, Robinson-Rechavi M (2009) Selectome: a database of positive selection. Nucleic Acids Research 37: D404–D407.
29. Ranwez V, Delsuc F, Ranwez S, Belkhir K, Tilak M-K, et al. (2007) {OrthoMaM:} A database of orthologous genomic markers for placental mammal phylogenetics. {BMC} Evolutionary Biology 7: 241. doi:10.1186/1471-2148-7-241.
30. Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: Munro HN, ed. Mammalian Protein Metabolism, Academic Press, Vol. 3. pp 21–132.
31. Tamura K (1992) Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C content biases. Molecular Biology and Evolution 9: 678–687.
32. Pollard KS, Salama SR, Lambert N, Lambot M-A, Coppens S, et al. (2006) An RNA gene expressed during cortical development evolved rapidly in humans. Nature 443: 167–172. doi:10.1038/nature05113.
33. Ohta T (1973) Slightly deleterious mutant substitutions in evolution. Nature 246: 96–98.
34. Barraclough TG, Fontaneto D, Ricci C, Herniou EA (2007) Evidence for inefficient selection against deleterious mutations in cytochrome oxidase I of asexual bdelloid rotifers. Molecular Biology and Evolution 24: 1952–1962.
35. Neiman M, Hehman G, Miller JT, Logsdon JM, Taylor DR (2010) Accelerated mutation accumulation in asexual lineages of a freshwater snail. Molecular biology and evolution 27: 954–963. doi:10.1093/molbev/msp300.
36. Pond SLK, Frost SDW, Muse SV (2005) HyPhy: hypothesis testing using phylogenies. Bioinformatics 21: 676–679. doi:10.1093/bioinformatics/bti079.
37. Dutheil J, Boussau B (2008) Non-homogeneous models of sequence evolution in the Bio++ suite of libraries and programs. BMC Evolutionary Biology 8: 255.
38. Escobar JS, Glémin S, Galtier N (2011) GC-Biased Gene Conversion Impacts Ribosomal DNA Evolution in Vertebrates, Angiosperms and Other Eukaryotes. Molecular Biology.
39. Duret L, Galtier N (2009) Biased Gene Conversion and the Evolution of Mammalian Genomic Landscapes. Annual Review of Genomics and Human Genetics 10: 285–311. doi:10.1146/annurev-genom-082908-150001.
40. Nickel GC, Tefft D, Adams MD (2008) Human PAML browser: a database of positive selection on human genes using phylogenetic methods. Nucleic acids research 36: D800.
41. Liberles D, Schreiber D, Govindarajan S, Chamberlin S, Benner S (2001) The Adaptive Evolution Database (TAED). Genome Biology 2.
42. Genome 10K Community of Scientists (2009) Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. The Journal of heredity 100: 659–674. doi:10.1093/jhered/esp086.
43. Dutheil J, Gaillard S, Bazin E, Glémin S, Ranwez V, et al. (2006) Bio++: a set of C++ libraries for sequence analysis, phylogenetics, molecular evolution and population genetics. BMC Bioinformatics.
44. Pupko T, Sharan R, Hasegawa M, Shamir R, Graur D (2003) Detecting excess radical replacements in phylogenetic trees. Gene 319: 127–135.
45. Galtier N, Boursot P (2000) A new method for locating changes in a tree reveals distinct nucleotide polymorphism vs divergence patterns in mouse mitochondrial control region. Journal of Molecular Evolution 50: 224–231. doi:10.1007/s002399910025.
46. Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. Journal of Molecular Evolution 17: 368–376.
47. De Magalhaes J, Costa J (2009) A database of vertebrate longevity records and their relation to other life-history traits. Journal of Evolutionary Biology 22: 1770–1774. doi:10.1111/j.1420-9101.2009.01783.x.
48. Team RDC (2008) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing 1: ISBN 3–900051-07–0.

Efficient Selection of Branch-Specific Models of Sequence Evolution

Julien Y. Dutheil,*,[1] Nicolas Galtier,[1] Jonathan Romiguier,[1] Emmanuel J.P. Douzery,[1] Vincent Ranwez,[1,2] and Bastien Boussau[3,4]

[1]Institut des Sciences de l'Évolution—Montpellier, Université Montpellier 2, Montpellier, France
[2]Montpellier SupAgro, UMR AGAP, Montpellier, France
[3]Laboratoire de Biométrie et Biologie Évolutive, Université de Lyon, Université Lyon 1, CNRS, UMR5558, Villeurbanne, France
[4]Department of Integrative Biology, University of California Berkeley

***Corresponding author:** E-mail: julien.dutheil@univ-montp2.fr.

Associate editor: Jeffrey Throne

Abstract

The analysis of extant sequences shows that molecular evolution has been heterogeneous through time and among lineages. However, for a given sequence alignment, it is often difficult to uncover what factors caused this heterogeneity. In fact, identifying and characterizing heterogeneous patterns of molecular evolution along a phylogenetic tree is very challenging, for lack of appropriate methods. Users either have to a priori define groups of branches along which they believe molecular evolution has been similar or have to allow each branch to have its own pattern of molecular evolution. The first approach assumes prior knowledge that is seldom available, and the second requires estimating an unreasonably large number of parameters. Here we propose a convenient and reliable approach where branches get clustered by their pattern of molecular evolution alone, with no need for prior knowledge about the data set under study. Model selection is achieved in a statistical framework and therefore avoids overparameterization. We rely on substitution mapping for efficiency and present two clustering approaches, depending on whether or not we expect neighbouring branches to share more similar patterns of sequence evolution than distant branches. We validate our method on simulations and test it on four previously published data sets. We find that our method correctly groups branches sharing similar equilibrium GC contents in a data set of ribosomal RNAs and recovers expected footprints of selection through dN/dS. Importantly, it also uncovers a new pattern of relaxed selection in a phylogeny of Mantellid frogs, which we are able to correlate to life-history traits. This shows that our programs should be very useful to study patterns of molecular evolution and reveal new correlations between sequence and species evolution. Our programs can run on DNA, RNA, codon, or amino acid sequences with a large set of possible models of substitutions and are available at http://biopp.univ-montp2.fr/forge/testnh.

Key words: molecular phylogenetics, maximum likelihood, ancestral character reconstruction, dN/dS, paml, selection.

Introduction

Living organisms show a striking diversity in size, life history, ecology, population structure, physiology, and cellular biology. This diversity propagates to the genome level: substantial between-species variations in base or amino acid compositions and in rates of sequence evolution have been documented (Hickey and Singer 2004; Bromham 2009; Lartillot and Poujol 2011). Understanding the links between these two kinds of variations—phenotypic versus genomic—is an important goal of current molecular evolutionary research (Boussau and Daubin 2010).

A popular approach to this question is to correlate phenotypic and sequence evolution across the branches of a phylogenetic tree (Yang 1998; Paland and Lynch 2006; Boussau et al. 2008). This typically requires identifying groups of branches (e.g., subtrees) sharing a common molecular evolutionary process. Usually this branch-clustering step is performed a priori and reflects existing knowledge (or hypothesis) about the major factors affecting sequence evolution in the group of organisms under study.

For instance, many studies of lineage-specific variations in selective pressure rely on PAML (Yang 2007) and use one of two distinct procedures: 1) the user either a priori defines clusters of branches that are expected to show a similar ratio of nonsynonymous to synonymous codon substitutions (dN/dS) or 2) assumes that each branch has an idiosyncratic dN/dS. The second option is discouraged by the PAML manual as it requires estimating one parameter per branch of the tree, which tends to be unstable when the tree is large. The first approach can only be applied in cases where prior knowledge, for instance on the phenotype of organisms, is available to cluster branches. Even in these few cases where phenotypes of extant organisms are well known and can be assumed to drive sequence evolution, it is often difficult to determine a priori how internal branches should be clustered as they correspond to organisms whose phenotypes can no longer be observed. An alternative approach is therefore desirable, which would avoid overparameterization and arbitrary grouping of branches prior to sequence analysis. With the increasing facility for gathering sequence data in living organisms of any sort, the tree-partitioning issue has lately become prominent in the molecular evolutionary literature and has motivated specific methodological developments (Jayaswal et al. 2011, Zhang et al. 2011).

We present a statistical approach to cluster branches in a phylogenetic tree using only sequence information. This approach aims at identifying the optimal partition of the set of branches in a likelihood framework according to Akaike information criterion (AIC) or Bayesian information criterion (BIC) and by design avoids overparametrization and subjective decisions from the user. The objective is to group branches along which sequence evolution has been similar, in terms of any set of descriptive statistics of the substitution matrix. These statistics can be, for instance, dN/dS, equilibrium GC content, or the ratio of conservative to nonconservative amino acid substitutions (Sainudiin et al. 2005). The partition of branches returned by our approach can be correlated to characteristics of the organisms under study in order to identify new links between phenotypic and genomic features and possibly deduce ancestral characters at internal nodes of the tree. Such an approach is similar in principle to the so-called local molecular clocks, where branches with similar rates are clustered (Yang and Yoder 2003; Aris-Brosou 2007; Drummond and Suchard 2010; Heath et al. 2011). In our case, however, we are interested in the differential rate of each type of substitution, not the global amount of substitutions.

In this manuscript, we first present a new heuristic algorithm to find the optimal partition of branches and estimate parameters of substitution matrices. This algorithm benefits from an initial step of substitution mapping (Nielsen 2002; Rodrigue et al. 2008; Minin and Suchard 2008) and is implemented in C++ using the Bio++ libraries (Dutheil et al. 2006). Then, using simulations, we show that our approach is both fast and accurate. We apply our methods to previously published data sets, compare it with other approaches, and demonstrate that our new algorithms allow one to reveal the phenotypic determinants of sequence evolution for a wide range of experimental conditions, including large data sets.

Materials and Methods

Definitions and Notations

We denote D_i the ith site of the data set, that is, a column of the alignment, and Θ the set of parameters of a given model of sequence evolution. We consider the tree topology as fixed. By convention, we consider top nodes to be closer to the leaves than bottom nodes.

Substitution Mapping

We count the number of substitutions that occurred on each branch of a phylogenetic tree and at each site in a sequence alignment. We extend the procedure described in Dutheil et al. (2005) by providing detailed counts for each type of substitution in lieu of the total number of substitutions, following work by Minin and Suchard (2008) and Hobolth and Stone (2009).

We recall that at each position i in the alignment, we can compute the substitution vector $V_i^s = (v_{i,1}^s, \ldots, v_{i,b}^s, \ldots, v_{i,m}^s)$, where $v_{i,b}^s$ is the posterior estimate of the number of substitutions of type s that occurred on branch b and m is the number of branches in the tree. Following Dutheil et al. (2005), $v_{i,b}^s$ is estimated by averaging over all possible ancestral states at top (x_q) and bottom (x_p) nodes of branch b:

$$v_{i,b}^s = \sum_{x_p}\sum_{x_q} \Pr(x_p, x_q | D_i, \Theta) \times n_{x_p,x_q}^s(t). \qquad (1)$$

In this equation, $\Pr(x_p, x_q|D_i, \Theta)$ is the joint probability of having state x_p at bottom node and state x_q at top node given the data and parameters. It is computed as follows (Galtier and Boursot 2000; Pupko et al. 2003; Dutheil et al. 2005):

$$\Pr(x_p, x_q|D_i, \Theta) = \frac{\Pr(x_p, x_q, D_i|\Theta)}{\Pr(D_i|\Theta)}. \qquad (2)$$

The denominator is the likelihood for site i (Felsenstein 1981), whereas the numerator is obtained in a very similar way but considering the ancestral states x_p and x_q as known in the Felsenstein recursion. Term $n_{x_p,x_q}^s(t)$ is the mean number of substitutions of type s that occurred on a branch of length t knowing initial state x_p and final state x_q. Several methods have been proposed to compute this mean number. Instead of the method of Dutheil et al. (2005), we use the uniformization method proposed by Hobolth and Stone (2009) and Tataru and Hobolth (2011) because it is exact, numerically more stable, and for each site and branch it returns an array containing counts for each type of substitution. For alphabets with high dimension like the codon alphabets, substitution types can be summed, for example, all synonymous substitutions and all nonsynonymous substitutions, generating two types of counts instead of $61 \times (61-1) = 3,660$. As shown in Dutheil et al. (2005), these equations can be easily extended to account for variation of the substitution rate across sites and do not depend on the model used to describe the rate variation. For codon models, a constant distribution of codon substitution rates was used, as in the PAML software, whereas for nucleotide sequences we used a gamma distribution of site-specific substitution rates.

We summed substitution counts obtained for each site of the alignment to obtain branch-wise counts for each type of substitution. We further pooled substitution counts depending on the biological question we addressed: A or $T \rightarrow G$ or C and G or $C \rightarrow A$ or T in order to study the variation of GC content in nucleotide sequence and synonymous versus nonsynonymous substitutions for studying the variation of selection regime in codon sequences. Had we chosen to use our method to study selection at the amino acid level, we could have pooled substitutions in conservative versus nonconservative substitutions.

The substitution mapping requires a model of sequence evolution and a phylogenetic tree to work with. Several works have shown that this procedure is robust to the input model of sequence evolution (see, for instance, Minin and Suchard 2008). We therefore fitted a Tamura (1992) (respectively Nielsen and Yang 1998) homogeneous model for the ribosomal RNA (rRNA) (respectively codon) data set and estimated all numerical parameters, that is, kappa, theta (respectively omega), and branch lengths. These parameters were then used for mapping substitutions. For the rRNA data set, substitution mapping was performed on a rooted tree as nonstationary models will be used in the model selection procedure.

Measuring Substitution Process Homogeneity between Branches

The total numbers of substitutions of each type, ν^s_b, were computed for each branch by summing over all sites:

$$\nu^s_b = \sum_i \nu^s_{i,b}. \tag{3}$$

We designed a multinomial likelihood-based measure of substitution process homogeneity between any two branches, here named b_1 and b_2. Under the null model of homogeneous process, the two vectors of counts are assumed to be drawn from a unique multinomial distribution in which the probabilities of each type of substitution, p_s, are shared by the two branches. The likelihood of the set of counts under the null model is

$$L_0 = \frac{\left(\nu_{b_1}+\nu_{b_2}\right)!}{\prod_s \left(\nu^s_{b_1}+\nu^s_{b_2}\right)!}\prod_s p_s^{\left(\nu^s_{b_1}+\nu^s_{b_2}\right)}, \tag{4}$$

where $\nu_{b_1} = \sum_s \nu^s_{b_1}$ is the total number of substitutions summed across categories $\left(\text{and similarly for } \nu_{b_2}\right)$.

The maximum likelihood estimates of the p_s's, to be used in equation (4), are

$$p_s = \frac{\nu^s_{b_1}+\nu^s_{b_2}}{\nu_{b_1}+\nu_{b_2}}. \tag{5}$$

Under the alternative model of heterogeneous process, the two vectors of counts are drawn from two distinct multinomial distributions, the probabilities of substitution types being different between branches. The likelihood of the data is now

$$L_1 = \frac{\nu_{b_1}!}{\prod_s \nu^s_{b_1}!}\prod_s p_{1s}^{\nu^s_{b_1}} \times \frac{\nu_{b_2}!}{\prod_s \nu^s_{b_2}!}\prod_s p_{2s}^{\nu^s_{b_2}} \tag{6}$$

and the maximum likelihood estimates of the p_{1s}'s and p_{2s}'s are

$$p_{1s} = \frac{\nu^s_{b_1}}{\nu_{b_1}}, \quad p_{2s} = \frac{\nu^s_{b_2}}{\nu_{b_2}}. \tag{7}$$

Twice the log-likelihood ratio between the two models is calculated and compared with a chi-squared distribution, the number of degrees of freedom being equal to the number of categories minus one. The resulting P value is considered as a measure of compatibility between the two considered branches. This measure accounts for the uncertainty due to stochastic errors in substitution counts.

This method compares substitution counts between branches. Substitution counts, however, are not a full description of the substitution process when sequences are not at compositional equilibrium. Consider, for instance, two branches in each of which exactly 10 $AT \longrightarrow GC$ and 10 $GC \longrightarrow AT$ changes were counted. Now suppose that the GC content of the considered sequence is 90% for branch 1 and 10% for branch 2. Despite having identical substitution counts, these two branches have distinctive evolutionary processes: the per AT site $AT \longrightarrow GC$ substitution rate, for instance, is nine times higher in branch 1 than in branch 2. To account for this effect, we define corrected counts ν'^s_b as

$$\nu'^s_b = \Lambda\frac{\nu^s_b}{k^s_b}, \tag{8}$$

where k^s_b is the number of positions in the sequence in branch b at which a substitution of category s could have occurred (e.g., for $AT \longrightarrow GC$ substitutions, the number of A and T positions). The k^s_b's are estimated by reconstructing the distribution of ancestral sequences at the parent node of the considered branch (Yang and Roberts 1995). In equation (8), $\Lambda = \sum \nu^s_b / \sum \frac{\nu^s_b}{k^s_b}$ is a scaling factor ensuring that the sum (over substitution categories) of the ν'^s_b's is equal to the sum of the ν^s_b's. In this study, the exact counts ν^s_b were used in equations (4)–(7) for comparisons between synonymous and nonsynonymous substitutions and the corrected counts ν'^s_b for comparisons between $AT \longrightarrow GC$ and $GC \longrightarrow AT$ substitutions.

Partitioning Branches

We developed a new hierarchical clustering procedure in order to define subsets of branches along the phylogenetic tree, based on their respective substitution processes, as inferred from branch-wise counts for each type of substitution. The procedure clusters branches of the tree following a neighbor-joining strategy. Each branch is initially assigned its own cluster, then the two most similar clusters of branches are repeatedly merged until only one cluster is left. The resulting tree is assigned branch lengths so that the height of inner nodes reflect the P value associated to the underlying clusters. More precisely:

Initialization. We start by associating each branch to their own subset. All pairs of single-branch subsets are tested for homogeneity of substitution process using the previously introduced branch–pair homogeneity test. The corresponding P values are stored.

Extension. The two subsets with the highest P value, $P_{\max}$, are gathered into a new subset. A new node in the clustering tree is created, with height equal to $\frac{(1-P_{\max})}{2}$. Substitution counts for the new subset are obtained by summing the counts for each individual subset. The new subset is tested against all other subsets using the multinomial test and the summed counts.

Termination. The procedure stops when there is only one cluster left.

The resulting clustering tree defines a hierarchy of nonhomogeneous models, from one substitution matrix per branch of the phylogenetic tree to one substitution matrix only for the whole phylogenetic tree. These models can be optimized in the maximum likelihood framework and compared using AIC or BIC (see below).

We introduced two modifications to improve the robustness of the above clustering algorithm to short branches. First, branches for which no substitutions (or by extension, for which less than a user-specified number of substitutions) are inferred are automatically clustered with their parent node in the phylogenetic tree. Second, we treat negative branch lengths in the clustering tree as equal to zero and the corresponding parent node as multifurcating, therefore defining more than one new cluster of branches compared with the immediately simpler model.

Finally, we introduced a variant of the clustering algorithm that only allows neighbor branches to be clustered. This was obtained by modifying the initialization step so that non-neighbor branches have a negative *P* value, which prevents them from being clustered at any step. Similarly, during the extension step, non-neighbor subsets are assigned a negative *P* value. We refer to this variant as the "join" model, whereas we refer to the original clustering without neighbor constraint as the "free" model.

Model Selection

Models are constructed from a set of branch clusters and optimized in the maximum likelihood framework. Models built from more than one subset are nonhomogeneous and are constructed as described in Dutheil and Boussau (2008). Here we consider nonhomogeneous models derived from a single type of substitution matrix (e.g., Tamura 1992; Nielsen and Yang 1998), and only branch lengths and parameters of these substitution matrices are allowed to vary on a per-branch basis (θ for Tamura 1992, ω for Nielsen and Yang 1998). Despite these restrictions, such models encompass the vast majority of nonhomogeneous models used in the literature.

Nested models are obtained by successively considering each node of the clustering tree in order of increasing height, starting from the root. The root node defines a bipartition (two subsets) of branches, and each descending node iteratively splits one of the previously defined subsets of branches. Nested models are compared on the basis of their maximum likelihood. We implemented two comparison procedures: AIC and BIC. Model exploration is stopped when the scores of *n* consecutive models are lower than the current best score, where *n* is a positive number defined by the user, or when all models have been tested.

Note that in the case of nonstationary models like the Galtier and Gouy (1998) nucleotide model, we tested both the homogeneous and the homogeneous nonstationary models, which has one extra parameter, the root equilibrium GC frequency.

Simulations

In order to assess the performance of the method (clustering + model selection), we conducted a simulation analysis under various models. We used a random tree containing 25 leaves, generated under a Yule distribution with the *rtree* command from the R package APE (Paradis et al. 2004). We randomly generated 18 nonhomogeneous models according to the free setting by splitting the tree in 1, 2, 3, 4, 5, or 10 subsets of branches, with three replicates in each case. Each branch was assigned one of the partitions randomly. We also generated 18 nonhomogeneous models according to the join setting, by picking a random subtree and assigning it to a partition number. We simulated codon sequences under the YN98 model of sequence evolution in which the omega (=d*N*/d*S*) parameter of each subset of branch was drawn from a gamma distribution with $\alpha = \beta = 0.5$. Six sequence alignments were generated for each of the 18 models, three with 300 positions and three with 1,000 positions, totalling 108 sequence alignments. On each data set, we fitted a YN98 homogenous model with the true phylogeny and used it to perform substitution mapping. The resulting substitution maps were used to obtain a tree of clustered branches which in turn was used to guide model selection using BIC. We used the free (respectively join) clustering algorithm on the data sets simulated under free (respectively join) model. Three additional nested models were tested after a local minimum was found, and the resulting global minimum used for creating partitions.

Programs and Examples

The testnh package is available at http://biopp.univ-montp2.fr/forge/testnh as source code and binary versions. The programs are written in standard C++ and compile on Linux, Windows, and MacOS systems and only depend on the Bio++ libraries (Dutheil et al. 2006). The package contains two major programs: mapnh for performing the substitution mapping and clustering of branches and partnh for fitting substitution models on the resulting subsets. All the data sets analyzed in this article are provided as example files in the source distribution. The package also contains the randnh program that was used to generate random models in the simulation analysis.

Data Sets

Tree of Life Concatenated RRNA *Alignment*

We used the rRNA alignment (concatenated small and large subunits) and corresponding phylogeny from Boussau and Gouy (2006) previously analyzed in Dutheil and Boussau (2008). The tree was midpoint-rooted. The alignment contains 527 complete sites for 92 sequences including 22 archaea, 34 bacteria, and 36 eukaryotes. These species come from a large variety of environments and show a wide range of optimal growth temperatures, which is known to affect rRNA GC content evolution in prokaryotes (Boussau and Gouy 2006).

Lysozyme Data Set

This is the data set provided together with the PAML software and described in Yang (1998). This data set was also analyzed by Zhang et al. (2011). We used the tree shown in figure 2 in Zhang et al. (2011) where the human branch was removed in order to compare results.

Daphnia Data Set

Accession numbers given in Paland and Lynch (2006) were used to download mitochondrial sequences from 28

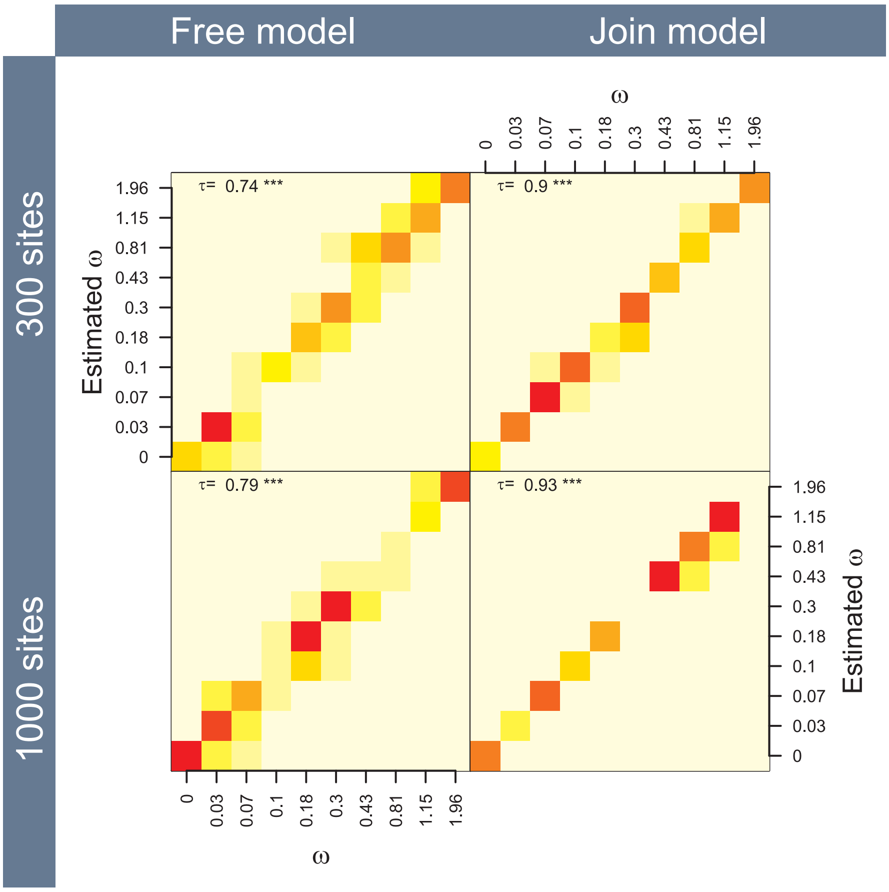


FIG. 1. Recovery of the branch-wise dN/dS (ω), displayed as a heat map. The x axis represents the true value of ω as used in the simulations, and the y axis displays the corresponding value estimated by the selected model. Colors describe the density of points, as 12 equispaced categories, from white (no point) to red (maximum number of points). Simulations with 1, 2, 3, 4, 5, and 10 random subsets are used (see Materials and Methods and supplementary fig. 1 for separate plotting). Values of the Kendall correlation test are reported for each model.

Daphnia pulex strains from GenBank (accession numbers DQ340817–DQ340843 and AF117817). Coding DNA sequences were extracted for genes *ATP6*, *ATP8*, *CO1*, *CO2*, *CO3*, *CYTB*, *ND1*, *ND2*, *ND3*, *ND4*, *ND4L*, *ND5*, and *ND6*, and concatenated. Alignments were done with Muscle (Edgar 2004) on amino acid sequences using Seaview (Gouy et al. 2010) and corrected by eye where necessary. This resulted in 3,681 codon sites.

Mantellid Frogs Data Set

The alignment and tree used in our analyses are as in Boussau et al. (2011) and can be downloaded from Treebase (http://purl.org/phylo/treebase/phylows/study/TB2:S11392).

Results

Simulations

Using simulated data with a codon model, we assessed the ability of our branch-clustering method to recover branch-specific parameter estimates. Figure 1 shows that the dN/dS values estimated by our clustering approach are accurately estimated, both in the free and in the join models. The join approach, however, shows more accurate estimates than the free approach. This was expected, as for the same number of partitions, the free model has more breakpoints (i.e. change in the substitution process) along the phylogeny than the join model, making it more difficult to recover for a data set of equivalent size. The larger data sets (1,000 sites) expectedly display a lower dispersion of estimates than the smaller ones (300 sites). The dispersion also increases with the number of partitions used in simulations (see Supplementary Material online), and the effect is stronger on the free model.

We also used these simulations to assess whether the number of clusters returned by the methods are meaningful. This is more complicated to evaluate as the number of clusters depends on the criterion used for model selection and the size of the data set as larger data sets have more power to discriminate small changes in the substitution process along the tree. Figure 2 displays the number of partitions recovered with the BIC criterion. The join approach performs slightly better than the *free* approach, which can again be explained by the fact that join data sets display less breakpoints than the free ones, given a certain number of partitions. Also apparent is that larger data sets (1,000 sites) tend to recover more clusters than smaller data sets (300 sites). In our simulation setup, this leads to an

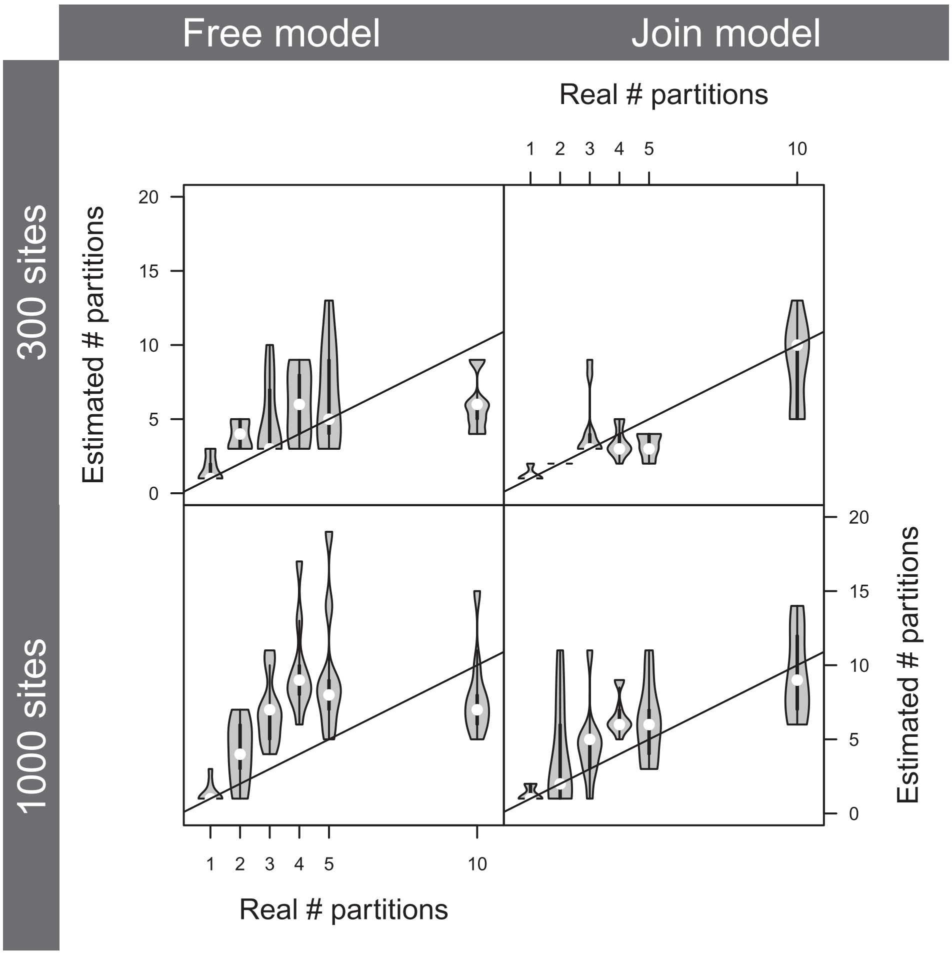


FIG. 2. Violin plots of the number of subsets retained under the BIC criterion as a function of the number of subsets used in the simulation procedure. Simulations with 1, 2, 3, 4, 5, and 10 random partitions were pooled (see Materials and Methods). The lines display the $y = x$ identity function.

overestimation of the number of clusters when the real number is small. As BIC is the most conservative criterion for model selection, this indicates that more permissive criteria like AIC or likelihood ratio test are likely to result in overparametrized models.

If there is an underlying discrete structure in the data, as in these simulations, the number of clusters inferred by our methods can provide insights into the biology of the organisms. In cases where the underlying structure is not discrete, but continuous, the partition itself is less relevant to the biology of the clade under study but should still be useful to decrease the noise in estimates of branch-wise parameters.

Heterogeneity in GC Content

The tree-of-life rRNA data set has been previously used to test branch-heterogeneous models in which the equilibrium GC content can vary among branches (Boussau and Gouy 2006; Dutheil and Boussau 2008). In prokaryotes, GC content in the stem portion of rRNA is correlated to optimal growth temperature (Galtier and Lobry 1997). Reconstructing GC content evolution therefore provides insight into phenotype evolution (Galtier et al. 1999; Boussau et al. 2008). The large number of leaves and relatively small number of complete sites in this data set make it challenging to estimate parameter-rich branch-heterogeneous models. Nonetheless, both the free and the join approaches recover extremely similar patterns of equilibrium GC content evolution along this tree (fig. 3). However, the free approach detects more changes between equilibrium GCs in clades, despite a lower number of clusters than the join approach (5 clusters for the free approach and 14 for the join). In addition, the join approach recovers more extreme values than the free approach. The differences between the two approaches in the branches leading from the root of the tree to its descendants are probably linked to different branch length estimates at the root. Such differences at the root between the two approaches are expected as it is known that it is very difficult to find the root of a phylogenetic tree using a branch-heterogeneous model of sequence evolution (Huelsenbeck et al. 2002; Boussau and Gouy 2006). Importantly, branches leading to thermophilic and hyperthermophilic species of bacteria and archaea, which live at high temperatures and have high GC contents in the stem portions of their rRNAs (Galtier and

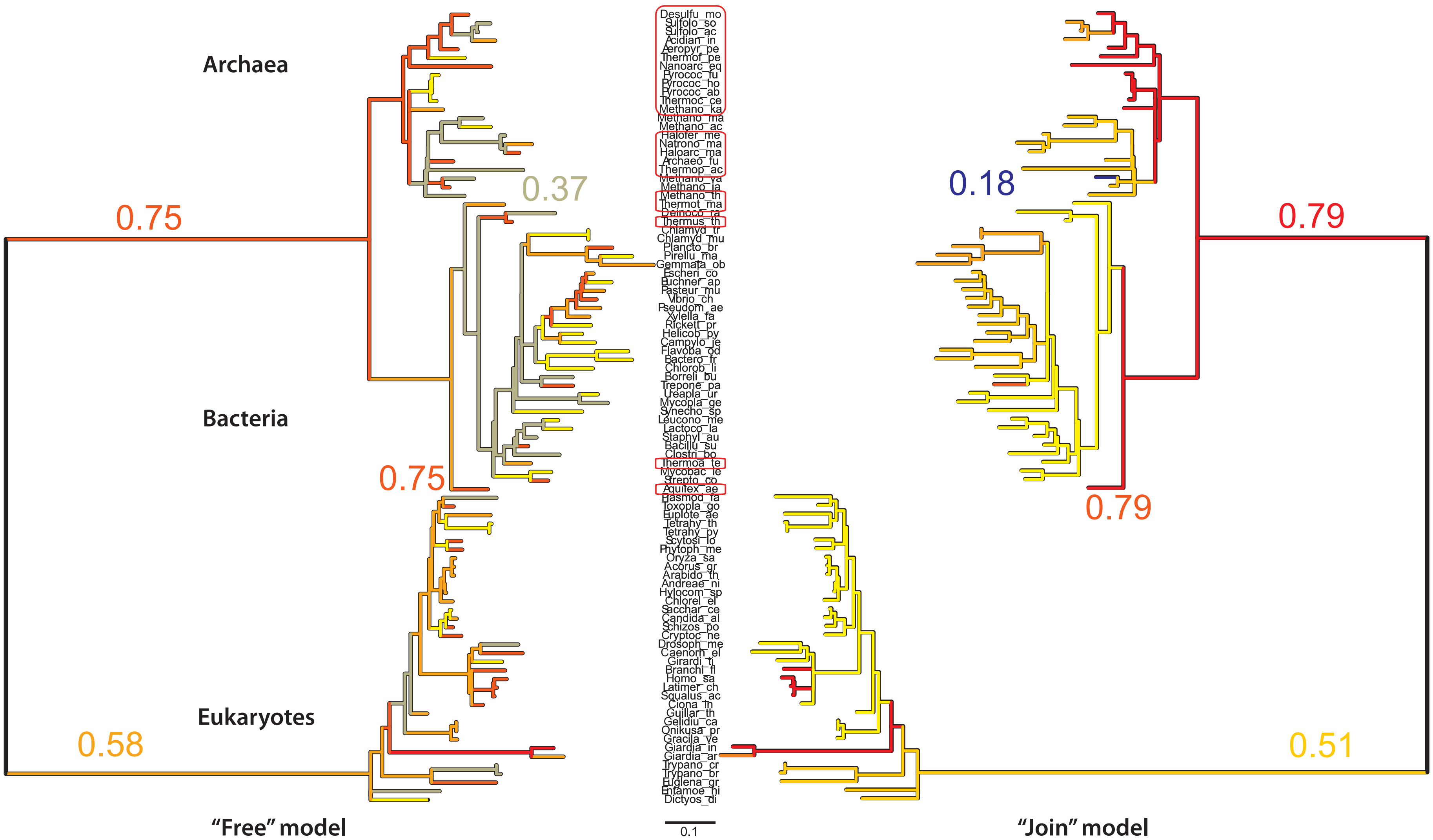


FIG. 3. Branch partition on the rRNA data set. Branches have been colored according to the equilibrium GC of the cluster they belong to (from blue for low GC to red for high GC, some values of interest are shown next to branches). Left: free model. Right: join model. Names of the thermophilic and hyperthermophilic species are circled in red. Branch lengths are proportional to the inferred numbers of substitutions per site (scale bar included).

Lobry 1997), are associated to high-equilibrium GC contents. The method also recovers high GC contents at the base of archaea and bacteria (Boussau et al. 2008) but does not seem to find evidence for later decreases in bacteria (Gaucher et al. 2008) and Archaea (Groussin and Gouy 2011), perhaps because the taxonomic sampling is inadequate to tackle such questions.

Expectedly, models obtained by our clustering methods have a much better BIC than a model (named 'general' here and in Dutheil and Boussau 2008) in which a specific stationary GC content parameter is associated to every single branch: the optimum free model has a log-likelihood of $-13,941$ and a BIC of 29,074 (182 branch lengths + transition/transversion ratio $[\kappa]$ + shape of gamma distribution of site-specific rate $[\alpha]$ + GC content at the root node $[\theta_{\text{root}}]$ + 5 partition-specific equilibrium GC contents = 190 parameters), the join model a log-likelihood of $-13,956$ and a BIC of 29,159 (199 parameters, 14 clusters of branches), whereas the general model a log-likelihood of $-13,821$ and a BIC of 29,942 (182 branch lengths + κ + α + θ_{root} + 182 branch-specific equilibrium GC contents = 367 parameters) (Dutheil and Boussau 2008). This suggests that our clustering methods indeed find models with a good balance between parameter richness and fit to the data. Because we estimated 20 more complex models than the optimum one to ensure that the optimum had really been found, 23 (free approach) and 27 (join approach) models have been fully optimized and compared using BIC during model selection. This is far smaller than the total number of models possible for this large tree and even much smaller than the number of branches in the tree (181). Figure 4 shows the optimization profile with the values of equilibrium GC contents (θ) during optimization of increasingly complex models. Expectedly, models chosen according to AIC use more parameters than models chosen according to BIC. For each criterion, the global minimum is reached after local minima, which stresses the need for our algorithm not to stop at the first minimum value encountered. Despite this, our algorithms are efficient and end in less than an hour on a desktop computer for a data set containing 92 sequences.

Heterogeneity in Selection: Selecting Codon Models

Sex and dN/dS *in Daphnia*

This data set of 28 mitochondrial sequences from *Daphnia pulex* contains 14 sexual and 14 asexual strains (respectively noted *Sn* and *An*, $n \in [1:14]$) and has been used to study the impact of recombination between mitochondrial genome and nuclear genome on coding sequence evolution (Paland and Lynch 2006). Asexual strains are thought to have repeatedly evolved from sexual ancestors. Paland and Lynch (2006) computed dN/dS separately for sexual and asexual lineages and found it was higher in asexual lineages, in agreement with the expectation that recombination should improve the efficiency of purifying selection. Figure 5 displays the optimization profiles with the corresponding values of ω during optimization, and figure 6 shows that the free approach recovers the dN/dS difference between sexual and asexual lineages. It clusters branches in three groups with dN/dS values of 0.10, 0.25, and 0.85. Among terminal branches, for which the sexual/asexual status of the strain can be observed, all four branches assigned to the cluster with the largest dN/dS lead to asexual organisms, whereas only three branches leading to asexual organisms are in the cluster with the lowest dN/dS. Among the branches leading to sexual organisms, none are assigned to the cluster with the highest dN/dS but 11 are assigned to the cluster with the lowest dN/dS. Internal branches tend to be assigned to the cluster with the lowest dN/dS, which confirms that asexual strains originated from sexual ancestors. The join approach groups all branches but one in a single low dN/dS cluster. One branch stands out in a high dN/dS cluster of its own, the branch leading to the asexual strain A13. This is an indication that the join approach here is not appropriate as it aims at grouping together neighboring branches, when sampling was intentionally designed such that sister taxa have contrasted life-history traits. This shows that the two methods presented here should be used in situations where their respective underlying hypotheses fit the data set under study.

Breeding System and dN/dS *in Mantellid Frogs*

Mantellid frogs of Madagascar have been used to study patterns of ecology-driven diversification (Vences et al. 2002) as well as patterns of mitochondrial genome evolution (Kurabayashi et al. 2008). Ecologically, some species of frogs breed in ponds, whereas others breed in streams. This difference may have left a trace in the pattern of diversification of a particular genus of Mantellid frogs, *Boophis*, where pond breeders tend to speciate less easily than stream breeders, presumably due to lower barriers to gene flow in pond breeders (Vences et al. 2002). Recently, the data set analyzed in the present article, and initially published in Kurabayashi et al. (2008), was used to find that increases in mitochondrial genome sizes occurred jointly with increases in dN/dS (Boussau et al. 2011). This indicates that important changes in genome structure may have fixed non adaptively.

We used our approach to cluster branches showing similar dN/dS along the tree of 17 Mantellid frogs. Figure 7 shows the best partition obtained with the free and join approach using BIC. Both partitions are highly similar with two and five clusters, respectively, and show a striking difference in dN/dS between stream breeders (circled) and pond breeders. Overall, pond breeders tend to display lower dN/dS (about 0.045) than stream breeders (0.08). This dN/dS difference can be explained by selectionist or neutralist hypotheses. We know of no reason to assume different selection regimes in the mitochondrial genomes of stream breeders compared with pond breeders. Instead, because pond breeders may suffer less barriers to gene flow, they may have larger effective population sizes than stream breeders. With larger effective population sizes, purifying selection would be more efficient in pond breeders than in stream breeders, and their dN/dS would be lower.

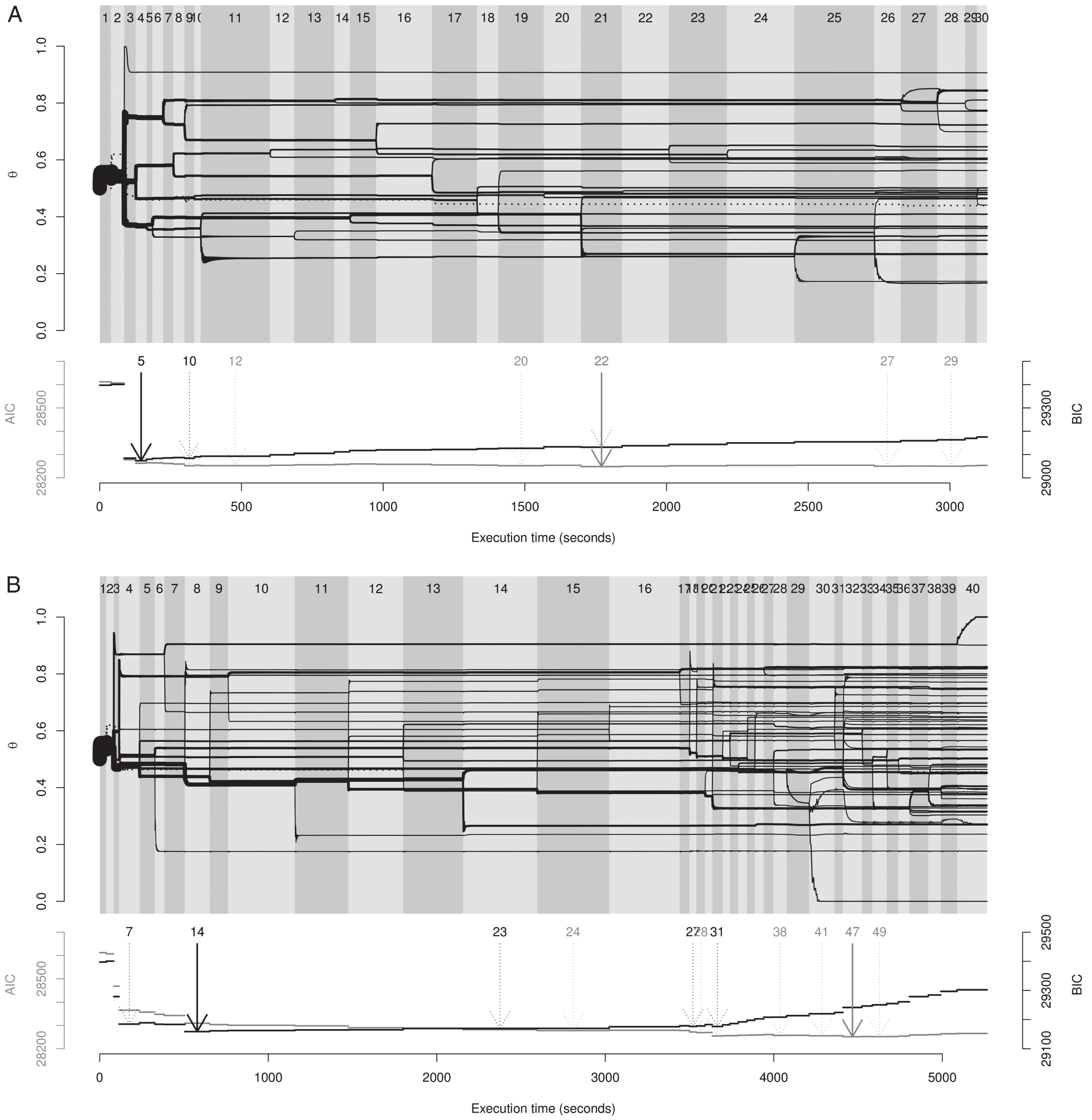


Fig. 4. Equilibrium GC content (θ_i parameters) estimation profiles for the rRNA data set. The values taken by the θ parameters during optimization (top panels) and the model comparison criteria (bottom panels) are plotted as a function of execution time in seconds, for both the free model (A) and the join model (B). Each distinct model is reported (top line numbers), from the simplest one (model 1, homogeneous, one partition with one θ on the left) to the most complex (models 30 and 40, more than 20 partitions/θ on the right). Top panels: line width depicts the size of the underlying partition. The dashed line shows the GC content at the root node of the tree. Bottom panels: AIC (gray) and BIC (black) values of corresponding models. Arrows depict local minima values, global minima being displayed with solid lines. Values on top of each arrow show the number of clusters in the corresponding model.

The optimal clusters found using BIC do not distinguish between branches with and without increases in mitochondrial sizes. This does not invalidate the results of Boussau et al. (2011) but instead illustrates the respective strengths of two different types of approaches. Hypothesis-driven approaches, as in Boussau et al. (2011), are very sensitive and therefore can reveal weak but significant effects. Exploratory approaches as used here cannot detect very subtle signals in the data but have the power to reveal strong signal of unsuspected but meaningful patterns of molecular evolution.

Innermost branches in the phylogenetic tree of Mantellids seem to be generally associated to larger values of dN/dS, both in the free and in the join approaches. This

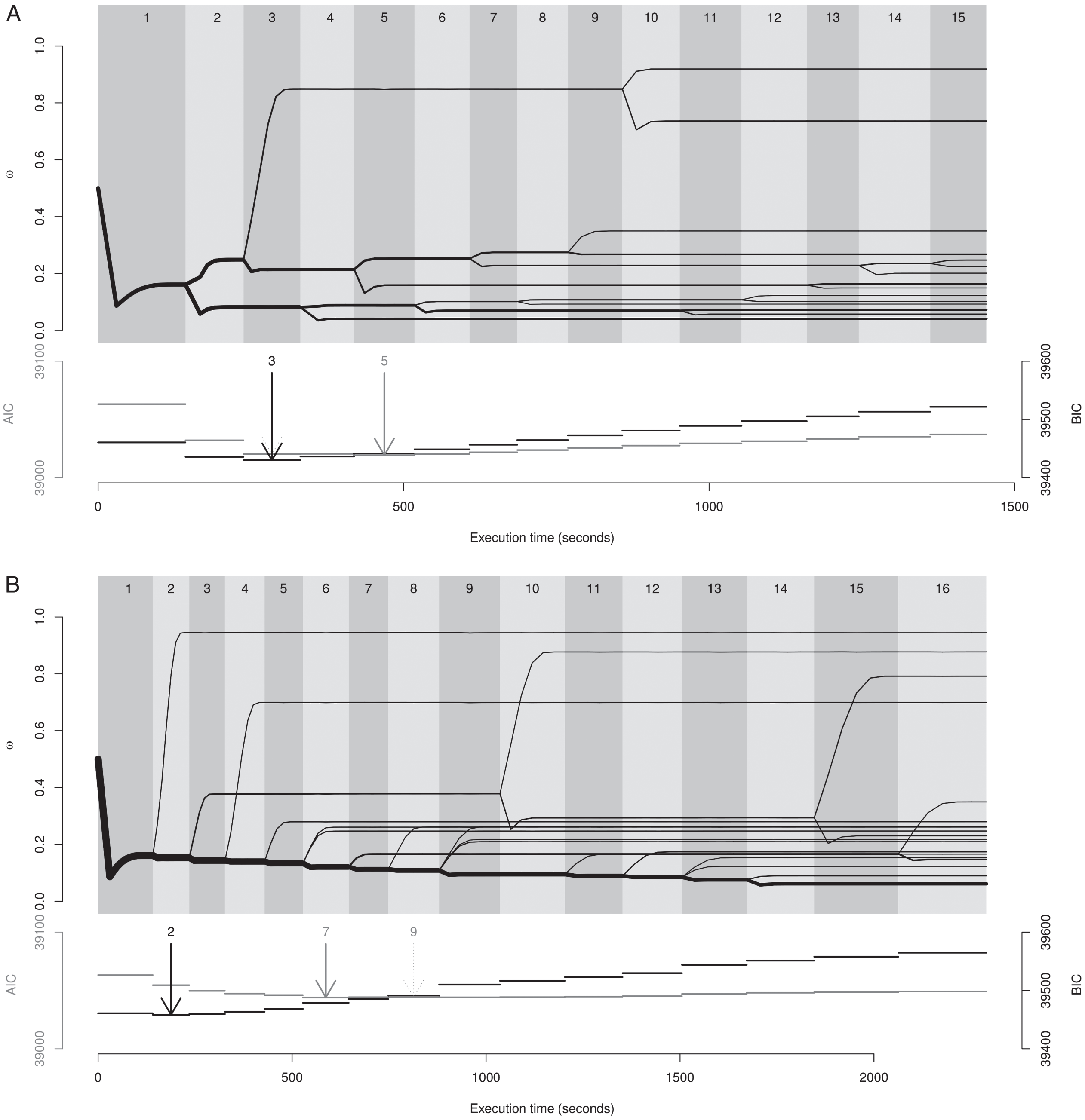


FIG. 5. dN/dS (ω_i parameters) estimation profiles for the Daphnia data set. The values taken by the ω parameters during optimization (top panels) and the model comparison criteria (bottom panels) are plotted as a function of execution time in seconds, for both the free model (A) and the join model (B). Other legends are similar to figure 4.

may be due to smaller effective population sizes in the ancestors of Mantellid frogs, about 20–50 Ma. This may also be indicative that dS is saturated on the most ancient branches, artificially inflating dN/dS. This second hypothesis would reconcile our results with the idea that Mantellid ancestors were pond breeders (Vences et al. 2002).

Assessing the Robustness of Selected Models

In the procedure we present, a homogeneous model is first used in order to perform substitution mapping, which is in turn used to cluster branches of the tree and guide the model selection procedure. The underlying rationale of this approach is the robustness of the substitution mapping procedure to the substitution model used (Minin and Suchard 2008). This robustness can be further assessed by performing a new substitution map from the selected model. This new a posteriori map can be used to obtain new branch-clustering trees and subsequent model selection. The resulting model can then be compared with the one previously found.

For codon data sets (Daphnia and Mantellid frogs), we find no difference in the selected models (with BIC criterion), with the exception of one node in the Daphnia data

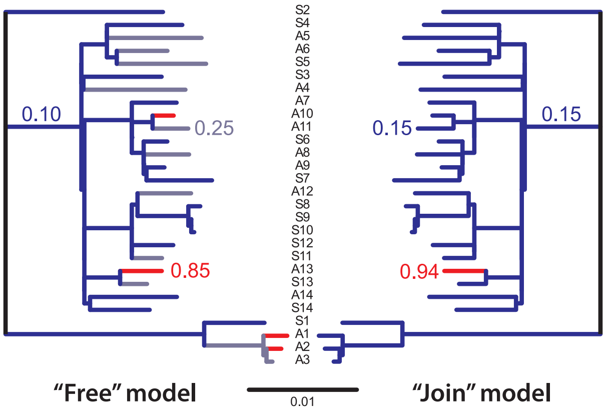


FIG. 6. Best partition proposed by the free and join models and according to BIC for the Daphnia data set. Strains named *An* are asexual, and strains named *Sn* are sexual. Branch lengths are proportional to the inferred numbers of substitutions per codon site (scale bar included).

set with a free clustering, moving from a partition with $\omega = 0.08$ to a partition with $\omega = 0.21$. For the rRNA data set, the free model gains one extra partition and the join model loses two. There are several differences between the two selected models in both cases, yet when results are compared on a per-node basis, resulting estimates of the θ parameters are quite similar (fig. 8). Varying nodes only move to the nearest cluster, reflecting uncertainty in the underlying value for the parameter, possibly due to a lack of signal in the data.

Discussion

We have presented an approach to cluster branches of a phylogenetic tree according to their pattern of sequence evolution. This approach provides models that accurately describe the evolution of a data set without overparametrization and bypasses preconception from the user. It can help in discovering new links between phenotype and sequence evolution.

Simulations and the four data sets we analyzed show that our method accurately clusters branches of a phylogenetic tree according to patterns of sequence evolution. This result represents a noticeable achievement as the challenge of finding an optimal partition of the branches of a tree is made difficult by the large combinatorial space that needs to be explored. An unrooted phylogenetic tree with 10 sequences contains 17 branches to cluster in 1–17 clusters, which can be done in more than 682 billion different ways (Zhang et al. 2011). Obviously all possible partitions cannot be tested, and heuristic algorithms have to be used. Such heuristics should only test a minute portion of the space of possible partitions to be fast but should not miss relevant partitions to provide biological insights.

Two recent works have shown that this challenge is currently generating a lot of interest and can now be tackled thanks to progresses in computing power (Zhang et al. 2011; Jayaswal et al. 2011). Zhang et al. (2011) developed a wrapper Perl script to produce option files to run PAML and test various partitions of branches in order to find branches showing similar d*N*/d*S*. Their most accurate algorithm starts by building a cluster around the single branch whose d*N*/d*S* is most different from the d*N*/d*S* of other branches. To find this branch, all models in which a branch is set apart from all other branches have to be optimized using PAML. The algorithm then clusters branches one at a time. Both the initial step and the recursive step can be long when the number of branches is large, and overall $O(n^2)$ models are optimized, with *n* the number of branches. We applied our methods to the lysozyme data set, originally studied by Yang (1998) and reanalyzed by Zhang et al. (2011). Using AIC, the free approach recovers their partition of the branches (not shown). Using BIC yields a lower number of clusters. However, branch-wise d*N*/d*S* values estimated using either AIC or BIC, and using either the free or join approaches, are robust and consistent with previous results (Yang 1998). Jayaswal et al. (2011) developed a heuristic algorithm in R to cluster branches based on the pattern of nucleotide substitutions rather than based on d*N*/d*S*. They use another type of algorithm, which starts from the most complex model in which each branch of the phylogenetic tree is associated to its own substitution matrix. Then, they iteratively cluster short branches or branches that have similar substitution matrices. In the end,

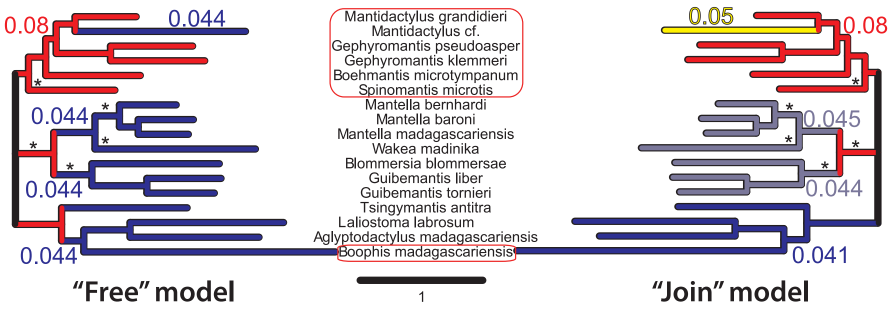


FIG. 7. Branch partition on the Mantellid frog data set. Branches have been colored according to the d*N*/d*S* of the cluster they belong to. d*N*/d*S* values are shown next to branches. Left: free model. Right: join model. Names of the species breeding in streams are circled in red; other species breed in ponds. Branch lengths are proportional to the inferred numbers of substitutions per codon site (scale bar included). Lineages where duplications have occurred are annotated with asterisks. Three of the five clusters recovered by the join approach have d*N*/d*S* values that are very close to each other: 0.041, 0.044, 0.045. The same branches in the free model are all found in the same cluster with a d*N*/d*S* of 0.044.

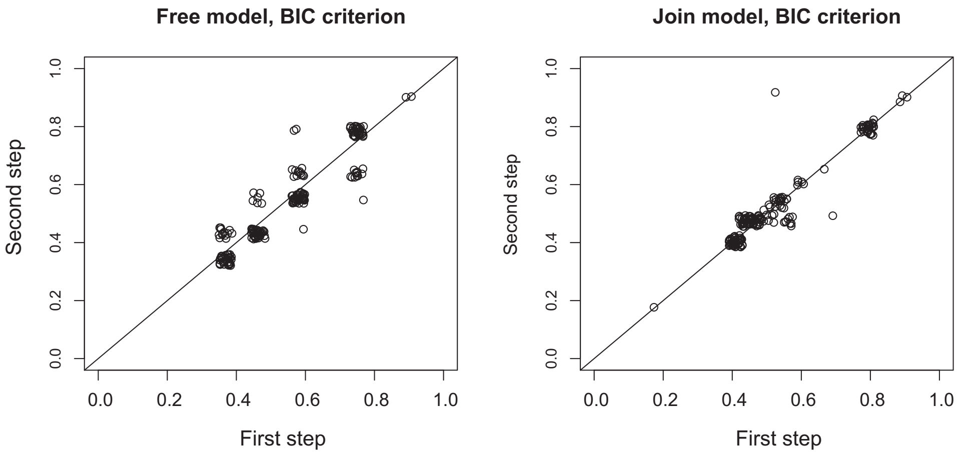


FIG. 8. Branch-specific equilibrium GC content estimates (θ parameters) from the best model according to BIC, after one step (x axis) or two steps (y axis) of model selection. Jitter was added in order to display overlapping points and do not correspond to real dispersion.

models ranging from the most complex, with one substitution matrix per branch, to the simplest, with one substitution matrix for all branches, are available, and the model providing the highest AIC or BIC is returned. This algorithm may be faster than Zhang et al.'s method as it requires optimizing only $O(n)$ models. However, the initial optimization of the parameters of the most complex model, the very same type of model discouraged in PAML's manual, may be difficult and costly. Both Zhang et al. (2011) and Jayaswal et al. (2011) approaches rely on a greedy heuristic algorithm to cluster branches, which cannot correct a mistake made at an early step. Therefore, if the initial decision to build clusters starting from the most distinct branch is not correct, or the estimation of substitution matrices in the most complex model is faulty, the resulting partition may not be correct.

Our approach improves upon these two early attempts in several respects. First, we cluster branches of the phylogenetic tree based on the substitution mapping procedure (Nielsen 2002; Minin and Suchard 2008), which has the advantage of being fast and robust to model misspecification (Minin and Suchard 2008). Substitution mapping provides counts of each type of substitution for each branch of the tree, based on a substitution model. Minin and Suchard (2008) have shown that substitution mapping was robust to the model used for mapping: even a simple homogeneous substitution model, where all branches share the same substitution matrix, can uncover accurate patterns of heterogeneity in the substitution process. Such a homogeneous substitution model also offers the advantage of being fast to fit, compared with parameter-rich, nonhomogeneous models used in the initial steps of the two previously mentioned approaches. The robustness of the mapping and subsequent clustering approach can be assessed a posteriori using the selected model in order to generate a new substitution map. In the data sets exemplified in this work, we showed that the resulting parameter estimates are quite robust to the initial model used for mapping substitution. However, it remains possible that for some data sets parameter estimates may be sensitive to the initial model. In such cases, successive iterations of mapping and model selection can be used to assess the uncertainty in branch-specific model attributions. Such an iterative approach can easily be performed with the TestNH package.

The robustness of the mapping procedure ensures that only meaningful clusters of branches are tested. This represents an advantage over the arbitrary starting point used by Zhang et al. (2011) and the complex and difficult to optimize starting point used by Jayaswal et al. (2011). Our procedure also prevents issues of overparametrization as it starts from models with small numbers of clusters. Counts of substitutions obtained for each branch are then used to partition branches in increasing numbers of subsets. From the resulting clustering tree, we design and fit a series of nested nonhomogeneous models, starting from the most simple one and progressively increasing the number of parameters, until it seems certain that improvement in AIC or BIC score can no longer be achieved. This procedure has the advantage that less than n optimizations are required (45 or 49 optimizations for the rRNA data set, containing 181 branches) and, perhaps more importantly, the models with the largest numbers of clusters, overparameterized and most costly to optimize, are often never optimized as the algorithm settles on models with small numbers of clusters (for instance, two to five dN/dS clusters in the tree of Mantellid frogs, which contains 31 branches). This is in sharp contrast to Jayaswal et al. (2011)'s approach where all the most parameter-rich models are necessarily optimized.

We propose in this work two clustering algorithms, corresponding to two distinct assumptions about character evolution. Choosing between these two methods mostly depends on the biological question underlying each specific analysis. If our approach is to be used mainly for avoiding overparametrization issues while estimating branch-specific parameters, we recommend using the free algorithm, as we expect it to provide better models (according to statistical criteria like BIC).

Another important feature of our algorithms is their generality as they do not depend on a particular sequence type or family of models. They are by construction not limited to dN/dS studies (Zhang et al. 2011) or to studies of compositional heterogeneity in DNA sequences (Jayaswal et al. 2011): built upon the Bio++ libraries (Dutheil et al. 2006), these programs can cluster branches based on any features of DNA, RNA, codon, or amino acid substitution matrices.

Finally, as shown in the simulations and on real data, these algorithms are fast despite the number of models fitted. This stems from our use of C++ for all steps of the code and of substitution mapping to restrain the number of models to optimize. Even with conservative parameters allowing to test a large set of models, the total execution times for the data sets exemplified in this work are of the order of 1–3 h on a 2.27-Ghz computer, which is comparable to the execution time of PAML with a branch model. Our results suggest that less conservative parameters for model exploration can be safely used: in our examples looking at only three models after a local AIC or BIC minimum has been found would still ensure that the global minimum is reached and would provide a significant gain in execution time.

We are confident that our two clustering approaches will be very useful to study patterns of molecular evolution since they can be used to cluster branches according to any statistics describing sequence evolution. For instance, one could cluster branches according to their ratio of transitions to transversions, according to both their equilibrium GC content and their dN/dS, or according to counts of all possible types of substitutions. Moreover, our programs can run on DNA, RNA, codon, or amino acid sequences, are efficient, and require very little input from the user. They could therefore be used on phylogenomic data sets to estimate genome-wide heterogeneity in sequence evolution, help reveal phenotypic determinants of sequence evolution, and consequently provide means to reconstruct phenotypic evolution along the tree of life.

Supplementary Material

Supplementary materials are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors would like to thank Asger Hobolth and Paula Tataru for sharing code on the computation of the conditional number of jumps of a Markov chain. We greatly benefited from some C++ code contributed by Laurent Guéguen. We would also like to thank two anonymous reviewers for very constructive comments on an earlier version of this manuscript. This work was supported by European Research Council grant 232971 ("PopPhyl") and the French Agence Nationale de la Recherche "Bioinformatique" (ANR-10-BINF-01 "Ancestrome"). B.B. was supported by a postdoctoral fellowship from the Human Frontier Science Program and the CNRS. This publication is contribution 2012-016 of the Institut des Sciences de l'Évolution de Montpellier (UMR 5554CNRS).

References

Aris-Brosou S. 2007. Dating phylogenies with hybrid local molecular clocks. *PLoS One* 2(9):e879.

Boussau B, Blanquart S, Necsulea A, Lartillot N, Gouy M. 2008. Parallel adaptations to high temperatures in the Archaean eon. *Nature* 456(7224):942–945.

Boussau B, Brown JM, Fujita MK. 2011. Nonadaptive evolution of mitochondrial genome size. *Evolution* 65(9):2706–2711.

Boussau B, Daubin V. 2010. Genomes as documents of evolutionary history. *Trends Ecol Evol.* 25(4):224–232.

Boussau B, Gouy M. 2006. Efficient likelihood computations with nonreversible models of evolution. *Syst Biol.* 55(5):756–768.

Bromham L. 2009. Why do species vary in their rate of molecular evolution? *Biol Lett.* 5(3):401–404.

Drummond AJ, Suchard MA. 2010. Bayesian random local clocks, or one rate to rule them all. *BMC Biol.* 8:114.

Dutheil J, Boussau B. 2008. Non-homogeneous models of sequence evolution in the Bio++ suite of libraries and programs. *BMC Evol Biol.* 8:255.

Dutheil J, Gaillard S, Bazin E, Glémin S, Ranwez V, Galtier N, Belkhir K. 2006. Bio++: a set of C++ libraries for sequence analysis, phylogenetics, molecular evolution and population genetics. *BMC Bioinform.* 7:188.

Dutheil J, Pupko T, Jean-Marie A, Galtier N. 2005. A model-based approach for detecting coevolving positions in a molecule. *Mol Biol Evol.* 22(9):1919–1928.

Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinform.* 5:113.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol.* 17(6):368–376.

Galtier N, Boursot P. 2000. A new method for locating changes in a tree reveals distinct nucleotide polymorphism vs. divergence patterns in mouse mitochondrial control region. *J Mol Evol.* 50(3): 224–231.

Galtier N, Gouy M. 1998. Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. *Mol Biol Evol.* 15(7):871–879.

Galtier N, Lobry JR. 1997. Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. *J Mol Evol.* 44(6):632–636.

Galtier N, Tourasse N, Gouy M. 1999. A nonhyperthermophilic common ancestor to extant life forms. *Science* 283(5399):220–221.

Gaucher EA, Govindarajan S, Ganesh OK. 2008. Palaeotemperature trend for Precambrian life inferred from resurrected proteins. *Nature* 451(7179):704–707.

Gouy M, Guindon S, Gascuel O. 2010. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol Biol Evol.* 27(2):221–224.

Groussin M, Gouy M. 2011. Adaptation to environmental temperature is a major determinant of molecular evolutionary rates in Archaea. *Mol Biol Evol.* 28(9):2661–2674.

Heath TA, Holder MT, Huelsenbeck JP. Forthcoming 2011. A Dirichlet process prior for estimating lineage-specific substitution rates. *Mol Biol Evol.*

Hickey DA, Singer GA. 2004. Genomic and proteomic adaptations to growth at high temperature. *Genome Biol.* 5(10):117.

Hobolth A, Stone EA. 2009. Simulation from endpoint-conditioned, continuous-time Markov chains on a finite state space, with applications to molecular evolution. *Ann Appl Stat.* 3:1204.

Huelsenbeck JP, Bollback JP, Levine AM. 2002. Inferring the root of a phylogenetic tree. *Syst Biol.* 51(1):32–43.

Jayaswal V, Ababneh F, Jermiin LS, Robinson J. 2011. Reducing model complexity of the general Markov model of evolution. *Mol Biol Evol.* 28(11):3045–3059.

Kurabayashi A, Sumida M, Yonekawa H, Glaw F, Vences M, Hasegawa M. 2008. Phylogeny, recombination, and mechanisms of stepwise mitochondrial genome reorganization in mantellid frogs from Madagascar. *Mol Biol Evol.* 25(5):874–891.

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

Minin VN, Suchard MA. 2008. Fast, accurate and simulation-free stochastic mapping. *Philos. Trans. R. Soc. B* 363(1512):3985–3995.

Nielsen R. 2002. Mapping mutations on phylogenies. *Syst Biol.* 51(5):729–739.

Nielsen R, Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* 148(3):929–936.

Paland S, Lynch M. 2006. Transitions to asexuality result in excess amino acid substitutions. *Science* 311(5763):990–992.

Paradis E, Claude J, Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20(2):289–290.

Pupko T, Sharan R, Hasegawa M, Shamir R, Graur D. 2003. Detecting excess radical replacements in phylogenetic trees. *Gene* 319:127–135.

Rodrigue N, Philippe H, Lartillot N. 2008. Uniformization for sampling realizations of Markov processes: applications to Bayesian implementations of codon substitution models. *Bioinformatics* 24(1):56–62.

Sainudiin R, Wong WSW, Yogeeswaran K, Nasrallah JB, Yang Z, Nielsen R. 2005. Detecting site-specific physicochemical selective pressures: applications to the Class I HLA of the human major histocompatibility complex and the SRK of the plant sporophytic self-incompatibility system. *J Mol Evol.* 60(3):315–326.

Tamura K. 1992. The rate and pattern of nucleotide substitution in *Drosophila* mitochondrial DNA. *Mol Biol Evol.* 9:814–825.

Tataru P, Hobolth A. 2011. Comparison of methods for calculating conditional expectations of sufficient statistics for continuous time Markov chains. *BMC Bioinform.* 12(1):465.

Vences M, Andreone F, Glaw F, Kosuch J, Meyer A, Schaefer HC, Veith M. 2002. Exploring the potential of life-history key innovation: brook breeding in the radiation of the Malagasy treefrog genus *Boophis*. *Mol Ecol.* 11(8):1453–1463.

Yang Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol Biol Evol.* 15(5):568–573.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 24(8):1586–1591.

Yang Z, Roberts D. 1995. On the use of nucleic acid sequences to infer early branchings in the tree of life. *Mol Biol Evol.* 12(3):451–458.

Yang Z, Yoder AD. 2003. Comparison of likelihood and Bayesian methods for estimating divergence times using multiple gene Loci and calibration points, with application to a radiation of cute-looking mouse lemur species. *Syst Biol.* 52(5):705–716.

Zhang C, Wang J, Xie W, Zhou G, Long M, Zhang Q. 2011. Dynamic programming procedure for searching optimal models to estimate substitution rates based on the maximum-likelihood method. *Proc Natl Acad Sci U S A.* 108(19):7860–7865.

Troisième partie

Conclusion

At the beginning of my thesis, I did not imagine that the study of GC-content would shed light on the evolutionary story of placental mammals. The initial objectives of this work was about molecular evolution, not species evolution. To better understand nucleotidic landscapes and biased gene conversion, mammalian genomes were considered as nothing more than a tool. Yet, at some point, GC-content turned out to be the tool and early mammalian evolution the subject.

This unexpected switch is, I think, one of the originalities of this work. Approaching mammal evolution through the lens of molecular evolution is not so common. From the viewpoint of a paleontologist, or even a phylogeneticist, using genomic GC-content to argue about the mammalian ancestry would probably look weird. However, in me opinion, using molecular data to explore the biology of our ancestors is a promising research avenue [Lartillot and Poujol, 2011]. This approach will probably lead to same controversies as for conflicting time divergence estimations, which have been opposing the molecular and morphologist communities for decades.

In my opinion, both fields can complement each other. Fossils provide snapshots of ancestral diversity at some point, but are not easily linked to modern species. On the other side, molecular estimations are only based on current species: they are more suitable to argue specifically about the common ancestors of modern species, but can not give trends about the whole past diversity, which includes extinct lineages. Accordingly to our results, the few cretaceous lineages that survived and gave birth to modern placental mammals were probably long-lived, even though long-lived animals probably represent a small proportion of the mammals living at this time, as fossils suggest.

However, for the moment, and as long as mammalian fossils from cretaceous are not assigned to Placentalia, direct comparisons between molecular estimations and paleontology remains impossible. Therefore, one perspective would be to test molecular estimations of life-history traits within a modern order, where fossils are abundant and less controversial.

From the mouse deer (0.8 kg) to the blue whale (180 000 kg), the Cetartiodactyla order appears as a good candidate. The fossil record is well known and contains a common ancestor which is supposed to be rabbit-sized (*Diacodexis*, [Rose, 1982]), whereas current species are mainly large-sized. Such a small ancestor is in agreement with our results, despite the fact that our four Cetartiodactyla species are relatively large (alpaca, pig, cow and dolphin). A focus on this order could confirm if parallel body mass increases occured in these different lineages, as suggested by these preliminary elements.

Molecular estimations of ancestral life-history traits could also be applied outside of mammals. Birds have a similar isochore structure, and are supposed to experience strong GC-biased gene conversion than mammals. Life-history traits estimations of the first bird lineages could give key hints about the origin of avian flight in relation to dinosaur evolution.

Furthermore, with the increasing number of molecular data, birds are supposed to enter sooner or later in the phylogenomic era. In order to resolve their hardest nodes, our results strongly suggest to avoid the phylogenetic signal provided by genes on microchromosomes. These small chromosomes (less than 20 Mb) are known to have a higher recombination rates, and to be the GC-richest pieces of their genome [Chicken and Sequencing, 2004]. According to our results, AT-rich macrochromosomes are expected to provide more reliable results. Because of this peculiar organization of bird karyotypes, conflicting phylogenetic signal could be common in this lineage. This is all the more important since ˜60 of the 80 avian chromosomes are microchromosomes, and contain most of the coding sequences (between 50 and 75% [McQueen et al., 1998, Burt, 2002]). Coding sequences are by far the biggest source of orthologous sequences used in phylogeny, but should be used with caution here.

The similarity between the isochore structure of birds and mammals brings us to the central question of the origin of isochores. The hypothesis of an amniote ancestry for isochores was recently challenged by the recent report of an absence of isochore in the genome of *Anolis carolinensis*, the first sequenced reptile genome [Fujita et al., 2011] (see Figure 2.10 in introduction). According to the authors, this fact can be explained by the independant apparition of this structuration in warm-blooded animals (the isochore as adaptation to homeothermy theory [Bernardi, 1985]) or by the stop and/or reversal of GC-biased gene conversion. However, our preliminary results presented in the introduction (Part 2.5.2 and 2.5.3) suggest that GC-biased gene conversion is still active in the anole lizard. Indeed, in agreement with biased gene conversion expectations, the average GC3% of genes on Anolis microchromosomes (55.6%) is superior to the average GC3% of those on macrochromosomes (42.9%). This is also true for the average equilibrium GC3, the GC-content that sequences would reach at equilibrium if patterns of substitutions remained constant over time, which predict an ongoing GC3 enrichment in microchromosomes (equilibrium at 59.6%) and an ongoing GC3 impoverisment in macrochromosomes (equilibrium at 34.4%). How to explain that ongoing biased gene conversion fails to maintain an isochore structure in Anolis, as it is the case in mammals and birds ? We suggest that the difference could be ascribed to mobile elements, reported in the green anole lizard genome to be young and diverse - more than in any other sequenced amniote genome [Alföldi et al., 2011, Novick et al., 2011]. Spreading fastly through the genome, they can rapidly modify a nucleotidic structuration slowly acquired by biased gene conversion. Furthermore, a fast turn-over of repeated elements could affect the distribution of recombination hostpots, leading to homogenization of the genomic GC-content. Such theories could be tested with more sequenced reptile genomes, which would certainly help to better understand the evolutionary forces underlying isochore birth and maintenance.

To conclude, I wish to share some more general personal reflections. Beyond all the academic knowledge, the main lesson I have drawn from this PhD thesis would be to be wary of seducing stories. From past studies or results found during this PhD thesis, biased gene conversion partly dismissed several attractive assumptions. (i) Isochores are probably not a wonderful convergent adaptation to homeothermy [Bernardi, 1985]. (ii) Human accelerated non-coding regions (HARs in the literature) are probably not what differentiates our brain from the other primates one [Duret, 2009]. (iii) The first placental mammals did probably not follow the most elegant Laurasia versus Gondwana vicariance scenario. (iv) Placental mammal ancestors were probably not tiny shrews that outperformed giant dinosaurs thanks to their ability to hide themselves from an asteroid impact.

All these stories share a common feature: they are appealing. In this regard, they could be viewed as succesful memes [Dawkins, 1976], ideas and concepts that remain in our culture in first place for their ability to spread from mind to mind. Even without clear-cut evidences, an adaptative (i)(ii) or biogeographical (iii) scenario is nearly always favoured. Story (iv) is by far the most widespread, and small rodent-like ancestors living side by side with dinosaurs are deeply embedded in our modern culture (see Annex 6.2 for some examples). Part of this success could be explained by the fact that mouse and shrew lifestyle is fundamentally different of human one. For many people, such an opposite lifestyle must be primitive, whereas becoming bigger and more longevive - more "human-like" - is seen as the logical next step of evolution. This view seems more confortable to many people, as it was for Haeckl (Figure 1.2 in Introduction), who made the mistake to consider evolution as successive stairs leading to human characteristics. With placental mammal ancestors more longevive than expected, the short-lived lifestyle would become a derived state, and perhaps the most successful one, indepedently acquired by Eulipotyphla, Afroinsectiphilia and the ubiquitous rodent order (more than 40% of current placental diversity).

Quatrième partie

Bibliographie

Bibliographie

[Alföldi et al., 2011] Alföldi, J., Di Palma, F., Grabherr, M., Williams, C., Kong, L., Mauceli, E., Russell, P., Lowe, C. B., Glor, R. E., Jaffe, J. D., Ray, D. A., Boissinot, S., Shedlock, A. M., Botka, C., Castoe, T. A., Colbourne, J. K., Fujita, M. K., Moreno, R. G., Ten Hallers, B. F., Haussler, D., Heger, A., Heiman, D., Janes, D. E., Johnson, J., De Jong, P. J., Koriabine, M. Y., Lara, M., Novick, P. A., Organ, C. L., Peach, S. E., Poe, S., Pollock, D. D., De Queiroz, K., Sanger, T., Searle, S., Smith, J. D., Smith, Z., Swofford, R., Turner-Maier, J., Wade, J., Young, S., Zadissa, A., Edwards, S. V., Glenn, T. C., Schneider, C. J., Losos, J. B., Lander, E. S., Breen, M., Ponting, C. P., and Lindblad-Toh, K. (2011). The genome of the green anole lizard and a comparative analysis with birds and mammals. *Nature*, 477(7366):5–9. 59, 66, 143

[Amrine-Madsen et al., 2003] Amrine-Madsen, H., Koepfli, K.-P., Wayne, R. K., and Springer, M. S. (2003). A new phylogenetic marker, apolipoprotein B, provides compelling evidence for eutherian relationships. *Molecular Phylogenetics and Evolution*, 28(2):225–240. 34

[Archibald et al., 2001] Archibald, J. D., Averianov, a. O., and Ekdale, E. G. (2001). Late Cretaceous relatives of rabbits, rodents, and other extant eutherian mammals. *Nature*, 414(6859):62–5. 24

[Archibald et al., 2010] Archibald, J. D., Clemens, W. A., Padian, K., Rowe, T., Macleod, N., Barrett, P. M., Gale, A., Holroyd, P., Sues, H.-D., Arens, N. C., Horner, J. R., Wilson, G. P., Goodwin, M. B., Brochu, C. A., Lofgren, D. L., Hurlbert, S. H., Hartman, J. H., Eberth, D. A., Wignall, P. B., Currie, P. J., Weil, A., Prasad, G. V. R., Dingus, L., Courtillot, V., Milner, A., Milner, A., Bajpai, S., Ward, D. J., and Sahni, A. (2010). Cretaceous extinctions: multiple causes. *Science*, 328(5981):973; author reply 975–6. 23

[Belle et al., 2002] Belle, E. M. S., Smith, N., and Eyre-Walker, A. (2002). Analysis of the phylogenetic distribution of isochores in vertebrates and a test of the thermal stability hypothesis. *Journal of Molecular Evolution*, 55(3):356–363. 47

[Bernardi, 1985] Bernardi, G. (1985). The Mosaic Genome of Warm-Blooded Vertebrates. *Science*, 228(4702):953–958. 44, 46, 47, 143

[Bernardi, 1990] Bernardi, G. (1990). Compositional transitions in the nuclear genomes of cold-blooded vertebrates. *Journal of Molecular Evolution*, 31(4):282–293. 46

[Bernardi, 2000] Bernardi, G. (2000). Isochores and the evolutionary genomics of vertebrates. *Gene*, 241(1):3–17. 46, 47

[Bernardi, 2007] Bernardi, G. (2007). The neoselectionist theory of genome evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 104(20):8385–8390. 47

[Biémont and Vieira, 2006] Biémont, C. and Vieira, C. (2006). Genetics: junk DNA as an evolutionary force. 41

[Bininda-Emonds et al., 2007] Bininda-Emonds, O. R. P., Cardillo, M., Jones, K. E., MacPhee, R. D. E., Beck, R. M. D., Grenyer, R., Price, S. a., Vos, R. a., Gittleman, J. L., and Purvis, A. (2007). The delayed rise of present-day mammals. *Nature*, 446(7135):507–12. 31

[Birdsell, 2002] Birdsell, J. A. (2002). Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. *Molecular Biology and Evolution*, 19(7):1181–1197. 53, 54

[Birney et al., 2004] Birney, E., Andrews, T. D., Bevan, P., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cuff, J., Curwen, V., Cutts, T., Down, T., Eyras, E., Fernandez-Suarez, X. M., Gane, P., Gibbins, B., Gilbert, J., Hammond, M., Hotz, H.-R., Iyer, V., Jekosch, K., Kahari, A., Kasprzyk, A., Keefe, D., Keenan, S., Lehvaslaiho, H., McVicker, G., Melsopp, C., Meidl, P., Mongin, E., Pettett, R., Potter, S., Proctor, G., Rae, M., Searle, S., Slater, G., Smedley, D., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Storey, R., Ureta-Vidal, A., Woodwark, K. C., Cameron, G., Durbin, R., Cox, A., Hubbard, T., and Clamp, M. (2004). An Overview of Ensembl. *Genome Research*, 14(5):925–928. 63

[Bromham, 2011] Bromham, L. (2011). The genome as a life-history character: why rate of molecular evolution varies between mammal species. *Philosophical Transactions of the Royal Society of London - Series B: Biological Sciences*, 366(1577):2503–2513. 69

[Brown and Jiricny, 1988] Brown, T. C. and Jiricny, J. (1988). Different base/base mispairs are corrected with different efficiencies and specificities in monkey kidney cells. *Cell*, 54(5):705–711. 51, 53, 54

[Burt and Trivers, 2006] Burt, A. and Trivers, R. (2006). *Genes in Conflict: The Biology of Selfish Genetic Elements*, volume 18. Belknap Press. 41, 55

[Burt, 2002] Burt, D. W. (2002). Origin and evolution of avian microchromosomes. *Cytogenetic and Genome Research*, 96(1-4):97–112. 142

[Capra and Pollard, 2011] Capra, J. A. and Pollard, K. S. (2011). Substitution Patterns Are GC-Biased in Divergent Sequences across the Metazoans. *Genome biology and evolution*, 3(0):516–527. 48, 59, 65

[Cargill et al., 1999] Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., Shaw, N., Lane, C. R., Lim, E. P., Kalyanaraman, N., Nemesh, J., Ziaugra, L., Friedland, L., Rolfe, A., Warrington, J., Lipshutz, R., Daley, G. Q., and Lander, E. S. (1999). Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genetics*, 22(3):231–238. 48

[Chen et al., 2007] Chen, J.-M., Cooper, D. N., Chuzhanova, N., Férec, C., and Patrinos, G. P. (2007). Gene conversion: mechanisms, evolution and human disease. *Nature Reviews Genetics*, 8(10):762–775. 51

[Chiari et al., 2012] Chiari, Y., Cahais, V., Galtier, N., and Delsuc, F. (2012). Phylogenomic analyses support the position of turtles as the sister group of birds and crocodiles (Archosauria). *BMC biology*, 10(1):65. 64

[Chicken and Sequencing, 2004] Chicken, I. and Sequencing, G. (2004). Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, 432(7018):695–716. 60, 142

[Churakov et al., 2009] Churakov, G., Kriegs, J. O., Baertsch, R., Zemann, A., Brosius, J., and Schmitz, J. (2009). Mosaic retroposon insertion patterns in placental mammals. *Genome Research*, 19(5):868–875. 33

[Ciccarelli et al., 2006] Ciccarelli, F. D., Doerks, T., Von Mering, C., Creevey, C. J., Snel, B., and Bork, P. (2006). Toward automatic reconstruction of a highly resolved tree of life. *Science*, 311(5765):1283–7. 21

[Costantini and Bernardi, 2008] Costantini, M. and Bernardi, G. (2008). Replication timing, chromosomal bands, and isochores. *Proceedings of the National Academy of Sciences of the United States of America*, 105(9):3433–3437. 44

[Costantini et al., 2006] Costantini, M., Clay, O., Auletta, F., and Bernardi, G. (2006). An isochore map of human chromosomes. *Genome Research*, 16(4):536–541. 45

[Dawkins, 1976] Dawkins, R. (1976). *The Selfish Gene*, volume 32. Oxford University Press. 38, 40, 56, 144

[Dawkins, 2004] Dawkins, R. (2004). *The Ancestor's Tale: A Pilgrimage to the Dawn of Evolution*. Houghton Mifflin Harcourt. 24, 34

[de Villena and Sapienza, 2001] de Villena, F. P.-M. and Sapienza, C. (2001). Recombination is proportional to the number of chromosome arms in mammals. *Mammalian Genome*. 60

[Delsuc et al., 2002] Delsuc, F., Scally, M., Madsen, O., Stanhope, M. J., De Jong, W. W., Catzeflis, F. M., Springer, M. S., and Douzery, E. J. P. (2002). Molecular phylogeny of living xenarthrans and the impact of character and taxon sampling on the placental tree rooting. *Molecular Biology and Evolution*, 19(10):1656–71. 34

[Doolittle and Sapienza, 1980] Doolittle, W. F. and Sapienza, C. (1980). Selfish genes, the phenotype paradigm and genome evolution. *Nature*, 284(5757):601–603. 41

[Duret, 2009] Duret, L. (2009). Mutation patterns in the human genome: more variable than expected. *PLoS biology*, 7(2):e1000028. 51, 61, 144

[Duret and Arndt, 2008] Duret, L. and Arndt, P. F. (2008). The impact of recombination on nucleotide substitutions in the human genome. {*PLoS*} *genetics*, 4(5). 44, 52

[Duret et al., 2008] Duret, L., Cohen, J., Jubin, C., Dessen, P., Goût, J.-F., Mousset, S., Aury, J.-M., Jaillon, O., Noël, B., Arnaiz, O., Bétermier, M., Wincker, P., Meyer, E., and Sperling, L. (2008). Analysis of sequence variability in the macronuclear DNA of Paramecium tetraurelia: A somatic view of the germline. *Genome research*, 18(4):585–596. 59

[Duret et al., 1995] Duret, L., Mouchiroud, D., and Gautier, C. (1995). Statistical analysis of vertebrate sequences reveals that long genes are scarce in GC-rich isochores. *Journal of Molecular Evolution*, 40(3):308–317. 44

[Duret et al., 2002] Duret, L., Semon, M., Piganeau, G., Mouchiroud, D., and Galtier, N. (2002). Vanishing GC-rich isochores in mammalian genomes. *Genetics*, 162(4):1837–1847. 48, 51

[Escobar et al., 2011] Escobar, J. S., Glémin, S., and Galtier, N. (2011). GC-Biased Gene Conversion Impacts Ribosomal DNA Evolution in Vertebrates, Angiosperms and Other Eukaryotes. *Molecular Biology*. 59, 65

[Eyre-Walker, 1993] Eyre-Walker, A. (1993). Recombination and mammalian genome evolution. *Proceedings of the Royal Society B: Biological Sciences*, 252(1335):237–243. 50, 51

[Eyre-Walker, 1999] Eyre-Walker, A. (1999). Evidence of selection on silent site base composition in mammals: potential implications for the evolution of isochores and junk DNA. *Genetics*, 152(2):675–683. 48, 51

[Eyre-Walker and Hurst, 2001] Eyre-Walker, A. and Hurst, L. D. (2001). The evolution of isochores. *Nature Reviews Genetics*, 2(7):549–555. 44, 48

[Feldhamer et al., 2007] Feldhamer, G. A., Drickamer, L. C., Vessey, S. H., Merritt, J. F., and Krajewski, C. (2007). *Mammalogy: Adaptation, Diversity, Ecology*. The Johns Hopkins University Press. 24, 34

[Felsenstein, 1988] Felsenstein, J. (1988). Phylogenies and Quantitative Characters. *Annual Review of Ecology and Systematics*, 19(1):445–471. 27

[Feschotte, 2008] Feschotte, C. (2008). Transposable elements and the evolution of regulatory networks. *Nature Reviews Genetics*, 9(5):397–405. 41

[Filipski, 1987] Filipski, J. (1987). Correlation between molecular clock ticking, codon usage fidelity of DNA repair, chromosome banding and chromatin compactness in germline cells. *FEBS Letters*, 217(2):184–186. 48

[Fryxell and Zuckerkandl, 2000] Fryxell, K. J. and Zuckerkandl, E. (2000). Cytosine deamination plays a primary role in the evolution of mammalian isochores. *Molecular Biology and Evolution*, 17(9):1371–1383. 48

[Fujita et al., 2011] Fujita, M. K., Edwards, S. V., and Ponting, C. P. (2011). The Anolis lizard genome: an amniote genome without isochores. *Genome biology and evolution*, pages evr072–. 59, 65, 68, 142

[Fullerton, 2001] Fullerton, S. M. (2001). Local rates of recombination are positively correlated with GC content in the human genome. *Mol. Biol. Evol.*, pages 1139–1142. 44, 52

[Galtier, 2003] Galtier, N. (2003). Gene conversion drives GC content evolution in mammalian histones. *Trends in Genetics*, 19(2):65–68. 51

[Galtier and Duret, 2007] Galtier, N. and Duret, L. (2007). Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. *Trends in genetics : TIG*, 23(6):273–7. 52, 57, 58

[Galtier et al., 2009] Galtier, N., Duret, L., Glémin, S., and Ranwez, V. (2009). GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. *Trends in Genetics*, 25(1):1–5. 56, 57

[Galtier and Lobry, 1997] Galtier, N. and Lobry, J. R. (1997). Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. *Journal of Molecular Evolution*, 44(6):632–636. 47

[Galtier et al., 2001] Galtier, N., Piganeau, G., Mouchiroud, D., and Duret, L. (2001). GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. 51

[Gerton et al., 2000] Gerton, J. L., DeRisi, J., Shroff, R., Lichten, M., Brown, P. O., and Petes, T. D. (2000). Global mapping of meiotic recombination hotspots and coldspots in the yeast Saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences of the United States of America*, 97(21):11383–90. 59

[Gibbs et al., 2004] Gibbs, R. A., Weinstock, G. M., Metzker, M. L., Muzny, D. M., Sodergren, E. J., Scherer, S., Scott, G., Steffen, D., Worley, K. C., Burch, P. E., Okwuonu, G., Hines, S., Lewis, L., DeRamo, C., Delgado, O., Dugan-Rocha, S., Miner, G., Morgan, M., Hawes, A., Gill, R., Celera, Holt, R. A., Adams, M. D., Amanatides, P. G., Baden-Tillson, H., Barnstead, M., Chin, S., Evans, C. A., Ferriera, S., Fosler, C., Glodek, A., Gu, Z., Jennings, D., Kraft, C. L., Nguyen, T., Pfannkoch, C. M., Sitter, C., Sutton, G. G., Venter, J. C., Woodage, T., Smith, D., Lee, H.-M., Gustafson, E., Cahill, P., Kana, A., Doucette-Stamm, L., Weinstock, K., Fechtel, K., Weiss, R. B., Dunn, D. M., Green, E. D., Blakesley, R. W., Bouffard, G. G., De Jong, P. J., Osoegawa, K., Zhu, B., Marra, M., Schein, J., Bosdet, I., Fjell, C., Jones, S., Krzywinski, M., Mathewson, C., Siddiqui, A., Wye, N., McPherson, J., Zhao, S., Fraser, C. M., Shetty, J., Shatsman, S., Geer, K., Chen, Y., Abramzon, S., Nierman, W. C., Havlak, P. H., Chen, R., Durbin, K. J., Egan, A., Ren, Y., Song, X.-Z., Li, B., Liu, Y., Qin, X., Cawley, S., Cooney, A. J., D'Souza,

L. M., Martin, K., Wu, J. Q., Gonzalez-Garay, M. L., Jackson, A. R., Kalafus, K. J., McLeod, M. P., Milosavljevic, A., Virk, D., Volkov, A., Wheeler, D. A., Zhang, Z., Bailey, J. A., Eichler, E. E., Tuzun, E., Birney, E., Mongin, E., Ureta-Vidal, A., Woodwark, C., Zdobnov, E., Bork, P., Suyama, M., Torrents, D., Alexandersson, M., Trask, B. J., Young, J. M., Huang, H., Wang, H., Xing, H., Daniels, S., Gietzen, D., Schmidt, J., Stevens, K., Vitt, U., Wingrove, J., Camara, F., Mar Albà, M., Abril, J. F., Guigo, R., Smit, A., Dubchak, I., Rubin, E. M., Couronne, O., Poliakov, A., Hübner, N., Ganten, D., Goesele, C., Hummel, O., Kreitler, T., Lee, Y.-A., Monti, J., Schulz, H., Zimdahl, H., Himmelbauer, H., Lehrach, H., Jacob, H. J., Bromberg, S., Gullings-Handley, J., Jensen-Seaman, M. I., Kwitek, A. E., Lazar, J., Pasko, D., Tonellato, P. J., Twigger, S., Ponting, C. P., Duarte, J. M., Rice, S., Goodstadt, L., Beatson, S. A., Emes, R. D., Winter, E. E., Webber, C., Brandt, P., Nyakatura, G., Adetobi, M., Chiaromonte, F., Elnitski, L., Eswara, P., Hardison, R. C., Hou, M., Kolbe, D., Makova, K., Miller, W., Nekrutenko, A., Riemer, C., Schwartz, S., Taylor, J., Yang, S., Zhang, Y., Lindpaintner, K., Andrews, T. D., Caccamo, M., Clamp, M., Clarke, L., Curwen, V., Durbin, R., Eyras, E., Searle, S. M., Cooper, G. M., Batzoglou, S., Brudno, M., Sidow, A., Stone, E. A., Payseur, B. A., Bourque, G., López-Otín, C., Puente, X. S., Chakrabarti, K., Chatterji, S., Dewey, C., Pachter, L., Bray, N., Yap, V. B., Caspi, A., Tesler, G., Pevzner, P. A., Haussler, D., Roskin, K. M., Baertsch, R., Clawson, H., Furey, T. S., Hinrichs, A. S., Karolchik, D., Kent, W. J., Rosenbloom, K. R., Trumbower, H., Weirauch, M., Cooper, D. N., Stenson, P. D., Ma, B., Brent, M., Arumugam, M., Shteynberg, D., Copley, R. R., Taylor, M. S., Riethman, H., Mudunuri, U., Peterson, J., Guyer, M., Felsenfeld, A., Old, S., Mockrin, S., and Collins, F. (2004). Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature*, 428(6982):493–521. 65

[Glémin et al., 2006] Glémin, S., Bazin, E., and Charlesworth, D. (2006). Impact of mating systems on patterns of sequence polymorphism in flowering plants. *Proceedings of the Royal Society B Biological Sciences*, 273(1604):3011–3019.

59

[Gregory et al., 2007] Gregory, T. R., Nicol, J. a., Tamm, H., Kullman, B., Kullman, K., Leitch, I. J., Murray, B. G., Kapraun, D. F., Greilhuber, J., and Bennett, M. D. (2007). Eukaryotic genome size databases. *Nucleic acids research*, 35(Database issue):D332–8. 64

[Hallström and Janke, 2008] Hallström, B. M. and Janke, A. (2008). Resolution among major placental mammal interordinal relationships with genome data imply that speciation influenced their earliest radiations. *BMC Evolutionary Biology*, 8(1):162. 33

[Hallström et al., 2007] Hallström, B. M., Kullberg, M., Nilsson, M. A., and Janke, A. (2007). Phylogenomic data analyses provide evidence that Xenarthra and Afrotheria are sister groups. *Molecular Biology and Evolution*, 24(9):2059–2068. 33

[Hamada et al., 2003] Hamada, K., Horiike, T., Ota, H., Mizuno, K., and Shinozawa, T. (2003). Presence of isochore structures in reptile genomes suggested by the relationship between GC contents of intron regions and those of coding regions. *Genes genetic systems*, 78(2):195–198. 47

[Hedges et al., 2006] Hedges, S. B., Dudley, J., and Kumar, S. (2006). Time-Tree: a public knowledge-base of divergence times among organisms. *Bioinformatics*, 22(23):2971–2972. 33

[Hellmann et al., 2003] Hellmann, I., Ebersberger, I., Ptak, S. E., Pääbo, S., and Przeworski, M. (2003). A neutral explanation for the correlation of diversity with recombination rates in humans. *The American Journal of Human Genetics*, 72(6):1527–1535. 48

[Huchon et al., 2002] Huchon, D., Madsen, O., Sibbald, M. J. J. B., Ament, K., Stanhope, M. J., Catzeflis, F., De Jong, W. W., and Douzery, E. J. P. (2002). Rodent phylogeny and a timescale for the evolution of Glires: evidence from an extensive taxon sampling using three nuclear genes. *Molecular Biology and Evolution*, 19(7):1053–65. 33

[Hughes et al., 1999] Hughes, S., Zelus, D., and Mouchiroud, D. (1999). Warm-blooded isochore structure in Nile crocodile and turtle. *Molecular Biology and Evolution*, 16(11):1521–1527. 47

[Hurst and Werren, 2001] Hurst, G. D. and Werren, J. H. (2001). The role of selfish genetic elements in eukaryotic evolution. *Nature Reviews Genetics*, 2(8):597–606. 41

[Hurst and Merchant, 2001] Hurst, L. D. and Merchant, A. R. (2001). High guanine-cytosine content is not an adaptation to high temperature: a comparative analysis amongst prokaryotes. *Proceedings of the Royal Society B: Biological Sciences*, 268(1466):493–7. 47

[Jø rgensen et al., 2006] Jø rgensen, F. G., Schierup, M. H., and Clark, A. G. (2006). Heterogeneity in regional GC content and differential usage of codons and. *Molecular Biology and Evolution*. 59

[Johnson, 2007] Johnson, L. J. (2007). The genome strikes back: The evolutionary importance of defence against mobile elements. *Evolutionary Biology*, 34(3-4):121–129. 41

[Jukes and Cantor, 1969] Jukes, T. H. and Cantor, C. R. (1969). Evolution of protein molecules. In Munro, H. N., editor, *Mammalian Protein Metabolism*, volume 3 of *Mammalian protein metabolism*, chapter 24, pages 21–132. Academic Press. 27

[Kazanskaya et al., 1997] Kazanskaya, O. V., Severtzova, E. A., Barth, K. A., Ermakova, G. V., Lukyanov, S. A., Benyumov, A. O., Pannese, M., Boncinelli, E., Wilson, S. W., and Zaraisky, A. G. (1997). Methylation patterns in the isochores of vertebrate genomes. *Gene*, 205(1-2):25–34. 44

[Keller et al., 2010] Keller, G., Adatte, T., Pardo, A., Bajpai, S., Khosla, A., and Samant, B. (2010). Cretaceous extinctions: evidence overlooked. 23

[Keller et al., 2004] Keller, G., Adatte, T., Stinnesbeck, W., Rebolledo-Vieyra, M., Urrutia Fucugauchi, J., Kramar, U., and Stüben, D. (2004). Chicxulub impact predates the K-T boundary mass extinction. *Proceedings of the National*

Academy of Sciences of the United States of America, 101(11):3753–3758. 23

[Kimura, 1968] Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature*. 38

[Kimura, 1983] Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge University Press. 37, 47

[King and Jukes, 1969] King, J. and Jukes, T. (1969). Non-darwinian evolution. *Science*. 38

[Kjer and Honeycutt, 2007] Kjer, K. M. and Honeycutt, R. L. (2007). Site specific rates of mitochondrial genomes and the phylogeny of eutheria. *BMC Evolutionary Biology*, 7(1):8. 33

[Konu and Li, 2002] Konu, O. and Li, M. D. (2002). Correlations between mRNA expression levels and GC contents of coding and untranslated regions of genes in rodents. *Journal of Molecular Evolution*, 54(1):35–41. 44

[Kriegs et al., 2006] Kriegs, J. O., Churakov, G., Kiefmann, M., Jordan, U., Brosius, J., and Schmitz, J. (2006). Retroposed Elements as Archives for the Evolutionary History of Placental Mammals. *PLoS Biology*, 4(4):e91. 33

[Kudla et al., 2004] Kudla, G., Helwak, A., and Lipinski, L. (2004). Gene conversion and GC-content evolution in mammalian Hsp70. *Molecular Biology and Evolution*, 21(7):1438–1444. 51

[Kuraku et al., 2006] Kuraku, S., Ishijima, J., Nishida-Umehara, C., Agata, K., Kuratani, S., and Matsuda, Y. (2006). cDNA-based gene mapping and GC3 profiling in the soft-shelled turtle suggest a chromosomal size-dependent GC bias shared by sauropsids. *Chromosome research an international journal on the molecular supramolecular and evolutionary aspects of chromosome biology*, 14(2):187–202. 47

[Lander et al., 2001] Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, a., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, a., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, a., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, a., Deadman, R., Deloukas, P., Dunham, a., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, a., Jones, M., Lloyd, C., McMurray, a., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, a., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. a., Mardis, E. R., Fulton, L. a., Chinwalla, a. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, a., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, a., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. a., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, a., Hattori, M., Yada, T., Toyoda, a., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, a., Platzer, M., Nyakatura, G., Taudien, S., Rump, a., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, a., Qin, S., Davis, R. W., Federspiel, N. a., Abola, a. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. a., Athanasiou, M., Schultz, R., Roe, B. a., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. a., Bateman, a., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. a., Kasif, S., Kaspryzk, a., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, a., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, a. F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, a., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, a., Wetterstrand, K. a., Patrinos, a., Morgan, M. J., de Jong, P., Catanese, J. J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y. J., and Szustakowki, J. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921. 38, 44, 60

[Lartillot and Poujol, 2011] Lartillot, N. and Poujol, R. (2011). A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters. *Molecular Biology and Evolution*, 28(1):729–744. 141

[Lercher and Hurst, 2002] Lercher, M. J. and Hurst, L. D. (2002). Human SNP variability and mutation rate are higher in regions of high recombination. *Trends in Genetics*, 18(7):337–340. 48

[Li et al., 2008] Li, M.-K., Gu, L., Chen, S.-S., Dai, J.-Q., and Tao, S.-H. (2008). Evolution of the isochore structure in the scale of chromosome: insight from the mutation bias and fixation bias. *Journal of Evolutionary Biology*, 21(1):173–182. 52

[Liu et al., 2001] Liu, F. G., Miyamoto, M. M., Freire, N. P., Ong, P. Q., Tennant, M. R., Young, T. S., and Gugel, K. F. (2001). Molecular and morphological supertrees for eutherian (placental) mammals. *Science*, 291(5509):1786–1789. 28

[Lunter et al., 2006] Lunter, G., Ponting, C. P., and Hein, J. (2006). Genome-Wide Identification of Human Functional DNA Using a Neutral Indel Model. *PLoS Computational Biology*, 2(1):11. 53

[Lynch et al., 2006] Lynch, M., Koskella, B., and Schaack, S. (2006). Mutation pressure and the evolution of organelle genomic architecture. *Science*, 311(5768):1727–30. 62

[Lynch and Walsh, 2007] Lynch, M. and Walsh, B. (2007). *The Origins of Genome Architecture*. Sinauer Associates Inc.,U.S. 43, 54

[Macleod et al., 1997] Macleod, N., Rawson, P., Forey, P., Banner, F., BoudagherFadel, M., Bown, P., Burnett, J., Chambers, P., Culver, S., Evans, S., Jeffery, C., Kaminski, M., Lord, A., Milner, A., Milner, A., Morris, N., Owen, E., Rosen, B., Smith, A., Taylor, P., Urquhart, E., and Young (1997). The Cretaceous-Tertiary biotic transition. *Journal of the Geological Society*, 154(2):265–292. 23

[Madsen et al., 2001] Madsen, O., Scally, M., Douady, C. J., Kao, D. J., DeBry, R. W., Adkins, R., Amrine, H. M., Stanhope, M. J., De Jong, W. W., and Springer, M. S. (2001). Parallel adaptive radiations in two major clades of placental mammals. *Nature*, 409(6820):610–614. 30, 34

[Mancera et al., 2008] Mancera, E., Bourgon, R., Brozzi, A., Huber, W., and Steinmetz, L. M. (2008). High-resolution mapping of meiotic crossovers and non-crossovers in yeast. *Nature*, 454(7203):479–485. 53

[Marais et al., 2001] Marais, G., Mouchiroud, D., and Duret, L. (2001). Does recombination improve selection on codon usage? Lessons from nematode and fly complete genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 98(10):5688–92. 59

[Marais et al., 2003] Marais, G., Mouchiroud, D., and Duret, L. (2003). Neutral effect of recombination on base composition in Drosophila. *Genetical Research*, 81(2):79–87. 59

[McCormack et al., 2012] McCormack, J. E., Faircloth, B. C., Crawford, N. G., Gowaty, P. A., Brumfield, R. T., and Glenn, T. C. (2012). Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis. *Genome research*. 34

[McQueen et al., 1998] McQueen, H. A., Siriaco, G., and Bird, A. P. (1998). Chicken microchromosomes are hyperacetylated, early replicating, and gene rich. *Genome Research*, 8(6):621–630. 142

[Meredith et al., 2011] Meredith, R. W., Janecka, J. E., Gatesy, J., Ryder, O. A., Fisher, C. A., Teeling, E. C., Goodbla, A., Eizirik, E., Simão, T. L. L., Stadler, T., Rabosky, D. L., Honeycutt, R. L., Flynn, J. J., Ingram, C. M., Steiner, C., Williams, T. L., Robinson, T. J., Burk-Herrick, A., Westerman, M., Ayoub, N. A., Springer, M. S., and Murphy, W. J. (2011). Impacts of the Cretaceous Terrestrial Revolution and KPg Extinction on Mammal Diversification. *Science*, 334(6055):521–4. 31, 33

[Meunier and Duret, 2004] Meunier, J. and Duret, L. (2004). Recombination drives the evolution of GC-content in the human genome. *Molecular Biology and Evolution*, 21(6):984–990. 51

[Montoya-Burgos et al., 2003] Montoya-Burgos, J. I., Boursot, P., and Galtier, N. (2003). Recombination explains isochores in mammalian genomes. *Trends in Genetics*, 19(3):128–130. 51, 53, 57

[Mouchiroud et al., 1988] Mouchiroud, D., Gautier, C., and Bernardi, G. (1988). The compositional distribution of coding sequences and DNA molecules in humans and murids. *Journal of Molecular Evolution*, 27(4):311–320. 65

[Murphy et al., 2001] Murphy, W. J., Eizirik, E., Johnson, W. E., Zhang, Y. P., Ryder, O. A., and O'Brien, S. J. (2001). Molecular phylogenetics and the origins of placental mammals. *Nature*, 409(6820):614–618. 30, 34

[Murphy et al., 2007] Murphy, W. J., Pringle, T. H., Crider, T. A., Springer, M. S., and Miller, W. (2007). Using genomic data to unravel the root of the placental mammal phylogeny. *Genome Research*, 17(4):413–421. 33

[Muyle et al., 2011] Muyle, A., Serres-Giardi, L., Ressayre, A., Escobar, J., and Glémin, S. (2011). GC-Biased Gene Conversion and Selection Affect GC Content in the Oryza Genus (rice). *Molecular Biology and Evolution*, 28(9):2695–2706. 59

[Nabholz et al., 2008] Nabholz, B., Glémin, S., and Galtier, N. (2008). Strong variations of mitochondrial mutation rate across mammals–the longevity hypothesis. *Molecular Biology and Evolution*, 25(1):120–30. 69

[Nagylaki, 1983] Nagylaki, T. (1983). Evolution of a finite population under gene conversion. *Proceedings of the National Academy of Sciences of the United States of America*, 80(20):6278–6281. 58

[Necşulea et al., 2011] Necşulea, A., Popa, A., Cooper, D. N., Stenson, P. D., Mouchiroud, D., Gautier, C., and Duret, L. (2011). Meiotic recombination favors the spreading of deleterious mutations in human populations. *Human mutation*, 32(2):198–206. 57

[Nikolaev et al., 2007] Nikolaev, S. I., Montoya-Burgos, J. I., Popadin, K., Parand, L., Margulies, E. H., and Antonarakis, S. E. (2007). Life-history traits drive the evolutionary rates of mammalian coding and noncoding genomic elements. *Proceedings of the National Academy of Sciences of the United States of America*, 104(51):20443–8. 34, 69

[Nishihara et al., 2009] Nishihara, H., Maruyama, S., and Okada, N. (2009). Retroposon analysis and recent geological data suggest near-simultaneous divergence of the three superorders of mammals. *Proceedings of the National Academy of Sciences of the United States of America*, 106(13):5235–40. 34

[Novacek, 1986] Novacek, M. J. (1986). The skull of Leptictid insectivorans and the higher-level classification of Eutheiran mammals. *Bulletin of the American Museum of Natural History*, 183(1):1–111. 34

[Novacek, 1992] Novacek, M. J. (1992). Mammalian phylogeny: shaking the tree. *Nature*, 356(6365):121–5. 28

[Novick et al., 2011] Novick, P. A., Smith, J. D., Floumanhaft, M., Ray, D. A., and Boissinot, S. (2011). The Evolution and Diversity of DNA Transposons in the Genome of the Lizard Anolis carolinensis. *Genome biology and evolution*, 3(0):1–14. 67, 68, 143

[Orgel and Crick, 1980] Orgel, L. E. and Crick, F. H. (1980). Selfish DNA: the ultimate parasite. *Nature*, 284(5757):604–7. 41

[Östergren, 1945] Östergren, G. (1945). Parasitic nature of extra fragment chromosomes. *Botaniska Notiser*. 40

[Perry and Ashworth, 1999] Perry, J. and Ashworth, A. (1999). Evolutionary rate of a gene affected by chromosomal position. *Current Biology*, 9(17):987–989. 52

[Pessia et al., 2012] Pessia, E., Popa, A., Mousset, S., Rezvoy, C., Duret, L., and Marais, G. A. B. (2012). Evidence for widespread GC-biased gene conversion in eukaryotes. *Genome Biology and Evolution*, 4(7):1–20. 59, 65

[Pollard et al., 2006] Pollard, D. A., Iyer, V. N., Moses, A. M., and Eisen, M. B. (2006). Widespread discordance of gene trees with species tree in Drosophila: evidence for incomplete lineage sorting. *PLoS genetics*, 2(10):e173. 58

[Popadin et al., 2007] Popadin, K., Polishchuk, L. V., Mamirova, L., Knorre, D., and Gunbin, K. (2007). Accumulation of slightly deleterious mutations in mitochondrial protein-coding genes of large versus small mammals. *Proceedings of the National Academy of Sciences of the United States of America*, 104(33):13390–5. 69

[Prabhakar et al., 2006] Prabhakar, S., Noonan, J. P., Pääbo, S., and Rubin, E. M. (2006). Accelerated evolution of conserved noncoding sequences in humans. *Science*, 314(5800):786. 58

[Prabhakar et al., 2008] Prabhakar, S., Visel, A., Akiyama, J. A., Shoukry, M., Lewis, K. D., Holt, A., Plajzer-Frick, I., Morrison, H., Fitzpatrick, D. R., Afzal, V., Pennacchio, L. A., Rubin, E. M., and Noonan, J. P. (2008). Human-specific gain of function in a developmental enhancer. *Science*, 321(5894):1346–1350. 58

[Prasad and Allard, 2008] Prasad, A. B. and Allard, M. W. (2008). Confirming the phylogeny of mammals by use of large comparative sequence data sets. *Molecular Biology and Evolution*, 25(9):1795. 33

[Ream et al., 2003] Ream, R. A., Johns, G. C., and Somero, G. N. (2003). Base compositions of genes encoding alpha-actin and lactate dehydrogenase-A from differently adapted vertebrates show no temperature-adaptive variation in G + C content. *Molecular Biology and Evolution*, 20(1):105–110. 47

[Robertson et al., 2004] Robertson, D. S., McKenna, M. C., Toon, O. B., Hope, S., and Lillegraven, J. A. (2004). Survival in the first hours of the Cenozoic. *Geological Society Of America Bulletin*, 116(5):760. 24, 34

[Rose, 1982] Rose, K. D. (1982). Skeleton of diacodexis, oldest known artiodactyl. *Science*, 216(4546):621–623. 142

[Schulte et al., 2010] Schulte, P., Alegret, L., Arenillas, I., Arz, J. a., Barton, P. J., Bown, P. R., Bralower, T. J., Christeson, G. L., Claeys, P., Cockell, C. S., Collins, G. S., Deutsch, A., Goldin, T. J., Goto, K., Grajales-Nishimura, J. M., Grieve, R. a. F., Gulick, S. P. S., Johnson, K. R., Kiessling, W., Koeberl, C., Kring, D. a., MacLeod, K. G., Matsui, T., Melosh, J., Montanari, A., Morgan, J. V., Neal, C. R., Nichols, D. J., Norris, R. D., Pierazzo, E., Ravizza, G., Rebolledo-Vieyra, M., Reimold, W. U., Robin, E., Salge, T., Speijer, R. P., Sweet, A. R., Urrutia-Fucugauchi, J., Vajda, V., Whalen, M. T., and Willumsen, P. S. (2010). The Chicxulub asteroid impact and mass extinction at the Cretaceous-Paleogene boundary. *Science (New York, N.Y.)*, 327(5970):1214–8. 23

[Sekita et al., 2008] Sekita, Y., Wagatsuma, H., Nakamura, K., Ono, R., Kagami, M., Wakisaka, N., Hino, T., Suzuki-Migishima, R., Kohda, T., Ogura, A., Ogata, T., Yokoyama, M., Kaneko-Ishino, T., and Ishino, F. (2008). Role of retrotransposon-derived imprinted gene, Rtl1, in the feto-maternal interface of mouse placenta. *Nature Genetics*, 40(2):243–248. 53

[Serres-Giardi et al., 2012] Serres-Giardi, L., Belkhir, K., David, J., and Glemin, S. (2012). Patterns and Evolution of Nucleotide Landscapes in Seed Plants. *the Plant Cell Online*, 24(4):1–20. 59

[Shoshani and McKenna, 1998] Shoshani, J. and McKenna, M. C. (1998). Higher taxonomic relationships among extant mammals based on morphology, with selected comparisons of results from molecular data. *Molecular Phylogenetics and Evolution*, 9(3):572–584. 28, 33

[Signor and Lipps, 1982] Signor, P. W. and Lipps, J. H. (1982). Sampling bias, gradual extinction patterns and catastrophes in the fossil record. *Geological Society of America Special Paper*, 190(190):291–296. 36

[Sinzelle et al., 2009] Sinzelle, L., Izsvák, Z., and Ivics, Z. (2009). Molecular domestication of transposable elements: from detrimental parasites to useful host genes. *Cellular and molecular life sciences CMLS*, 66(6):1073–1093. 41

[Smit, 1999] Smit, A. (1999). Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Current opinion in genetics & development*, 9(6):657–63. 46

[Smith et al., 2010] Smith, F. a., Boyer, A. G., Brown, J. H., Costa, D. P., Dayan, T., Ernest, S. K. M., Evans, A. R., Fortelius, M., Gittleman, J. L., Hamilton, M. J., Harding, L. E., Lintulaakso, K., Lyons, S. K., McCain, C., Okie, J. G., Saarinen, J. J., Sibly, R. M., Stephens, P. R., Theodor, J., and Uhen, M. D. (2010). The evolution of maximum body size of terrestrial mammals. *Science*, 330(6008):1216–9. 24, 34

[Smith and Eyre-Walker, 2001] Smith, N. G. C. and Eyre-Walker, A. (2001). Synonymous Codon Bias Is Not Caused by Mutation Bias in G+C-Rich Genes in Humans. *Mol. Biol. Evol.*, 18(6):982–986. 48

[Song et al., 2012] Song, S., Liu, L., Edwards, S. V., and Wu, S. (2012). Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. *Proceedings of the National Academy of Sciences*. 33

[Spencer et al., 2006] Spencer, C. C. A., Deloukas, P., Hunt, S., Mullikin, J., Myers, S., Silverman, B., Donnelly, P., Bentley, D., and McVean, G. (2006). The Influence of Recombination on Human Genetic Diversity. *PLoS Genetics*, 2(9):11. 48, 51

[Springer et al., 2003] Springer, M. S., Murphy, W. J., Eizirik, E., and O'Brien, S. J. (2003). Placental mammal diversification and the Cretaceous-Tertiary boundary. *Proceedings of the National Academy of Sciences of the United States of America*, 100(3):1056–61. 31

[Springer et al., 2004] Springer, M. S., Stanhope, M. J., Madsen, O., and de Jong, W. W. (2004). Molecules consolidate the placental mammal tree. *Trends in ecology & evolution*, 19(8):430–8. 29, 32

[SUEOKA, 1962] SUEOKA, N. (1962). On the genetic basis of variation and heterogeneity of DNA base composition. *Proceedings of the National Academy of Sciences of the United States of America*, 48:582–92. 43

[Sumiyama and Saitou, 2011] Sumiyama, K. and Saitou, N. (2011). Loss-of-function mutation in a repressor module of human-specifically activated enhancer HACNS1. *Molecular biology and evolution*, 28(11):3005–7. 58

[Surtees et al., 2004] Surtees, J. A., Argueso, J. L., and Alani, E. (2004). Mismatch repair proteins: key regulators of genetic recombination. *Cytogenetic and Genome Research*, 107(3-4):146–159. 52

[Suzuki et al., 2007] Suzuki, S., Ono, R., Narita, T., Pask, A. J., Shaw, G., Wang, C., Kohda, T., Alsop, A. E., Marshall Graves, J. A., Kohara, Y., Ishino, F., Renfree, M. B., and Kaneko-Ishino, T. (2007). Retrotransposon Silencing by DNA Methylation Can Drive Mammalian Genomic Imprinting. *PLoS Genetics*, 3(4):7. 53

[Tavaré, 1986] Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences*, 17(2):57–86. 27

[Waddell et al., 2001] Waddell, P. J., Kishino, H., and Ota, R. (2001). A phylogenetic foundation for comparative mammalian genomics. *Genome informatics International Conference on Genome Informatics*, 12(0919-9454 LA - eng PT - Journal Article SB - IM):141–154. 33

[Waddell and Shelley, 2003] Waddell, P. J. and Shelley, S. (2003). Evaluating placental inter-ordinal phylogenies with novel sequences including RAG1, gamma-fibrinogen, ND6, and mt-tRNA, plus MCMC-driven nucleotide, amino acid, and codon models. *Molecular Phylogenetics and Evolution*, 28(2):197–224. 34

[Warren et al., 2008] Warren, W. C., Hillier, L. W., Marshall Graves, J. A., Birney, E., Ponting, C. P., Grützner, F., Belov, K., Miller, W., Clarke, L., Chinwalla, A. T., Yang, S.-P., Heger, A., Locke, D. P., Miethke, P., Waters, P. D., Veyrunes, F., Fulton, L., Fulton, B., Graves, T., Wallis, J., Puente, X. S., López-Otín, C., Ordóñez, G. R., Eichler, E. E., Chen, L., Cheng, Z., Deakin, J. E., Alsop, A., Thompson, K., Kirby, P., Papenfuss, A. T., Wakefield, M. J., Olender, T., Lancet, D., Huttley, G. A., Smit, A. F. A., Pask, A., Temple-Smith, P., Batzer, M. A., Walker, J. A., Konkel, M. K., Harris, R. S., Whittington, C. M., Wong, E. S. W., Gemmell, N. J., Buschiazzo, E., Vargas Jentzsch, I. M., Merkel, A., Schmitz, J., Zemann, A., Churakov, G., Kriegs, J. O., Brosius, J., Murchison, E. P., Sachidanandam, R., Smith, C., Hannon, G. J., Tsend-Ayush, E., McMillan, D., Attenborough, R., Rens, W., Ferguson-Smith, M., Lefèvre, C. M., Sharp, J. A., Nicholas, K. R., Ray, D. A., Kube, M., Reinhardt, R., Pringle, T. H., Taylor, J., Jones, R. C., Nixon, B., Dacheux, J.-L., Niwa, H., Sekita, Y., Huang, X., Stark, A., Kheradpour, P., Kellis, M., Flicek, P., Chen, Y., Webber, C., Hardison, R., Nelson, J., Hallsworth-Pepin, K., Delehaunty, K., Markovic, C., Minx, P., Feng, Y., Kremitzki, C., Mitreva, M., Glasscock, J., Wylie, T., Wohldmann, P., Thiru, P., Nhan, M. N., Pohl, C. S., Smith, S. M., Hou, S., Nefedov, M., De Jong, P. J., Renfree, M. B., Mardis, E. R., and Wilson, R. K. (2008). Genome analysis of the platypus reveals unique signatures of evolution. *Nature*, 453(7192):175–183. 62

[Watson and Crick, 1953] Watson, J. and Crick, F. (1953). Molecular structure of nucleic acids. *Nature*. 24

[Webster et al., 2006] Webster, M. T., Axelsson, E., and Ellegren, H. (2006). Strong regional biases in nucleotide substitution in the chicken genome. *Molecular Biology and Evolution*, 23(6):1203–1216. 51

[Webster and Smith, 2004] Webster, M. T. and Smith, N. G. C. (2004). Fixation biases affecting human SNPs. *Trends in Genetics*, 20(3):122–126. 48, 51

[Werren, 2011] Werren, J. H. (2011). Selfish genetic elements, genetic conflict, and evolutionary innovation. *Proceedings of the National Academy of Sciences of the United States of America*, 108 Suppl:10863–70. 41

[Werren et al., 1988] Werren, J. H., Nur, U., and Wu, C. I. (1988). Selfish genetic elements. *Trends in Ecology & Evolution*, 3(11):297–302. 41

[Whitby, 2005] Whitby, M. C. (2005). Making crossovers during meiosis. *Biochemical Society Transactions*, 33(Pt 6):1451–1455. 49

[Wildman et al., 2007] Wildman, D. E., Uddin, M., Opazo, J. C., Liu, G., Lefort, V., Guindon, S., Gascuel, O., Grossman, L. I., Romero, R., and Goodman, M. (2007). Genomics, biogeography, and the diversification of placental mammals. *Proceedings of the National Academy of Sciences of the United States of America*, 104(36):14395–400. 33

[Williams, 1966] Williams, G. C. (1966). *Adaptation and Natural Selection*, volume 1996 re-pr of *Princeton science library*. Princeton University Press. 40, 56

[Wilson and Reeder, 2005] Wilson, D. E. and Reeder, D. M. (2005). *Mammal species of the world: a taxonomic and geographic reference*, volume 1. Johns Hopkins University Press. 22

[Wolfe et al., 1989] Wolfe, K. H., Sharp, P. M., and Li, W. H. (1989). Mutation rates differ among regions of the mammalian genome. *Nature*, 337(6204):283–285. 48

[Yang and Rannala, 2012] Yang, Z. and Rannala, B. (2012). Molecular phylogenetics: principles and practice. *Nature Reviews Genetics*, 13(May):303–314. 25, 27

Cinquième partie

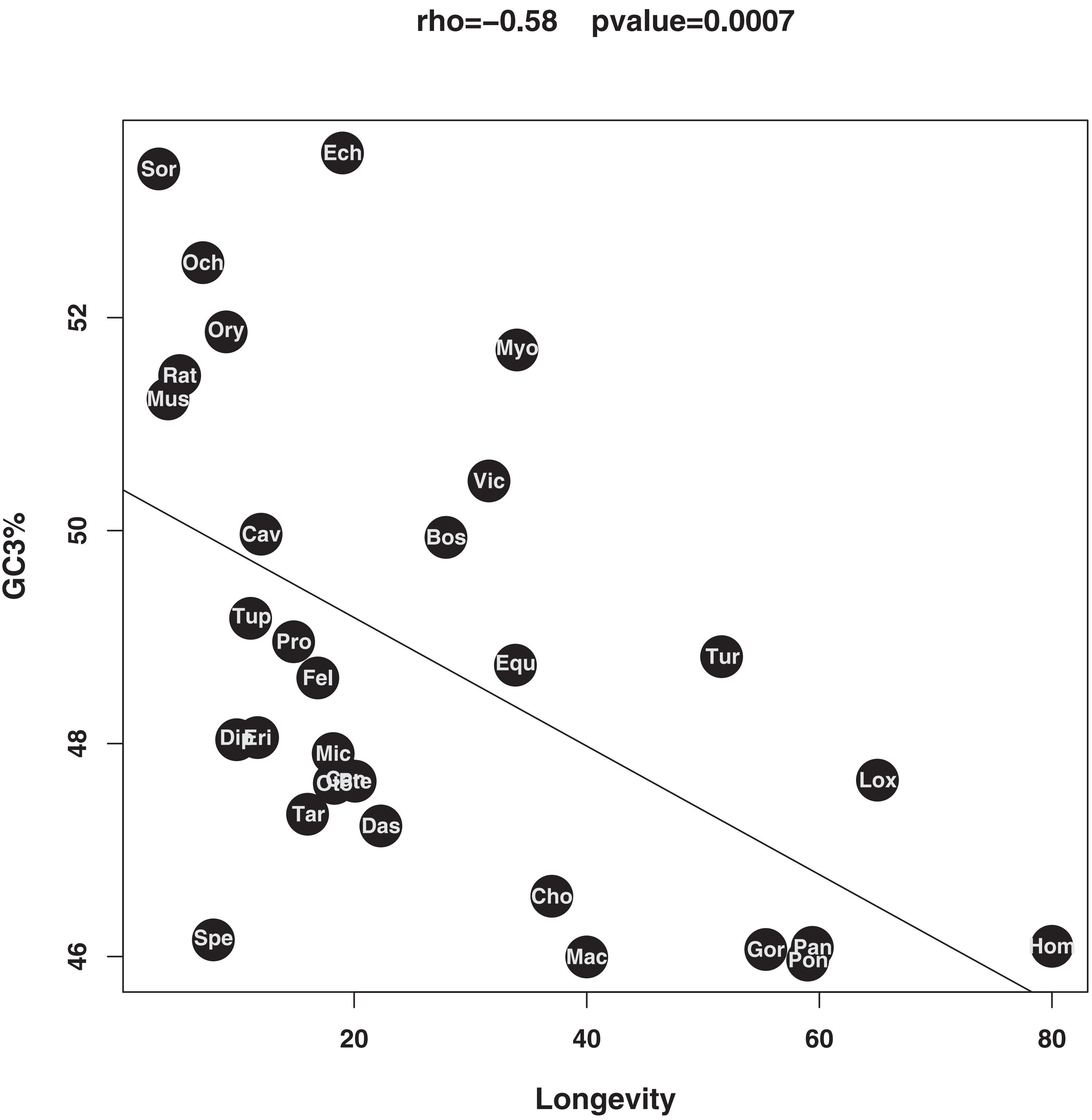
Annexes

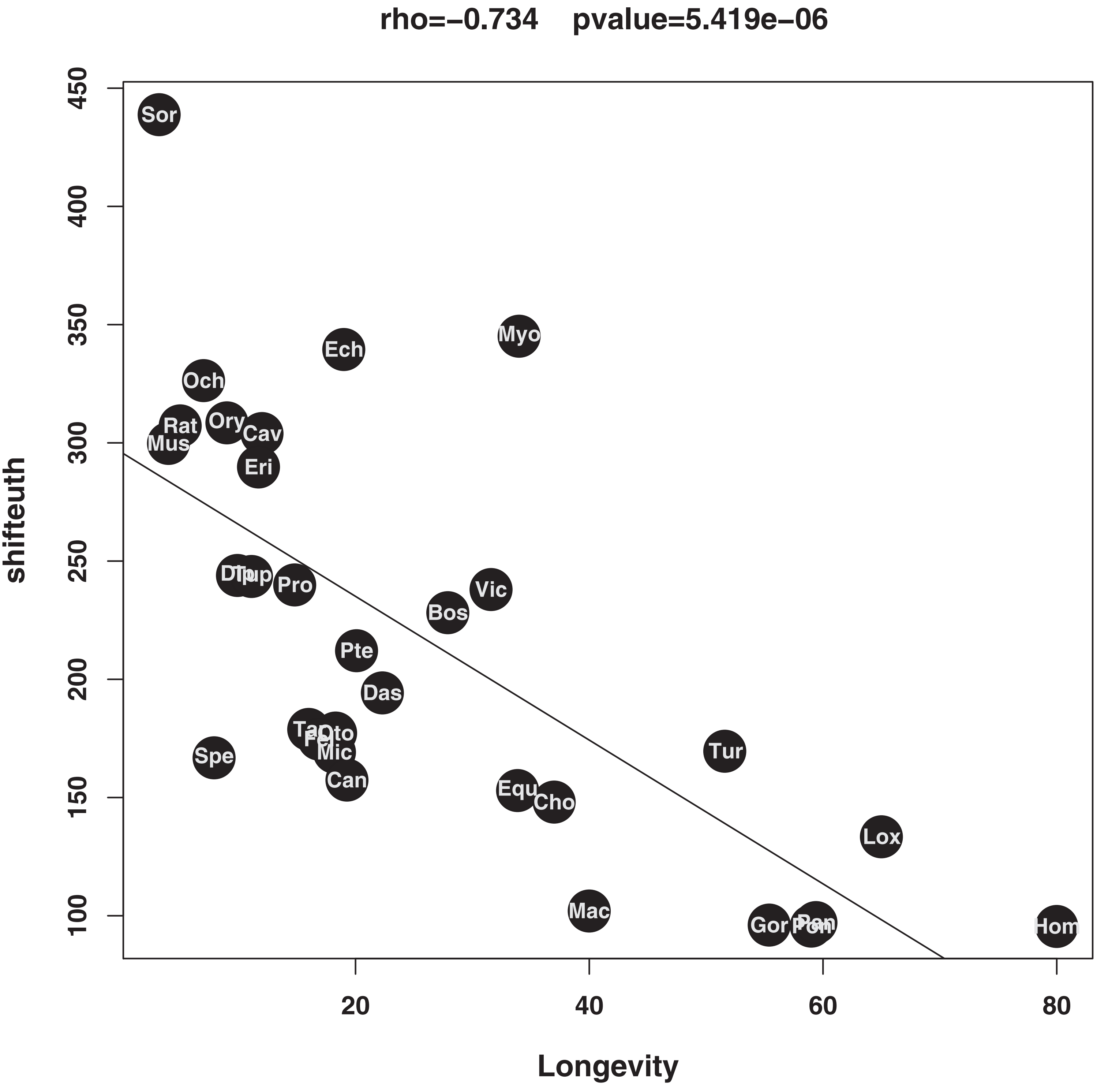
6.1 Annexes Chapitre 3

- Supplemental data : Table S1
- Correlation between longevity and GC3%
- Correlation between longevity and Di,anc (=shifteuth)
- Poster smbe 2010

Species	Common name	Abbrev.	Mean GC%	Standard deviation GC%	Di,anc
Choloepus hoffmanni	Sloth	Cho	41.12	6.98	96.12
Dasypus novemcinctus	Armadillo	Das	42.02	7.78	118.39
Echinops telfairi	Tenrec	Ech	45.69	7.59	223.08
Loxodonta africana	Elephant	Lox	42.36	7.09	96.69
Procavia capensis	Hyrax	Pro	42.99	6.85	141.49
Tupaia belangeri	Tree shrew	Tup	43.25	7.75	149.58
Homo sapiens	Human	Hom	42.08	7.55	70.56
Pan troglodytes	Chimp	Pan	41.97	7.47	71.12
Gorilla gorilla	Gorilla	Gor	42.45	7.52	74.85
Pongo pygmaeus	Orangutan	Pon	41.99	7.48	69.3
Macaca mulatta	Macaque	Mac	42.03	7.47	73.37
Tarsius syrichta	Tarsier	Tar	42.15	7.49	128.13
Microcebus murinus	Mouse lemur	Mic	42.44	7.74	103.49
Otolemur garnettii	Bushbaby	Oto	42.32	6.73	117.61
Oryctolagus cuniculus	Rabbit	Ory	45.34	8.65	224.62
Ochotona princeps	Pika	Och	45.61	8.06	215.92
Spermophilus tridecemlineatus	Squirrell	Spe	41.71	7.14	124.03
Cavia porcellus	Guinea pig	Cav	43.14	7.60	171.15
Dipodomys ordii	Kangaroo rat	Dip	42.65	7.65	149.14
Rattus norvegicus	Rat	Rat	44.39	5.97	206.18
Mus musculus	Mouse	Mus	44.27	6.31	193.5
Erinaceus europeaus	Hedgehog	Eri	42.02	7.80	162.73
Sorex araneus	Shrew	Sor	43.56	9.08	200.24
Bos taurus	Cow	Bos	42.86	7.54	116.12
Tursiops truncatus	Dolphin	Tur	42.8	7.46	98.35
Vicugna pacos	Alpaca	Vic	42.75	6.91	134.78
Myotis lucifugus	Microbat	Myo	44.2	8.10	183.28
Pteropus vampyrus	Megabat	Pte	41.93	7.35	112.66
Equus caballus	Horse	Equ	43.32	7.62	109.23
Canis familiaris	Dog	Can	43.93	8.58	156.95
Felis catus	Cat	Fel	43.62	7.12	138.52
Eutherian ancestor	-	-	41.85	8.34	-

Table S1 : Evolutionary dynamics of the GC-content of noncoding flanking regions in 33 mammals


rho=−0.58   pvalue=0.0007
GC3%
Longevity
46
48
50
52
20
40
60
80
Ech
Sor
Och
Ory
Myo
Rat
Mus
Vic
Cav
Bos
Tup
Pro
Equ
Tur
Fel
Dip
Eri
Mic
Ote
Pte
Lox
Tar
Das
Cho
Spe
Mac
Gor
Pan
Pon
Hom


rho=−0.734   pvalue=5.419e−06
shifteuth
Longevity
450
400
350
300
250
200
150
100
20
40
60
80
Sor
Myo
Ech
Och
Rat
Ory
Cav
Mus
Eri
Dip
Tup
Pro
Vic
Bos
Pte
Das
Tar
Fel
Oto
Mic
Spe
Can
Tur
Equ
Cho
Lox
Mac
Gor
Pon
Pan
Hom

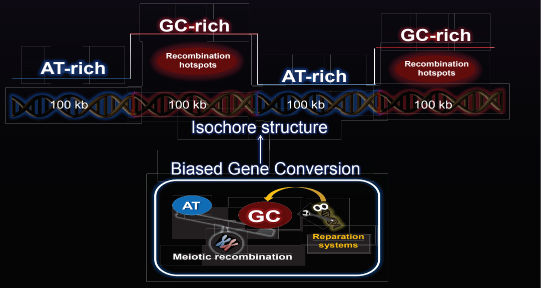
Contrasted GC-content dynamics across 33 mammalian genomes : influence of body mass and chromosome size

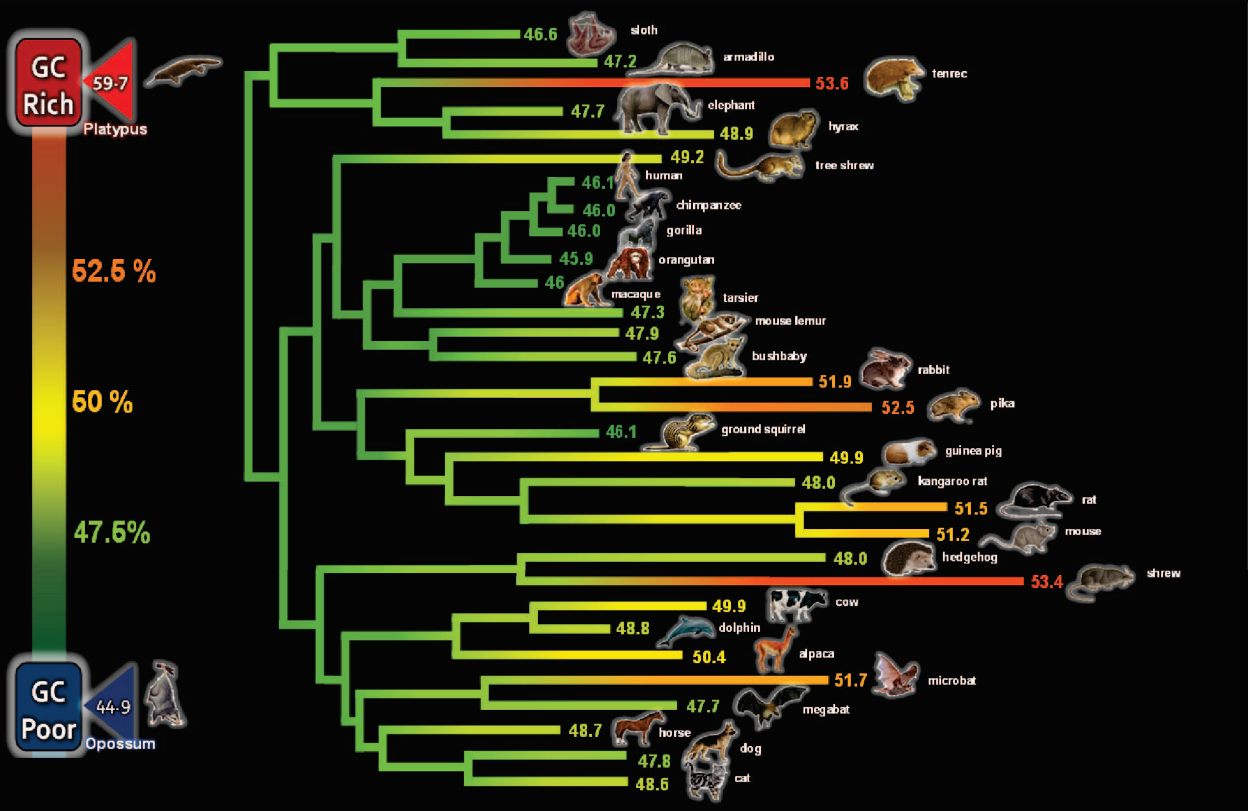
Romiguier Jonathan, Ranwez Vincent, Douzery Emmanuel, Galtier Nicolas

2010 Genome Research (accepted)

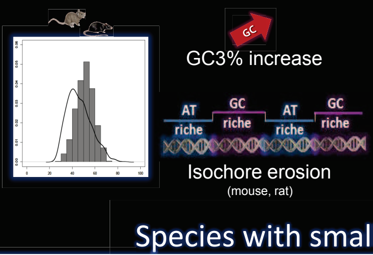
Abstract

The origin, evolution and functional relevance of isochores (genomic variations in GC-content) is a long-debated topic, especially in mammals. Such variations in GC-content could be explained by recombination hotspots, where reparation systems could promote the fixation of GC alleles : The Biased Gene Conversion Model. Most of the existing literature, however, has focused on a small number of model species and/or limited sequence datasets. We analyzed more than 1000 orthologous genes in 33 fully-sequenced mammalian genomes, reconstructed their ancestral isochore organization in the maximum likelihood framework, and explored the evolution of third-codon position GC-content in representatives of 16 orders and 27 families. We showed that the previously-reported erosion of GC-rich isochores is not a general trend. Several species (e.g. shrew, microbat, tenrec, rabbit) have independently undergone a marked increase in GC-content, with a widening gap between the GC-poorest and GC-richest classes of genes. The intensively studied apes and (especially) murids do not reflect the general placental pattern. We correlated GC-content evolution with species life-history traits and cytology. Significant effects of body mass and genome size were detected, with each being consistent with the GC-biased gene conversion model.

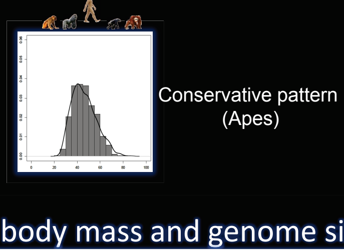




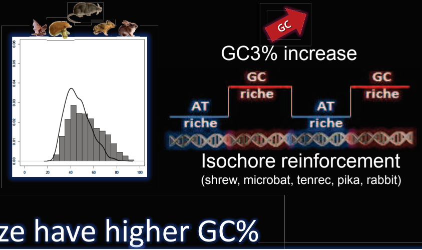
The color gradient between red and green reflects respectively hight and low estimated average GC3 on 1138 ortologous genes. Branch lengths quantify the amount of GC3 divergence.
As shown in the 3 graph below, there are 3 major pattern of GC3% evolution. Black curve reflects the estimated ancestral GC3% distribution, grey histogram reflects the observed distribution of extant species



Isochore erosion
(mouse, rat)

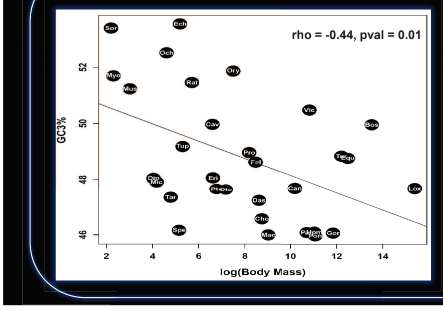


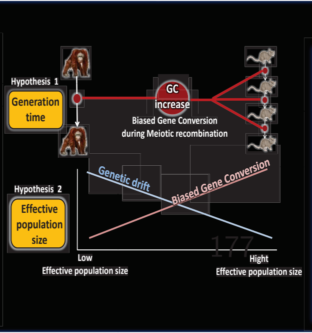
Conservative pattern
(Apes)



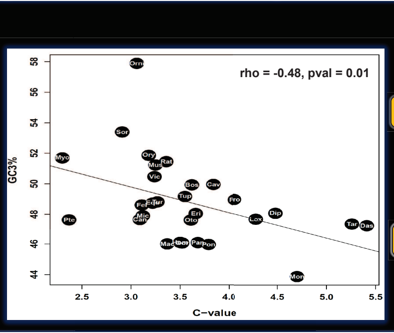
Isochore reinforcement
(shrew, microbat, tenrec, pika, rabbit)

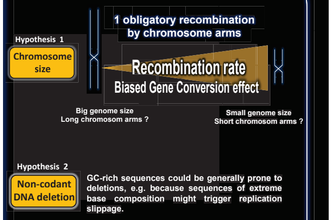
Species with small body mass and genome size have higher GC%





Hypothesis 1
Generation time
GC increase
Biased Gene Conversion during Meiotic recombination
Hypothesis 2
Effective population size
Genetic drift
Biased Gene Conversion
Low Effective population size
Hight Effective population size



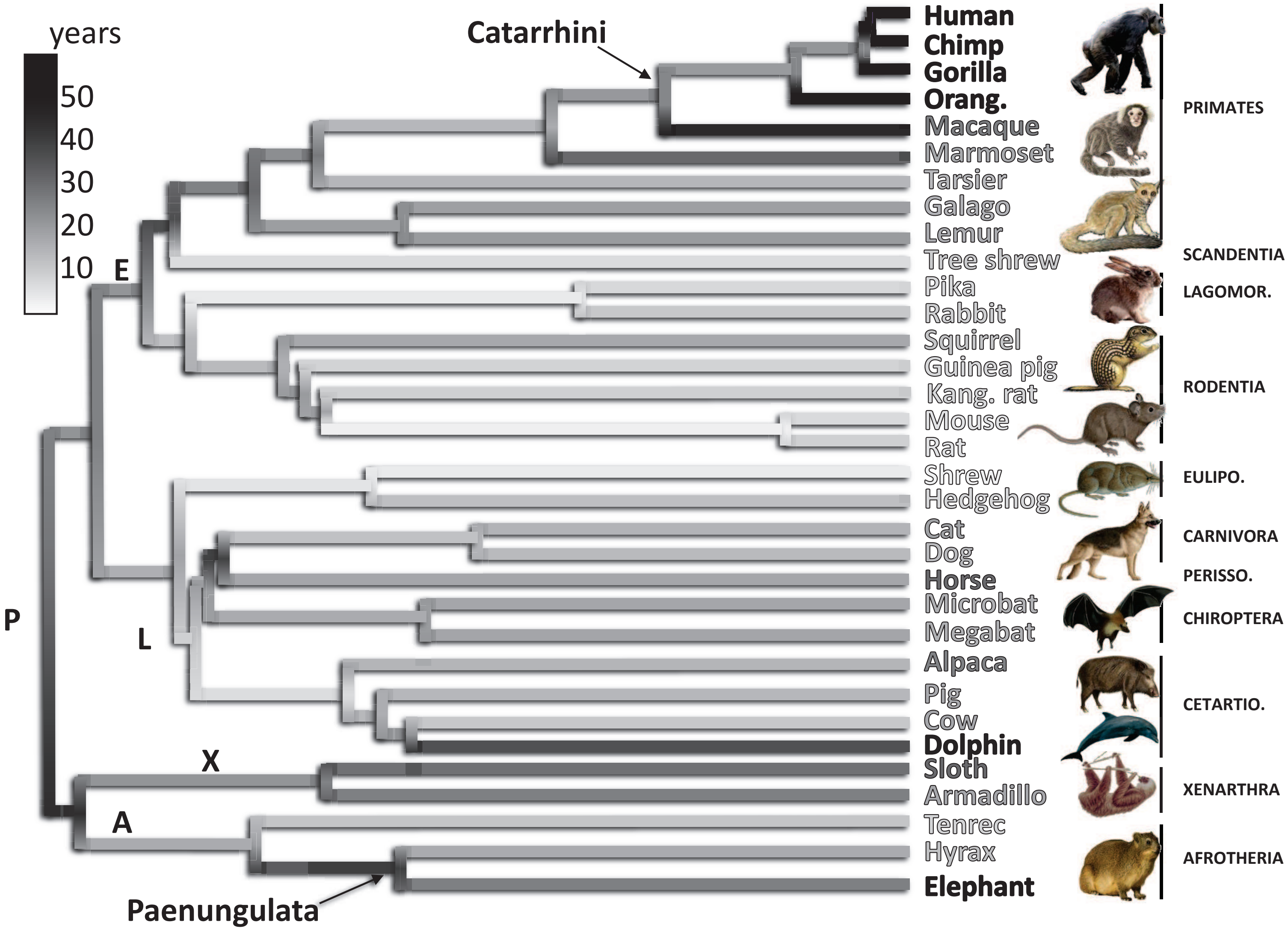


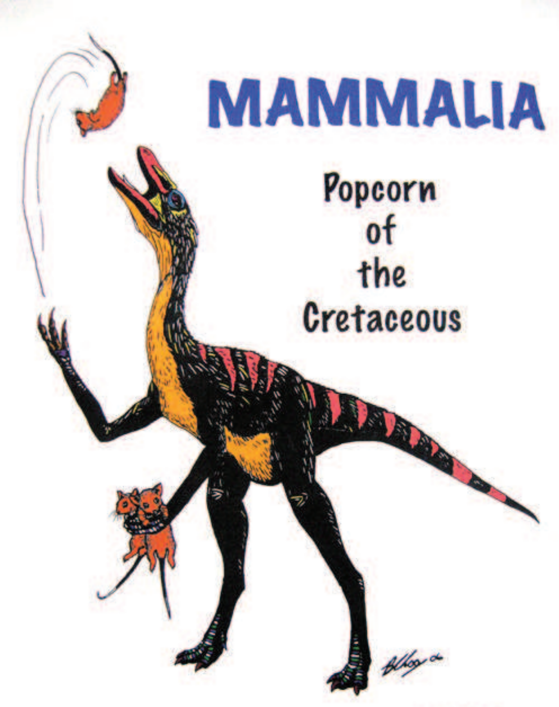
Hypothesis 1
Chromosome size
1 obligatory recombination by chromosome arms
Recombination rate
Biased Gene Conversion effect
Big genome size Long chromosom arms ?
Small genome size Short chromosom arms ?
Hypothesis 2
Non-codant DNA deletion

GC-rich sequences could be generally prone to deletions, e.g. because sequences of extreme base composition might trigger replication slippage.

6.2 Annexes Chapitre 4

- Supplemental data : Figure 1 in black and white (for color blind).
- Small cretaceous mammal "memes".
- Large audience article in "Newscientist" : http://www.newscientist.com/article/dn22343-mammals-ancestor-was-not-as-puny-as-we-thought.html


years
50
40
30
20
10
Catarrhini
Human
Chimp
Gorilla
Orang.
Macaque
Marmoset
Tarsier
Galago
Lemur
Tree shrew
Pika
Rabbit
Squirrel
Guinea pig
Kang. rat
Mouse
Rat
Shrew
Hedgehog
Cat
Dog
Horse
Microbat
Megabat
Alpaca
Pig
Cow
Dolphin
Sloth
Armadillo
Tenrec
Hyrax
Elephant
PRIMATES
SCANDENTIA
LAGOMOR.
RODENTIA
EULIPO.
CARNIVORA
PERISSO.
CHIROPTERA
CETARTIO.
XENARTHRA
AFROTHERIA
E
L
P
X
A
Paenungulata


MAMMALIA
Popcorn of the Cretaceous

“No daughter of mine's going to date a mammal. They have no future.”
WWW.CHUCKLEADUCK.COM
©2011

...TRY TO TAKE OVER THE WORLD!!!
METEOR MAGNET
65 million years ago - Late Cretaceous

SUSPECTS
- 001
13010 - 002
13010 - 003

SAVE YOURSELF MAMMAL! WE WILL FEND OFF THE ASTEROIDS!
NEVER FORGET
12/21/65000000 b.c.
memecenter.com MemeCenter

Home | Life | News | Back to article

Mammals' ancestor was not as puny as we thought

11:00 07 October 2012 by **Michael Marshall**
For similar stories, visit the **Evolution** Topic Guide

The long and short of it is that things remain contentious *(Image: Louie Psihoyos/Science Faction/SuperStock)*

The common ancestor of modern mammals was tiny and shrewlike, living unobtrusively in the shadow of the dinosaurs – or so we thought. A genetic analysis now suggests it may have been more like a small monkey in size.

Fossils indicate that some larger mammals shared the dinosaurs' world, but palaeontologists think that they all disappeared alongside the giant reptiles. Only tiny mammals survived, giving rise to all modern forms.

Nicolas Galtier of the Institute of Evolutionary Sciences in Montpellier, France, begs to differ. With colleagues, he used common features in the genomes of 36 modern mammals to sketch out the genome of the creature from which they descended.

Reconstructing the detailed genome is impossible, but Galtier managed to recover two of its properties. In modern mammals, these properties are correlated with body size and lifespan. Galtier's results suggest the ancestor of modern mammals weighed at least a kilogram, and lived over 25 years'.

Michael Novacek of the American Museum of Natural History in New York is sceptical, given what the fossil record tells us. Several modern mammal groups such as rodents emerged after the dinosaur extinction, and the fossils show their first members were small. "There's no question about that," says Novacek.

But Galtier points out that the fossil record is incomplete. A large mammal ancestor that ultimately gave rise to all modern mammal groups, including the rodents, might simply have failed to fossilise.

Journal reference: *Molecular Biology and Evolution*, doi.org/jfk

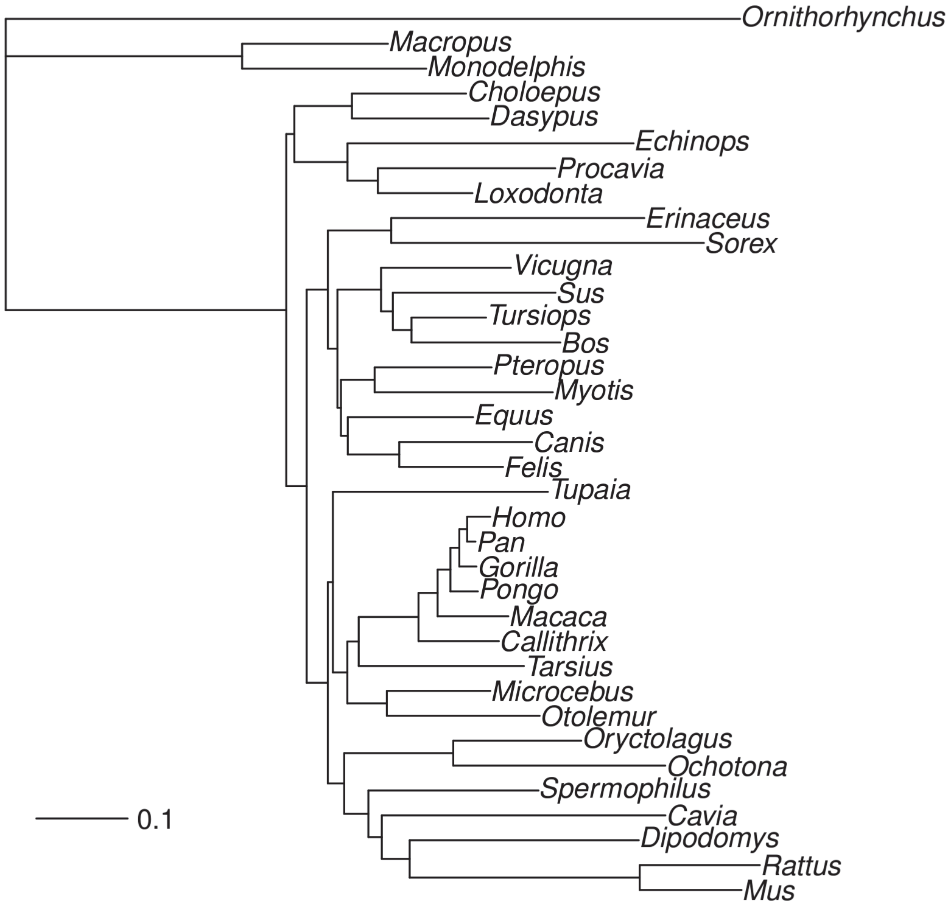
J'aime 77 PRINT SEND SHARE

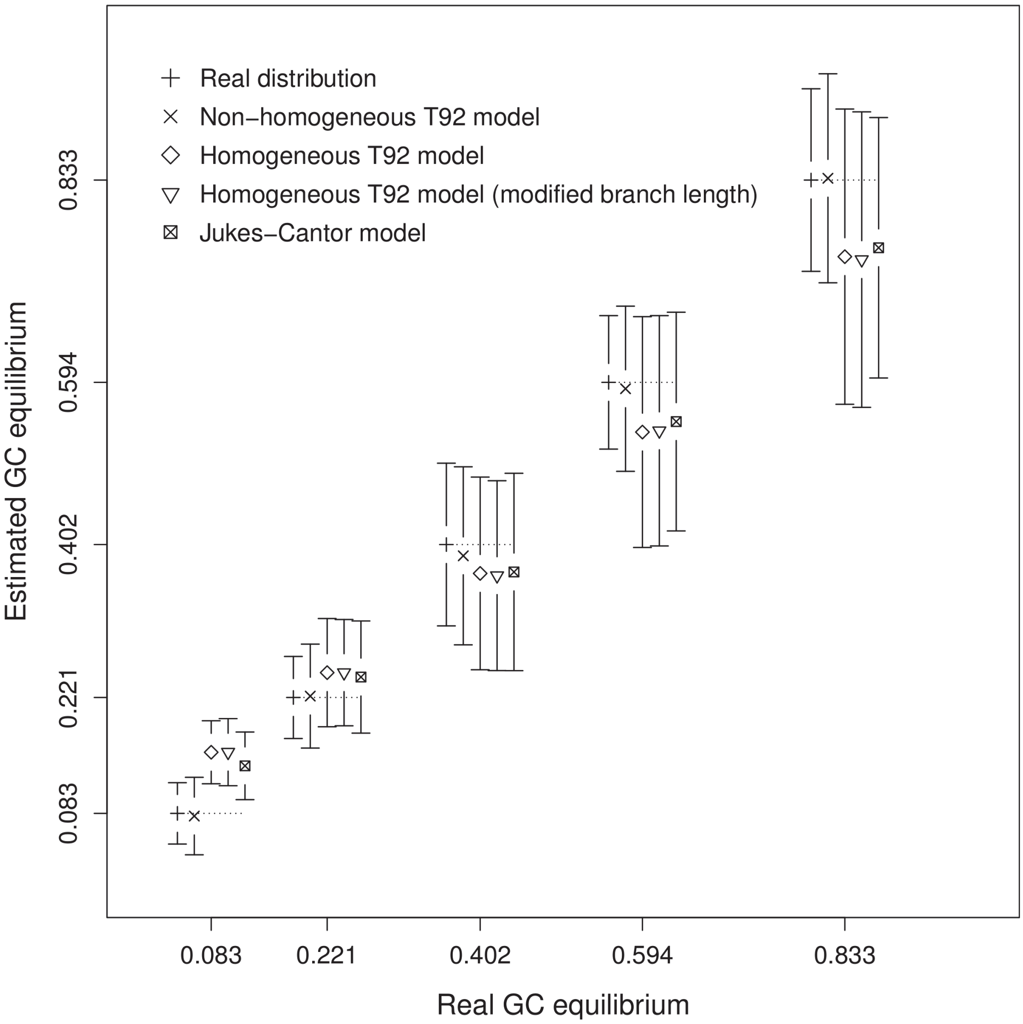
Back to article

PRINT SEND SHARE

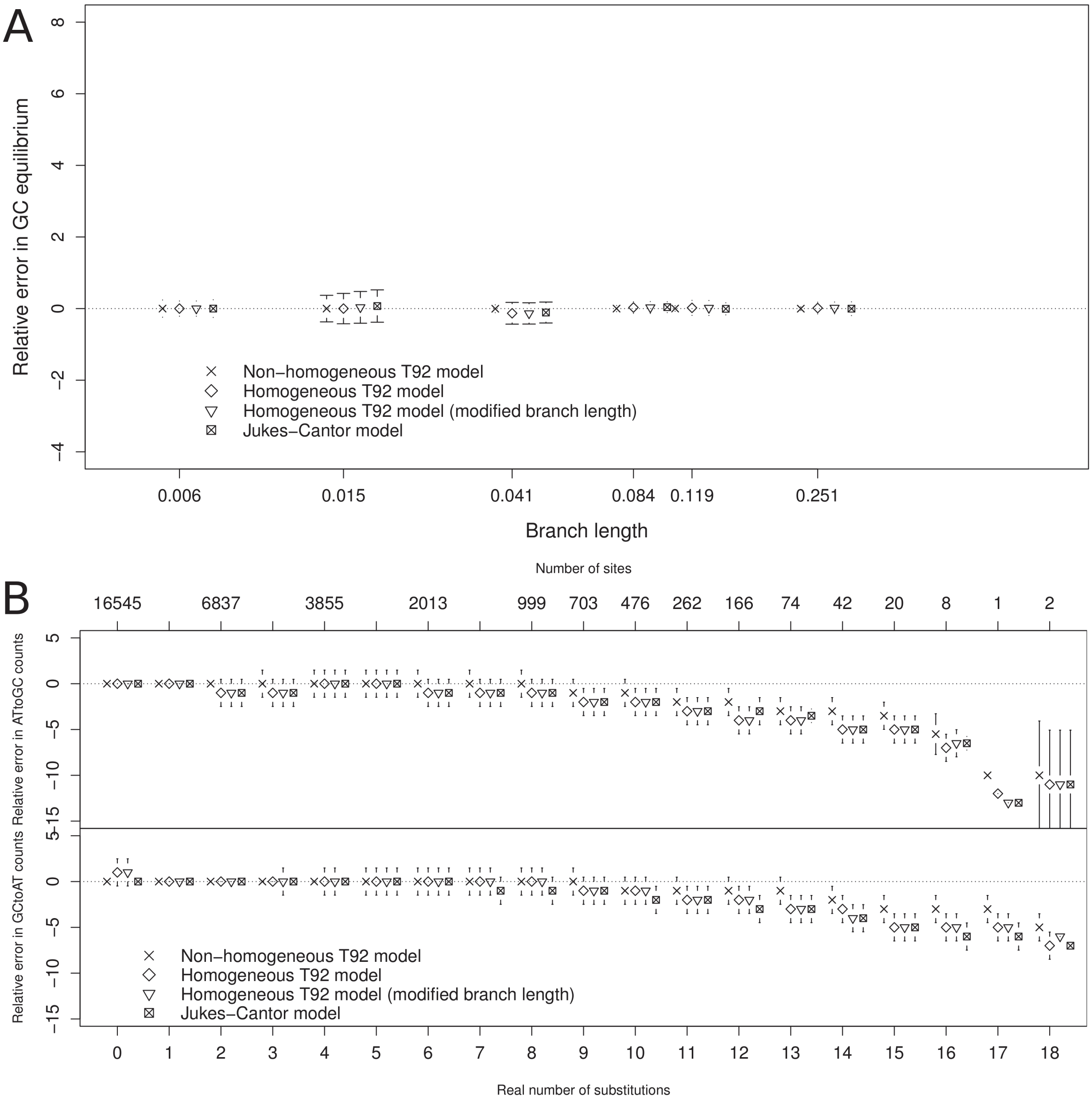
6.3 Annexes Chapitre 6

- Figure S1 : Tree used in simulation.
- Figure S2 : Estimated GC-equilibrium apprimation under various substitution mapping conditions compared with real GC equilibrium approximation from simulations.
- Figure S3 : Relative error in GC-equilibrium estimation by branch and by sites.


Ornithorhynchus
Macropus
Monodelphis
Choloepus
Dasypus
Echinops
Procavia
Loxodonta
Erinaceus
Sorex
Vicugna
Sus
Tursiops
Bos
Pteropus
Myotis
Equus
Canis
Felis
Tupaia
Homo
Pan
Gorilla
Pongo
Macaca
Callithrix
Tarsius
Microcebus
Otolemur
Oryctolagus
Ochotona
Spermophilus
Cavia
Dipodomys
Rattus
Mus
0.1



+ Real distribution
× Non−homogeneous T92 model
◇ Homogeneous T92 model
▽ Homogeneous T92 model (modified branch length)
⊠ Jukes−Cantor model
Estimated GC equilibrium
0.083
0.221
0.402
0.594
0.833
Real GC equilibrium
0.083
0.221
0.402
0.594
0.833


A
Relative error in GC equilibrium
8
6
4
2
0
−2
−4
Non−homogeneous T92 model
Homogeneous T92 model
Homogeneous T92 model (modified branch length)
Jukes−Cantor model
0.006
0.015
0.041
0.084
0.119
0.251
Branch length
Number of sites
B
16545
6837
3855
2013
999
703
476
262
166
74
42
20
8
1
2
Relative error in ATtoGC counts
Relative error in GCtoAT counts
5
0
−5
−10
−15
Non−homogeneous T92 model
Homogeneous T92 model
Homogeneous T92 model (modified branch length)
Jukes−Cantor model
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
Real number of substitutions