Estimation et analyse du taux de substitution adaptatif chez les animaux

Marjolaine Rousselle

THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE L'UNIVERSITÉ DE MONTPELLIER

En génétique et génomique

École doctorale GAIA

Unité de recherche ISEM

Estimation et analyse du taux de substitution adaptatif chez les animaux

Présentée par Marjolaine ROUSSELLE

Le 26 novembre 2018

Sous la direction de Nicolas GALTIER et Benoit NABHOLZ

Devant le jury composé de

Gwenaël PIGANEAU, Directeur de recherche, CNRS	Rapporteur
Adam EYRE-WALKER, Professor, University of Sussex	Rapporteur
Carole SMADJA, Chargée de recherche, CNRS	Examinatrice
Laurent DURET, Directeur de recherche, CNRS	Président du jury
Nicolas GALTIER, Directeur de recherche, CNRS	Directeur
Benoit NABHOLZ, Maître de conférences, Université de Montpellier	Co-directeur

UNIVERSITÉ DE MONTPELLIER

Estimation et analyse du taux de substitution adaptatif chez les animaux

Résumé : Comprendre les déterminants du taux d'adaptation est une question primordiale en évolution moléculaire. En particulier, l'influence de la taille efficace de population sur la sélection positive, ainsi que la nature des changements d'acides aminés qui mènent à de l'adaptation sont des questions encore débattues. Pour y répondre, la méthode DFE-α, dérivée du test fondateur de McDonald & Kreitman, est un outil puissant pour mesurer le taux de substitution adaptatif. Elle est néanmoins sensible à certains biais. Au cours de cette thèse, nous avons identifié deux biais majeurs de cette méthode, les fluctuations de long-terme du régime de sélection-dérive via des fluctuations démographiques, et la conversion génique biaisée vers GC (gBGC). Via des simulations, nous avons montré que divers scénarios plausibles de fluctuations démographiques peuvent mener à une sur-estimation du taux de substitution adaptatif. Nous avons aussi obtenu des indications empiriques que le régime de sélection-dérive récent ne reflète pas le régime de sélection-dérive de long-terme chez diverses espèces animales, ce qui représente une violation d'une hypothèse forte de la méthode DFE-α. D'autre part, nous avons montré que la gBGC entraîne une sur-estimation du taux de substitution adaptatif chez les primates et les oiseaux. Via un jeu de données de neuf taxons de métazoaires et un total de 40 espèces, nous avons d'une part initié une analyse visant à identifier la nature des changements d'acides aminés qui mènent à l'adaptation, et montré que les changements radicaux sont soumis à une plus forte sélection purificatrice que les changements conservatifs. D'autre part, nous avons pu évaluer le lien entre la taille efficace et le taux de substitution adaptatif tout en prenant en compte les deux sources de biais explorées précédemment. Nous avons mis en évidence pour la première fois une relation négative entre le taux de substitution adaptatif et des traits d'histoire de vie représentatifs de la taille de population de long-terme. Ce résultat va à l'encontre de l'hypothèse canonique d'une adaptation plus efficace en grandes populations.

Mots clefs : Adaptation moléculaire – test de McDonald & Kreitman – animaux – effets en fitness des mutations – conversion génique biaisée vers GC – fluctuations démographiques

Estimation and analysis of the adaptive substitution rate in animals

Abstract: Understanding the determinants of the adaptive substitution rate is a central question in molecular evolution. In particular, the influence of the effective population size N_e on positive selection as well as the nature of amino acid changes that lead to adaptation are still debated. The DFE-α method, which was derived from the seminal McDonald & Kreitman test, is a powerful tool for estimating the adaptive substitution rate. However, it is sensitive to various sources of bias. In this thesis, we identified two major sources of bias of this test, long-term fluctuations of the selective-drift regime through demographic fluctuations, and GC-biased gene conversion (gBGC). Using simulations, we showed that under plausible scenarios of fluctuating demography, the DFE-α method can lead to a severe over-estimation of the adaptive substitution rate. We also showed that polymorphism data reflect a transient selective-drift regime which is unlikely to correspond to the average regime experienced by genes and genomes during the long-term divergence between species. This violates an important assumption of the DFE-α method. Our results also indicate that gBGC leads to an over-estimation of the adaptive substitution rate in primates and birds. Using a dataset of nine metazoan taxa for a total of 40 species, we started an analysis aiming at identifying the type of amino acid changes that are more prone to adaptation, and evaluated the link between N_e and the adaptive substitution rate while accounting for the two sources of bias previously explored. We reveal for the first time a negative relationship between the adaptive substitution rate and life-history traits representative of long-term N_e. This result is in contradiction with the widespread hypothesis that adaptation is more efficient in large populations.

Key words : Molecular adaptation – McDonald & Kreitman test – animals – fitness effects of mutations – GC-biased gene conversion – demographic fluctuations

Remerciements

Je vais commencer par remercier les deux personnes qui ont été les plus importantes pour la bonne réalisation de ce travail et le bon déroulement de ces trois dernières années, Nicolas et Benoit. Au cours de ces trois ans, j'ai eu maintes occasions de partager avec d'autres thésards les difficultés de la thèse, et ainsi de réaliser à quel point mes directeurs m'ont facilité la tâche par rapport à l'ensemble de mes collègues en thèse. Malgré leur emploi du temps chargé à tous les deux, Nicolas étant directeur adjoint et futur directeur de l'ISEM et Benoit étant Maître de conférence, ils ont toujours pris le temps de répondre à mes questions, de m'aider à interpréter mes résultats parfois insolites et à comprendre et manipuler le C++, et de réviser rapidement et de manière toujours constructive mes manuscrits. Mes interactions régulières avec Nicolas ont le don de me redonner confiance en ce que je fais et renouveler mon enthousiasme pour ma thèse. Ils m'ont de plus encouragé à assister et/ou participer à de nombreuses conférences, me finançant même un voyage jusqu'au Japon cette année.

Au sein de l'équipe, trois autres membres ont été, ou sont encore, des piliers de ce travail de thèse : Émeric Figuet, Paul Simion et Marie-Ka Tilak. Émeric et Paul ont tout deux contribué à préparer des données de vertébrés en suivant un protocole très fastidieux de génotypage, et m'ont apporté un soutien scientifique en plus de leur amitié. Marie-Ka a été mon guide tout au long de ma préparation du jeu de données « capture », me prenant sous son aile alors que je n'avais pratiquement jamais manipulé de pipette de ma vie. Elle s'est montré extrêmement pédagogue, et tolérante lorsque, en fin de semaine, j'anéantissais toute une journée de travail en confondant deux lignes d'une plaque de 96 puits. Pour cela, je les remercie fortement tous les trois.

J'aimerais également remercier mes deux stagiaires, Alexandre Laverré et Hilde Schneemann, qui ont tous les deux effectué un travail formidable et gardé le moral même devant les difficultés techniques que nous avons rencontré au cours de leurs projets, ainsi que Maeva Mollion (et son encadrant Thomas Bataillon), pour sa collaboration à mon projet de simulations et pour m'avoir fait découvrir « Exploding kittens », jeu à destination d'amoureux de chats psychopathes.

Je voudrais aussi remercier Nicolas Bierne, Roger de Villa, Nicolas Faivre, Marion Ballenghien et Lise Dupont pour m'avoir fourni des échantillons soit de tissu, soit d'ADN. Je remercie aussi les membres de mes comités de thèse, Laurent Duret et Vincent Ranwez.

Je remercie aussi les autres membres de l'équipe PhylEvolMol, avec une mention particulière à Sylvain Glémin qui m'a assisté dans mes tentatives de coder mon tout premier programme en C++.

Au sein de l'équipe, mes co-bureau méritent une mention spéciale : Grand Yoann, et par la suite Thibault, ont grandement contribué au maintien de ma santé mentale en m'aidant régulièrement à

gérer mes problèmes de bio-informatiques, ou de vocabulaire, ou encore de démarches administratives. Andrea, et surtout, surtout Clémentine, m'ont quant à elles été d'une aide vitale pour faire baisser mon niveau de stress, déjà élevé en situation normale, qui a subit quelques pics réguliers à certains (nombreux) moments clés de ma thèse. Quand Clémentine dit « Ne t'inquiètes pas, tous les thésards passent par là, c'est normal », on la croit. Ce soutien mental a aussi été prodigué par les autres thésards de l'ISEM, que je remercie : Maud, Maeva, Sergio, Alexis, Alice, Mine, Rémi, Cécile, Maxime, Quentin, Paul, Alain, Laura, Alex, et Manon. Manon mérite des remerciements pour elle toute seule : ayant commencé en même temps, nous avons vécu et partagé tous les moments de la thèse. Je me suis rendue compte en observant des doctorants plus « isolés » que moi à quel point il est vital d'avoir la possibilité de partager son expérience avec d'autres thésards, et nos discussion quasi quotidiennes avec Manon ont été d'une énorme aide tout au long de ma thèse. Elle a contribué à ma participation active à la vie de l'ISEM, via l'organisation du WE d'intégration des doctorants ou la JDD. Enfin, sur un autre registre, Manon m'a fait découvrir un nombre considérable de choses, du Block Out à DnD, m'a appris à rester réveillée après 4h du matin grâce à la danse hawaïenne, et a même presque réussi à me faire aimer la trans.

Je voudrais aussi remercier Pierrick, Sandrine, Benoit, et Mélanie (qui a aussi traqué les fautes d'orthographes de mon introduction!) qui ont été mes guides tout au long de mon monitorat et m'ont transmis le goût de l'enseignement.

Enfin, sur le plan personnel, sans lequel il n'y aurait pas de plan professionnel, je remercie ma famille, qui m'a permis de relativiser et de re-découvrir régulièrement que les difficultés dans la thèse n'avaient rien de dramatique, et qu'il y a plus important dans la vie qu'un programme C++ qui retourne un « segmentation fault ».

Je remercie aussi mes colocs, Johanne et Claire, puis Alain, Marianne, Valentin, Charly et Iago (que je remercie spécialement pour l'aide qu'il m'a apporté en ce qui concerne la programmation), ainsi que tous mes amis, Marianne, Estelle, Morgane (x2), Tiphanie, Claire,Vincent, Théo, Félise, Thibault, Coco, Julie, Olympe, et les filles de l'ISC, pour tous les moments qu'ils m'ont apporté et qui m'ont permis de profiter à fond de ces trois années malgré le travail intensif.

Enfin, cela vient en dernier mais c'est peut-être la personne qui mérite le plus ma gratitude, je remercie Yo. En plus de m'avoir supportée dans mes moments d'indignation sur des thèmes divers et variés, rassurée dans mes moments de stress sur mon avenir professionnel, relu la quasi-totalité de mon manuscrit et écouté patiemment mes réflexions embrouillées sur mes résultats compliqués, il m'a surtout donné confiance en moi, dans tous les aspects de ma vie. En partageant tous les moments d'évasion et de déconnexion de la thèse, qui sont salutaires pour rester motivé, il a fait de ces trois ans une période de bonheur et d'épanouissement.

Et en plus maintenant, je sais skier.

Liste des publications

Rousselle M, Faivre N, Ballenghien M, Galtier N, Nabholz B. 2016. Hemizygosity enhances purifying selection: lack of fast-Z evolution in two Satyrine butterflies. Genome Biology and Evolution 8:3108–3119.

Galtier N, Roux C, Rousselle M, Romiguier J, Figuet E, Glemin S, Bierne N, Duret L. 2017. Codon usage bias in animals: disentangling the effects of natural selection, effective population size and GC-biased gene conversion. Molecular Biology and Evolution 35:1092-1103.

Rousselle M, Mollion M, Nabholz B, Bataillon T, Galtier N. 2018. Overestimation of the adaptive substitution rate in fluctuating populations. Biology Letters 14:20180055.

Rousselle M, Laverré A, Figuet E, Nabholz B, Galtier N. 2018. Influence of recombination and GC-biased gene conversion on the adaptive and non-adaptive substitution rate in mammals *vs*. birds. Molecular Biology and Evolution (major revisions).

Pinharanda A , Rousselle M, Martin SH, Hanly JJ , Davey JW, Kumar S, Galtier N & Jiggins CD. 2018. Sexually dimorphic expression and transcriptome evolution provides mixed evidence for a fast-Z effect in *Heliconius.* Journal of Evolutionary Biology (under revision).

Table des matières

INTRODUCTION

I. Adaptation à l'échelle moléculaire : concepts et théories

1. L'évolution des molécules pour répondre aux questions évolutives

L'évolution moléculaire est une discipline récente relativement à d'autres disciplines de la biologie évolutive. Elle doit son développement à l'émergence de techniques telles que l'électrophorèse en 1960 pour les protéines, le séquençage Sanger en 1977 pour l'ADN, ou l'analyse en série de l'expression des gènes (SAGE) en 1995 pour l'ARN. Née de la rencontre entre la biologie moléculaire et la biologie évolutive, elle est par construction une discipline duelle : à mon sens, son objectif est à la fois de comprendre les mécanismes moléculaires à l'origine de l'évolution des génomes, transcriptomes et protéomes, mais aussi d'utiliser l'information contenue dans ces molécules pour retracer l'histoire évolutive des individus, des populations ou des espèces.

Pour ce dernier point, l'apport de l'information contenue dans les molécules a été une avancée majeure. En effet, les caractères morphologiques étudiés jusqu'alors, s'ils ont permis le développement des théories initiales de l'évolution et plus tard de mettre en place les premières phylogénies, présentent néanmoins certains écueils, tels que la présence d'homoplasies du fait des convergences évolutives liées à des modes de vie communs (Murphy et al. 2004). De plus, ces caractères sont en nombre limité. Par exemple, les premières phylogénies des mammifères actuels, ainsi que les phylogénies de mammifères fossiles, ont été établies grâce aux ossements fossiles et à des caractères anatomiques. Mais la comparaison de millions de paires de bases de séquences génomiques, rendues accessibles grâce au développement des Nouvelles Technologies de Séquençage (Next Generation Sequencing, NGS) a permis d'améliorer considérablement le pouvoir de résolution phylogénétique et la compréhension des mécanismes évolutifs (Murphy et al. 2001). En effet, l'étude des molécules, que ce soit ADN, ARN ou protéines, a permis de confirmer ou d'infirmer les théories existantes, mais aussi l'émergence de nouvelles théories soutenues par des arguments plus tangibles, telles que celles développées dans la partie suivante. En particulier, les données de séquence se sont avérées peu conformes au paradigme Darwinien considérant l'adaptation comme moteur de l'évolution, comme nous le verrons par la suite. Pour cette raison, les théories qui ont émergé à cette époque s'articulent en grande partie autour d'une question centrale en évolution, et qui est également l'objet de cette thèse : celle de la place de la sélection naturelle *vs.* la place des processus neutres dans l'évolution des génomes, et par extension, des espèces. La difficulté de la résolution de cette question réside en grande partie de la dualité de la discipline évoquée précédemment. D'un côté, nous utilisons les molécules d'ADN, ARN ou d'acides aminés pour découvrir et mesurer les forces évolutives agissant sur les individus. Mais d'un autre côté nous

devons avant – ou parfois en parallèle - comprendre les mécanismes moléculaires à l'œuvre au sein des génomes afin d'identifier les sources de biais ou effets confondants potentiels.

2. Adaptationnisme ou neutralisme ?

a. Émergence des théories de génétique des populations

Pour que les premiers modèles de génétique des populations émergent, il aura fallu la réconciliation entre les découvertes de la fin du XIXe siècle en génétique et le paradigme darwinien de la sélection naturelle. Au début du XXe siècle, on voit émerger un consensus interdisciplinaire, résumé dans ce qu'on appelle la théorie synthétique de l'évolution menée entre autres par R.A. Fisher, J.B.S Haldane, Sewall Wright, Theodosius Dobzhansky, Ernst Mayr, G.G. Simpson et Julian Huxley. Cette synthèse permet le développement de principes et modèles en génétique des populations où la sélection naturelle est le principal mécanisme à l'œuvre. En s'appuyant sur ce principe, John B.S. Haldane fait des prédictions sur les patrons d'évolution moléculaire en calculant le taux de substitution attendu chez les mammifères. En 1957, il estime que la limite supérieure de ce taux de substitution est de une sur 300 générations, au vue de la taille moyenne d'un génome de mammifère. Pour parvenir à ce résultat, il se base sur le « coût de la sélection », qui signifie que pour qu'un nouvel allèle adaptatif se fixe dans la population il faut que les individus non-porteurs de cet allèle disparaissent, ce qui constitue effectivement un coût pour la population, et mène à une limitation du nombre de substitutions possibles au sein d'une espèce en un nombre de générations données (Haldane 1957). Il ne considère ainsi que des substitutions avantageuses, et néglige l'effet de la dérive et les mutations neutres et faiblement délétères.

Quelques années plus tard, les biochimistes Linus Pauling et Emile Zuckerkandl énoncent une hypothèse qui diffère beaucoup de celle de Haldane, l'hypothèse de l'horloge moléculaire, selon laquelle les gènes évoluent selon un taux constant au cours du temps (Zuckerkandl and Pauling 1965). Cette découverte est d'une grande importance pour le développement de la théorie neutre plus tard énoncée par Kimura (voir plus loin). En effet, ce postulat est peu compatible avec l'hypothèse de la sélection naturelle comme principal, voire seul, moteur de l'évolution puisque les événements d'adaptation sont dépendants à la fois des changements environnementaux - qui provoquent des variations des contraintes sélectives - et de l'apparition des mutations avantageuses - qui relève d'un processus aléatoire.

Cette théorie s'appuie sur les premières comparaisons de séquences protéiques entre espèces proches, par exemple celle de l'hémoglobine chez un certain nombre de vertébrés. Ces comparaisons ont permis d'observer des différences entre espèces, c'est-à-dire des substitutions, et ainsi de pouvoir les compter. Associées aux données paléontologiques, ces nouvelles données ont permis d'obtenir des taux de substitution empiriques, qui se sont révélés être relativement constants, appuyant donc la théorie de l'horloge moléculaire (Salser et al. 1976). À la lumière de ces nouvelles données de séquences, d'autres arguments peu compatibles avec les attendus de la théorie néo-darwiniste se sont accumulés. Notamment, l'estimation du taux de substitution chez les mammifères s'est avéré plusieurs centaines de fois plus élevé que l'estimation de Haldane. De plus, les premières comparaisons de séquences d'ARN messagers ont montré que les substitutions en troisième position des codons, dont les ~2/3 sont synonymes, et donc supposément sélectivement neutres ou à effets faibles, sont prépondérantes par rapport aux changements d'acide aminés (Kimura 1977).

Ces premières comparaisons de séquences protéiques entre espèces se sont également accompagnées du développement de la technique d'électrophorèse sur les enzymes, ce qui a rendu possible l'évaluation de la variabilité de ces dernières entre les individus de la même espèce, et révélé la prévalence du polymorphisme au sein des gènes (Harris 1966, Lewontin and Hubby 1966). Encore une fois, ce résultat est difficilement compatible avec l'hypothèse de la sélection naturelle comme seule moteur de l'évolution, puisqu'on s'attendrait dans ce cas à ce que le niveau de polymorphisme reste faible car les mutants favorables sont rapidement fixés, et les mutants défavorables éliminés.

Ces derniers résultats sont à l'origine du développement d'une théorie radicalement différente de la théorie synthétique de l'évolution, la théorie neutraliste de l'évolution proposée par Kimura dans les années soixante (Kimura 1968), qui alimentera les débats dans les décennies suivantes. Cette théorie stipule que la plupart des variations observables au niveau moléculaire sont « neutres » du point de vue de la sélection, c'est-à-dire qu'elles ne confèrent ni un avantage ni un désavantage aux individus qui les portent. Plus important encore, elle introduit le fait que c'est en priorité la dérive génétique, et non la sélection naturelle, qui détermine les changements de fréquences des allèles au cours des générations (Kimura 1983). En parallèle, King et Jukes développent également des idées qui rejoignent celles de Kimura, mais basées sur des constatations plus biochimiques (King and Jukes 1969). Dans un article scientifique de 1969 appelé "Non-Darwinian Evolution", ils mettent en évidence une relation négative entre l'importance fonctionnelle d'une protéine ou d'un site au sein d'une protéine, et son taux d'évolution, et

introduisent la notion de « contrainte sélective ». Ainsi, ils montrent que les mutations qui arrivent à fixation sont principalement des mutations qui affectent peu la séquence protéique ou la survie de l'individu, alors que les autres mutations sont sous contrainte forte et rapidement éliminées par la sélection purificatrice.

Ces articles ont reçu beaucoup de critiques à leur parution, mais cela semble avoir plutôt contribué à regrouper et faire connaître largement les arguments neutralistes, puisque Kimura a par la suite ajouté à son approche certains arguments de King et Jukes (1969) et défendu ce point de vue au cours des décennies qui suivirent. Malgré ce que l'on peut penser au premier abord, cette théorie ne nie pas l'existence de la sélection naturelle, mais ne la considère pas comme le seul ou principal moteur de l'évolution moléculaire, pour y préférer une combinaison entre dérive génétique et sélection positive et purificatrice. Kimura précise d'ailleurs que le mot « neutre » n'est pas à prendre au pied de la lettre, que les mutations ne sont peut-être pas neutres en soi, mais se comportent comme telles du fait que leur destin est contrôlé par la dérive génétique et non la sélection quand elles ne sont pas d'effet suffisamment fort (Kimura and Ohta 1971, Kimura 1983).

De cette théorie réunie entre Kimura, Jukes et King vient le postulat que les mutations adaptatives sont très rares dans les génomes. Premièrement parce qu'une grande proportion des génomes, notamment chez les eucaryotes, est non fonctionnelle et n'est pas impliquée dans la construction des phénotypes (Lynch and Walsh 2007). Et deuxièmement, parce qu'une mutation qui change une séquence protéique a plus de chance de détériorer la protéine que de l'améliorer, les séquences codantes étant de manière générale assez optimisées et sous fortes contraintes. La sélection naturelle est donc peu à peu passée d'une force perçue comme « motrice », résumée uniquement à la sélection positive qui pousse les mutations avantageuses à fixation, à une force perçue comme un frein à l'accumulation des mutations désavantageuses (Lynch and Walsh 2007).

Les avancées en termes de développements mathématiques, ainsi que l'accroissement des données de séquences disponibles et l'accumulation des éléments de compréhension du fonctionnement des génomes, ont mis en évidence certaines failles dans le raisonnement qui a conduit aux idées neutralistes. Kimura et Otha introduisent donc par la suite la notion de mutations « presque neutres » (Kimura and Ohta 1971, Ohta 1973). Ce terme désigne les mutations dont le coefficient de sélection est non-nul, mais inférieur ou proche du ratio $1/N_e$, où N_e est la taille efficace de la population dans laquelle la mutation ségrège. La taille efficace désigne l'effectif d'une population « idéale » de Wright-Fisher (c'est-à-dire d'effectif constant, non structurée, à générations non-chevauchantes et sans sélection) qui présenterait les mêmes caractéristiques

génétiques que la population observée, notamment en termes d'intensité de la dérive. En d'autres termes, il s'agit du nombre d'individus qui participent à la production de la génération suivante. Si le coefficient de sélection d'une mutation, noté s, est petit face au coefficient de la dérive, $1/(2N_e)$ (pour un organisme diploïde), alors le destin de la mutation sera dicté par les fluctuations aléatoires de la reproduction, c'est-à-dire la dérive, et non pas par la sélection. Cette nouvelle conceptualisation est moins stricte que la théorie neutraliste initiale, et explique l'existence d'une relation négative entre le taux d'évolution des protéines et la taille efficace, un des résultats majeurs en évolution moléculaire (Nikolaev et al. 2007; Popadin et al. 2007; Lartillot and Delsuc 2012; Romiguier et al. 2014; Figuet et al. 2016). En effet, ces mutations « presque neutres » sont en grande partie des mutations qualifiées de « faiblement délétères » (Ohta 1973). Si en grande taille de population ces mutations sont efficacement éliminées par la sélection purificatrice, elles peuvent se maintenir en ségrégation voire arriver à fixation en petite taille de population du fait de la dérive qui diminue l'efficacité de la sélection, ce qui accroît le taux d'évolution protéique. Selon la théorie neutre, en revanche, aucune relation n'est attendue entre taille efficace et taux d'évolution, car le nombre de mutations qui apparaissent dans une population est égal au produit du taux de mutation μ par le nombre de copie d'allèles qui vont passer à la génération suivante, c'est-à-dire deux fois la taille efficace N_e pour un diploïde. Or, la probabilité de fixation d'une mutation neutre qui apparaît est égale à $1/(2N_e)$ pour un diploïde. Ainsi, le taux de fixation des mutations neutres est égal à : $t\mu 2Ne\frac{1}{2Ne}$, c'est-à-dire μt. Il est donc effectivement indépendant de la taille efficace (Kimura 1968). Enfin, si la majorité des mutations sont sélectionnées positivement, la relation entre taille efficace et taux d'évolution protéique est positive. Ainsi, cette nouvelle notion de mutation « presque neutre » est d'une importance capitale, puisque leur prise en compte permet d'expliquer des résultats empiriques, et nous verrons par la suite que les négliger au contraire peut être une importante source de biais dans certains tests en évolution moléculaire.

Pour résumer, ces théories successives mettent en évidence le formidable apport des données moléculaires à l'étude des mécanismes à l'origine de l'évolution, et la compréhension du fonctionnement des génomes. Elles illustrent également l'objectif central de la discipline qu'est l'évolution moléculaire, qui est d'identifier les patrons de variations génomiques et comprendre leur origine en termes de forces évolutives, ainsi que leur statut fonctionnel.

b. Une question toujours débattue ?

Dans un contexte qualifié de « néo-Darwinisme », où l'évolution est centrée sur le principe de sélection naturelle, la théorie neutre de l'évolution a provoqué beaucoup de débats et d'échanges parfois vifs entre scientifiques (Gillespie 1991). Mais est-ce un débat légitime, surtout depuis l'atténuation de la théorie neutre par la théorie quasi-neutre ? Kimura lui même, ainsi que beaucoup d'évolutionnistes actuels s'accordent à dire que les deux théories sont compatibles. Seulement, la sélection naturelle perd son caractère de force évolutive prépondérante et devient une force parmi d'autres au nombre desquels on compte en particulier les facteurs stochastiques telle que la dérive génétique. Le rôle de la sélection naturelle n'est néanmoins pas remis en question par la théorie neutraliste. Ainsi, beaucoup de chercheurs ne voient plus le débat « adaptationnistes *vs.* neutralistes » comme une dichotomie, mais comme un continuum entre stricte neutralité et sélection déterministe en passant par les mutations à effets sélectifs faibles ($s \sim 1/N_e$).

Pourtant, il persiste dans le domaine une forte dualité dans la façon de voir les choses. Certains « héritiers » de la pensée de Kimura, sont très méfiants de ce qu'on appelle communément en anglais « the adaptive story telling », c'est-à-dire des cas où les résultats sont expliqués par un scénario adaptatif alors que les hypothèses neutres alternatives n'ont pas été correctement rejetées. Cette méfiance s'appuie sur certains cas où le scénario adaptatif a été le premier avancé, puis a été par la suite démenti pour invoquer le rôle d'un processus neutre. Par exemple, les effets de la conversion génique biaisée vers GC (voir partie **II. 3. c.**) ont souvent été expliqués par des hypothèses adaptationnistes avant que l'on ne découvre l'existence d'un biais de réparation des séquences lié à la recombinaison. La structuration du génome en isochore (variations à large échelle de la composition en GC à travers le génome) chez les oiseaux et certains mammifères (Bernardi et al. 1985) a entre autre été liée à la plus forte stabilité thermique des paires GC que AT, qui seraient ainsi sélectionnées chez les poikilothermes (Belle et al. 2002). Un autre exemple est l'identification de zones à taux d'évolution accéléré dans le génome humain par rapport aux autres lignées (human accelerated regions en anglais, ou HARs), qui ont été considérées comme des zones sous sélection positive qui auraient contribué au développement de caractéristiques spécifiques au cerveau humain (Pollard et al. 2006). Il semblerait en fait que ces « HARs », ainsi que la structure en isochores soient plus probablement le résultat de la conversion génique biaisée vers GC (Galtier and Duret 2007). Pour pallier à ces problèmes, l'hypothèse neutre a été étendue pour incorporer des scénarios plus réalistes, comme l'existence de la conversion génique biaisée vers GC ou de variations démographiques, ce qui a permis la production de nombreuses prédictions théoriques neutres facilement comparables aux données observées, notamment en ce qui concerne le niveau et les

fréquences des variations intra-spécifiques (Hahn, 2018). Ainsi, la théorie neutre a permis le développement d'un « modèle nul » de l'évolution moléculaire, c'est-à-dire une hypothèse nulle ou neutraliste qu'il faut tester et rejeter avant de pouvoir évoquer la sélection comme mécanisme à l'origine du patron observé. Certains modèles ne se basent d'ailleurs que sur des forces non-adaptatives pour expliquer l'évolution de la structure des génomes (Lynch and Conery 2003, Lynch et al. 2006). De tels modèles nuls ont le mérite de pouvoir véritablement prouver un scénario adaptationniste, puisqu'ils offrent une hypothèse alternative aisément identifiable.

Malgré cela, d'autres évolutionnistes se placent en nette opposition à la théorie neutraliste, se basant sur les nombreuses preuves de l'influence de la sélection positive dans les génomes, apparues à la lumière des récents développements techniques (notamment en termes de séquençage) et résultats empiriques obtenus depuis les années 80. En effet, ils réfutent la pertinence de ce modèle nul en avançant un manque de pouvoir explicatif. Ainsi, encore très récemment, Kern & Hahn (2018) ont publié une revue soutenant que le modèle neutre de l'évolution était basé sur des résultats très préliminaires et qu'on peut aujourd'hui nier son caractère universel. Ils remettent également en cause l'utilisation du modèle neutraliste comme hypothèse de base de certains tests (comme le test de McDonald & Kreitman, voir plus loin), pointant du doigt l'ambivalence de reposer sur l'hypothèse neutraliste afin de la rejeter (ils reconnaissent malgré tout l'absence d'alternative vraiment satisfaisante) (Kern and Hahn 2018). Il existe en effet de nombreuses preuves de l'influence de la sélection positive, que ce soit en régions codantes ou non codantes, en termes de mutations ponctuelles ou de variations du nombre de copies de gènes (Kern & Hahn 2018). Malgré tout je pense que ces nombreuses preuves d'adaptations moléculaires ne remettent pas en cause la théorie quasi-neutre de Kimura et Ohta, et leur postulat que les mutations ou variations avantageuses soient rares dans les génomes. Sur de longues échelles de temps, il est normal d'observer une accumulation de ces mutations avantageuses puisqu'elles ont justement la capacité de se fixer au sein des espèces, et donc d'être observées aujourd'hui dans les données de divergence. Néanmoins, elles restent rares à un moment donné dans le temps en comparaison des mutations neutres ou faiblement délétères, comme le démontre l'existence d'une relation négative entre le taux de substitution protéique et la taille efficace (Nikolaev et al. 2007; Popadin et al. 2007; Lartillot and Delsuc 2012; Romiguier et al. 2014; Figuet et al. 2016). Le fait de baser un test de détection de la sélection positive sur une comparaison à l'attendu neutre est, à mes yeux, loin de constituer un « inconvénient », comme Kern & Hahn l'ont mentionné. Comme l'a souligné Martin Kreitman dès 1996 dans un article appelé « The neutral theory is dead. Long live the neutral theory », si le pouvoir explicatif de la théorie neutre n'est pas applicable à tous les mécanismes et phénomènes à l'échelle génomique, l'attendu neutre, utilisé en tant que modèle nul, est le moyen le

plus efficace de prouver l'existence d'un phénomène d'adaptation. Il souligne donc l'importance de l'utilisation de l'hypothèse neutre comme contrôle, et prédit qu'elle restera au centre de la quête vers la compréhension de l'évolution moléculaire (« Despite limitations in the applicability of the neutral theory, it is likely to remain an integral part of the quest to understand molecular evolution », (Kreitman 1996)), prédiction qui s'est avérée vérifiée aujourd'hui.

Ainsi, pour la majorité des chercheurs se penchant sur la question de la place de la sélection naturelle *vs.* les processus neutres dans les génomes, il semble principalement subsister de ce débat une question quantitative quant à la proportion de mutations adaptatives *vs.* non adaptatives en ségrégation ou fixées dans les génomes, et au lien entre cette proportion et les traits d'histoire de vie des espèces étudiées. Cela met en évidence l'importance du développement des méthodes d'estimation de la distribution des effets en fitness des mutations, ainsi que des taux de substitution adaptatifs et non-adaptatifs, qui sont illustrées dans la partie suivante.

3. Comment mesurer la sélection positive et la distribution des effets en fitness des mutations?

a. Détecter les loci sous sélection

Pour détecter les loci sous sélection, un ensemble de méthodes se basent sur la théorie de la coalescence, un modèle rétrospectif de génétique des populations, dont l'objectif est de remonter de génération en génération l'évolution de tous les allèles d'un gène donné de tous les individus d'une population, jusqu'à une seule copie ancestrale, appelée ancêtre commun le plus récent. La séparation de l'ancêtre commun en deux copies distinctes s'appelle un « évènement de coalescence ». Les tests basés sur ce principe s'appuient sur les prédictions théoriques du processus de coalescence, prédictions qui peuvent ensuite être confrontées aux données (Wright 1938). Typiquement, la sélection ou les variations démographiques distordent les généalogies prédites par le modèle de coalescence, et ainsi, distordent les spectres de fréquence alléliques (SFS). Ces SFS comptabilisent le nombre de sites polymorphes (i.e. SNPs pour « Single Nucleotide Polymorphisms ») en ségrégation pour chaque classe de fréquence de l'échantillon (soit 2n-1 classes de fréquences pour un échantillon contenant n individus diploïdes et 2n copies de chaque allèle). Ainsi, détecter ces écarts entre les prédictions et le patron observé peut permettre d'identifier les forces évolutives impactant les séquences étudiées. Par exemple, pour analyser facilement les distorsions des SFS par rapport à l'attendu, les tests se basent sur des statistiques résumées, comme le *D* de Tajima (Tajima 1989), la statistique F de Fu & Li (Fu and Li 1993) ou la statistique H de

Fay & Wu (Fay and Wu 2000). Premièrement, le *D* de Tajima se base sur une comparaison entre deux mesures du polymorphisme, appelées respectivement p et π, où p est le nombre de sites polymorphes en ségrégation sur le nombre de sites total, et π est le nombre moyen de différences entre les individus d'un échantillon sur le nombre de sites total. Sous l'hypothèse neutre, ces deux quantités sont égales, alors que sous sélection purificatrice ou positive, ou en cas de contraction ou d'expansion de la taille de population, elles diffèrent (en effet, elle ne donnent pas le même poids aux SNPs en faible fréquence). Deuxièmement, le F de Fu & Li vise à détecter les contractions démographiques via l'analyse des singletons (polymorphismes en fréquence 1/2n dans l'échantillon de diploïdes de taille n). Et troisièmement, le H de Fay & Wu a pour objectif de détecter les locus sous sélection positive via la détection des traces de balayages sélectifs dans les génomes, en se basant sur les différences entre les proportions d'allèles présents en fortes fréquences et ou en fréquences intermédiaires.

Par ailleurs, d'autres méthodes se basent sur les propriétés des loci en liaison pour inférer les zones du génome où se sont produits des événements d'adaptation. En effet, si les substitutions adaptatives sont fréquentes, alors on attend que la diversité génétique au niveau des sites en liaison décroisse (un phénomène appelé balayage sélectif). Cette réduction de la diversité génétique dépend du ratio du coefficient de sélection de la mutation avantageuse sur le taux de recombinaison. Grâce à cette propriété théorique, il est donc possible d'estimer le taux et la force de la sélection positive dans les génomes via la comparaison de la diversité génétique aux sites neutres en liaison (Booker et al. 2017).

Un autre ensemble de tests se base sur la comparaison des sites synonymes et non-synonymes. Les sites synonymes sont considérés comme neutres, et servent donc de contrôle en comparaison des sites non-synonymes qui eux sont sous sélection positive et purificatrice. En particulier, l'utilisation du ratio dN/dS permet de tester l'hypothèse neutre (évoquée précédemment) et mesurer la direction principale de la sélection si elle agit sur les séquences étudiées. Il s'agit du ratio du taux de substitution non-synonyme (où un changement non-synonyme est un changement nucléotidique au sein d'un gène codant qui entraîne un changement d'acide aminé dans la protéine) et du taux de substitution synonyme (où un changement synonyme est un changement nucléotidique au sein d'un gène codant qui n'entraîne pas un changement d'acide aminé dans la protéine du fait de la redondance du code génétique). Le ratio dN/dS peut être obtenu de deux façons différentes. Premièrement, en comptant le nombre de changements observés entre deux séquences codantes orthologues d'un alignement, et en comptant le nombre de sites synonymes et non-synonymes qui composent les séquences, c'est-à-dire le nombre d'opportunités d'un changement synonyme ou

non-synonyme (qui peut varier en fonction de la composition nucléotidique des séquences). Pour calculer ce nombre de sites, il est judicieux de prendre en compte le taux de transitions et transversions, ou encore l'usage inégal de certains codons synonymes. Deuxièmement, le dN/dS peut être calculé en utilisant des méthodes plus sophistiquées, fournissant des estimations par maximum de vraisemblance (Goldman and Yang 1994). Ces méthodes se basent sur une matrice de transition entre les 61 codons non-terminaux, et les paramètres κ (le ratio de transition/transversion), π_j (la fréquence d'équilibre du codon j) et ω (la probabilité d'observer des substitutions non-synonymes relativement aux substitutions synonymes, c'est-à-dire le dN/dS estimé). Si ces substitutions non-synonymes sont neutres, alors le ratio dN/dS vaut 1. Au contraire, si les gènes sont sous sélection purifiante, alors les mutations non-synonymes sont délétères ou faiblement délétères, et sont éliminées par la sélection et n'arrivent pas à fixation. Ainsi, elles ne participent pas à la divergence et le le dN/dS est plus petit que 1. Enfin, si le gène est sous sélection positive, alors il y a un excès de mutations avantageuses qui arrivent à fixation, et le dN/dS est alors plus grand que 1. Ce raisonnement a été utilisé pour la première fois par Nei et Gojobori en 1986, et ensuite repris dans de nombreuses études (par exemple Ford 2001, Johnson and Seger 2001, Lukens and Doebley 2001, et Welch and Meselson 2001). Néanmoins, cette vision des choses est un peu simpliste, et surtout très conservative. Un ratio dN/dS inférieur à 1 ne signifie pas forcément une absence de substitutions avantageuses. En effet, souvent au sein d'une protéine, seule une zone limitée est sous sélection positive (typiquement la zone correspondant au site actif de la protéine), alors que le reste est sous sélection purifiante, ce qui entraîne un dN/dS plus petit que 1 alors que certains sites sont sous sélection positive. Il faut que les épisodes d'adaptation soient fréquents au sein d'un gène pour pouvoir observer un ratio plus grand que 1. Cela explique que seulement 1 % des gènes orthologues entre l'humain et le chimpanzé montrent un ratio dN/dS significativement plus grand que 1 et que la plupart des gènes présentent un ratio inférieur à 0,3 (Waterson et al. 2005). Ce résultat est valable dans un grand nombre d'autres organismes, des métazoaires aux plantes (**Figure 1** pour une illustration chez *Arabidopsis*) (Consortium 2004; Waterson et al. 2005; Consortium 2007; Yang and Gaut 2011). Cela indique qu'une large majorité des mutations non-synonymes sont délétères et ne parviennent pas à fixation.

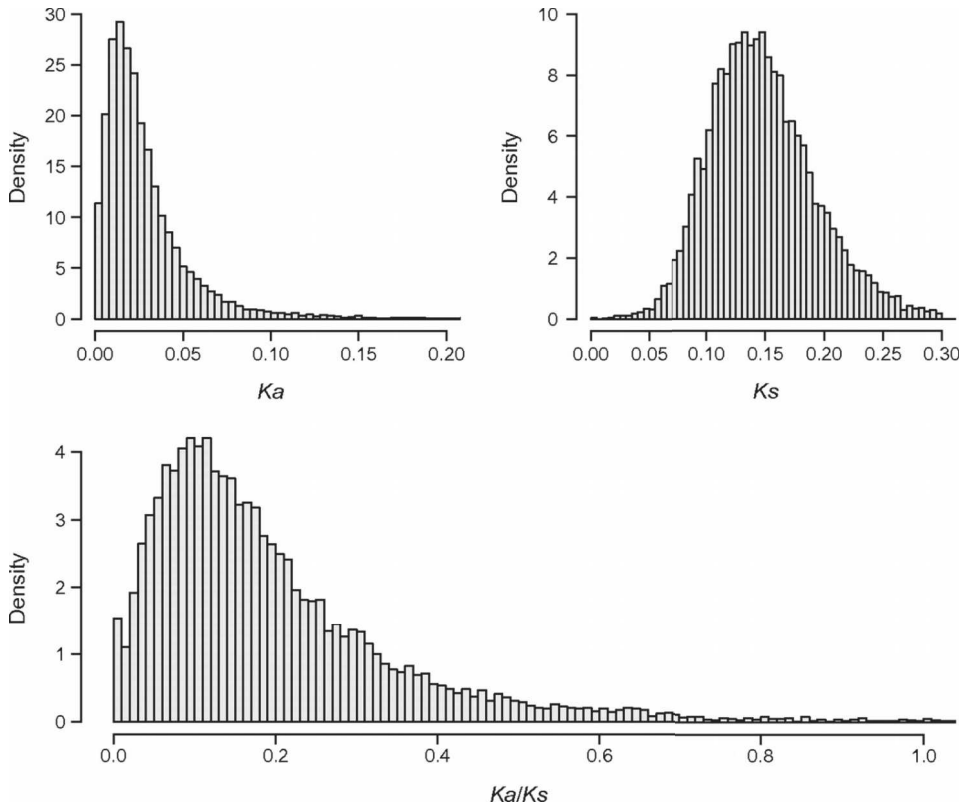


Figure 1 : Distribution du dN (aussi noté Ka), dS (aussi noté Ks) et dN/dS (Ka/Ks) chez les espèces du genre *Arabidopsis* (issu de Yang and Gaut 2011).

Ainsi, les méthodes d'estimation du ratio dN/dS par maximum de vraisemblance présentent l'avantage de permettre l'implémentation d'un test pour détecter les gènes ou codons sous sélection positive, i.e. qui présentent un $\omega>1$ (Yang 2000, Pond and Frost 2005). Néanmoins, d'autres phénomènes que la sélection peuvent être à l'origine d'une élévation du dN/dS, tels que le phénomène de biais d'usage du code ou la conversion génique biaisée vers GC (voir parie **II. 3. b.** et **c.**, et Ratnakumar et al. 2010). La comparaison des sites synonymes et non-synonymes peut aussi se faire en utilisant des données de polymorphisme, via le ratio π_n/π_s (où π_n est le nombre moyen de différences non-synonymes entre les individus de l'échantillon, et π_s est le nombre moyen de différences synonymes). L'interprétation de ce ratio est assez similaire à celle du dN/dS, mais il existe deux différences. D'une part, le π_n est susceptible de contenir une fraction non négligeable de mutations faiblement délétères, car même si elles sont maintenues à faible fréquence par la sélection purificatrice, elles contribuent au polymorphisme (ceci d'autant plus que N_e est faible). D'autre part, les mutations positives étant rares, et rapidement fixées par la sélection positive, elles contribuent très peu au polymorphisme échantillonné à un instant donné. Ainsi, des gènes qui présentent un ratio π_n/π_s plus grand que 1 sont souvent des gènes contenant plusieurs sites sous sélection balancée, du fait d'un avantage hétérozygote ou d'un phénomène de fréquence dépendance au sein de la

population (typiquement, ces rares cas sont des gènes du complexe majeur histocompatibilité, ou plus généralement liés à l'immunité (Hughes and Nei 1988)).

Enfin, l'utilisation des données de polymorphisme associées aux données de divergence a permis de développer d'autres tests, basés sur les prédictions des modèles neutres de génétique des populations, permettant de mesurer l'adaptation au sein des génomes ou au niveau d'un locus. En 1987, Hudson et al. développèrent le test HKA. Il se base sur l'idée assez simple que la sélection directionnelle entraîne une diminution de la diversité génétique par rapport à un locus neutre, et ainsi la comparaison du niveau de polymorphisme de différents loci peut renseigner sur les contraintes sélectives de ces loci. Mais des variations du taux de mutation à travers le génome peuvent aussi expliquer des différences de diversité entre loci, ainsi Hudson et al. développèrent l'idée d'utiliser la divergence entre espèce, via l'utilisation d'un groupe externe, comme une mesure du taux de mutation des différents loci (sous réserve que la divergence soit neutre) (Hudson et al. 1987). Par la suite, une approche plus précise et réservée aux séquences codantes, le test de McDonald & Kreitman, a été développée. Basée également sur la comparaison entre diversité intra-spécifique et inter-spécifique, cette approche utilise les changements synonymes comme référence neutre, ce qui constitue une nette amélioration par rapport au test HKA (voir partie **II. 1.**) (McDonald and Kreitman 1991).

b. Mesurer la distribution des effets en fitness des mutations

Estimer la distribution des effets en fitness des mutations (DFEM) constitue un enjeu primordial pour valider ou infirmer le modèle presque neutre de l'évolution. On classe souvent les mutations en grandes catégories, telles que délétères, faiblement délétères, neutres et adaptatives, mais il existe bien sûr un continuum d'effets en fitness des mutations, qu'il est intéressant de quantifier afin de pouvoir tester des affirmations telle que "…the great majority of evolutionary changes at the molecular level…are caused not by Darwinian selection but by random drift of selectively neutral or nearly neutral mutants" (Kimura 1983), et ainsi avoir une réelle idée de ce que « great majority » signifie exactement. Plusieurs approches ont été développées pour déterminer cette distribution, à la fois de manière expérimentale, théorique et analytique, leur point commun étant le besoin de données en très grande quantité.

Une des méthodes qui semble la plus directe est d'induire, ou laisser apparaître, des nouvelles mutations dans des populations suivies au laboratoire et d'identifier les effets sur la fitness de ces mutations. Il apparaît donc immédiatement que cette méthode présente de grandes

difficultés. En effet, il faut laisser évoluer les organismes sur de nombreuses générations, et disposer de populations d'assez grande taille efficace. Cela limite donc ces expérimentations aux micro-organismes. Devant ces difficultés, une autre approche expérimentale a été développée, l'expérience d'accumulation de mutations. Il faut initialement disposer d'un grand nombre de lignées génétiquement identiques, qu'on laisse ensuite accumuler indépendamment des mutations sur plusieurs générations. Il faut prendre garde à maintenir les populations à faibles effectifs, et dans des conditions idéales, afin de minimiser l'impact de la sélection purificatrice qui éliminerait les mutations délétères, empêchant ainsi leur observation. Ensuite, les valeurs sélectives des sous-populations sont mesurées (en général, elle décroît à mesure que l'expérimentation progresse du fait de l'accumulation des mutations délétères) et utilisées pour inférer la DFEM. Les méthodes permettant ces inférences sont néanmoins coûteuses en temps et en matériel, et présentent quelques limitations. D'une part, le taux de mutation est un facteur confondant et doit donc être estimé en même temps que la DFEM au cours de l'expérimentation. De plus, pour des questions de temps les mutations sont souvent provoquées (via des agents chimiques ou l'insertion d'éléments transposables par exemple), ce qui peut induire une DFEM différente de la DFEM en conditions naturelles. D'autre part, la DFEM estimée par cette approche est biaisée vers des mutations qui ont un fort effet sur la fitness, puisque pour mesurer leur effet, on doit pouvoir observer une augmentation ou une diminution du taux de croissance, et seules celles qui ont un effet détectable sur les phénotypes sont considérées comme non-neutres (Eyre-Walker and Keightley 2007). C'est pourquoi, en parallèle de ces tests expérimentaux, des méthodes se basant sur les séquences d'ADN ont été développées.

Ces méthodes fonctionnent via une comparaison de la divergence et/ou du polymorphisme à deux types de sites : ceux qui sont neutres, typiquement les sites introniques ou synonymes au sein des séquences codantes, et ceux qui sont potentiellement sous sélection, c'est à dire les sites non-synonymes des séquences codantes (comme les méthodes qui testent l'hypothèse de neutralité, cf. partie précédente). Ces méthodes se basent sur l'idée que le temps de séjour d'une mutation, ou la probabilité qu'elle ségrège avec une certaine fréquence, ainsi que sa probabilité de fixation, dépendent de son coefficient de sélection, s, et de la taille efficace de population, N_e. Si s est petit face à $1/N_e$, alors la mutation est effectivement neutre dans le sens où la dérive, et non la sélection, déterminera ses variations de fréquence. Du fait de ce dernier principe, les premières méthodes développées avaient pour but de comparer le niveau de contrainte des gènes codants entre des espèces à tailles efficaces différentes, afin d'estimer la forme de la DFEM au sein de ces espèces, et en particulier la prévalence des mutations faiblement délétères (Eyre-Walker et al. 2002, Piganeau and Eyre-Walker 2003, Sawyer et al. 2003, Loewe and Charlesworth 2006, Eyre-Walker

and Keightley 2007). Ces méthodes présentaient elles aussi certaines limitations, telles qu'un manque de pouvoir statistique du fait de l'utilisation de statistiques résumées comme le π_s et π_n pour le polymorphisme, et le dN et le dS pour la divergence.

Ainsi pour améliorer l'estimation de la DFEM, en tout cas la partie négative et neutre, une approche alternative consiste à ajuster un modèle de DFEM par maximum de vraisemblance aux spectres de fréquence allélique (Site Frequency Spectrum, SFS) synonymes et non-synonymes observés (**Figure 2**) (Sawyer et al. 1987 ; Eyre-Walker et al. 2006). Ces spectres résument l'information contenue dans le polymorphisme de manière beaucoup plus détaillée qu'une statistique telle que le ratio π_n/π_s, puisqu'ils comptabilisent le nombre de SNPs pour chaque classe de fréquence de l'échantillon.

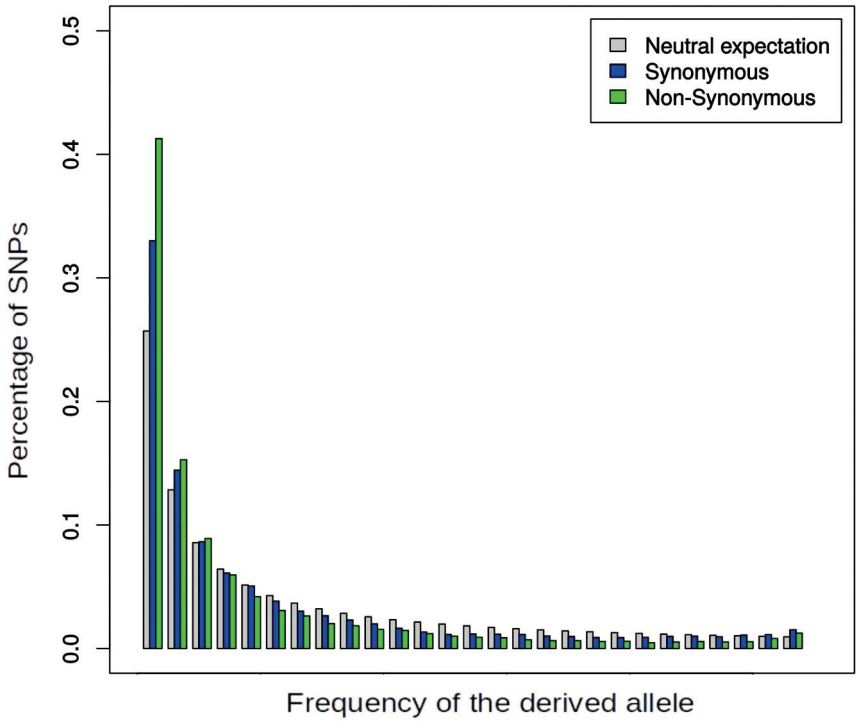


Figure 2 : Spectre de fréquences alléliques synonymes et non-synonymes observé chez 14 individus *Homo sapiens*, ainsi que le spectre neutre théorique attendu dans une population de Wright-Fisher.

Le spectre synonyme est considéré comme neutre, et on peut le comparer aux attendus provenant des modèles de génétique des populations établis précédemment, qui prédisent la forme d'un SFS réellement neutre dans une population panmictique de taille constante (**Figure 2**). Ainsi, on peut corriger le spectre synonyme observé pour tous les effets non-sélectifs qui peuvent affecter les SFS, tels que des variations démographiques, et reporter ces corrections sur le spectre non-synonyme observé. Ensuite, la comparaison entre le spectre synonyme et non-synonyme fournit des informations quant à la quantité de mutations faiblement délétères. En effet, ces mutations sont maintenues à faible fréquence par la sélection purificatrice, et vont donc s'accumuler dans les premières catégories du spectre non-synonyme, i.e. les classes de basses-fréquences. Ces classes vont donc être enrichies en SNPs par rapport aux mêmes classes de fréquence du spectre synonyme. Au delà de ce raisonnement intuitif, il existe des prédictions quantitatives sur l'attendu du nombre de SNPs non-synonymes en ségrégation à une certaine fréquence i, dans une population de taille efficace N_e, connaissant la DFEM. Cela se formalise comme ceci :

$$\hat{P_N}[i] = 2L_N N_e \mu \int_{-1}^{1} \phi(s) g(N_e, s, i, n) ds, \quad (1)$$

où $\phi(s)$ représente la distribution des coefficients de sélection s, c'est-à-dire la DFEM, L_N est le nombre de sites non-synonymes, μ le taux de mutation, et $g(N_e, s, i, n)$ représente la probabilité qu'une mutation de coefficient de sélection s ségrège à une fréquence i dans un échantillon de taille n.

$$g(N_e, s, i, n) = \int_{0}^{1} h(N_e, s, x) q(i, n, x) dx, \quad (2)$$

où $h(N_e, s, x)$ représente le temps de séjour relatif à une fréquence x d'une mutation de coefficient s, et $q(i, n, x)$ représente la probabilité qu'un allèle en ségrégation en fréquence x dans la population soit observé en fréquence i dans un sous-échantillon de taille n.
Le pendant synonyme (neutre) de l'équation (1) se formalise comme ceci :

$$\widehat{P_S}[i] = 4L_S N_e \mu / i, \quad (3)$$

où L_S est le nombre de sites synonymes.

(équations modifiées d'après Wright (1938))

Ainsi, grâce aux équations (1) et (3), et en faisant une hypothèse préalable sur le type de distribution de la DFEM, il est possible d'ajuster un modèle aux SFS observés et ainsi d'estimer les paramètres de cette DFEM par maximum de vraisemblance. Plusieurs distributions différentes ont

été utilisées pour modéliser la DFEM (distribution normale, exponentielle, ou Gamma), et préférées selon leur adéquation aux jeux de données. Néanmoins, la distribution Gamma semble être aujourd'hui la plus largement utilisée, à la fois parce qu'elle s'ajuste bien aux données dans de nombreux cas (Piganeau and Eyre-Walker 2003, Loewe and Charlesworth 2006, Galtier 2016), et parce qu'elle peut prendre une grande variété de formes différentes selon les deux paramètres qui la définissent : la forme et l'échelle.

Cette méthode présente deux avantages majeurs. Premièrement, elle autorise le taux de mutation et la taille efficace à varier entre les différents gènes. Et deuxièmement, elle permet la prise en compte implicite de contractions ou d'expansions démographiques récentes, de différentes manières possibles (voir partie **II. 2. c**). Néanmoins, elle fait aussi certaines hypothèses, comme l'absence de phénomène de dominance ou d'épistasie, et surtout, qu'il n'y ait pas de liaison génétique entre les différents loci (Eyre-Walker et al. 2006). En outre, il persiste certaines limitations à cette méthode. D'une part, elle nécessite des sites qui servent de référence neutre, or on est rarement sûr de la neutralité, même pour des éléments souvent considérés comme tels, tels que les sites synonymes ou les séquences introniques. En effet, les sites synonymes peuvent être soumis à un biais d'usage du code qui semble être, dans certains cas, adaptatif (voir partie **II. 3. b.**). Les séquences introniques, quant à elles, se sont révélées potentiellement soumises à sélection chez certains organismes (Shabalina and Kondrashov 1999, Bergman & Kreitman 2001, Keightley & Gaffney 2003, Andolfatto 2005, Haddrill et al. 2005). Ainsi, si les sites considérés comme neutres sont en réalité sous sélection, l'estimation de la DFEM peut être biaisée de manière difficilement prédictible.

Les méthodes décrites ci-dessus ne concernent que l'estimation des DFEM négatives. Il a été longtemps considéré que les mutations adaptatives sont trop rares pour apparaître dans les SFS (McDonald and Kreitman 1991), mais il apparaît désormais que cette hypothèse est réductrice, comme discuté ci-après. De plus, les méthodes empiriques ne sont pas toujours très efficaces pour estimer cette partie positive. En effet, d'une part les mutations observées dans les expériences d'accumulation de mutations par exemple ne sont pas vraiment des mutations *de novo*, mais des mutations qui ont échappé à une perte stochastique durant les quelques générations suivant leur apparition. D'autre part, beaucoup de mutations adaptatives sont perdues quand elles sont à faible fréquence (Haldane 1927). Estimer la DFEM positive à partir de données de séquences apparaît donc comme un exercice encore plus délicat que la DFEM négative. On dispose pourtant de prédictions théoriques quant à la DFEM positive : Orr (2003) a démontré théoriquement que, sous une hypothèse très peu restrictive, la DFEM positive attendue d'un gène suit une distribution

exponentielle (Orr 2003). Ainsi, certaines méthodes utilisent les spectres de fréquence allélique, parfois en association avec les données de divergence, pour estimer les paramètres d'une distribution exponentielle des effets en fitness positive (Galtier 2016, Tataru et al. 2017). Ces méthodes font l'hypothèse, contrairement au test de McDonald & Kreitman, qu'il y a effectivement des mutations avantageuses qui ségrègent de manière non négligeable dans les données de polymorphisme résumées par les spectres de fréquence allélique. Cette hypothèse est soutenue par le fait que la présence de mutations faiblement délétères devrait également impliquer la présence de mutations faiblement avantageuses (Charlesworth and Eyre-Walker 2007).

II. Estimer le taux de substitution adaptatif d'une espèce

1. Le test de McDonald & Kreitman

a. Méthode historique

En 1991, McDonald & Kreitman proposent un test statistique de l'hypothèse neutre de l'évolution des protéines basé sur la comparaison entre taux de substitution non-synonyme et taux de substitution synonyme, à la fois entre les espèces et entre les individus au sein des espèces. Ils se basent sur l'hypothèse selon laquelle si toutes les substitutions et mutations sont neutres, alors le ratio du taux de substitution non-synonyme sur le taux de substitution synonyme (dN/dS) est le même que le ratio de changements intra-spécifiques non-synonymes sur synonymes (appelé p_n/p_s). Ici, la divergence et le polymorphisme synonymes sont utilisés comme contrôle, afin de calibrer les deux ratios.

Au contraire, si les mutations et substitutions non-synonymes sont positivement ou négativement sélectionnées, on s'attend à une différence entre les deux ratios. En effet, ils s'appuient sur l'hypothèse neutraliste selon laquelle les mutations avantageuses sont très rares et se fixent vite. Ainsi, ces dernières sont très peu présentes à un moment donné au sein des mutations non-synonymes en ségrégation. Par contre, elles s'accumulent dans la divergence non-synonyme. Les mutations délétères, quant à elles, apparaissent en proportion plus importantes, et selon la force de la dérive peuvent demeurer un certain nombre de générations en ségrégation. En revanche, elles n'arrivent souvent pas à fixation puisqu'elles sont éliminées avant par la sélection purificatrice. Ainsi, si l'on observe un ratio dN/dS plus grand que le ratio p_n/p_s, la différence est attribuée à la présence de substitutions avantageuses. Si le ratio p_n/p_s est plus grand que le ratio dN/dS, alors la

différence est attribuée à la présence de mutations faiblement délétères. Ainsi, on peut estimer l' « index de neutralité » ou *NI*, qui indique la direction et la force de l'écart à la neutralité :

$$NI = (p_n/p_s)/(dN/dS). \quad (4)$$

Cette équation est également applicable en utilisant le ratio π_n/π_s au lieu du ratio p_n/p_s. π_n et π_s représentent respectivement le nombre moyen de différences synonymes et non-synonymes entre deux chromosomes d'un échantillon. La différence entre ces deux ratios réside dans leur sensibilité à la présence des allèles rares, p étant peu sensible puisque tous les SNPs ont le même poids quelle que soit leur fréquence alors que π y est très sensible.

Ce test est à l'origine du développement par Smith & Eyre-Walker d'une « extension » du test de McDonald & Kreitman, selon leurs propres mots, qui permet d'obtenir la proportion de substitutions non-synonymes avantageuses, appelée α, telle que :

$$\alpha = 1 - (dS * p_n/dN * p_s) \quad (5)$$
$$\text{ou } \alpha = 1 - NI.$$

(Smith & Eyre-Walker, 2002)

En découle ensuite le taux de substitution adaptatif, noté ω_a, où

$$\omega_a = \alpha * (dN/dS), \quad (6)$$

ainsi que le taux de substitution non-adaptatif ω_{na}, où

$$\omega_{na} = (1 - \alpha) * (dN/dS). \quad (7)$$

Dans leur article fondateur de 1991, McDonald & Kreitman appliquèrent leur nouvelle méthode sur le gène *Adh* codant pour la protéine alcool-déshydrogénase, chez trois espèces de *Drosophila* : *D. melanogaster, D. simulans* et *D. yakuba.* Ils conclurent par un rejet de l'hypothèse neutre, puisqu'ils identifièrent proportionnellement plus de substitutions non-synonymes que de polymorphismes non-synonymes qu'ils attribuèrent à la présence de substitutions adaptatives. Ce résultat fut ensuite validé, à plus grande échelle, puisque Smith & Eyre-Walker (2002) estimèrent un α de 45 % au cours de la divergence entre *D. simulans* et *D. yakuba* sur un jeu de données de 43 gènes. Peu de temps après, les premières estimations entre l'homme et le chimpanzé ont indiqué un

α non-significativement différent de zéro dans plusieurs études (Waterson et al. 2005, Zhang and Li 2005), suggérant un impact très faible de la sélection positive dans la lignée humaine. Ces premiers résultats préliminaires ont été à l'origine de l'hypothèse d'une corrélation entre taille efficace et taux de substitution adaptatif (voir partie **III. 1. b.**).

b. Les problèmes liés au test de McDonald & Kreitman

Ce test est sensible à plusieurs sources d'erreurs, qui avaient d'ailleurs été initialement identifiées par McDonald & Kreitman. Ces sources incluent en particulier la présence de mutations faiblement délétères en ségrégation de manière substantielle et les variations de taille de population efficace, mais aussi la variation du taux de mutation et des histoires de coalescence de gènes à travers le génome. En règle général, ces sources de biais causent une sous-estimation de α. Prenons par exemple la présence de mutations faiblement délétères, souvent considérée comme le plus important problème potentiel du test de McDonald & Kreitman : ces mutations se maintiennent en ségrégation du fait de la dérive, et ainsi contribuent au p_n ou π_n. En revanche, elles arrivent rarement à fixation, du fait de la sélection purificatrice qui les maintient à faible ou moyenne fréquence, et ainsi elles ne contribuent pas au dN. La différence entre ratio dN/dS et p_n/p_s (ou π_n/π_s) est alors réduite et ne représente pas exactement la fraction adaptative des substitutions. Dans certains cas, la présence de mutations faiblement délétères peut même induire des estimations négatives de α, ce qui n'a pas de sens biologiquement (Deinum et al. 2015). Le problème posé par les variations de taille efficace est lié à cette dernière explication. En effet, une réduction de N_e par exemple entraîne une baisse de l'efficacité de la sélection purificatrice, et donc favorise le maintient en ségrégation de mutations dont le coefficient de sélection devient petit face à $1/N_e$. De manière plus générale, si le régime de sélection n'a pas été le même entre la période durant laquelle s'est construite la divergence (c'est-à-dire depuis la divergence entre les deux espèces étudiées), et la période durant laquelle s'est construit le polymorphisme (c'est-à-dire le temps qu'il faut pour remonter à l'ancêtre commun à tous les allèles présents dans l'échantillon, qu'on estime en moyenne à $4N_e$ générations (Tajima 1983)), alors on s'attend à ce que le test ne soit pas valide. Ainsi, les variations récentes aussi bien que passées sont importantes à prendre en compte, mais les variations de long-terme sont très difficiles à retracer comme leur signal est effacé des données de séquences (voir partie **II. 3. a.**). Ces problèmes ont été mis en évidence théoriquement en 2008 par Charlesworth & Eyre-Walker, puis empiriquement en 2009 par Parsch et al. (2009), via une comparaison des estimations de α dans deux populations de *D. melanogaster* dont l'une a subi un goulot d'étranglement récent. Ils montrent que la population avec une taille de population réduite a en effet un α plus faible que

l'autre, du fait à la fois de la présence de plus de mutations faiblement délétères et de la réduction de puissance statistique due à la baisse de variation intra-spécifique (Parsch et al. 2009).

2. Complexifier la méthode pour corriger ses biais

a. Éliminer les mutations faiblement délétères

Afin d'éviter les biais issus de la présence de mutations faiblement délétères en ségrégation, qui entraîne une sous-estimation de α, Fay et al. (2001) proposèrent une solution consistant à éliminer des jeux de données les SNPs ségrégeant à faible fréquence. En effet, les mutations faiblement délétères ont une forte probabilité d'être maintenues à faible fréquence par la sélection purificatrice (Fay et al. 2001). Cette procédure s'est révélée en effet efficace pour diminuer la sous-estimation de α, puisque les estimations sans SNPs à faible fréquence se sont avérées plus élevées qu'en conservant ces SNPs dans les jeux de données, et ce pour différents jeux de données (Fay et al. 2001, Bierne and Eyre-Walker 2004, Charlesworth and Eyre-Walker 2006). Mais cette méthode présente d'importantes limites. En effet, il semble difficile d'estimer le seuil de fréquence à partir duquel on peut penser éliminer toutes les mutations faiblement délétères; ce seuil est donc arbitraire. De plus, il est possible que des mutations faiblement délétères ségrègent à de plus fortes fréquences (même si la probabilité est faible), en particulier dans les zones de faible recombinaison où les mutations faiblement délétères peuvent être associées à une mutation avantageuse et monter en fréquence par auto-stop génétique (Messer and Petrov 2013). Cette mutation pourra ensuite mettre du temps à revenir à faible fréquence. Néanmoins, l'idée d'utiliser l'information de la fréquence des SNPs plutôt que de se concentrer uniquement sur une statistique résumée telle que le p ou le π est à l'origine des derniers développements les plus précis en manière d'estimation de α (voir partie **II. 2. c.**).

b. L'index de neutralité

D'autres problèmes du test de McDonald & Kreitman, moins souvent évoqués, résident dans le fait que α ou *NI* sont estimées via un ratio de deux ratios. Ainsi, ces deux statistiques peuvent être biaisées ou avoir une très forte variance surtout quand les comptes de SNPs ou de substitutions sont faibles (quand les estimations sont faites gène par gène par exemple) (Stoletzki & Eyre-Walker 2010). Cela signifie entre autre que ce biais dépend de la taille de l'échantillon étudié, ce qui pose un problème majeur. De plus, ce double ratio exclut les gènes pour lesquels dN ou p_S

valent zéro. Pour pallier à ces problèmes, Stoletzki & Eyre-Walker ont développé une autre statistique appelée « Direction of selection » (DoS), qui se définit comme ceci :

$$DoS = \frac{dN}{dN + dS} - \frac{p_n}{p_n + p_s} \quad (8)$$

(Stoletzki and Eyre-Walker 2010)

Cette statistique mesure, comme son nom l'indique, la direction de la sélection prépondérante qui agit sur un locus, selon son signe, et l'intensité de la sélection, selon sa valeur absolue. Le DoS attendu est positif quand le gène est sous sélection positive, nul s'il n'y a que des mutations neutres, et négatif en présence de mutation faiblement délétères.

Cette statistique présente l'avantage de pouvoir être estimée à partir d'un très petit nombre de gènes (voir même un seul), ce qui peut être utile si on veut faire le lien entre les effets en fitness des mutations et le rôle fonctionnel du gène. Les méthodes présentées dans la partie suivante ne le permettent pas, par exemple. Néanmoins, le DoS reste comme le *NI* ou le α estimés classiquement, sensible à la présence de mutations faiblement délétères en ségrégation.

c. Utilisation des spectres de fréquences alléliques

Les deux solutions exposées précédemment fournissent des améliorations face à certains biais de la méthode initiale de McDonald & Kreitman, sans être toutefois satisfaisantes face au plus gros problème de ce test : la présence de mutations faiblement délétères en ségrégation. Comme souligné pour la première fois par Fay et al. (2001), l'utilisation de l'information de la fréquence des SNPs synonymes et non-synonymes semble être la bonne voie pour pallier à ce problème. Ainsi à partir de 2007 des méthodes basées sur les spectres de fréquence allélique se sont développées (Eyre-Walker et al. 2006; Keightley and Eyre-Walker 2007; Boyko et al. 2008; Eyre-Walker and Keightley 2009). Ces méthodes visent à estimer les paramètres de la DFEM négative (et parfois positive, Galtier 2016 et Tataru et al. 2017) par maximum de vraisemblance comme présenté dans la partie **I. 3. b,** en ajustant à la fois un modèle de polymorphisme et un modèle de divergence aux données. Une fois les paramètres de la DFEM négative établis, il est possible d'en déduire le nombre attendu de mutations délétères ou neutres qui vont arriver à fixation. Autrement dit, il est possible d'inférer le taux de substitution non-adaptatif (ω_{na}) attendu sous une certaine DFEM. Si on

reprend le formalisme exposé dans la partie **I. 3. b.** (équations 1, 2 et 3), on peut écrire le nombre de substitutions non-synonymes non-adaptatives attendu comme suit :

$$\widehat{D_N^{na}} = 2L_N N_e t\mu \int_{-\infty}^{0} \phi(s) f(N_e, s) ds, \quad (9)$$

où L_N est le nombre de sites non-synonymes, t est le temps de divergence de la branche considérée, et $\phi.f(Ne, s)$ représente la probabilité de fixation d'une mutation de coefficient de sélection s dans une population de taille N_e. ω_{na} est alors obtenu comme suit :

$$\omega_{na} = (\widehat{D_N^{na}}/L_N)/(D_S/L_S), \quad (10)$$

où D_S est le nombre de substitutions synonymes observé, et L_S le nombre de sites synonymes.
La comparaison de ω_{na} attendu et du ratio dN/dS observé fourni alors des informations quant au taux de substitution adaptatif ω_a, puisque $dN/dS = \omega_{na} + \omega_a$ (Boyko et al. 2008, Eyre-Walker & Keightley 2009, Galtier 2016, Tataru 2017). On a alors :

$$\omega_a = (d_N - \widehat{d_N^{na}})/d_S \quad (11)$$

$$\text{et } \alpha = (d_N - \widehat{d_N^{na}})/d_N \quad (12)$$

Certaines variantes de cette méthode incluent également une partie positive au modèle de DFEM, modélisée par une distribution exponentielle (Galtier 2016, Tataru et al. 2017).

Un récapitulatif des différentes méthodes successives et leurs principaux avantages et sources de biais est présenté dans le **Tableau 1**.

Méthode	Description principale et variantes	Caractéristiques principales
Test historique de McDonald & Kreitman (et extension de Smith & Eyre-Walker (2002) pour estimer α et ω_a)	$\alpha = 1 - (dS * p_n/dN * p_s)$ou $\alpha = 1 - (dS * \pi_n/dN * \pi_s)$ et	Donne une estimation de α et ω_a très sensible à la présence de **mutations faiblement délétères** et à des **variations démographiques récentes et de long-terme**. Biais lié au fait que le test repose sur un **ratio de ratio**.
	Idem, mais p_n et π_n sont estimés en éliminant au préalable les SNPs à faible fréquence.	Réduit un peu le problème de mutations délétères en ségrégation **Choix du seuil de fréquence arbitraire**.
DoS	$DoS = \frac{dN}{dN + dS} - \frac{p_n}{p_n + p_s}$	Ne repose **plus sur un ratio de ratio** **Est estimable gène par gène**.
Méthode DFE-α	1. Estimation des paramètres de la DFE négative à partir des SFS synonymes et non-synonymes 2. Estimation du taux de substitution non-synonyme non-adaptatif ($\widehat{D_N^{na}}$) grâce aux paramètres de la DFE 3. Comparaison de $\widehat{D_N^{na}}$ et du dN/dS observé : $\omega_a = (d_N - \widehat{d_N^{na}})/d_S$ et $\alpha = (d_N - \widehat{d_N^{na}})/d_N$	Permet la prise en compte des **mutations faiblement délétères** et des **variations démographiques récentes.** **Est difficilement estimable gène à gène.** Fourni des estimations sensibles aux **fluctuations démographiques de long-terme.**
	1. Estimation des paramètres de la DFE négative **et positive** à partir des SFS synonymes et non-synonymes. 2. Estimation du taux de substitution non-synonyme non-adaptatif ($\widehat{D_N^{na}}$) grâce aux paramètres de la DFE 3. Comparaison de $\widehat{D_N^{na}}$ et du dN/dS observé : $\omega_a = (d_N - \widehat{d_N^{na}})/d_S$ et $\alpha = (d_N - \widehat{d_N^{na}})/d_N$	Permet la prise en compte des **mutations faiblement délétères** et des **variations démographiques récentes.** **Est difficilement estimable gène à gène.** Fourni des estimations sensibles aux **fluctuations démographiques de long-terme.**
	1. Estimation des paramètres de la DFE négative **et positive** à partir des SFS synonymes et non-synonymes et **du dN/dS observé.** 2. Estimation du taux de substitution non-synonyme non-adaptatif ($\widehat{D_N^{na}}$) **et adaptatif (**$\widehat{D_N^{a}}$**)** grâce aux paramètres de la DFE 3. $\omega_a = \widehat{D_N^{a}}/d_S$ et $\alpha = \widehat{D_N^{a}}/d_N$	Permet la prise en compte des **mutations faiblement délétères** et des **variations démographiques récentes.** Fourni des estimations un peu moins sensibles aux fluctuations démographiques de long-terme. **Est difficilement estimable gène à gène.**
	1. Estimation des paramètres de la DFE négative **et positive** à partir des **SFS synonymes et non-synonymes uniquemen**t. 2. Estimation du taux de substitution non-synonyme non-adaptatif ($\widehat{D_N^{na}}$) **et adaptatif** ($\widehat{D_N^{a}}$) grâce aux paramètres de la DFE 3. $\omega_a = \widehat{D_N^{a}}/d_S$ et $\alpha = \widehat{D_N^{a}}/d_N$	Permet la prise en compte des **mutations faiblement délétères** et des **variations démographiques récentes et de long-terme.** **Est difficilement estimable gène à gène.** Nécessite des jeux de données de **très grande qualité.**

Tableau 1 : Récapitulatif des différentes méthodes dérivées du test de McDonald & Kreitman.

3. Des facteurs de biais qui subsistent ?

a. Les variations démographiques récentes et passées

Ainsi, la méthode présentée ci-dessus résout théoriquement le problème de la présence de mutations faiblement délétères en ségrégation. Malgré tout, ce test fait aussi l'hypothèse que le régime de sélection est resté constant pendant la période étudiée, c'est-à-dire depuis la séparation entre l'espèce focale, dont on a les données de polymorphisme, et l'espèce utilisée comme groupe externe pour estimer la divergence (**Figure 3**).

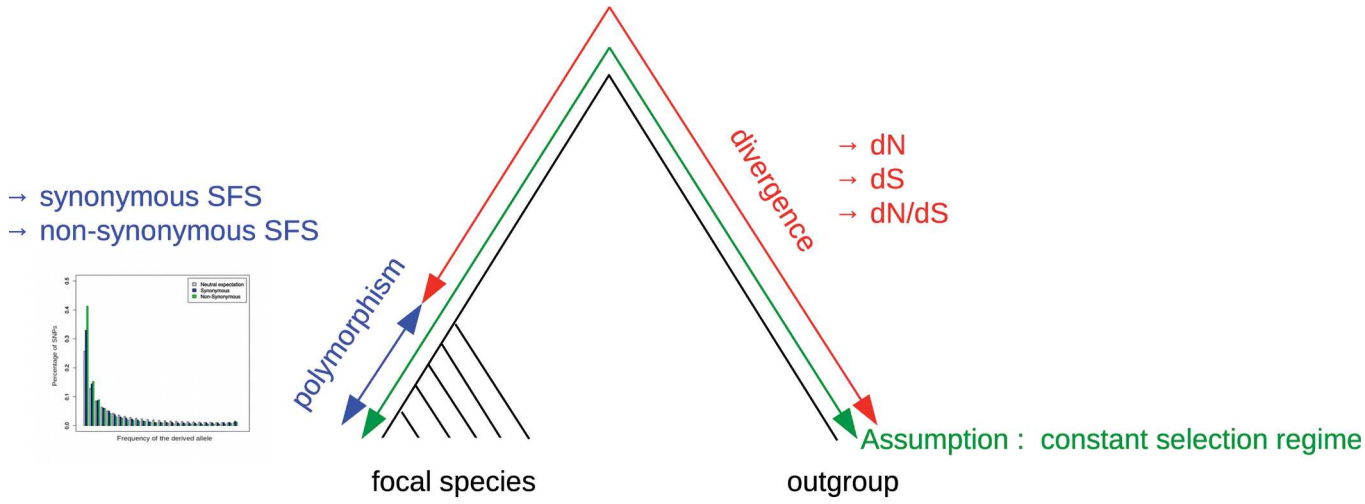


Figure 3 : **Structure du jeu de données nécessaire pour la méthode DFE-α.**

Or, des variations du régime de sélection peuvent se produire, puisque de simples variations démographiques provoquent un changement de l'efficacité de la sélection. Ces variations peuvent survenir dans le passé lointain, influençant donc la divergence, mais pas le polymorphisme. En effet, si les variations sont trop anciennes, les traces génomiques de la variation ne sont plus présentes dans les données de polymorphisme. Le temps de coalescence moyen de la population pour un locus neutre vaut $4N_e$ générations (Tajima 1983). Des fluctuations démographiques antérieures à cette date seront très difficiles à détecter et à prendre en compte. Les variations démographiques peuvent également se produire dans la période récente ($<\sim 4N_e$). Alors, elles influencent le polymorphisme et non la divergence. Ce cas là est plus facile à prendre en compte,

car le signal de telles variations est toujours présent dans les données, et influence notamment la forme des SFS. Ainsi, il est possible de prendre en compte ces fluctuations démographiques récentes dans la méthode présentée ici, et ce de deux manières différentes. La première consiste à autoriser dans le modèle l'ajustement aux données de polymorphisme à un changement de N_e à un temps donné, qui a pu conduire à un changement du régime de sélection (Boyko et al. 2008**,** Eyre-Walker and Keightley 2009). La seconde consiste à capturer les écarts du spectre synonyme dues aux variations démographiques par rapport au spectre neutre théorique d'une population de Wright-Fisher à l'équilibre (c'est-à-dire une population théorique idéale, de taille constante, à générations non-chevauchantes, et à l'équilibre mutation-dérive, donc sans sélection) via des paramètres de nuisances appelés r_i (Eyre-Walker et al. 2006, Galtier 2016). Il existe un tel paramètre pour chaque catégorie de fréquence du spectre. Ces paramètres sont utilisés ensuite pour corriger à la fois le spectre synonyme et le spectre non-synonyme. Ces paramètres ont moins de sens biologique qu'un modèle de changement de N_e, mais ils ont l'avantage d'être très flexibles pour prendre en compte beaucoup d'effets à l'origine de distorsions des spectres par rapport à l'attendu. En effet, ils sont supposés capturer aussi bien des distorsions dues aux variations démographiques, à de la structure de population ou à la présence de SNPs mal orientés dans les spectres (Eyre-Walker et al. 2006, Galtier 2016). Néanmoins, ils ne sont efficaces que si l'effet pour lequel ils corrigent distord de la même façon les spectres observés synonyme et non-synonyme.

Ainsi, si les méthodes actuelles peuvent corriger avec succès les effets de variations démographiques récentes, il reste néanmoins la possibilité que le régime de sélection ne soit pas le même entre la période durant laquelle s'est construite la divergence et la période plus récente, représentée par les données de polymorphisme. Une solution à ce problème potentiel serait d'estimer le taux de substitution adaptatif en utilisant une seule source d'information. Tataru et al. (2017) ont proposé récemment une méthode pour estimer α, ω_a et ω_{na} à partir des données de polymorphisme uniquement (Tataru et al. 2017). Cette méthode s'est avérée efficace pour estimer α dans les cas où la DFEM n'est pas la même entre l'espèce focale et le groupe externe. Cette étude démontre aussi que ne pas prendre en compte la présence de mutations avantageuses dans les SFS peut mener à une estimation biaisée de la DFEM négative. Néanmoins, les estimations de α via cette méthode présentent une forte variance, mettant en évidence le fait que cette méthode nécessite de grands jeux de données de très bonne qualité pour être efficace.

b. Le biais d'usage du code

Comme déjà mentionné, un point commun à toutes les méthodes d'estimation de DFEM et du taux de substitution adaptatif est la nécessité de disposer de sites neutres qui servent de référence. Très souvent, les sites neutres utilisés sont les sites synonymes. Mais ces sites peuvent être soumis à sélection, un phénomène appelé la sélection traductionnelle. Il a été observé à travers l'arbre du vivant que certains codons codant pour un même acide aminé sont plus fréquemment présents dans les séquences codantes que les autres. Chez certaines espèces, ces codons préférés correspondent aux ARN de transfert (ARNt) les plus abondants dans les cellules. Ce constat a mené à l'hypothèse qu'il s'agit d'une co-adaptation permettant d'optimiser l'efficacité de la traduction, car le ribosome peut progresser plus vite sur les brins d'ARN messagers, puisque sa progression semble être inversement proportionnelle à la concentration dans la cellule d'ARNt complémentaires à une séquence donnée. Dans des cellules où la quantité de ribosomes est limitée, il apparaît comme primordial de les utiliser avec efficacité. De plus, l'utilisation de codons synonymes correspondant aux ARNt les plus abondants permet également de limiter les risques d'erreurs de traduction, qui peuvent s'avérer coûteuses pour la cellule si elles génèrent des protéines non-fonctionnelles, du fait de la dépense d'énergie inutile, voire la toxicité de certaines protéines non-fonctionnelles qui ne se replient pas correctement. Au vu de ces constats, on peut faire l'hypothèse que plus un gène est fortement exprimé, plus on peut s'attendre à un fort biais d'usage du code car la pression de sélection traductionnelle est plus forte, ce qui permet de tester cette hypothèse. Les résultats empiriques ont mis en évidence l'existence de ce phénomène chez *Escherichia coli*, *Saccharomyces cerevisiae* (Ikemura 1985), chez le nématode *Caenorhabditis elegans* (Duret and Mouchiroud 1999), chez *Daphnia pulex* (Lynch et al. 2017), et enfin chez certaines Drosophiles (Shields et al. 1988; Akashi 1994; Bierne and Eyre-Walker 2006). Une autre hypothèse adaptationniste propose que le biais d'usage du code est sous sélection pour favoriser la stabilité thermodynamique des molécules d'ARN messagers (Chamary and Hurst 2005).

Dans les espèces où la sélection sur l'usage des codons synonymes se produit effectivement, les sites synonymes sont sous contrainte et présentent donc moins de substitutions que s'ils étaient vraiment neutres. Cela entraîne une sous-estimation du taux de substitution neutre, et ainsi une sous-estimation du niveau de contrainte sur les changements non-synonymes. Si la sélection traductionnelle n'est plus à l'œuvre, en revanche, le niveau de contrainte sur la protéine est surestimé (Eyre-Walker et al. 2002).

Alternativement, le biais d'usage du code peut être dû à un biais mutationnel, ou encore à la conversion génique biaisée vers GC (gBGC, voir partie suivante). Des résultats récents suggèrent que la sélection traductionnelle n'est efficace que dans les espèces à grande taille de population (Galtier et al. 2018). Par exemple, chez les Mammifères (faible N_e), aucun lien entre l'usage des codons synonymes et le niveau d'expression des gènes, ni de claire adéquation des codons préférés et des ARNt les plus abondants n'ont été détectés (Urrutia and Hurst 2001; Duret 2002; Sémon et al. 2005). Ceci peut s'expliquer par le fait que l'avantage procuré par une mutation ponctuelle vers un codon préféré a un coefficient de sélection très faible. Ainsi, il est négligeable face au coefficient de dérive, $1/N_e$, si N_e est assez petit. Par conséquent, le biais d'usage du code chez les grands vertébrés ou les insectes sociaux, c'est-à-dire chez des espèces à N_e faible (parfois 100 fois moins que chez la drosophile (Galtier et al. 2018)), semble être en majorité dû à la gBGC qui favorise les allèles G et C et les rend plus fréquents en troisième positions des codons, créant donc un biais dans leur usage (Galtier et al. 2018). Chez ces espèces, il n'y a donc pas de risque de sous-estimer le taux de substitution neutre du fait du biais d'usage du code. En revanche, la gBGC est elle aussi susceptible d'entraîner un biais dans les méthodes d'estimation de DFEM et du taux de substitution adaptatif.

c. La conversion génique biaisée vers GC

La conversion génique biaisée vers GC est un biais de fixation au profit des allèles G et C, et au détriment des allèles A et T, qui est causé par un biais du système de réparation de l'ADN qui œuvre lors des événements de recombinaison méiotique. La recombinaison implique la formation d'un hétéroduplex, c'est-à-dire l'association de brins d'ADN de chromosomes maternel et paternel. Cet ADN de mélange présentera donc des mésappariements aux positions hétérozygotes. Ces mésappariements sont détectés par la machinerie de réparation de l'ADN, qui va alors rétablir une paire de nucléotides complémentaires en corrigeant un des deux brins. Chez de nombreuses espèces, cette réparation est biaisée vers G et C: en cas de mésappariement de type T-G, par exemple, c'est l'allèle T qui va préférentiellement être remplacé par un nucléotide C, et non G par un A (**Figure 4**). De ce fait, les allèles G et C ont une plus forte probabilité d'être transmis à la descendance, et ont donc une plus forte probabilité de fixation, tout comme une mutation positivement sélectionnée. C'est pour cela que l'on considère que la gBGC mime les effets de la sélection positive (Eyre-Walker 1993; Galtier et al. 2001; Glémin et al. 2015). L'effet de la gBGC est bien-sûr prépondérant dans les régions chromosomiques fortement recombinantes.

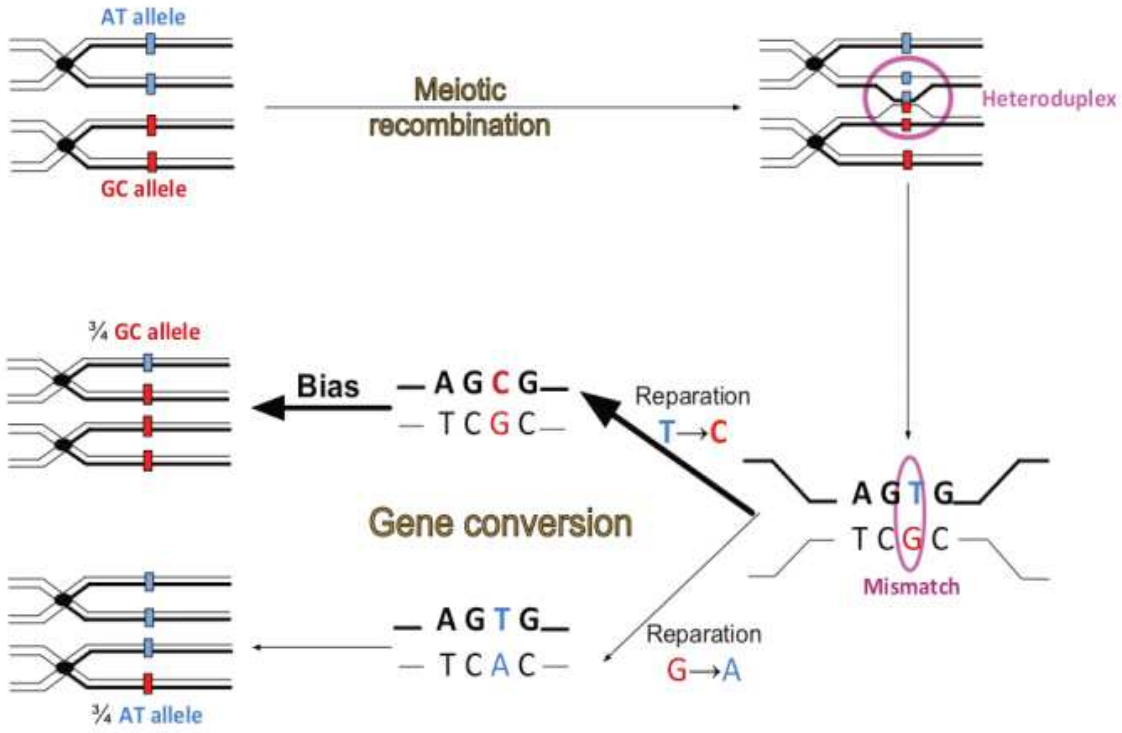


Figure 4 : Mécanisme de la conversion génique biaisée vers GC.

La gBGC pourrait entraîner une surestimation de α et ω_a car les allèles G et C fixés du fait de la gBGC peuvent être confondus avec des allèles fixés par sélection positive, un biais que l'on peut qualifier de « biologique ». Mais la gBGC peut aussi être à l'origine d'un autre biais plus « technique », au sens où le modèle sous-jacent aux méthodes de type McDonald-Kreitman ne la considère pas. Plusieurs études semblent indiquer que les effets de ce mécanisme ne sont pas les mêmes sur les sites synonymes et les sites non-synonymes. En effet, la gBGC influence les ratios tels que le dN/dS. Chez les primates, par exemple, la gBGC entraîne une élévation du dN/dS des gènes préalablement identifiés comme ayant un taux d'évolution accéléré (Galtier et al. 2009, Berglund et al. 2009, Ratnakumar et al. 2010, Kostka et al. 2012). Chez les oiseaux, ainsi que pour 17 gènes de 73 espèces placentaires, la gBGC semble entraîner au contraire une baisse du ratio dN/dS (Lartillot 2012; Bolívar et al. 2016). Ainsi, comme la méthode DFE-α est basée sur la comparaison des sites synonymes et non-synonymes au niveau de la divergence et du polymorphisme, la gBGC constitue potentiellement un facteur de biais important à prendre en compte.

Ceci est d'autant plus vrai que ce mécanisme semble être très répandu au sein des métazoaires, avec quelques remarquables exceptions (la moule *Mytilus galloprovincialis* ou le tunicier *Ciona robusta* (Galtier et al. 2018)). Il a cependant été expérimentalement démontré chez plusieurs espèces comme les levures (Mancera et al. 2008, Lesecque et al. 2013), les humains (Williams et al. 2015, Halldorsson et al. 2016), les gobe-mouches (Smeds et al. 2016) ou encore les daphnies (Keith et al. 2016), et de nombreuses indications suggèrent son existence dans d'autres métazoaires et plantes (Eyre-Walker 1999, Montoya-Burgos et al. 2003, Meunier and Duret 2004, Webster and Smith 2004, Spencer 2006, Webster et al. 2006, Escobar et al. 2011, Muyle et al. 2011, Kent et al. 2012, Pessia et al. 2012, Wallberg et al. 2015, Galtier et al. 2018, Long et al. 2018, Smith et al. 2018), ainsi que chez de nombreuses bactéries (Lassalle et al. 2015).

La gBGC est un facteur de biais d'autant plus difficile à prendre en compte qu'elle a potentiellement des effets variables entre les espèce où elle est présente. D'une part, l'intensité de la gBGC à travers le génome dépend du paysage de recombinaison, dont la dynamique varie selon les taxons. Par exemple, la durée de vie et le taux de renouvellement des points-chauds de recombinaison sont différents entre la plupart des mammifères qui possèdent le gène PRDM9 et les oiseaux qui ne l'ont pas (Mugal et al. 2013, Lesecque et al. 2014, Baker et al. 2015, Singhal et al. 2015, Latrille et al. 2017). Toutes les espèces ne sont donc pas à l'équilibre en terme de contenu en GC, et sont en outre à différentes distances de cet équilibre. D'autre part, on s'attend à ce que l'intensité de la gBGC, tout comme l'intensité de la sélection, dépende de la taille efficace des populations (Nagylaki 1983), qui elle même est variable entre espèces. Ces considérations et l'influence de la gBGC sur l'estimation des statistiques clés de la méthode DFE-α font l'objet du **Chapitre 2** de cette thèse.

III. Quels sont les déterminants du taux de substitution adaptatif et de la distribution des effets en fitness des mutations ?

1. Que peut-on apprendre de la génomique comparative ?

a. Le résultat de l'approche de Mc Donad & Kreitman et ses dérivés

Pour comprendre les facteurs qui déterminent la proportion de mutations adaptatives *vs.* non adaptatives en ségrégation ou fixées dans les génomes, un des meilleurs outils semble être la génomique comparative : il s'agit de l'étude comparative de la structure et fonction des génomes de différentes espèces. Les méthodes issues du test de McDonald & Kreitman ont été appliquées à un

grand nombre d'espèces, d'autant plus que le développement de ces méthodes s'est accompagné d'un accroissement très rapide des données de séquences disponibles, du fait des progrès constants des techniques de séquençage associés à la baisse des coûts, ainsi que des outils bio-informatiques pour les analyser. C'est cet outil d'analyse puissant qui a permis de mettre en évidence un important résultat empirique en évolution moléculaire : il existe, au moins chez les vertébrés, une relation positive entre le ratio dN/dS et la masse des espèces et leur longévité (Nikolaev et al. 2007; Popadin et al. 2007; Lartillot and Delsuc 2012; Romiguier et al. 2014; Figuet et al. 2016). Ces deux traits d'histoire de vie sont utilisés comme des marqueurs de la taille efficace des populations. Les espèces grandes et longévives ont en effet tendance à avoir des tailles de population plus petites que les petites espèces à durée de vie plus courte. Ce résultat confirme l'attendu théorique selon lequel il y a une corrélation négative entre le taux d'évolution protéique et la taille efficace si les nouvelles mutations non-synonymes sont en grande partie faiblement délétères, et ainsi, corrobore la théorie quasi neutre de l'évolution d'Ohta et Kimura (Kimura and Ohta 1971, Ohta 1973).

En ce qui concerne l'estimation de α, si les premiers résultats se sont cantonnés aux espèces modèles chez les animaux (genre *Drosophila* et espèce humaine (Fay et al. 2001, Clark et al. 2003, Bierne and Eyre-Walker 2004, Waterson et al. 2005, Bustamante et al. 2005, Welch 2006), on dispose aujourd'hui d'un panel beaucoup plus large de résultats parmi les métazoaires (Liti et al. 2009; Halligan et al. 2010; Carneiro et al. 2012; Tsagkogeorga et al. 2012; Loire et al. 2013; Galtier 2016), et même les plantes, en plus limité ((Barrier et al. 2003, Gossmann et al. 2010). Une large partie des résultats de α obtenus jusqu'à présent est recueillie dans le **Tableau 2**. Les premières estimations de α vont de 41 à 94 % pour le genre *Drosophila*, alors qu'elle vont de de 0 à un maximum de 35 % chez l'humain. Plus tard, ces résultats ont été confirmés par des méthodes plus sophistiquées, indiquant un α proche de zéro chez les grands singes (Boyko et al. 2008, Hvilsom et al. 2012), et autour de 50 % chez *Drosophila*.

Espèce focale	Groupe externe	Méthode	α	Référence
Homo sapiens	Old world monkeys	MK	0,35	Fay et al. 2001
Arabidopsis thaliana	*Arabidopsis lyrata*	MK	0	Bustamante et al. 2002
Drosophila simulans	*Drosophila yakuba*	MK	0,45	Smith & Eyre-Walker 2002
Drosophila simulans	*Drosophila melanogaster*	MK	0,94	Sawer et al. 2003
Drosophila simulans	*Drosophila melanogaster*	MK	0,43	Bierne & Eyre-Walker 2004
Drosophila melanogaster	*Drosophila simulans*	MK	0,45	Bierne & Eyre-Walker 2004
Homo sapiens	Mouse	MK	0	Zhang & Li 2005
Homo sapiens	Old world monkeys	MK	0	Zhang & Li 2005
Homo sapiens	*Pan troglodyres*	MK	0-0,09	Waterson et al. 2005
Homo sapiens	*Pan troglodyres*	MK	0,2	Zhang & Li 2005
Homo sapiens	*Pan troglodyres*	MK	0,06	Bustamante et al. 2005
Drosophila simulans	*Drosophila yakuba*	MK	0,41	Welch 2006
Escherichia coli	*Salmonella enterica*	MK	0,56	Charlesworth & Eyre-Walker 2006
Salmonella enterica	*Escherichia coli*	MK	0,34	Charlesworth & Eyre-Walker 2006
Zea mays	*Sorghum bicolora*	Méthode DFE-α	-0,25	Gossman et al. 2010
Boechera stricta	*Arabidopsis thaliana*	Méthode DFE-α	-1,75 ; -0,92	Gossman et al. 2010
Oryza sativa	*Sorghum bicolora*	Méthode DFE-α	-1,81 ; -0,16	Gossman et al. 2010
Arabidopsis lyrata	*Arabidopsis thaliana*	Méthode DFE-α	-0,65 ; -0,17	Gossman et al. 2010
Populus balsamifera	*Populus trichocarpa*	Méthode DFE-α	-0,03 ; 0,02	Gossman et al. 2010
Schiedea globosa	*Schiedea adamantis*	Méthode DFE-α	-0,05 ; 0,34	Gossman et al. 2010
Mus musculus castaneus	*Mus famulus*	Méthode DFE-α	0,57	Halligan et al. 2010
Mus musculus castaneus	*Rattus rattus*	Méthode DFE-α	0,44	Halligan et al. 2010
Ciona intestinalis B	Ciona intestinalis A	MK	0,54	Tsagkogeorga et al. 2012
Oryctolagus cuniculus	*Lepus granatensis*	MK	~0,4	Carneiro et al. 2012
Oryctolagus cuniculus	*Lepus granatensis*	Méthode DFE-α	~0,65	Carneiro et al. 2012
Chelonoidis nigra	*Chelonoidis carbonaria*	MK	<0	Loire et al. 2013
Chelonoidis nigra	*Chelonoidis carbonaria*	Méthode DFE-α	0,13	Loire et al. 2013
Rattus norvegicus	*Rattus rattus*	Méthode DFE-α	<0	Deinum et al. 2015
Abatus cordatus	*Abatus agassizi*	Méthode DFE-α	0,28	Galtier 2016
Allolobophora chlorotica	*Aporrectodea icterica*	Méthode DFE-α	0,73	Galtier 2016
Aptenodytes patagonicus	*Aptenodytes forsteri*	Méthode DFE-α	0,93	Galtier 2016
Armadillidium vulgare	*Armadillidium nasatum*	Méthode DFE-α	0,73	Galtier 2016
Artemia franciscana	*Artemia sinica*	Méthode DFE-α	0,48	Galtier 2016
Caenorhabditis brenneri	*Caenorhabditis sp.10*	Méthode DFE-α	0,83	Galtier 2016
Camponotus ligniperdus	*Camponotus aethiops*	Méthode DFE-α	0,32	Galtier 2016
Chelonoidis nigra	*Chelonoidis carbonaria*	Méthode DFE-α	0,56	Galtier 2016
Chlorocebus aethiops	*Macaca mulatta*	Méthode DFE-α	0,43	Galtier 2016
Ciona intestinalis A	*Ciona intestinalis B*	Méthode DFE-α	0,5	Galtier 2016
Ciona intestinalis B	*Ciona intestinalis A*	Méthode DFE-α	0,76	Galtier 2016
Crepidula fornicata	*Crepidula plana*	Méthode DFE-α	0,73	Galtier 2016

Culex pipiens	*Culex torrentium*	Méthode DFE-α	0,74	Galtier 2016
Echinocardium mediterraneum	*Echinocardium cordatum B2*	Méthode DFE-α	0,66	Galtier 2016
Emys orbicularis	*Trachemys scripta*	Méthode DFE-α	0,69	Galtier 2016
Eudyptes moseleyi	*Pygoscelis papua*	Méthode DFE-α	0,86	Galtier 2016
Eulemur coronatus	*Eulemur mongoz*	Méthode DFE-α	0,3	Galtier 2016
Eulemur mongoz	*Eulemur coronatus*	Méthode DFE-α	0,39	Galtier 2016
Eunicella cavolinii	*Eunicella verrucosa*	Méthode DFE-α	0,32	Galtier 2016
Galago senegalensis	*Nycticebus coucang*	Méthode DFE-α	0,29	Galtier 2016
Halictus scabiosae	*Halictus simplex*	Méthode DFE-α	0,68	Galtier 2016
Hippocampus guttulatus	*Hippocampus kuda*	Méthode DFE-α	0,23	Galtier 2016
Homo sapiens	*Pan troglodytes*	Méthode DFE-α	0,24	Galtier 2016
Lepus granatensis	*Lepus americanus*	Méthode DFE-α	0,77	Galtier 2016
Lineus longissimus	*Lineus ruber*	Méthode DFE-α	0,73	Galtier 2016
Macaca mulatta	*Chlorocebus aethiops*	Méthode DFE-α	0,24	Galtier 2016
Melitaea cinxia	*Melitaea didyma*	Méthode DFE-α	0,67	Galtier 2016
Messor barbarus	*Messor structor*	Méthode DFE-α	0,69	Galtier 2016
Microtus arvalis	*Microtus glareolus*	Méthode DFE-α	0,6	Galtier 2016
Mytilus galloprovincialis	*Mytilus californianus*	Méthode DFE-α	0,7	Galtier 2016
Necora puber	*Carcinus aestuarii*	Méthode DFE-α	0,64	Galtier 2016
Nycticebus coucang	*Galago senegalensis*	Méthode DFE-α	0,18	Galtier 2016
Ophioderma longicauda L1	*Ophioderma longicauda L3*	Méthode DFE-α	0,73	Galtier 2016
Ostrea edulis	*Ostrea chilensis*	Méthode DFE-α	0,51	Galtier 2016
Pan troglodytes	*Homo sapiens*	Méthode DFE-α	0,34	Galtier 2016
Parus caeruleus	*Parus major*	Méthode DFE-α	0,93	Galtier 2016
Physa acuta	*Physa gyrina*	Méthode DFE-α	0,62	Galtier 2016
Propithecus coquereli	*Varecia variegata variegata*	Méthode DFE-α	0,75	Galtier 2016
Reticulitermes grassei	*Reticulitermes flavipes*	Méthode DFE-α	0,51	Galtier 2016
Sepia officinalis	*Sepiella japonica*	Méthode DFE-α	0,62	Galtier 2016
Thymelicus lineola	*Thymelicus sylvestris*	Méthode DFE-α	0,68	Galtier 2016
Thymelicus sylvestris	*Thymelicus lineola*	Méthode DFE-α	0,66	Galtier 2016
Varecia variegata variegata	*Propithecus coquereli*	Méthode DFE-α	0,58	Galtier 2016
Parus major	*Taeniopygia guttata*	Méthode DFE-α	0,48	Corcoran et al. 2017
Taeniopygia guttata	*Parus major*	Méthode DFE-α	0,64	Corcoran et al. 2017
Ficedula albicollis	*Taeniopygia guttata/gallus gallus*	Méthode DFE-α	0,18	Bolivar et al. 2018

Tableau 2 : **Liste (non exhaustive) des estimations de α obtenues jusqu'à présent pour diverses espèces et différentes méthodes au sein des séquences codantes, dans l'ordre chronologique.**
« MK » signifie l'approche de Smith & Eyre-Walker (2002) et ses dérivés les plus basiques.

D’après les estimations obtenues chez d’autres espèces, il semble qu’on puisse les séparer en deux groupes distincts, un comprenant les espèces à α faible (<50%), et un les espèces à α fort (>50%). Le premier groupe contient les tortues des Galapagos (Loire et al. 2013), les levures (Liti et al. 2009), et neuf espèces de plantes (Gossmann et al. 2010), et le second contient la souris (Halligan et al. 2010), le rat (Deinum et al. 2015), le lapin (Carneiro et al. 2012), les ascidies (Tsagkogeorga et al. 2012), et enfin les entérobactéries (Charlesworth and Eyre-Walker 2006). Ces organismes sont très différents, notamment en termes de traits d’histoire de vie, et ainsi ces résultats ont le potentiel de nous aider dans la compréhension des facteurs qui mènent à un fort taux d’adaptation, ou au contraire à un faible taux d’adaptation. Notamment, le premier contraste entre les résultats humains *vs. Drosophila* a mené à l’hypothèse selon laquelle un des déterminants majeurs de α est la taille efficace N_e, α étant plus fort dans les espèces à forte N_e, et faible dans les espèces à faible N_e. Cette hypothèse a été quelque peu renforcée par l’observation d’un faible α chez les tortues des Galapagos qui évoluent à une faible taille de population, et un α plus fort chez les souris, lapins, ascidies et bactéries qui sont de manière générale des espèces considérées comme évoluant à plus forte N_e. Malgré tout, cette hypothèse mérite de plus amples considérations, à la fois théoriques et empiriques, ce qui fait l’objet de la partie suivante.

b. Un lien entre la taille efficace et le taux de substitution adaptatif ?

En théorie, il existe plusieurs raisons de penser qu’il existe effectivement un lien entre la taille efficace et le taux de substitution adaptatif (Eyre-Walker 2006, Gossmann et al. 2012, Lanfear et al. 2014). Premièrement, plus la taille de population est grande, plus la probabilité d’apparition d’une mutation avantageuse est forte. Deuxièmement, les mutations avantageuses *de novo* d'effet modéré ont une plus forte probabilité de fixation en forte taille de population, car les effets de la dérive sont amoindris et ainsi, la perte stochastique de la mutation alors qu’elle ségrège encore à faible fréquence est moins probable. D’autre part, si s est grand devant $1/N_e$, la probabilité de fixation d’un allèle de coefficient de sélection s peut être approximée à 2s, c’est-à-dire indépendant de la taille efficace (Haldane 1927, Fisher 1931, Wright 1931), et il y a en grande population une plus grande proportion de mutations qui ont un coefficient de sélection $s>>1/N_e$. Ces deux arguments prédisent donc un lien positif entre N_e et le taux de substitution adaptatif si l’adaptation est limitée par l’apport de nouvelles mutations adaptatives. On peut également penser que les grandes populations ont une plus grande diversité génétique, et ainsi ont une plus grande probabilité de posséder des allèles qui s’avéreront adaptatifs en cas de changement d’environnement. Ces considérations théoriques sont soutenues par une analyse comparative de 13 espèces eucaryotes où une corrélation positive a été établie entre ω_a et N_e (Gossmann et al. 2012). Une telle corrélation a

été également établie chez plusieurs espèces de tournesols de tailles efficaces variées (Strasburg et al. 2010).

Néanmoins, cette conclusion concernant l'existence d'une relation positive entre N_e et ω_a est encore débattue, aussi bien sur le plan théorique qu'empirique, par des résultats contradictoires. En ce qui concerne la théorie, l'existence d'une relation positive entre N_e et ω_a n'est valide que si l'apport de mutations adaptatives est limitant, ce qui est difficile à établir. Gillespie souligne dans son livre « The cause of molecular evolution » (1991) que le taux de substitution peut également être déterminé non pas par l'apport de nouvelles mutations adaptatives et la dérive génétique, mais pas le taux de variation de l'environnement (Gillespie 1991). Venant appuyer cette hypothèse, Lourenço et al. (2013) montrent via des simulations utilisant le formalisme du modèle géométrique de Fisher que le taux de variations environnementales et la complexité (correspondant au nombre de dimensions qui permettent de résumer tous les traits phénotypiques qui ont un effet sur la fitness) ont une influence plus forte sur le taux de substitution adaptatif que N_e (Lourenço et al. 2013). En ce qui concerne les résultats empiriques, certains organismes se sont révélés très tôt être des contre-exemples de l'hypothèse initiale, tels que la levure *Saccharomyces paradoxus*, espèce à forte N_e qui montre un α faible (Liti et al. 2009). De plus, plusieurs études n'ont pas pu mettre en évidence une corrélation entre ω_a et N_e : chez neuf espèces de plantes (Gossmann et al. 2010), et 44 paires d'animaux non-modèles (Galtier et al. 2016).

Ainsi, s'il est avéré qu'il existe un lien fort entre N_e et l'intensité de la sélection purificatrice, illustrée par la corrélation entre le dN/dS et certains traits d'histoire de vie liés à la taille efficace, l'existence d'un lien entre N_e et l'intensité de la sélection positive reste controversé.

2. Les déterminants de la distribution des effets en fitness des mutations

Comme illustré dans la partie précédente, un autre outil puissant pour déterminer les facteurs à l'origine des différences entre espèces est l'utilisation de modèles théoriques tels que le modèle géométrique de Fisher. En utilisant des données de polymorphisme de *Drosophila melanogaster* et d'*Homo sapiens*, Huber et al. (2017) montrent que le modèle s'ajustant le mieux à ces données est le modèle géométrique de Fisher, et que sous ce modèle, les facteurs qui déterminent le mieux la DFEM sont la complexité et la taille efficace de long-terme en lien avec la distance de la population par rapport à l'optimum de fitness (Huber et al. 2017), rejoignant les résultats de Lourenço et al. (2013). On peut donc faire plusieurs prédictions quant à la DFEM d'une espèce en fonction de ces trois facteurs : premièrement, plus une population est proche de son

optimum, plus une nouvelle mutations aura une forte probabilité d'être délétère. Deuxièmement, plus la taille de long-terme est faible, plus la proportion de mutations qui sont adaptatives est grande. En effet, une taille de population faible sur le long-terme implique la fixation de nombreuses mutations faiblement délétères par dérive, éloignant la population de son optimum, ce qui entraîne de plus nombreuses opportunités pour des mutations compensatoires adaptatives. Et enfin, plus l'organisme est complexe, plus une mutation a de chance de « casser » un mécanisme délicat, et donc d'être délétère (Huber et al. 2017).

Cette dernière notion de complexité, malgré son utilisation assez répandue, reste assez difficile à définir et à mesurer sur de réels organismes. Dans le cadre du modèle géométrique de Fisher, la complexité représente le nombre de traits sous sélection (Lourenço et al. 2011). En général, la notion de complexité est associée au nombre d'éléments phénotypiques indépendants qui caractérisent l'organisation d'un organisme à toutes les échelles. Mais comment mesurer ce paramètre empiriquement ? Au niveau moléculaire, le nombre de gènes peut sembler être un bon proxy de cette notion de complexité. En effet, il est courant de s'imaginer que les eucaryotes ont plus de gènes que les procaryotes, et au sein des eucaryotes, que les animaux ont plus de gènes que les plantes, et les vertébrés plus de gènes que les invertébrés. Néanmoins ces considérations ont été complètement réfutées par les données empiriques, qui montrent par exemple qu'il y a plus de gènes chez *Arabidopsis thaliana* et *Caenorhabditis elegans* que chez *Drosophila melanogaster* (Szathmáry et al. 2001). Il a également été suggéré que la complexité « biologique » pourrait être mieux représentée par le nombre total d' « états transcriptomiques » pouvant être générés à partir du génome d'un organisme, c'est-à-dire le set complet de transcrits (Claverie 2001). Enfin, certaines études suggèrent que le nombre d'interactions entre protéines pourrait être un bon proxy de la complexité (Stumpf et al. 2008). Malgré ces tentatives pour trouver un proxy approprié, cette notion reste floue.

Ces résultats suggèrent que le coefficient de sélection d'une même mutation peut varier au cours du temps si la complexité ou la taille de population de long-terme varie. Ils sont ainsi difficilement conciliables avec le postulat usuel que les gènes aux propriétés fonctionnelles les plus importantes sont les plus contraints, et donc qu'ils sont fortement conservés entre espèces (Huber et al. 2017).

IV. Adaptation des séquences codantes au niveau fonctionnel

1. La contrainte du niveau d'expression

Jusqu'alors, j'ai exposé des statistiques qui s'estiment à l'échelle d'un génome codant, ou au moins d'un groupe de plusieurs gènes. Mais les différents gènes d'un génome n'ont pas tous la même importance dans les processus cellulaires, et ne subissent pas tous les même contraintes sélectives. Un des éléments qui atteste de la variabilité de l'importance des gènes est leur patron d'expression. Comme ce patron d'expression est variable à la fois dans le temps (puisque certains gènes s'expriment différemment selon les stades de développement de l'organisme) et dans l'espace (certains gènes peuvent s'exprimer différemment selon le tissu cellulaire considéré), il peut se définir via deux paramètres, sa largeur (définie par le nombre de tissus ou de stades de développement dans lequel un gène est exprimé) et son niveau (quantité moyenne ou maximale d'ARNm produits). Pour obtenir ces deux mesures, on peut se baser sur la technique RNAseq (séquençage de l'ARN messager (ARNm)) de différents tissus, à différents stades de développement. Il s'agit du séquençage à haut débit des ADN complémentaires (ADNc) issus de la transcription inverse des ARNm présents dans l'échantillon, via des techniques telles qu'Illumina. La couverture des différents gènes obtenue à l'issu de ce séquençage informe de la quantité d'ARNm présents dans l'échantillon à la base, via des statistiques telles que le FPKM (Fragments Per Kilobase Million), RPKM (Reads Per Kilobase Million) ou encore TPM (Transcripts Per Kilobase Million). Ces mesures restent délicates, d'une part car les techniques peuvent introduire des biais (par exemple, l'étape de PCR nécessaire avant le séquençage en RNAseq introduit des « doublons » appelés « duplicats de PCR » qui peuvent mener à une surestimation du niveau d'expression). D'autre part, il n'est pas toujours facile de contrôler tous les facteurs environnementaux qui peuvent influer sur le niveau d'expression, comme l'âge, le sexe, voire l'alimentation ou le moment de la journée de l'échantillonnage.

Mais quel est le lien entre le niveau d'expression et la contrainte sélective qui s'applique sur un gène ? Il faut d'abord considérer que l'expression des gènes est un processus très régulé, par divers mécanismes, qui est lui même sous contrainte sélective. En effet, la production d'ARN et surtout de protéines représente un coût énergétique pour l'organisme. Cela explique l'existence des mécanismes tels que la sélection traductionnelle exposée dans la partie **II. 3. b.** qui vise à maximiser l'efficacité de la traduction et d'éviter la production de protéines non-fonctionnelles inutiles, voire toxiques pour la cellule. D'un autre côté, il a été montré que le niveau d'expression explique jusqu'à 30 % de la variabilité de la vitesse d'évolution entre les gènes au sein du génome

de la levure (Drummond et al. 2005), ce qui en fait le principal déterminant du taux d'évolution. Cette relation existe aussi chez les vertébrés (Hastings 1996) dont des mammifères (Duret and Mouchiroud 1999), et chez des bactéries (Rocha and Danchin 2004). Plusieurs hypothèses ont été avancées pour expliquer ce résultat. Les premières explications de l'existence de cette corrélation sont liées à la largeur du patron d'expression des gènes, en particulier le nombre de tissus différents dans lequel ils sont exprimés. D'une part, plus ce nombre de tissus ou types cellulaires est grand, plus les protéines issues de ces gènes sont exposées à des environnements différents, et entrent donc probablement en interaction avec des éléments cellulaires plus variés et ce, dans des conditions physico-chimiques diverses. Ainsi, elles ont probablement une plus forte densité de résidus sous contrainte sélective forte (Hastings 1996). D'autre part, il est considéré qu'une mutation qui affecte l'efficacité d'une protéine aura moins d'effet sur le phénotype si elle s'exprime dans une zone limitée de l'organisme que si elle est exprimée dans un grand nombre de tissus. Ainsi, la sélection purificatrice sera plus forte sur les mutations au sein des gènes largement exprimés car elles auront un effet plus fort sur la valeur sélective d'un individu. Sans exclure ces deux hypothèses, des indications empiriques semblent suggérer que c'est le niveau de transcription des gènes en lui même et non pas l'abondance des protéines ou la largeur du spectre d'expression qui est plus fortement corrélé au ratio dN/dS. Pour expliquer cela une troisième hypothèse a été avancée, appelée « robustesse traductionnelle » (Drummond et al. 2005). Cette hypothèse stipule que les séquences sont sous contraintes fortes pour éviter au maximum les erreurs générées lors de l'étape de traduction. En effet, les erreurs d'incorporation des ribosomes sont de l'ordre de 5 sur 10000. Ces erreurs, en plus de produire des protéines non-fonctionnelles inutilement coûteuses, peuvent altérer le repliement des protéines qui peuvent devenir toxiques pour la cellule, car elles dévoilent des résidus normalement enfouis qui peuvent entrer en interaction avec d'autres éléments cellulaires et déstabiliser toute la cellule. Le coût potentiel pour la cellule est d'autant plus élevé qu'un gène est fortement exprimé. Ainsi les gènes fortement exprimés sont sous contrainte forte pour assurer un repliement correct malgré les fréquentes erreurs de traduction (Drummond et al. 2005). Alternativement, Gout et al. (2010) proposent que le niveau d'expression d'un gène est le résultat d'un compromis entre le coût énergétique de l'expression et le bénéfice qu'apporte l'expression du gène, qui s'accroissent tous les deux en fonction de l'abondance de la protéine. Ainsi, la corrélation entre niveau d'expression et taux d'évolution des gènes résulte de la sélection contre les mutations qui affectent la fonction de la protéine, et non pas sa toxicité comme dans les hypothèses précédentes, puisque le coût énergétique reste le même alors que le bénéfice qu'apporte l'expression du gène diminue (Gout et al. 2010; Zhang and Yang 2015).

Ainsi, il semble que le niveau d'expression puisse constituer un déterminant très important de la vitesse d'évolution des gènes, plus important même que le rôle fonctionnel des protéines pour lesquelles ils codent. Cela en fait un facteur confondant important à prendre en compte lorsque l'on compare des statistiques telles que le ratio dN/dS ou les taux de substitution adaptatif ou non-adaptatif entre gènes.

2. Les contraintes sélectives varient au sein d'un gène

La variabilité des contraintes sélectives existe également à une échelle encore plus petite, au sein même des gènes. En effet, différentes contraintes s'appliquent d'une part à des positions qui correspondent à des zones directement impliquées dans la fonction de la protéine, appelées sites actifs, d'autre part aux zones qui permettent aux protéines de se replier correctement pour que les bons résidus soient enfouis ou au contraire exposés en surface de la molécule, et enfin, aux zones de surface de la protéine, qui n'ont pas de rôle fort dans la fonction ou l'élaboration de la structure tertiaire de la protéine, mais sont à même de gagner/perdre des fonctions de liaison à d'autres molécules (Echave et al. 2016). Ainsi, la structure de la protéine joue un rôle important dans le taux d'évolution de la séquence d'ADN, puisque le taux de substitution à une position particulière d'un gène dépend de la contrainte sélective qui opère à la position correspondante de la protéine liée à ce gène. On peut parler de deux types de contraintes différentes, la contrainte structurelle, et la contrainte fonctionnelle (Echave et al. 2016). Ces réflexions théoriques sur la variabilité de la vitesse d'évolution au sein d'un gène ont été corroborées par des observations empiriques, facilitées par le développement des modèles d'estimation du ratio dN/dS par site par maximum de vraisemblance mentionnés dans la partie **I. 3. a.** Par exemple, chez les protéines globulaires, les zones exposées en surface évoluent plus vite que les zones enfouies, car ces dernières contiennent plus de résidus en étroite interaction entre eux, et un seul changement déstabiliserait donc l'ensemble (Goldman et al. 1998). Le contraste entre les zones enfouies et de surface des protéines peut s'expliquer par plusieurs facteurs que l'on peut mesurer sur différents résidus des protéines :

-l'accès au solvant : il s'agit de la propriété la plus visiblement liée à la position en surface ou enfouie des résidus, les résidus externes étant en contact avec l'environnement cellulaire, notamment l'eau, alors que les résidus internes sont souvent hydrophobes, et plus conservés.

-la « densité » : elle mesure à quel point un résidu entre en contact avec peu ou beaucoup d'autres résidus de la protéine. Plus la structure tertiaire de la protéine est compacte, plus un résidu sera en contact avec beaucoup d'autres et ainsi sera plus contraint en termes de propriétés physico-chimiques pour assurer des liaisons avec des résidus variés.

-la flexibilité : les protéines ne sont pas statiques, mais sont souvent soumises à des fluctuations conformationnelles, notamment pour exposer leur site actif avant qu'un substrat puisse y être ligué. Un site d'une région très flexible est probablement plus tolérant aux mutations qu'un site d'une région moins flexible d'une protéine.

Ces considérations s'appuient sur l'hypothèse que les taux d'évolution dépendent des contraintes subies par les différents sites des protéines en fonction de leur importance fonctionnelle et structurale, et ainsi, que ces taux sont constants dans le temps. Néanmoins, nous avons vu dans la partie **III. 2.** que l'étude des déterminants de la DFEM suggère plutôt que les effets en fitness des mutations, qui influencent le taux d'évolution, peuvent varier au cours du temps au gré des changements de taille efficace. Ainsi, les taux d'évolution par site varient au cours du temps, un phénomène appelé l'hétérotachie (Lopez et al. 2002), ajoutant un niveau de complexité supplémentaire à l'étude des taux d'évolution par site.

3. Les contraintes sélectives varient selon les types de mutations

Depuis le début de cette introduction, j'ai différencié les mutations synonymes et non-synonymes. Dans les organismes étudiés dans cette thèse, c'est-à-dire les métazoaires, les mutations synonymes sont considérées comme neutres, ou à effets sélectifs faibles en cas de sélection sur l'usage du code, phénomène mentionné dans la partie **II. 3. b.**, alors que les mutations non-synonymes sont plus souvent soumises à sélection. Cela semble néanmoins beaucoup moins vrai chez d'autres eucaryotes non-métazoaires, chez les bactéries ou encore les virus. Ces organismes ont en général une taille de population beaucoup plus élevée, et ainsi les mutations synonymes qui optimisent ou au contraire détériorent l'efficacité de la traduction ou la stabilité des ARN messagers sont plus efficacement sélectionnées ou contre-sélectionnées. Certaines études ont même pu montrer de forts effets délétères, voire létaux, de mutations synonymes chez les virus (Cuevas et al. 2012). Les mutations non-synonymes, au contraire, peuvent être neutres, délétères ou faiblement délétères, ou avantageuses chez tous les organismes. Au sein de ces mutations, il est possible de différencier deux types de mutations qui sont potentiellement soumises à des niveaux de contrainte différents. Il s'agit des mutations appelées conservatives et radicales. Les mutations conservatives provoquent un changement d'un acide aminé vers un autre acide aminé qui possède les mêmes propriétés physico-chimiques, telles le volume, la polarité et la charge. Au contraire, une mutation non-synonyme est considérée comme radicale si elle provoque un changement d'une ou plusieurs de ces propriétés physico-chimiques (**Figure 5**).

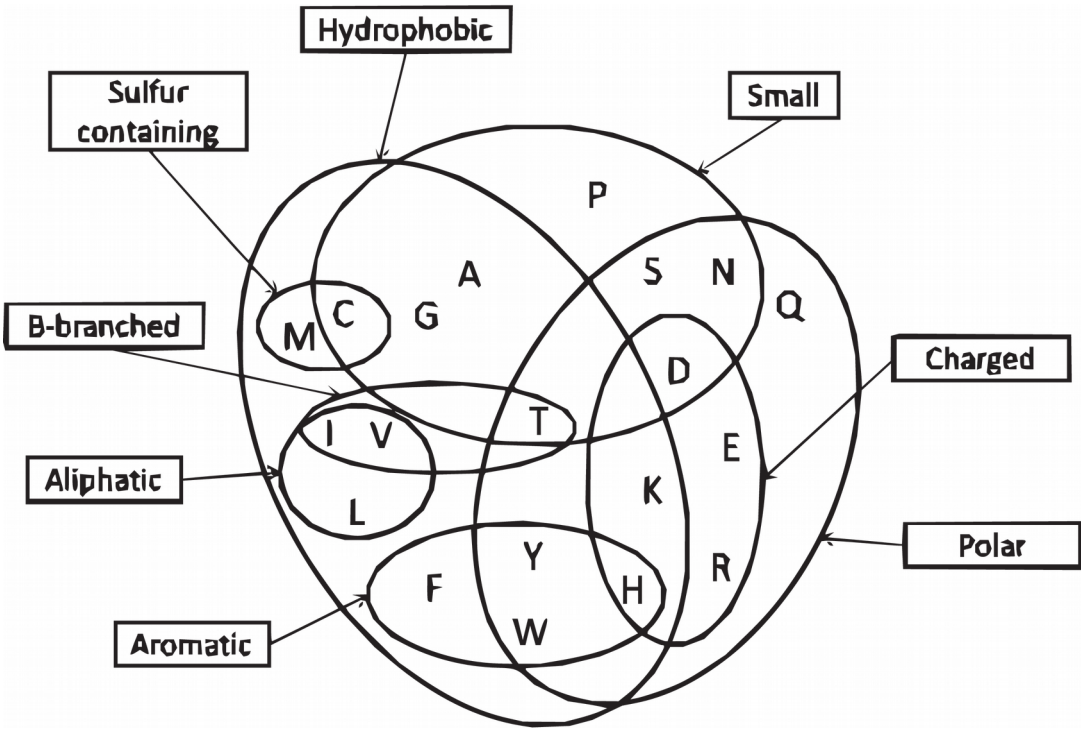


Figure 5 : Schéma de la classification des différents acides aminés en fonction de leurs propriétés physico-chimiques souvent utilisée pour la définition des substitutions radicales et conservatives.

Cette classification fut utilisée pour éviter le problème de saturation des substitutions synonymes lors de comparaison de ratio dN/dS entre des orthologues d'espèces très éloignées (Gojobori 1983; Smith and Smith 1996). L'utilisation du ratio Kr/Kc, où Kr représente le taux de substitution radical et Kc le taux de substitution conservatif, se base sur l'hypothèse qu'il devrait se comporter de la même façon que le ratio dN/dS. En effet, on s'attend à ce que les effets en fitness des mutations radicales soient plus forts que les effets en fitness des mutations conservatives, puisqu'un changement physico-chimique a une plus grande chance d'altérer, en bien ou en mal, la fonction de la protéine, tout comme un changement non-synonyme a plus de chance d'altérer la protéine qu'un changement synonyme. Ainsi pour les deux ratios, le numérateur représente une classe de mutation supposée sous plus forte sélection positive et négative que le dénominateur. Tout comme le dN/dS, le ratio Kr/Kc a donc été utilisé afin de détecter les gènes sous sélection positive (Hughes et al. 1990; Hughes 1992; Rand et al. 2000), un ratio significativement élevé étant le signe d'un signal de sélection positive. Le ratio Kr/Kc présente l'avantage, face au ratio dN/dS, que la différence de pression de sélection purificatrice entre les mutations radicales et conservatives est

théoriquement plus faible qu'entre les mutations non-synonymes et synonymes. Ainsi, la proportion de substitutions adaptatives nécessaire à engendrer un ratio plus grand que un est plus faible pour le Kr/Kc que pour le dN/dS, ce qui fait du ratio Kr/Kc une statistique moins conservative. Estimer correctement ce ratio nécessite néanmoins de prendre des précautions : d'une part, il est primordial de contrôler pour le nombre d'opportunités de mutations radicales et conservatives, les mutations *de novo* conservatives étant plus fréquentes que les radicales (Goldberg and Wittes 1966; Epstein 1967). D'autre part, Dagan et al. (2002) montrèrent que les estimations du ratio Kr/Kc sont fortement influencées par le ratio de transition/transversion, souvent appelé κ. En effet, négliger ce paramètre mène à une sous-estimation à la fois du taux de substitution radical et conservatif (Dagan et al. 2002), car les transversions entraînent en général des changements qui altèrent plus les propriétés physico-chimiques des acides aminés, c'est-à-dire qui sont plus souvent radicales, et sont moins fréquentes que les transitions (Zhang 2000).
Un autre problème est de choisir la bonne classification de changements radicaux et conservatifs, c'est-à-dire de savoir quelles propriétés physico-chimiques affectent le plus les fonctions des protéines.

Plusieurs études ont pu vérifier que le ratio Kr/Kc est corrélé au ratio dN/dS (Zhang 2000). Néanmoins, Figuet et al. (2014) montrent que les ratios Kr/dS et Kc/dS répondent comme le dN/dS aux traits d'histoire de vie liés à la taille efficace (Figuet et al. 2014), ce qui indique que les deux types de mutations contiennent des mutations faiblement délétères, ce qui peut compliquer l'interprétation du ratio Kr/Kc. Malgré tout, le Kr/Kc lui-même semble également être lié aux traits d'histoire de vie, puisqu'il a été estimé plus faible chez les rongeurs que chez les primates (Zhang 2000), et corrélé à la masse corporelle chez les oiseaux (Weber et al. 2014, Figuet et al. 2014). Ces deux types de mutations ont également été utilisées dans le cadre du test de McDonald & Kreitman, par Smith (2003) qui estima le taux de substitution adaptatif et non-adaptatif des mutations radicales et conservatives chez la drosophile (Smith 2003). Ses résultats indiquent que les mutations radicales sont soumises à une sélection purificatrice plus forte que les conservatives, mais pas à une sélection positive significativement différente des conservatives. Néanmoins, ces résultats ont pu être fortement influencés par la présence de mutations faiblement délétères en ségrégation qui n'ont pas été prises en compte dans la méthode utilisée dans cette étude. Ainsi, l'étude des taux de substitutions et mutations radicales et conservatives présentent encore le potentiel d'améliorer notre compréhension sur la nature de la sélection sur la fonction des protéines.

V. Objectifs de la thèse

Ce travail de thèse se focalise sur les méthodes dérivées du test de McDonald & Kreitman qui visent à estimer le taux de substitution adaptatif, et plus particulièrement sur les derniers développements maintenant regroupés sous le terme « DFE-α » (Tataru et al. 2017). Ce thème central est étudié de deux manières différentes qui illustrent bien la dualité du domaine de l'évolution moléculaire. Le premier objectif est d'étudier l'influence de sources de biais qui subsistent dans ce test malgré les améliorations successives, et de tenter de les prendre en compte. Une fois ces facteurs de biais pris en compte, le second objectif est d'utiliser ce test sur un grand jeu de données d'espèces métazoaires afin d'identifier les déterminants du taux de substitution adaptatif. Ainsi, le but de cette thèse est à la fois de comprendre comment certains mécanismes influencent les statistiques résumées que nous utilisons, mais aussi d'estimer ces statistiques sur des données moléculaires pour comprendre comment, au niveau moléculaire, et en quelle proportion les espèces s'adaptent à leur environnement.

Le premier objectif se décompose en deux questions :

(i) Comment la méthode DFE-α est-elle influencée par les fluctuations de taille de population dans le passé lointain ? Nous avons expliqué dans la partie **II. 3. a.** que cette méthode fait l'hypothèse que le régime de sélection est resté constant pendant la période étudiée, c'est-à-dire depuis la séparation entre l'espèce focale, dont on a les données de polymorphisme, et l'espèce utilisée comme groupe externe pour estimer la divergence. Or, des variations démographiques de long-terme peuvent provoquer un changement de régime de sélection entre la période durant laquelle les substitutions se sont accumulées dans la divergence et la période plus récente durant laquelle le polymorphisme se construit. Deux résultats empiriques en particulier ont motivé cette question. D'une part, une étude révèle que l'application de la méthode de Eyre-Walker & Keightley (2009) sur douze exomes du rat brun *Rattus norvegicus* génère une valeur de α négatif (Deinum et al. 2015). Les auteurs expliquent ce résultat en invoquant le fort goulot d'étranglement qui a eu lieu il y a approximativement 20 000 ans au sein de cette espèce. D'autre part, dans un projet récent, mon directeur de thèse Nicolas Galtier a observé en utilisant des paires d'espèces, que selon l'espèce choisie comme groupe externe pour estimer la divergence et l'espèce focale utilisée, on peut estimer des α différents (Galtier et al. 2016). Nous avons tenté de répondre à cette question dans le **Chapitre 1**, et avons développé une méthode et un jeu de donnée structuré de manière à pouvoir circonvenir ce biais présentés dans le **Chapitre 3**.

(ii) Comment la méthode DFE-α est-elle influencée par la gBGC ? Et plus particulièrement, peut-on vérifier l'hypothèse selon laquelle la gBGC, en mimant la sélection positive, peut engendrer une

surestimation du taux de substitution adaptatif ? Nous avons choisi deux taxons pour répondre à cette question, les primates et les oiseaux, du fait de l'apparent contraste de l'effet de la gBGC sur le ratio dN/dS entre ces deux taxons rapporté dans la littérature au moment du début de mon projet (Berglund et al. 2009; Galtier et al. 2009; Ratnakumar et al. 2010; Kostka et al. 2012; Bolívar et al. 2016). Nous avons tenté de répondre à cette question dans le **Chapitre 2**.

Le second objectif de cette thèse se décompose lui aussi en deux sous-questions :

(i) Peut-on trouver un lien entre la taille efficace et le taux de substitution adaptatif en prenant en compte de possibles fluctuations démographiques de long-terme, ainsi que les effets de la gBGC ? Cette question a été motivée par la mise en évidence d'une absence de corrélation significative entre un proxy de N_e et le taux de substitution adaptatif sur un large jeu de données de métazoaires (Galtier 2016), en désaccord avec les attendus théoriques qui prédisent une corrélation positive. Pour répondre à cette question, nous avons élaboré un jeu de données de neuf taxons de métazoaires (dont cinq nouvellement produits au laboratoire via la technique de capture) dans lequel nous disposons de données pour plus de cinq individus dans plusieurs espèces du taxon (quatre à six espèces). Les résultats sont présentés dans le **Chapitre 3**.

(ii) Quelle est la nature des changements d'acides aminés qui mènent à de l'adaptation ? Via l'utilisation de la méthode DFE-α sur des changements non-synonymes uniquement radicaux ou conservatifs, nous avons évalué la contribution de ces deux types de changements à l'évolution adaptative et non-adaptative des protéines chez les animaux. Si l'un des deux types de changements est plus particulièrement ciblé par la sélection positive, alors le taux de substitution adaptatif mesuré en utilisant seulement ce type de changement non-synonyme pourrait se révéler une statistique en mesure de mieux répondre à des variations inter-espèces de taille efficace, rejoignant ainsi la sous-question précédente. En ce qui concerne ces questions, nous ne disposons que de résultats préliminaires, qui sont présentés dans le **Chapitre 4** et accompagnés des perspectives pour la suite de ce projet.

VI. English version of the thesis objectives:

This work focuses on the methods derived from the McDonald & Kreitman test which aims at estimating the adaptive substitution rate. In particular, we address the last methods called « DFE-α » (Tataru et al. 2017) via two different angles to fulfill two objectives. The first objective is to study the influence of some sources of bias that remain in the test despite the numerous improvements since the initial McDonald & Kreitman approach. The second objective is to use this test while accounting for those source of bias on a wide range of metazoan species in order to uncover the determinants of the adaptive substitution rate. Thus we try in this work to understand

how some genomic mechanisms influence the summary statistics that we use, and in the mean time we try o estimate those summary statistics to determine how and in what proportion species adapt to their environment at the molecular level.

The first mentioned goal is composed of two distinct questions :

(i) How is the DFE-α method influenced by past fluctuations in population size ? As mentioned in this introduction, the DFE-α method makes the assumption that the selection regime has remained constant over the period under study, i.e. since the divergence between the focal species and the outgroup species. However, demographic fluctuations can lead to a change in the selective regime between the period during which the divergence builds up, and the period during which the polymorphism builds up. In particular, two empirical results led us to address this question : on the one hand, a study focusing on the brown rat *Rattus norvegicus* showed that the method as presented in Eyre-Walker and Keightley 2009 yields a negative value of α (Deinum et al. 2015), which is interpreted as the result of a bottleneck that happened approximately 20,000 yeas ago. On the other hand, my supervisor Nicolas Galtier observed that in a species pair, depending on which species is used as an outgourp and which is the focal one, the DFE-α method yields different results (Galtier et al. 2016). In view of these results, we try to analyze how the method is influenced by plausible demographic scenarios of long-term fluctuations in population size in **Chapter 1** of this manuscript. We also developed a method and a data-set structured so as to circumvent this issue, presented in **Chapter 3**.

(ii) How is the DFE-α method influenced by gBGC ? In particular, we ask if we can validate that gBGC, by mimicking positive selection, can lead to the over-estimation of the adaptive substitution rate. We chose two taxa to answer this question, primates (as representative of mammals) and birds. This choice was motivated by the contrast between the two taxa in terms of the influence of gBGC in the dN/dS ratio reported in the literature (Berglund et al. 2009; Galtier et al. 2009; Ratnakumar et al. 2010; Kostka et al. 2012; Bolívar et al. 2016). We attempt to address this question in **Chapter 2** of this thesis.

The second objective of this thesis is also composed of two distinct questions :

(i) Can we detect a relationship between the effective population size and the adaptive substitution rate if we take into account possible long-term fluctuations in population size and gBGC ? This question was motivated by the absence of positive correlation between π_s and the adaptive substitution rate in a wide range of metazoan brought out in Galtier (2016), which is contrary to theoretical expectations that predict a positive correlation. To answer this question, we built a dataset of nine metazoan taxa (including five newly produced in our lab by exon capture), in which

we have coding sequence data of at least five individuals in four to six species per taxa. Results are presented in **Chapter 3**.

(ii) What is the nature of amino-acid changes leading to adaptation ? We used the DFE-α method using only non-synonymous radical or conservative changes to estimate the contribution of those two types of changes to adaptative and non-adaptive protein evolution. This question may also be connected to the previous one, as if one of the two types of changes is particularly prone to produce adaptive changes, thus estimating the adaptive substitution rate using only this type of amino-acid substitution may lead to a statistic that could correlate better to N_e. Even if we have not fully address this last question yet, we have already performed some preliminary data analyses. Our first results allow a few perspectives to be proposed, as presented in the **Chapter 4**.

VII. Références

Akashi H. 1994. Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. Genetics 136:927–935.

Andolfatto P. 2005. Adaptive evolution of non-coding DNA in *Drosophila*. Nature 437:1149.

Baker CL, Kajita S, Walker M, Saxl RL, Raghupathy N, Choi K, Petkov PM, Paigen K. 2015. PRDM9 drives evolutionary erosion of hotspots in *Mus musculus* through haplotype-specific initiation of meiotic recombination. PLoS genetics 11:e1004916.

Barrier M, Bustamante CD, Yu J, Purugganan MD. 2003. Selection on rapidly evolving proteins in the *Arabidopsis* genome. Genetics 163:723–733.

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. Journal of molecular evolution 55:356–363.

Berglund J, Pollard KS, Webster MT. 2009. Hotspots of biased nucleotide substitutions in human genes. PLoS biology 7:e1000026.

Bergman CM, Kreitman M. 2001. Analysis of conserved noncoding DNA in *Drosophila* reveals similar constraints in intergenic and intronic sequences. Genome Research 11:1335–1345.

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

Bierne N, Eyre-Walker A. 2004. The genomic rate of adaptive amino acid substitution in *Drosophila*. Molecular Biology and Evolution 21:1350–1360.

Bierne N, EYRE-WALKER A. 2006. Variation in synonymous codon use and DNA polymorphism within the *Drosophila* genome. Journal of evolutionary biology 19:1–11.

Bolívar P, Mugal CF, Nater A, Ellegren H. 2016. Recombination rate variation modulates gene sequence evolution mainly via GC-biased gene conversion, Not Hill–Robertson Interference, in an Avian System. Molecular Biology and Evolution 33:216–227.

Booker TR, Jackson BC, Keightley PD. 2017. Detecting positive selection in the genome. BMC biology 15:98.

Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, Lohmueller KE, Adams MD, Schmidt S, Sninsky JJ, Sunyaev SR. 2008. Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS genetics 4:e1000083.

Bustamante CD, Nielsen R, Sawyer SA, Olsen KM, Purugganan MD, Hartl DL. 2002. The cost of inbreeding in *Arabidopsis*. Nature. 416:531.

Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, Glanowski S, Tanenbaum DM, White TJ, Sninsky JJ, Hernandez RD. 2005. Natural selection on protein-coding genes in the human genome. Nature 437:1153.

Carneiro M, Albert FW, Melo-Ferreira J, Galtier N, Gayral P, Blanco-Aguiar JA, Villafuerte R, Nachman MW, Ferrand N. 2012. Evidence for widespread positive and purifying selection

across the European rabbit (*Oryctolagus cuniculus*) genome. Molecular Biology and Evolution 29:1837–1849.

Chamary J, Hurst LD. 2005. Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. Genome Biology 6:R75.

Charlesworth J, Eyre-Walker A. 2006. The rate of adaptive evolution in enteric bacteria. Molecular Biology and Evolution 23:1348–1356.

Charlesworth J, Eyre-Walker A. 2007. The other side of the nearly neutral theory, evidence of slightly advantageous back-mutations. Proceedings of the National Academy of Sciences 104:16992–16997.

Clark AG, Glanowski S, Nielsen R, Thomas PD, Kejariwal A, Todd MA, Tanenbaum DM, Civello D, Lu F, Murphy B. 2003. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. Science 302:1960–1963.

Claverie J-M. 2001. What if there are only 30,000 human genes? Science 291:1255–1257.

Consortium D 12 G. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. Nature 450:203.

Consortium RGSP. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature 428:493.

Cuevas JM, Domingo-Calap P, Sanjuán R. 2012. The fitness effects of synonymous mutations in DNA and RNA viruses. Molecular Biology and Evolution 29:17–20.

Dagan T, Talmor Y, Graur D. 2002. Ratios of radical to conservative amino acid replacement are affected by mutational and compositional factors and may not be indicative of positive darwinian selection. Molecular Biology and Evolution 19:1022–1025.

Deinum EE, Halligan DL, Ness RW, Zhang Y-H, Cong L, Zhang J-X, Keightley PD. 2015. Recent evolution in *Rattus norvegicus* is shaped by declining effective population size. Molecular Biology and Evolution 32:2547–2558.

Drummond DA, Raval A, Wilke CO. 2005. A single determinant dominates the rate of yeast protein evolution. Molecular biology and evolution 23:327–337.

Duret L. 2002. Evolution of synonymous codon usage in metazoans. Current opinion in genetics & development 12:640–649.

Duret L, Mouchiroud D. 1999. Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*. Proceedings of the National Academy of Sciences 96:4482–4487.

Echave J, Spielman SJ, Wilke CO. 2016. Causes of evolutionary rate variation among protein sites. Nature Reviews Genetics 17:109–121.

Epstein CJ. 1967. Non-randomness of ammo-acid changes in the evolution of homologous proteins. Nature 215:355.

Escobar JS, Glémin S, Galtier N. 2011. GC-biased gene conversion impacts ribosomal DNA evolution in vertebrates, angiosperms, and other eukaryotes. Molecular Biology and Evolution 28:2561–2575.

Eyre-Walker A. 1993. Recombination and mammalian genome evolution. Proceedings of the Royal Society of London 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. 2006. The genomic rate of adaptive evolution. Trends in Ecology & Evolution 21:569–575.

Eyre-Walker A, Keightley PD. 2007. The distribution of fitness effects of new mutations. Nature Reviews Genetics 8:610.

Eyre-Walker A, Keightley PD. 2009. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. Molecular Biology and Evolution 26:2097–2108.

Eyre-Walker A, Keightley PD, Smith NGC, Gaffney D. 2002. Quantifying the slightly deleterious mutation model of molecular evolution. Molecular Biology and Evolution 19:2142–2149.

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

Fay JC, Wu C-I. 2000. Hitchhiking under positive darwinian selection. Genetics 155:1405–1413.

Fay JC, Wyckoff GJ, Wu C-I. Positive and negative selection on the Human genome. :8.

Figuet E, Ballenghien M, Romiguier J, Galtier N. 2014. Biased gene conversion and GC-content evolution in the coding sequences of reptiles and vertebrates. Genome Biology and Evolution 7:240–250.

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

Fisher RA. 1931. XVII.—The distribution of gene ratios for rare mutations. Proceedings of the Royal Society of Edinburgh 50:204–219.

Ford MJ. 2001. Molecular evolution of transferrin: evidence for positive selection in salmonids. Molecular Biology and Evolution 18:639–647.

Fu Y-X, Li W-H. 1993. Statistical tests of neutrality of mutations. Genetics 133:693–709.

Galtier N. 2016. Adaptive Protein Evolution in animals and the effective population size hypothesis. PLOS Genetics 12:e1005774.

Galtier N, Duret L. 2007. Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. Trends in Genetics 23:273–277.

Galtier N, Duret L, Glémin S, Ranwez V. 2009. GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. Trends in Genetics 25:1–5.

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

Galtier N, Roux C, Rousselle M, Romiguier J, Figuet E, Glémin S, Bierne N, Duret L, Singh N. 2018. Codon usage bias in animals: disentangling the effects of natural selection, effective population size, and GC-biased gene conversion. Molecular Biology and Evolution 35:1092–1103.

Gillespie JH. 1994. The causes of molecular evolution. Oxford University Press

Glémin S, Arndt PF, Messer PW, Petrov D, Galtier N, Duret L. 2015. Quantification of GC-biased gene conversion in the human genome. Genome Research 25:1215–1228.

Gojobori T. 1983. Codon substitution in evolution and the" saturation" of synonymous changes. Genetics 105:1011–1027.

Goldberg AL, Wittes RE. 1966. Genetic code: aspects of organization. Science 153:420–424.

Goldman N, Thorne JL, Jones DT. 1998. Assessing the impact of secondary structure and solvent accessibility on protein evolution. Genetics 149:445–458.

Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Molecular Biology and Evolution 11:725–736.

Gossmann TI, Keightley PD, Eyre-Walker A. 2012. The effect of variation in the effective population size on the rate of adaptive molecular evolution in eukaryotes. Genome Biology and Evolution 4:658–667.

Gossmann TI, Song B-H, Windsor AJ, Mitchell-Olds T, Dixon CJ, Kapralov MV, Filatov DA, Eyre-Walker A. 2010. Genome wide analyses reveal little evidence for adaptive evolution in many plant species. Molecular Biology and Evolution 27:1822–1832.

Gout J-F, Kahn D, Duret L, Consortium PP-G. 2010. The relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution. PLoS Genetics 6:e1000944.

Haddrill PR, Charlesworth B, Halligan DL, Andolfatto P. 2005. Patterns of intron sequence evolution in *Drosophila* are dependent upon length and GC-content. Genome biology 6:R67.

Haldane JBS. 1927. A mathematical theory of natural and artificial selection, part V: selection and mutation. In: Mathematical Proceedings of the Cambridge Philosophical Society. Vol. 23. Cambridge University Press. p. 838–844.

Halldorsson BV, Hardarson MT, Kehr B, Styrkarsdottir U, Gylfason A, Thorleifsson G, Zink F, Jonasdottir Adalbjorg, Jonasdottir Aslaug, Sulem P. 2016. The rate of meiotic gene conversion varies by sex and age. Nature genetics 48:1377.

Halligan DL, Oliver F, Eyre-Walker A, Harr B, Keightley PD. 2010. Evidence for pervasive adaptive protein evolution in wild mice. PLoS Genetics 6:e1000825.

Hastings KE. 1996. Strong evolutionary conservation of broadly expressed protein isoforms in the troponin I gene family and other vertebrate gene families. Journal of Molecular Evolution 42:631–640.

Huber CD, Kim BY, Marsden CD, Lohmueller KE. 2017. Determining the factors driving selective effects of new nonsynonymous mutations. Proceedings of the National Academy of Sciences 114:4465–4470.

Hudson RR, Kreitman M, Aguadé M. 1987. A test of neutral molecular evolution based on nucleotide data. Genetics 116:153–159.

Hughes AL. 1992. Positive selection and interallelic recombination at the merozoite surface antigen-1 (MSA-1) locus of *Plasmodium falciparum*. Molecular Biology and Evolution 9:381–393.

Hughes AL, Nei M. 1988. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature 335:167.

Hughes AL, Ota T, Nei M. 1990. Positive Darwinian selection promotes charge profile diversity in the antigen-binding cleft of class I major-histocompatibility-complex molecules. Molecular Biology and Evolution 7:515–524.

Hvilsom C, Qian Y, Bataillon T, Li Y, Mailund T, Sallé B, Carlsen F, Li R, Zheng H, Jiang T. 2012. Extensive X-linked adaptive evolution in central chimpanzees. Proceedings of the National Academy of Sciences 109:2054–2059.

Ikemura T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. Molecular Biology and Evolution 2:13–34.

Johnson KP, Seger J. 2001. Elevated rates of nonsynonymous substitution in island birds. Molecular Biology and Evolution 18:874–881.

Keightley PD, Eyre-Walker A. 2007. Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. Genetics 177:2251–2261.

Keightley PD, Gaffney DJ. 2003. Functional constraints and frequency of deleterious mutations in noncoding DNA of rodents. Proceedings of the National Academy of Sciences 100:13402–13406.

Keith N, Tucker AE, Jackson CE, Sung W, Lledó JIL, Schrider DR, Schaack S, Dudycha JL, Ackerman M, Younge AJ. 2016. High mutational rates of large-scale duplication and deletion in *Daphnia pulex*. Genome research 26:60–69.

Kent CF, Minaei S, Harpur BA, Zayed A. 2012. Recombination is associated with the evolution of genome structure and worker behavior in honey bees. Proceedings of the National Academy of Sciences 109:18012–18017.

Kern AD, Hahn MW. 2018. The neutral theory in light of natural selection. Molecular Biology and Evolution 35:1366–1371.

Kimura M. 1983. The neutral theory of molecular evolution. Cambridge University Press

Kimura M, Ohta T. 1971. On the rate of molecular evolution. Journal of Molecular Evolution 1:1–17.

Kostka D, Hubisz MJ, Siepel A, Pollard KS. 2012. The role of GC-biased gene conversion in shaping the fastest evolving regions of the human genome. Molecular Biology and Evolution 29:1047–1057.

Kreitman M. 1996. The neutral theory is dead. Long live the neutral theory. Bioessays 18:678–683; discussion 683.

Lanfear R, Kokko H, Eyre-Walker A. 2014. Population size and the rate of evolution. Trends in Ecology & Evolution 29:33–41.

Lartillot N. 2012. Interaction between selection and biased gene conversion in mammalian protein-coding sequence evolution revealed by a phylogenetic covariance analysis. Molecular Biology and Evolution 30:356–368.

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

Lassalle F, Périan S, Bataillon T, Nesme X, Duret L, Daubin V. 2015. GC-content evolution in bacterial genomes: the biased gene conversion hypothesis expands. PLoS Genetics 11:e1004941.

Latrille T, Duret L, Lartillot N. 2017. The Red Queen model of recombination hot-spot evolution: a theoretical investigation. Philosophical Transactions of the Royal Society B 372:20160463.

Lesecque Y, Glémin S, Lartillot N, Mouchiroud D, Duret L. 2014. The red queen model of recombination hotspots evolution in the light of archaic and modern human genomes. PLoS genetics 10:e1004790.

Liti G, Carter DM, Moses AM, Warringer J, Parts L, James SA, Davey RP, Roberts IN, Burt A, Koufopanou V. 2009. Population genomics of domestic and wild yeasts. Nature 458:337.

Loewe L, Charlesworth B. 2006. Inferring the distribution of mutational effects on fitness in *Drosophila*. Biology Letters 2:426–430.

Loire E, Chiari Y, Bernard A, Cahais V, Romiguier J, Nabholz B, Lourenço JM, Galtier N. 2013. Population genomics of the endangered giant Galápagos tortoise. Genome Biology 14:R136.

Long H, Sung W, Kucukyildirim S, Williams E, Miller SF, Guo W, Patterson C, Gregory C, Strauss C, Stone C, et al. 2018. Evolutionary determinants of genome-wide nucleotide composition. Nature Ecology & Evolution 2:237–240.

Lopez P, Casane D, Philippe H. 2002. Heterotachy, an important process of protein evolution. Molecular Biology and Evolution 19:1–7.

Lourenço J, Galtier N, Glémin S. 2011. Complexity, pleiotropy, and the fitness effect of mutations. Evolution 65:1559–1571.

Lourenço JM, Glémin S, Galtier N. 2013. The rate of molecular adaptation in a changing environment. Molecular Biology and Evolution 30:1292–1301.

Lukens L, Doebley J. 2001. Molecular evolution of the teosinte branched gene among maize and related grasses. Molecular Biology and Evolution 18:627–638.

Lynch M, Conery JS. 2003. The Origins of Genome Complexity. Science 302:1401–1404.

Lynch M, Gutenkunst R, Ackerman M, Spitze K, Ye Z, Maruki T, Jia Z. 2017. Population Genomics of *Daphnia pulex*. Genetics 206:315.

Lynch M, Koskella B, Schaack S. 2006. Mutation pressure and the evolution of organelle genomic architecture. Science 311:1727–1730.

Lynch M, Walsh B. 2007. The origins of genome architecture. Sinauer Associates Sunderland (MA)

Mancera E, Bourgon R, Brozzi A, Huber W, Steinmetz LM. 2008. High-resolution mapping of meiotic crossovers and non-crossovers in yeast. Nature 454:479–485.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in *Drosophila*. Nature 351:652–654.

Messer PW, Petrov DA. 2013. Frequent adaptation and the McDonald–Kreitman test. PNAS 110:8615–8620.

Meunier J, Duret L. 2004. Recombination drives the evolution of GC-content in the human genome. Molecular Biology and Evolution 21:984–990.

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

Mugal CF, Arndt PF, Ellegren H. 2013. Twisted signatures of GC-biased gene conversion embedded in an evolutionary stable karyotype. Molecular Biology and Evolution 30:1700–1712.

Murphy WJ, Eizirik E, Johnson WE, Zhang YP, Ryder OA, O'brien SJ. 2001. Molecular phylogenetics and the origins of placental mammals. Nature 409:614.

Muyle A, Serres-Giardi L, Ressayre A, Escobar J, Glémin S. 2011. GC-biased gene conversion and selection affect GC-content in the *Oryza* genus (rice). Molecular Biology and Evolution 28:2695–2706.

Nagylaki T. 1983. Evolution of a finite population under gene conversion. Proceedings of the National Academy of Sciences 80:6278–6281.

Nikolaev SI, Montoya-Burgos JI, Popadin K, Parand L, Margulies EH, Program NI of HISCCS, Antonarakis SE. 2007. Life-history traits drive the evolutionary rates of mammalian coding and noncoding genomic elements. PNAS 104:20443–20448.

Ohta T. 1973. Slightly deleterious mutant substitutions in evolution. Nature 246:96.

Orr HA. 2003. The distribution of fitness effects among beneficial mutations. Genetics 163:1519–1526.

Parsch J, Zhang Z, Baines JF. 2009. The influence of demography and weak selection on the McDonald–Kreitman test: an empirical study in *Drosophila*. Molecular Biology and Evolution 26:691–698.

Pessia E, Popa A, Mousset S, Rezvoy C, Duret L, Marais GAB. 2012. Evidence for widespread GC-biased gene conversion in eukaryotes. Genome Biology and Evolution 4:675–682.

Piganeau G, Eyre-Walker A. 2003. Estimating the distribution of fitness effects from DNA sequence data: Implications for the molecular clock. PNAS 100:10335–10340.

Pollard KS, Salama SR, King B, Kern AD, Dreszer T, Katzman S, Siepel A, Pedersen JS, Bejerano G, Baertsch R. 2006. Forces shaping the fastest evolving regions in the human genome. PLoS genetics 2:e168.

Pond SLK, Frost SD. 2005. Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. Bioinformatics 21:2531–2533.

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. PNAS 104:13390–13395.

Rand DM, Weinreich DM, Cezairliyan BO. 2000. Neutrality tests of conservative-radical amino acid changes in nuclear-and mitochondrially-encoded proteins. Gene 261:115–125.

Ratnakumar A, Mousset S, Glémin S, Berglund J, Galtier N, Duret L, Webster MT. 2010. Detecting positive selection within genomes: the problem of biased gene conversion. Philosophical Transactions of the Royal Society B: Biological Sciences 365:2571–2580.

Razeto-Barry P, Díaz J, Vásquez RA. 2012. The nearly neutral and selection theories of molecular evolution under the fisher geometrical framework: substitution rate, population size, and complexity. Genetics 191:523–534.

Rocha EP, Danchin A. 2004. An analysis of determinants of amino acids substitution rates in bacterial proteins. Molecular Biology and Evolution 21:108–116.

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

Salser W, Bowen S, Browne D, el-Adli F, Fedoroff N, Fry K, Heindell H, Paddock G, Poon R, Wallace B, et al. 1976. Investigation of the organization of mammalian chromosomes at the DNA sequence level. Federation Proceedings 35:23–35.

Sawyer SA, Dykhuizen DE, Hartl DL. 1987. Confidence interval for the number of selectively neutral amino acid polymorphisms. Proceedings of the National Academy of Sciences 84:6225–6228.

Sawyer SA, Kulathinal RJ, Bustamante CD, Hartl DL. 2003. Bayesian analysis suggests that most amino acid replacements in *Drosophila* are driven by positive selection. Journal of Molecular Evolution. 57 Suppl 1:S154-164.

Sémon M, Lobry JR, Duret L. 2005. No evidence for tissue-specific adaptation of synonymous codon usage in humans. Molecular Biology and Evolution 23:523–529.

Shabalina SA, Kondrashov AS. 1999. Pattern of selective constraint in *C. elegans* and *C. briggsae* genomes. Genetics Research 74:23–30.

Shields DC, Sharp PM, Higgins DG, Wright F. 1988. “ Silent” sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. Molecular Biology and Evolution 5:704–716.

Singhal S, Leffler EM, Sannareddy K, Turner I, Venn O, Hooper DM, Strand AI, Li Q, Raney B, Balakrishnan CN. 2015. Stable recombination hotspots in birds. Science 350:928–932.

Smeds L, Mugal CF, Qvarnström A, Ellegren H. 2016. High-resolution mapping of crossover and non-crossover recombination events by whole-genome re-sequencing of an avian pedigree. PLoS genetics 12:e1006044.

Smith JM, Smith NH. 1996. Synonymous nucleotide divergence: what is “saturation”? Genetics 142:1033–1036.

Smith NG, Eyre-Walker A. 2002. Adaptive protein evolution in *Drosophila*. Nature, 415 :1022.

Smith NGC. 2003. Are Radical and Conservative Substitution Rates Useful Statistics in Molecular Evolution? Journal of Molecular Evolution 57:467–478.

Smith TCA, Arndt PF, Eyre-Walker A. 2018. Large scale variation in the rate of germ-line de novo mutation, base composition, divergence and diversity in humans. PLOS Genetics 14:e1007254.

Spencer CCA. 2006. Human polymorphism around recombination hotspots. Portland Press Limited

Stoletzki N, Eyre-Walker A. 2010. Estimation of the neutrality index. Molecular Biology and Evolution 28:63–70.

Strasburg JL, Kane NC, Raduski AR, Bonin A, Michelmore R, Rieseberg LH. 2010. Effective population size is positively correlated with levels of adaptive divergence among annual sunflowers. Molecular Biology and Evolution 28:1569–1580.

Stumpf MP, Thorne T, de Silva E, Stewart R, An HJ, Lappe M, Wiuf C. 2008. Estimating the size of the human interactome. Proceedings of the National Academy of Sciences 105:6959–6964.

Szathmáry E, Jordán F, Pál C. 2001. Can Genes Explain Biological Complexity? Science 292:1315–1316.

Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics 105:437–460.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595.

Tataru P, Mollion M, Glémin S, Bataillon T. 2017. Inference of Distribution of Fitness Effects and Proportion of Adaptive Substitutions from Polymorphism Data. Genetics 207:1103–1119.

Tsagkogeorga G, Cahais V, Galtier N. 2012. The population genomics of a fast evolver: high levels of diversity, functional constraint, and molecular adaptation in the tunicate *Ciona intestinalis*. Genome Biology and Evolution 4:852–861.

Urrutia AO, Hurst LD. 2001. Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection. Genetics 159:1191–1199.

Wallberg A, Glémin S, Webster MT. 2015. Extreme recombination frequencies shape genome variation and evolution in the honeybee, *Apis mellifera.* PLoS genetics 11:e1005189.

Waterson RH, Lander ES, Wilson RK. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437:69.

Weber CC, Nabholz B, Romiguier J, Ellegren H. 2014. Kr/Kc but not dN/dS correlates positively with body mass in birds, raising implications for inferring lineage-specific selection. Genome Biology 15:542.

Webster MT, Axelsson E, Ellegren H. 2006. Strong regional biases in nucleotide substitution in the chicken genome. Molecular Biology and Evolution 23:1203–1216.

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

Welch DBM, Meselson MS. 2001. Rates of nucleotide substitution in sexual and anciently asexual rotifers. Proceedings of the National Academy of Sciences 98:6720–6724.

Williams AL, Genovese G, Dyer T, Altemose N, Truax K, Jun G, Patterson N, Myers SR, Curran JE, Duggirala R. 2015. Non-crossover gene conversions show strong GC bias and unexpected clustering in humans. Elife 4.

Wright S. 1931. Evolution in Mendelian populations. Genetics 16:97–159.

Yang L, Gaut BS. 2011. Factors that contribute to variation in evolutionary rate among *Arabidopsis* genes. Molecular Biology and Evolution 28:2359–2369.

Yang Z. 2000. Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus A. Journal of Molecular Evolution 51:423–432.

Zhang J. 2000. Rates of conservative and radical nonsynonymous nucleotide substitutions in mammalian nuclear genes. Journal of Molecular Evolution 50:56–68.

Zhang J, Yang J-R. 2015. Determinants of the rate of protein sequence evolution. Nature Reviews Genetics 16:409–420.

Zhang L, Li W-H. 2005. Human SNPs reveal no evidence of frequent positive selection. Molecular Biology and Evolution 22:2504–2507.

Zuckerkandl E, Pauling L. 1965. Evolutionary divergence and convergence in proteins. Evolving genes and proteins 97–166.

Context of the PhD project and manuscript structure

This PhD project has been proposed by Nicolas Galtier after I did my Master project co-supervised by Benoit Nabholz and him. This Master project aimed at comparing the efficiency of purifying and positive selection between the Z-chromosome and autosomes in two Satyrine butterflies. I published the results of this project during my first year of PhD in an article entitled « Hemizygosity enhances purifying selection: lack of Fast-Z evolution in two Satyrine butterflies » in *Genome Biology and Evolution* (Rousselle et al. 2016). After this, I had the opportunity to contribute to a related project led by Ana Pinharanda in Jiggins lab, who explored the same questions in *Heliconius* butterfies. This also led to an article called « Sexually dimorphic gene expression and transcriptome evolution provides mixed evidence for a fast-Z effect in *Heliconius* », which is under revision in *Journal of Evolutionary Biology.* In these two projects, I discovered, alongside bio-informatics, the DFE-α method. In the mean time, Nicolas published a study exploring the relationship between the adaptive substitution rate, ω_a, and the effective populations size N_e approximated by the synonymous genetic diversity in 44 species pairs (Galtier 2016). Contrary to theoretical expectations, he did not detect a positive relationship between ω_a and N_e. This study represents the starting point of my PhD project. It was originally entitled "Adaptive evolution in animals: effective population size, GC-biased gene conversion and fitness effects of mutations", and contrary to a lot of thesis, this title is still pretty relevant in view of the work I carried out. I chose however to change the title for it to be more informative of the questions I tried to answer, and not the different topics I addressed. I organized the next chapters of this manuscript roughly following the chronology of my PhD, adjusting only the third and fourth chapter to facilitate the comprehension and follow the order in which I would like to later publish the corresponding articles. The articles that constitute the four chapters are preceded by short contextualiations that present the goals of each studies and the reasons that led us to address them, as well as details regarding the work I did myself *vs.* the work for which I received help or which was carried out by the two internship students I supervised with Nicolas during my PhD.

References:

Galtier N. 2016. Adaptive protein evolution in animals and the effective population size hypothesis. PLOS Genetics 12:e1005774.

Rousselle M, Faivre N, Ballenghien M, Galtier N, Nabholz B. 2016. Hemizygosity enhances purifying selection: lack of fast-Z evolution in two Satyrine butterflies. Genome Biology and Evolution 8:3108–3119.

Chapitre 1

Overestimation of the adaptive substitution rate in fluctuating populations

In this first chapter, we evaluated the impact of long-term demographic fluctuations on the estimation of the adaptive substitution rate (ω_a) and the proportion of adaptive substitutions (α) via simulations.

This project emerged in a context where the estimation methods of the adaptive substitution rate have improved since the seminal McDonald & Kreitman test (McDonald and Kreitman 1991), in particular in terms of recent population size changes and presence of slightly deleterious mutations. Those two sources of bias impact the π_n/π_s ratio and the site frequency spectra. They lead to the violation of one of the assumptions of the test, which is that the regime of selection/drift – i.e. the distribution of fitness effects and the effective population size – has remained constant since the divergence of the species under study to the moment data were sampled.

Nevertheless, those new developments (in particular the most sophisticated method first introduced by Eyre-Walker et al. 2006 then Eyre-Walker & Keithgley, 2009 (Eyre-Walker et al. 2006; Eyre-Walker and Keightley 2009) only partially correct for the violation of this assumption. Indeed, they only correct for recent variations of the regime of selection/drift, which are still traceable from polymorphism data. They do not account for long-term variations that affect divergence and not polymorphism, and for good reason, all signal of such variations being erased from coding sequences.

Despite the fact that those variations are not traceable in polymorphism data, several empirical results indicate that long-term fluctuations in the regime of selection/drift can lead to errors in the estimation of ω_a. For instance, a study reveals that in the brown rat *Rattus norvegicus*, Eyre-Walker & Keightley's method yielded a negative value of α (Deinum et al. 2015). This result is interpreted as the consequence of the strong bottleneck that happened around 20 000 years ago.
Besides, in a recent project Nicolas Galtier estimated ω_a and α in 44 species pairs and observed that the results changed depending on which species was used as the outgroup and which used at the ingroup (Galtier 2016).

These unexpected empirical results led us to investigate more deeply this source of bias by measuring its direction and extent via simulations. The use of simulations allowed us to obtain the true values of α that can then be compared to estimations of several implementation of the so-called DFE-α method (Tataru et al. 2017).

When we conceived this project, Maeva Mollion, a PhD student of the University of Aarhus working under the supervision of Thomas Bataillon, chose the Molecular Evolution and Phylogeny team of ISEM to carry out a four month-long research project. She had previously worked with Paula Tataru and Thomas Bataillon on a new implementation of the DFE-α method (Tataru et al. 2017), so it seemed relevant to include her in this project. She then took part in the generation of simulations and their analysis.

References :

Deinum EE, Halligan DL, Ness RW, Zhang Y-H, Cong L, Zhang J-X, Keightley PD. 2015. Recent evolution in *Rattus norvegicus* is shaped by declining effective population size. Molecular Biology and Bvolution 32:2547–2558.

Eyre-Walker A, Keightley PD. 2009. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. Molecular Biology and Evolution 26:2097–2108.

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

Galtier N. 2016. Adaptive protein evolution in animals and the effective population size hypothesis. PLOS Genetics 12:e1005774.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in Drosophila. Nature 351:652–654.

Tataru P, Mollion M, Glémin S, Bataillon T. 2017. Inference of distribution of fitness effects and proportion of adaptive substitutions from polymorphism data. Genetics 207:1103–1119.

BIOLOGY LETTERS

rsbl.royalsocietypublishing.org

Research

Cite this article: Rousselle M, Mollion M, Nabholz B, Bataillon T, Galtier N. 2018 Overestimation of the adaptive substitution rate in fluctuating populations. *Biol. Lett.* **14**: 20180055.
http://dx.doi.org/10.1098/rsbl.2018.0055

Received: 25 January 2018
Accepted: 18 April 2018

Subject Areas:
evolution, molecular biology

Keywords:
molecular adaptation, simulations, coding sequence evolution, fitness effect of mutations, effective population size

Author for correspondence:
Marjolaine Rousselle
e-mail: marjolaine.rousselle@umontpellier.fr

Electronic supplementary material is available online at https://dx.doi.org/10.6084/m9.figshare.c.4084511.

THE ROYAL SOCIETY PUBLISHING

Molecular evolution

Overestimation of the adaptive substitution rate in fluctuating populations

Marjolaine Rousselle[1], Maeva Mollion[2], Benoit Nabholz[1], Thomas Bataillon[2] and Nicolas Galtier[1]

[1]UMR-5554 Institut des Sciences de l'Evolution, CNRS, Université de Montpellier, IRD, EPHE, Montpellier, France
[2]BiRC-Bioinformatics Research Center, Aarhus University, Aarhus, Denmark

MR, 0000-0003-3930-0732; BN, 0000-0003-0447-1451; TB, 0000-0002-4730-2538; NG, 0000-0002-0479-4878

Estimating the proportion of adaptive substitutions (α) is of primary importance to uncover the determinants of adaptation in comparative genomic studies. Several methods have been proposed to estimate α from patterns polymorphism and divergence in coding sequences. However, estimators of α can be biased when the underlying assumptions are not met. Here we focus on a potential source of bias, i.e. variation through time in the long-term population size (N) of the considered species. We show via simulations that ancient demographic fluctuations can generate severe overestimations of α, and this is irrespective of the recent population history.

1. Introduction

The proportion of amino acid substitutions that are adaptive, α, is an important parameter routinely inferred in population genomic studies. Methods for estimating α usually rely on the McDonald & Kreitman principle, which is based on the comparison of polymorphic and fixed mutations at synonymous versus non-synonymous positions [1–5]. These methods, however, make various assumptions that are not always met in real data. One of them is that the stringency of purifying selection against deleterious mutations has been constant over the considered time period, so that the rate of non-adaptive (neutral and slightly deleterious) substitutions during divergence, ω_{na}, can be estimated based on polymorphism data. Sequence divergence, however, builds up across long periods of time. If the selection regime has changed in the past, so that polymorphism data are not representative of the average long-term process governing rates of divergence, then the estimation of α can be biased [6].

Importantly, the strength of purifying selection is determined by population size (N) [7], which likely varies in time. The geographical range of Palaearctic species, for instance, has expanded and contracted due to the alternation of glacial and interglacial periods during the Quaternary [8]. Species adapted to warm habitats, hereafter called 'temperate', have mainly subsisted in refuges during glacial periods, where their census population sizes were presumably reduced. Conversely, species adapted to cold habitats, hereafter called 'alpine', have presumably occupied larger ranges during cold periods than during warm ones [9,10]. Both types of species are particularly prone to exhibit discrepancies between current and average long-term N, which might bias estimations of α based on both divergence and polymorphism data that have not been generated under the same selection regime. More specifically, if the current population size is lower than the average long-term population size, we expect an underestimation of this statistic due to the presence of slightly deleterious mutations that are now segregating in polymorphism data, but have not contributed to divergence. If, however,

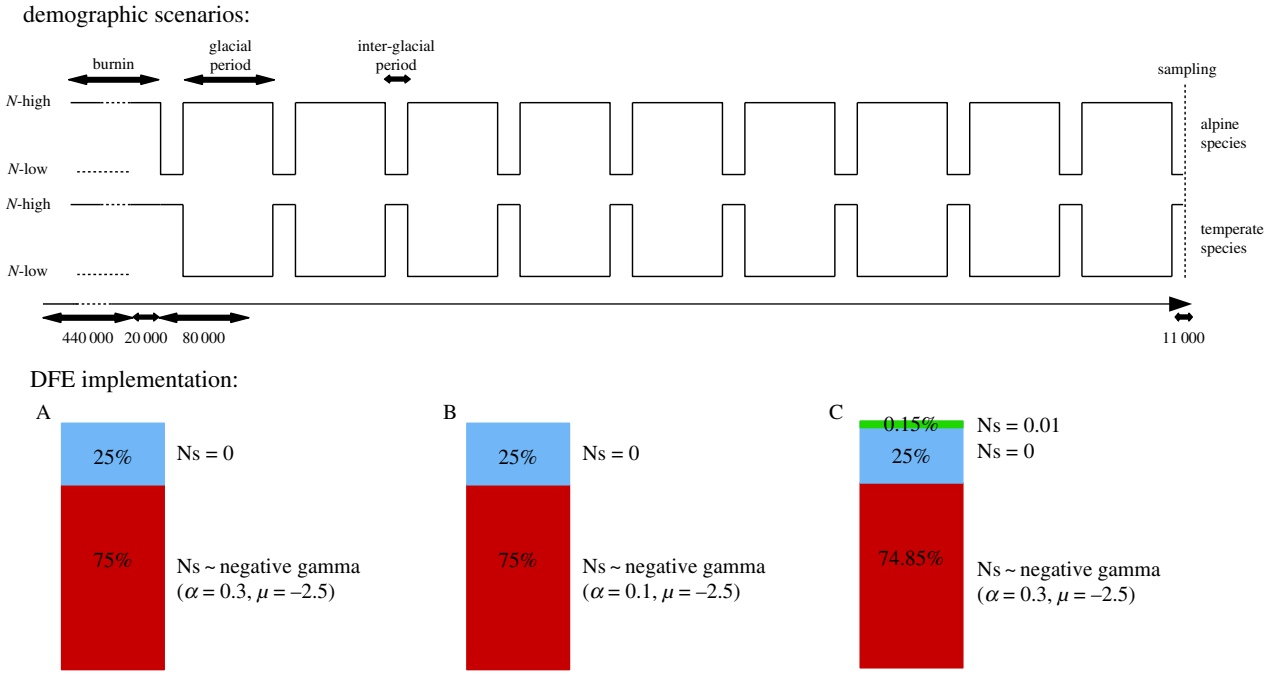


Figure 1. Demographic scenario and distribution of fitness effect (DFE) parameters used in the simulations. *N*-high is set to 10^5 individuals, and *N*-low is set to 10^4 individuals for the reference scenarios, and 10^3 individuals for the 'severe bottlenecks' scenarios. DFE 'A' describes the reference DFE we use in the simulations, which did not include a positive component. DFE 'B' presents a change in the shape of the negative DFE, modelled as a gamma distribution, and DFE 'C' contains a positive DFE component. α and μ, respectively, stand for the shape and the mean parameter of a gamma distribution. (Online version in colour.)

current population size is higher than the average long-term population size, then we expect an overestimation of α due to an inflated rate of fixation of slightly deleterious mutations in the past, whereas these have been efficiently selected against recently.

Methods taking into account recent *N* changes that affect polymorphism in the estimation process have been developed and proved to perform well [2]. In contrast, ancient changes that affect divergence have been rarely investigated. Modelling single changes in *N*, Eyre-Walker showed that in presence of slightly deleterious mutations, an increase in *N* in the past could yield spurious evidence for positive selection, whereas a decrease in *N* can either increase or decrease α depending on when it happened [6].

To further explore the extent of the error one can make when estimating genome-wide adaptive rates, we simulated coding sequence evolution under plausible scenarios involving fluctuating *N* along with linkage effects, and tested the performance of the most recent implementations of the McDonald–Kreitman approach [4,5]. Our results reveal a substantial overestimation of α and of the adaptive rate ω_a in most scenarios, calling for a re-examination of the interpretation of the high values of α often reported based on these methods.

2. Material and methods

(a) Generating simulated data of temperate and alpine species

We simulated the evolution of coding sequences in a single population evolving forward in time using SLIM V2 [11]. We considered panmictic populations of diploid individuals whose genomes consisted of 1500 coding sequences, each of 999 base pairs. The mutation rate was set to 2.2×10^{-9} per base pair per generation [12] and the recombination rate to 1×10^{-8} per base pair per generation [13]. Simulations differed from one another with respect to the demographic scenario and the assumed distribution of the fitness effect of mutations (DFE) as shown in figure 1. The alternation of a high (10^5) and a low *N* (10^4) was assumed to follow quaternary climatic cycles, with temperate species having a high *N* during interglacial periods, and alpine having a high *N* during glacial periods. Each combination of parameters was replicated 50 times. SLIM allows tracking of all mutations arising during a simulation, along with their associated fitness effect and frequency, which can be 1 if the mutation has reached fixation. Each mutation that arose during a simulation was categorized as either synonymous (if the fitness effect was zero) or non-synonymous (if the fitness effect was different from zero). For each replicate simulation, we retrieved all fixed and segregating mutations and their population frequencies, and we computed non-synonymous and synonymous divergence and the unfolded site frequency spectra (SFS). An unfolded SFS is a vector of $2n-1$ entries corresponding to the counts of SNPs where the absolute frequency of the derived allele is 1, 2, …, $2n-1$, respectively, in a sample of n diploid individuals. Here samples of size $n = 10$ individuals were considered. From SLIM output we also calculated for each simulation the true (realized) values of the proportion of adaptive substitutions α, the rate of adaptive substitutions ω_a and the rate of non-adaptive substitutions ω_{na}.

(b) Computing estimates of α, ω_a and ω_{na}

We estimated α, ω_a and ω_{na} from simulated SFS and divergence data using two distinct programs introduced by Galtier [4] and Tataru *et al.* [5], denoted by G and T underscripts, respectively. Both programs re-implement and extend a method of estimation of the adaptive rate introduced by Eyre-Walker & Keightley [2]. Distinct DFE models were considered and fitted to SFS and divergence data. The model 'Gamma' includes the fitting of a gamma distribution for neutral and deleterious mutations, along with a fixed class of strongly adaptive substitution, whereas under model 'GammaExpo' the DFE includes both a negative (gamma) and a positive (exponential) component [4,5] (see electronic

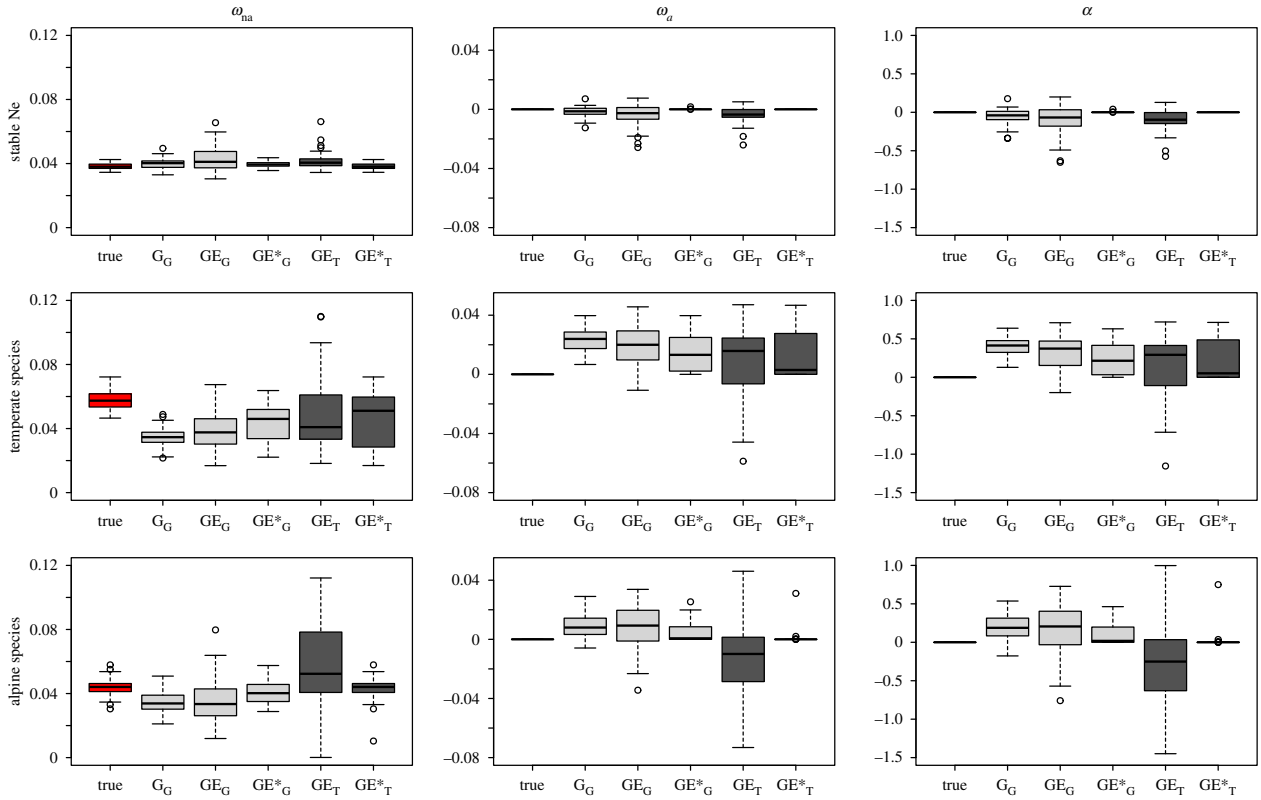


Figure 2. ω_{na}, ω_a and α estimates according to different models (implementation of Galtier in light grey and implementation of Tataru in dark grey: 'G' stands for Gamma, 'GE' for GammaExpo and 'GE*' for GammaExpo*) for three main treatments: stable population size, temperate species and alpine species with the simulated DFE 'A' in figure 1, i.e. which does not include adaptive mutation. (Online version in colour.)

supplementary methods). Two approaches were taken for estimating ω_a, ω_{na} and α posterior to model fitting. In the first approach [2], the expected ω_{na} is estimated from the inferred DFE, whereas ω_a and α are obtained by subtracting ω_{na} from the observed dN/dS ratio ($dN/dS = \omega_a + \omega_{na}$ and $\omega_a = \alpha^* dN/dS$). In the second approach [4,5], divergence data were only used at the DFE model fitting step, ω_{na}, ω_a and α being estimated directly from the inferred DFE (see electronic supplementary methods). This approach is only applicable to DFE models including an explicit adaptive DFE component, such as GammaExpo. We called this procedure 'GammaExpo*'. When estimating DFE model parameters, we jointly accounted for demographic effects by using nuisance parameters, which correct each class of frequency of the synonymous and non-synonymous SFS relative to the neutral expectations in an equilibrium Wright–Fisher population [14].

3. Results and discussion

(a) Influence of past demographic fluctuations on the estimations of the parameters of adaptation

We report a substantial overestimation of α in the great majority of scenarios of fluctuating N, especially for temperate species, whereas methods tended to reliably infer positive selection when N was stable (figure 2 and table 1). In the worst case (inference made with the model $Gamma_G$, for a temperate species and simulation under a scenario with no adaptive evolution (DFE implementation 'A'), all the replicates yielded estimated α values above 0.13, 54% of which being above 0.4, whereas the true α equalled 0 (figure 2). The overestimation of α for temperate species is caused by an underestimation of the non-adaptive substitution rate ω_{na}, indicating that this result is indeed due to the presence of slightly deleterious mutations. As for alpine species, for which the last episode of demographic fluctuation is a population decrease, we still observed inflated α values with some models, contrary to the findings of [6]. This might be because here the average expected coalescence time is longer than the last inter-glacial period (11 000 generations), so that polymorphism reflects a period that may span several cycles of fluctuations. Thus the long-term population size, which can be approximated by the harmonic mean through time, may not be greater than the recent N even in alpine species (see electronic supplementary material, table S1).

(b) Comparison of the different methods of inference of α

Figure 2 shows that the five tested inference models are overall consistent, all showing an overestimation of α when there are past fluctuations of N. A significantly positive correlation was found between the two pairs of inference models GammaExpo and GammaExpo* of the implementations of Galtier and Tataru [4,5] ($R = 0.59$ and $R = 0.73$ respectively, and electronic supplementary material, figure S1). However, the models differed in terms of variance of the estimations and in the extent of the overestimation they produced. The model performing the best was GammaExpo*. This model uses the observed non-synonymous divergence count as a parameter when fitting the DFE to the data, and estimates ω_a from the inferred DFE. Consequently, the discrepancies between divergence and polymorphism data that could be entailed by past demographic fluctuations are mitigated, as the parameters of the DFE that lead to the estimation of ω_a are adjusted from the two sources of data. Ideally, using solely polymorphism

Table 1. Statistic estimates according to different models for additional treatments with different DFE implementations (see the DFE descriptions in figure 1), different recombination rate and severeness of bottlenecks.

	scenario	true α	α_{GammG}	dN/dS	$\pi n/\pi s$	true ω_a	ω_{aGammG}	true ω_{na}	$\omega_{naGammaG}$
temperate species	DFE C	0.58	0.61	0.15	0.079	0.085	0.090	0.062	0.057
		[0.52;0.66]	[0.40;0.77]	[0.12;0.18]	[0.063;0.097]	[0.069;0.10]	[0.056;0.13]	[0.046;0.081]	[0.038;0.079]
	DFE B	0	0.16	0.34	0.34	0	0.054	0.34	0.28
		[0;0]	[−0.030;0.33]	[0.31;0.37]	[0.29;0.38]	[0;0]	[−0.0094;0.11]	[0.31;0.37]	[0.22;0.32]
	DFE A	0	0.69	0.099	0.056	0	0.069	0.10	0.030
	N-low = 1000	[0;0]	[0.53;0.97]	[0.082;0.12]	[0.044;0.067]	[0;0]	[0.046;0.096]	[0.082;0.12]	[0.0034;0.043]
alpine species	DFE C	0.83	0.81	0.27	0.071	0.23	0.22	0.048	0.051
		[0.79;0.87]	[0.73;0.98]	[0.25;0.30]	[0.055;0.084]	[0.20;0.25]	[0.12;0.28]	[0.034;0.059]	[0.0068;0.071]
	DFE B	0	0.084	0.31	0.33	0	0.027	0.31	0.28
		[0;0]	[−0.074;0.24]	[0.28;0.34]	[0.30;0.36]	[0;0]	[−0.022;0.075]	[0.28;0.34]	[0.24;0.32]
	DFE A	0	0.48	0.067	0.054	0	0.032	0.067	0.035
	N-low = 1000	[0;0]	[0.29;0.72]	[0.055;0.077]	[0.042;0.065]	[0;0]	[0.019;0.051]	[0.055;0.077]	[0.019;0.048]
stable *N*	DFE C	0.73	0.65	0.23	0.10	0.17	0.15	0.062	0.080
		[0.72;0.75]	[0.60;0.72]	[0.22;0.24]	[0.092;0.11]	[0.16;0.18]	[0.14;0.16]	[0.059;0.065]	[0.064;0.094]
	DFE A	0	0.0022	0.034	0.056	0	0.00019	0.034	0.034
	higher recombination rate	[0;0]	[−0.21;0.15]	[0.030;0.037]	[0.053;0.059]	[0;0]	$[-6.5 \times 10^{-3};5.5 \times 10^{-3}]$	[0.030;0.037]	[0.029;0.038]
	DFE C	0.90	0.87	0.41	0.069	0.37	0.36	0.040	0.051
	higher recombination rate	[0.90;0.91]	[0.85;0.90]	[0.40;0.42]	[0.064;0.074]	[0.36;0.38]	[0.34;0.38]	[0.037;0.043]	[0.044;0.060]

data, such as developed in [5], removes all risks of discrepancies, but such a method needs high quality datasets, and estimation relying exclusively on SFS data pays a penalty because estimates will have inherently more sampling variance [5]. Nevertheless, with such a method, α estimates will reflect the recent proportion of adaptive substitutions, contrary to the α estimated by the other models.

(c) Control analyses

We explored how estimation of α and ω_a is influenced by changes in the DFE, as well as in recombination rate and the intensity of the bottlenecks (table 1). When adaptive mutation was part of the simulation, we observed a slight overestimation of α only for temperate species. This suggests that the estimation methods are particularly prone to strong biases particularly when the true adaptive substitution rate is low.

Additionally, we tested the influence of the shape of the negative part of the DFE modelled with a gamma distribution. Using a shape parameter of 0.3 instead of 0.1 decreased the true value of α and lead to a less severe overestimation of the adaptive rate (table 1). This is likely due to the fact that, as the shape parameter approaches zero, the assumed DFE becomes closer and closer to the neutral model—i.e. with a low prevalence of slightly deleterious mutations—so that dN/dS is less and less sensitive to N.

We tested the influence of the recombination rate by increasing it by a factor of 10 under stable N. This decreased the bias and reduced the variance of the estimations of α substantially for the simulations using the DFE 'C', but not for those using DFE 'A', where the bias was already very low (table 1).

Finally, we also tested the effect of the intensity of the fluctuations, by decreasing the population size to 1000 individuals in the bottlenecks. As expected, the extent of the overestimation was stronger, with estimations of α reaching a mean of 0.69 for temperate species.

4. Conclusion

We showed that under plausible demographic scenarios involving fluctuations in population size, current inference methods can severely overestimate α. This upward bias is exacerbated when the true adaptive substitution rate is low, when there are a lot of slightly deleterious mutations, and recombination rate is low. The upward bias is mainly caused by ancient demographic events which are likely to be essentially undetectable from current polymorphism data. Without any specific clue regarding the long-term demographic history of a population, we therefore call for caution when interpreting the results of such methods, especially when the estimated α is high.

Methods relying on an explicit model of adaptive evolution performed better in the case of population fluctuations and should probably be developed further in order to overcome this problem.

Data accessibility. The unfolded SFS and divergence counts of each replicate of the different simulation scenarios we generated are deposited in the Dryad Digital Repository (http://dx.doi.org/10.5061/dryad.85qb2r1) [15].

Authors' contributions. M.R. and M.M. performed the simulations, analysis and interpretations, under the supervision of N.G., B.N. and T.B. M.R. drafted the article, and M.M., N.G., B.N. and T.B. revised it critically for important intellectual content. M.R., M.M., B.N., T.B. and N.G. all approve the final version and all aspects of the work, and agree to be accountable for all aspects of the work, the content therein and approve the final version of the manuscript.

Competing interests. We have no competing interests.

Funding. This work was supported by Agence Nationale de la recherche grant no. ANR-15-CE12-0010 'DarkSideOfRecombination'.

Acknowledgements. We thank Yoann Anciaux and Iago Bonicci for their help.

References

1. McDonald JH, Kreitman M. 1991 Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* **351**, 652–654. (doi:10.1038/351652a0)
2. Eyre-Walker A, Keightley PD. 2009 Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Mol. Biol. Evol.* **26**, 2097–2108. (doi:10.1093/molbev/msp119)
3. Messer PW, Petrov DA. 2013 Frequent adaptation and the McDonald–Kreitman test. *Proc. Natl Acad. Sci. USA* **110**, 8615–8620. (doi:10.1073/pnas.1220835110)
4. Galtier N. 2016 Adaptive protein evolution in animals and the effective population size hypothesis. *PLoS Genet.* **12**, e1005774. (doi:10.1371/journal.pgen.1005774)
5. Tataru P, Mollion M, Glémin S, Bataillon T. 2017 Inference of distribution of fitness effects and proportion of adaptive substitutions from polymorphism data. *Genetics* **207**, 1103–1119. (doi:10.1534/genetics.117.300323)
6. Eyre-Walker A. 2002 Changing effective population size and the McDonald–Kreitman test. *Genetics* **162**, 2017–2024.
7. Kimura H. 1962 On the probability of fixation of mutant genes in a population. *Genetics* **47**, 713–719.
8. Hewitt GM. 1999 Post-glacial re-colonization of European biota. *Biol. J. Linn. Soc.* **68**, 87–112. (doi:10.1111/j.1095-8312.1999.tb01160.x)
9. Stewart JR, Lister AM. 2001 Cryptic northern refugia and the origins of the modern biota. *Trends Ecol. Evol.* **16**, 608–613. (doi:10.1016/S0169-5347(01)02338-2)
10. Parmesan C. 2006 Ecological and evolutionary responses to recent climate change. *Annu. Rev. Ecol. Evol. Syst.* **37**, 637–669. (doi:10.1146/annurev.ecolsys.37.091305.110100)
11. Haller BC, Messer PW. 2017 SLiM 2: flexible, interactive forward genetic simulations. *Mol. Biol. Evol.* **34**, 230–240. (doi:10.1093/molbev/msw211)
12. Kumar S, Subramanian S. 2002 Mutation rates in mammalian genomes. *Proc. Natl Acad. Sci. USA* **99**, 803–808. (doi:10.1073/pnas.022629899)
13. Stapley J, Feulner PG, Johnston SE, Santure AW, Smadja CM. 2017 Variation in recombination frequency and distribution across eukaryotes: patterns and processes. *Phil. Trans. R. Soc. B.* **372**, 20160455. (doi:10.1098/rstb.2016.0455)
14. Eyre-Walker A, Woolfit M, Phelps T. 2006 The distribution of fitness effects of new deleterious amino acid mutations in humans. *Genetics* **173**, 891–900. (doi:10.1534/genetics.106.057570)
15. Rousselle M, Mollion M, Nabholz B, Bataillon T, Galtier N. 2018 Data from: Overestimation of the adaptive substitution rate in fluctuating populations. Dryad Digital Repository. (doi:10.5061/dryad.85qb2r1)

Chapitre 2

Influence of recombination and GC-biased gene conversion on the adaptive and non-adaptive substitution rate in mammals *vs.* birds

As the previous one, this second chapter is about a source of bias that can affect the DFE-α method, GC-biased gene conversion (gBGC).

Since the years 2000, many studies have alerted about the danger of gBGC to “mimic” the effects of positive selection, leading us to make the assumption that the estimations of α and ω_a may be impacted by this evolutionary force. The influence of gBGC on the dN/dS and π_n/π_s ratios has been recently assessed in several species (Galtier and Duret 2007; Berglund et al. 2009; Ratnakumar et al. 2010; Bolívar et al. 2016; Corcoran et al. 2017; Bolívar et al. 2018), and the results seem to differ between taxonomic groups, even if they are not easily comparable between studies. The discrepancies may be caused by the differences between the explored taxonomic groups in terms of recombination landscape dynamics. Besides, when we started this project, the impact of gBGC on the estimation of α and ω_a had never been assessed. This lead us to elaborate this project that has two objectives. In one hand we aimed at confirming or invalidating the results of previous studies as for the impact of gBGC on the dN/dS and π_n/π_s ratios by investigating two taxonomic groups that are different in terms of recombination landscape dynamics, namely mammals and birds. In the other hand, we aimed at exploring the impact of gBGC on the DFE-α method in these two groups.

In view of the existing questions as well as the multi-specific and multi-individual datasets available in the public databases, we chose to investigate Galloanserae (birds related to chicken and duck) and catharrine primates in this project. Those datasets have been later included in the projects presented in **Chapter 3** and **Chapter 4**. The arrival in January 2017 of Alexandre Laverré, a student in first year of master in evolution, was the opportunity to start this project, with the help of Émeric Figuet who prepared the data, first of Galloanserae, then later for primates. Alexandre worked on the Galloanserae dataset until the end of his four month-long internship under the shared supervision of Nicolas and I. I then resumed his analyses and analyzed the primate dataset.

I had the opportunity to present this work at two international conferences : first at the SMBE conference in July 2018 at Yokohama, Japan, and then at the joint congress « Evolution » in August 2018 in Montpellier.

On the same thematic, I also participated in a project involving the estimation of the impact of gBGC on allele frequency spectra in a large number of animal-scale taxonomic groups (Galtier et al. 2018).

References :

Berglund J, Pollard KS, Webster MT. 2009. Hotspots of biased nucleotide substitutions in human genes. PLoS biology 7:e1000026.

Bolívar P, Mugal CF, Nater A, Ellegren H. 2016. Recombination rate variation modulates gene sequence evolution mainly via GC-biased gene conversion, not Hill–Robertson interference, in an avian system. Molecular Biology and Evolution 33:216–227.

Bolívar P, Mugal CF, Rossi M, Nater A, Wang M, Dutoit L, Ellegren H. Biased inference of selection due to GC-biased gene conversion and the rate of protein evolution in flycatchers when accounting for it. Molecular Biology and Evolution.

Corcoran P, Gossmann TI, Barton HJ, Slate J, Zeng K. 2017. Determinants of the efficacy of natural selection on coding and noncoding variability in two passerine species. Genome Biology and Evolution 9:2987–3007.

Galtier N, Duret L. 2007. Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. Trends in Genetics 23:273–277.

Galtier N, Roux C, Rousselle M, Romiguier J, Figuet E, Glémin S, Bierne N, Duret L. 2018. Codon usage bias in animals: disentangling the effects of natural selection, effective population size and GC-biased gene conversion. Molecular Biology and Evolution 35:1092-1103.

Ratnakumar A, Mousset S, Glémin S, Berglund J, Galtier N, Duret L, Webster MT. 2010. Detecting positive selection within genomes: the problem of biased gene conversion. Philosophical Transactions of the Royal Society B: Biological Sciences 365:2571–2580.

Manuscript type: Article

Section: Discoveries

Title: Influence of recombination and GC-biased gene conversion on the adaptive and non-adaptive substitution rate in mammals *vs.* birds

Auhors: Rousselle M[1], Laverré A[1], Figuet E[1], Nabholz B[1], Galtier N[1].

[1]UMR 5554 Institut des Sciences de l'Evolution, CNRS, Université de Montpellier, IRD, EPHE, Place E. Bataillon, Montpellier, France

Corresponding author: Marjolaine Rousselle

marjolaine.rousselle@umontpellier.fr

ABSTRACT

Recombination is expected to affect functional sequence evolution in several ways. On one hand, recombination is thought to improve the efficiency of multi-locus selection by dissipating linkage disequilibrium. On the other hand, natural selection can be counteracted by recombination-associated transmission distorters such as GC-biased gene conversion (gBGC), which tends to promote G and C alleles irrespective of their fitness effect in high-recombining regions. It has been suggested that gBGC might impact coding sequence evolution in vertebrates, and particularly the ratio of non-synonymous to synonymous substitution rates (dN/dS). Distinctive gBGC patterns, however, have been reported in mammals and birds, maybe reflecting the documented contrasts in evolutionary dynamics of recombination rate between these two taxa. Here, we explore how recombination and gBGC affect coding sequence evolution in mammals and birds by analyzing proteome-wide data in six species of Galloanserae (fowls) and six species of catarrhine primates. We estimated the dN/dS ratio and rates of adaptive and non-adaptive evolution in bins of genes of increasing recombination rate, separately analyzing AT→GC, GC→AT and G↔C/A↔T mutations. We show that in both taxa, recombination and gBGC entail a decrease in dN/dS. Our analysis indicates that recombination enhances the efficiency of purifying selection by lowering Hill-Robertson effects, while gBGC leads to an overestimation of the adaptive rate of AT→GC mutations. Finally, we report a mutagenic effect of recombination, which is independent of gBGC.

INTRODUCTION

Understanding the relative importance of natural selection versus non-adaptive forces is a central question in molecular evolution (Kimura 1983, Gillespie 1991). Over the past recent years, a number of methods and statistics have been developed to assess the efficacy of positive and purifying selection.

Many of these methods are based on the comparison of the non-synonymous and synonymous mutation and substitution rates. Non-synonymous changes are supposedly under selective effects, whereas synonymous mutations are used as a control for non selective processes. Statistics commonly used to estimate the extent of selective pressure acting at the sequence level include dN/dS, the ratio of non-synonymous over synonymous substitution rate (between species), and π_n/π_s, the ratio of non-synonymous over synonymous nucleotide diversity (within species). Combining divergence and polymorphism data can provide a way to disentangle adaptive from non-

adaptive effects (McDonald & Kreitman 1991) and estimate the proportion of amino acid substitutions that resulted from positive selection – a proportion called α (Smith and Eyre-Walker 2002). The most recent versions of this approach, grouped under the name "DFE-α" (Eyre-Walker et al. 2006,Keightley and Eyre-Walker 2007, Eyre-Walker and Keightley 2009,Galtier 2016, Tataru et al. 2017), extract information on the distribution of the fitness effect (DFE) of non-synonymous mutations from the joint analysis of the synonymous and non-synonymous site frequency spectra (SFS). The expected dN/dS under near neutrality is deduced from the analysis of polymorphism data, and the difference between observed and expected dN/dS provides estimates for α and for the per synonymous substitution rate of adaptive and non-adaptive amino-acid substitution, respectively $\omega_a=\alpha(dN/dS)$ and $\omega_{na}=(1-\alpha)(dN/dS)$.

The methods reviewed above rely on the assumption that only drift and mutation determine the synonymous component, while drift, mutation and selection determine the non-synonymous component. However, coding sequences may be affected by other forces, such as selection on codon usage and GC-biased gene conversion (gBGC), which can modify the expectations regarding the dN/dS ratio, the π_n/π_s ratio and the DFE-α method (Galtier et al. 2009, Berglund et al. 2009, Ratnakumar et al. 2010, Bolívar et al. 2016, Corcoran et al. 2017). Here, we focus on gBGC as a potential source of bias in the estimation of the rate of adaptive and non-adaptive amino acid substitution.

gBGC originates from a repair bias during meiotic recombination that results in a distorted segregation favoring G and C over A and T alleles in highly recombining regions (Eyre-Walker 1993, Galtier et al. 2001, Glémin et al. 2015). A large body of literature provides evidence for gBGC in a wide range of organisms (Eyre-Walker 1999, Montoya-Burgos et al. 2003, Meunier and Duret 2004, Webster and Smith 2004, , Spencer 2006, Webster et al. 2006, Mancera et al. 2008, Escobar et al. 2011, Pessia et al. 2012, Lesecque et al. 2013, Williams et al. 2015, Halldorsson et al. 2016, Smeds et al. 2016, Keith et al. 2016 , Long et al. 2018, Galtier et al. 2018, Smith et al. 2018). This meiotic distortion both mimics positive selection by increasing the fixation probability of G or C (i.e., S) over A or T (i.e., W) neutral alleles (Galtier and Duret 2007, Berglund et al. 2009, Ratnakumar et al. 2010), and promotes the fixation of slightly deleterious GC alleles (Necşulea et al. 2011, Duret and Galtier 2009, Glémin 2010, Lachance and Tishkoff 2014). A striking example of the latter effect is the mouse *Fxy* gene. This gene, which was recently (<1-3 millions years ago) translocated in the house mouse *Mus musculus* from the X-specific region to the highly recombining pseudoautosomal region (PAR), experienced a dramatic increase in W→S substitution

rate in its part overlapping the PAR, at both coding and non coding sites. This resulted in a >100-fold increase in amino acid substitution rate in the *M. musculus* lineage, illustrating how gBGC can promote the fixation of otherwise counter-selected W→S mutations (Perry and Ashworth 1999, Montoya-Burgos et al. 2003). Ratnakumar et al. (2010) showed that gBGC significantly affects the evolution of functional coding sequences in mammals, and can lead to patterns of evolution that can be mistaken for positive selection (Ratnakumar et al. 2010). Besides, it has been shown that gBGC can elevate the dN/dS ratio locally in specific genes in primates (Galtier et al. 2009, Berglund et al. 2009, Ratnakumar et al. 2010, Kostka et al. 2012) indicating that gBGC may not impact the evolution of selected *vs.* neutral sites in the same way.

Somehow, the pattern reported in mammals does not seem to be observed in birds. An analysis of >8000 genes in the *Ficedula* flycatcher lineage indicated that recombination and gBGC tend to decrease the dN/dS ratio (Bolívar et al. 2016). Another study focusing on passenger and band-tailed pigeons found a higher dN/dS ratio for substitutions opposed by gBGC and a lower one for substitutions promoted by gBGC (Murray et al. 2017), a result also confirmed by (Bolívar et al. 2016). Besides, Corcoran et al. (2017) showed in great tits and zebra finches that ignoring the effect of gBGC can bias estimates of the DFE, α and ω_a.

Interestingly, there are reasons to suspect that mammals and birds differ with respect to gBGC dynamics, due to a fundamental difference between these two taxa in the way recombination is controlled. In many mammals recombination hotspot location is determined by the PRDM9 gene (Baudat et al. 2010, Myers et al. 2010, Parvanov et al. 2010, Sandor et al. 2012, Stevison et al. 2016, Baker et al. 2017). The PRDM9 protein binds DNA through a highly variable tandem array of zinc fingers, and this participates to recruitment of the protein complex initiating recombination via double strand break. PRDM9 evolves very rapidly (Oliver et al. 2009, Berg et al. 2011) and its binding motif experiences frequent changes (Brick et al. 2012), which result in a rapid turn-over in hotspot location, as demonstrated in primates and rodents (Lesecque et al. 2014, Baker et al. 2015, Latrille et al. 2017).

Birds, in contrast, lack the PRDM9 gene (Baker et al. 2017). In this group, recombination hotspots seem to be mainly located upstream of genes. A recent study in flycatchers reports a correlation between hotspot location and CpG islands, CpG islands being themselves often located in promoter regions (Kawakami et al. 2017). Recombination rate is also linked to chromosome size (Hillier et al.2004). The lack of PRDM9 and the conserved karyotype of birds probably explain that the location of recombination hotspots is conserved across species (Mugal et al. 2013, Singhal et al.

104 2015). gBGC may thus have a particularly strong effect on bird genome evolution by persistently
105 acting on specific genomic regions over long periods of time (Mugal et al. 2013). Moreover,
106 phylogenetic analyses indicate that GC-content at putatively neutral sites is still increasing in avian
107 genomes and has not yet reached its equilibrium (Webster et al. 2006, Nabholz et al. 2011, Weber et
108 al. 2014). In contrast, GC-content at putatively neutral sites seems to be decreasing in primates
109 (Duret et al. 2006). The distance between current GC-content and equilibrium GC-content has been
110 shown to affect the estimation of the dN/dS ratio. Indeed, Bolívar el al. (2016) showed that the GC-
111 content at 4-fold degenerated sites is further away from equilibrium than at 0-fold sites in
112 flycatchers, leading to a stronger impact of gBGC on synonymous than on non-synonymous
113 substitutions, which entails a decrease of the dN/dS ratio (Bolívar et al. 2016).

114 The above reviewed literature suggests that the distinctive pattern of coding sequence evolution in
115 mammals *vs.* birds could be mediated by gBGC, and explained by the contrasted dynamics of
116 recombination landscape between the two groups. This, however, is only a hypothesis requiring
117 further corroboration. First, the forces underlying this contrasted pattern are still difficult to
118 understand theoretically (Galtier et al. 2009, Bolívar et al. 2016). Secondly, the dN/dS ratio has
119 been measured and linked to gBGC using methods and gene sets that differ between studies, which
120 can greatly influence the results. In particular, model choice and assumptions have been shown to
121 potentially bias the estimation of dS, dN and dN/dS (Guéguen and Duret 2017). Thirdly, we still
122 lack a clear picture on how gBGC affects estimates of α, ω_a and ω_{na}, besides an indication that not
123 controlling for the effects of gBGC can lead to an overestimation of α (Corcoran et al. 2017).
124 Finally, yet another level of complexity is added by the fact that recombination is expected to affect
125 the synonymous and non-synonymous substitution rate irrespective of gBGC. Recombination (i)
126 could be mutagenic (Pratto et al. 2014, Arbeithuber et al. 2015), and (ii) is expected to enhance the
127 efficiency of natural selection by breaking linkage and Hill-Robertson interference (HRI, Hill and
128 Robertson 1966).

129 Here we investigated the influence of recombination and gBGC on adaptive and non-adaptive
130 coding sequence evolution using two datasets composed of six species of catarrhine primates
131 (mammals, with PRDM9) and six species of Galloanserae (birds, without PRDM9). We provide a
132 detailed comparison of the influence of gBGC on coding sequence evolution by separately
133 analyzing changes promoted by gBGC ($W \rightarrow S$), changes countered by gBGC ($S \rightarrow W$) and changes
134 supposedly unaffected by gBGC, i.e. GC-conservative ones ($A \leftrightarrow T$, $C \leftrightarrow G$). In both groups, we find
135 that recombination strongly influences the synonymous substitution rate and the dN/dS ratio,

particularly so for W→S changes, presumably reflecting the combined effect of gBGC and HRI. Contrary to what the current literature suggests, we report a roughly similar pattern in primates and Galloanserae, both showing a decrease of dN/dS with GC3. However, the shape of the relationship differs between the two taxa, likely reflecting differences in the dynamics of the recombination landscape. The analysis of GC-conservative synonymous substitutions reveals the existence of a mutagenic effect of recombination, which may also concern the other mutation types. Finally, we found that gBGC may lead to an overestimation of the adaptive substitution rate in both taxa.

RESULTS

1. Primates and Galloanserae alignments and SNP calling

We used six species of primates (*Homo sapiens*, *Pan troglodytes*, *Papio anubis*, *Pongo abelii*, *Gorilla gorilla*, *Macaca mulatta)* and six species of Galloanserae (*Meleagris gallopavo*, *Phasianus colchicus*, *Pavo cristatus*, *Numida meleagris*, *Anas platyrhynchos*, *Anser cygnoides)* in which we could find either genomic or transcriptomic data in at least five individuals (Prado-Martinez et al. 2013, Teixeira et al. 2015, Wright et al. 2015, Xue et al. 2016). Orthogroups prediction yielded 8604 orthogroups in primates and 4439 orthogroups in Galloanserae. For each orthogroup, we estimated branch specific dN, dS and dN/dS ratio per category of mutation (W→S, S→W, GC-conservative) and estimated ancestral sequences at each internal node of the species tree. Synonymous and non-synonymous Single Nucleotide Polymorphisms (SNPs) were called from polymorphism data and oriented using the predicted ancestral sequences, and classified as synonymous *vs.* non-synonymous, and W→S, S→W or GC-conservative. The same genes were thus used for divergence and polymorphism analysis.

As a control for orientation errors we masked all sites in the alignment containing at least one CpG site. GC-conservative SNPs were far less numerous that W→S and S→W SNPs, with W→S and S→W SNPs representing on average 90% of the 6447 (average per species) synonymous SNPs and 78% of the 2750 (average per species) non-synonymous SNPs.

2. Correlation between per-gene recombination rate and GC3

Using two available recombination maps (one for *Homo sapiens* and one for *Gallus gallus*, see Material & Methods) and the R package MareyMap, we estimated the per gene recombination rate

(r) by comparing the genetic map with the physical position of genes. Mean r was quite different between the two species (r=1.39 cM/MB with 95% confidence intervals of [0.52;2.93] in *H. sapiens* and r=3.98 cM/MB with 95% confidence intervals of [0.73;12.43] in *G. gallus*). GC3 and r were significantly correlated in both species, with Spearman correlation coefficients of 0.39 (p-value<2.2e-16) and 0.24 (p-value<2.2e-16) for *G. gallus* and *H. sapiens* respectively (**Figure S1**). We mainly used GC3 as a proxy of long-term recombination rate throughout the rest of the study.

Additionally, we estimated the correlation between GC3 and GC3*, the equilibrium GC-content at third codon position estimated under a model assuming non-stationary evolution. We found a significant positive correlation in Galloanserae (Spearman's R=0.38, p-value<2.2e-16) and in primates but the correlation was weaker in the latter (Spearman's R=0.05, p-value=4.6e-7). This is congruent with the suggestion that the recombination/gBGC landscape remains stable over long periods of time in birds (Mugal et al. 2013, Singhal et al. 2015), but evolves rapidly in primates (Lesecque et al. 2014).

3. Influence of GC3 level and recombination rate on divergence estimates

We binned orthogroups in ten sets of genes of even size sorted by increasing GC3 or r and compared the lineage specific dN, dS and dN/dS ratio estimated assuming non-stationarity of base composition across bins (all changes). **Figure 1** shows that in primates, the dN/dS ratio is negatively correlated with GC3 (this holds true when we use r instead of GC3, **Figure S2**). This effect was significant in all primate species but *Pan troglodytes* (see **Table 1**). The relationship between dN/dS and GC3 was also negative in all six species of Galloanserae, but non-significant (**Table 1**). Besides, the shape of the decreasing relationship between dN/dS and GC3 was quite different between the two taxonomic groups. When we considered r instead of GC3 (**Table S3**), similar results were obtained, but significance was only reached in *Pongo abelii* after applying a False Discovery Rate (FDR) correction.

To better understand these results, we separately analyzed W→S, S→W and GC-conservative substitutions. We see in **Figure 2** that in both groups, the dN/dS ratio calculated from W→S substitutions ($dN/dS_{[W\rightarrow S]}$) decreases with GC3, an effect that is strong and significant in all twelve species (**Table 1**). This decrease in $dN/dS_{[W\rightarrow S]}$ with GC3 is due to a significant increase of $dS_{[W\rightarrow S]}$ for all species. $dN_{[W\rightarrow S]}$, in contrast, is only marginally influenced by GC3. When considering r instead of GC3, $dS_{[W\rightarrow S]}$ still strongly increases, but surprisingly, $dN_{[W\rightarrow S]}$ also increases with r, an

effect that is significant for half of the species. $dN/dS_{[W \rightarrow S]}$ also decreases with r, and significantly so in eight species out of twelve (**Table S3**).

This suggests that gBGC has a strong influence on $dN/dS_{[W \rightarrow S]}$ by dramatically enhancing the fixation rate of W→S synonymous mutations, and much less that of W→S non-synonymous mutations. HRI may be the mechanism that explains the decoupling between $dS_{[W \rightarrow S]}$ and $dN_{[W \rightarrow S]}$: as recombination increases, the increased efficiency of purifying selection presumably prevents slightly deleterious W→S mutations to come to fixation, thus counteracting the effect of gBGC on $dN_{[W \rightarrow S]}$. Alternatively, this could reflect a stronger effect of gBGC on synonymous sites compared to non-synonymous sites because of a difference between the current GC-content and the equilibrium GC between sites – we further discuss this hypothesis below.

In contrast, no strong relationship was detected between $dN/dS_{[S \rightarrow W]}$ and GC3 or r. Interestingly, $dS_{[S \rightarrow W]}$ significantly increased with GC3 or r in primates (**Figure 2**, **Table 1**). Under the hypothesis that both HRI and gBGC shapes the relationship between GC3 or r and divergence, we would expect both $dN_{[S \rightarrow W]}$ and $dS_{[S \rightarrow W]}$ to decrease with GC3 or r. To better understand the determinants of this surprising pattern, we analyzed the GC-conservative pattern of substitution. We found that $dN/dS_{[GC\text{-}conservative]}$ globally decreases with GC3, significantly so in all primates but one, and in two Galloanserae. This is in agreement with the hypothesis that HRI affects the rate of fixation of slightly deleterious non-synonymous mutations. Interestingly, we found that in most species $dS_{[GC\text{-}conservative]}$, and to a lesser extent $dN_{[GC\text{-}conservative]}$, are positively correlated with GC3 and r (**Table 1**), which seems to imply the existence of a substantial mutagenic effect of recombination. This might explain why we do not observe a negative relationship between $dN_{[S \rightarrow W]}$ and $dS_{[S \rightarrow W]}$ and GC3.

We tested the robustness of these results to the codon model used in the estimation process of branch lengths and substitution parameters by reproducing the same divergence analysis with a model assuming stationarity of base composition (Guéguen and Duret 2017). The results were very consistent between the two models (see **Figure S3**, **Figure S4** and **Table S4** and **S5**).

4. Influence of GC3 level and recombination rate on polymorphism estimates

Within each species, SNPs were called from polymorphism data and classified as synonymous *vs.* non-synonymous, and W→S *vs.* S→W *vs*. GC-conservative (**Table S2**). We split the dataset in three bins of genes of even number of genes (**Figure 3**) and ten bins of even number of SNPs (**Table 2**)

sorted according to GC3. We see in **Figure 3** and **Table 2** that $\pi_n/\pi_{s[W\rightarrow S]}$ decreases with GC3, an effect which is significant in nine out of twelve species. This seems to be due both to a decrease of $\pi_{n[W\rightarrow S]}$ and an increase of $\pi_{s[W\rightarrow S]}$ with GC3 (**Table 2**). S→W mutations and GC-conservative mutations show basically the same pattern, although less markedly than for W→S. $\pi_n/\pi_{s[GC\text{-}conservative]}$ seems to decrease with GC3 as seen in **Figure 3**, even if the correlation between $\pi_n/\pi_{s[GC\text{-}conservative]}$ and GC3 is significant for only four species. **Figure 3** (A and B) shows that the decay of π_n/π_s with GC3 is steepest for W→S mutations, intermediate for GC-conservative mutations, and hardly significant for S→W mutations. These results are consistent with the divergence pattern, the GC-conservative pattern being intermediate between the W→S and S→W one. We found that in all twelve species W→S mutations segregate at a higher mean frequency than S→W and GC-conservative ones (**Figure 3**, **C** and **D**), confirming the influence of gBGC (mean difference between W→S and S→W synonymous average SNP frequency is 0.069 (SD=0.052), mean difference between W→S and GC-conservative synonymous average SNP frequency is 0.064 (SD=0.026), mean difference between W→S and S→W non-synonymous average SNP frequency is 0.041 (SD=0.029), mean difference between W→S and GC-conservative non-synonymous average SNP frequency is 0.047 (SD=0028).

To minimize risks of orientation errors, we removed columns of the alignment containing a least one CpG site. Removing them (i.e. on average 6.3% of the sites in primates, and 5.9% in Galloanserae) drastically reduced π_s for each species (almost two times on average) and also led to the reduction of the significance level of the previous pattern for all mutation categories, but did not change our conclusions regarding π_n/π_s (**Table S6**). However, the tendency for $\pi_{s[S\rightarrow W]}$ and $\pi_{s[W\rightarrow S]}$ to increase with GC3 (significantly for half of the species, both primates and birds), is far less present after masking CpG sites. SFS with and without CpG sites are shown in supplementary **Figures S8 to S73**.

Splitting the dataset in five bins instead of ten yielded qualitatively similar results (**Table S7**).

5. Influence of GC3 level on α, ω_a and ω_{na}

We estimated α, ω_a and ω_{na} on the whole sample of gene considering only GC-conservative mutations, S→W, W→S or all mutations at once using two different models for the DFE, namely GammaZero and GammaExpo (see Material & Methods). **Figure 4** shows that both $\alpha_{[W\rightarrow S]}$ and $\omega_{a[W\rightarrow S]}$ were higher than $\alpha_{[S\rightarrow W]}$ and $\omega_{a[S\rightarrow W]}$ in ten out of twelve species (binomial test p-value=0.038) (see also **Figure S5**). $\alpha_{[all]}$ and $\omega_{a\ [all]}$ were also higher than $\alpha_{[GC\text{-}conservative]}$ and $\omega_{a[GC\text{-}}$

$_{conservative]}$ in a majority of species, even if this was not significant (**Figure 4** and **Figure S5**). This indicates that gBGC could lead to an overestimation of the adaptive substitution rate. We checked that these results are robust to the number of individuals included in the analysis (**Figure S6**).

Splitting the dataset in ten bins of genes of even number of SNPs sorted according to their GC3 level, we estimated the correlation between GC3 and α, ω_a and ω_{na}. Our analysis did not allow to detect any significant effect of GC3 on the estimates of ω_a or α for any of the models (**Table 3, Table S8**). For W→S and GC-conservative mutations, the correlation coefficient between α and GC3 was positive in a majority of species, but the relationship between ω_a and GC3 was not consistent across species. The fact that $\omega_{a[GC\text{-}conservative]}$ is positively correlated to GC3 in nine species out of twelve might indicate that an increased recombination rate leads to a greater efficiency of positive selection, but the effect is tenuous. The relation between ω_{na} and GC3 was found to be negative in all species when considering all mutations types together, and negative in ten out of twelve species for GC-conservative mutations, which might indicate that an increased recombination rate leads to a greater efficiency of purifying selection. However, this analysis is limited by a lack of statistical power due to the splitting of the dataset in different bins of genes, resulting in a large sampling variance of estimates of α, ω_a and ω_{na}. Splitting the dataset in five bins yielded qualitatively similar results (**Table S9**).

DISCUSSION

Here we assessed the impact of recombination rate and gBGC on coding sequence evolution in two taxonomic groups, primates and Galloanserae, with contrasted recombination dynamics. We addressed this question by comparing estimates of dN/dS, π_n/π_s, α, ω_a and ω_{na} between bins of genes with different recombination rate and GC3, and by separately analyzing S→W, W→S and GC-conservative changes.

Recombination influences divergence and polymorphism in a roughly similar way in birds and primates

One of the most striking results we obtained is that all the measured variables, whether they are based on divergence data, polymorphism data, or both, were similarly influenced by recombination rate in the two taxonomic groups – despite some notable differences in the shape of the relationship between dN/dS and GC3 (see below). In particular, we showed for the first time that the dN/dS ratio

in primates decreases with increasing GC3 and r, a result previously reported in passerines (Bolívar et al. 2016) and that we confirm here in Galloanserae, despite the lack of significance of the signal. Previous studies in primates have indicated that gBGC promotes a local increase in the dN/dS ratio and can mislead the inference of positive selection (Galtier et al. 2009, Berglund et al. 2009, Ratnakumar et al. 2010, Kostka et al. 2012). Those studies focused on a small subset of genes *a priori* identified on the basis of their high dN/dS. Our results indicate that the positive effect of gBGC on dN/dS is only local/transient and that, in contrast, the global pattern is a negative relationship between recombination rate and dN/dS in primates. A significant negative correlation between dN/dS and the equilibrium GC-content has also been observed in a dataset of 17 nuclear protein-coding genes in 73 placental taxa (Lartillot 2012), consistent with our results. As for birds, our results are consistent with previous analyses in flycatchers (Bolívar et al. 2016).

That said, our analysis revealed some differences between the two taxonomic groups. Most importantly, the shape of the relationship between dN and dN/dS and GC3 differs between primates and Galloanserae (**Figure 1**). We observe that in Galloanserae, dN varies non-monotonically with GC3. dN decreases with GC3 at low GC3 values, but increases with GC3 at high GC3 values. dN/dS also shows a sharp decline at low GC3 values then plateaus. The initial decrease of dN and dN/dS seems to depict the effect of Hill-Robertson interference: in very low recombining regions, selection against deleterious mutations is poorly efficient. When recombination increases, selection becomes more efficient, until a recombination rate threshold is reached, above which interferences become negligible (Kliman and Hey 1993, Comeron and Kreitman 2002, Corbett-Detig et al. 2015). The increase in dN with GC3 above this threshold can be interpreted as reflecting the mutagenic effect of recombination, as it affects both $dN_{[W \rightarrow S]}$ and $dN_{[GC\text{-}conservative]}$ and is not perceptible in the dN/dS analysis (**Figure 2**).

Interestingly, the six species of primates show a different pattern, i.e., a gradual decrease of dN/dS with GC3. We suggest that this might be explained by the variation in time of recombination map in primates due to the presence of PRDM9. Let us assume, as discussed above, that HRI affects dN when the recombination rate is below some threshold (low-r state), but negligibly so when the recombination rate is above this threshold (high-r state). If r varies in time, then the time-averaged HRI effect for a given gene is expected to reflect the proportion of time spent by this gene in the low-r state as species were diverging. Said differently, we suggest that GC3 and dN/dS are expected to recapitulate the long-term effect of gBGC and HRI, respectively, which in primates vary continuously across genes due to the temporal dynamics of recombination rate.

316 Additionally, some species of primates show an increase of $dN_{[S\rightarrow W]}$ and $dS_{[S\rightarrow W]}$ with GC3 and r,
317 whereas this is not observed in birds (**Table 1** and **Figure 2**). This could reflect an effect of back
318 S→W mutations after gBGC has been turned off due to the shifting of recombination hotspot
319 location in primates – a process that might be less prevalent in birds due to the stability of the
320 recombination landscape. This is an interesting hypothesis that would deserve to be investigated
321 further.

322 Bolívar et al. (2016), finally, concluded that the impact of gBGC on the dN/dS ratio may be mainly
323 governed by the difference between the current GC-content and GC*. They suggest that gBGC
324 leads to a reduced dN/dS in high recombining regions in the flycatcher lineage due to current GC-
325 content being lower than GC*, and more so at synonymous than non-synonymous sites. Here we
326 found a decreasing GC3 in primates (GC3*-GC3~ -0.14 [-0.5;0.25] on average) and an increasing
327 one in Galloanserae (GC3*-GC3~ 0.05 (-0.31;0.28] on average), a result consistent with previous
328 studies (Duret et al. 2006, Nabholz et al. 2011, Weber et al. 2014). These differences may contribute
329 to the observed differences of behavior of the dN/dS ratio between primates and Galloanserae.

330 **Selection *vs.* gBGC: who wins?**

331 Here, we show that there is a stronger influence of gBGC on synonymous sites than on non-
332 synonymous sites. To explain this result we suggest that another evolutionary force may
333 compensate for the effects of gBGC on non-synonymous sites. One good candidate here is Hill-
334 Robertson interference, as suggested by our results concerning GC-conservative changes. Indeed,
335 $\pi_n/\pi_{s[GC\text{-}conservative]}$, as well as $dN/dS_{[GC\text{-}conservative]}$ and the non-adaptive substitution rate $\omega_{na[GC\text{-}conservative]}$,
336 decrease with GC3 in most species – although not always significantly. This is consistent with the
337 HRI hypothesis, and with previous studies empirically demonstrating a link between recombination
338 rate and genetic diversity in primates and birds (Spencer et al. 2006, Mugal, Nabholz, et al. 2013,
339 Corbett-Detig et al. 2015). The intensity of the decrease of π_n/π_s and ω_{na} with r is not the same for all
340 mutations types, though, suggesting that gBGC also plays a role here. Anyway, this result confirms
341 that in both taxonomic groups purifying selection is more efficient in highly recombining regions of
342 the genome.
343 Alternatively, Bolívar et al. (2016) suggested that the influence of gBGC on neutrally and selected
344 sites mainly depends on current synonymous and non-synonymous GC-content, and the distance to
345 their respective equilibrium (GC*). They found that in flycatchers the relationship between dN/dS
346 and recombination rate mainly reflect the greater distance between current and equilibrium GC-

content at 4-fold than at 0-fold degenerated sites (Bolívar et al. 2016). Using GC2 as a proxy for the non-synonymous GC-content, we found that in the two taxonomic groups current non-synonymous GC-content is far away from its equilibrium. In particular, in Galloanserae, we estimated a greater distance between GC and GC* at non-synonymous sites than at synonymous sites, suggesting that Bolívar's explanation does not apply here.

gBGC has been termed the "Achilles' heel" of the genome. It has been shown that an elevated recombination rate could locally decrease the efficiency of purifying selection due to the fixation of W→S deleterious mutations through gBGC (Dreszer et al. 2007, Duret and Galtier 2009). Glémin (2010) developed a population genetics model including gBGC and showed that the interaction between gBGC and selection has important consequences on load and inbreeding depression. Here we show that this effect, which can be locally strong in the vicinity of recently active recombination hotspots, does not dominate when one considers all genes and a longer time scale. We suggest that recombination influences synonymous and non-synonymous substitution rates via a combination of the effects of gBGC and HRI. The former is demonstrated by the distinctive patterns we report between W→S and S→W changes, and the latter by the existence of an effect of GC3 or r on $dN/dS_{[GC\text{-}conservative]}$.

Interestingly, we observe that the dN/dS ratio varies much between bins of genes and categories of changes – up to a factor of four (**Figure 2**). These differences are presumably independent of the selective constraints applying on the corresponding genes. This implies that controlling for recombination rate is of utmost importance when using dN/dS as a proxy for the extent of selective pressure acting on a sequence.

A mutagenic effect of recombination ?

We report a positive correlation between $dS_{[GC\text{-}conservative]}$ and both r and GC3, which reveals the existence of a mutagenic effect of recombination. A similar result was reported in flycatcher (Bolívar et al. 2016), but little discussed. In humans, several studies have previously reported such a phenomenon (Pratto et al. 2014, Arbeithuber et al. 2015, Smith et al. 2018). Pratto et al. (2014) specifically examined the mutation process around recombination hotspots by analyzing rare variants. They found that G↔A, C↔T, and G↔C mutations were enriched around recombination hotspots, G→A and C→T being the most frequent (Pratto et al. 2014). Arbeithuber et al. (2015) reported from sperm typing analysis that mutations appearing simultaneously with cross-over

events are enriched in S→W changes. More recently, Smith et al. (2018) analyzed father/mother/child trios and showed that both the S→W and the GC-conservative mutation rate are positively correlated with recombination rate, while results were less consistent across data set as far as the W→S rate was concerned.

In view of these recent results, it is quite plausible that the effect we detect on GC-conservative mutations is driven by a G↔C and A↔T mutagenic effect of recombination. Additionally, the fact that we do not detect a negative relationship between GC3 and $dS_{[S\rightarrow W]}$ or $dN_{[S\rightarrow W]}$ may be due to the fact that S→W mutations are submitted to an enhanced recombination-linked mutagenic effect as reported in Pratto et al. 2014 and Arbeithuber et al. 2015, which counterbalances the effect of gBGC. Finally, we cannot exclude the existence of an enhanced mutation associated with recombination regarding W→S mutations, as Pratto et al. (2014) detected such an effect. However, it seems weaker than the S→W mutation bias in view of the results of Arbeithuber et al. (2015) and Smith et al. (2018).

Influence of gBGC on the adaptive substitution rate

The comparison of ω_a and α, computed from the total set of genes for different mutation categories reveals that estimates of the adaptive substitution rate are lower for S→W mutations than for W→S ones, in agreement with the suggestion that gBGC mimics positive selection in increasing the fixation probability of neutral and slightly deleterious W→S mutations. Additionally, we found that ω_a and α were often higher when considering all mutations than when considering only GC-conservative ones, confirming that gBGC tends to entail an increase of the estimated adaptive substitution rate. This is in line with what was observed in great tits and zebra finch, where using only CG-conservative mutations led to a decreased estimate of α (Corcoran et al. 2017). Our analysis, however, did not allow to accurately validate the prediction that HRI leads to positive selection being less efficient in low recombining regions compared to highly recombining regions (whereas this was shown in a fungal pathogen (Grandaubert et al. 2018) and in *Drosophila* (Castellano et al. 2016)) - probably due to a lack of power.

Our results regarding the influence of recombination and gGC on ω_a and α may be explained by the fact that the estimation process is sensitive to some technical biases, in particular orientations errors (in this case, low frequency W→S SNPs mis-attributed as high frequency S→W SNPs and conversely). As some of the SFS in Galloanserae showed an unusual shape (*M. gallopavo*, P. *cristatus*, *N. meleagris*, *A. cygnoides*, **Figures S8** to **S73**), we took particular care to remove any

sources of such errors. We incorporated the so-called r_i's nuisance parameters (Eyre-Walker et al. 2006, Eyre-Walker and Keightley 2009), to be optimized along side with DFE parameters in the DFE-α method (see Methods). These parameters are intended to capture a wide range of effect that would distort the shape of SFS, including orientation errors, as soon as they influence similarly the synonymous and non-synonymous SFS. We also removed CpG sites from the sequences, as CpG hypermutability may make unfolded SFS very sensible to polarization errors. This did not significantly affect the results (**Figure S8** to **S73**), suggesting that the pattern we are reporting regarding ω_a and α reflects a real effect of gBGC on coding sequence evolution.

Conclusions

Our analysis revealed a substantial effect of recombination and gBGC on the rate of coding sequence evolution in primates and Galloanserae, with dN/dS varying by up to 4-fold between categories of genes and base changes irrespective of gene function. We report an increase in dS and a decrease in dN/dS with GC3 and recombination rate, which are particularly strong as far as W→S mutations are concerned, demonstrating a combined influence of gBGC and HRI. This pattern is reported in mammals as well as in birds, despite some differences in the dynamic of the influence of GC3 or r. This suggests that the presence/absence of PRDM9 is not as strong a predictor of the long-term evolutionary pattern of coding sequences as we hypothesized, but that it may still lead to differences in the dynamic of the impact of recombination on coding sequences between primates and birds. Overall, our analysis demonstrates a complex effect of recombination on molecular evolution, which should be appropriately taken into account when interpreting patterns of coding sequence variation among genes and genomes.

MATERIAL & METHODS

1. Sequence data

We used six species of primates and six species of Galloanserae for which we had more than 5 individuals: *Homo sapiens* (19 individuals), *Pan troglodytes* (20 individuals), *Papio anubis* (5 individuals), *Pongo abelii* (10 individuals), *Gorilla gorilla* (20 individuals), *Macaca mulatta* (19 individuals), and *Meleagris gallopavo* (10 individuals), *Phasianus colchicus* (10 individuals), P*avo cristatus* (10 individuals), *Numida meleagris* (10 individuals), *Anas platyrhynchos* (10 individuals), *Anser cygnoides* (10 individuals).

For primates reference genomes, assemblies and annotations files were downloaded from Ensembl (release 89). We kept only '*CDS*' reports in the annotations files, corresponding to coding exons, which were annotated with the automatic *ensembl* annotation pipeline, and the havana team for *Homo sapiens*. Raw genomic reads for each primate individuals were retrieved from various Bioproject of SRA (see **Table S1**). We used trimmomatic to remove Illumina adapters and trimmed low-quality reads (i.e. with an average base quality below 20), and kept only reads longer than 50bp.

For Galloanserae, we retrieved RNA-seq reads from SRA Bioproject PRJNA271731, generated in a previous study (Wright et al. 2015).We used trimmomatic to remove Illumina adapters as well as reads with a quality below 30. We constructed *de novo* transcriptome assemblies for each species following strategies B in (Cahais et al. 2012), using Abyss (Simpson et al. 2009) and Cap3 (Huang and Madan 1999). Open reading frames (ORFs) were predicted using the Trinity package (Grabherr et al. 2011). Contigs carrying ORF shorter than 150 bp were discarded.

2. SNP calling

Primates reads were mapped using Burrow Wheeler Aligner (BWA) software (Li and Durbin 2009) on the complete reference assembly (version 0.7.12-r1039). We filtered out hits with mapping quality below 20 and removed duplicates, and we extracted mapping hits corresponding to regions containing coding sequences according to the annotated reference assembly. This was done to avoid calling SNPs on the whole genome, which is heavily time consuming and useless in the present context. We called SNPs using a pipeline based on GATK (v3.8-0-ge9d80683). Roughly, this pipeline comprised two rounds of variant calling separated by a base quality score recalibration. Variant calling was first run on every individuals from every species using HaplotypeCaller (--emitRefConfidence GVCF --genotyping_mode DISCOVERY -hets 0.001). The variant callings from all individuals of a given species were then used to produce a joint genotype using GenotypeGVCFs. Indels in the resulting vcf files were then filtered out using vcftools. The distributions of various parameters associated with SNPs were then used to set several hard thresholds (i.e. Quality by Depth < 3.0; Fisher Strand > 10; Strand Odds Ratio > 3.0; MQRootMeanSquare < 50; MQRankSum < -0.5; ReadPosRankSum < -2.0) in order to detect putative SNP-calling errors using VariantFiltration. This erroneous SNPs were then used for base quality score recalibration of the previously created mapping files using BaseRecalibrator. These

mappings with re-calibrated quality scores were then used to re-call variants (HaplotypeCaller), to re-produce a joint genotype (GenotypeGVCFs, --allsites) and to re-set empirical hard thresholds (i.e. same values as above, except for Quality by Depth < 5.0). The obtained vcf files were converted to fasta files using custom python scripts while discarding exons found on both mitochondrial and sexual chromosomes and while filtering out additional SNPs. We removed SNPs with a too high coverage (thresholds were empirically set for each species), with a too low coverage (i.e. 10x per individual) and with a too low genotype quality per individual (i.e. less than 30).

For Galloanserae, filtered RNA-seq reads were mapped to predicted cDNAs with BWA (Li and Durbin 2009). Contigs with a per individual average coverage below x2,5 were discarded. Diploid genotypes were called according to the method described in Tsagkogeorga et al. 2012 (model M1) via a the software reads2snps. This software calls a genotype at each site with a minimum of 10 reads and calculates the posterior probability of each possible genotype in the maximum likelihood framework. Genotypes supported by a posterior probability higher than 95% are retained, otherwise missing data is called. We used version of the method which accounts for between-individual, within-species contamination as introduced in Ballenghien et al. (2017), using the -contam=0.1 option, which means assuming that up to 10% of the reads assigned to one specific sample may actually come from a distinct sample, and only validating genotypes robust to this source of uncertainty.

3. Orthology prediction

For primates, we extracted the 1-to-1 orthologous prediction of the six species from the OrthoMaM database (Ranwez et al. 2007, Douzery et al. 2014).

For Galloanserae, we translated the obtained CDS into proteins and predicted orthology via OrthoFinder (Emms and Kelly 2015) that uses a proteic BLAST (Basic Local Alignment Search Tool). For this specific step, we added coding sequences of the chicken *Gallus gallus*, extracted from the reference genome inEnsembl (release 89). Indeed, RNA-seq assemblies are very fragmented, and chicken CDS were on average longer than contigs assembled via the *de novo* transcriptome assemblies. Including long CDS allowed OrthoFinder to group several RNA-seq contigs of a same species into one orthogroup, allowing us to concatenate such contigs into one longer sequence after checking that they were not overlapping and thus improving the orthologous detection. We kept only orthogroups that included all species.

For both groups, we aligned the orthologous sequences via MACSE (Multiple Alignment for Coding SEquences (Ranwez et al. 2011).

4. Per gene recombination rate computation

We used the R package MareyMap (R version 3.4.3 (30/11/2017) to perform a Loess interpolation method (LOcally WEighted Scatterplot Smoothing) (Rezvoy et al. 2007) with a span of 0.2 on two recombination maps (one for *Homo sapiens* and one for *Gallus gallus*) to estimate the per gene recombination rate (r) by comparing the genetic map with the physical position of genes. We used the recombination map of *Homo sapiens* from the Phase 2 HapMap project (HapMap release 22, NCBI 36) that comprises over 3.1 millions SNPs *(The International HapMap Consortium Consortium 2007)*, and the recombination map of *Gallus gallus* that comprises 9.268 SNPs (Groenen et al. 2009).

5. Divergence and polymorphism statistics computation

For both taxonomic groups, we used the bppml program and a modified version of mapNH [http://biopp.univ-montp2.fr/forge/testnh] (Romiguier et al. 2012, Guéguen et al. 2013) to estimate the synonymous and non-synonymous substitution rate (dS, the number of synonymous substitutions per synonymous site and dN, the number of non-synonymous substitutions per non-synonymous site) per branch by substitution mapping under the Nielsen–Yang model (Nielsen and Yang 1998). We used the tree topologies presented in supplementary **Figure S7**.

We tested and compared both a model assuming a stationary base composition and a model assuming a non-stationary base composition. This was motivated by the results of Guéguen & Duret 2017 stating that estimates of dN, dS and dN/dS can be biased when using standard methods assuming sequence stationarity, this bias being influenced by the evolution of GC3 in particular (Guéguen and Duret 2017). We did so for three categories of substitution: W→S, S→W and GC-conservative.

For each sequence we estimated the non-synonymous and synonymous number of sites for each of those categories to normalize substitutions counts, using an in-house script that counts up mutational opportunities of each mentioned category of mutation under a neutral model assuming a transition-transversion ratio. The principle of this count is as follow: for each site, there are three possible alternative states which are examined. We estimate the probability to mutate to either of the three possible states using a ratio of transition over transversion parameter (estimated from the data). We then we add up those probabilities across sites, separating possible changes that are

synonymous from possible changes that are non-synonymous along the gene. When counting only W→S (or S→W or GC-conservative) sites, we use the same strategy but restrict the counts to the relevant alternative states.

We then computed dN, dS and dN/dS estimates for bins of genes defined according to GC3 level or per-gene recombination rate (ten bins of even size) by summing substitutions and number of sites across genes and then dividing the sum of substitutions by the sum of number of sites. Ninety-five percent confidence intervals were determined by bootstrapping genes (1000 replicates).
For each alignment, we estimated ancestral sequences at each node of the tree with the Bio ++-based SeqAncestor program (Guéguen et al. 2013). Ancestral sequences were then used to orientate mutations, so that we could then compute non-synonymous (π_n) and synonymous (π_s) nucleotide diversity and site frequency spectra for three bins of genes of even size and ten bins of genes of even number of SNPs, and each mutation category, via an in-house script. Ninety-five percent confidence intervals on π_n and π_s were determined by bootstrapping genes (100 replicates). We applied the same protocol after removing columns of the alignments that contains at least one CpG site.

6. Adaptive and non-adaptive substitution rate computation

We estimated α, ω_a and ω_{na} using the method of (Eyre-Walker and Keightley 2009) as implemented in Galtier 2016. It models the distribution of the fitness effect (DFE) of deleterious non-synonymous mutations as a negative Gamma distribution, which is fitted to the synonymous and non-synonymous site frequency spectra (SFS) computed for a set of genes.
We accounted for recent effects (demographic or other) that could distort the SFS relative to the neutral expectation in an equilibrium Wright–Fisher population by adjusting the so-called r_i's nuisance parameters alongside with the DFE parameters. They are adjusted for each allele frequency class in the SFS, and multiply both the synonymous and the non-synonymous expected number of SNP (Eyre-Walker et al. 2006).

We computed α, ω_a and ω_{na} for each terminal branches of the tree for primates and Galloanserae, for ten bins of genes of even number of SNPs sorted depending on their GC3, and each category of mutation and substitution (W→S, S→W and GC-conservative). We tested two different models to fit the DFE that are described in Galtier 2016. Briefly, the GammaZero models a negative DFE only, as a Gamma distribution. The GammaExpo model contains a negative Gamma DFE as well as

a positive DFE modeled as an exponential. Confidence intervals correspond to the maximum likelihood confidence intervals computed during the optimization step using the Newton-Raphson method (defined as values of α and ω_a for which the log-likelihood was within two units of its maximum).

Acknowledgments

We thank Lurent Duret for his pieces of advice, and Paul Simion, Yoann Anselmetti and Thibault Leroy for their help with bio-informatics.

Funding information

This work was supported by Agence Nationale de la recherche grant no. ANR-15-CE12-0010 'DarkSideOfRecombination'.

Figures legends:

Figure 1: dN, dS and dN/dS ratio against GC3 for each species for all substitutions taken together. Statistics are estimated under a model assuming base composition non-stationarity.

Figure 2: dN, dS and dN/dS ratio against GC3 for each species and each type of substitutions (W→S, S→W and GC-conservative substitutions). Statistics are estimated under a model assuming base composition non-stationarity.

Figure 3: π_n/π_s ratio against GC3 for each species (A.: primates, B.: Galloanserae) and each type of mutations (W→S, S→W and GC-conservative mutations) and average non-synonymous (C.) and synonymous (D.) allele frequency for each species for W→S, S→W and GC-conervative SNPs (statistics estimated without masking CpG sites).

Figure 4: α and ω_a estimates for each species and each type of mutations (all mutations, W→S, S→W and GC-conservative mutations) using all genes.
Statistics are obtained using the model "GammaZero".

Table 1: Spearman correlation coefficients between GC3 and divergence estimates obtained with a model assuming non-stationarity.
Significance levels are showed with * (* p-value< 0.05, ** p-value < 0.01) before the FDR correction (False Discovery Rate), and with two shades of red (if the correlation is positive) or green (if the correlation in negative) after FDR correction (light :p-value< 0.05, dark : p-value< 0.01).

Table 2: Spearman correlation coefficients between GC3 and π_n, π_s, π_n/π_s, obtained without masking CpG sites.

Significance levels are showed with * (* p-value< 0.05, ** p-value < 0.01) before the FDR correction, and with two shades of red (if the correlation is positive) or green (if the correlation in negative) after FDR correction (light :p-value< 0.05, dark : p-value< 0.01).

Table 3: Spearman correlation coefficients between GC3 and α, ω_a and ω_{na} estimates obtained obtained using the model "GammaZero".

Significance levels are showed with * (* p-value< 0.05, ** p-value < 0.01) before the FDR correction, and with two shades of red (if the correlation is positive) or green (if the correlation in negative) after FDR correction (light:p-value< 0.05, dark : p-value< 0.01).

Supplementary material:

Table S1: Details of the Bioprojects used in this study to retrieve reads data.

Table S2: Numbers of SNPs of different category in the different species without masking CpG sites.

Table S3: Spearman correlation coefficients between r and divergence estimates obtained with a model assuming non-stationarity.

Significance levels are showed with * (* p-value< 0.05, ** p-value < 0.01) before the FDR correction, and with two shades of red (if the correlation is positive) or green (if the correlation in negative) after FDR correction (light:p-value< 0.05, dark : p-value< 0.01).

Table S4 : Spearman correlation coefficients between GC3 and divergence estimates obtained with a model assuming stationarity.

Significance levels are showed with * (* p-value< 0.05, ** p-value < 0.01) before the FDR correction, and with two shades of red (if the correlation is positive) or green (if the correlation in negative) after FDR correction (light:p-value< 0.05, dark: p-value< 0.01).

Table S5: Spearman correlation coefficients between r and divergence estimates obtained with a model assuming stationarity.

Significance levels are showed with * (* p-value< 0.05, ** p-value < 0.01) before the FDR correction, and with two shades of red (if the correlation is positive) or green (if the correlation in negative) after FDR correction (light:p-value< 0.05, dark : p-value< 0.01).

Table S6: Spearman correlation coefficients between GC3 and π_n, π_s, π_n/π_s obtained after masking CpG sites from the alignments.

Significance levels are showed with * (* p-value< 0.05, ** p-value < 0.01) before the FDR correction, and with two shades of red (if the correlation is positive) or green (if the correlation in negative) after FDR correction (light:p-value< 0.05, dark : p-value< 0.01).

Table S7: Spearman correlation coefficients between GC3 and π_n, π_s, π_n/π_s.

The dataset was split in five bins of increasing GC3 and equal number of SNPs.

Significance levels are showed with * (* p-value< 0.05, ** p-value < 0.01) before the FDR correction, and with two shades of red (if the correlation is positive) or green (if the correlation in negative) after FDR correction (light:p-value< 0.05, dark : p-value< 0.01).

Table S8: Spearmancorrelation coefficients between GC3 and α, ω_a and ω_{na} estimates for the "GammaExpo" model.

Significance levels are showed with * (* p-value< 0.05, ** p-value < 0.01) before the FDR correction, and with two shades of red (if the correlation is positive) or green (if the correlation in negative) after FDR correction (light:p-value< 0.05, dark : p-value< 0.01).

Table S9: Spearman correlation coefficients between GC3 and α, ω_a and ω_{na} estimates for the "GammaZero" model.

The dataset was split in five bins of increasing GC3 and equal number of SNPs.

Significance levels are showed with * (* p-value< 0.05, ** p-value < 0.01) before the FDR correction, and with two shades of red (if the correlation is positive) or green (if the correlation in negative) after FDR correction (light:p-value< 0.05, dark : p-value< 0.01).

Figure S1: Spearman correlation between GC3 and r obtained with for *G. gallus* and *H. sapiens* respectively.

Figure S2: dN, dS and dN/dS ratio against r for each species for all substitution type taken together. Statistics are estimated under a model assuming base composition non-stationarity (above: primates, below: Galloanserae).

Figure S3: dN, dS and dN/dS ratio against GC3 for each species for all substitution type taken together. Statistics are estimated under a model assuming base composition stationarity (above: primates, below: Galloanserae).

Figure S4: dN, dSand dN/dS ratio against GC3 for each species and each type of substitutions ($W \rightarrow S$, $S \rightarrow W$ and GC-conservative substitution). Statistics are estimated under a model assuming base composition stationarity (above: primates, below: Galloanserae).

Figure S5: α and ω_a estimates for each species and each type of mutations (all mutations, $W \rightarrow S$, $S \rightarrow W$ and GC-conservative) using all genes.

Statistics are obtained using the model "GammaExpo".

Figure S6: α, ω_a and ω_{na} estimates for each species and each type of mutations (all mutations, W→S, S→W and GC-conservative) using all genes and using only five randomly chosen individuals of each species.

Statistics are obtained using the model “GammaZero”.

Figure S7: Tree topologies used to map substitution in primates (A) and Galloanserae (B).

Figure S8-73: Synonymous and non-synonymous site frequency spectra for all species obtained without or with a masking of CpG sites.

Bibliography :

Arbeithuber B, Betancourt AJ, Ebner T, Tiemann-Boege I. 2015. Crossovers are associated with mutation and biased gene conversion at recombination hotspots. Proc. Natl. Acad. Sci. U. S. A. 112:2109–2114.

Baker CL, Kajita S, Walker M, Saxl RL, Raghupathy N, Choi K, Petkov PM, Paigen K. 2015. PRDM9 drives evolutionary erosion of hotspots in Mus musculus through haplotype-specific initiation of meiotic recombination. PLoS Genet. 11:e1004916.

Baker Z, Schumer M, Haba Y, Bashkirova L, Holland C, Rosenthal GG, Przeworski M. 2017. Repeated losses of PRDM9-directed recombination despite the conservation of PRDM9 across vertebrates. Elife 6.

Ballenghien M, Faivre N, Galtier N. 2017. Patterns of cross-contamination in a multispecies population genomic project: detection, quantification, impact, and solutions. BMC Biol. [Internet] 15. Available from: http://bmcbiol.biomedcentral.com/articles/10.1186/s12915-017-0366-6

Baudat F, Buard J, Grey C, Fledel-Alon A, Ober C, Przeworski M, Coop G, De Massy B. 2010. PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. Science 327:836–840.

Berg IL, Neumann R, Sarbajna S, Odenthal-Hesse L, Butler NJ, Jeffreys AJ. 2011. Variants of the protein PRDM9 differentially regulate a set of human meiotic recombination hotspots highly active in African populations. Proc. Natl. Acad. Sci. 108:12378–12383.

Berglund J, Pollard KS, Webster MT. 2009. Hotspots of biased nucleotide substitutions in human genes. PLoS Biol. 7:e1000026.

Bolívar P, Mugal CF, Nater A, Ellegren H. 2016. Recombination Rate Variation Modulates Gene Sequence Evolution Mainly via GC-Biased Gene Conversion, Not Hill–Robertson Interference, in an Avian System. Mol. Biol. Evol. 33:216–227.

Brick K, Smagulova F, Khil P, Camerini-Otero RD, Petukhova GV. 2012. Genetic recombination is directed away from functional genomic elements in mice. Nature 485:642–645.

Cahais V, Gayral P, Tsagkogeorga G, Melo-Ferreira J, Ballenghien M, Weinert L, Chiari Y, Belkhir K, Ranwez V, Galtier N. 2012. Reference-free transcriptome assembly in non-model animals from next-generation sequencing data. Mol. Ecol. Resour. 12:834–845.

Castellano, D., Coronado-Zamora, M., Campos, J. L., Barbadilla, A., & Eyre-Walker, A. 2015. Adaptive evolution is substantially impeded by Hill–Robertson interference in *Drosophila*. Mol. Biol. Evol. 33:442-455.

Comeron JM, Kreitman M. 2002 Population, Evolutionary and Genomic Consequences of Interference Selection. Genetics 161 :389-410.

Consortium ICGS. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature 432:695.

Consortium IH. 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851.

Corbett-Detig RB, Hartl DL, Sackton TB. 2015. Natural Selection Constrains Neutral Diversity across A Wide Range of Species. PLOS Biol. 13:e1002112.

Corcoran P, Gossmann TI, Barton HJ, Slate J, Zeng K. 2017. Determinants of the Efficacy of Natural Selection on Coding and Noncoding Variability in Two Passerine Species. Genome Biol. Evol. 9:2987–3007.

Douzery EJ, Scornavacca C, Romiguier J, Belkhir K, Galtier N, Delsuc F, Ranwez V. 2014. OrthoMaM v8: a database of orthologous exons and coding sequences for comparative genomics in mammals. Mol. Biol. Evol. 31:1923–1928.

Dreszer TR, Wall GD, Haussler D, Pollard KS. 2007. Biased clustered substitutions in the human genome: The footprints of male-driven biased gene conversion. Genome Res. 17:1420–1430.

Duret L, Eyre-Walker A, Galtier N. 2006. A new perspective on isochore evolution. Gene 385:71–74.

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

Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 16:157.

Eyre-Walker A. 1993. Recombination and mammalian genome evolution. Proc. R. Soc. Lond. B 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, Keightley PD. 2009. Estimating the Rate of Adaptive Molecular Evolution in the Presence of Slightly Deleterious Mutations and Population Size Change. Mol. Biol. Evol. 26:2097–2108.

Eyre-Walker A, Woolfit M, Phelps T. 2006. The Distribution of Fitness Effects of New Deleterious Amino Acid Mutations in Humans. Genetics 173:891–900.

Galtier N. 2016. Adaptive Protein Evolution in Animals and the Effective Population Size Hypothesis. PLOS Genet. 12:e1005774.

Galtier N, Duret L. 2007. Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. TRENDS Genet. 23:273–277.

Galtier N, Duret L, Glémin S, Ranwez V. 2009. GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. Trends Genet. 25:1–5.

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

Galtier N, Roux C, Rousselle M, Romiguier J, Figuet E, Glémin S, Bierne N, Duret L. Codon usage bias in animals: disentangling the effects of natural selection, effective population size and GC-biased gene conversion. Mol. Biol. Evol. [Internet]. Available from: https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msy015/4829954

Gillespie JH. 1994. The causes of molecular evolution. Oxford University Press on Demand

Glémin S. 2010. Surprising fitness consequences of GC-biased gene conversion: I. Mutation load and inbreeding depression. Genetics 185:939–959.

Glémin S, Arndt PF, Messer PW, Petrov D, Galtier N, Duret L. 2015. Quantification of GC-biased gene conversion in the human genome. Genome Res. 25:1215–1228.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29:644.

Grandaubert J, Dutheil JY, Stukenbrock EH. 2018. The genomic determinants of adaptive evolution in a fungal pathogen. bioRxiv. 1:176727.

Groenen MA, Wahlberg P, Foglio M, Cheng HH, Megens H-J, Crooijmans RP, Besnier F, Lathrop M, Muir WM, Wong GK-S. 2009. A high-density SNP-based linkage map of the chicken genome reveals sequence features correlated with recombination rate. Genome Res. 19:510–519.

Guéguen L, Duret L. 2017. Unbiased Estimate Of Synonymous And Non-Synonymous Substitution Rates With Non-Stationary Base Composition. bioRxiv:124925.

Guéguen L, Gaillard S, Boussau B, Gouy M, Groussin M, Rochette NC, Bigot T, Fournier D, Pouyet F, Cahais V. 2013. Bio++: efficient extensible libraries and tools for computational molecular evolution. Mol. Biol. Evol. 30:1745–1750.

Halldorsson BV, Hardarson MT, Kehr B, Styrkarsdottir U, Gylfason A, Thorleifsson G, Zink F, Jonasdottir Adalbjorg, Jonasdottir Aslaug, Sulem P. 2016. The rate of meiotic gene conversion varies by sex and age. Nat. Genet. 48:1377.

Hill WG, Robertson A. 1966. The effect of linkage on limits to artificial selection. Genet. Res. 8:269–294.

Huang X, Madan A. 1999. CAP3: A DNA sequence assembly program. Genome Res. 9:868–877.

Kawakami T, Mugal CF, Suh A, Nater A, Burri R, Smeds L, Ellegren H. 2017. Whole-genome patterns of linkage disequilibrium across flycatcher populations clarify the causes and consequences of fine-scale recombination rate variation in birds. Mol. Ecol. 26:4158–4172.

Keightley PD, Eyre-Walker A. 2007. Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. Genetics 177:2251–2261.

Keith N, Tucker AE, Jackson CE, Sung W, Lledó JIL, Schrider DR, Schaack S, Dudycha JL, Ackerman M, Younge AJ. 2016. High mutational rates of large-scale duplication and deletion in Daphnia pulex. Genome Res. 26:60–69.

Kimura M. 1983. The neutral theory of molecular evolution. Cambridge University Press.

Kliman RM, Hey J. 1993. Reduced natural selection associated with low recombination in Drosophila melanogaster. Mol Biol Evol 10:1239–1258.

Kostka D, Hubisz MJ, Siepel A, Pollard KS. 2012. The role of GC-biased gene conversion in shaping the fastest evolving regions of the human genome. Mol. Biol. Evol. 29:1047–1057.

Lachance J, Tishkoff SA. 2014. Biased gene conversion skews allele frequencies in human populations, increasing the disease burden of recessive alleles. Am. J. Hum. Genet. 95:408–420.

Lartillot N. 2012. Interaction between selection and biased gene conversion in mammalian protein-coding sequence evolution revealed by a phylogenetic covariance analysis. Mol. Biol. Evol. 30:356–368.

Latrille T, Duret L, Lartillot N. 2017. The Red Queen model of recombination hot-spot evolution: a theoretical investigation. Phil Trans R Soc B 372:20160463.

Lesecque Y, Glémin S, Lartillot N, Mouchiroud D, Duret L. 2014. The red queen model of recombination hotspots evolution in the light of archaic and modern human genomes. PLoS Genet. 10:e1004790.

Lesecque Y, Mouchiroud D, Duret L. 2013. GC-Biased Gene Conversion in Yeast Is Specifically Associated with Crossovers: Molecular Mechanisms and Evolutionary Significance. Mol. Biol. Evol. 30:1409–1419.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25:1754–1760.

Mancera E, Bourgon R, Brozzi A, Huber W, Steinmetz LM. 2008. High-resolution mapping of meiotic crossovers and non-crossovers in yeast. Nature 454:479–485.

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.

Mugal CF, Arndt PF, Ellegren H. 2013. Twisted Signatures of GC-Biased Gene Conversion Embedded in an Evolutionary Stable Karyotype. Mol. Biol. Evol. 30:1700–1712.

Mugal CF, Nabholz B, Ellegren H. 2013. Genome-wide analysis in chicken reveals that local levels of genetic diversity are mainly governed by the rate of recombination. BMC Genomics 14:86.

Murray GG, Soares AE, Novak BJ, Schaefer NK, Cahill JA, Baker AJ, Demboski JR, Doll A, Da Fonseca RR, Fulton TL. 2017. Natural selection shaped the rise and fall of passenger pigeon genomic diversity. Science 358:951–954.

Myers S, Bowden R, Tumian A, Bontrop RE, Freeman C, MacFie TS, McVean G, Donnelly P. 2010. Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. Science 327:876–879.

Nabholz B, Künstner A, Wang R, Jarvis ED, Ellegren H. 2011. Dynamic Evolution of Base Composition: Causes and Consequences in Avian Phylogenomics. Mol. Biol. Evol. 28:2197–2210.

Necşulea A, Popa A, Cooper DN, Stenson PD, Mouchiroud D, Gautier C, Duret L. 2011. Meiotic recombination favors the spreading of deleterious mutations in human populations. Hum. Mutat. 32:198–206.

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.

Oliver PL, Goodstadt L, Bayes JJ, Birtle Z, Roach KC, Phadnis N, Beatson SA, Lunter G, Malik HS, Ponting CP. 2009. Accelerated evolution of the Prdm9 speciation gene across diverse metazoan taxa. PLoS Genet. 5:e1000753.

Parvanov ED, Petkov PM, Paigen K. 2010. Prdm9 controls activation of mammalian recombination hotspots. Science 327:835–835.

Perry J, Ashworth A. 1999. Evolutionary rate of a gene affected by chromosomal position. Curr. Biol. 9:987-S3.

Pessia E, Popa A, Mousset S, Rezvoy C, Duret L, Marais GAB. 2012. Evidence for Widespread GC-biased Gene Conversion in Eukaryotes. Genome Biol. Evol. 4:675–682.

Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, Veeramah KR, Woerner AE, O'Connor TD, Santpere G. 2013. Great ape genetic diversity and population history. Nature 499:471.

Ranwez V, Delsuc F, Ranwez S, Belkhir K, Tilak M-K, Douzery EJ. 2007. OrthoMaM: a database of orthologous genomic markers for placental mammal phylogenetics. BMC Evol. Biol. 7:241.

Ranwez V, Harispe S, Delsuc F, Douzery EJ. 2011. MACSE: Multiple Alignment of Coding SEquences accounting for frameshifts and stop codons. PloS One 6:e22594.

Ratnakumar A, Mousset S, Glémin S, Berglund J, Galtier N, Duret L, Webster MT. 2010. Detecting positive selection within genomes: the problem of biased gene conversion. Philos. Trans. R. Soc. B Biol. Sci. 365:2571–2580.

Rezvoy C, Charif D, Guéguen L, Marais GA. 2007. MareyMap: an R-based tool with graphical interface for estimating recombination rates. Bioinformatics 23:2188–2189.

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

Sandor C, Li W, Coppieters W, Druet T, Charlier C, Georges M. 2012. Genetic variants in REC8, RNF212, and PRDM9 influence male recombination in cattle. PLoS Genet. 8:e1002854.

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I. 2009. ABySS: a parallel assembler for short read sequence data. Genome Res. 19:1117–1123.

Singhal S, Leffler EM, Sannareddy K, Turner I, Venn O, Hooper DM, Strand AI, Li Q, Raney B, Balakrishnan CN. 2015. Stable recombination hotspots in birds. Science 350:928–932.

Smeds L, Mugal CF, Qvarnström A, Ellegren H. 2016. High-resolution mapping of crossover and non-crossover recombination events by whole-genome re-sequencing of an avian pedigree. PLoS Genet. 12:e1006044.

Smith NGC, Eyre-Walker A. 2002. Adaptive protein evolution in *Drosophila*. Nature 415:1022–1024.

Smith TCA, Arndt PF, Eyre-Walker A. 2018. Large scale variation in the rate of germ-line de novo mutation, base composition, divergence and diversity in humans. PLOS Genet. 14:e1007254.

Spencer CC, Deloukas P, Hunt S, Mullikin J, Myers S, Silverman B, Donnelly P, Bentley D, McVean G. 2006. The influence of recombination on human genetic diversity. PLoS Genet. 2:e148.

Stevison LS, Woerner AE, Kidd JM, Kelley JL, Veeramah KR, McManus KF, Bustamante CD, Hammer MF, Wall JD. 2016. The Time Scale of Recombination Rate Evolution in Great Apes. Mol. Biol. Evol. 33:928–945.

Tataru P, Mollion M, Glémin S, Bataillon T. 2017. Inference of Distribution of Fitness Effects and Proportion of Adaptive Substitutions from Polymorphism Data. Genetics 207:1103–1119.

Teixeira JC, de Filippo C, Weihmann A, Meneu JR, Racimo F, Dannemann M, Nickel B, Fischer A, Halbwax M, Andre C. 2015. Long-term balancing selection in LAD1 maintains a missense trans-species polymorphism in humans, chimpanzees, and bonobos. Mol. Biol. Evol. 32:1186–1196.

Tsagkogeorga G, Cahais V, Galtier N. 2012. The population genomics of a fast evolver: high levels of diversity, functional constraint, and molecular adaptation in the tunicate Ciona intestinalis. Genome Biol. Evol. 4:852–861.

Weber CC, Boussau B, Romiguier J, Jarvis ED, Ellegren H. 2014. Evidence for GC-biased gene conversion as a driver of between-lineage differences in avian base composition. Genome Biol. 15:549.

Webster MT, Axelsson E, Ellegren H. 2006. Strong Regional Biases in Nucleotide Substitution in the Chicken Genome. Mol. Biol. Evol. 23:1203–1216.

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

Williams AL, Genovese G, Dyer T, Altemose N, Truax K, Jun G, Patterson N, Myers SR, Curran JE, Duggirala R. 2015. Non-crossover gene conversions show strong GC bias and unexpected clustering in humans. Elife 4.

Wright AE, Harrison PW, Zimmer F, Montgomery SH, Pointer MA, Mank JE. 2015. Variation in promiscuity and sexual selection drives avian rate of Faster-Z evolution. Mol. Ecol. 24:1218–1235.

Xue C, Raveendran M, Harris RA, Fawcett GL, Liu X, White S, Dahdouli M, Deiros DR, Below JE, Salerno W. 2016. The population genomics of rhesus macaques (Macaca mulatta) based on whole-genome sequences. Genome Res. 26:1651–1662.

Species	dN				dS				dN/dS			
	all	WS	SW	GC-conservative	all	WS	SW	GC-conservative	all	WS	SW	GC-conservative
M. mulatta	0.867**	0.915**	0.115	0.952**	1**	1**	1**	0.952**	-0.952**	-0.988**	-0.769*	-0.855**
H. sapiens	0.879**	0.139	0.818**	0.721*	0.988**	0.976**	0.964**	0.988**	-0.709*	-0.952**	0.188	-0.721*
G. gorilla	-0.103	-0.515	0.030	-0.067	0.939**	0.891**	0.539	0.903**	-1**	-0.976**	-0.539	-0.939**
P. troglodytes	0.927**	0.818**	0.939**	0.867**	1**	0.988**	0.891**	0.988**	-0.564	-0.939**	0.648*	-0.636
P. anubis	-0.467	-0.612	-0.636	0.091	0.867**	0.952**	0.661*	0.769*	-1**	-1**	-0.879**	-0.939**
P. abelii	0.830**	-0.356	0.830**	0.891**	1**	1**	1**	1**	-1**	-1**	-0.2	-0.939**
M. gallopavo	-0.369	-0.491	-0.042	-0.164	0.479	0.794**	-0.285	0.297	-0.564	-0.903**	0.309	-0.6
N. meleagris	-0.127	-0.806**	0.115	0.103	0.636	0.952**	-0.297	0.6	-0.430	-0.927**	0.236	-0.467
P. cristatus	-0.236	-0.648*	0.042	0.467	0.369	0.964**	0.552	0.769*	-0.273	-0.927**	-0.176	-0.624
P. colchicus	0.055	-0.212	0.321	0.079	0.745*	0.891**	0.345	0.455	-0.491	-0.806**	0.067	-0.297
A. cygnoides	0.588	0.624	0.467	0.673*	0.879**	0.988**	0.709	0.964**	-0.479	-0.939**	-0.151	-0.891**
A. platyrhynchos	0.527	0.479	0.503	0.612	0.891**	0.976**	0.333	0.927**	-0.527	-0.964**	0.055	-0.927**

Table 1: Spearman correlation coefficients between GC3 and divergence estimates obtained with a model assuming non-stationarity.

Significance levels are showed with * (* p-value< 0.05, ** p-value < 0.01) before the FDR correction (False Discovery Rate), and with two shades of red (if the correlation is positive) or green (if the correlation in negative) after FDR correction (light:p-value< 0.05, dark : p-value< 0.01).

Species	π_n				π_s				π_n/π_s			
	all	WS	SW	GC-conservative	all	WS	SW	GC-conservative	all	WS	SW	GC-conservative
M. mulatta	0,22	-0.34	0.34	0.23	0,90**	0.842**	0.6	0.70*	-0,77*	-0.78*	-0.22	-0.36
H. sapiens	0,61	0.50	0.69*	0.13	0,93**	0.38	0.83**	0.73*	-0,69*	-0.01	-0.45	-0.45
G. gorilla	0,18	0.12	0.12	0.18	-0,43	-0.57	0.28	-0.01	0,35	0.39	-0.054	0.2
P. troglodytes	-0,63	-0.79**	0.09	-0.39	0,91**	0.56	0.64*	-0.41	-0,96**	-0.86**	-0.66*	0.24
P. anubis	-0,5	-0.83**	0.06	0.2	0,93**	0.50	0.64*	0.57	-0,84**	-0.80**	-0.43	-0.29
P. abelii	0,63	0.15	0.68*	0.442	0,93**	0.47	0.93**	0.90**	-0,72*	-0.44	-0.44	-0.76*
M. gallopavo	-0,1	-0.78*	0.17	-0.55	0,83**	0.90**	0.24	-0.32	-0,75*	-0.93**	0.11	-0.33
N. meleagris	0,71*	-0.44	0.91**	0.47	0,96**	0.96**	0.92**	0.95**	-0,85**	-0.89**	-0.46	-0.61
P. cristatus	-0,12	-0.44	0.05	-0.45	0,86**	0.87**	0.90**	0.51	-0,88**	-0.83**	-0.55	-0.81**
P. colchicus	-0,7*	-0.90**	0.33	-0.76*	0,90**	0.63	0.73*	-0.56	-0.96**	-0.91**	-0.26	-0.56
A. cygnoides	0,34	-0.33	0.52	-0.22	0,91**	0.96**	0.90**	0.80**	-0,81**	-0.85**	-0.15	-0.78*
A. platyrhynchos	0,87**	-0.31	0.86**	0.52	0,89**	0.92**	0.89**	0.96**	-0,95**	-0.92**	-0.80**	-0.89**

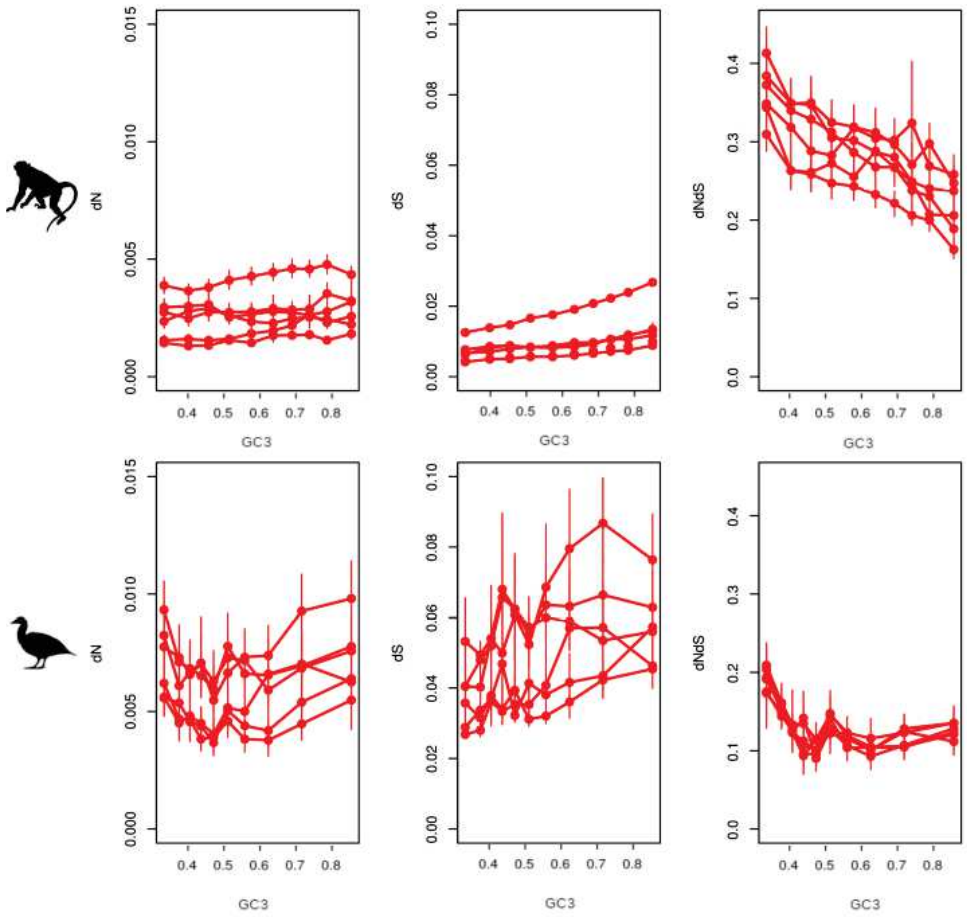
Table 2: Spearman correlation coefficients between GC3 and π_n, π_s, π_n/π_s, obtained without masking CpG sites.

Significance levels are showed with * (* p-value< 0.05, ** p-value < 0.01) before the FDR correction, and with two shades of red after FDR correction (light :p-value< 0.05, dark : p-value< 0.01).

Species	α				ω_a				ω_{na}			
	all	WS	SW	GC-conservative	all	WS	SW	GC-conservative	all	WS	SW	GC-conservative
M. mulatta	0,41	0.24	0.0060	0.490	0,2	-0.24	-0.030	0.466	-0,27	-0.21	0.0424	-0.46
H. sapiens	0,28	0.32	0.35	0.28	0,33	0.24	0.357	0.381	-0,2	-0.34	-0.21	-0.38
G. gorilla	-0,06	0.12	0.52	0.042	-16	-0.090	0.49	0.17	-0,21	-0.054	-0.50	-0.17
P. troglodytes	0,38	0.393	0.381	0.35	0,6	0.187	0.49	0.35	-0,41	-0.33	-0.15	-0.35
P. anubis	0,09	0.0424	0.22	0.57	-0,08	0.22	0.187	0.57	-0,3	-0.15	-0.33	-0.57
P. abelii	0,26	0.22	0.32	-0.078	0,23	0.15	0.27	-0.29	-0,41	-0.2	-0.39	0.29
M. gallopavo	0,45	-0.20	-0.031	0.082	0,1	-0.15	0.0060	-0.28	-0,5	0.012	-0.11	-0.096
N. meleagris	0,03	0.143	-0.14	-0.13	-0,16	0.078	-0.30	-0.16	-0,13	0.056	0.14	0.093
P. cristatus	0,52	0.13	-0.032	0.30	0,62	-0.066	-0.22	0.22	-0,70*	-0.10	0.0064	-0.27
P. colchicus	0,04	0.20	-0.18	0.079	-0,39	-0.06	-0.16	0.030	-0,32	-0.27	0.018	-0.06
A. cygnoides	0,03	-0.055	-0.28	0.10	-0,18	0.006	-0.29	0.13	-0,19	-0.11	-0.10	-0.12
A. platyrhynchos	0,04	0.06	-0.03	0.57	-0,24	-0.53	-0.16	0.57	-0,56	-0.38	0.019	-0.57

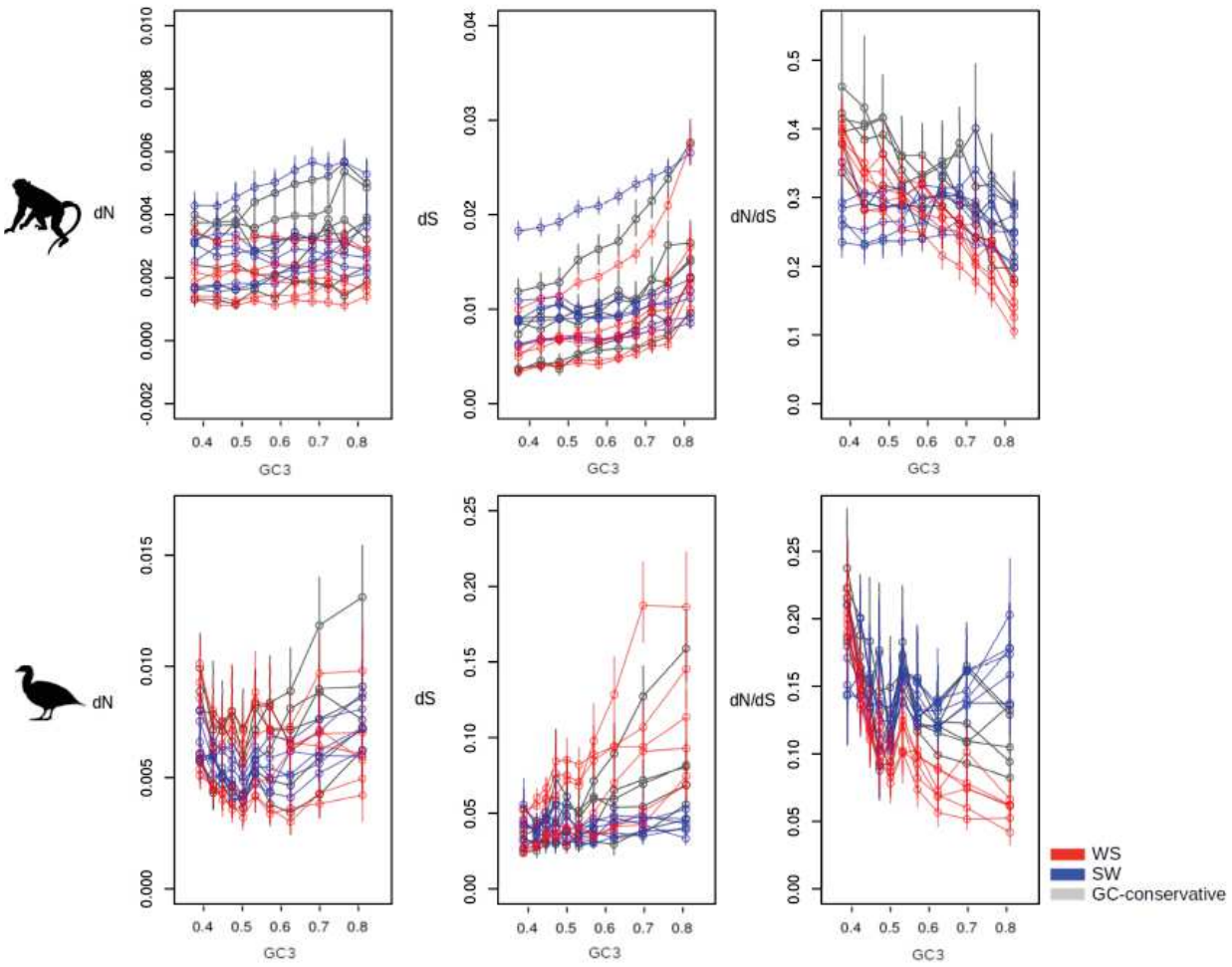
Table 3: Spearman correlation coefficients between GC3 and α, ω_a and ω_{na} estimates obtained without masking CpG sites from the alignments.

α, ω_a and ω_{na} are obtained using the model GammaZero. Significance levels are showed with * (* p-value< 0.05, ** p-value < 0.01) before the FDR correction, and with two shades of red after FDR correction (light:p-value< 0.05, dark : p-value< 0.01).

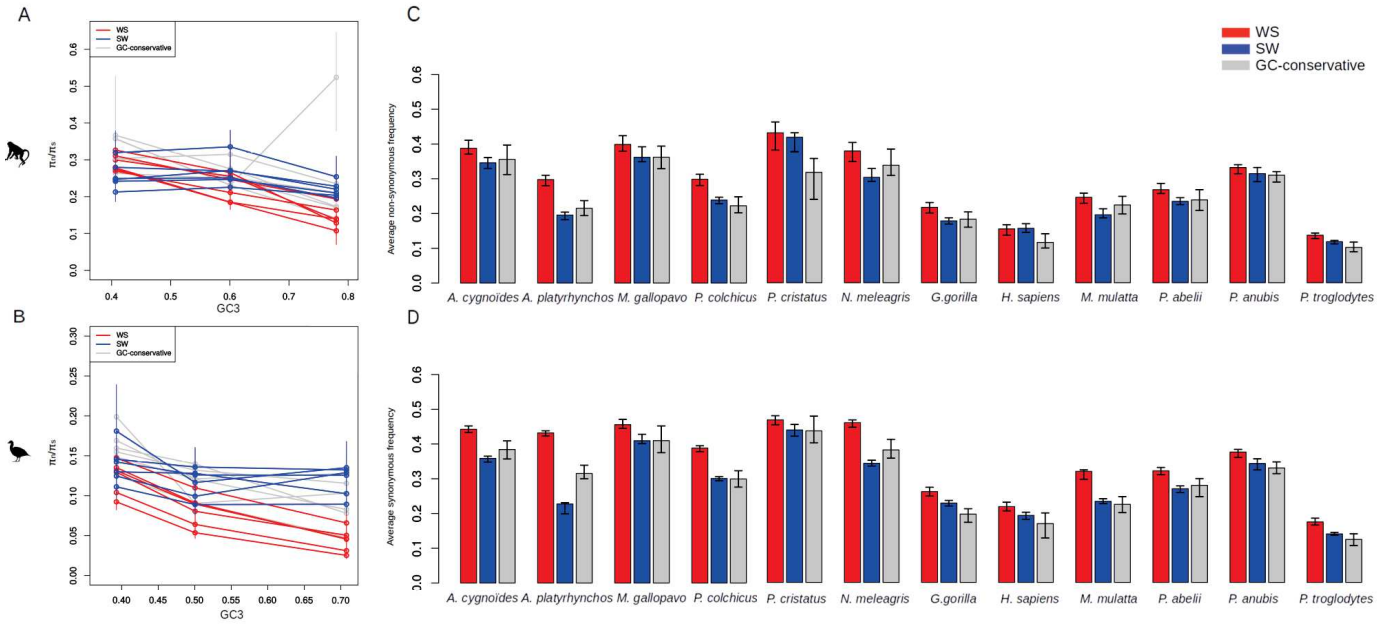


673 **Figure 1:** dN, dS and dN/dS ratio against GC3 for each species for all substitutions taken together.

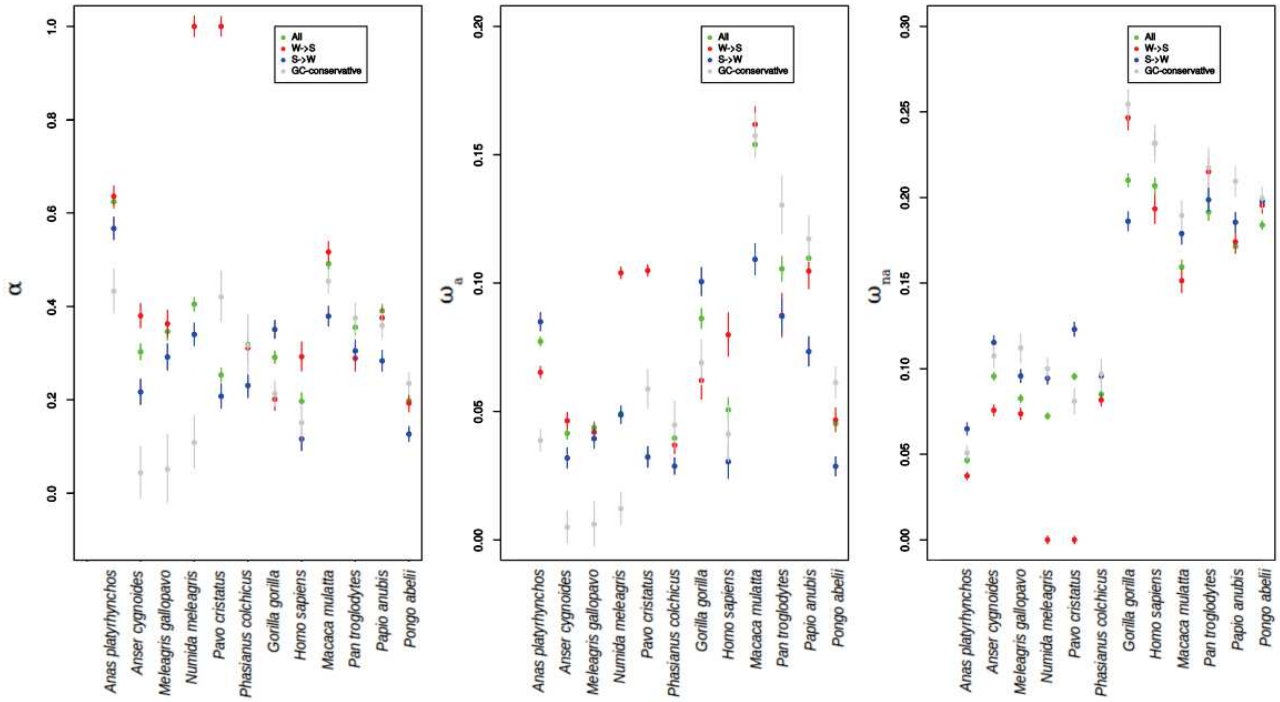
674 Statistics are estimated under a model assuming base composition non-stationarity.



675 **Figure 2:** dN, dS and dN/dS ratio against GC3 for each species and each type of substitutions (W→S, S→W
676 and GC-conservative substitutions).
677 Statistics are estimated under a model assuming base composition non-stationarity.



678 **Figure 3:**π_n/π_s ratio against GC3 for each species (A.: primates, B.: Galloanserae) and each type of mutations
679 (W→S, S→W and GC-conservative mutations) and average non-synonymous (C.) and synonymous (D.)
680 allele frequency for each species for W→S, S→W and GC-conervative SNPs (statistics estimated without
681 masking CpG sites)**.**



682 **Figure 4:** α, ω_a and ω_{na} estimates for each species and each type of mutations (all mutations, W→S, S→W
683 and GC-conservative) using all genes.
684 Statistics are obtained using the model "GammaZero".

Chapitre 3

Improved estimation of the adaptive substitution rate uncovers a negative relationship with the effective population size

The study presented in this chapter first aims at accounting for the two sources of bias of the DFE-α method identified in the two first chapters of this thesis, and second at testing the existence of a positive relationship between the adaptive substitution rate ω_a and the effective population size N_e.

This project was motivated by the discrepancies between empirical results concerning the existence of such a relationship. Indeed, the recent analysis of 44 species pairs failed to find this relationship (Galtier 2016), whereas Gossman et al. (2012) found a positive relationship between ω_a and N_e. In the first chapter, we showed that in case of long-term demographic fluctuations, estimations of α and ω_a can be severely overestimated. We also showed in the second chapter that GC-biased gene conversion can also lead to overestimations of these two statistics. Thus, those two sources of bias may explain the discrepancies reported in the literature.

To correctly estimate ω_a and correlate it to N_e, we had to generate a dataset meeting certain requirements : it had to include several species with coding sequence data for about ten individuals per species. To estimate divergence statistics, we needed for each species at least one closely related species. In addition, to take into account potential long-term fluctuations in selective and drift regime, which are by definition untraceable in polymorphism data of a single species, we chose to sample several closely related species within a taxonomic group (four to six species). Indeed, we make the hypothesis that the present variation in population sizes between closely related species represents well the possible range of population size fluctuations that one population experienced during the time period of its divergence with sister species. Thus, we have the potential to use this information to understand or correct our estimates of α and ω_a.

The dataset consists of newly generated data as well as data from public databases. The data from public databases (vertebrate species: primates, fowls, passerines and muroides) were retrieved and prepared (assembly, genotyping and sometimes orthogroups detection) by two computer engineers who have succeeded in my team, Émeric Figuet and Paul Simion, between my second and third year of thesis (muroides data are being finalized and are not included in the article yet). We started to prepare the newly generated dataset (invertebrate species: mussels, earth worms, ribbon worms, butterflies and ants) much sooner, in my fist year of PhD, as it was a big work that took me a lot of time. After I collected the samples (already extracted DNA or tissues), and in order to minimize the costs, we chose to use the capture technique, which allows to target the sequencing to a subset of predefined genes. I performed the DNA extractions, the libraries preparation and the capture experiment for about 250 samples under the supervision of Marie-Ka Tilak, engineer in

charge of my team's molecular biology laboratory. The sequencing took place in the middle of my second year, and then I processed the data myself.

I present the results of this study under the form of an article, even if this work is not totally complete and ready to be submitted to a scientific journal. I included some remarks as for the possible improvements of the dataset and analysis.

References :

Galtier N. 2016. Adaptive Protein evolution in animals and the effective population size hypothesis. PLOS Genetics 12:e1005774.

Gossmann TI, Keightley PD, Eyre-Walker A. 2012. The effect of variation in the effective population size on the rate of adaptive molecular evolution in Eukaryotes. Genome Biology and Evolution 4:658–667.

Title : Improved estimation of the adaptive substitution rate uncovers a negative relationship with the effective population size.

Authors : Rousselle M[1], Simion P[1], Tilak MK[1], Figuet E[1], Nabholz B[1], Galtier N[1].

[1]ISEM, Univ Montpellier, CNRS, EPHE, IRD, Montpellier, France.

Corresponding author: Marjolaine Rousselle

marjolaine.rousselle@umontpellier.fr

ABSTRACT

Estimating the adaptive amino acid substitution rate (ω_a) in various species is of primary importance to uncover the determinants of adaptation, and test the theoretical existence of a link between ω_a and the effective population size (N_e). Since the seminal test of McDonald & Kreitman (1991), methods for estimating ω_a have improved, taking in consideration some sources of bias, notably the presence of slightly deleterious mutations and recent changes in effective population size. Nevertheless, a number of intriguing empirical results suggests that not all sources of bias have been removed yet. Recently, it was suggested that various evolutionary processes, such as GC-biased gene conversion (gBGC) and ancient demographic fluctuations can mislead methods of estimation of the adaptive substitution rate if not taken into account. These remaining sources of bias may explain the lack of strong empirical support for a positive relationship between ω_a and N_e in the existing literature. In this study we re-examine this relationship accounting for the two above-mentioned sources of bias. The problem of gBGC is circumvented by only using GC-conservative mutations and substitutions, whereas the extent of long-term population size fluctuations is approached via the variance in estimated drift rate among species from the same taxon. Using coding sequence polymorphism data in 40 species from eight families of animals, including 23 newly generated invertebrates species, we report a negative relation between ω_a and various proxies of long-term N_e. This result is in contradiction with the common suggestion that ω_a should be positively related to N_e, but is consistent with theoretical predictions obtained under Fisher's geometrical model.

INTRODUCTION

A long standing question in evolution is whether the adaptive substitution rate is influenced by the effective population size, N_e, and how. Theory predicts that there is a positive correlation between the adaptive substitution rate, that we note ω_a, and N_e (Kimura 1983) if adaptation is limited by the supply of mutations. This is because the fixation rate of an adaptive mutation is expected to be proportional to $\mu N_e s$, where μ is the mutation rate and s is the selection coefficient of the mutation. Indeed, the fixation probability of a new advantageous mutation of selection coefficient s is proportional to $N_e s/N$ (where N is the census population size), if $N_e s >> 1$ and s is small, and the rate at which new mutations occur is $N\mu$ (Kimura 1983). Moreover, if N_e is large, there is a higher proportion of mutations with $N_e s >> 1$ (Gossmann et al. 2012).

The empirical evidence for a positive correlation between N_e and the adaptive substitution rate, however, is so far equivocal. On one hand, a study focusing on six closely related species of sunflowers, and an other one focusing in 13 pairs of eukaryotic species found a positive significant relationship between Ne and ω_a (Strasburg et al. 2010, Gossmann et al. 2012). On the other hand, one study focusing on two species of *Drosophila* and another one comparing 44 non-model animal species pairs failed to detect a positive significant relationship between N_e and ω_a (Jensen and Bachtrog 2011; Galtier 2016). These contrasted results demonstrate the uncertainty that still prevails on the question.

To explain the absence of relationship between N_e and ω_a in their results, Galtier (2016) invoked the fact that in small-N_e species, proteins tend to be less adapted (i.e., further away from their fitness optimum). These proteins would therefore have more opportunity for adaptation than proteins in large-N_e species. Indeed, theoretical work based on Fisher's geometrical model (FGM) suggests that the proportion of new adaptive mutations is mainly determined by the distance of the population to the fitness optimum, which itself depends on the long-term effective population size (Lourenço et al. 2013; Huber et al. 2017). Small populations tend to accumulate slightly deleterious substitutions, which creates a potential for compensatory, beneficial mutations. In this framework, Huber et al. (2017) estimated that 14% of new non-synonymous mutations are beneficial in humans, but only 1.5 % in *Drosophila* (Huber et al. 2017). This difference in distribution of fitness effects (DFE) between small- and large-N_e populations might somehow compensate for the overall larger number of mutations and higher fixation probability of beneficial mutations in large-N_e species. Besides,

this theoretical work based on Fisher's geometrical model assumes that there is a single-peaked fitness landscape, and predicts that the distribution of fitness effects of mutations (DFE) depends on N_e, which is in contradiction with the arguments underlying the theory of a positive link between N_e and ω_a. If these hypotheses of Fisher's geometrical model are correct, we can expect adaptive mutations to be more common in small-N_e species, leading to a negative correlation between N_e and ω_a.

Besides this theoretical discussion, the discrepancies between empirical studies also raise the question of the biases that could affect the methods used to estimate ω_a. All the studies mentioned above rely on McDonald & Kreitman-like approaches, now known under the term "DFE-α" methods (Tataru et al. 2017) – because they allow the estimation, alongside with ω_a, of the DFE parameters as well as the statistic α, which is the proportion of amino acid substitutions that are adaptive. These methods rely on the comparison of the rate of non-adaptive substitution (ω_{na}), estimated from polymorphism data in one focal species, and the observed dN/dS ratio, the ratio of non-synonymous substitution rate over synonymous substitution rate, which is typically calculated by comparing the focal species to an outgroup. Then, one can deduced $\omega_a = dN/dS - \omega_{na}$, and $\alpha = \frac{dN/dS - \omega_{na}}{dN/dS}$. The DFE-α method implicitly makes the assumption that the regime of selection/drift has been constant over the considered time period, i.e. since the divergence between the focal and outgroup species. If however the selection/drift regime has changed (through a change in effective population size, for instance) between the period during which divergence was built and the period during which polymorphism was built, then one can expect overestimation or underestimation of ω_a (Eyre-Walker 2002; Rousselle et al. 2018). The most recent versions of the DFE-α method allow to take into account recent N_e changes that affect polymorphism in the estimation process (Eyre-Walker et al. 2006; Eyre-Walker and Keightley 2009; Messer and Petrov 2013). In contrast, ancient N_e changes that affect divergence are virtually impossible to trace, as all the signal of such changes was lost through time. We showed in a previous study based on simulations that ancient demographic fluctuations can lead to severely overestimated values of α and ω_a, an upward bias which is exacerbated when the true adaptive substitution rate is low (Rousselle et al. 2018). Additionally, it has been shown by modeling single changes in N_e that in presence of slightly deleterious mutations, an increase in N_e in the past could yield spurious evidence for positive selection, whereas a decrease in N_e could either increase or decrease α (Eyre-Walker 2002). Both the studies of Gossman et al. (2012) and Galtier (2016) used methods allowing

for recent changes in N_e, but we argue that ancient fluctuations in N_e may blur the signal of the relationship between the two statistics.

Another source that may bias the DFE-α method is GC-biased gene conversion (gBGC). gBGC is a fixation bias in favor of G and C alleles around recombination hotspots, which is known to be at work in a wide range of organisms (Eyre-Walker 1999; Montoya-Burgos et al. 2003; Meunier and Duret 2004; Spencer 2006; Webster et al. 2006; Mancera et al. 2008; Escobar et al. 2011; Pessia et al. 2012; Lesecque et al. 2013; Lassalle et al. 2015; Keith et al. 2016; Smeds et al. 2016; Galtier et al. 2018). Several studies focusing on birds and more recently on primates indicate that gBGC may lead to overestimation (Corcoran et al. 2017; Rousselle et al. under revision) or underestimation of ω_a (Bolívar et al. 2018). Interestingly, gBGC does not affect genomic evolution with the same intensity in all organisms (Galtier et al. 2018). Therefore, the signal of a correlation between N_e and ω_a may be blurred if gBGC does not influence the estimation of ω_a the same way, or to the same extent, in different taxa.

In this study, we built a data set comprising eight taxonomic groups of animals with coding sequence polymorphism data in four to six species per group. Our goal with this new design is to provide an estimate of the adaptive rate for each group, taking information on the intensity and impact of demographic fluctuations from the among-species variance in estimated drift rate. We introduce two different strategies for estimating one adaptive substitution rate per taxonomic group accounting for ancient demographic fluctuations, while also correcting for gBGC by considering only GC-conservative changes – i.e. A<->T and C<->G mutations/substitutions. One strategy consists in pooling the synonymous and non-synonymous SFS from each species within a group as well as the divergence counts of the whole subtree of the group, and then applying the “classical” approach. The second one consists in computing the arithmetic mean of ω_{na} across species within a group, and then subtracting this average to the total dN/dS ratio of the subtree to obtain the adaptive substitution rate. This strategy is motivated by the hypothesis that the present variation in population sizes between closely related species represents well the possible range of population size fluctuations that one population experienced during the time period of its divergence with sister species. Both strategies uncover for the first time a negative relationship between ω_a and life history traits previously identified as good indicators of long-term N_e. This is at odds with the predominant theory underlying the link between N_e and ω_a, but it is supported by recent theoretical predictions obtained under Fisher’s geometrical model.

MATERIAL & METHODS

1. Dataset

We assembled a dataset of coding sequence polymorphism in 40 species from eight taxonomic groups, each group including 4-6 closely-related species (Table S1). The eight taxa we analyzed are Catharrhni (Mammalia, hereafter called “primates”), Passeriformes (Aves, hereafter called “passerines”), Galloanserae (Aves, hereafter called “fowls”), Lumbricidae (Annelida, hereafter called “earth worms”), *Lineus* (Nemertea, hereafter called “ribbon worms”), *Mytilus* (Mollusca, hereafter called “mussels”), Satyrini (Lepidoptera, hereafter called “butterflies”), and *Formica* (Hymenoptera, hereafter called “ants”).

We retrieved genomic and exomic data for six species of primates with more than five individuals: *Homo sapiens* (19 individuals), *Pan troglodytes* (20 individuals), *Papio anubis* (5 individuals), *Pongo abelii* (10 individuals), *Gorilla gorilla* (20 individuals), *Macaca mulatta* (19 individuals) from four different Bioprojects (see Table S1).

We retrieved genomic data for five species of passerines with at least ten individuals: *Ficedula albicollis* (20 individuals), *Taeniopygia guttata* (20 individuals), *Geospiza difficilis* (8 individuals), *Parus major* (20 individuals) and *Corvus sp.* (10 individuals from both *C. corone* or *C. cornix*) from different Bioprojects (see **Table S1**).

We retrieved RNA-seq reads from six species of fowls: *Meleagris gallopavo* (10 individuals), *Phasianus colchicus* (11 individuals), P*avo cristatus* (10 individuals), *Numida meleagris* (10 individuals), *Anas platyrhynchos* (10 individuals), *Anser cygnoides* (10 individuals) from SRA Bioproject PRJNA271731 (Wright et al. 2015).

The other five datasets were newly generated via coding sequence capture (see below). We gathered tissue samples or DNA samples for at least eight individuals per species and four or five species per taxonomic groups. Reference transcriptomes were obtained from previously published RNA-seq data in one species per taxonomic group (Romiguier et al. 2014; Rousselle et al. 2016; Ballenghien et al. 2017; Galtier et al. 2018). Details of the species and individuals number are presented in **Table S1**.

2. Multiplexed target capture experiment

DNA from whole animal body (ants), body section (earth worms, ribbon worms), mantle (mussels) or head/thorax (butterflies) was extracted using DNAeasy Blood and Tissue kit (QIAGEN)

following the manufacturer instructions. About 3 µg of total genomic DNA were sheared for 20 min using an ultrasonic cleaning unit (Elmasonic One). Illumina libraries were constructed for all samples following the classical protocol involving blunt-end repair, adapter ligation, and adapter fill-in steps as developed by Meyer and Kircher (2010) and adapted in Tilak et al. (2015).

To perform target capture, we randomly chose contigs out of the five reference transcriptomes (*Maniola jurtina* for butterflies, *Lineus longissimus* for ribbon worms, *Mytilus galloprovincialis* for mussels, *Allobophora chlorotica L1* for earth worms, and *Formica cunicularia* for ants) in order to reach 2Mb of total sequence length per taxon (~2000 contigs). 100nt-long baits corresponding to these sequences were synthesized by MYbaits (Ann Arbor, MI, USA), with an average cover of 3X.

We then performed multiplexed target capture following the MYbaits targeted enrichment protocol: about 5 ng of each library were PCR-dual-indexed using Taq Phusion (Phusion High-Fidelity DNA Polymerase Thermo Scientific) or KAPA HiFi (2× KAPA HiFi HotStart ReadyMix KAPABIOSYSTEMS) polymerases. We used primers developed in Rohland and Reich (2012). Indexed libraries were purified using AMPure (Agencourt) with a ratio of 1.6, quantified with Nanodrop ND-800, and pooled in equimolar ratio. We had a total of 96 combinations of indexes, and two Illumina lanes, for a total of 244 individuals. This means that we had to index two (rarely three) individuals with the same combination to be sequenced in the same lane. When this was necessary, we chose individuals very distantly related (i.e. from different taxa). 150 ng of each indexed libraries were pooled with the baits according to species and putative phylogenetic distance (i.e. the species whose transcriptome was used to design the baits), resulting in 24 pools of 7 to 15 libraries. This allowed us to adjust the hybridization temperature protocol for each pool depending on the putative phylogenetic distance between the libraries and the baits. We then prepared between 100 and 500 ng of DNA of each pool in 7 µl that we used for hybridization reactions following the MYbaits targeted enrichment protocol. For pools containing samples libraries corresponding to individuals of the species used to design baits, we used a temperature of 65°C during 22 h, and for the other ones we ran the hybridization reactions for 16 h at 65°C, 2 h at 63°C, 2 h at 61°C and 2 h at 59°C. Following hybridization, the reactions were cleaned according to the kit protocol with 200 µL of wash buffers, and hot washes were performed at 65°C or 59°C depending on the pools. The enriched pools were then PCR-amplified for 14 to 16 cycles depending on the pools, after removal of the streptavidin beads, and PCR products were purified using AMPure (Agencourt) with a ratio of 1.6. 150 ng of each enriched pools were mixed in two pools and paired-end sequenced on two Illumina HiSeq® 2500 lines. Illumina sequencing and demultiplexing were subcontracted.

3. Assembly and genotyping

For RNA-seq data (fowls), we used trimmomatic (Bolger et al. 2014) to remove Illumina adapters and reads with a quality score below 30. We constructed *de novo* transcriptome assemblies for each species following strategies B in (Cahais et al. 2012), using Abyss (Simpson et al. 2009) and Cap3 (Huang and Madan 1999). Open reading frames (ORFs) were predicted using the Trinity package (Grabherr et al. 2011). Contigs carrying ORF shorter than 150 bp were discarded. Filtered RNA-seq reads were mapped to this assembly using BWA (Li and Durbin 2009). Contigs with a coverage across all individual below 2.5Xn (where n is the number of individuals) were discarded. Diploid genotypes were called according to the method described in Tsagkogeorga et al. (2012) and Gayral et al. (2013)(model M1) via the software reads2snps. This method calculates the posterior probability of each possible genotype in the maximum likelihood framework. Genotypes supported by a posterior probability higher than 95% are retained, otherwise missing data is called.

For primates and passerines reference genomes, assemblies and annotations files were downloaded from Ensembl (release 89) and NCBI (see Table S1). We kept only '*CDS*' reports in the annotations files, corresponding to coding exons, which were annotated with the automatic Ensembl annotation pipeline, and the havana team for *Homo sapiens*. We used trimmomatic to remove Illumina adapters, to trim low-quality reads (i.e. with an average base quality below 20), and to keep only reads longer than 50bp. Reads were mapped using BWA program (Li and Durbin 2009) on the complete reference assembly (version 0.7.12-r1039). We filtered hits with mapping score below 20 and removed duplicates, and we extracted mapping hits corresponding to regions containing coding sequences according to the annotated reference assembly. We called SNP using GATK (v3.8-0-ge9d80683). The pipeline comprised two rounds, each including a variant calling with HaplotypeCaller, joint genotyping with GenotypeGVCFs, filtering of the SNPs with VariantFiltration, with a base quality score recalibration between the two rounds. The obtained vcf were then converted into fasta alignments using custom python scripts.

For reads generated through target capture experiment, we cleaned reads with trimmomatic to remove Illumina adapters and reads with a quality score below 30. For each species, we chose the individual with the highest coverage and constructed *de novo* assemblies using the same strategy than for fowls. Reads of each individuals were then mapped to the newly generated assemblies for each species, using BWA (Li and Durbin 2009). Diploid genotypes were called using the same protocol as in fowls. We used a version of the SNP calling method which accounts for between-individual, within-species contamination as introduced in Ballenghien et al. (2017) (see the following section). As the newly generated assemblies likely contained intronic sequences, the

predicted cDNAs were compared to the reference transcriptome using blastn searches, with a threshold of e-value of 10e-15. We used an in-house script to remove any incongruent correspondence or inconsistent overlap between sequences from the transcriptomic references and the predicted assemblies, and removed six base pairs at each extremity of the resulting predicted exonic sequences. These high-confidence exonic sequences were used for downstream analyses.

3. Contamination detection and removal

For the newly generated data set, we performed two steps of contamination detection. First, we used the program CroCo to detect inter-specific contamination in the *de novo* assembly generated after exon capture (Simion et al. 2018). CroCo is a database-independent tool designed to detect and remove cross-contaminations in assembled transcriptomes of distantly related species. This program classifies predicted cDNA in five categories, “clean”, “dubious”, “contamination”, “low coverage” and “high expression”.

Secondly, we used a version of the SNP calling method which accounts for between-individual, within-species contamination as introduced in Ballenghien et al. (2017), using the -contam=0.1 option. This means assuming that up to 10% of the reads assigned to one specific sample may actually come from a distinct sample, and only validating genotypes robust to this source of uncertainty.

4. Orthology prediction

In primates, we extracted predicted one-to-one orthology groups across the six species from the OrthoMaM database (Ranwez et al. 2007, Douzery et al. 2014).

In fowls and passerines, we translated the obtained CDS into proteins and predicted orthology using OrthoFinder (Emms and Kelly 2015). In fowls, full coding sequences from the well-annotated chicken genome (Ensembl release 89) were added to the dataset prior to orthology prediction, then discarded. We kept only orthogroups that included all species. We aligned the orthologous sequences with MACSE (Multiple Alignment for Coding SEquences (Ranwez et al. 2011).

In each of earth worms, ribbon worms, mussels, butterflies and ants, orthogroups were created via a a blastn similarity search between predicted exonic sequences and reference transcriptomes. In each taxon, we concatenated the predicted exonic sequences of each species that matched the same ORF from the reference transcriptome and aligned these using MACSE.

5. Divergence analysis

We estimated lineage specific dN/dS ratio using bppml (version 2.4) and MapNH (version 2.3.2) (Guéguen et al. 2013), the former for estimating each branch length and the latter for mapping substitutions on species specific branches.

Tree topologies were obtained from the literature (**Table S2**). In passerines, fowls, and primates, we kept only alignments comprising all the species. In the other groups we also kept alignments comprising all species but one.

We also estimated dN/dS ratios at group level by adding up substitution counts across branches of the trees, including internal branches.

To account for gBGC, we modified the MapNH software such that only GC-conservative substitutions were recorded (Rousselle et al., under revision). We estimated the non-synonymous and synonymous number of GC-conservative sites per coding sequence using an in-house script (Rousselle et al., under revision). We could then compute the dN/dS ratio only for GC-conservative substitutions.

6. Polymorphism analysis

For each taxon, we estimated ancestral sequences at each internal node of the tree with the Bio++ program SeqAncestor (Guéguen et al. 2013). The ancestral sequences at each internal node were used to orientate single nucleotide polymorphisms (SNPs) of species that descend from this node. We then computed non-synonymous (π_n) and synonymous (π_s) nucleotide diversity and unfolded synonymous and non-synonymous site frequency spectra both using all mutations and only GC-conservative mutations using in-house scripts as in Galtier (2016).

7. DFE-α method

We estimated α, ω_a and ω_{na} using the approach of Eyre-Walker and Keightley (2009) as implemented in Galtier (2016) (program Grapes v.1.0). It models the distribution of the fitness effects (DFE) of deleterious and neutral non-synonymous mutations as a negative gamma distribution, which is fitted to the synonymous and non-synonymous site frequency spectra (SFS) computed for a set of genes. This estimated DFE is then used to deduce the expected dN/dS under near-neutrality. The difference between observed and expected dN/dS provides an estimate of the proportion of adaptive non-synonymous substitutions, α. The per mutation rate of adaptive and non-

adaptive amino acid substitution were then obtained as following: $\omega_a = \alpha(dN/dS)$ and $\omega_{na} = (1-\alpha)(dN/dS)$. We computed these statistics for each species using the per branch dN/dS ratio, using either all mutations and substitutions, or only GC-conservative mutations and substitutions. We used the model "GammaExpo" (Galtier 2016) as it has proven to best fit the data in Galtier (2016). Briefly, in this model the DFE includes both a negative (gamma) and a positive (exponential) component. When estimating DFE model parameters, we accounted for recent demographic effects by using nuisance parameters, which correct each class of frequency of the synonymous and non-synonymous SFS relative to the neutral expectation in an equilibrium Wright–Fisher population (Eyre-Walker et al. 2006).

We also estimated α, ω_a and ω_{na} at the group level. In one hand, we pooled species specific SFS from each group, and used the dN/dS ratio of the total tree of each taxon. We did so following the unweighted and unbiased strategy of James et al. (2013), which combines polymorphism data across species with equal weights. Briefly, we divided the synonymous and non-synonymous number of SNPs of each category of the SFS of each species by the total number of SNPs of the species, then we summed those normalized numbers across species (James et al. 2016) and finally we transformed those sums so that the total number of SNPs of the pooled SFS matches the total number of SNPs across species. In the other hand, we simply computed the arithmetic mean of ω_{na} across species within a taxonomic group to obtain a non-adaptive substitution rate at the group level. We then subtracted this average to the dN/dS ratio for the total tree of each taxon to obtain an adaptive substitution rate at the group level.

8. Life history traits variables

Five life history traits were retrieved from the literature for each species: adult size (i.e. the average length of adults), body mass (i.e. the mean body mass of adults' wet weights), fecundity (i.e. the number of offspring released per day), longevity (i.e. the maximal recorded longevity in years), and propagule size (i.e. the size of the juvenile or egg or larva when leaving parents or group of relatives) (**Table S3**). In case of social insects and birds, parental care is provided to juveniles until they reach adult size so in these cases, propagule size is similar to adult size.

RESULTS & DISCUSSION

1. Exon capture control for contamination

Exon capture was achieved in a total of 244 individuals from 23 species. We obtained sufficient data for 225 of them (~92%). We obtained an average coverage of 9X in ants, 23X in butterflies, 10X in earth worms, 28X in ribbon worms and 26X in mussels (average of per species medians). The percentage of targeted coding sequences for which at least one contig was recovered varied from 31.9% to 88.2% across species (median=78.8%, **Table 1**).

Species	Number of targeted transcripts with a blastn hit in the *de novo* assembly	Total number of targeted transcripts	Percentage of targeted transcripts present in the *de novo* assembly
F. fusca	1427	1810	78.8
F. sanguinea	1396	1810	77.1
F. pratensis	1398	1810	77.2
F. cunicularia	1406	1810	77.7
F. selysi	1073	1810	59.3
M. jurtina	1921	2235	86.0
M.galathea	1713	2235	76.6
P. tithonus	1823	2235	81.6
P. bathseba	1864	2235	83.4
A. hyperanthus	1772	2235	79.3
A. chlorotica L1	2293	2955	77.6
A. chlorotica L2	2315	2955	78.3
A. chlorotica L4	1732	2955	58.6
A. icterica	2321	2955	78.5
L. terrestris	943	2955	31.9
L. sanguineus	1251	1725	72.5
L. ruber	1521	1725	88.2
L. lacteus	1516	1725	87.9
L. longissimus	1505	1725	87.2
M. galloprovincialis	1820	2181	83.4
M. edulis	1721	2181	78.9
M. trossulus	1740	2181	79.8
M. californianus	1808	2181	82.9

Table 1: Summary of the number of targeted transcripts recovered in the capture experiment.

This resulted in between 991 and 1261 orthogroups in ants, between 1561 and 1671 orthogroups in butterflies, between 796 and 1460 orthogroups in earth worms, between 1049 and 1413 orthogroups in ribbon worms, and 1525 orthogroups in mussels. The differences in terms of number of orthogroups comes from the fact that we not only kept orthogroups with all species but also

orthogroups with all species but one to estimate dN/dS value for each terminal branches in order to maximize the number of substitutions.

We evaluated inter-specific contamination using the software CroCo (**Figure S1**). Overall, inter-groups connections in **Figure S1** indicate a low level of cross-contamination: when there are connections between taxonomic groups, they concern on average 38 contigs identified as contaminants, the worst case being the 172 contigs identified as contaminants between the assembly of *L. sanguineus* and *M. galloprovincialis*. The connections between assemblies of closely related species are very likely to be false positive, especially since the intensity of the intra-taxon connections is congruent with the phylogenetic distance between species within taxa, and since we did not grouped our molecular work per taxonomic group. All the contigs identified as potential contaminants were excluded from the dataset in downstream analyzes.

For the other datasets obtained from the literature, we obtained 8604 orthogroups in primates, 4439 orthogroups in fowls, and 6755 orthgroups in passerines.

SNPs counts obtained after genotyping for all species used in the analysis are summed up in **Table S4**. We obtained only three species with less than a thousand SNPs, the minimum being 153 for *L. longissimus*, for which we could recover data for only six individuals. This may not be sufficient to properly estimate α, ω_a and ω_{na}, and cast the doubt on the estimates of *L. longissimus*, *F. selysi* and *A. chlorotica L2*, however the number of SNPs for the rest of the species is very satisfying.

In conclusion, the capture experiment seems to be suitable to recover population coding sequence data for several closely related species. This associated with the growing number of public datasets allowed us to build a satisfying dataset with 40 species and eight taxonomic groups. We hope to further improve this dataset by adding a Muroides and a *Drosophila* group.

2. Influence of GC-biased gene conversion

We evaluated the influence of gBGC on estimates of ω_a by comparing the estimates obtained for each species using all types of mutations and substitutions *vs.* only GC-conservative ones. We observed that in a majority of cases, ω_a estimates obtained using all mutations and substitutions are higher than those estimated using only GC-conservative mutations and substitutions (26 out of 40 one-sided binomial test p-value: 0.04). Two studies previously identified a similar effect of gBGC in primates and birds (Corcoran et al. 2017, Rousselle et al. under revision). However, the fact that this effect is only observed in roughly two thirds of the species may reflect the fact that all the species we used here may not be affected by gBGC with the same intensity. Galtier et al. (2018) showed that gBGC is present in a wide range of metazoans with some remarkable exceptions

(Galtier et al. 2018). Consequently, gBGC may have confounded the relationship between N_e and ω_a in previous studies if the estimates of ω_a were overestimated in some species but not other ones.

3. What do McDonald & Kreitman-like approaches really measure ?

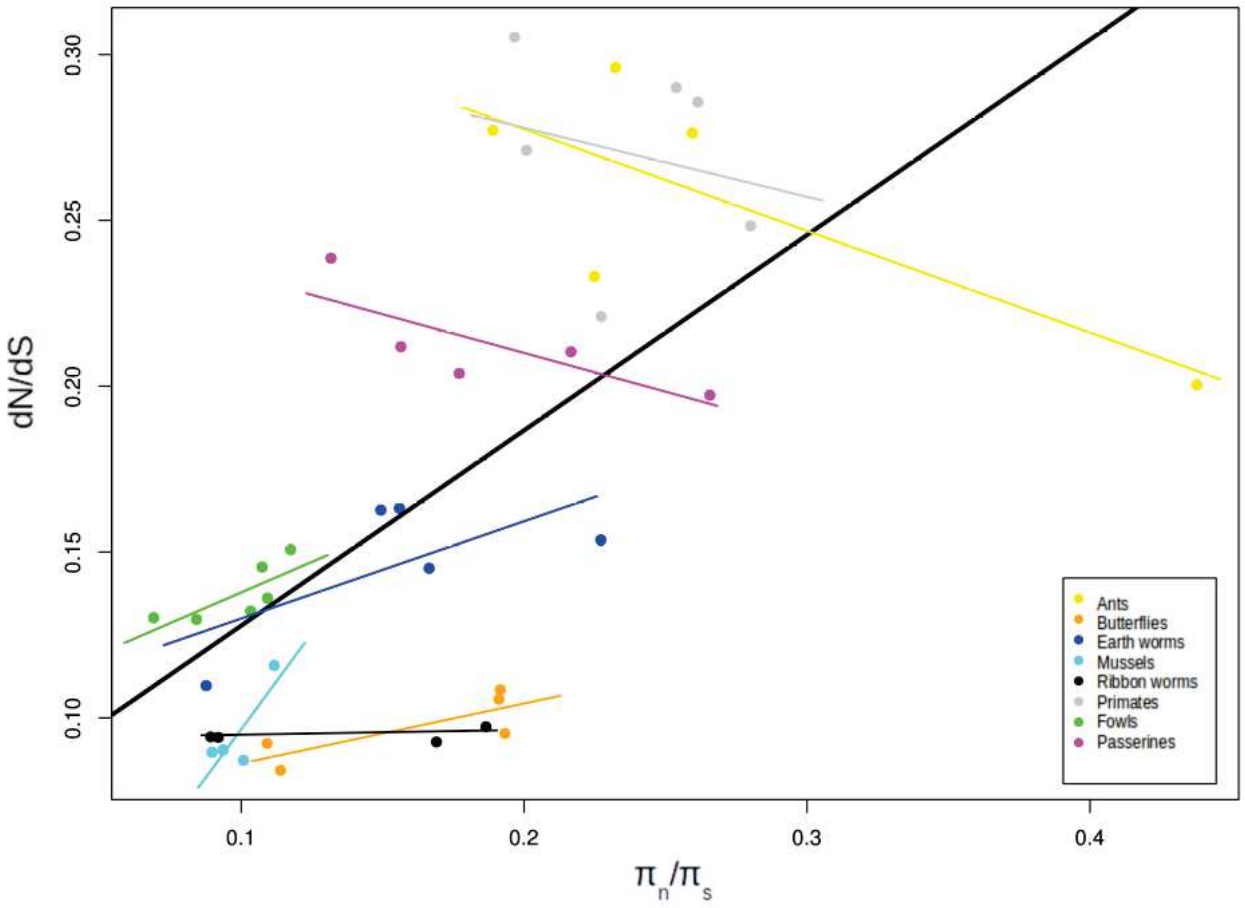


Figure 1 : dN/dS ratio and π_n/π_s relationship.
Linear regression r^2=0.37, p-value=2.9e-05.

There is a significant positive relationship between the dN/dS ratio and the π_n/π_s ratio (regression test r^2=0.36, p-value=2.89e-05, **Figure 1**). This indicates that coding sequence evolution is dominated by nearly neutral processes. The nearly neutral theory of molecular evolution predicts that small populations should carry more deleterious mutations at a faster rate than large populations, both polymorphic and fixed (Ohta 1973). Adaptive evolution, on the other hand, would be expected to decouple dN/dS from π_n/π_s since beneficial mutations contribute to divergence to a much greater extent than to polymorphism (McDonald and Kreitman 1991).

The correlation between dN/dS and π_n/π_s seems to be mainly driven by the inter-group signal. When we explored the link between the dN/dS ratio and the π_n/π_s ratio within groups, we obtained no consistent results, with correlations coefficients being either positive or negative (Spearman correlation test : ants : R=-0.5, p-value=0.45, butterflies:R=0.6, p-value= 0.35, earth worms :R=0.2, p-value=0.78, ribbon worms :R=0.2, p-value=0.92, mussels :R=0.4, p-value=0.75, primates :R=-0.37, p-value=0.49, fowls:R=0.89, p-value=0.033, passerines : R=-0.9, p-value=0.083).

We plotted the standard deviation of dN/dS, π_n/π_s, ω_{na} and ω_a, computed between species for each group (**Figure 2**). We found that the variance of π_n/π_s, ω_{na} and ω_a between closely related species is substantially greater than the variance in dN/dS, which tends to be remarkably homogeneous within groups.

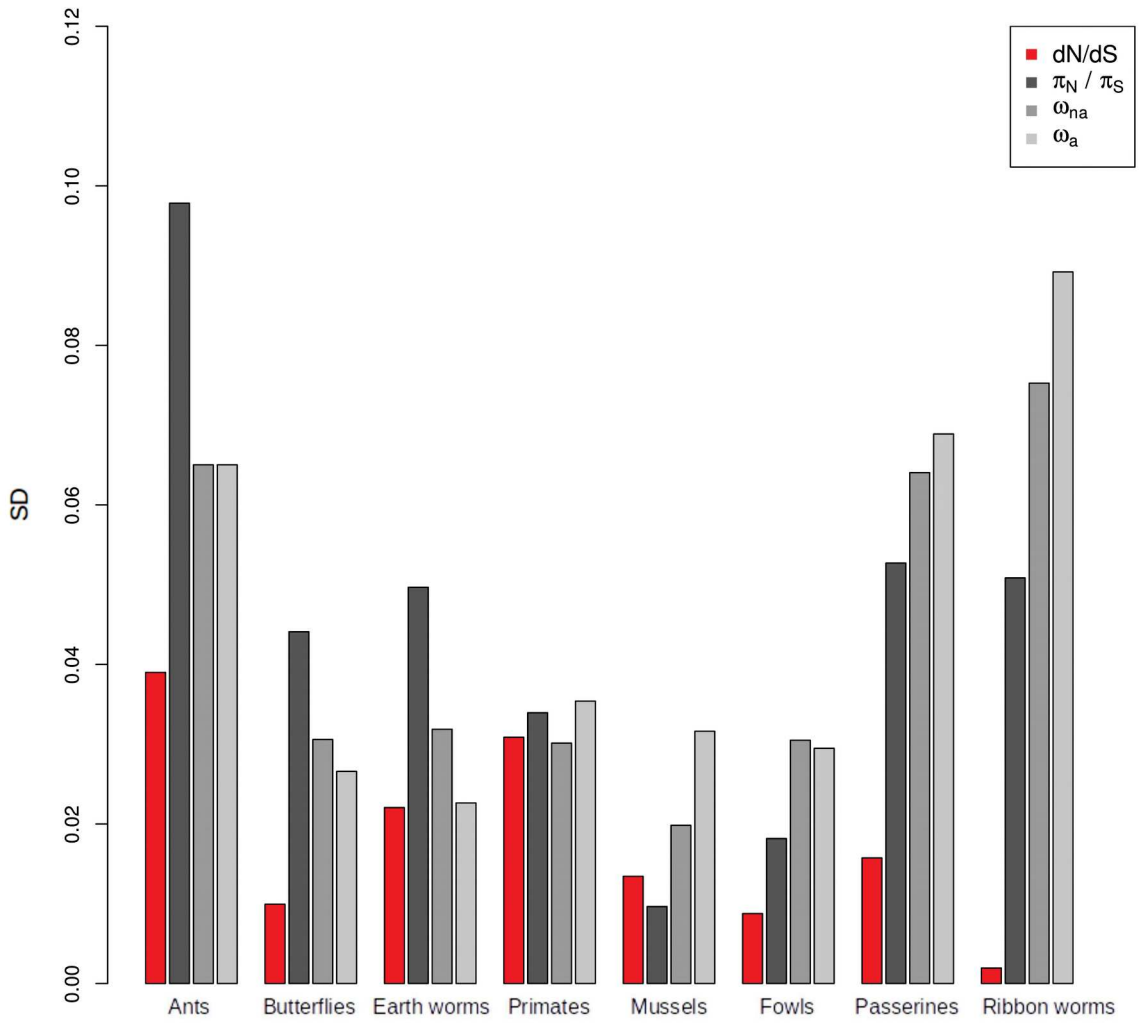


Figure 2 : Standard deviation of dN/dS, π_n/π_s, ω_{na} and ω_a within each taxonomic group.

dN/dS ratios being so similar between closely related species supports the suggestion that these species have probably been undergoing similar amount of adaptive evolution since their divergence, and have similar long-term effective population size – which does not imply constant population size. In contrast, the elevated within-group variance in π_n/π_s, ω_{na} and ω_a as well as the fact that there is no consistent correlation between the dN/dS ratio and the π_n/π_s ratio between closely related species (**Figure 1**) presumably reflects temporary and recent changes in N_e that are unrelated to the long-term rate of adaptation. Yet in the McDonald-Kreitman framework the adaptive substitution rate ω_a is deduced from the difference between dN/dS and ω_{na}, the latter being inferred from polymorphism data. This suggests that between-species differences in ω_a might mainly reflect changes in N_e that influence polymorphism data rather than between-lineages differences in the rate of adaptive evolution. This issue relates to previous studies which showed that a change in the selection/drift regime between recent and more ancient past can lead to spurious estimations of α and ω_a (Eyre-Walker, 2002; Rousselle et al. 2018).

Figure 2 suggests that divergence data might primarily reflect the average long-term selection/drift regime, whereas polymorphism data from any particular species reflects a recent, transient selection/drift regime, which might differ substantially from the long-term average. If it was true, then using the DFE-α method for individual species may then not be a good strategy to correctly estimate their adaptive substitution rate. We suggest rather using a sample of closely related species, and estimating a global adaptive substitution rate for the whole group. Indeed, the variation in N_e between species conveys information on the extent of past fluctuations in N_e. Incorporating this variation in the estimation process could help to overcome the issues of different selection/drift regimes between divergence and polymorphism data. In the next section, we introduce two approaches for estimating the adaptive rate from multi-species polymorphism and divergence data.

4. Measuring the adaptive substitution rate from multi-species polymorphism data

Our first estimates of the adaptive substitution rate at group level, which we call $\omega_{a[P]}$, was obtained by pooling the synonymous and non-synonymous SFS from each species within a group. We applied this strategy using only GC-conservative mutations and substitutions. We then explored the relationship between $\omega_{a[P]}$ estimates and life history traits known to be correlated to the long-term effective population size (propagule size, adult size, longevity, body mass and fecundity) (Romiguier et al. 2014; Figuet et al. 2016). We also used as a proxy for N_e the synonymous genetic diversity π_s averaged across closely related species. Note that π_s is expected to be linearly correlated

with the long-term N_e under the assumption that the mutation rate is constant across species, which seems plausible for closely related species. Surprisingly, we detected a significant negative relationship between $\omega_{a[P]}$ and π_s (r^2=0.86 and p-value=0.0006), as well as a marginally significant positive relationship between $\omega_{a[P]}$ and $\log_{10}$ transformed propagule size (regression test, r^2=0.40 and p-value=0.056) (**Figure 3**).

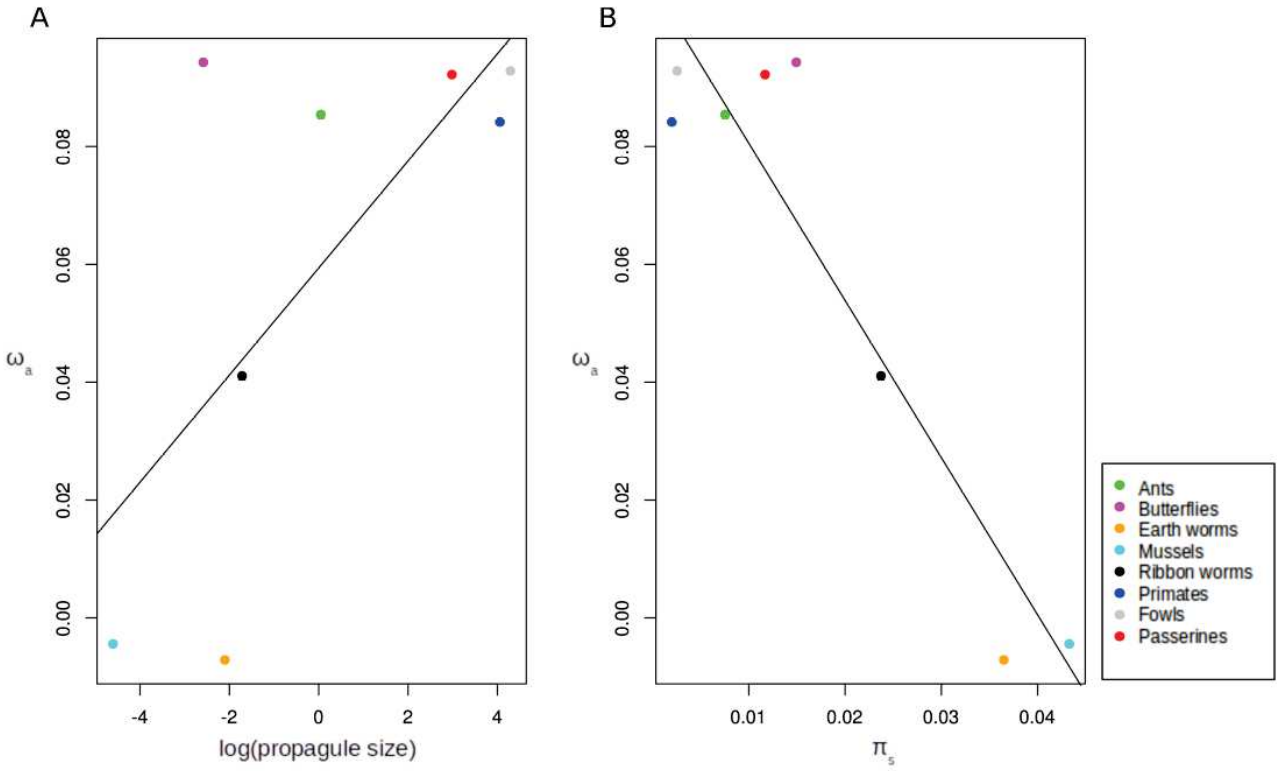


Figure 3 : Relationship between $\omega_{a[P]}$ and propagule size (A) and ω_a and π_s (B).
$\omega_{a[P]}$ is estimated with all mutations on pooled SFS (one per species).
A : Relationship between $\omega_{a[P]}$ and $\log_{10}$(propagule size) (cm).
B : Relationship between $\omega_{a[P]}$ and π_s.

We then estimated an adaptive substitution rate representative of all the species of one taxonomic group, that we call $\omega_{a[A]}$, by using the arithmetic mean of ω_{na} across species within a taxonomic group, and then subtracting this average to the total dN/dS ratio of the subtree of the group to obtain $\omega_{a[A]}$. We justify this strategy by making the hypothesis that the present variation in population sizes between closely related species represents well the possible range of population size fluctuations that one population experienced during the time period of its divergence with its sister species. Under our model, ω_{na} for an individual species is :

$$\omega_{na} = (\widehat{D_N^{na}}/L_N)/(D_S/L_S), \quad (1)$$

where $\widehat{D_N^{na}}$ is the expected number of non-synonymous substitutions:

$$\widehat{D_N^{na}} = 2L_N N_e t\mu \int_{-\infty}^{0} \phi(s) f(N_e, s) ds \quad (2)$$

where L_N is the number of non-synonymous sites, t is the divergence time, $\phi(s)$ is the fixation probability of a mutation with a selection coefficient s, and $f(Ne, s)$ is the DFE, in a population of size N_e.

Then, if we have sampled n closely related species, and we call N_{e1}, N_{e2}, N_{e3}, … N_{en} their respective present effective population size, then, making the hypothesis that the shape of the DFE and the mutation rate μ remains constant in time, one can express the expected number of non-synonymous substitutions, $\widehat{D_N^{na}}$, as :

$$\widehat{D_N^{na}} = \frac{t}{n} \sum_{i}^{n} 2L_N N_{ei} \mu \int_{-\infty}^{0} \phi(s) f(N_{ei,s}) ds \quad (3)$$

Equation (3) is equivalent to assuming that, as the considered species were diverging, they have randomly fluctuated between the n regimes of selection/drift we currently observe, spending the same amount of time in the n regimes. Under this assumption, the group level $\widehat{D_N^{na}}$ is simply the arithmetic mean of $\widehat{D_N^{na}}$ estimated across individual species. Then, using L_N, L_S and D_S of the total subtree of the considered species, we can use the arithmetic mean of ω_{na} across individual species, $\omega_{na[A]}$, as representative of the average non-adaptive selective regime during their divergence. Subtracting $\omega_{na[A]}$ from the dN/dS ratio estimated using all branches of the tree of the group, we obtain an estimate of the adaptive substitution rate for the whole group, $\omega_{a[A]}$.

We applied this strategy using only GC-conservative mutations and substitutions, and explored the relationship between $\omega_{a[A]}$ estimates and life history traits and π_s averaged across closely related species. Here again, we detected a significant positive relationship between $\omega_{a[A]}$ and $\log_{10}$ transformed propagule size (regression test, r^2=0.42 and p-value=0.049), as well as $\log_{10}$ transformed adult size (regression test, r^2=0.67 and p-value=0.0077) and $\log_{10}$ transformed longevity (regression test, r^2=0.59 and p-value=0.027). We also detected a significant negative relationship between $\omega_{a[A]}$ and π_s (regression test, r^2=0.47 and p-value=0.037) (**Figure 4**).

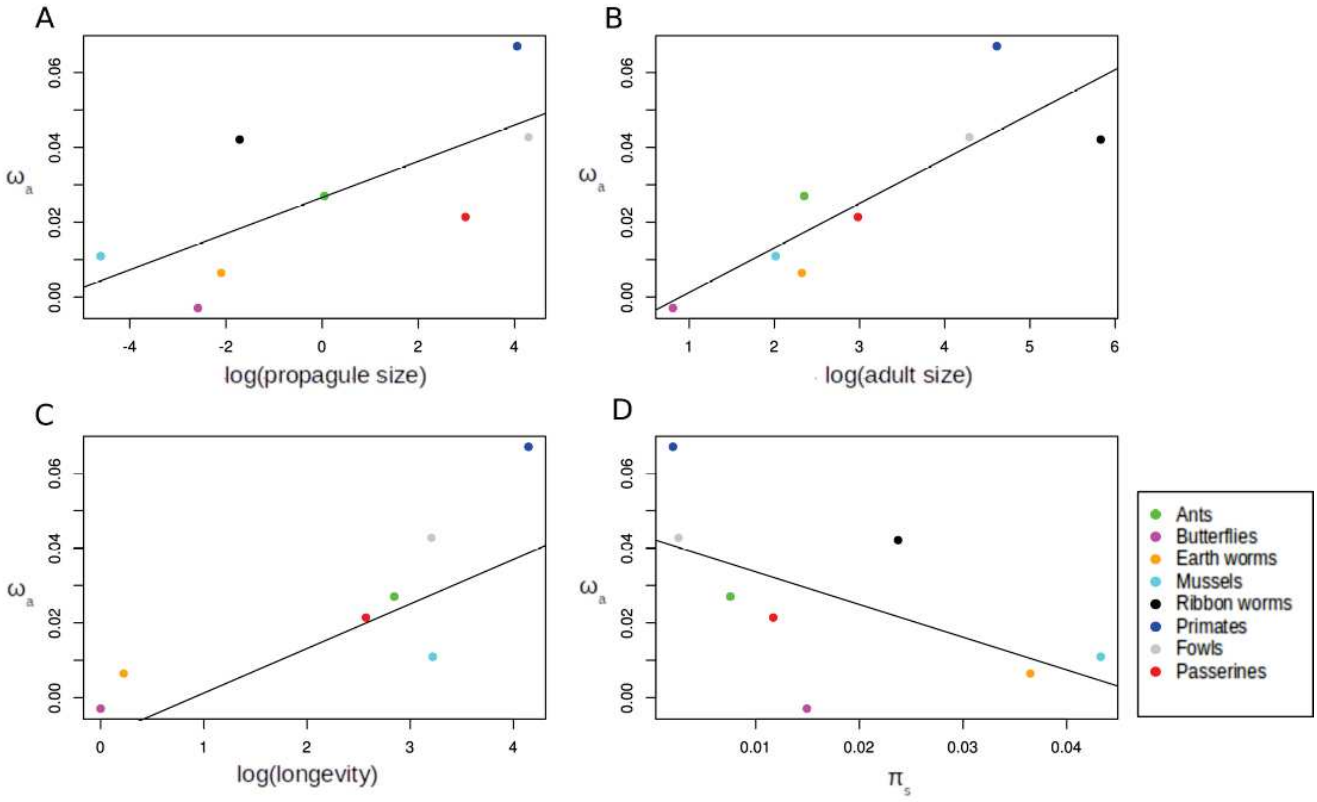


Figure 4 : Relationship between $\omega_{a[A]}$ and life history traits and π_s.
$\omega_{a[A]}$ is estimated using the average of ω_{na} estimates (estimated considering only GC-conservative mutations) across species within one taxonomic group and subtracting it to the dN/dS ratio estimated considering all branches within a tree for one taxonomic group.
A : Relationship between $\omega_{a[A]}$ and $\log_{10}$(propagule size) (cm).
B : Relationship between $\omega_{a[A]}$ and $\log_{10}$(adult size) (cm).
C : Relationship between $\omega_{a[A]}$ and $\log_{10}$(longevity) (years)).
D : Relationship between $\omega_{a[A]}$ and π_s.

To obtain this result, we took into account discrepancies between divergence data and polymorphism data using two different strategies that seems satisfying. However, one better way to correct the DFE-α method for this violation of its assumptions would consist in directly incorporate the observed variation of DFE within closely related species in the estimation process. One perspective of this work is then to assess the extent of demographic fluctuations that is necessary to yields the observed variation of DFE (or more simply the π_n/π_s ratio) between closely related species, and then develop the DFE-α method to account for this variation. Alternatively, a recently developed method allows the estimation of α and ω_a using polymorphism data alone (Tataru et al. 2017). This method does not rely on the assumption that the DFE is shared between the recent history of the ingroup and the past history since the divergence with the outgroup. However, the

estimation of α and ω_a by this method has a different meaning than the usual one, as it represents the rate of adaptive evolution of the species during its recent history, and not the one of its long-term history. This method also present the drawback that it require very high quality datasets. It may not be suitable for very low diversity species.

We also estimated the relationship between ω_a and life history traits related to N_e and π_s for individual species. We found a significant positive correlation between ω_a and adult size, and body mass, but no significant relationship with propagule size, fecundity, longevity or π_s (**Figure S4**). Note that we did not take into account the non-independence between ω_a and π_s throughout this study. We may consider re-running the analysis by splitting polymorphism data in two group, one being used to estimate ω_a and the other to estimate π_s. However, Galtier (2016) found no difference in his results whether or not controlling for non-independence between ω_a and π_s.

5. More adaptation in small populations?

We used two approaches to estimate the adaptive substitution rate at group level, and both support a negative relationship between ω_a and N_e. These results apparently contradicts the predominant theory suggesting that adaptation is more efficient in large population and previous empirical results (Gossmann et al. 2012). Our results, however, are compatible with theoretical predictions obtained under the Fisher's geometrical model (FGM) (Lourenço et al. 2013, Huber et al. 2017), which has been found to best explain the differences in DFE between human and *Drosophila* (Huber et al. 2017). Under this model, three different aspects of species biology and ecology have been identified as determinants of the adaptive substitution rate or the proportion of adaptive mutations, which may explain the negative relationship we observe between N_e and ω_a.

First, the mean distance of the population to the fitness optimum could be modulated by the long-term population size. Indeed, under FGM, the proportion of beneficial mutations increases with the distance to the optimum. Populations evolving under small long-term Ne are further away from the optimum related to larger populations, due to increased fixation of deleterious mutations at equilibrium, so they are predicted to undergo a larger proportion of beneficial mutations (Huber et al. 2017).

The second factor of importance is the rate of environmental change. The simulations of Lourenço et al. (2013) confirmed that under FGM and a moving optimum, the adaptive substitution rate increases linearly as a function of the rate of environmental change. It is not directly obvious why small-population size species would have a higher rate of environmental change, whether it be abiotic or biotic, and this is very difficult to evaluate empirically. One may speculate, however, that

species with long generation time may undergo a higher environmental change per generation and generation time is negatively correlated to population size in amniotes (Chao and Carr 1993).

The third variable of interest is species complexity. In FGM, complexity represents the dimensionality of the phenotypic space. Lourenço et al. (2013) suggested that this affects the adaptive substitution rate more strongly than does the effective population size : the adaptive substitution rate increases as a function of organismal complexity. Indeed, the probability for a new mutation to be in the optimal direction decreases as the number of potential directions increases. Therefore, the average adaptive walk in a high-dimensional space takes more steps than the average adaptive walk in a low-dimensional space, thus increasing the number of adaptive substitutions of small effect (Orr 2000; Lourenço et al. 2013). Here again, it is far from being straightforward that primates and birds would be more complex than mussels and worms, for instance. Complexity sensu FGM is very hard to quantify in a biologically relevant way. Different measures such as genome size, number of genes or proteins, number of protein-protein interactions and number of cell types have been used and seems to indicate that mammals are more complex than insects for instance (Valentine et al. 1994, Stumpf et al. 2008), which would be consistent with the idea of a greater complexity of species with smaller N_e. Additionally, Fernández and Lynch showed that the accumulation of mildly deleterious mutations in populations of small size induces secondary selection for protein–protein interactions that stabilize key gene functions, yielding a plausible mechanism for the emergence of molecular complexities (Fernández and Lynch 2011). If the number of protein-protein interactions is a relevant measure of proteome complexity, then this might contribute to explain why we find that the rate of adaptive substitution is higher in small-N_e species than in large-N_e species.

Another explanation to our results would be that small-N_e species adapt by a lot of small steps, whereas large-N_e species adapt with a few mutations of stronger fitness effect. As the DFE-α method estimates the number of adaptive steps and not their associated fitness effects, we expect a negative relation between N_e and ω_a if this hypothesis is confirmed. It is actually corroborated by FGM, under the hypothesis that high complexity species have on average lower population size, as discussed above (Orr 2000; Lourenço et al. 2013). Besides, it has been shown for instance that the human genome is submitted to many weak selective sweeps that are only detectable when averaging across many instances, suggesting that there may be many small-scale adaptive steps in humans (Enard et al. 2014).

Our results associated with the theoretical work based on FGM from Lourenço et al. (2013) and Huber et al. (2017) seem to suggest that one assumption underlying the main theory of the existence of a positive relationship between ω_a and N_e is not met, which is that the DFE should be

521 independent from N_e. Our results instead suggest that beneficial mutations could be more common
522 in small-N_e species than in large-N_e species, despite the fact that small-N_e species present a smaller
523 absolute number of total mutations. Additionally, they suggest that this larger proportion of
524 beneficial mutations in smaller populations would compensate for their lower fixation probability
525 due to increased drift.
526 Despite the fact that expectations of FGM are coherent with our empirical results, we need to keep
527 in mind that some of its assumptions (the one-peak shape of the fitness landscape and the universal
528 pleiotropy assumption, for instance), may not be clearly appropriate to model evolving proteins
529 (Lourenço et al. 2013).

530 **CONCLUSION**

531 In this study, we first explored how some biases can influence the results of the DFE-α method. We
532 showed that gBGC is likely to lead to the overestimation of ω_a in a lot of species of our sample, but
533 not all, leading to potential blurring when exploring the relationship between N_e and ω_a. We also
534 showed that divergence data is likely to represent an average of the selection/drift regime to which
535 the population has been submitted since its divergence with its sister species, whereas
536 polymorphism data represents a punctual selection/drift regime which is very unlikely to correspond
537 to the averaged selection/drift regime of the divergence time, which violates one assumption of the
538 DFE-α method.
539 Taking into account gBGC and differences in the DFE between the recent history of the ingroup and
540 the long-term history since the divergence with the outgroup, we obtain for the first time a negative
541 relationship between ω_a and life history traits indicators of long-term N_e. This result is in
542 contradiction with the hypothesis that DFE is independent from N_e, as this assumption underlies the
543 theory of the existence of a positive relationship between ω_a and N_e. In contrast, it is consistent with
544 the predictions that species with low long-term N_e experience a higher proportion of beneficial *de*
545 *novo* mutations (Huber et al. 2017), and that species with high complexity and high rate of
546 environmental changes have a higher adaptive substitution rate (Lourenço et al. 2013).
547 We recommend great prudence when interpreting absolute values yielded by the DFE-α method on
548 individual species whose polymorphism is unlikely to have been generated under the same DFE
549 than divergence data.

Supplementary Material:

Table S1 : Details of the species used in this study and numbers of individuals for each species.

Table S2 : Sources of the tree topologies of each taxonomic group used to estimate branch length and map substitutions.

Table S3: Values and sources of the life history traits used in this study.

Table S4: SNPs counts for each species.

SNPs counts are not integers because they corresponds to SNPs that are present in our SFS, where we chose a sample size (i.e. the number of categories of the SFS) lower that 2*the number of individuals. This is to compensate the uneven coverage between individuals that results in some sites in some individuals not to be genotyped. We chose sample sizes that maximize the number of SNPs in each SFS.

Figure S1: Network of contaminants in the *de novo* assembly generated after exon capture (Croco).

Figure S2: Relationship between ω_a estimated for each species using GC-conservative mutations and life history traits (log_{10} transformed) and π_s.

Acknowledgments

We thank Iago Bonicci for the homemade program that allowed us to remove intronic sequences of the contigs obtained during the capture experiment by identifying any incongruent correspondence or inconsistent overlap on both the transcriptomic reference and the assembly of the capture experiment contigs. I also thank Yoann Anselmetti and Thibault Leroy for their frequent help, and Yoann Anciaux for sharing his experience with Fisher's geometrical model.

Funding information

This work was supported by Agence Nationale de la recherche grant no. ANR-15-CE12-0010 'DarkSideOfRecombination'.

574 **References:**

Ballenghien M, Faivre N, Galtier N. 2017. Patterns of cross-contamination in a multispecies population genomic project: detection, quantification, impact, and solutions. BMC Biology 15:25.

Bolívar P, Mugal CF, Rossi M, Nater A, Wang M, Dutoit L, Ellegren H. 2018. Biased inference of selection due to GC-biased gene conversion and the rate of protein evolution in flycatchers when accounting for it. Molecular Biology and Evolution.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114-2120.

Cahais V, Gayral P, Tsagkogeorga G, Melo-Ferreira J, Ballenghien M, Weinert L, Chiari Y, Belkhir K, Ranwez V, Galtier N. 2012. Reference-free transcriptome assembly in non-model animals from next-generation sequencing data. Molecular ecology resources 12:834–845.

Chao L, Carr DE. 1993. The molecular clock and the relationship between population size and generation time. Evolution 47:688–690.

Corcoran P, Gossmann TI, Barton HJ, Slate J, Zeng K. 2017. Determinants of the efficacy of natural selection on coding and noncoding variability in two passerine species. Genome Biology and Evolution 9:2987–3007.

Douzery EJ, Scornavacca C, Romiguier J, Belkhir K, Galtier N, Delsuc F, Ranwez V. 2014. OrthoMaM v8: a database of orthologous exons and coding sequences for comparative genomics in mammals. Molecular Biology and Evolution 31:1923–1928.

Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome biology 16:157.

Enard D, Messer PW, Petrov DA. 2014. Genome-wide signals of positive selection in human evolution. Genome research.

Escobar JS, Glémin S, Galtier N. 2011. GC-biased gene conversion impacts ribosomal DNA evolution in vertebrates, angiosperms, and other eukaryotes. Molecular Biology and Evolution 28:2561–2575.

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. 2002. Changing effective population size and the McDonald-Kreitman test. Genetics 162:2017–2024.

Eyre-Walker A, Keightley PD. 2009. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. Molecular Biology and Evolution 26:2097–2108.

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

Fernández A, Lynch M. 2011. Non-adaptive origins of interactome complexity. Nature 474:502.

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

Galtier N. 2016. Adaptive protein evolution in animals and the effective population size hypothesis. PLOS Genetics 12:e1005774.

Galtier N, Roux C, Rousselle M, Romiguier J, Figuet E, Glémin S, Bierne N, Duret L. 2018. Codon usage bias in animals: disentangling the effects of natural selection, effective population size, and GC-Biased gene conversion. Molecular Biologu and Evolution 35:1092–1103.

Gayral P, Melo-Ferreira J, Glémin S, Bierne N, Carneiro M, Nabholz B, Lourenco JM, Alves PC, Ballenghien M, Faivre N. 2013. Reference-free population genomics from next-generation transcriptome data and the vertebrate–invertebrate gap. PLoS genetics 9:e1003457.

Gossmann TI, Keightley PD, Eyre-Walker A. 2012. The effect of variation in the effective population size on the rate of adaptive molecular evolution in eukaryotes. Genome Biology and Evolution 4:658–667.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature biotechnology 29:644.

Guéguen L, Gaillard S, Boussau B, Gouy M, Groussin M, Rochette NC, Bigot T, Fournier D, Pouyet F, Cahais V. 2013. Bio++: efficient extensible libraries and tools for computational molecular evolution. Molecular Biologu and Evolution 30:1745–1750.

Huang X, Madan A. 1999. CAP3: A DNA sequence assembly program. Genome research 9:868–877.

Huber CD, Kim BY, Marsden CD, Lohmueller KE. 2017. Determining the factors driving selective effects of new nonsynonymous mutations. Proceedings of the National Academy of Sciences 114:4465–4470.

Jensen JD, Bachtrog D. 2011. Characterizing the influence of effective population size on the rate of adaptation: Gillespie's Darwin domain. Genome Biology and Evolution 3:687–701.

Keith N, Tucker AE, Jackson CE, Sung W, Lledó JIL, Schrider DR, Schaack S, Dudycha JL, Ackerman M, Younge AJ. 2016. High mutational rates of large-scale duplication and deletion in *Daphnia pulex*. Genome research 26:60–69.

Kimura M. 1983. The neutral theory of molecular evolution. Cambridge University Press

Lassalle F, Périan S, Bataillon T, Nesme X, Duret L, Daubin V. 2015. GC-content evolution in bacterial genomes: the biased gene conversion hypothesis expands. PLoS Genetics 11:e1004941.

Lesecque Y, Mouchiroud D, Duret L. 2013. GC-biased gene conversion in yeast is specifically associated with crossovers: molecular mechanisms and evolutionary significance. Molecular Biologu and Evolution 30:1409–1419.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25:1754–1760.

Lourenço J, Galtier N, Glémin S. 2011. Complexity, pleiotropy, and the fitness effect of mutations. Evolution 65:1559–1571.

Lourenço JM, Glémin S, Galtier N. 2013. The rate of molecular adaptation in a changing environment. Molecular Biologu and Evolution 30:1292–1301.

Mancera E, Bourgon R, Brozzi A, Huber W, Steinmetz LM. 2008. High-resolution mapping of meiotic crossovers and non-crossovers in yeast. Nature 454:479–485.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in *Drosophila*. Nature 351:652–654.

Messer PW, Petrov DA. 2013. Frequent adaptation and the McDonald–Kreitman test. PNAS 110:8615–8620.

Meunier J, Duret L. 2004. Recombination drives the evolution of GC-content in the human genome. Molecular Biologu and Evolution 21:984–990.

Meyer M, Kircher M. 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harbor Protocols 2010:pdb. prot5448.

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

Ohta T. 1973. Slightly deleterious mutant substitutions in evolution. Nature 246:96.

Orr HA. 2000. Adaptation and the cost of complexity. Evolution 54:13–20.

Pessia E, Popa A, Mousset S, Rezvoy C, Duret L, Marais GAB. 2012. Evidence for widespread GC-biased gene conversion in eukaryotes. Genome Biology and Evolution 4:675–682.

Ranwez V, Delsuc F, Ranwez S, Belkhir K, Tilak M-K, Douzery EJ. 2007. OrthoMaM: a database of orthologous genomic markers for placental mammal phylogenetics. BMC evolutionary biology 7:241.

Ranwez V, Harispe S, Delsuc F, Douzery EJ. 2011. MACSE: Multiple Alignment of Coding SEquences accounting for frameshifts and stop codons. PloS one 6:e22594.

Rohland N, Reich D. 2012. Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. Genome research.

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

Rousselle M, Faivre N, Ballenghien M, Galtier N, Nabholz B. 2016. Hemizygosity enhances purifying selection: lack of fast-Z evolution in two Satyrine butterflies. Genome Biology and Evolution 8:3108–3119.

Rousselle M, Mollion M, Nabholz B, Bataillon T, Galtier N. 2018. Overestimation of the adaptive substitution rate in fluctuating populations. Biology Letters 14:20180055.

Simion P, Belkhir K, François C, Veyssier J, Rink JC, Manuel M, Philippe H, Telford MJ. 2018. A software tool ‘CroCo’detects pervasive cross-species contamination in next generation sequencing data. BMC biology 16:28.

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I. 2009. ABySS: a parallel assembler for short read sequence data. Genome research 19:1117–1123.

Smeds L, Mugal CF, Qvarnström A, Ellegren H. 2016. High-resolution mapping of crossover and non-crossover recombination events by whole-genome re-sequencing of an avian pedigree. PLoS genetics 12:e1006044.

Spencer CCA. 2006. Human polymorphism around recombination hotspots. Portland Press Limited.

Strasburg JL, Kane NC, Raduski AR, Bonin A, Michelmore R, Rieseberg LH. 2010. Effective population size is positively correlated with levels of adaptive divergence among annual sunflowers. Molecular Biology and Evolution 28:1569–1580.

Stumpf MP, Thorne T, de Silva E, Stewart R, An HJ, Lappe M, Wiuf C. 2008. Estimating the size of the human interactome. Proceedings of the National Academy of Sciences 105:6959–6964.

Tataru P, Mollion M, Glémin S, Bataillon T. 2017. Inference of distribution of fitness effects and proportion of adaptive substitutions from polymorphism data. Genetics 207:1103–1119.

Tilak M-K, Justy F, Debiais-Thibaud M, Botero-Castro F, Delsuc F, Douzery EJP. 2015. A cost-effective straightforward protocol for shotgun Illumina libraries designed to assemble complete mitogenomes from non-model species. Conservation Genetic Resources 7:37–40.

Tsagkogeorga G, Cahais V, Galtier N. 2012. The population genomics of a fast evolver: high levels of diversity, functional constraint, and molecular adaptation in the tunicate *Ciona intestinalis*. Genome Biology and Evolution 4:852–861.

Valentine JW, Collins AG, Meyer CP. 1994. Morphological complexity increase in metazoans. Paleobiology 20:131–142.

Webster MT, Axelsson E, Ellegren H. 2006. Strong regional biases in nucleotide substitution in the chicken genome. Molecular Biology and Evolution 23:1203–1216.

Wright AE, Harrison PW, Zimmer F, Montgomery SH, Pointer MA, Mank JE. 2015. Variation in promiscuity and sexual selection drives avian rate of faster-Z evolution. Molecular ecology 24:1218–1235.

Chapitre 4

The nature of molecular adaptation: dissecting the contribution of radical and conservative amino acid substitutions to adaptive and non-adaptive protein evolution in animals.

This last chapter presents the results of a student of Master « MEME » (Erasmus Mundus Master Programme in Evolutionary Biology), Hilde Schneemann, who worked under the shared supervision of Nicolas and I. It is not a complete work, but consists in preliminary results obtained during the three-month long internship of Hilde, as well as perspectives for future work to achieve this project in view of a publication.

The objective here is to include an additional aspect of sequence evolution in the DFE-α method : physiochemical properties of amino acid changes. Indeed, including this aspect could help refining our comprehension of the mode of adaptation, and in particular if proteins adaptation proceeds via changes of small or big fitness effects. This question has been inspired from a previous study of Smith (2003), where the McDonald & Kreitman test is used to estimate α for substitutions and mutations defined as conservative or radical in *Drosophila* (where a conservative change is a change from an amino acid towards an amino acid with similar physiochemical properties, and a radical change is a change from an amino acid towards an amino acid with different physiochemical properties). This study does not provide a clear conclusion as for the role of radical or conservative changes in adaptive evolution, which may be due to the fact that its results are based on a version of the McDonald & Kreitman approach that is sensible to the presence of slightly deleterious mutations. Thus, we aim in this project at improving the method used to estimate α and ω_a for radical and conservative mutations via the use of the DFE-α method, as well as increasing the dataset in terms of species.

Hilde's work in this respect has suffered from a lot of technical issues, in particular regarding the software allowing to map radical and conservative substitutions. Indeed, new tools to achieve this have proven to be imperfect at the time we conducted the analysis. The correction of these errors opens news perspectives to improve this work. Besides, after we obtained our first results, a new study came out that brings both answers and new insights for this project (Bergman and Eyre-Walker 2018). Thus, I present here preliminary results but also ideas as for future analyzes that seem relevant to meet our objectives.

References :

Bergman J, Eyre-Walker A. 2018. Does adaptive protein evolution proceed by large or small steps at the amino acid level? bioRxiv:379073.

Smith NGC. 2003. Are radical and conservative substitution rates useful statistics in molecular evolution? Journal of Molecular Evolution 57:467–478.

Title: The nature of molecular adaptation: dissecting the contribution of radical and conservative amino acid substitutions to adaptive and non-adaptive protein evolution in animals.

Authors : Rousselle M[1], Schneemann HFH[1], Simion P[1], Figuet E[1], Nabholz B[1], Galtier N[1].

[1]ISEM, Univ Montpellier, CNRS, EPHE, IRD, Montpellier, France.

Corresponding author: Marjolaine Rousselle

marjolaine.rousselle@umontpellier.fr

ABSTRACT

The central topic of this thesis is the proportion of amino acid differences observed between species that corresponds to adaptive events *vs*. quasi neutral changes without functional impact. A rich literature attempts to place the evolutionary trajectory of a genome on the neutral-adaptive continuum by estimating the proportion of adaptive substitutions, and the great majority of the studies use all amino acid changes without distinction. However, the biochemical properties of amino acids provide a prior relative to the fitness effects of different types of changes. Changes between amino acid which are different in terms of volume, charge and/or polarity (called radical) are more likely to impact the function of a protein than changes between amino acids that are similar in terms of those properties (called conservative). This project aims at taking into account the biochemical properties of amino acids in the analysis of the adaptive substitution rate of proteins. This implies to identify the most relevant classification of amino acid changes, which could then allow us to assess both if radical substitutions are under stronger purifying selection than conservative substitutions, as expected, and if the adaptive substitution rate is higher for radical than for conservative substitutions. Identifying the type of amino acid changes that are particularly prone to strong adaptive fitness effects may also help us understand the determinants of the adaptive substitution rate and the impact of the effective population size. The following presents preliminary results obtained using datasets introduced in the previous chapters, as well as ideas for future developments that would allow us to answer those questions.

INTRODUCTION

Quantifying the mode and abundance of positive selection driving divergence between species is central to the study of evolution, but this question remains a topic of debate due to the difficulty to disentangle non-selective and selective processes at the genomic level. Moving forward from the dichotomous neutralist-selectionist debate, there is now a rich literature that attempts to place the evolutionary trajectory of a genome on the neutral-adaptive continuum by estimating the proportion of adaptive substitutions (Fay et al. 2001; Barrier et al. 2003; Clark et al. 2003; Bierne and Eyre-Walker 2004; Bustamante et al. 2005; Waterson et al. 2005; Welch 2006; Boyko et al. 2008; Liti et al. 2009; Gossmann et al. 2010; Halligan et al. 2010; Carneiro et al. 2012; Hvilsom et al. 2012; Tsagkogeorga et al. 2012; Loire et al. 2013; Galtier 2016). Elaborating from the seminal work of McDonald & Kreitman (McDonald and Kreitman 1991), current methods for estimating the proportion of adaptive substitutions and the adaptive substitution rate are based on the estimation of parameters of the distribution of the fitness effects (DFE) of non-synonymous mutations extracted from the joint analysis of the synonymous and non-synonymous site frequency spectra (SFS). The expected dN/dS under near neutrality, called ω_{na}, is deduced from the estimated DFE, and the difference between observed and expected dN/dS provides an estimate of the per synonymous substitution rate of adaptive amino acid substitution, called ω_a, with $\omega_a=\alpha(dN/dS)$ and $\omega_{na}=(1-\alpha)(dN/dS)$ (Eyre-Walker et al. 2006; Keightley and Eyre-Walker 2007; Eyre-Walker and Keightley 2009; Galtier 2016; Tataru et al. 2017). The assumed distribution of fitness effects estimated in this method does not take into account the biochemistry of amino acid replacements and their effects on protein structure and function. However, the physiochemical properties of amino acid replacements can have important consequences on the protein function. Considering the physiochemical properties of amino acids may help understand with a higher resolution the mode and determinants of adaptation. In particular, it may help determine if adaptive evolution proceeds by large or small steps at the amino acid level.

This is why some authors suggested to differentiate non-synonymous mutations between two categories: radical and conservative. Radical mutations are the ones that change the physiochemical properties of the amino acids, whereas conservative mutations conserve those properties. Usually, relevant physiochemical properties are considered to be the volume, charge and polarity. Several classifications have been proposed and used over different studies (Sneath 1966; Epstein 1967; Clarke 1970; Grantham 1974; Goodman and Moore 1977; Miyata et al. 1979; Smith 2003; Sainudiin et al. 2005; Yampolsky and Stoltzfus 2005). Notably, the matrices of physiochemical distances developed by Grantham's (1974) and Miyata et al.'s (1979) are the most

commonly used, but some measures of dissimilarity between amino acids are based on the empirical probability of an amino acid being replaced by another (Tang et al. 2004). The difficulty of choosing a classification is emphasized by the fact that the various classifications do not agree with one another.

Initially, this differentiation between radical and conservative was proposed to replace the dN/dS ratio, to avoid the saturation issue associated with the use of synonymous substitutions between distantly related species (Gojobori 1983; Smith and Smith 1996), as well as providing a less conservative test for adaptive evolution. Indeed, the rationale behind the use of radical and conservative substitution was that radical substitutions are more likely to affect protein structure and function, both positively and negatively. Therefore, in both dN/dS and Kr/Kc ratio (i.e. the ratio of radical substitution rate on conservative substitution rate), the numerator represents the class of mutations supposedly of stronger selective effect than the denominator. As the difference in terms of selective pressures between radical and conservative is in theory lower than between synonymous and non-synonymous substitutions, genes under positive selection should more easily show a significantly elevated Kr/Kc ratio than dN/dS ratio.

Under the reasoning that radical changes have stronger fitness effects than conservative changes, exploring the adaptive and non-adaptive substitution rate of radical and conservative changes may help determine if adaptive evolution proceeds by large or small steps, a long standing question in evolution. Theoretical work based on the Fisher's geometric model suggests that most advantageous mutations would be of small effect (Fisher 1931). However, more recent work in this framework suggests that this may depends on the distance of the population to its fitness optimum, which itself depends of its long-term population size for instance, as the further a population is from this optimum, the larger the effects are on average (MacLean and Buckling 2009; Huber et al. 2017). Besides, this theoretical work concerns *de novo* mutations, but it is unclear whether the pattern is the same for mutations that have reached fixation. To answer this question of the relative contribution of large and small effect mutations to the adaptive substitution rate, the two types of mutations have previously been used as part of the McDonald & Kreitman test in *Drosophila* (Smith 2003). If, as predicted, radical mutations are shown to be more strongly deleterious than conservative ones, this study does not find that they are more strongly or often adaptive. However, this result has been obtained without any consideration of segregating slightly deleterious mutations, which can bias the estimates. Very recently (after we started this project), a study also focusing on *Drosophila* showed that both the adaptive and non-adaptive evolution of proteins is

dominated by substitutions between amino acids that are more similar (Bergman and Eyre-Walker 2018), in agreement with some theoretical work.

Large populations are expected to have more efficient purifying and positive selection, and a larger supply of beneficial mutations. Consequently, theory predicts a positive correlation between the effective population size (N_e) and the adaptive substitution rate (Lanfear et al. 2014). However, there is very few empirical support for this correlation (Strasburg et al. 2010; Gossmann et al. 2012, Galtier 2016), and we even show opposite results in the previous chapter (**Chapter 3**). If one of the two types of changes, radical or conservative, is particularly prone to produce adaptive changes, thus estimating the adaptive substitution rate using only this type of amino acid substitutions may reveal a potential effect of N_e.

Thus, this project aims at answering three questions :
(i) Are radical substitutions under stronger purifying selection than conservative substitutions ?
(ii) Is the adaptive rate higher for radical than for conservative substitutions ?
(iii) Can the adaptive substitution rate be estimated more reliably when distinguishing radical from conservative substitutions ?

To answer these questions, we have for the moment conducted a preliminary study using five datasets comprising catarrhine primates, fowls (Galloanserae), earth worms (Lumbricidae), ants (Formicinae) and butterflies (Satyrinae). Despite our results presented in **Chapter 3**, showing that estimates of the DFE-α method on individual species may not be accurate estimates of the adaptive substitution rate, we conducted a “classic” DFE-α approach to start with. We estimated site frequency spectra, nucleotide diversities and substitution rates (total, adaptive and non-adaptive) differentiating radical and conservative non-synonymous mutations using a standard classification (Sainudiin et al. 2005).

Our results indicate that purifying selection is stronger on radical mutations, which may mean that these have stronger deleterious fitness effects. However, our results do not allow to conclude regarding the adaptive fitness effects of radical *vs.* conservative mutations, and using only radical or conservative mutations in the estimation of the adaptive substitution rate did not uncover any relationship between this statistic and N_e. Many aspects of this work can be improved, as discussed, which might hopefully lead to clearer answers to our questions.

MATERIAL & METHODS

1. Data

For this project, we used five datasets representing five taxonomic groups : catarrhine primates, fowls (Galloanserae), earth worms (Lumbricidae), ants (Formicinae) and butterflies (Satyrinae). Details of data generation (for earth worms, butterflies and ants) and recovering from existing databases (for catarrhine primates and fowls) are presented in **Chapter 3**. We plan to apply the following protocol on all the taxonomic groups available, i.e. all the datasets used in **Chapter 3**.

2. Classification of amino acid changes between radical and conservative

Polarity and volume have been shown to be among the most important properties for amino acid substitution frequency (Smith 2003; Sainudiin et al. 2005). For our preliminary study, radical and conservative substitutions were defined based on a combination of polarity and volume, following Sainudiin et al. 2005. The grouping was: large polar, {Y, W, H, K, R, E, Q}; small polar, {T, D, N, S, C}; small non-polar, {A, G, P, V}; large non-polar, {L, I, F, M}. Any change within these four categories was classified as conservative, and any change between these categories was classified as radical. Initially, we wanted to test several classifications to identify the more relevant. However, the tool we use to map substitution (MapNH, Guéguen et al. 2013a) was only available with the classification following Sainudiin et al. (2005). A new version of this program that allows the user to customize the classification is available since January 2018 (Bio++ version 2.4.0.), but we detected errors in this new version. Those errors has been reported and when they will be fixed, the new version of MapNH will allow to chose any classification (see **Discussion and perspectives**).

3. Divergence estimates computation

We estimated branch lengths of the five phylogenetic trees using bppml (version 2.4) and we mapped synonymous, radical and conservative substitutions with MapNH (version 2.3.2) (Guéguen et al. 2013b) on species specific branches. Tree topologies were obtained from the literature (see **Material & Method** section of **Chapter 3**). For fowls and primates, we kept only alignments comprising all the species. Butterflies, ants and earth worms datasets are smaller, therefore we also kept alignments comprising all species but one for those datasets. We controlled for the number of radical and conservative sites by counting the number of opportunities for all the sites in a given

sequence to mutate to a synonymous, radical or conservative change considering a κ parameter which represents the ratio of transition over transversion as measured by the bppml software. Indeed, it has been shown that ignoring this parameter could lead to underestimation of Kr and Kc (Dagan et al. 2002), as transversions lead to radical changes more often than transitions (Zhang 2000). We estimated Kr/Kc ratio, as well as Kr/dS and Kc/dS by summing substitutions and opportunities of one type across genes and then computing the ratio of the sums. Ninety-five per cent confidence intervals were estimated by bootstrapping genes (1000 replicates).

4. Polymorphism estimates computation

For each taxon, we estimated ancestral sequences at each internal node of the tree from the alignments with Bio ++ software SeqAncestor (Guéguen et al. 2013b). The ancestral sequences at each internal node were used to orientate mutations of species that descend from this node. We computed unfolded site frequency spectra (SFS) for synonymous, conservative and radical SNPs, as well as the average number of synonymous, radical and conservative SNPs between two copies of a gene within a species. We divided these average numbers of SNPs by the numbers of synonymous, radical or conservative sites computed in the same manner than for divergence estimates to compute synonymous (π_s), radical (π_R) and conservative (π_C) nucleotide diversity.

5. Computation of the adaptive and non-adaptive substitution rate for radical and conservative substitutions

We estimated the adaptive (ω_a) and non-adaptive (ω_{na}) substitution rate of radical and conservative substitutions using the method of Eyre-Walker and Keightley (2009) as implemented in Galtier 2016. It models the distribution of the fitness effects (DFE) of non-synonymous (here either radical or conservative) mutations as a negative gamma distribution for deleterious mutations and an exponential distribution for adaptive mutations, which is fitted to the synonymous and radical (or conservative) site frequency spectra (SFS) (model “GammaExpo” in Galtier 2016). The estimated DFE is then used to deduce the expected Kr/dS (or Kc/dS) under near-neutrality. The difference between observed and expected Kr/dS (or Kc/dS) provides an estimate of the proportion of adaptive radical (or conservative) substitutions, α_R (or α_C). The per mutation rate of adaptive and non-adaptive amino acid substitution were then obtained as following: $\omega_{a[R]} = \alpha_R(Kr/dS)$ and $\omega_{na[R]} = (1-\alpha_R)(Kr/dS)$ ($\omega_{a[C]} = \alpha_C(Kc/dS)$ and $\omega_{na[C]} = (1-\alpha_C)(Kc/dS)$). When estimating DFE model parameters, we accounted for demographic effects by fitting nuisance parameters, which correct

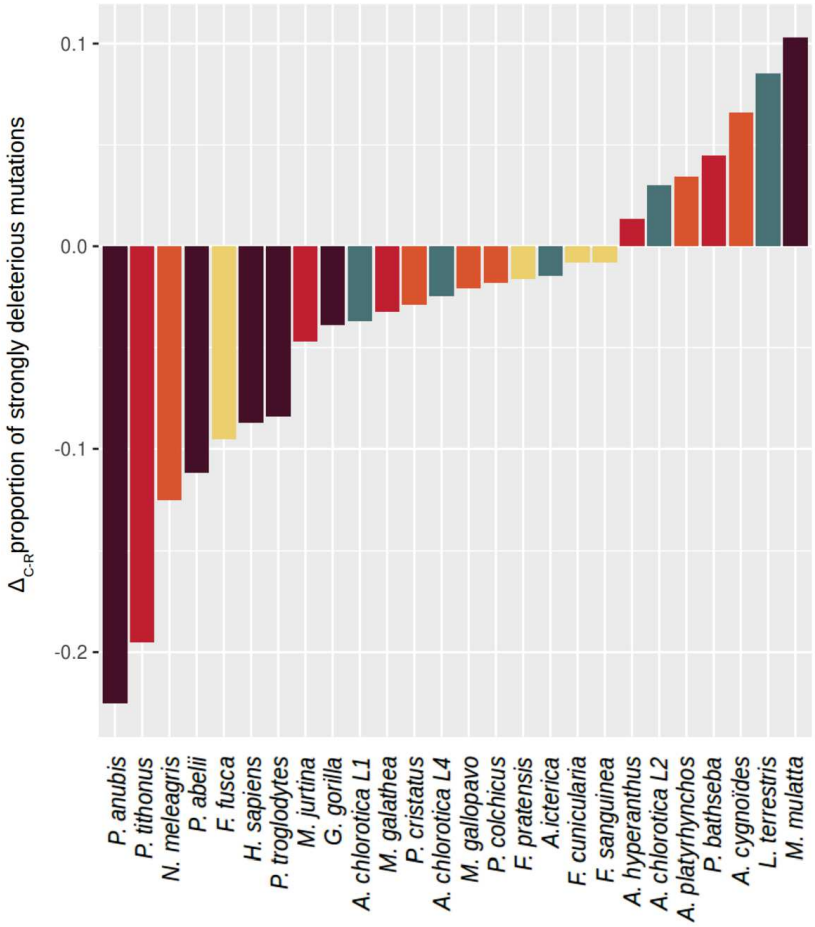
each class of frequency of the synonymous and radical or conservative SFS relative to the neutral expectation in an equilibrium Wright–Fisher population (Eyre-Walker et al. 2006).

PRELIMINARY RESULTS

1. Are radical mutations under stronger purifying selection than conservative ones ?

Comparing the estimated DFE of radical and conservative mutations, we found that the proportion of strongly deleterious mutations is larger for radical mutations than conservative mutations in a significant majority of species (19 out of 26 , binomial test p-value: 1.45e-02, **Figure 1**). Comparing Kr/dS with Kc/dS and π_R/π_S with π_C/π_S revealed that, for both divergence and polymorphism, the conservative ratio is consistently higher than its radical counterpart in a majority of species. π_C/π_S was larger than π_R/π_S in a significant majority of species (26 out of 27 species, one-sided binomial test p-value: 2.09e-07). The same holds true for Kc/dS and Kr/dS (21 out of 27 species, one-sided binomial test p-value: 2.82e-06) (**Figure 2**). We also found that $\omega_{na[C]}$ is higher than $\omega_{na[R]}$ in 21 out of 27 species (one-sided binomial test p-value: 1.45e-02). However, confidence intervals often overlap, hence this difference was significant in only 11 species.

This tendency is generally observed within each taxon, as $\omega_{na[C]}$ is higher than $\omega_{na[R]}$ in all ants, all butterflies, all worms but one and all primates but one. These findings provide strong support for the hypothesis that purifying selection acts more strongly on radical mutations than on conservative ones. Furthermore, this finding shows that the classification of radical and conservative used here captures some biochemical properties relevant to negative selection.



199 **Figure 1: Difference between the proportion of strongly deleterious conservative mutations**
200 **and the proportion of strongly deleterious radical mutations.**
201 Colors correspond to the taxonomic group (brown: primates, red: butterflies, orange: fowls, yellow:
202 ants, blue: earth worms).

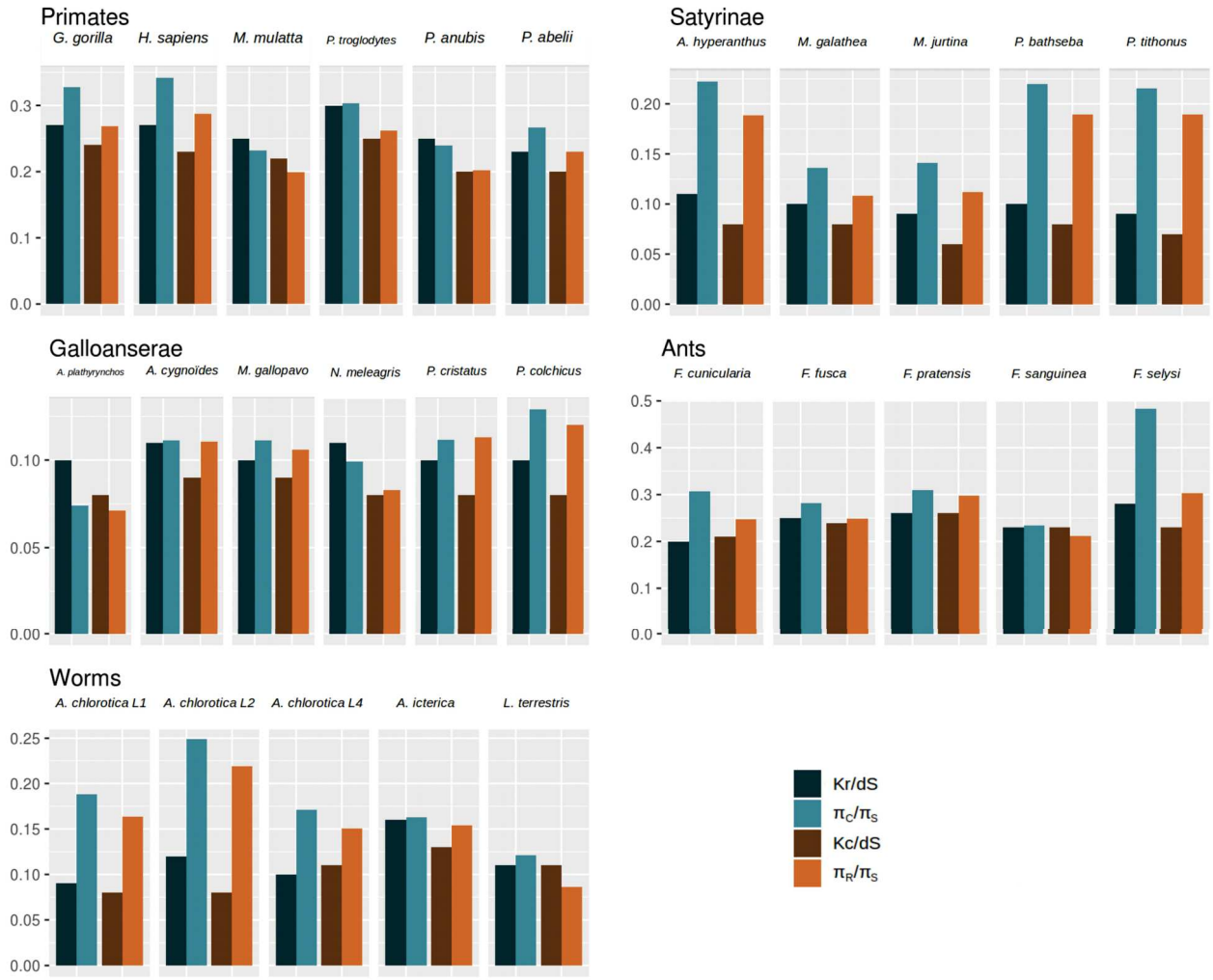
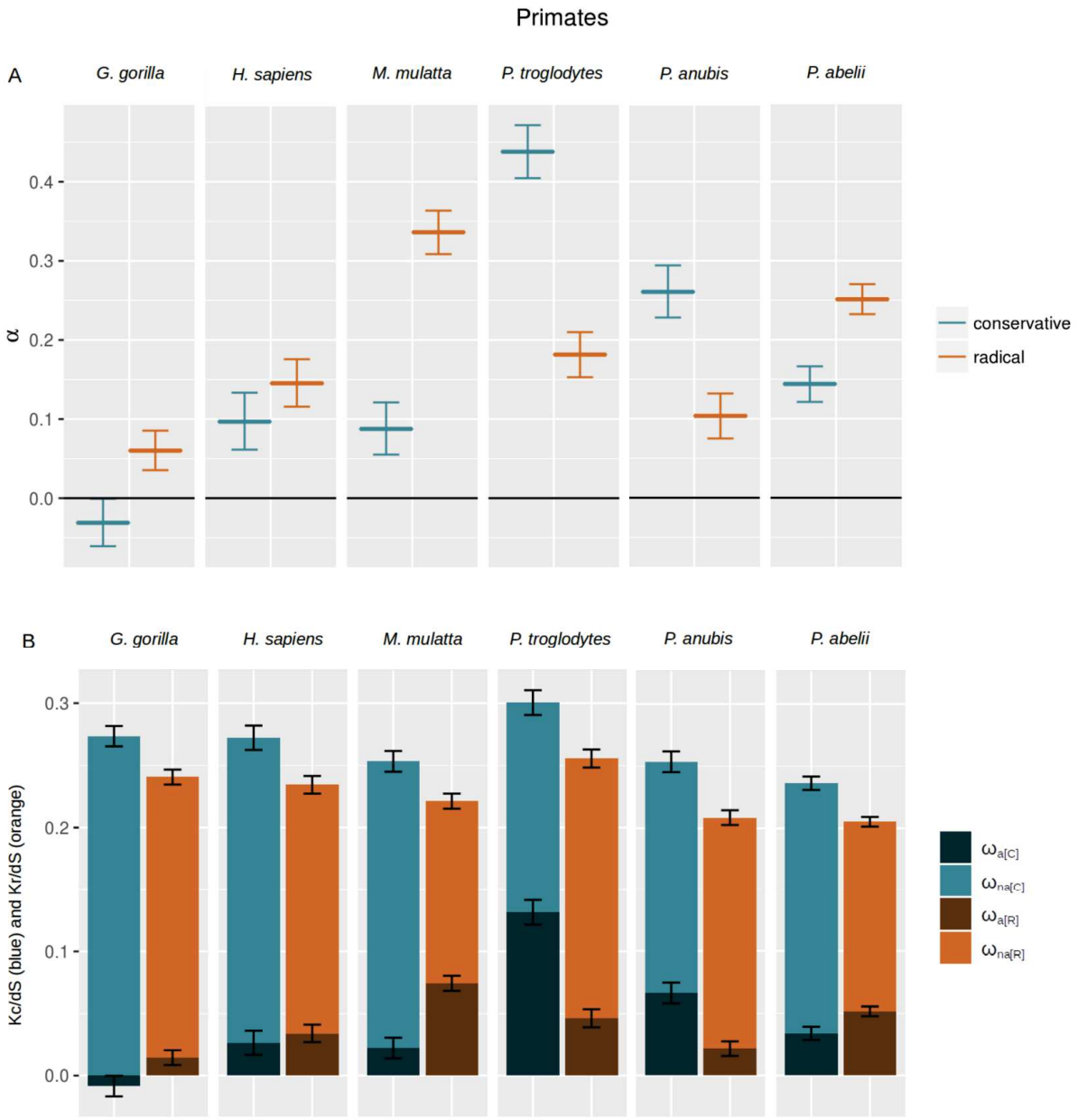


Figure 2: Comparison of Kr/dS, Kc/dS, π_R/π_S and π_C/π_S for all species.

2. Is the adaptive substitution rate higher for radical than for conservative substitutions ?

Primates

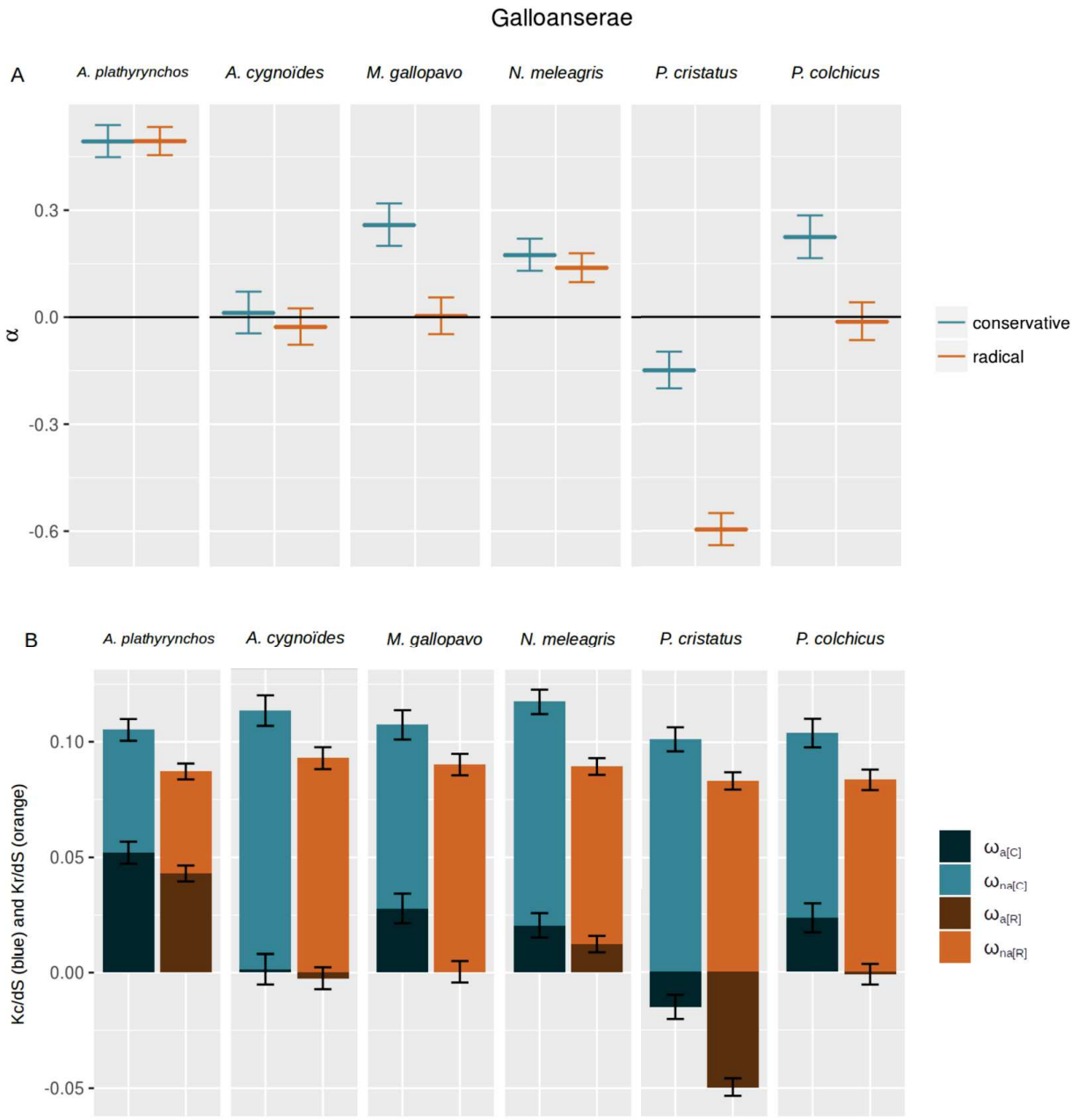
In primates, we found no consistent pattern in the difference between α_C and α_C (**Figure 3**). In two primate species, *Macaca mulatta* and *Pongo abelii*, α_R is significantly higher than α_C. In these two cases, $\omega_{a[R]}$ is also significantly higher than $\omega_{a[C]}$, which means that the effect on the proportion α is not solely due to a difference in ω_{na}. Conversely, in *Pan troglodytes* and *Papio anubis* α_R is significantly lower than α_C which can similarly be attributed to a corresponding difference between $\omega_{a[R]}$ and $\omega_{a[C]}$. In *Homo sapiens* there is no significant difference between $\omega_{a[R]}$ and $\omega_{a[C]}$. Eventually, in *Gorilla gorilla*, α_C and $\omega_{a[C]}$ are estimated to be negative, which makes the comparison difficult to interpret.



214 **Figure 3: Comparison of α_C and α_R (A) and contribution of $\omega_{a[R]}$ and $\omega_{na[R]}$ as well as $\omega_{a[C]}$ and**
215 **$\omega_{na[C]}$ to respectively Kr/dS and Kc/dS (B) in the primate dataset.**

216 ***Fowls***

217 In fowls, we observed a different pattern : we found that α_C is higher than α_R for all species
218 except *Anas platyrhynchos*, for which they are identical (**Figure 4**). For *Anser cygnoides*, *Pavo*
219 *cristatus* and *Phasianus colchicus*, one or both adaptive substitution rate were estimated to be
220 negative. This corresponds to an excess of π_R/π_S and π_C/π_S over Kr/dS and Kc/dS in these species
221 (**Figure 2**).



222 **Figure 4: Comparison of α_C and α_R (A) and contribution of $\omega_{a[R]}$ and $\omega_{na[R]}$ as well as $\omega_{a[C]}$ and**
223 **$\omega_{na[C]}$ to respectively Kr/dS and Kc/dS (B) in the fowls dataset.**

Butterflies, worms and ants

In butterflies and worms we see no consistent difference between α_R and α_C, similar to primates (i.e. α_R is sometimes higher and sometimes lower than α_C) (**Figure 5**). In contrast, α_R is consistently higher than α_C in ants, and $\omega_{a[R]}$ is also consistently higher than $\omega_{a[C]}$. Nevertheless, those datasets comprise less genes and less SNPs than the two previous ones, and this leads to confidence intervals to overlap substantially, indicating that the significance of a difference in α or ω_a cannot be asserted. Here also we obtained some negative values of α and ω_a, these negative values corresponding to an excess of π_R/π_S and π_C/π_S over Kr/dS and Kc/dS (**Figure 2**).

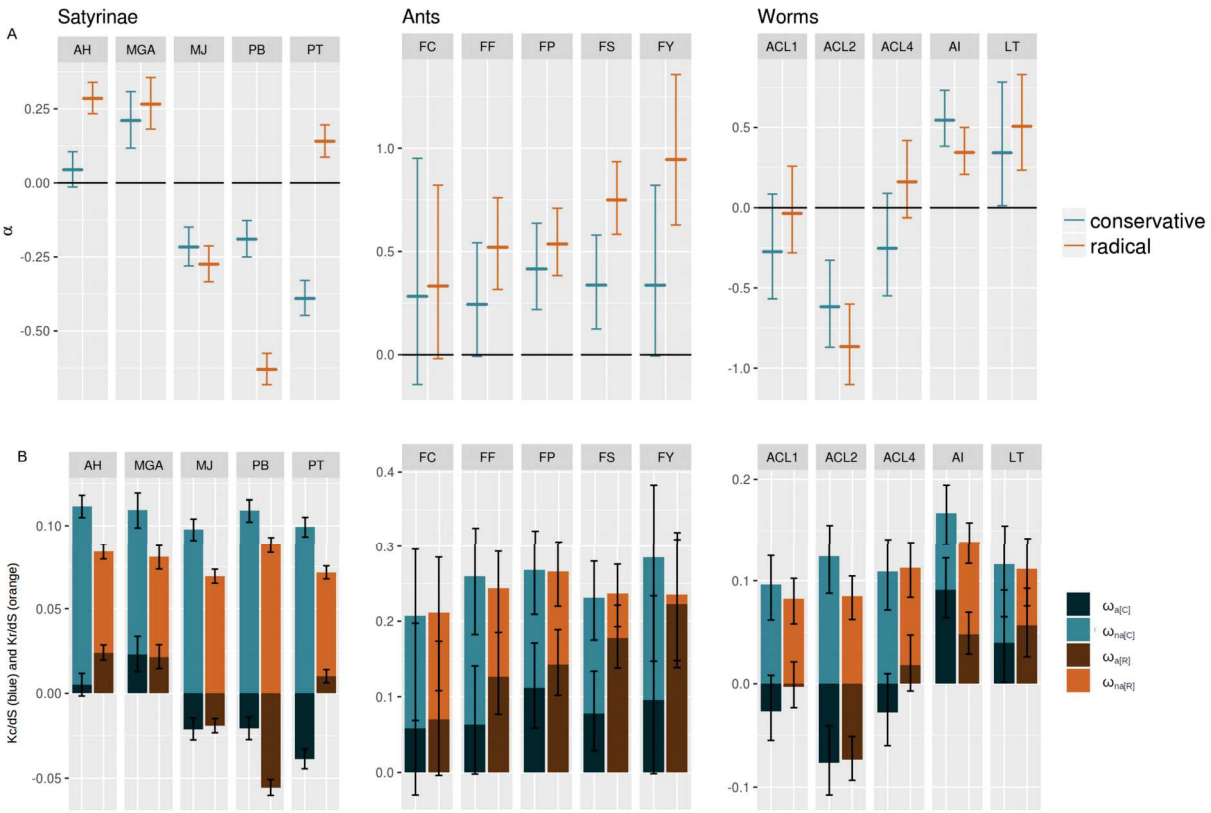


Figure 5: Comparison of α_C and α_R (A) and contribution of $\omega_{a[R]}$ and $\omega_{na[R]}$ as well as $\omega_{a[C]}$ and $\omega_{na[C]}$ to respectively Kr/dS and Kc/dS (B) for butterflies, earth worms and ants.

Species names are simplified as follows: AH: *A. hyperanthus*; MJ: *M. jurtina*; MGA: *M. galathea*; PB: *P. bathseba*; PT: *P. tithonus*; FS: *F. sanguinea*; FC: *F. cunicularia*; FY: *F. selysi*; FP: *F. pratensis*; FF: *F. fusca*; ACL1: *A. chlorotica L1*; ACL2: *A. chlorotica L2*;ACL4: *A. chlorotica L4*; AI: *A. icterica*; LT: *L. terrestris*.

The results presented here suggest that there is no consistent difference across taxa between the adaptive substitution rate of radical and conservative mutations (**Figure 3-5**). We could detect a significantly higher $\omega_{a[R]}$ than $\omega_{a[C]}$ only in *Aphantopus hyperantus*, *Macaca mulatta* and *Pongo abelii*, and a significantly higher $\omega_{a[C]}$ than $\omega_{a[R]}$ *Pan troglodytes* and *Papio anubis*.

3. Is there a link between N_e and the adaptive substitution rate of radical or conservative mutations?

Two-sided Spearman rank test revealed no significant correlation between π_S and $\omega_{a[C]}$ or $\omega_{a[R]}$ (p-values of 0.89 and 0.28 respectively, **Figure 6**). We found a negative significant correlation between π_S and $\omega_{na[R]}$ but not for $\omega_{na[C]}$ (p-values of 0.039 and 0.14 respectively).

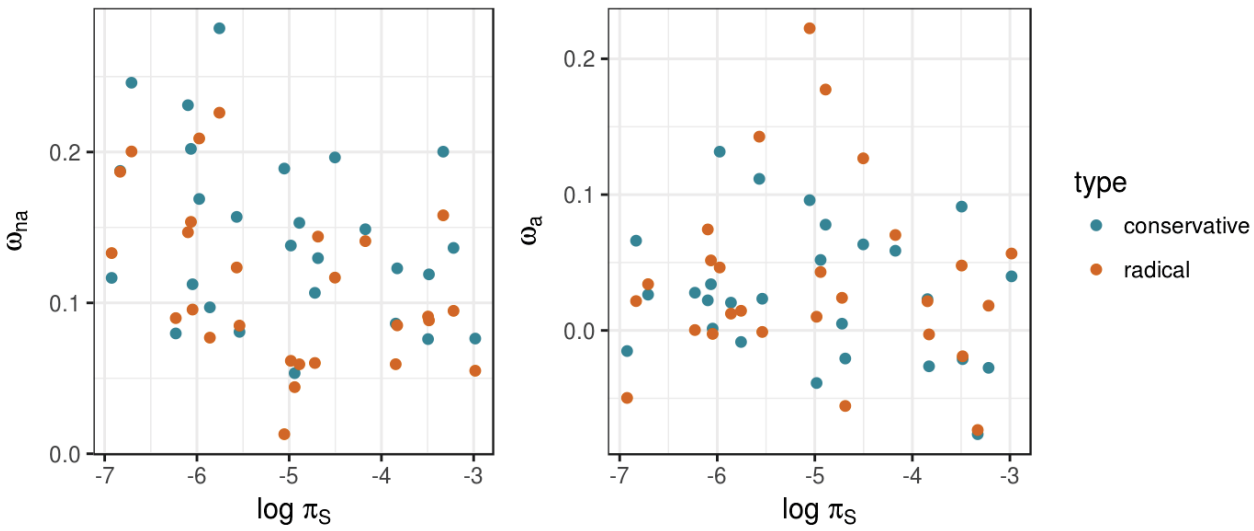


Figure 6 : Relationship between ω_{na} and $\log_{10}$ transformed π_S (left), and ω_a and $\log_{10}$ transformed π_S (right) using either only radical or conservative non-synonymous substitutions for each species of the dataset.

DISCUSSION AND PERSPECTIVES

1. Radical amino acid changes are under stronger purifying selection but not positive selection in most species

Our results provide support for the hypothesis that radical mutations are more often deleterious than conservative ones, and that they are submitted to stronger purifying selection, which confirms previous results (Zhang 2000; Bergman and Eyre-Walker 2018). This means that the way we classified radical and conservative mutations captures properties that affect protein functions.

However, our results do not validate that radical mutations have stronger adaptive fitness effects or higher adaptive substitution rate, or the opposite. We do not find a consistent pattern between the taxa we investigated, neither within the taxa between the different species. This could indicate that the classification used here does not consider amino acid properties that are relevant regarding adaptation. For instance, in Bergman and Eyre-Walker (2018), where they could conclude that adaptation of protein coding sequences is dominated by amino acid changes that are of small effect, they did not classify amino acid changes between radical and conservative, but estimated a distance in terms of volume and polarity of each amino acid pair separated by one mutation step. We discuss below other possibilities to classify radical and conservative amino acid changes that may better capture properties that are relevant regarding adaptation (**part 3** of this discussion).

Our preliminary analysis surprisingly yielded a lot of negative values of α and ω_a. This is usually interpreted as the sign that some of the assumptions of the method are not met in real data. Here we found that π_R/π_S and π_C/π_S ratios frequently exceed Kr/dS and Kc/dS. If, for instance, the selection regime has recently been relaxed, there will be an excess of slightly deleterious mutations in the SFS and π_R/π_S and π_C/π_S will be high. As a result, the estimated $\omega_{na[R]}$ and $\omega_{na[C]}$ will be inflated relative to the observed Kr/dS and Kc/dS, resulting in low or negative values of α and ω_a. However, in the light of results presented in **Chapter 3**, some of those negative values seem to need re-analysis. Indeed, in **Chapter 3** only three species (two butterflies and one ant) yielded negative values when considering all amino acid changes together, whereas here five species show both $\omega_{a[C]}$ and $\omega_{a[R]}$ values that are negative. Additionally we showed that using the DFE-α method on individual species whose polymorphism data is unlikely to have been generated under the same DFE than divergence data may lead to spurious estimations. We thus consider to follow the same

strategies as presented in **Chapter 3** to obtain one radical adaptive substitution rate estimate and one conservative adaptive substitution rate estimate per taxonomic group.

2. The relationship between N_e and the adaptive substitution rate

As discussed in **Chapter 3**, a positive correlation between N_e and the adaptive substitution rate may be expected for two reasons : first, the fixation probability of a new adaptive mutation of coefficient of selection s is proportional to N_e if s is small (Kimura 1983). Second, there is a higher supply of mutations in large populations, and among these, a higher proportion of mutations are expected to be effectively adaptive (i.e. $N_es >>1$) (Gossmann et al. 2012). However, empirical evidence for such correlation has proven to be difficult to establish. Gossmann et al. (2012) found no significant correlation between N_e and ω_a in 13 pairs of species of eukaryotes, but they found a positive correlation between the two variables within three taxonomic groups (Plants, Drosophilidae and Mammals) (Gossmann et al. 2012). Galtier (2016) found no significant correlation between N_e and ω_a in 44 pairs of non-model animals (Galtier 2016). Finally, we detected in **Chapter 3** a negative relation between ω_a and proxies of long-term N_e. Here, no correlation between population size (measured via the proxy π_s) and the adaptive substitution rate for either radical nor conservative substitutions was detected in our analysis. Furthermore, population size was found to be correlated with the non-adaptive substitution rate only for radical substitutions. These correlations may be improved by several means: first, estimates of ω_a or ω_{na} and π_s are not independent because both are estimated from the synonymous SFS. To correct for this issue, we may use the strategy used in Gossmann et al. (2012), where synonymous sites were randomly split in two halves, one being used to estimate π_s, and the other one to estimate α, ω_a and ω_{na}. We may also, as in **Chapter 3**, use proxies of N_e other than π_s, like propagule size, longevity or body mass. Finally, we may want to take into account the phylogenetic non-independence among species in the correlation.

3. What is the best classification of amino acid changes ?

At the beginning of this project, we planned to test several classifications of radical *vs.* conservative changes. Indeed, we thought that we had a tool to map substitutions of any type onto a tree, as the new version of MapNH allows the user to specify any list of codon pairs to be mapped separately. However, this new version has proven to be imperfect at the time this project was lead by our student, Hilde. This led us to conduct the preliminary analysis using the classification implemented in an older version of MapNH following the model « Polarity and/or volume » of (Sainudiin et al. 2005; Guéguen et al. 2013b)

However, other classifications may yields different results. We may use only polarity or volume, or a combination of charge, polarity and volume all at once. Alternatively, Grantham suggested the use of a matrix of distances between each amino acid pairs (a physiochemical measure) which identifies the chemical factors (here again charge, volume and polarity) that individually correlate best with evolutionary exchangeability of protein residues (Grantham 1974). By the use of such a distance matrix, one can either determine a threshold above which an amino acid change is considered as radical and below which it is defined as conservative, or estimate a statistic (like ω_a) for every amino acid pair. This way, a correlation between ω_a for each amino acid pair and each distance can be computed. This strategy has been used in Bergman and Eyre-Walker (2018), for instance, with another matrix than the one established by Grantham though.

Eventually, one may estimate a distance between each amino acid pair using the observed frequency of the substitution between them in a given dataset, under the hypothesis than the most radical a mutation is, the less frequently it can reach fixation. The risk is to obtain very few SNPs in the categories of the more radical changes. To overcome this issue, we suggest that we could pool SFS from closely related species to increase the number of SNPs in SFS of each category, as we did in the project presented in **Chapter 3**.

4. What are the next steps?

The analyses of the present study have been conducted by a student during an internship of four month, under my supervision and Nicolas' one. In view of the tight schedule, these analyses yielded only preliminary results that require to be deepened. In particular, it seems of utmost importance to test other ways to classify amino acid changes, whether in two categories or in as many categories as they are of amino acid pairs separated by one mutational step. For this, we plan to test the classifications mentioned in the previous section.

The numerous negative values of α that we obtained also require further explanations, and we may want to conduct this analysis again to see if it still hold true, as well as regrouping data of closely related species following the strategies presented in the previous chapter.

Additionally, there may be differences in the mode of adaptation in distinct taxonomic group with different life history traits, which may lead to a different distribution of radical or conservative amino acid changes in adaptive evolution. We already started to investigate different taxonomic groups, but we may want to increase the sampling of such groups, as in **Chapter 3** of this thesis. Indeed, as we mentioned above, we found for the moment no correlation between population size

(measured via the proxy π_s) and the adaptive substitution rate, despite the separation of radical and conservative substitutions. Combining the strategy used in **Chapter 3** to alleviate the risk of biases in the estimation of α and the separation in several categories of amino acid changes may enhance our understanding of the determinants of the adaptive substitution rate.

Acknowledgments

We thank Yoann Anselmetti and Thibault Leroy for their frequent help with bio-informatics.

Funding information

This work was supported by Agence Nationale de la recherche grant no. ANR-15-CE12-0010 'DarkSideOfRecombination'.

References:

Barrier M, Bustamante CD, Yu J, Purugganan MD. 2003. Selection on rapidly evolving proteins in the *Arabidopsis* genome. Genetics 163:723–733.

Bergman J, Eyre-Walker A. 2018. Does adaptive protein evolution proceed by large or small steps at the amino acid level? BioRxiv:379073.

Bierne N, Eyre-Walker A. 2004. The genomic rate of adaptive amino acid substitution in *Drosophila*. Molecular Biology and Evolution 21:1350–1360.

Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, Lohmueller KE, Adams MD, Schmidt S, Sninsky JJ, Sunyaev SR. 2008. Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS genetics 4:e1000083.

Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, Glanowski S, Tanenbaum DM, White TJ, Sninsky JJ, Hernandez RD. 2005. Natural selection on protein-coding genes in the human genome. Nature 437:1153.

Carneiro M, Albert FW, Melo-Ferreira J, Galtier N, Gayral P, Blanco-Aguiar JA, Villafuerte R, Nachman MW, Ferrand N. 2012. Evidence for widespread positive and purifying selection across the European rabbit (*Oryctolagus cuniculus*) genome. Molecular Biology and Evolution 29:1837–1849.

Clark AG, Glanowski S, Nielsen R, Thomas PD, Kejariwal A, Todd MA, Tanenbaum DM, Civello D, Lu F, Murphy B. 2003. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. Science 302:1960–1963.

Clarke B. 1970. Selective constraints on amino acid substitutions during the evolution of proteins. Nature 228:159–160.

Dagan T, Talmor Y, Graur D. 2002. Ratios of radical to conservative amino acid replacement are affected by mutational and compositional factors and may not be indicative of positive darwinian selection. Molecular Biology and Evolution 19:1022–1025.

Epstein CJ. 1967. Non-randomness of ammo-acid changes in the evolution of homologous proteins. Nature 215:355.

Eyre-Walker A, Keightley PD. 2009. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. Molecular Biology and Evolution 26:2097–2108.

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

Fay JC, Wyckoff GJ, Wu C-I. Positive and negative selection on the human genome. Genetics 158:1227–1234.

Fisher RA. 1931. XVII.—The distribution of gene ratios for rare mutations. Proceedings of the Royal Society of Edinburgh 50:204–219.

Galtier N. 2016. Adaptive protein evolution in animals and the effective population size hypothesis. PLOS Genetics 12:e1005774.

Gojobori T. 1983. Codon substitution in evolution and the" saturation" of synonymous changes. Genetics 105:1011–1027.

Goodman M, Moore GW. 1977. Use of Chou-Fasman amino acid conformational parameters to analyze the organization of the genetic code and to construct protein genealogies. Journal of Molecular Evolution 10:7–47.

Gossmann TI, Keightley PD, Eyre-Walker A. 2012. The effect of variation in the effective population size on the rate of adaptive molecular evolution in eukaryotes. Genome Biology and Evolution 4:658–667.

Gossmann TI, Song B-H, Windsor AJ, Mitchell-Olds T, Dixon CJ, Kapralov MV, Filatov DA, Eyre-Walker A. 2010. Genome wide analyses reveal little evidence for adaptive evolution in many plant species. Molecular Biology and Evolution 27:1822–1832.

Grantham R. 1974. Amino acid difference formula to help explain protein evolution. Science 185:862–864.

Guéguen L, Gaillard S, Boussau B, Gouy M, Groussin M, Rochette NC, Bigot T, Fournier D, Pouyet F, Cahais V. 2013a. Bio++: efficient extensible libraries and tools for computational molecular evolution. Molecular Biology and Evolution 30:1745–1750.

Guéguen L, Gaillard S, Boussau B, Gouy M, Groussin M, Rochette NC, Bigot T, Fournier D, Pouyet F, Cahais V. 2013b. Bio++: efficient extensible libraries and tools for computational molecular evolution. Molecular Biology and Evolution 30:1745–1750.

Halligan DL, Oliver F, Eyre-Walker A, Harr B, Keightley PD. 2010. Evidence for pervasive adaptive protein evolution in wild mice. PLoS Genetics 6:e1000825.

Huber CD, Kim BY, Marsden CD, Lohmueller KE. 2017. Determining the factors driving selective effects of new nonsynonymous mutations. Proceedings of the National Academy of Sciences 114:4465–4470.

Hvilsom C, Qian Y, Bataillon T, Li Y, Mailund T, Sallé B, Carlsen F, Li R, Zheng H, Jiang T. 2012. Extensive X-linked adaptive evolution in central chimpanzees. Proceedings of the National Academy of Sciences 109:2054–2059

Keightley PD, Eyre-Walker A. 2007. Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. Genetics 177:2251–2261.

Kimura M. 1983. The neutral theory of molecular evolution. Cambridge University Press

Lanfear R, Kokko H, Eyre-Walker A. 2014. Population size and the rate of evolution. Trends in Ecology & Evolution 29:33–41.

Liti G, Carter DM, Moses AM, Warringer J, Parts L, James SA, Davey RP, Roberts IN, Burt A, Koufopanou V. 2009. Population genomics of domestic and wild yeasts. Nature 458:337.

Loire E, Chiari Y, Bernard A, Cahais V, Romiguier J, Nabholz B, Lourenço JM, Galtier N. 2013. Population genomics of the endangered giant Galápagos tortoise. Genome Biology 14:R136.

MacLean RC, Buckling A. 2009. The distribution of fitness effects of beneficial mutations in *Pseudomonas aeruginosa*. PLoS genetics 5:e1000406.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in *Drosophila*. Nature 351:652–654.

Miyata T, Miyazawa S, Yasunaga T. 1979. Two types of amino acid substitutions in protein evolution. Journal of Molecular Evolution 12:219–236.

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. Journal of Molecular Evolution 60:315–326.

Smith JM, Smith NH. 1996. Synonymous nucleotide divergence: what is “saturation”? Genetics 142:1033–1036.

Smith NGC. 2003. Are radical and conservative substitution rates useful statistics in molecular evolution? Journal of Molecular Evolution 57:467–478.

Sneath PHA. 1966. Relations between chemical structure and biological activity in peptides. Journal of theoretical biology 12:157–195.

Strasburg JL, Kane NC, Raduski AR, Bonin A, Michelmore R, Rieseberg LH. 2010. Effective population size is positively correlated with levels of adaptive divergence among annual sunflowers. Molecular Biology and Evolution 28:1569–1580.

Tang H, Wyckoff GJ, Lu J, Wu C-I. 2004. A universal evolutionary index for amino acid changes. Molecular Biology and Evolution 21:1548–1556.

Tataru P, Mollion M, Glémin S, Bataillon T. 2017. Inference of distribution of fitness effects and proportion of adaptive substitutions from polymorphism data. Genetics 207:1103–1119.

Tsagkogeorga G, Cahais V, Galtier N. 2012. The population genomics of a fast evolver: high levels of diversity, functional constraint, and molecular adaptation in the tunicate *Ciona intestinalis*. Genome Biology and Evolution 4:852–861.

Waterson RH, Lander ES, Wilson RK. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437:69.

Welch JJ. 2006. Estimating the genomewide rate of adaptive protein evolution in *Drosophila*. Genetics 173:821–837.

Yampolsky LY, Stoltzfus A. 2005. The exchangeability of amino acids in proteins. Genetics 170:1459–1472.

Zhang J. 2000. Rates of conservative and radical nonsynonymous nucleotide substitutions in mammalian nuclear genes. Journal of Molecular Evolution 50:56–68.

DISCUSSION

The role of adaptive *vs.* non-adaptive forces that underly genome evolution lies at the heart of many questions in evolutionary biology. This is why detecting genetic signature of natural selection, typically by comparing the pattern of polymorphism and divergence across different species is of utmost importance. The McDonald & Kreitman test (McDonald and Kreitman 1991) and its derivatives exploit this comparison between polymorphism and divergence to provide an estimate of the adaptive substitution rate of a species. However, this comparison is only relevant if all the evolutionary forces influencing the genome have remained constant during the whole period under consideration, and impact divergence data and polymorphism data in the same way. This is why in this work, we explored how long-term demographic fluctuations and GC-biased gene conversion can lead to biases in the estimation of the adaptive substitution rate using the DFE-α method, one of the latest derivative of the McDonald & Kreitman test. We then accounted for those biases and estimated the adaptive amino acid substitution rate in a wide range of metazoan species datasets in order to uncover the determinants of the adaptive substitution rate. In particular, we investigated the relationship between the effective population size and the adaptive substitution rate, and the nature of amino acid changes particularly prone to contribute to adaptation. In this last part of the manuscript, we discuss the different results we obtained and their consequences as well as limitations and ideas to improve this work and go further .

I. Exploring the biases in the DFE-α method

a. Long-term fluctuations in population size

Two methods to take into account recent population size changes that would affect polymorphism data and consequently distort the synonymous and non-synonymous SFS have been proposed in the last years (Keightley and Eyre-Walker 2007; Eyre-Walker and Keightley 2009). First, a demographic change can be modeled explicitly by introducing three additional parameters to be estimated alongside the DFE parameters in the likelihood framework of the DFE-α method: N_1, N_2 and t_2, where N_1 is the equilibrium size of the population, N_2 is the new population size after a step change and t_2 is the number of generations the population stayed at size N_2 before the sampling of individuals to generate polymorphism data (Keightley and Eyre-Walker 2007). Second, instead of explicit parameters, nuisance parameters usually called r_i's can be also estimated alongside the DFE parameters. Those parameters have already been introduced several times in this manuscript as we chose them to correct the observed SFS in all our projects. We chose them because despite their lack of biological meaning, they can capture a wide range of effects that could distort the SFS, as long as these effects impact the non-synonymous and the synonymous SFS in the same manner

(Eyre-Walker and Keightley 2009). The two methods have proven to be very efficient to correct for recent variation in population size, and r_i's are also considered to correct for orientation errors and the effect of linkage (Tataru et al. 2017). However, this does not correct for long-term variations in population size, which yet are likely to be widespread, due, for instance, to the alternation of glacial and interglacial periods during the Quaternary (Hewitt 1999). In this thesis, we first showed via simulations that plausible demographic scenarios involving long-term fluctuations in population size can lead to overestimation of the adaptive substitution rate, even if the population size at recent time (i.e. when polymorphism was built) has remained constant (**Chapter 1**). Additionally, we present in **Chapter 3** indirect empirical evidence that the average regime of selection/drift that a species experienced during its divergence with its sister species is unlikely to be well depicted by polymorphism data, even when we correct the SFS using r_i's parameters. Those two pieces of argument are then congruent, and emphasize the need to account for long-term demographic fluctuations in the DFE-α method. This however seems very difficult to achieve, as current methods to infer population size variation rely on the use of polymorphism data and could not estimate long-term fluctuations.

In **Chapter 3**, we proposed two strategies to circumvent this issue. They both make strong assumptions, the question being if these assumptions are more questionable or not than those of the classical DFE-α approach. They also both require a dataset with several closely related species (the more the better, here we used between four and six species), which is more expensive and time consuming than the classical approach. This is because they both rely on an estimation of the adaptive substitution rate representative of the whole group of closely related species sampled. Indeed, the first strategy consists in pooling the synonymous and non-synonymous SFS across closely related species, and the second strategy consists in averaging the non-adaptive substitution rate estimates of each species, and subtracting this average to the dN/dS ratio of the whole tree of the group to obtain one estimate of the adaptive substitution rate for the group. The two strategies require the hypothesis that the shape of the DFE and the mutation rate have remained constant as species were diverging, and that only N_e varies through time. The hypothesis that the shape of the DFE have remained constant is supported by the fact that the different shapes estimated for each closely related species are similar: the fit of the model when allowing each species to have an individual shape is not significantly better than the fit of the model when fixing one common shape in all species (this shape being chosen as the one obtained from pooled SFS within taxonomic groups) in all species but five of the dataset used in **Chapter 3** (likelihood ratio tests).

Another possibility to circumvent the bias caused by long-term fluctuations of the regime of selection/drift would be to abandon the idea of relying on a comparison between divergence and polymorphism. Tataru et al. (2017) developed a new implementation of the DFE-α method that allows the estimation of the adaptive substitution rate using polymorphism data alone, thus hypothesizing that the unfolded non-synonymous SFS contains usable information on segregating beneficial mutations. They indeed show that not accounting for the effect of beneficial mutations in the SFS could lead to biases in the estimation process, and they identify this as being due to the method inferring too much slightly deleterious mutations that there really are when not accounting for the presence of segregating beneficial mutations. However, this method may underestimate the amount of strongly advantageous mutations that are unlikely to be present in polymorphism data. Moreover, this require high quality/quantity of data and may be only suitable for highly polymorphic species. Even if we did not included it in the published article, we actually tested this model on our simulated data generated as part as the project presented in **Chapter 1** and we obtained with this model the highest variance in estimates of α and ω_a among replicates of the same demographic scenario. Finally, estimates based on this method may require a different interpretation than estimates obtained via the classical approach, as it is representative of the expected amount of adaptive substitution under the recent regime of selection/drift, which is different from the amount of adaptive substitution fixed during the divergence between two sister species.

Simulations may help assessing the relevance and performance of the two strategies we used in **Chapter 3**. In particular, we would like to measure the extent of population size variation that are required to reached the observed variation in π_s between closely related species. Simulation is a very powerful tool to assess the influence of one source of bias at a time. For instance, we could decouple the effect of long-term fluctuations in population size from other effects we proved to be also the cause of overestimation of the adaptive substitution rate, such as gBGC.

b. GC-biased gene conversion and Hill-Robertson interference

In **Chapter 2**, we analyzed the influence of gBGC in the dN/dS ratio, π_n/π_s ratio as well as the results of the DFE-α method. We reveal a combined influence of gBGC and Hill-Robertson interference on the dN/dS ratio that leads to a decrease of this ratio with increasing recombination rate irrespective of gene function in both primates and birds. We also report a mutagenic effect of recombination in the two groups, which is independent of gBGC. Last but not least, our analysis indicates that gBGC leads to an overestimation of the adaptive substitution rate of AT→GC mutations, which entails an increase of the overall adaptive substitution rate compared to the

adaptive substitution rate of GC-conservative mutations in most of the species we investigated. This result may be a confirmation of the widespread hypothesis that gBGC mimics the effects of positive selection: the overestimation would then be a direct effect of AT→ GC mutations pushed by gBGC being mistakenly considered as beneficial mutations (Galtier and Duret 2007; Ratnakumar et al. 2010). The reason why the effects of gBGC would lead to such a bias is not obvious: indeed, gBGC is expected to affect non-synonymous as well as synonymous positions, so if the neutral regime is evaluated using synonymous positions, why an extra amount of positive selection would be inferred is not intuitive. However, there are actually two reasons for that. First, the extent of the impact of gBGC on synonymous and non-synonymous sites depends on the distance between current GC-content to its equilibrium (denoted GC*) at the two types of sites. Under a simple model of DFE of non-synonymous mutations, Bolivar et al. (2016) indeed found that due to interaction with selection, GC* at non-synonymous sites is lower than GC* at synonymous sites, and that depending on the current GC-content and how far it is from GC* at the two types of sites, the dN/dS ratio can be influenced by current GC3 without invoking Hill-Robertson interference (Bolívar et al. 2016). This indicates that gBGC impacts differently synonymous and non-synonymous sites. As such the non-synonymous and synonymous SFS are not distorted similarly by gBGC, and controlling for neutral effects using synonymous sites does not accurately correct the non-synonymous SFS, which could then lead to false inference of positive selection. Second, Galtier et al. (2009) showed, also via a model, that gBGC promotes the fixation of weakly deleterious AT→GC non-synonymous mutations. Indeed, gBGC acts like if the DFE of AT→GC non-synonymous mutations was shifted towards higher N_es. A number of weakly deleterious mutations are then expected to behave as if adaptive, and depending on the shape of the negative DFE, this can lead to a decrease of the proportion of slightly deleterious mutations in the "shifted" AT→GC non-synonymous DFE. Methods will then infer lower non-adaptive substitution rates, and thus higher adaptive substitution rates of AT→GC mutations (Galtier et al. 2009). This seems to be what we observe in our study, as we found that non-adaptive substitution rate of AT→GC mutations are lower than non-adaptive substitution rate of GC-conservative mutations for all but one species.

It is yet difficult to exclude that our results are not due to methodological issues. The observed overestimation of the adaptive substitution rate could originate from several factors, such as orientation errors, or indirect effects of gBGC that would affect the estimation process of the DFE-α method because of its impact on the summary statistics we use, or because we use r_i's parameters that could capture some of the effects of gBGC on the SFS for instance. This question is really difficult to answer with the tools that we currently have. One way to improve our analysis on this matter would be to directly include the estimation of the parameter $B= 4N_eb$, where b is the

gBGC coefficient, in the estimation process of the adaptive substitution rate. This require to modify the structure of the data, including three synonymous and three non-synonymous SFS, one for AT→ GC alleles, one for GC → AT alleles and one for GC-conservative alleles. I have started to develop this with the help of Sylvain Glémin who developed a method allowing the estimation of B from synonymous SFS (Glémin et al. 2015; Clément et al. 2017).

It is also difficult to extend our results to other species. If the same pattern than ours have been detected in great tits (Corcoran et al. 2017), opposite results have recently been obtained in flycatchers (Bolívar et al. 2018) for reasons that are hard to identify. As for other taxa than primates and birds, no results are available to our knowledge. The analysis of the SFS of 30 species of eight metazoan phyla revealed that a majority are affected by gBGC, with some exceptions (Ostrea *edulis* for instance) (Galtier et al. 2018). In view of the lack of information in many animal species, controlling for gBGC by using only GC-conservative mutations seems to be the most relevant strategy.

GC-conservative mutations are expected to only reflect the effects of recombination different from gBGC, i.e. linked selection and a potential mutagenic effect. We showed that the synonymous substitution rate increases with GC3 and the recombination rate, leading us to the conclusion that recombination is mutagenic in both taxonomic groups, in agreement with some studies focusing on human (Pratto et al. 2014; Arbeithuber et al. 2015; Smith et al. 2018). In spite of this, the dN/dS ratio of GC-conservative mutations appears to be decreasing with GC3 and recombination rate. We interpreted this as the consequence of reduced linkage with increasing recombination rate, which enhances the efficiency of selection against slightly deleterious mutations. Alternatively, this could be due to an effect of linkage on neutral divergence. The existence of such an effect has been debated, as theoretical work predicts that in absence of ancestral polymorphism, there are no such effect (Birky and Walsh 1988; Cutler 1998), whereas empirical results indicate that linkage may influence neutral divergence (Phung et al. 2016). Indeed, the results of Phung et al. (2016) indicate that neutral divergence between closely related species (e.g. human-primate) is positively correlated with human recombination rate, even after controlling for gBGC. If this originates from the contribution of ancestral polymorphism to divergence, it should concern terminal branches, not internal ones. In primates and birds, we estimated the substitution rate of synonymous GC-conservative substitutions on the internal branches of the subtrees of the two taxonomic groups, and found again a strong positive correlation between this statistic and GC3 (results not shown). This confirms that at least part of the increase we observed in

the substitution rate of synonymous GC-conservative substitutions is due to recombination being mutagenic, and also that recombination enhances the efficiency of purifying selection.
However, we failed to detect any significant increase of the efficiency of positive selection with increasing recombination rate, whereas it has been empirically documented, for instance in the fungal pathogen *Zymoseptoria tritici* (Grandaubert et al. 2018). This may be due to the reduced number of GC-conservative SNPs we obtained when we split our dataset in bins of genes of increasing GC3. As suggested by reviewers of our article, we split our dataset in bins of equal number of SNPs instead of equal number of genes in **Chapter 2**, but still we did not detect any significant increase of the efficiency of positive selection with increasing recombination rate, even though the correlation coefficients between GC3 and the adaptive substitution rate are consistently positive in all species (results not shown).

c. Are there some remaining biases and limitations?

We here discuss additional sources of bias, some specific of our work ans some specific of the DFE-α method. In regards to our work, there may be two limitations.

First, our analysis may be sensible to orientation errors, in particular when we analyze separately AT→ GC and GC→ AT mutations. Throughout the thesis, we orientated mutations by estimating an ancestral sequence at each node of the tree of the focal group, which is slightly better than using just one outgroup species. Moreover, the nuisance parameters we used to correct neutral forces affecting both the synonymous and non-syonymous SFS (r_i's) may also capture some of the distortions of the spectra that would be due to orientation errors. Finally, when evaluating the impact of gBGC on the DFE-α method, we reproduced our classical analysis by removing CpG sites from our alignments. CpG hypermutability is likely to make unfolded SFS particularly sensitive to polarization errors (Glémin et al. 2015). However, removing CpG sites did not changed our results or the shape of the SFS. Despite those efforts, polarizations errors may be one of the source of the high variance we detect between species or between methods or models. To our surprise, it did not explained the unusual shapes of the SFS of some Galloanserae commented in **Chapter 2** (see **Annexe 2**), for which we did not find any satisfying explanation.

Second, throughout the thesis we did not account for the fact that some substitutions are in fact polymorphisms. This can bias the estimates of the dN/dS ratio, and thus bias the estimates of the adaptive substitution rate, in particular if the nucleotide divergence between the species is low relative to within species variation (Keightley and Eyre-Walker 2012; Mugal et al. 2013). We could

consider correcting our estimations of the divergence counts to check that our results are not influenced by such a bias. Additionally, estimates of the dN/dS ratio may be less accurately estimated in short than in long branches, because i) a deleterious mutation segregating at the time of the species divergence is more likely to be segregating in one lineage and lost in the other than a neutral mutation, and ii) beneficial mutations fix more rapidly than neutral mutations (Bierne and Eyre-Walker 2004), which will lead to an inflated dN/dS ratio. However, in our main analysis (**Chapter 3**), we used the total dN/dS of the tree of each groups, which include a relatively long branch for at least one species in all groups as well as internal branches, so this is not likely to be a major issue in our study.

One assumption of the general DFE-α method is the independence between sites. This is however hardly the case unless recombination is uniformly high. Messer and Petrov (2013) simulated the evolution of genes located on a single chromosome and then applied methods that correct the synonymous and non-synonymous SFS for demographic effects to measure its adaptive substitution rate, like we did in **Chapter 1**. They showed that the estimations of the adaptive substitution rate remain pretty accurate even without independence between sites, even if the mean strength of purifying selection and inferred demography are erroneous (Messer and Petrov 2013). However, the use of nuisance parameters instead of explicit demographic parameters to correct the SFS relative to the neutral expectation may be better suited to capture the effects of linkage, thus alleviating its influence on the results of the test, may it be the adaptive or non-adaptive substitution rate or the inferred parameters of the DFE (Tataru et al. 2017).

An other difficulty of the DFE-α method is that it requires to make an assumption regarding the type of distribution of the DFE. As mentioned in introduction (part **I. 3. b.**), several distributions are classically employed: normal, exponential or gamma, or beta (Piganeau and Eyre-Walker 2003; Loewe and Charlesworth 2006; Galtier 2016). In this work we only tested a gamma distribution to model the negative DFE and an exponential distribution to model the adaptive DFE. However we have little information on the real distribution of the DFE of any of the species, and we even do not know if the real DFE of all species is well depicted by the same type of distribution. We may consider several distributions for each species or each group, and chose each time the one that best explains the data.

II. Is there a link between N_e and the adaptive substitution rate ?

Results presented in **Chapter 3** of this thesis indicate that when accounting for two important biases of the DFE-α method, there is a negative relationship between ω_a and life history traits indicative of long-term N_e (propagule size, adult size, longevity and fecundity), as well as π_s (even if we did not control for the non-independence between ω_a and π_s). This relationship is in contradiction with the hypothesis that the DFE is independent from N_e, and that adaptation is limited by the supply of new mutations, two hypothesis underpinning the expectation of a positive relationship between ω_a and N_e (Kimura 1983; Gossmann et al. 2012). It is however congruent with theoretical predictions of Fisher's geometrical model (FGM), which suggest that species with low long-term N_e may experience a higher proportion of beneficial *de novo* mutations (Huber et al. 2017). This is because species evolving under small population sizes are expected to be further away from their fitness optimum, on average, than species evolving under larger population sizes. Additionally, other theoretical results obtained under FGM indicate that species with a high complexity and experiencing a high rate of environmental change will experience a higher adaptive substitution rate (Lourenço et al. 2013), and there are indications that high complexity and high rate of environmental change may be associated with low population size. In one hand, Fernández and Lynch showed that the accumulation of mildly deleterious mutations in populations of small size which are far from their fitness optimum induces secondary selection for protein–protein interactions that stabilize key gene functions, leading to higher complexity, under the hypothesis that organismal complexity may be accurately measured by the number of protein–protein interactions (Fernández and Lynch 2011). In the other hand, the rate of environmental change is measured per generation, and generation time is negatively correlated to population size in amniotes (Chao and Carr 1993). Because of this, the rate of environmental change may be on average higher in species of small population size.

Linking the results and predictions of FGM to our empirical results implies to make the hypothesis that the theoretical conclusions made under the FGM framework are directly transposable to protein evolution. Yet, some of the assumptions of FGM, such as the one-peak shape of the fitness landscape and the universal pleiotropy assumption may not be clearly appropriate to model evolving proteins (Lourenço et al. 2013). Nevertheless, some studies developed models for the phenotypic basis of protein evolution that show strong parallels with FGM (DePristo et al. 2005; Weinreich and Knies 2013), implying that essential elements of protein evolution are captured by FGM. Besides, the results of Huber et al. (2017) indicate a link between long-term population size and the proportion of beneficial mutations, but not with the adaptive substitution rate. The rate of

substitution depends on the rate of mutation as well as the probability of fixation. Thus our results imply that despite a higher probability of fixation in large N_e species due to decreased drift, the larger proportion of adaptive mutations in small-N_e species may be enough to lead to a higher adaptive substitution rate. This is not incompatible with the prediction that adaptive walks are more efficient, in terms of rate of fitness increase, in large than in small populations (Orr 2000). This alongside with our results would imply that small-N_e species adapt by a lot of small steps, whereas large-N_e species adapt with a few mutations of stronger fitness effect. The preliminary project presented in **Chapter 4** may be particularly suitable to investigate this question and help us conclude as for the causes of the negative relationship between the adaptive substitution rate and the population size. Indeed, it consists in comparing the adaptive substitution rate of radical *vs*. conservative amino acids changes (or between amino acids changes classified according to the physiochemical similarities between the two amino acid of the pair) which can be used as a proxy of small (conservative) or large (radical) fitness effects. Thanks to the dataset presented in **Chapter 3**, we have the potential to compare the rate of adaptive substitution of radical *vs*. conservative amino acids changes between species with different population sizes.

III. Intra-genome comparisons: the best usage of the DFE-α method?

In this thesis, we only used the DFE-α method on the whole set of available coding sequences to estimate the species specific, or the taxon specific, adaptive substitution rate in order to uncover the determinants of the differences between species or taxa. Nevertheless, the DFE-α method can also be used to compare genes within species. This has been done for instance to compare the efficiency of positive and purifying selection on autosomes *vs*. sex chromosomes. Indeed, on one hand, due to hemizygozity of sex chromosomes in heterogametic species, selection for or against recessive mutations is supposed to be more efficient. On the other hand, the effective population size of sex chromosome is expected to be smaller than the effective population size of autosomes, which counteracts the effects of hemizygozity. We explored this question in two Satyrinae butterflies, as part of my Master project. We found that the two species do not experience a significantly higher efficiency of positive selection in sex chromosomes, but we revealed an increased efficiency of purifying selection against recessive deleterious mutations in Z-linked genes using the DFE-α method (Rousselle et al. 2016, **Annexe 4**). This strategy has also been used on *Heliconius* butterflies as part of a collaboration with Ana Pinharanda (Butterflies genetic groups, university of Cambridge) (Pinharanda et al. 2018), and in moths (Sackton et al. 2014).

Comparing the results of the DFE-α method between genes within species has been used to identify genes, and by extension, functions, evolving under positive selection. For instance, Enard et al. (2016) compared the adaptive substitution rate of proteins identified as virus-interacting proteins (VIPs) relative to other proteins (using the simple McDonald & Kreitman approach as presented in Smith and Eyre-Walker (2002) and the "asymptotic MK test" presented in (Messer and Petrov (2013)). They found that the proportion of adaptive substitutions is significantly higher in VIPs *vs*. non-VIPs.

We argue that comparisons of the adaptive substitution rate between genes within a genome may be the safest usage of the DFE-α method, as long as the absolute values of the proportion of adaptive substitutions or the adaptive substitution rate are not commented, but only relatives differences between bins of genes under analysis. Indeed, it allows avoiding some biases, in particular the one we identified as the most problematic because the hardest to correct, i.e. the difference in regime of selection/drift represented by polymorphism data *vs*. divergence data, as this bias should be the same in all bins of genes. However, correcting for the presence of slightly deleterious mutations as well as gBGC is still required in this type of comparisons. Besides, genes within a genome may not all have the same effective population size, which may lead to biases in the method. Indeed, genes with a high effective population size are expected to have a high synonymous diversity, but low non-synonymous diversity, which may lead to underestimations of the non-adaptive substitution rate, and thus an overestimation of the adaptive substitution rate (Smith and Eyre-Walker 2002; Welch 2006). However, Eyre-Walker & Keigthley (2009) compared the estimates of the proportion of adaptive substitution rate between an approach where data are combined across genes and an approach allowing each genes to have their own population size (Eyre-Walker et al. 2006) in *Drosophila*, and found that the results are almost identical (Eyre-Walker and Keightley 2009). This suggests that combining genes with different effective population size will not lead to serious biases. Consequently, using the DFE-α method on GC-conservative mutations while correcting the synonymous and non-synonymous SFS for effects that would distort them compare to the neutral expectation (via the nuisance parameters r_i's, for instance) seems to be an accurate way to compare the level of positive selection between different groups of genes within a genome.

IV. Conclusion

The main take-home message of this thesis is a message of caution towards the absolute values yielded by the DFE-α method when the biases produced by gBGC and discrepancies

between the recent and long-term regime of selection/drift are not properly accounted for. If the bias of gBGC can be easily accounted for by using only GC-conservative mutations and substitutions, it is not the case for discrepancies between the recent and long-term regime of selection/drift. We proposed two strategies towards this goal, but they require some assumptions as well as a relatively heavy dataset, and they do not allow to differentiate the adaptive substitution rate between closely related species. In this respect, we argue that the DFE-α method should not be estimated on single species without new developments that would accurately correct this bias. If this can be achieved, then comparative genomics will have the potential to help us understand the determinants of adaptive evolution, and in particular assess if the negative relationship between N_e and the adaptive substitution rate we uncovered in this thesis is confirmed, and if yes, what are the mechanisms that underly such a relationship. To help answer this last question, we think that exploring the sizes of fitness effects of mutations in small *vs.* large-N_e species is a very promising track.

V. References

Arbeithuber B, Betancourt AJ, Ebner T, Tiemann-Boege I. 2015. Crossovers are associated with mutation and biased gene conversion at recombination hotspots. Proceedings of the National Academy of Sciences of the United States of America 112:2109–2114.

Bierne N, Eyre-Walker A. 2004. The genomic rate of adaptive amino acid substitution in *Drosophila*. Molecular Biology and Evolution 21:1350–1360.

Birky CW, Walsh JB. 1988. Effects of linkage on rates of molecular evolution. Proceedings of the National Academy of Sciences 85:6414–6418.

Bolívar P, Mugal CF, Nater A, Ellegren H. 2016. Recombination rate variation modulates gene sequence evolution mainly via GC-biased gene conversion, not Hill–Robertson interference, in an avian system. Molecular Biology and Evolution 33:216–227.

Bolívar P, Mugal CF, Rossi M, Nater A, Wang M, Dutoit L, Ellegren H. Biased inference of selection due to GC-biased gene conversion and the rate of protein evolution in flycatchers when accounting for it. Molecular Biology and Evolution.

Chao L, Carr DE. 1993. The molecular clock and the relationship between population size and generation time. Evolution 47:688–690.

Clément Y, Sarah G, Holtz Y, Homa F, Pointet S, Contreras S, Nabholz B, Sabot F, Sauné L, Ardisson M. 2017. Evolutionary forces affecting synonymous variations in plant genomes. PLoS genetics 13:e1006799.

Corcoran P, Gossmann TI, Barton HJ, Slate J, Zeng K. 2017. Determinants of the efficacy of natural selection on coding and noncoding variability in two passerine species. Genome Biology and Evolution 9:2987–3007.

Cutler DJ. 1998. Clustered mutations have no effect on the overdispersed molecular clock: a response to Huai and Woodruff. Genetics 149:463–464.

DePristo MA, Weinreich DM, Hartl DL. 2005. Missense meanderings in sequence space: a biophysical view of protein evolution. Nature Reviews Genetics. 6:678–687.

Eyre-Walker A, Keightley PD. 2009. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. Molecular Biology and Evolution 26:2097–2108.

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

Fernández A, Lynch M. 2011. Non-adaptive origins of interactome complexity. Nature 474:502.

Galtier N. 2016. Adaptive protein evolution in animals and the effective population size hypothesis. PLOS Genetics 12:e1005774.

Galtier N, Duret L. 2007. Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. Trends in Genetics 23:273–277.

Galtier N, Duret L, Glémin S, Ranwez V. 2009. GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. Trends in Genetics 25:1–5.

Galtier N, Roux C, Rousselle M, Romiguier J, Figuet E, Glémin S, Bierne N, Duret L. 2018. Codon usage bias in animals: disentangling the effects of natural selection, effective population size, and GC-Biased gene conversion. Molecular Biology and Evolution 35:1092–1103.

Glémin S, Arndt PF, Messer PW, Petrov D, Galtier N, Duret L. 2015. Quantification of GC-biased gene conversion in the human genome. Genome Research 25:1215–1228.

Gossmann TI, Keightley PD, Eyre-Walker A. 2012. The effect of variation in the effective population size on the rate of adaptive molecular evolution in eukaryotes. Genome Biology and Evolution 4:658–667.

Grandaubert J, Dutheil JY, Stukenbrock EH. 2018. The genomic determinants of adaptive evolution in a fungal pathogen. bioRxiv:176727.

Hewitt GM. 1999. Post-glacial re-colonization of European biota. Biological Journal of the Linnean Society 68:87–112.

Huber CD, Kim BY, Marsden CD, Lohmueller KE. 2017. Determining the factors driving selective effects of new nonsynonymous mutations. Proceedings of the National Academy of Sciences 114:4465–4470.

Keightley PD, Eyre-Walker A. 2007. Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. Genetics 177:2251–2261.

Keightley PD, Eyre-Walker A. 2012. Estimating the rate of adaptive molecular evolution when the evolutionary divergence between species is small. Journal of molecular evolution 74:61–68.

Kimura M. 1983. The neutral theory of molecular evolution. Cambridge University Press

Loewe L, Charlesworth B. 2006. Inferring the distribution of mutational effects on fitness in *Drosophila*. Biology Letters 2:426–430.

Lourenço JM, Glémin S, Galtier N. 2013. The rate of molecular adaptation in a changing environment. Molecular Biology and Evolution 30:1292–1301.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in *Drosophila*. Nature 351:652–654.

Messer PW, Petrov DA. 2013. Frequent adaptation and the McDonald–Kreitman test. PNAS 110:8615–8620.

Mugal CF, Wolf JB, Kaj I. 2013. Why time matters: codon evolution and the temporal dynamics of dN /dS. Molecular Biology and Evolution 31:212-231.

Orr HA. 2000. Adaptation and the cost of complexity. Evolution 54:13–20.

Phung TN, Huber CD, Lohmueller KE. 2016. Determining the effect of natural selection on linked neutral divergence across species. PLOS Genetics 12:e1006199.

Piganeau G, Eyre-Walker A. 2003. Estimating the distribution of fitness effects from DNA sequence data: Implications for the molecular clock. PNAS 100:10335–10340.

Pinharanda A, Rousselle M, Martin SH, Hanly JJ, Davey JW, Kumar S, Galtier N, Jiggins CD. 2018. Sexually dimorphic gene expression and transcriptome evolution provides mixed evidence for a fast-Z effect in *Heliconius*. bioRxiv:380030.

Pratto F, Brick K, Khil P, Smagulova F, Petukhova GV, Camerini-Otero RD. 2014. Recombination initiation maps of individual human genomes. Science 346:1256442.

Ratnakumar A, Mousset S, Glémin S, Berglund J, Galtier N, Duret L, Webster MT. 2010. Detecting positive selection within genomes: the problem of biased gene conversion. Philosophical Transactions of the Royal Society B: Biological Sciences 365:2571–2580.

Rousselle M, Faivre N, Ballenghien M, Galtier N, Nabholz B. 2016. Hemizygosity enhances purifying selection: lack of fast-Z evolution in two Satyrine butterflies. Genome Biology and Evolution 8:3108–3119.

Sackton TB, Corbett-Detig RB, Nagaraju J, Vaishna L, Arunkumar KP, Hartl DL. 2014. Positive selection drives faster-Z evolution in silkmoths. Evolution 68:2331–2342.

Smith NGC, Eyre-Walker A. 2002. Adaptive protein evolution in *Drosophila*. Nature 415:1022–1024.

Smith TCA, Arndt PF, Eyre-Walker A. 2018. Large scale variation in the rate of germ-line *de novo* mutation, base composition, divergence and diversity in humans. PLOS Genetics 14:e1007254.

Tataru P, Mollion M, Glémin S, Bataillon T. 2017. Inference of distribution of fitness effects and proportion of adaptive substitutions from polymorphism data. Genetics 207:1103–1119.

Weinreich DM, Knies JL. 2013. Fisher's geometric model of adaptation meets the functional synthesis: data on pairwise epistasis for fitness yields insights into the shape and size of phenotype space. Evolution 67:2957–2972.

Welch JJ. 2006. Estimating the genomewide rate of adaptive protein evolution in *Drosophila*. Genetics 173:821–837.

Annexe 1
Supplementary Material of Chapter 1

Electronic Supplementary Methods : Adaptive rate estimation in the extended McDonald-Kreitman framework

The non-synonymous substitution rate, dN, can be seen as the sum of an adaptive and a non-adaptive component:

$dN = dN_a + dN_{na}$

Dividing by the synonymous substitution rate, dS, we obtain:

$dN/dS = dN_a/dS + dN_{na}/dS$

Let us define ω_a, ω_{na} and α as:

$\omega_a = dN_a/dS$
$\omega_{na} = dN_{na}/dS$
$\alpha = dN_a/dN$

Here we provide an overview of various options in two recently developed programs aiming at estimating these quantities in the maximum likelihood framework based on polymorphism and divergence data. The main idea is that polymorphism informs on ω_{na}, and divergence on the sum of ω_a and ω_{na}. See Eyre-Walker and Keightley (2009), Galtier (2016) and Tataru et al. (2017) for a detailed description of the underlying theory.

1. Data

Polymorphism data: synonymous and non-synonymous site frequency spectra (SFS_S and SFS_N), i.e., observed distribution of derived allele frequencies across synonymous and non-synonymous segregating sites, respectively.

Divergence data: observed number of fixed synonymous and non-synonymous differences between species, divided by the number of synonymous and nonsynonymous sites, respectively (dS and dN).

2. Model

Model parameters:
- Θ: population mutation rate

Overestimation of the adaptive substitution rate in fluctuating populations

Marjolaine Rousselle*[1], Maeva Mollion[2], Benoit Nabholz[1], Tomas Bataillon[2] and Nicolas Galtier[1]

[1]UMR-5554 Institut des Sciences de l'Evolution, CNRS, Université de Montpellier, IRD, EPHE, Place E. Bataillon, Montpellier, France

[2]BiRC-Bioinformatics Research Center, Department of Genetics and Ecology, University of Aarhus, Aarhus, Denmar

- Φ: distribution of the scaled fitness effect *S* of mutations (DFE)
- T: neutral divergence
- **R**: nuisance parameters

3. Likelihood calculation

See Eyre-Walker and Keightley (2009), Galtier (2016), Tataru et al. (2017). The expected SFS_S, SFS_N, dS and dN are obtained from the population genetic theory. Observed counts are assumed to follow Poisson distributions of means equal to the expectations.

4. *Procedures for estimating wna, wa and alpha*

a. Gamma model

Only neutral or deleterious effects are explicitly modelled, as proposed by Eyre-Walker and Keightley (2009). This corresponds to model "Gamma" in Galtier 2016 and the current manuscript, and to p_b=0 in Tataru et al. 2017. The assumption is that beneficial mutations are of sufficiently large effect to negligibly contribute to polymorphism. In this case the estimation procedure is as follows:
- fit Φ, Θ, and **R** to SFS_S and SFS_N
- calculate the expected dN/dS from parameter estimates; this corresponds to the estimated ω_{na}
- estimate ω_a by substracting estimated ω_{na} from observed dN/dS
- estimate α as the ratio of estimated ω_a over observed dN/dS

b. GammaExpo model

Here the effects of beneficial mutations are explicitly modelled (Galtier 2016, Tataru et al. 2017). Two options are considered. Option 1 is to estimate ω_{na}, ω_a and α the same way as in section 4a above. The other option is to:
- fit Φ, Θ, **R** and T to SFS_S, SFS_N, dS and dN
- calculate the expected ω_a and ω_{na} from parameter estimates; this implies to define adaptive substitutions; here we called adaptive the mutations of population scaled effect *S* above 5.
- calculate α as $\omega_a/(\omega_{na}+\omega_a)$

This option is called "+A" in Galtier (2016), α_{DFE} in Tataru et al. (2107) and GammaExpo* in the current manuscript.

Tataru et al. (2017) also consider the option of not using divergence data at all, *i.e.*, fit model to SFS only, then estimate ω_a, ω_{na} and α from estimated model parameters. This option was not considered in the current study.

Overestimation of the adaptive substitution rate in fluctuating populations

Marjolaine Rousselle*[1], Maeva Mollion[2], Benoit Nabholz[1], Tomas Bataillon[2] and Nicolas Galtier[1]

[1]UMR-5554 Institut des Sciences de l'Evolution, CNRS, Université de Montpellier, IRD, EPHE, Place E. Bataillon, Montpellier, France

[2]BiRC-Bioinformatics Research Center, Department of Genetics and Ecology, University of Aarhus, Aarhus, Denmar

	Scenario	Actual population size	Long-term population size
Temperate species	soft bottlenecks	100.000	12.190
	severe bottlenecks	100.000	1.240
Alpine species	soft bottlenecks	10.000	35.710
	severe bottlenecks	1.000	4.800

Table S1: Actual and long-term population size in the different scenarios tested in the study.

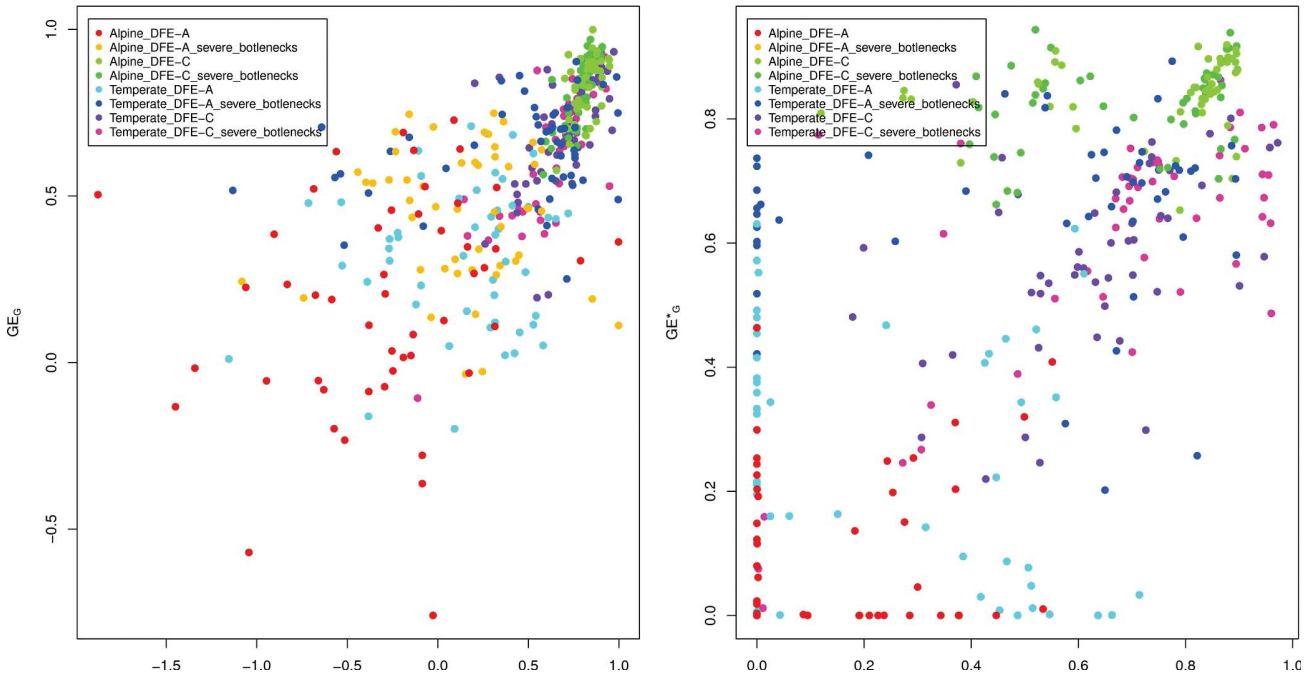


Figure S1: Correlation between the estimates of the two implementations of Galtier 2016 and Tataru et al. 2017 for the models GammaExpo and GammaExpo*.

Overestimation of the adaptive substitution rate in fluctuating populations

Marjolaine Rousselle*[1], Maeva Mollion[2], Benoit Nabholz[1], Tomas Bataillon[2] and Nicolas Galtier[1]

[1]UMR-5554 Institut des Sciences de l'Evolution, CNRS, Université de Montpellier, IRD, EPHE, Place E. Bataillon, Montpellier, France

[2]BiRC-Bioinformatics Research Center, Department of Genetics and Ecology, University of Aarhus, Aarhus, Denmar

```
#Stable Ne-DFE A:
// set up a
initialize()
{
// set the overall mutation rate
initializeMutationRate(2.2e-8);
initializeSex("A");
// two types of mutations: deleterious and neutral.
initializeMutationType("m1", 0.5, "f", 0.0);
initializeMutationType("m2", 0.5, "g", -2.5, 0.3);
// g1 genomic element type: coding:
initializeGenomicElementType("g1", c(m1, m2), c(0.25,0.75));
// chromosome consisting of 1500 genomic elements of type g1 , with gaps between
them at regular intervals
for (index in 1:1500)
initializeGenomicElement(g1, index*1000, index*1000 + 499);
// uniform recombination rate
initializeRecombinationRate(1e-7);
}
1 { sim.addSubpop("p1", 10000); }
10000 late() { p1.outputSample(20); }
100000 late() { p1.outputSample(20); }
100000 late() { sim.outputFixedMutations(); }

#Temperate species-DFE A:
// set up a
initialize()
{
// set the overall mutation rate
initializeMutationRate(2.2e-8);
initializeSex("A");
// two types of mutations: deleterious and neutral.
initializeMutationType("m1", 0.5, "f", 0.0);
initializeMutationType("m2", 0.5, "g", -2.5, 0.3);
// g1 genomic element type: coding:
initializeGenomicElementType("g1", c(m1, m2), c(0.25,0.75));
// chromosome consisting of 1500 genomic elements of type g1 , with gaps between
them at regular intervals
for (index in 1:1500)
initializeGenomicElement(g1, index*1000, index*1000 + 499);
// uniform recombination rate
initializeRecombinationRate(1e-7);
}
1 { sim.addSubpop("p1", 10000); }
// burnin
44000 late() { p1.outputSample(20); }
46000 { p1.setSubpopulationSize(1000); }
54000 { p1.setSubpopulationSize(10000); }
56000 { p1.setSubpopulationSize(1000); }
64000 { p1.setSubpopulationSize(10000); }
```

Overestimation of the adaptive substitution rate in fluctuating populations

Marjolaine Rousselle*[1], Maeva Mollion[2], Benoit Nabholz[1], Tomas Bataillon[2] and Nicolas Galtier[1]

[1]UMR-5554 Institut des Sciences de l'Evolution, CNRS, Université de Montpellier, IRD, EPHE, Place E. Bataillon, Montpellier, France

[2]BiRC-Bioinformatics Research Center, Department of Genetics and Ecology, University of Aarhus, Aarhus, Denmar

```
66000 { p1.setSubpopulationSize(1000); }
74000 { p1.setSubpopulationSize(10000); }
76000 { p1.setSubpopulationSize(1000); }
84000 { p1.setSubpopulationSize(10000); }
86000 { p1.setSubpopulationSize(1000); }
94000 { p1.setSubpopulationSize(10000); }
96000 { p1.setSubpopulationSize(1000); }
104000 { p1.setSubpopulationSize(10000); }
106000 { p1.setSubpopulationSize(1000); }
114000 { p1.setSubpopulationSize(10000); }
116000 { p1.setSubpopulationSize(1000); }
124000 { p1.setSubpopulationSize(10000); }
126000 { p1.setSubpopulationSize(1000); }
134000 { p1.setSubpopulationSize(10000); }
135100 late() { p1.outputSample(20); }
135100 late() { sim.outputFixedMutations(); }

#Alpine species-DFE A:
// set up a
initialize()
{
// set the overall mutation rate
initializeMutationRate(2.2e-8);
initializeSex("A");
// two types of mutations: deleterious and neutral.
initializeMutationType("m1", 0.5, "f", 0.0);
initializeMutationType("m2", 0.5, "g", -2.5, 0.3);
// g1 genomic element type: coding:
initializeGenomicElementType("g1", c(m1, m2), c(0.25,0.75));
// chromosome consisting of 1500 genomic elements of type g1 , with gaps between
them at regular intervals
for (index in 1:1500)
initializeGenomicElement(g1, index*1000, index*1000 + 499);
// uniform recombination rate
initializeRecombinationRate(1e-7);
}
1 { sim.addSubpop("p1", 10000); }
// burnin
44000 late() { p1.outputSample(20); }
44000 { p1.setSubpopulationSize(1000); }
46000 { p1.setSubpopulationSize(10000); }
54000 { p1.setSubpopulationSize(1000); }
56000 { p1.setSubpopulationSize(10000); }
64000 { p1.setSubpopulationSize(1000); }
66000 { p1.setSubpopulationSize(10000); }
74000 { p1.setSubpopulationSize(1000); }
76000 { p1.setSubpopulationSize(10000); }
84000 { p1.setSubpopulationSize(1000); }
86000 { p1.setSubpopulationSize(10000); }
94000 { p1.setSubpopulationSize(1000); }
```

Overestimation of the adaptive substitution rate in fluctuating populations

Marjolaine Rousselle*[1], Maeva Mollion[2], Benoit Nabholz[1], Tomas Bataillon[2] and Nicolas Galtier[1]

[1]UMR-5554 Institut des Sciences de l'Evolution, CNRS, Université de Montpellier, IRD, EPHE, Place E. Bataillon, Montpellier, France

[2]BiRC-Bioinformatics Research Center, Department of Genetics and Ecology, University of Aarhus, Aarhus, Denmar

```
96000 { p1.setSubpopulationSize(10000); }
104000 { p1.setSubpopulationSize(1000); }
106000 { p1.setSubpopulationSize(10000); }
114000 { p1.setSubpopulationSize(1000); }
116000 { p1.setSubpopulationSize(10000); }
124000 { p1.setSubpopulationSize(1000); }
126000 { p1.setSubpopulationSize(10000); }
134000 { p1.setSubpopulationSize(1000); }
135100 late() { p1.outputSample(20); }
135100 late() { sim.outputFixedMutations(); }
```

Box S1: SLIM command lines for the main demographic scenarios (stable N, temperate species and alpine species with a soft bottleneck) and DFE-A. All the simulations have been rescaled by a factor ten after checking that it will not affect the result, for time saving considerations.

Marjolaine Rousselle*[1], Maeva Mollion[2], Benoit Nabholz[1], Tomas Bataillon[2] and Nicolas Galtier[1]

[1]UMR-5554 Institut des Sciences de l'Evolution, CNRS, Université de Montpellier, IRD, EPHE, Place E. Bataillon, Montpellier, France

[2]BiRC-Bioinformatics Research Center, Department of Genetics and Ecology, University of Aarhus, Aarhus, Denmar

Annexe 2
Supplementary Material of Chapter 2

Supplementaty Material

Species	Bioproject	data_type	number_of_individuals	publication
Gorilla gorilla	PRJNA189439	Genome	20	Prado-Martinez et al. 2013
Homo sapiens	PRJEB8350	Exome	19	Teixeira et al. 2015
Pan troglodytes	PRJEB8350	Exome	20	Teixeira et al. 2015
Papio anubis	PRJNA54005	Genome	5	unpublished baboon genome project
Pongo abelii	PRJNA189439 and PRJEB1675	Genome	10	Prado-Martinez et al. 2013 and -
Macaca mulatta	PRJNA251548	Exome	20	Xue et al. 2016
Meleagris gallopavo	PRJNA271731	RNA_seq	10	Wright et al. 2015
Phasianus colchicus	PRJNA271731	RNA_seq	10	Wright et al. 2015
Pavo cristatus	PRJNA271731	RNA_seq	10	Wright et al. 2015
Numida meleagris	PRJNA271731	RNA_seq	7	Wright et al. 2015
Anas platyrhynchos	PRJNA271731	RNA_seq	10	Wright et al. 2015
Anser cygmoides	PRJNA271731	RNA_seq	10	Wright et al. 2015

Table S1: Details of the Bioprojects used in this study to retrieve reads data.

species	dN				dS				dNdS			
	all	WS	SW	GC-conservative	all	WS	SW	GC-conservative	all	WS	SW	GC-conservative
M. mulatta	0.867**	0.794**	0.782*	0.769*	0.83**	0.830**	0.939**	0.806**	-0.685*	-0.782*	-0.479	-0.588
H. sapiens	0.806**	0.855**	0.515	0.806**	0.988**	0.867**	0.927**	0.915**	-0.685*	-0.467	-0.321	-0.818**
G. gorilla	0.224	0.539	-0.127	0.248	0.818**	0.769*	0.624	0.782*	-0.624	-0.758*	-0.515	-0.576
P. troglodytes	0.782*	0.648*	0.842**	0.696*	0.976**	0.988**	0.952**	0.842**	-0.430	-0.842**	0.224	-0.176
P. anubis	0.636**	0.758*	0.406	0.563	0.818**	0.830**	0.782*	0.830**	-0.721*	-0.515	-0.418	-0.672*
P. abelii	0.624	0.358	0.685*	0.624	0.830**	0.806**	0.806**	0.685*	-0.927**	-0.879**	-0.467	-0.624
M. gallopavo	0.273	0.164	-0.127	0.176	0.455	0.794**	-0.552	0.745*	-0.006	-0.733*	0.418	-0.345
N. meleagris	0.685*	0.830**	0.272	0.006	0.915**	0.988**	-0.2	0.709*	-0.612	-0.879**	0.358	-0.6
P. cristatus	0.818**	0.758*	0.394	0.733*	0.818**	0.842**	-0.261	0.624	-0.261	-0.818**	0.442	-0.091
P. colchicus	0.552	0.176	0.709*	0.515	0.733*	0.830**	0.418	0.673*	-0.503	-0.636*	-0.151	-0.358
A. cygnoides	0.745*	0.830**	0.564	0.696*	0.891**	0.915**	-0.285	0.6	0.176	-0.358	0.430	0.054
A. platyrhynchos	0.564	0.164	0.745*	0.576	0.757*	0.891**	-0.115	0.867**	-0.091	-0.794**	0.539	-0.733*

Table S4: Spearman correlation coefficients between r and divergence estimates obtained with a model assuming non-stationarity.

Significance levels are showed with * (* p-value< 0.05, ** p-value < 0.01) before the FDR correction, and with two shades of red (if the correlation is positive) or green (if the correlation in negative) after FDR correction (light :p-value< 0.05, dark : p-value< 0.01).

Supplementaty Material

species	dN				dS				dNdS			
	all	WS	SW	GC-conservative	all	WS	SW	GC-conservative	all	WS	SW	GC-conservative
M. mulatta	0.903**	0.952**	-0.139	0.927**	0.939**	0.988**	0.055	0.939**	-0.855**	-0.988**	-0.418	-0.964**
H. sapiens	0.879**	0.491	0.842**	0.915**	1**	0.988**	0.758*	1**	-0.588	-0.988**	0.588	-0.636
G. gorilla	-0.442	-0.067	-0.782	0.188	0.612	0.927**	-0.6	0.418	-0.976**	-0.989**	-0.939**	-0.915**
P. troglodytes	0.915**	0.673*	0.939**	0.891**	1**	1**	0.867**	1**	#VALEUR !	-0.976**	0.782*	-0.552
P. anubis	-0.503	-0.321	-0.915**	0.697*	0.939**	1**	-0.891**	0.939**	-0.976**	-0.951**	-0.564	-0.988**
P. abelii	0.891**	0.406	0.769*	0.939**	1**	1**	0.988**	0.988*	-0.976**	-0.988**	0.673*	-0.952**
M. gallopavo	-0.467	-0.588	-0.188	-0.673*	0.309	0.915**	-0.745*	-0.588	-0.709*	-0.903**	0.3812	0.479
N. meleagris	-0.285	-0.6	0.042	-0.539	0.685*	0.964**	-0.903**	-0.563	-0.430	-0.915**	0.624	0.358
P. cristatus	-0.248	-0.394	0.236	-0.491	0.394	0.976**	-0.079	-0.503	-0.551	-0.952**	0.285	0.418
P. colchicus	-0.067	-0.030	0.2	-0.418	0.806**	0.976**	0.212	-0.503	-0.636	-0.939**	0.091	0.527
A. cygnoides	0.576	0.745*	0.273	-0.139	0.867**	1**	-0.127	-0.067	-0.6	-0.915**	0.442	0.176
A. platyrhynchos	0.527	0.648*	0.345	-0.091	0.891**	0.988**	-0.624	0.054	-0.6	-0.951**	0.564	-0.212

Table S5 : Spearman correlation coefficients between GC3 and divergence estimates obtained with a model assuming stationarity.

Significance levels are showed with * (* p-value< 0.05, ** p-value < 0.01) before the FDR correction, and with two shades of red (if the correlation is positive) or green (if the correlation in negative) after FDR correction (light :p-value< 0.05, dark : p-value< 0.01).

species	dN				dS				dNdS			
	all	WS	SW	GC-conservative	all	WS	SW	GC-conservative	all	WS	SW	GC-conservative
M. mulatta	0.685*	0.673*	0.479	0.697*	0.794**	0.782*	0.673*	0.830**	-0.855**	-0.964**	-0.564	-0.636
H. sapiens	0.564	0.636	0.418	0.539	0.855**	0.758*	0.842**	0.709*	-0.806**	-0.745*	-0.285	-0.769*
G. gorilla	0.709*	0.818**	0.115	0.806**	0.927**	0.903**	0.733*	0.915**	-0.867**	-0.660*	-0.842**	-0.758*
P. troglodytes	0.721*	0.721*	0.745*	0.672*	0.952**	1**	0.964**	0.879**	-0.212	-0.879**	0.236	-0.564
P. anubis	0.261	0.442	0.321	0.455	0.758*	0.879**	0.661*	0.673*	-0.769*	-0.830**	-0.479	-0.794**
P. abelii	0.6	0.624	0.685*	0.806**	0.842**	0.891**	0.697*	0.842**	-0.927**	-0.927**	-0.018	-0.769*
M. gallopavo	-0.164	-0.079	-0.333	-0.564	0.503	0.782*	-0.721*	-0.782*	-0.612	-0.782*	0.321	0.152
N. meleagris	-0.382	0.369	-0.467	-0.769*	0.806**	0.988**	-0.673*	-0.685*	-0.612	-0.903**	0.176	-0.018
P. cristatus	0.491	0.733*	0.067	0.055	0.758*	0.915**	-0.394	-0.491	-0.297	-0.867**	0.564	0.721*
P. colchicus	-0.018	-0.006	0.381	-0.418	0.830**	0.891**	0.224	-0.685*	-0.552	-0.636	-0.164	0.103
A. cygnoides	0.612	0.745*	0.285	-0.127	0.867**	0.952**	-0.842**	-0.564	0.091	-0.552	0.661*	0.224
A. platyrhynchos	0.176	0.394	0.709*	-0.321	0.636	0.903**	-0.2	0.685*	-0.636	-0.855**	0.6	-0.769*

Table S6: Spearman correlation coefficients between r and divergence estimates obtained with a model assuming stationarity.

Significance levels are showed with * (* p-value< 0.05, ** p-value < 0.01) before the FDR correction, and with two shades of red (if the correlation is positive) or green (if the correlation in negative) after FDR correction (light :p-value< 0.05, dark : p-value< 0.01).

Supplementaty Material

species	π_n			π_s			π_n/π_s		
	WS	SW	GC-conservative	WS	SW	GC-conservative	WS	SW	GC-conservative
M. mulatta	-0.939**	-0.454	-0.527	-0.236	0.406	0.236	-0.915**	-0.333	-0.442
H. sapiens	-0.3	-0.030	-0.483	0.483	0.139	0.3	-0.216	0.103	-0.7*
G. gorilla	-0.757*	-0.6	-0.595	-0.612	-0.612	0.071	-0.115	-0.830**	-0.833*
P. troglodytes	-0.951**	-0.745*	-0.683	-0.648*	-0.575	-0.383	-0.890**	-0.733*	-0.283
P. anubis	-0.890**	-0.563	-0.818**	-0.418	0.163	0.333	-0.684*	-0.660*	-0.527
P. abelii	-0.612	-0.272	-0.103	-0.296	0.151	-0.151	-0.248	-0.806**	-0.212
M. gallopavo	-0.854**	0.151	-0.478	-0.333	0.224	-0.2	-0.696*	0.139	-0.284
N. meleagris	-0.672*	0.866**	-0.321	0.709*	0.963**	0.381	-0.842**	-0.430	-0.624
P. cristatus	-0.793**	-0.527	-0.454	0.684*	0.745*	0.490	-0.903**	-0.539	-0.539
P. colchicus	-0.890**	0.418	-0.806**	-0.539	0.527	-0.539	-0.939**	-0.345	-0.745*
A. cygnoides	-0.709*	-0.296	-0.733*	0.890**	0.296	0.418	-0.818**	-0.163	-0.781*
A. platyrhynchos	-0.878**	0.951**	0.309	0.963**	0.987**	0.987**	-0.975**	-0.636	-0.830**

Table S7: Spearman correlation coefficients between GC3 and π_n, π_s, π_n/π_s estimates obtained after masking CpG sites from the alignments.

Significance levels are showed with * (* p-value< 0.05, ** p-value < 0.01) before the FDR correction, and with two shades of red (if the correlation is positive) or green (if the correlation in negative) after FDR correction (light :p-value< 0.05, dark : p-value< 0.01).

species	α			ω_a			ω_{na}		
	WS	SW	GC-conservative	WS	SW	GC-conservative	WS	SW	GC-conservative
M. mulatta	0.672*	0.0181	0.381	0.090	-0.151	0.006	-0.939**	-0.466	-0.587
H. sapiens	0.515	0.878**	0.393	-0.042	0.878*	0.2	-0.6	-0.709*	-0.393
G. gorilla	0.716*	0.284	0.384	0.533	0.284	0.266	-0.833**	-0.503	-0.433
P. troglodytes	0.454	0.563	-0.212	0.127	0.696*	-0.393	-0.587	-0.430	0.187
P. anubis	0.115	-0.163	0.139	-0.418	-0.163	-0.2	-0.6	-0.103	-0.139
P. abelii	0.066	0.684*	0.272	-0.284	0.721*	0.103	-0.503	-0.660*	-0.503
M. gallopavo	0.684*	-0.078	0.284	0.321	-0.151	0.127	-0.818**	0.224	-0.345
N. meleagris	0.648*	0.624	0.624	-0.163	0.563	0.466	-0.903**	-0.721*	-0.721*
P. cristatus	-0.066	0.393	0.078	-0.187	0.393	-0.248	-0.563	-0.454	-0.345
P. colchicus	0.833*	0.030	0.757*	0.833*	-0.030	0.745*	-0.952**	-0.054	-0.721*
A. cygnoides	0.757*	-0.212	0.139	0.357	-0.236	-0.503	-0.878**	0.296	-0.284
A. platyrhynchos	0.612	0.430	-0.187	-0.903**	0.430	-0.733*	-0.963**	-0.503	-0.563

Table S8: Spearman correlation coefficients between GC3 and α, ω_a and ω_{na} estimates for the GammaExpomodel.

Significance levels are showed with * (* p-value< 0.05, ** p-value < 0.01) before the FDR correction, and with two shades of red (if the correlation is positive) or green (if the correlation in negative) after FDR correction (light :p-value< 0.05, dark : p-value< 0.01).

Supplementaty Material

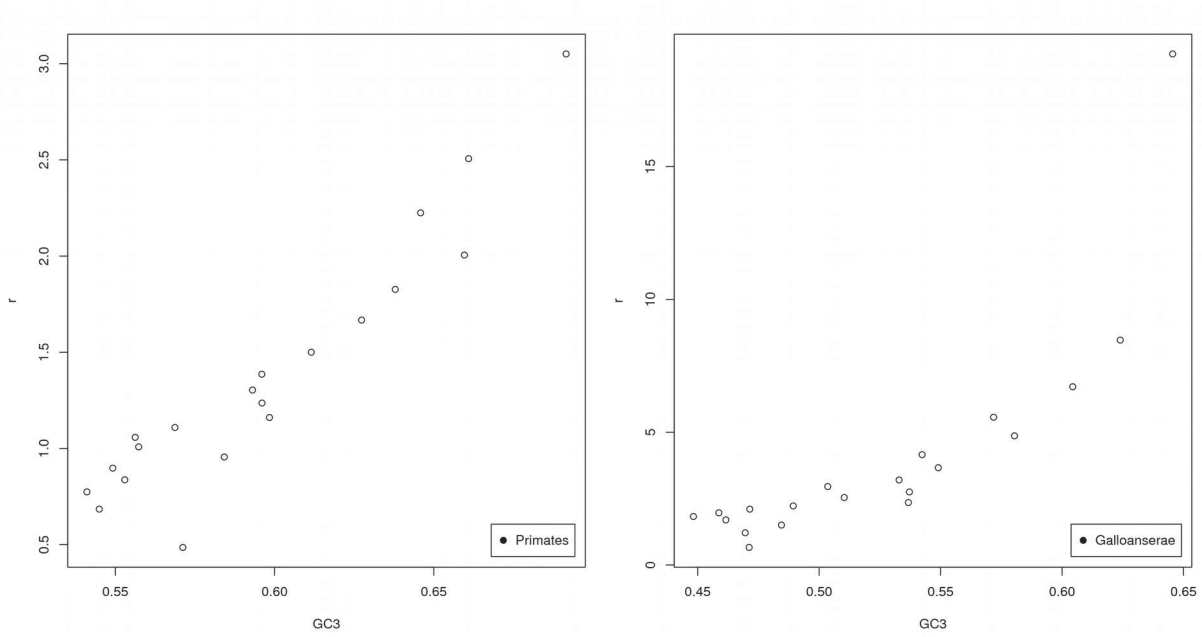


Figure S1: Correlation between GC3 and r obtained with Spearman correlation for *H. sapiens* (left) and *G. gallus* (right).

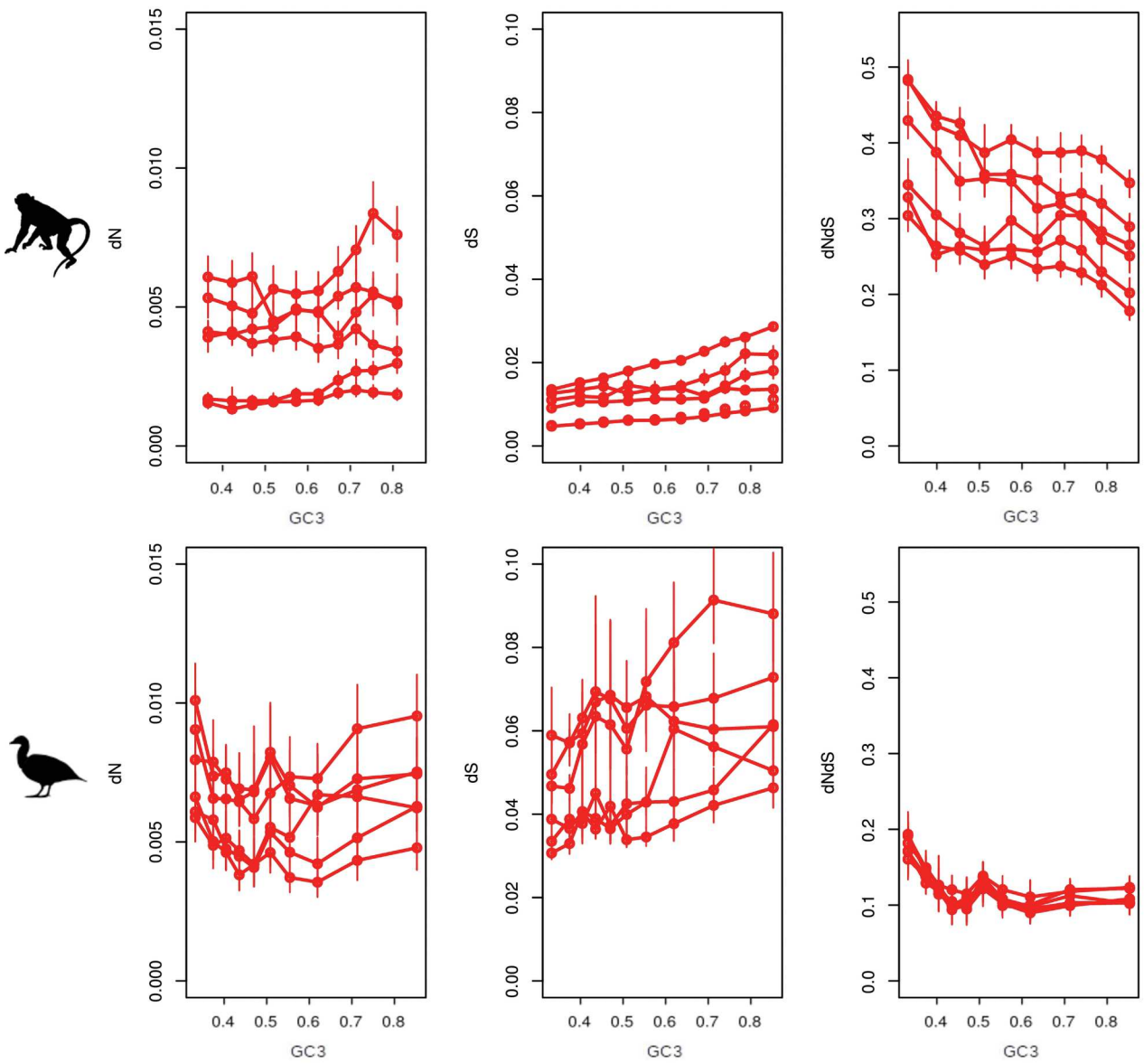


Figure S2: dN, dS and dN/dS ratio against GC3 for each species for all substitution type taken together.
Statistics are estimated under a model assuming base composition stationarity (above: primates, below: fowls).

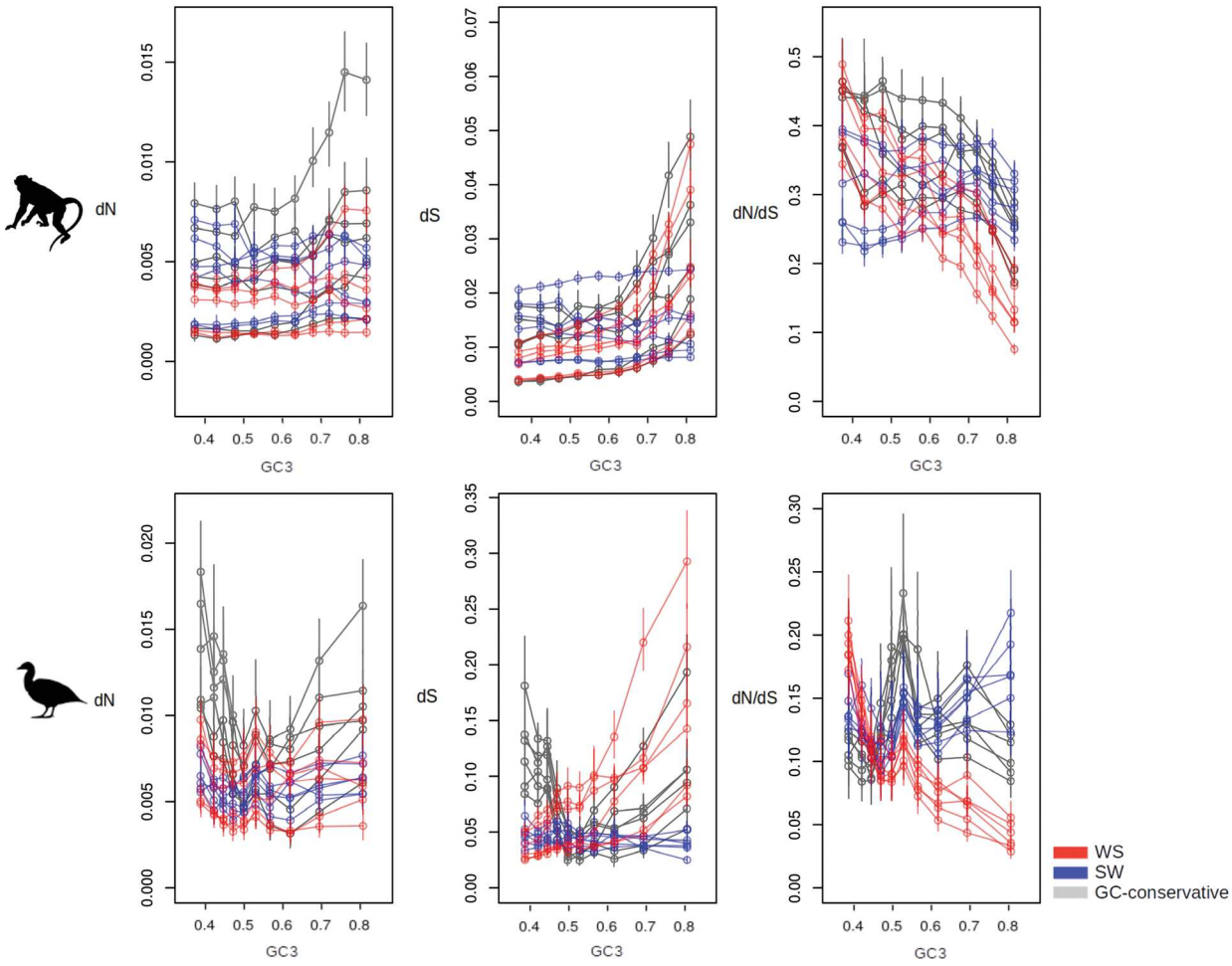


Figure S3: dN, dS and dN/dS ratio against GC3 for each species and each type of substitutions (WS, SW and GC-conservative substitution).
Statistics are estimated under a model assuming base composition stationarity (above: primates, below: fowls).

Supplementary Material

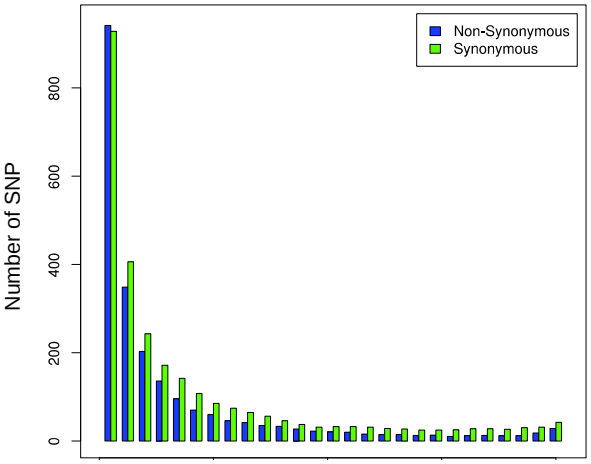


Figure S4 : Site frequency spectra of *Homo sapiens* with all mutation type without masking CpG sites.

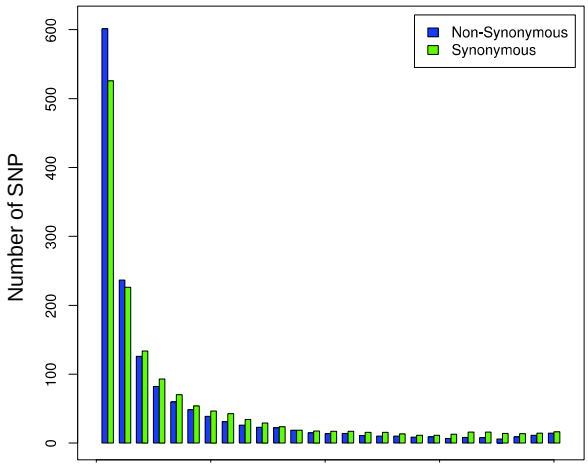


Figure S5 : Site frequency spectra of *Homo sapiens* with all mutation type with a masking of CpG sites.

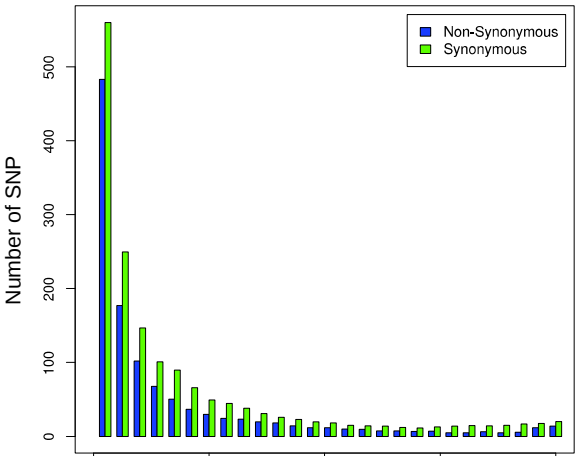


Figure S6 : Site frequency spectra of *Homo sapiens* with SW mutation type without masking CpG sites.

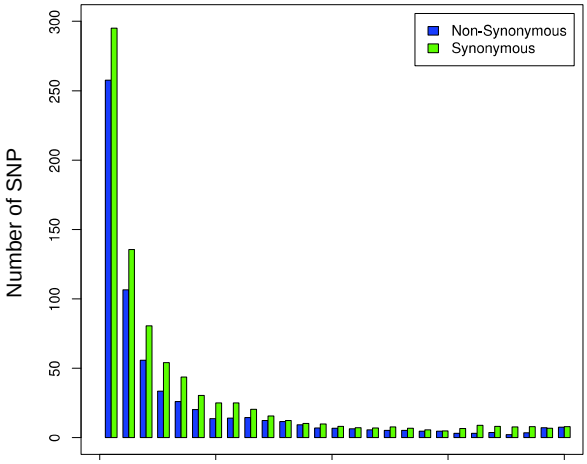


Figure S7 : Site frequency spectra of *Homo sapiens* with SW mutation type with a masking of CpG sites.

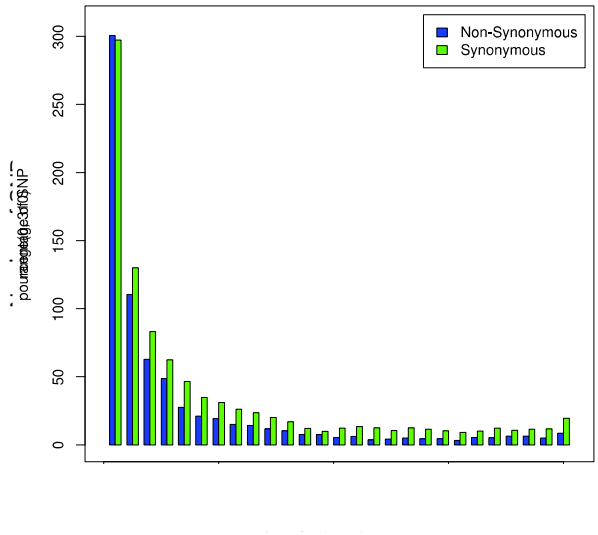


Figure S8 : Site frequency spectra of *Homo sapiens* with WS mutations without masking CpG sites.

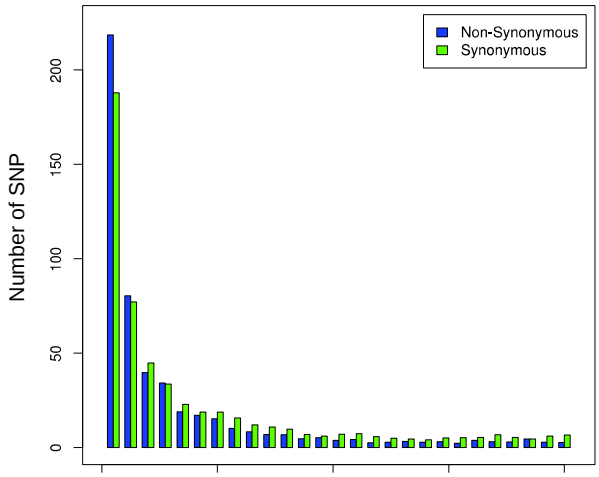


Figure S9 : Site frequency spectra of *Homo sapiens* with WS mutations with a masking of CpG sites.

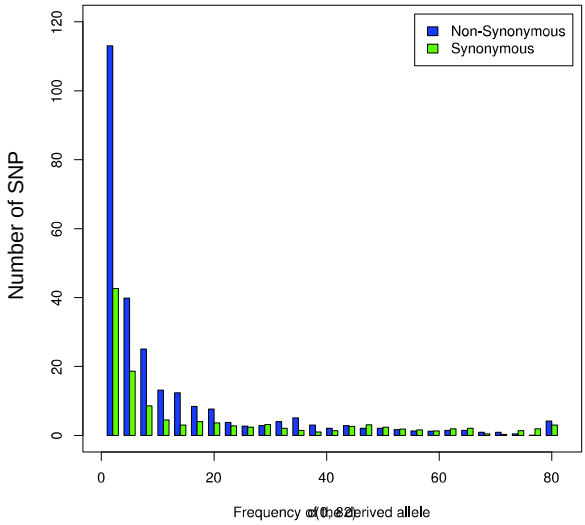


Figure S10 : Site frequency spectra of *Homo sapiens* with GC-conservative mutations without masking CpG sites.

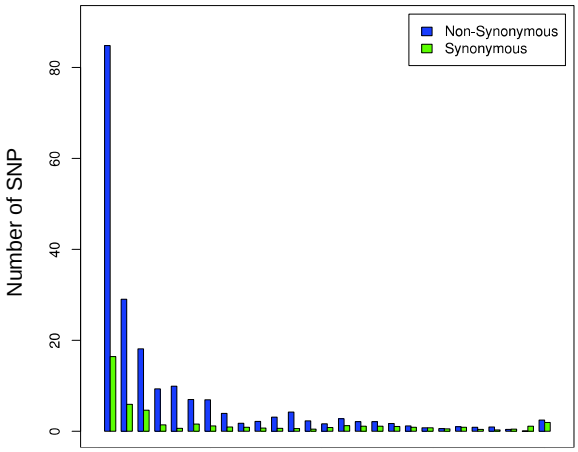


Figure S11 : Site frequency spectra of *Homo sapiens* with GC-conservative mutations with a masking of CpG sites.

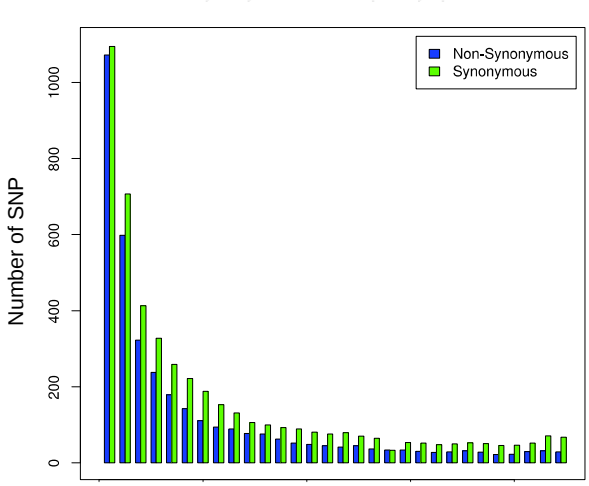


Figure S12 : Site frequency spectra of *Gorilla gorilla* with all mutation type without masking CpG sites.

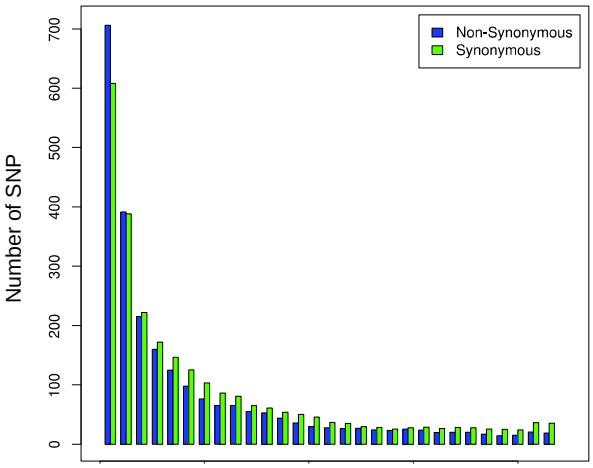


Figure S13 : Site frequency spectra of *Gorilla gorilla* with all mutation type with a masking of CpG sites.

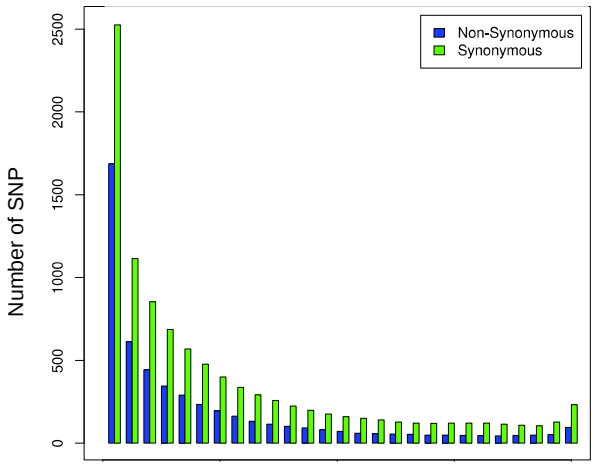


Figure S14 : Site frequency spectra of *Macaca mulatta* with all mutation type without maski

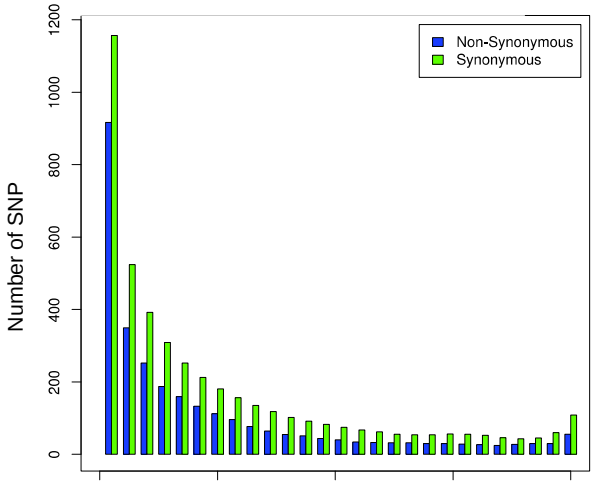


Figure S15 : Site frequency spectra of *Macaca mulatta* with all mutation type with a masking of C

Supplementary Material

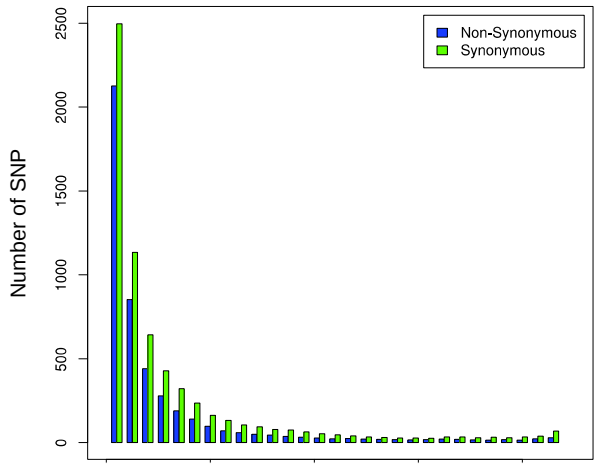


Figure S16 : Site frequency spectra of *Pan troglodytes* with all mutation type without masking CpG sites.

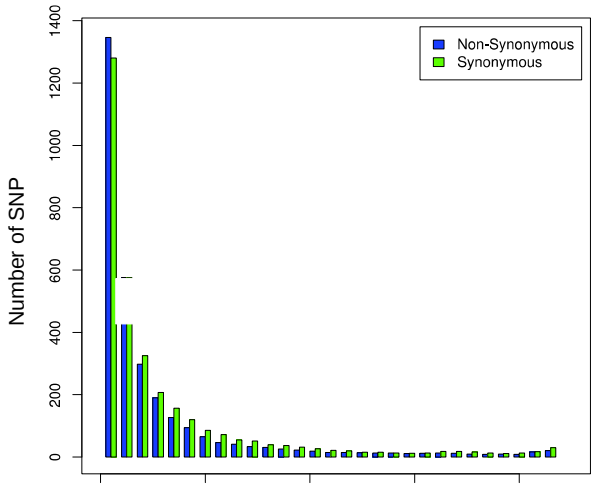


Figure S17 : Site frequency spectra of *Pan troglodytes* with all mutation type with a masking of CpG sites.

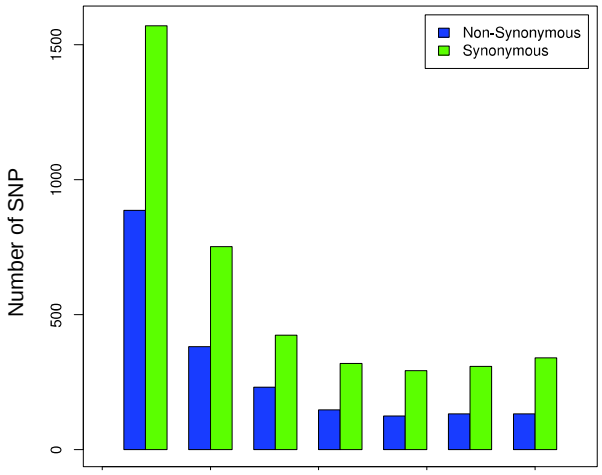


Figure S18 : Site frequency spectra of *Papio anubis* with all mutation type without masking CpG sites.

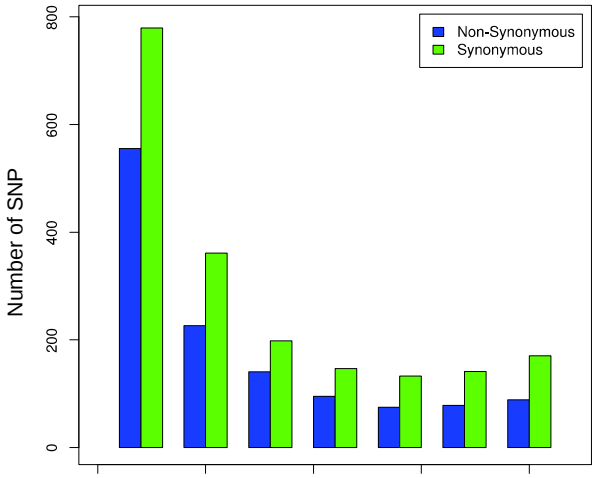


Figure S19 : Site frequency spectra of *Papio anubis* with all mutation type with a masking of CpG sites.

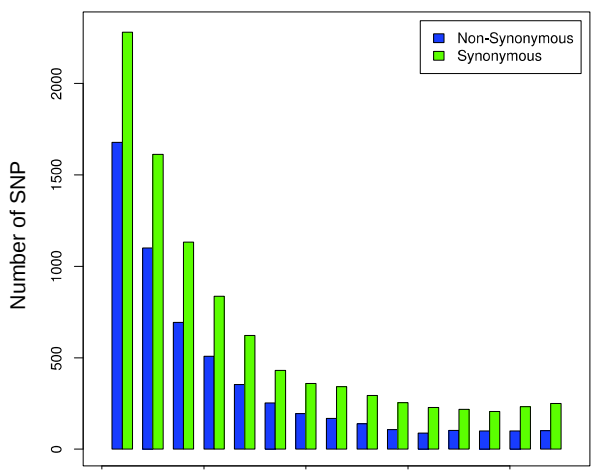


Figure S20 : Site frequency spectra of *Pongo abelii* with all mutation type without masking CpG sites.

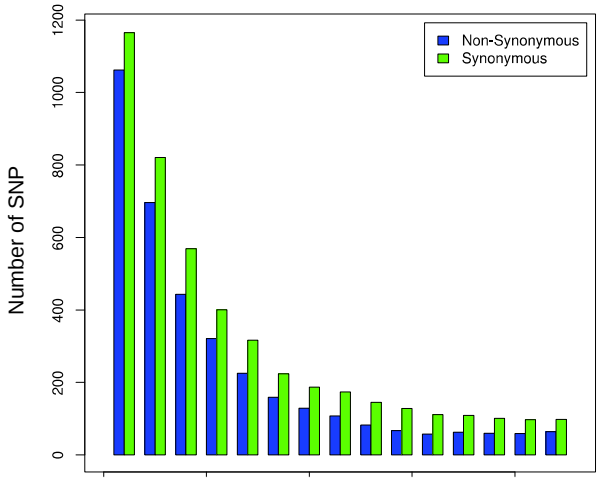


Figure S21 : Site frequency spectra of *Pongo abelii* with all mutation type with a masking of CpG sites.

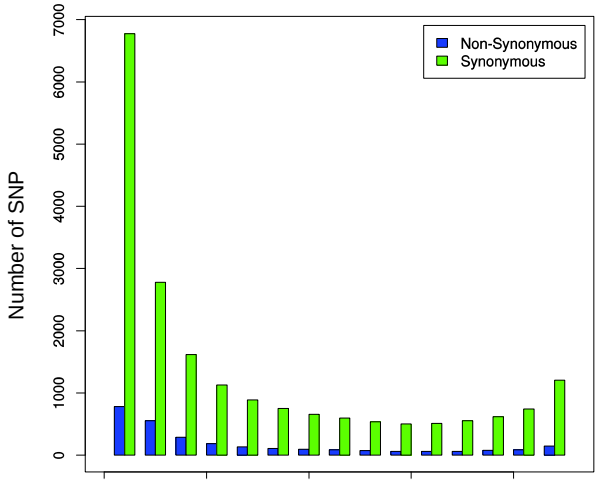


Figure S22 : Site frequency spectra of *Anas platyrhynchos* with all mutation type without masking CpG sites.

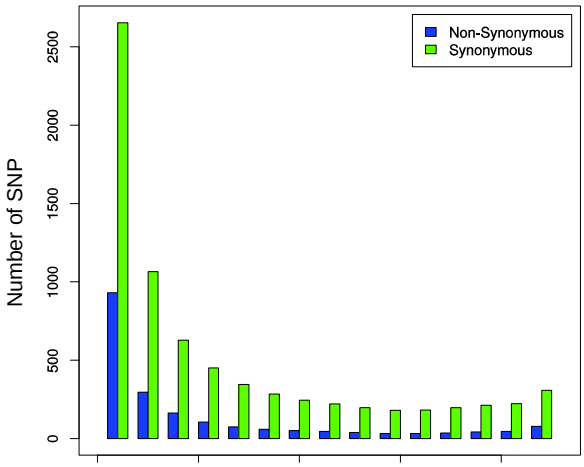


Figure S23 : Site frequency spectra of *Anas platyrhynchos* with all mutation type with a masking of CpG sites.

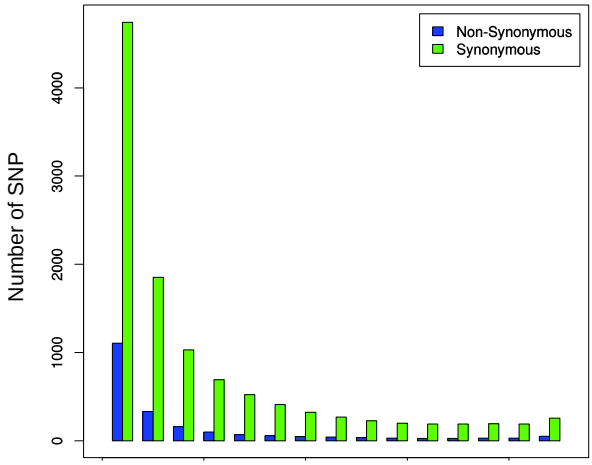


Figure S24 : Site frequency spectra of *Anas platyrhynchos* with SW mutations without masking CpG sites.

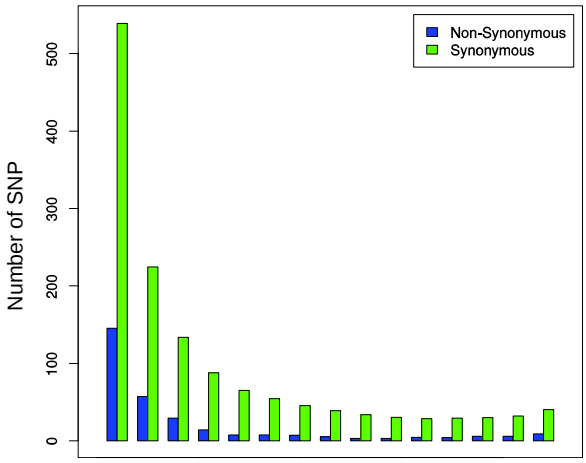


Figure S25 : Site frequency spectra of *Anas platyrhynchos* with SW mutations with a masking of CpG sites.

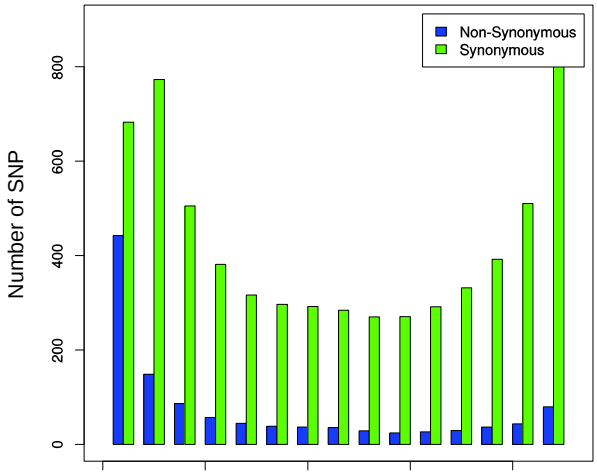


Figure S26 : Site frequency spectra of *Anas platyrhynchos* with WS mutations without masking CpG sites.

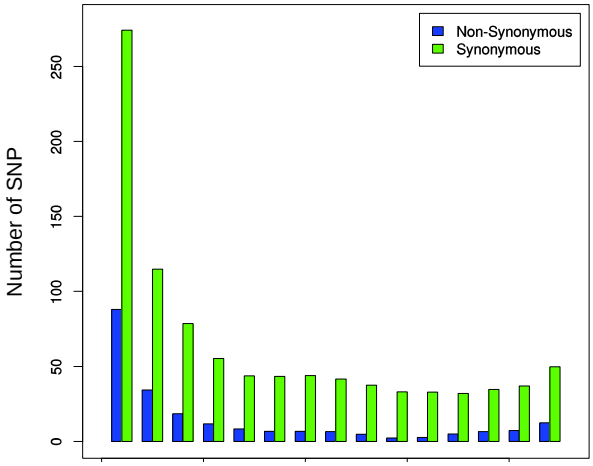


Figure S27 : Site frequency spectra of *Anas platyrhynchos* with WS mutations with a masking of CpG sites.

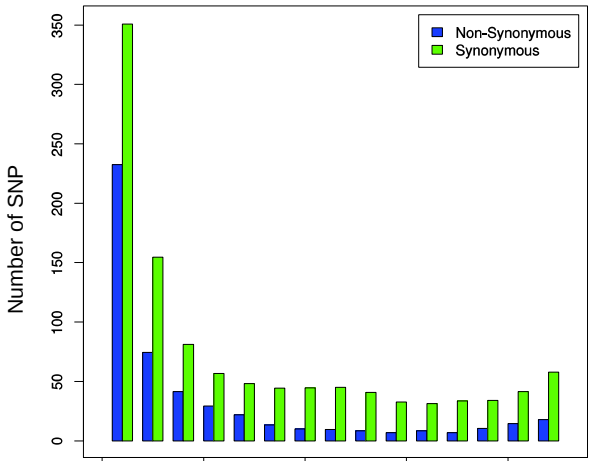


Figure S28 : Site frequency spectra of *Anas platyrhynchos* with GC-conservative mutations without masking CpG sites.

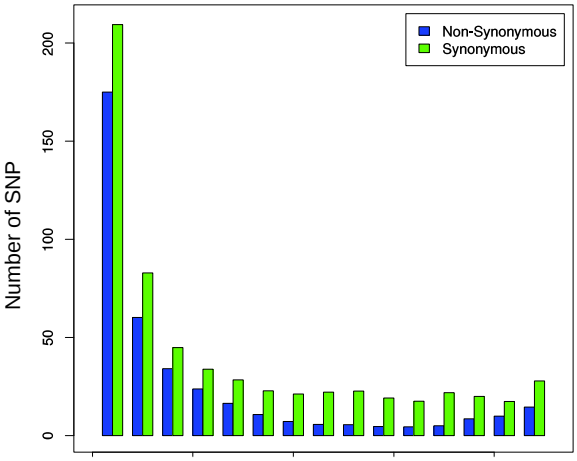


Figure S29 : Site frequency spectra of *Anas platyrhynchos* with GC-conservative mutations with a masking of CpG sites.

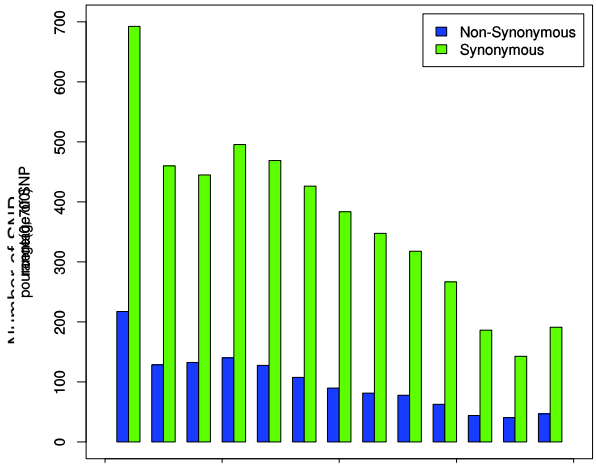


Figure S30 : Site frequency spectra of *Anser cygnoides* with all mutation type without masking CpG sites.

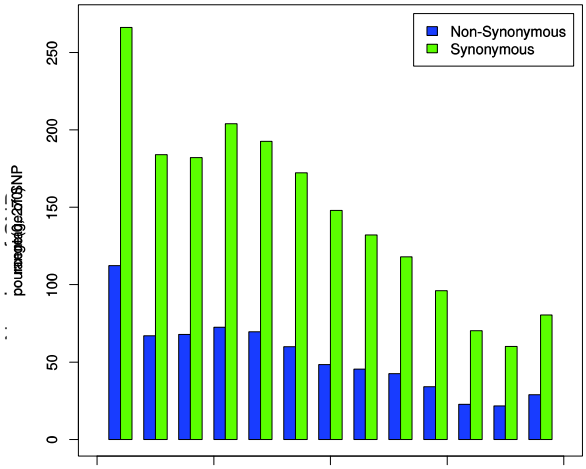


Figure S31 : Site frequency spectra of *Anser cygnoides* with all mutation type with a masking of CpG sites.

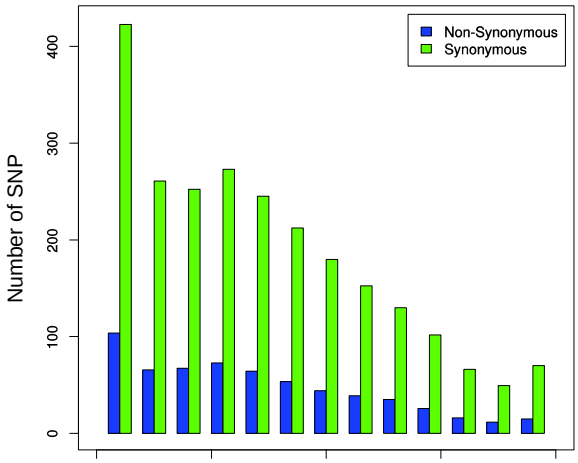


Figure S32 : Site frequency spectra of *Anser cygnoides* with SW mutations without masking CpG sites.

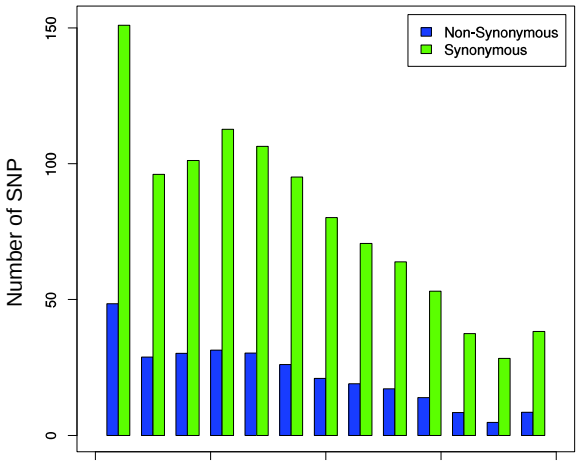


Figure S33 : Site frequency spectra of *Anser cygnoides* with SW mutations with a masking of CpG sites.

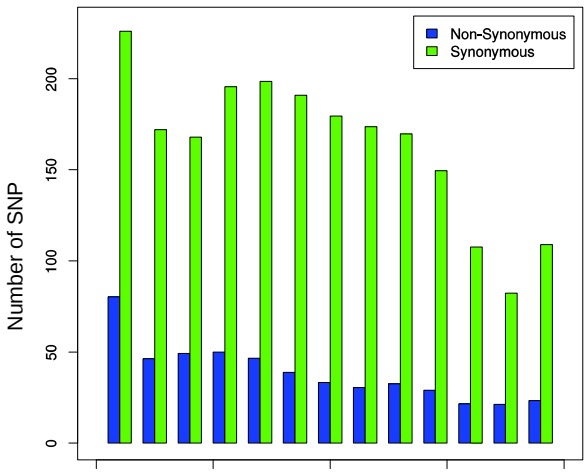


Figure S34 : Site frequency spectra of *Anser cygnoides* with WS mutations without masking CpG sites.

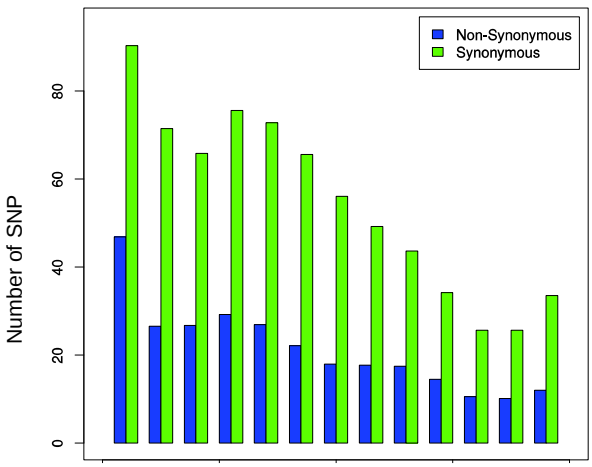


Figure S35: Site frequency spectra of *Anser cygnoides* with WS mutations with a masking of CpG sites.

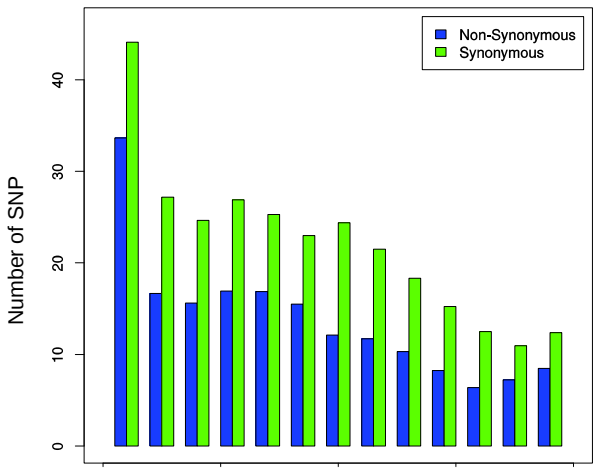


Figure S36 : Site frequency spectra of *Anser cygnoides* with GC-conservative mutations without masking CpG sites.

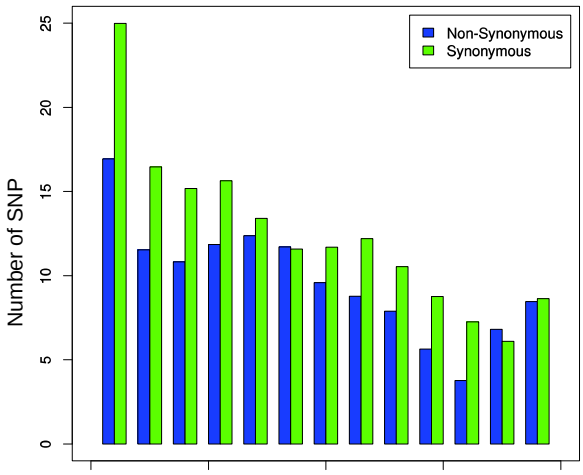


Figure S37: Site frequency spectra of *Anser cygnoides* with GC-conservative mutations with a masking of CpG sites.

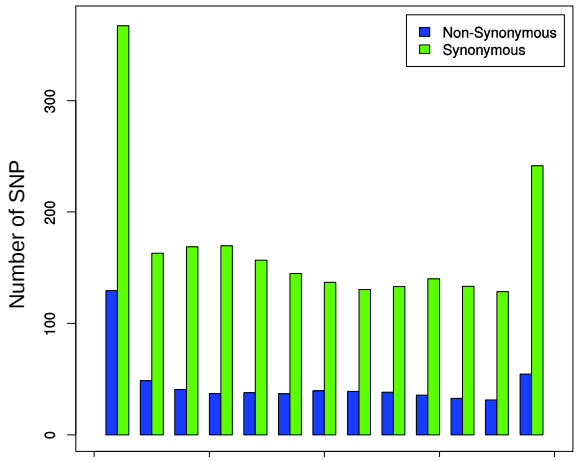


Figure S38 : Site frequency spectra of *Pavo cristatus* with all mutation type without masking CpG sites.

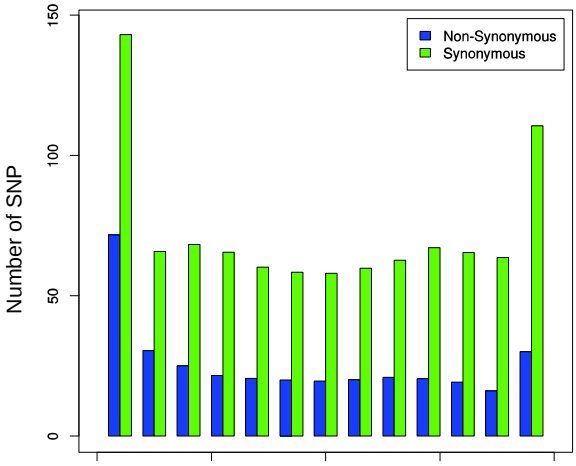


Figure S39: Site frequency spectra of *Pavo cristatus* with all mutation type with a masking of CpG sites.

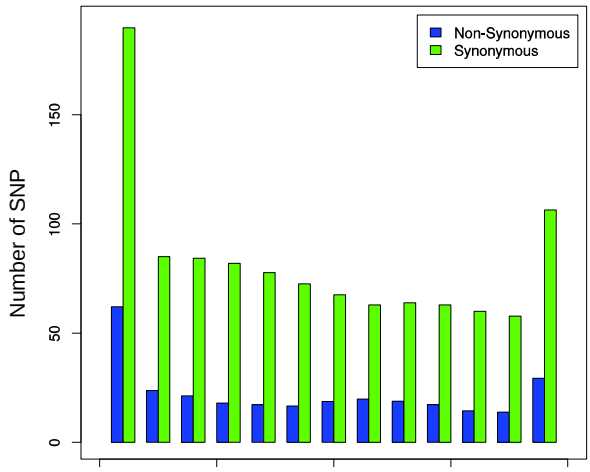


Figure S40 : Site frequency spectra of *Pavo cristatus* with SW mutations without masking CpG sites.

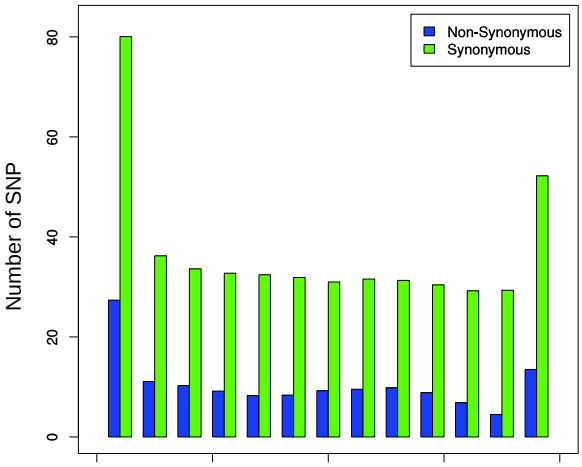


Figure S41: Site frequency spectra of *Pavo cristatus* with SW mutations with a masking of CpG sites.

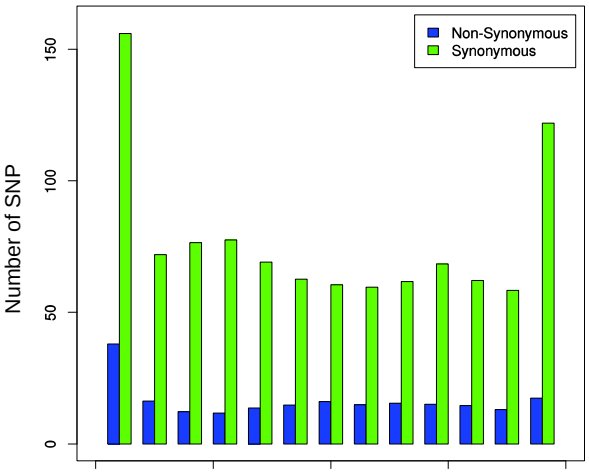


Figure S42 : Site frequency spectra of *Pavo cristatus* with WS mutations without masking CpG sites.

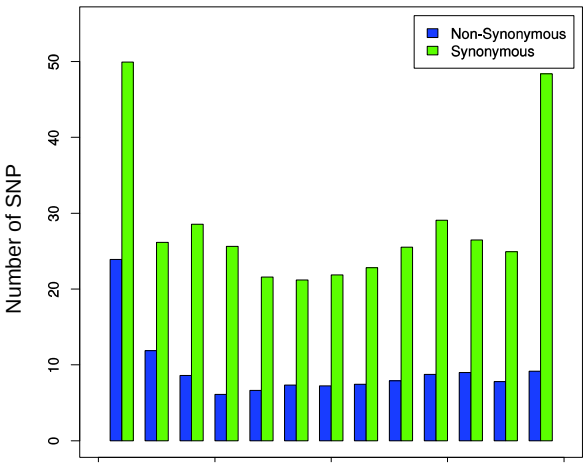


Figure S43: Site frequency spectra of *Pavo cristatus* with WS mutations with a masking of CpG sites.

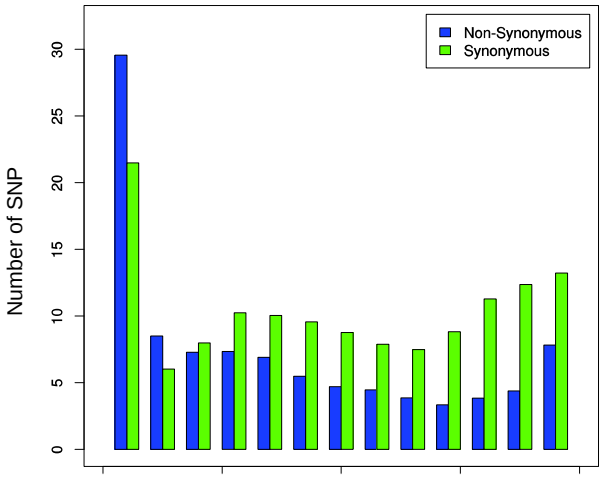


Figure S44 : Site frequency spectra of *Pavo cristatus* with GC-conservative mutations without masking CpG sites.

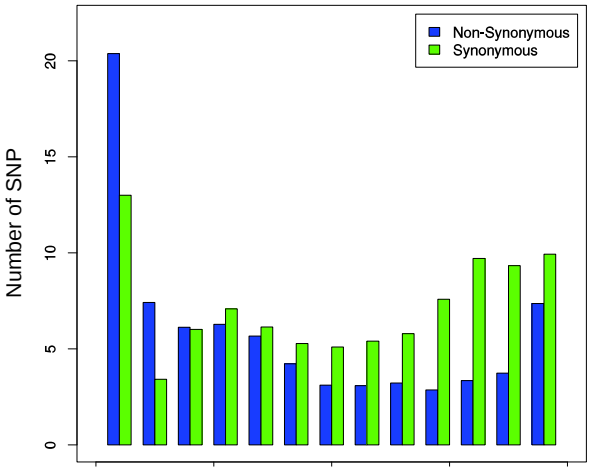


Figure S45: Site frequency spectra of *Pavo cristatus* with GC-conservative mutations with a masking of CpG sites.

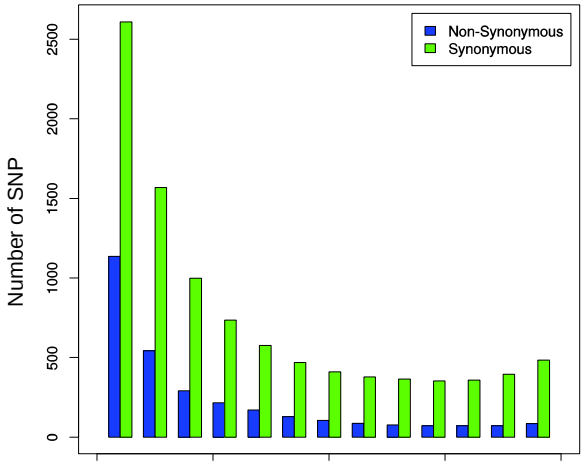


Figure S46 : Site frequency spectra of *Phasianus colchicus* with all mutation type without masking CpG sites.

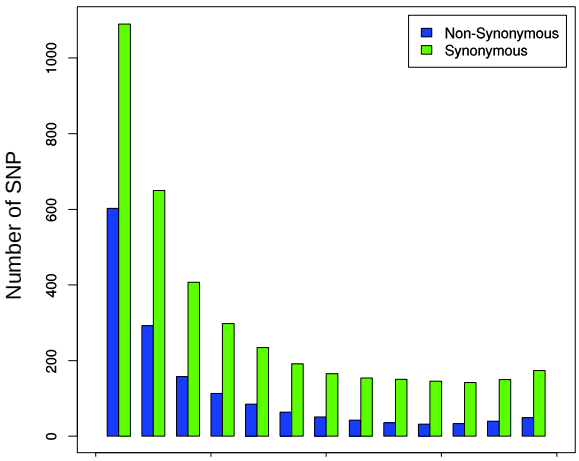


Figure S47: Site frequency spectra of *Phasianus colchicus* with all mutation type with a masking of CpG sites.

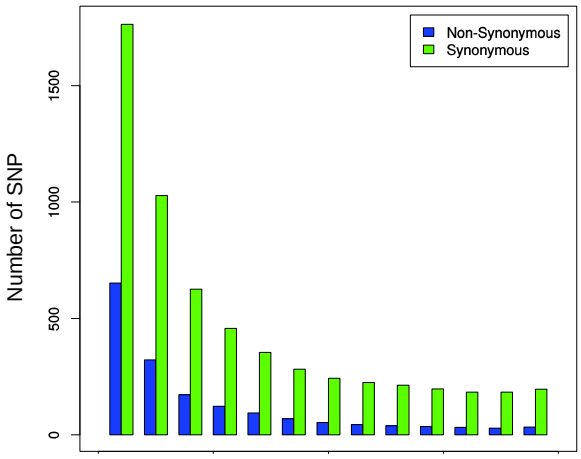


Figure S48 : Site frequency spectra of *Phasianus colchicus* with SW mutations without masking CpG sites.

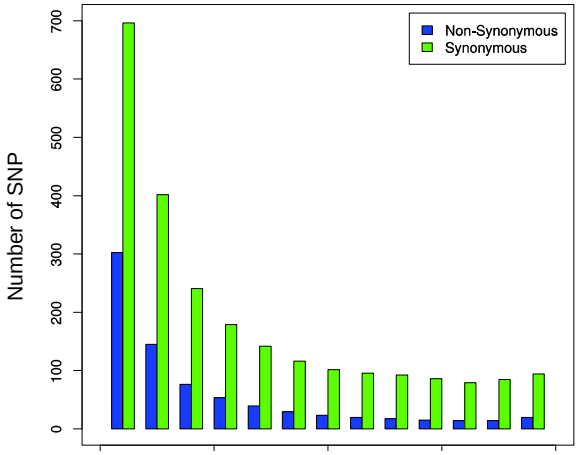


Figure S49: Site frequency spectra of *Phasianus colchicus* with SW mutations with a masking of CpG sites.

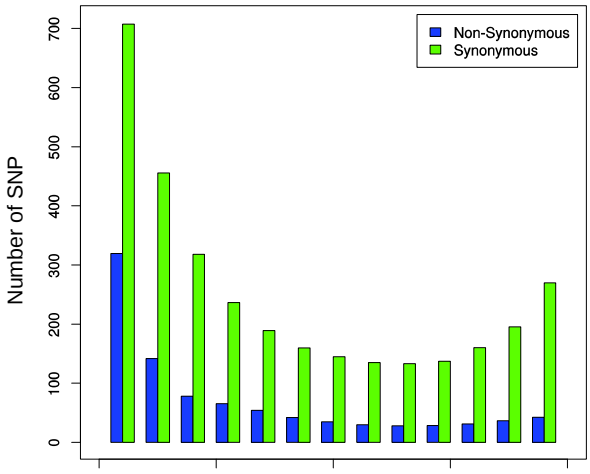


Figure S50 : Site frequency spectra of *Phasianus colchicus* with WS mutations without masking CpG sites.

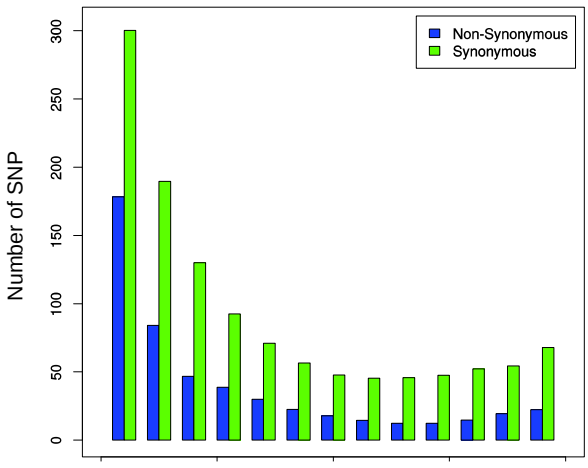


Figure S51: Site frequency spectra of *Phasianus colchicus* with WS mutations with a masking of CpG sites.

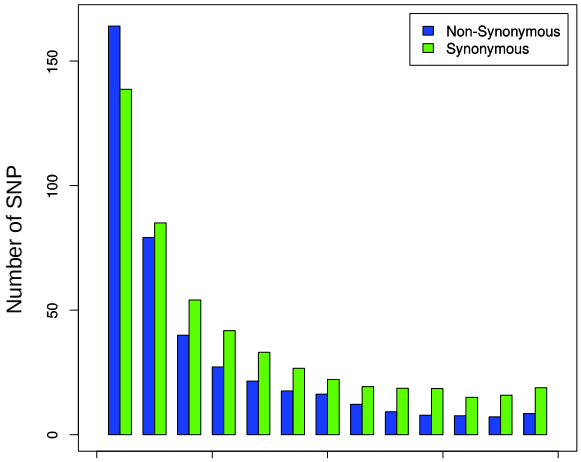


Figure S52 : Site frequency spectra of *Phasianus colchicus* with GC-conservative mutations without masking CpG sites.

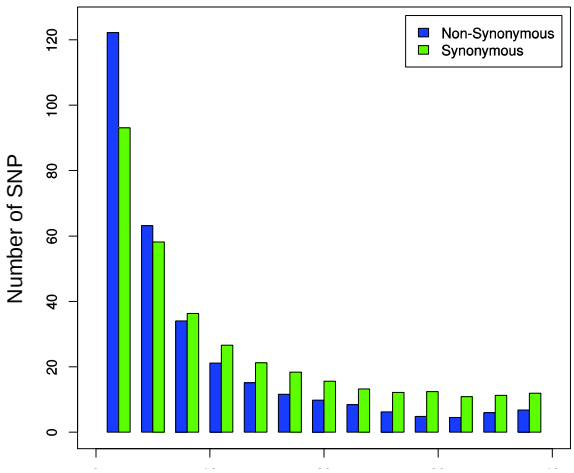


Figure S53: Site frequency spectra of *Phasianus colchicus* with GC-conservative mutations with a masking of CpG sites.

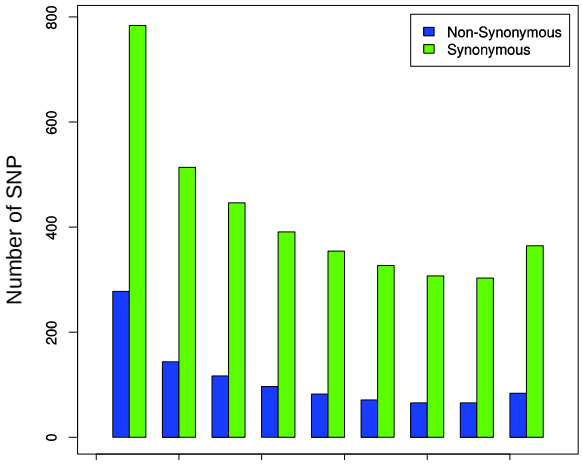


Figure S54 : Site frequency spectra of *Meleagris gallopavo* with all mutation type without masking CpG sites.

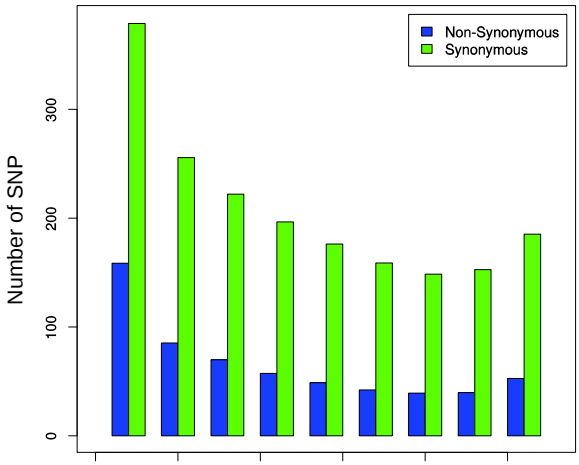


Figure S55: Site frequency spectra of *Meleagris gallopavo* with all mutation type with a masking of CpG sites.

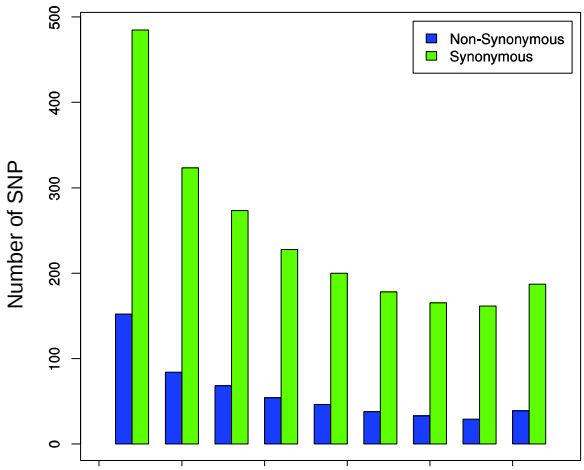


Figure S56 : Site frequency spectra of *Meleagris gallopavo* with SW mutations without masking CpG sites.

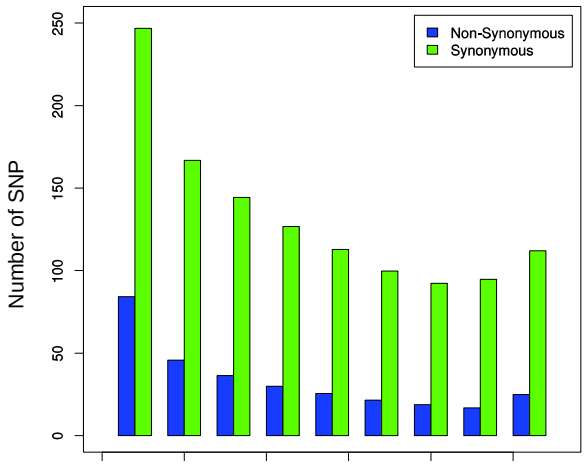


Figure S57: Site frequency spectra of *Meleagris gallopavo* with SW mutations with a masking of CpG sites.

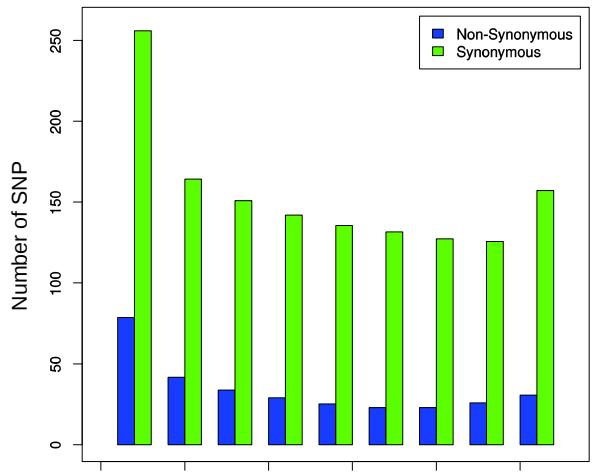


Figure S58 : Site frequency spectra of *Meleagris gallopavo* with WS mutations without masking CpG sites.

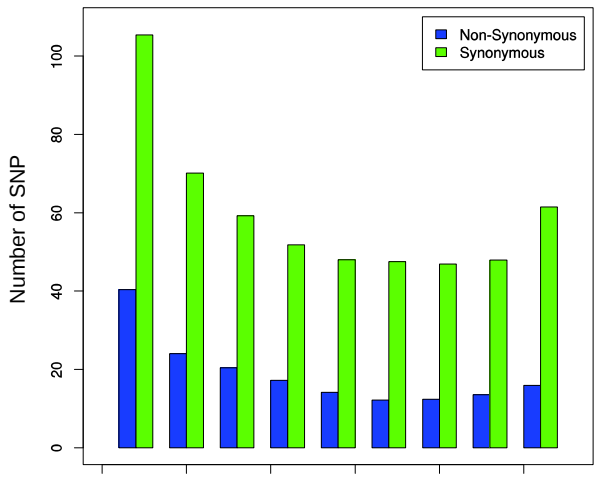


Figure S59: Site frequency spectra of *Meleagris gallopavo* with WS mutations with a masking of CpG sites.

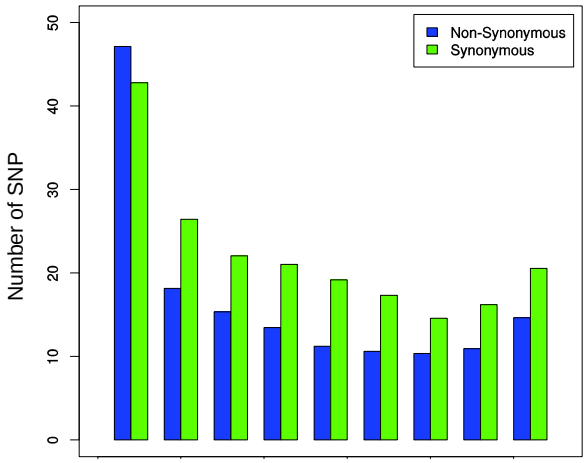


Figure S60 : Site frequency spectra of *Meleagris gallopavo* with GC-conservative mutations without masking CpG sites.

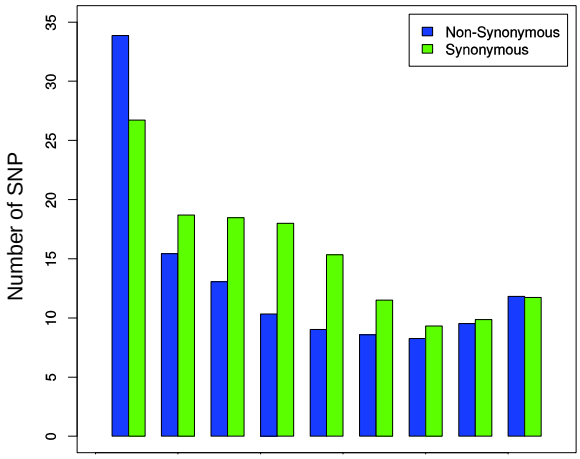


Figure S61: Site frequency spectra of *Meleagris gallopavo* with GC-conservative mutations with a masking of CpG sites.

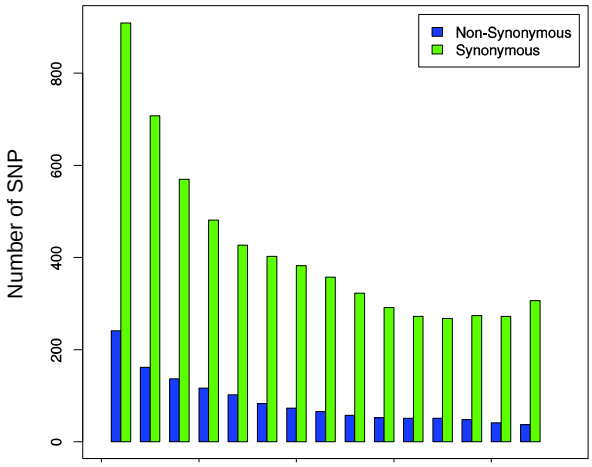


Figure S62 : Site frequency spectra of *Numida gallopavo* with all mutation type without masking CpG sites.

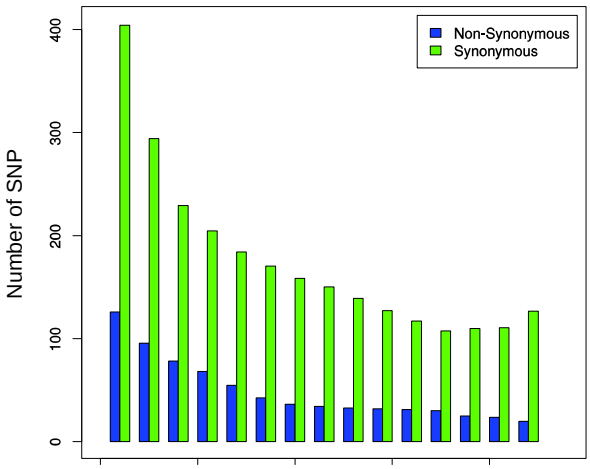


Figure S63: Site frequency spectra of *Numida meleagris* with all mutation type with a masking of CpG sites.

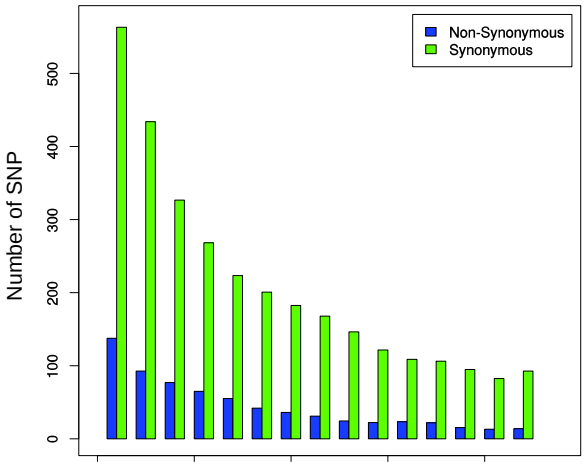


Figure S64 : Site frequency spectra of *Numida gallopavo* with SW mutations without masking CpG sites.

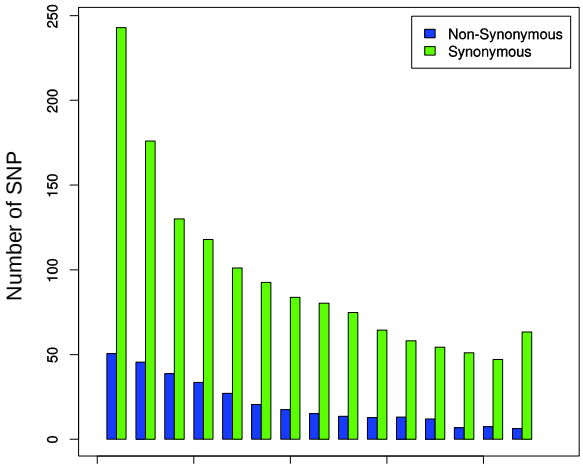


Figure S65: Site frequency spectra of *Numida meleagris* with SW mutations with a masking of CpG sites.

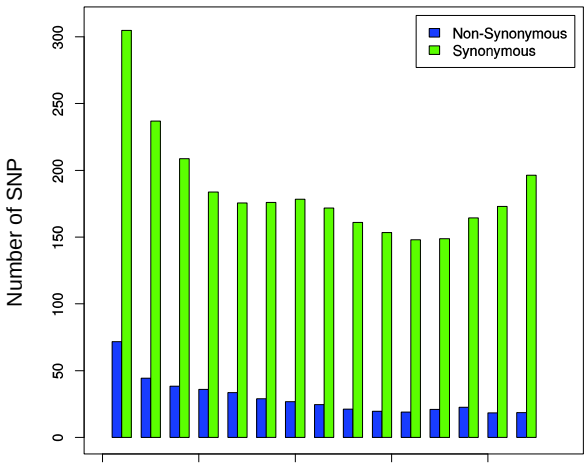


Figure S66 : Site frequency spectra of *Numida gallopavo* with WS mutations without masking CpG sites.

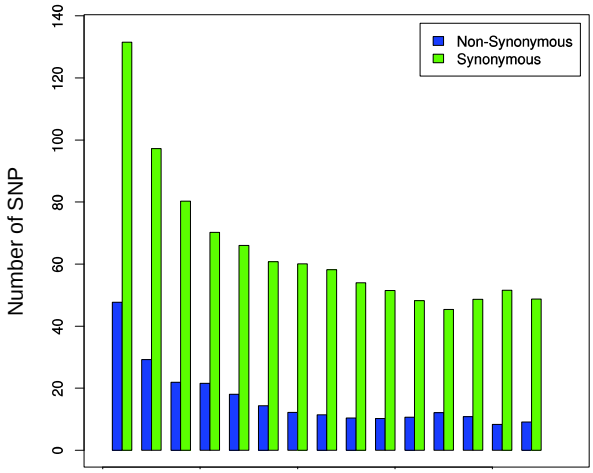


Figure S67: Site frequency spectra of *Numida meleagris* with WS mutations with a masking of CpG sites.

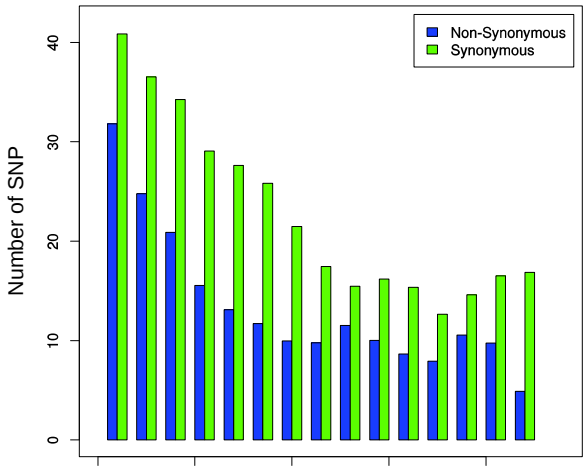


Figure S68 : Site frequency spectra of *Numida gallopavo* with GC-conservative mutations without masking CpG sites.

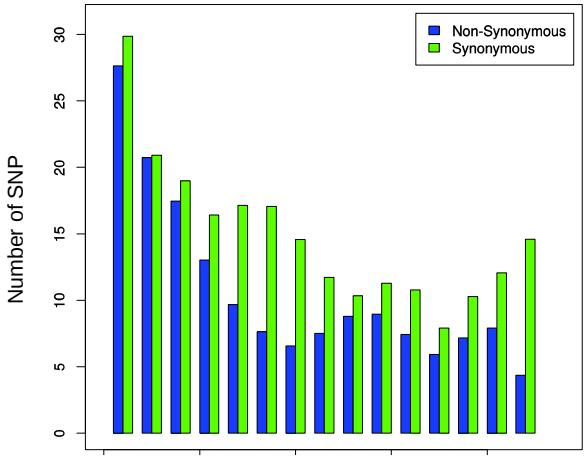


Figure S69: Site frequency spectra of *Numida meleagris* with GC-conservative mutations with a masking of CpG sites.

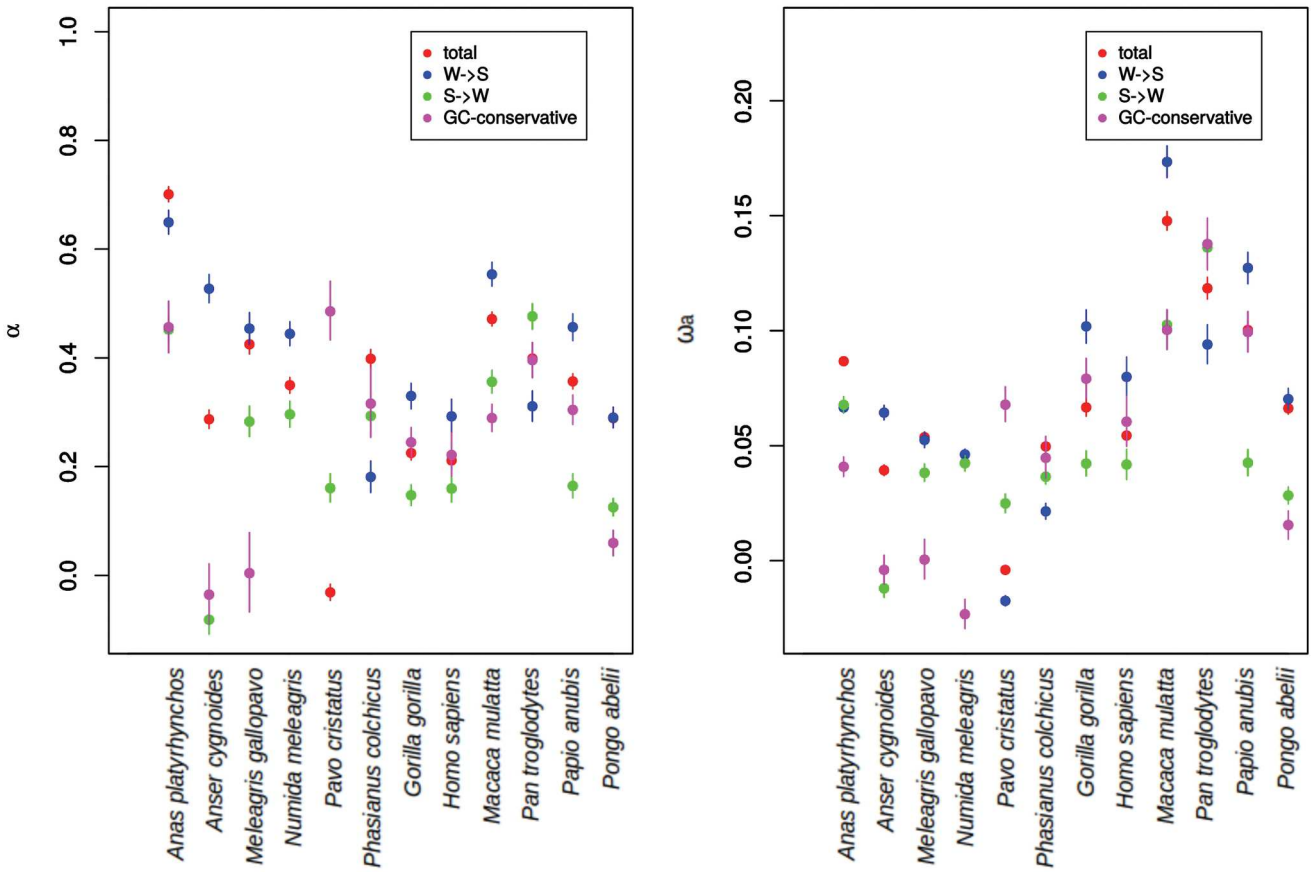


Figure S70: α and ω_a estimates for each species and each type of mutations (all mutations, W→S, S→W and GC-conservative) using all genes.

Statistics are obtained using the model "GammaExpo".

Annexe 3
Supplementary Material of Chapter 3

Supplementaty Material

Species	Bioproject	data_type	Number of individuals	Reference	Reference genome
Gorilla gorilla	PRJNA189439	Genome	20	Prado-Martinez et al. 2013	Ensembl (release 89)
Homo sapiens	PRJEB8350	Exome	19	Teixeira et al. 2015	Ensembl (release 89)
Pan troglodytes	PRJEB8350	Exome	20	Teixeira et al. 2015	Ensembl (release 89)
Papio anubis	PRJNA54005	Genome	5	unpublished baboon genome project	Ensembl (release 89)
Pongo abelii	PRJNA189439 and PRJEB1675	Genome	10	Prado-Martinez et al. 2013	Ensembl (release 89)
Macaca mulatta	PRJNA251548	Exome	20	Xue et al. 2016	Ensembl (release 89)
Meleagris gallopavo	PRJNA271731	RNA-seq	10	Wright et al. 2015	NA
Phasianus colchicus	PRJNA271731	RNA-seq	10	Wright et al. 2015	NA
Pavo cristatus	PRJNA271731	RNA-seq	10	Wright et al. 2015	NA
Numida meleagris	PRJNA271731	RNA-seq	7	Wright et al. 2015	NA
Anas platyrhynchos	PRJNA271731	RNA-seq	10	Wright et al. 2015	NA
Anser cygnoides	PRJNA271731	RNA-seq	10	Wright et al. 2015	NA
Ficedula albicollis	PRJEB2984	Genome	20	Ellegren et al. 2012	NCBI FicAlb1.5
Geospiza difficilis	PRJNA263122	Genome	8	Lamichhaney et al. 2015	NCBI Geofor1.0
Parus major	PRJNA304164	Genome	20	Kim et al. 2018	//ftp.ncbi.nlm.nih.gov/genomes/all/GCF/001/522/545/GCF_001522545.2_Parus_major1.1/GCF_001522545.2_Parus_major1.1_genomic.gff.gz
Corvus sp.	PRJEB9057	Genome	10	Vijay et al. 2017	//ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/738/735/GCF_000738735.2_ASM73873v2/GCF_000738735.2_ASM73873v2_genomic.gff.gz
Taniopygia guttata	PRJEB10586	Genome	20	Singhal et al. 2016	//ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/151/805/GCF_000151805.1_Taeniopygia_guttata-3.2.4/GCF_000151805.1_Taeniopygia_guttata-3.2.4_genomic.gff.gz
Maniola jurtina	NA	target capture	20	newly generated	NA
Melanargia galathea	NA	target capture	10	newly generated	NA
Aphantopus hyperantus	NA	target capture	7	newly generated	NA
Pyronia tithonus	NA	target capture	7	newly generated	NA
Pyronia bathseba	NA	target capture	8	newly generated	NA
Formica sanguinea	NA	target capture	10	newly generated	NA
Formica cunicularia	NA	target capture	8	newly generated	NA
Formica selysi	NA	target capture	5	newly generated	NA
Formica pratensis	NA	target capture	8	newly generated	NA
Formica fusca	NA	target capture	10	newly generated	NA
Allolobophira chlorotica L1	NA	target capture	19	newly generated	NA
Allolobophira chlorotica L2	NA	target capture	8	newly generated	NA
Allolobophira chlorotica L4	NA	target capture	9	newly generated	NA
Aporrectodea icterica	NA	target capture	10	newly generated	NA
lumbricus terrestris	NA	target capture	9	newly generated	NA
Lineus lacteus	NA	target capture	9	newly generated	NA
Lineus sanguineaus	NA	target capture	9	newly generated	NA
Lineus longissimus	NA	target capture	6	newly generated	NA
Lineus ruber	NA	target capture	8	newly generated	NA
Mytilus galloprovincialis	NA	target capture	9	newly generated	NA
Mytilus californianus	NA	target capture	16	newly generated	NA
Mytilus edulis	NA	target capture	10	newly generated	NA
Mytilus trossulus	NA	target capture	10	newly generated	NA

Supplementaty Material

References :

Ellegren H, Smeds L, Burri R, Olason PI, Backström N, Kawakami T, Künstner A, Mäkinen H, Nadachowska-Brzyska K, Qvarnström A, et al. 2012. The genomic landscape of species divergence in *Ficedula* flycatchers. Nature 491:756–760.

Kim J-M, Santure AW, Barton HJ, Quinn JL, Cole EF, Great Tit HapMap Consortium, Visser ME, Sheldon BC, Groenen MAM, van Oers K, et al. 2018. A high-density SNP chip for genotyping great tit (*Parus major*) populations and its application to studying the genetic architecture of exploration behaviour. Molecular Ecology Resources 18:877–891.

Lamichhaney S, Berglund J, Almén MS, Maqbool K, Grabherr M, Martinez-Barrio A, Promerová M, Rubin C-J, Wang C, Zamani N, et al. 2015. Evolution of Darwin's finches and their beaks revealed by genome sequencing. Nature 518:371–375.

Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, Veeramah KR, Woerner AE, O'Connor TD, Santpere G. 2013. Great ape genetic diversity and population history. Nature 499:471.

Singhal S, Leffler EM, Sannareddy K, Turner I, Venn O, Hooper DM, Strand AI, Li Q, Raney B, Balakrishnan CN. 2015. Stable recombination hotspots in birds. Science 350:928–932.

Teixeira JC, de Filippo C, Weihmann A, Meneu JR, Racimo F, Dannemann M, Nickel B, Fischer A, Halbwax M, Andre C. 2015. Long-term balancing selection in LAD1 maintains a missense trans-species polymorphism in humans, chimpanzees, and bonobos. Molecular Biology and Evolution 32:1186–1196.

Vijay N, Weissensteiner M, Burri R, Kawakami T, Ellegren H, Wolf JBW. 2017. Genome-wide signatures of genetic variation within and between populations - a comparative perspective. Available from: http://biorxiv.org/lookup/doi/10.1101/104604

Wright AE, Harrison PW, Zimmer F, Montgomery SH, Pointer MA, Mank JE. 2015. Variation in promiscuity and sexual selection drives avian rate of Faster-Z evolution. Molecular ecology 24:1218–1235.

Table S1 : Details of the species used in this study and numbers of individuals for each species.

Supplementaty Material

taxonomic group	tree topologies references
Catarrhine primates	P. Perelman *et al.*, A Molecular Phylogeny of Living Primates. *PLOS Genetics* **7**, e1001342 (2011).
Galloanserae	A. E. Wright *et al.*, Variation in promiscuity and sexual selection drives avian rate of Faster-Z evolution. *Molecular Ecology* **24**, 1218-1235 (2015).
Passeriformes	Barker FK, Barrowclough GF, Groth JG. 2002. A phylogenetic hypothesis for passerine birds: taxonomic and biogeographic implications of an analysis of nuclear DNA sequence data. Proceedings of the Royal Society B: Biological Sciences 269:295–308.
Mussels	Distel DL. 2000. Phylogenetic Relationships among Mytilidae (Bivalvia): 18S rRNA Data Suggest Convergence in Mytilid Body Plans. Molecular Phylogenetics and Evolution 15:25–33.
Satyrinae butterflies	C. Peña *et al.*, Higher level phylogeny of Satyrinae butterflies (Lepidoptera: Nymphalidae) based on DNA sequence data. *Molecular Phylogenetics and Evolution* **40**, 29-49 (2006).
Formica ants	J. Romiguier, J. Rolland, C. Morandin, L. Keller, Phylogenomics of palearctic Formica species suggests a single origin of temporary parasitism and gives insights to the evolutionary pathway toward slave-making behaviour. *BMC evolutionary biology* **18**, 40 (2018).
Earth worms	R. A. King, A. L. Tibble, W. O. C. Symondson, Opening a can of worms: unprecedented sympatric cryptic diversity within British lumbricid earthworms. *Molecular Ecology* **17**, 4684-4698 (2008).
Nemertea	Thollesson, Mikael, and Jon L. Norenburg. "Ribbon worm relationships: a phylogeny of the phylum Nemertea." Proceedings of the Royal Society of London B: Biological Sciences 270.1513 (2003): 407-415.

Table S2 : Sources of the tree topologies of each taxonomic group used to estimate branch length and map substitutions.

Supplementaty Material

Species	Propagule size (cm)	source
Formica fusca	14	Forel 1890
Formica sanguinea	10	Forel 1909
Formica cunicularia	8.5	Collingwood 1979
Formica pratensis	10.4	Bolton 1995
Formica selysi	9.5	Seifert 2002
Melanargia galathea	0.102	García-Barros 2000
Maniola jurtina	0.0535	García-Barros 2000
Aphantopus hyperantus	0.0792	García-Barros 2000
Pyronia tithonus	0.0628	García-Barros 2000
Pyronia bathseba	0.0802	García-Barros 2000
Mytilus californianus	0.01	Bayne et al. 1983
Mytilus trossulus	0.01	Bayne et al. 1983
Mytilus galloprovincialis	0.01	Bayne et al. 1983
Mytilus edulis	0.01	Bayne et al. 1983
Allolobophira chlorotica L1	0.0238	Eijsackers 2011
Allolobophira chlorotica L2	0.0238	Eijsackers 2011
Allolobophira chlorotica L4	0.0238	Eijsackers 2011
Aporrecta icterica	0.0411	Eijsackers 2011
Lumbricus terrestris	0.5	Cloudsley-Thompson and Sankey 1961
Lineus lacteus	0.02	Bierne, 1983
Lineus longissimus	0.02	Bierne, 1983
Lineus sanguineaus	0.02	Bierne, 1983
Lineus ruber	0.5	Bierne, 1983
Homo sapiens	93.4	De Magalhaes and Costa 2009
Pan troglodytes	45.12	De Magalhaes and Costa 2009
Gorilla gorilla	78	De Magalhaes and Costa 2009
Papio anubis	55.95	De Magalhaes and Costa 2009
Pongo abelii	42.48	De Magalhaes and Costa 2009
Macaca mulatta	31.1	De Magalhaes and Costa 2009
Anas platyrhynchos	55	Del Hoyo et al. 1992
Anser cygnoides	87	Del Hoyo et al. 1992
Meleagris gallopavo	90	Del Hoyo et al. 1992
Numida meleagris	53	Del Hoyo et al. 1992
Pavo cristatus	95	Del Hoyo et al. 1992

Phasianus colchicus	57.5	Del Hoyo et al. 1992
Parus major	13.5	Del Hoyo et al. 1992
Ficedula albicollis	13	Del Hoyo et al. 1992
Corvus sp.	50.5	Del Hoyo et al. 1992
Geospiza difficilis	11.5	Del Hoyo et al. 1992
Taniopygia guttata	10	Del Hoyo et al. 1992
Rattus norvegicus	13.717	De Magalhaes and Costa 2009
Microtus arvalis	7.184	De Magalhaes and Costa 2009
Microtus ochrogaster	10.6	De Magalhaes and Costa 2009
Mus musculus musculus	6.06	De Magalhaes and Costa 2009
Mus spretus	6.37	Inferred from De Magalhaes and Costa 2009

Species	**adult size (cm)**	**source**
Formica fusca	14	Forel 1890
Formica sanguinea	10	Forel 1909
Formica cunicularia	8.5	Collingwood 1979
Formica pratensis	10.4	Bolton 1995
Formica selysi	9.5	Seifert 2002
Melanargia galathea	2.59	García-Barros 2000
Maniola jurtina	2.58	García-Barros 2000
Aphantopus hyperantus	2.12	García-Barros 2000
Pyronia tithonus	1.89	García-Barros 2000
Pyronia bathseba	2.04	García-Barros 2000
Mytilus californianus	7.5	MArine Life Information Network, 2006
Mytilus trossulus	7.5	MArine Life Information Network, 2006
Mytilus galloprovincialis	7.5	MArine Life Information Network, 2006
Mytilus edulis	7.5	MArine Life Information Network, 2006
Allolobophira chlorotica L1	5.5	The Trustees of the Natural History Museum,2010
Allolobophira chlorotica L2	5.5	The Trustees of the Natural History Museum,2010

Supplementaty Material

Allolobophira chlorotica L4	5.5	The Trustees of the Natural History Museum,2010
Aporrecta icterica	9.5	Sims and Gerard 1985
Lumbricus terrestris	25	Cloudsley-Thompson and Sankey 1961
Lineus lacteus	17.5	Gontcharoff 1951
Lineus longissimus	1000	Gontcharoff 1951
Lineus sanguineus	NA	NA
Lineus ruber	5	Gontcharoff 1951, Bierne 1970
Homo sapiens	163	Ogden et al. 2004
Pan troglodytes	79.6	Jones et al. 2009
Gorilla gorilla	137.5	Wood 1979
Papio anubis	85	Fleagle 2013
Pongo abelii	83	Groves 1971
Macaca mulatta	55.5	Jones et al. 2009
Anas platyrhynchos	55	Del Hoyo et al. 1992
Anser cygnoides	87	Del Hoyo et al. 1992
Meleagris gallopavo	90	Del Hoyo et al. 1992
Numida meleagris	53	Del Hoyo et al. 1992
Pavo cristatus	95	Del Hoyo et al. 1992
Phasianus colchicus	57.5	Del Hoyo et al. 1992
Parus major	13.5	Del Hoyo et al. 1992
Ficedula albicollis	13	Del Hoyo et al. 1992
Corvus sp.	50.5	Del Hoyo et al. 1992
Geospiza difficilis	11.5	Del Hoyo et al. 1992
Taniopygia guttata	10	Del Hoyo et al. 1992
Rattus norvegicus	21.5	Burton and Burton 2002
Microtus arvalis	11.1	Jones et al. 2009
Microtus ochrogaster	15.2	Jones et al. 2009
Mus musculus musculus	8	Berry 1970
Mus spretus	8.6	Palomo et al. 2009

Species	**body mass (g)**	**source**
Formica fusca	NA	NA
Formica sanguinea	NA	NA
Formica cunicularia	NA	NA

Formica pratensis	0.0119	Keller & Passera, 1989
Formica selysi	NA	NA
Melanargia galathea	NA	NA
Maniola jurtina	0.05	Svärd &Wiklund,1989
Aphantopus hyperantus	0.0376	Svärd &Wiklund,1989
Pyronia tithonus	0,04	Corbet, 2000
Pyronia bathseba	NA	NA
Mytilus californianus	37.5	MArine Life Information Network, 2006
Mytilus trossulus	37.5	MArine Life Information Network 2006
Mytilus galloprovincialis	37.5	MArine Life Information Network, 2006
Mytilus edulis	37.5	MArine Life Information Network, 2006
Allolobophira chlorotica L1	0.3	Butt 1997
Allolobophira chlorotica L2	0.3	Butt 1997
Allolobophira chlorotica L4	0.3	Butt 1997
Aporrecta icterica	0.95	Bouché 1972
Lumbricus terrestris	7.5	Quillin, 1999
Lineus lacteus	NA	NA
Lineus longissimus	NA	NA
Lineus sanguineaus	NA	NA
Lineus ruber	NA	NA
Homo sapiens	62000	De Magalhaes and Costa 2009
Pan troglodytes	45000	De Magalhaes and Costa 2009
Gorilla gorilla	93000	De Magalhaes and Costa 2009
Papio anubis	14700	De Magalhaes and Costa 2009
Pongo abelii	45000	De Magalhaes and Costa 2009
Macaca mulatta	8240	De Magalhaes and Costa 2009
Anas platyrhynchos	1027	De Magalhaes and Costa 2009
Anser cygnoides	3150	De Magalhaes and Costa 2009
Meleagris gallopavo	4000	De Magalhaes and Costa 2009
Numida meleagris	1479	De Magalhaes and Costa 2009
Pavo cristatus	3375	De Magalhaes and Costa 2009
Phasianus colchicus	999	De Magalhaes and Costa 2009
Parus major	17	Del Hoyo et al. 1992

Ficedula albicollis	12	Del Hoyo et al. 1992
Corvus sp.	499	Del Hoyo et al. 1992
Geospiza difficilis	16.15	Del Hoyo et al. 1992
Taniopygia guttata	10	Del Hoyo et al. 1992
Rattus norvegicus	320	De Magalhaes and Costa 2009
Microtus arvalis	27.5	De Magalhaes and Costa 2009
Microtus ochrogaster	50	De Magalhaes and Costa 2009
Mus musculus musculus	20.5	De Magalhaes and Costa 2009
Mus spretus	17	Palomo et al. 2009

Species	**Fecundity (number of offspring per year)**	**source**
Formica fusca	NA	NA
Formica sanguinea	NA	NA
Formica cunicularia	NA	NA
Formica pratensis	NA	NA
Formica selysi	NA	NA
Melanargia galathea	NA	NA
Maniola jurtina	NA	NA
Aphantopus hyperantus	140	Lafranchis et al. 2015
Pyronia tithonus	125	Lafranchis et al. 2015
Pyronia bathseba	NA	NA
Mytilus californianus	110000	MArine Life Information Network, 2006
Mytilus trossulus	110000	MArine Life Information Network, 2006
Mytilus galloprovincialis	110000	MArine Life Information Network, 2006
Mytilus edulis	110000	MArine Life Information Network, 2006
Allolobophira chlorotica L1	0.74	Edwards & Bohlen 1996
Allolobophira chlorotica L2	0.74	Edwards & Bohlen 1996
Allolobophira chlorotica L4	0.74	Edwards & Bohlen 1996
Aporrecta icterica	2.67	Booth et al. 2000
Lumbricus terrestris	NA	NA

Lineus lacteus	NA	NA
Lineus longissimus	NA	NA
Lineus sanguineaus	NA	NA
Lineus ruber	NA	NA
Homo sapiens	0.0008219178	De Magalhaes and Costa 2009
Pan troglodytes	0.0005479452	De Magalhaes and Costa 2009
Gorilla gorilla	0.0008219178	De Magalhaes and Costa 2009
Papio anubis	0.002191781	De Magalhaes and Costa 2009
Pongo abelii	0.0005479452	De Magalhaes and Costa 2009
Macaca mulatta	0.002739726	De Magalhaes and Costa 2009
Anas platyrhynchos	0.02465753	De Magalhaes and Costa 2009
Anser cygnoides	NA	NA
Meleagris gallopavo	0.03013699	De Magalhaes and Costa 2009
Numida meleagris	0.02465753	De Magalhaes and Costa 2009
Pavo cristatus	0.01369863	De Magalhaes and Costa 2009
Phasianus colchicus	0.03013699	De Magalhaes and Costa 2009
Parus major	0.0205	Tomás et al. 2012
Ficedula albicollis	0.0178	Gill and Donsker 2017
Corvus sp.	0.01068	Holyoak 1967
Geospiza difficilis	0.0329	Grant and Grant 1980
Taniopygia guttata	0.0151	Olson et al. 2014
Rattus norvegicus	0.1003562	De Magalhaes and Costa 2009
Microtus arvalis	0.0768	De Magalhaes and Costa 2009
Microtus ochrogaster	0.04164384	De Magalhaes and Costa 2009
Mus musculus musculus	0.104	De Magalhaes and Costa 2009
Mus spretus	NA	NA

Species	**Longevity (years)**	**source**
Formica fusca	20	Personnal communication
Formica sanguinea	20	Personnal communication
Formica cunicularia	20	Personnal communication
Formica pratensis	6	Personnal communication
Formica selysi	20	Personnal communication
Melanargia galathea	1	Lafranchis et al. 2015
Maniola jurtina	1	Lafranchis et al. 2015

Supplementaty Material

Aphantopus hyperantus	1	Lafranchis et al. 2015
Pyronia tithonus	1	Lafranchis et al. 2015
Pyronia bathseba	NA	NA
Mytilus californianus	25	Bayne and Bayne 1976
Mytilus trossulus	25	Bayne and Bayne 1976
Mytilus galloprovincialis	25	Bayne and Bayne 1976
Mytilus edulis	25	Bayne and Bayne 1976
Allolobophira chlorotica L1	1.25	Edwards & Bohlen 1996
Allolobophira chlorotica L2	1.25	Edwards & Bohlen 1996
Allolobophira chlorotica L4	1.25	Edwards & Bohlen 1996
Aporrecta icterica	NA	NA
Lumbricus terrestris	NA	NA
Lineus lacteus	NA	NA
Lineus longissimus	NA	NA
Lineus sanguineaus	NA	NA
Lineus ruber	NA	NA
Homo sapiens	123	De Magalhaes and Costa 2009
Pan troglodytes	59.4	De Magalhaes and Costa 2009
Gorilla gorilla	60.1	De Magalhaes and Costa 2009
Papio anubis	37.5	De Magalhaes and Costa 2009
Pongo abelii	59	De Magalhaes and Costa 2009
Macaca mulatta	40	De Magalhaes and Costa 2009
Anas platyrhynchos	29.1	De Magalhaes and Costa 2009
Anser cygnoides	31	De Magalhaes and Costa 2009
Meleagris gallopavo	13	De Magalhaes and Costa 2009
Numida meleagris	NA	NA
Pavo cristatus	23.2	De Magalhaes and Costa 2009
Phasianus colchicus	27	De Magalhaes and Costa 2009
Parus major	15.4	De Magalhaes and Costa 2009
Ficedula albicollis	9.8	De Magalhaes and Costa 2009
Corvus sp.	19.2	De Magalhaes and Costa 2009
Geospiza difficilis	9	Oschadleus et al. 2016
Taniopygia guttata	12	De Magalhaes and Costa 2009

Rattus norvegicus	3.8	De Magalhaes and Costa 2009
Microtus arvalis	4.8	De Magalhaes and Costa 2009
Microtus ochrogaster	5.3	De Magalhaes and Costa 2009
Mus musculus musculus	4	De Magalhaes and Costa 2009
Mus spretus	NA	NA

References :

Bayne BL, Salkeld PN, Worrall CM. 1983. Reproductive effort and value in different populations of the marine mussel, *Mytilus edulis* L. Oecologia 59:18–26.

Bayne Brian Leicester, Bayne Brian L. 1976. Marine mussels: their ecology and physiology. Cambridge University Press.

Berry RJ. 1970. The natural history of the house mouse. Field studies 3:219–262.

Bierne J. 1970. Recherches sur la différenciation sexuelle au cours de l'ontogenèse et de la régénération chez le némertien *Lineus ruber*. Compte rendu de l'Académie des sciences de Paris 259:4841-4843.

Bierne J (1983) Oogenesis Oviposition Oosorption. Reproductive Biology of Invertebrates 146–167.

Bolton B. 1995. A new general catalogue of the ants of the world. Harvard University Press.

Booth LH, Heppelthwaite VJ, O'halloran K. 2000. Growth, development and fecundity of the earthworm *Aporrectodea caliginosa* after exposure to two organophosphates.New Zealand Plant Protection 53:221–225.

Bouché MB. 1972. Lombriciens de France: écologie et systématique. Quae.

Burton M, Burton R. 2002. International Wildlife Encyclopedia.

Butt KR. 1997. Reproduction and growth of the earthworm *Allolobophora chlorotica* (Savingy, 1826) in controlled environments. Pedobiologia 41:369–374.

Cloudsley-Thompson JL, Sankey J. 1961. Land invertebrates. A guide to British worms, molluscs and arthropods (excluding insects). Methuen & Co.

Collingwood CA. 1979. The Formicidae (Hymenoptera) of Fennoscandia and Denmark. Scandinavian Science Press.

Corbet SA. 2000. Butterfly nectaring flowers: butterfly morphology and flower form. Entomologia Experimentalis et Applicata, 96:289-298.

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

Del Hoyo J, Elliot A, Sargatal J. 1992. Handbook of the Birds of the World. Barcelona. Lynx Editions.

Edwards CA, Bohlen P J. 1996. Biology and ecology of earthworms. Chapman and Hall.

Eijsackers H. 2011. Earthworms as colonizers of natural and cultivated soil environments. Applied Soil Ecology 50:1–13.

Fleagle JG. 2013. Primate adaptation and evolution. Academic Press.

Forel A. 1890. Fourmis de Tunisie et de l'Algérie orientale. Antbase.org.

Supplementaty Material

Forel A. 1909. Fourmis d'Espagne. Antbase.org.

García-Barros E. 2000. Egg size in butterflies (Lepidoptera: Papilionoidea and Hesperiidae): a summary of data. Journal of Research on the Lepidoptera 35:90–136.

Gill F, Donsker D. 2017. IOC World Bird List (v 7.2), 10.14344/IOC. ML

Gontcharoff M. 1951. Biologie de la régénération et de la reproduction chez quelques Lineidae de France. Annales des Sciences Naturelles, Zoologie 13:149–235.

Grant PR, Grant BR. 1980. The breeding and feeding characteristics of Darwin's finches on Isla Genovesa, Galapagos. Ecological Monographs 50:381–410.

Groves CP. 1971. *Pongo pygmaeus*. Mammalian species:1–6.

Holyoak D. 1967. Breeding biology of the Corvidae. Bird Study 14:153–168.

Jones KE, Bielby J, Cardillo M, Fritz SA, O'Dell J, Orme CDL, Safi K, Sechrest W, Boakes EH, Carbone C. 2009. PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals. Ecology 90:2648–2648.

Keller, L. and Passera, L., 1989. Size and fat content of gynes in relation to the mode of colony founding in ants (Hymenoptera; Formicidae). Oecologia 802 :236-240.

Lafranchis, T., Jutzeler, D., Guillosson, J.Y., Kan, P. and Kan, B., 2015. La vie des papillons: écologie, biologie et comportement des Rhopalocères de France. Diatheo.

MArine Life Information Network. MarLIN BIOTIC (Biological Traits Information Catalogue).(2006). at <www.marlin.ac.uk/biotic/>

Ogden CL, Fryar CD, Carroll MD, Flegal KM. 2004. Mean body weight, height, and body mass index: United States 1960-2002. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics Washington, DC

Olson CR, Wirthlin M, Lovell PV, Mello CV. 2014. Proper care, husbandry, and breeding guidelines for the zebra finch, *Taeniopygia guttata*. Cold Spring Harbor Protocols 2014.

Oschadleus HD, Schultz B and Schultz SJ. 2016. Longevity of the Helmeted Guineafowl *Numida meleagris*. Biodiversity Observations,1-3.

Palomo LJ, Justo ER, Vargas JM. 2009. *Mus spretus* (Rodentia: muridae). Mammalian species:1–10.

Quillin KJ. 1999. Kinematic scaling of locomotion by hydrostatic animals: ontogeny of peristaltic crawling by the earthworm *Lumbricus terrestris*. Journal of Experimental Biology, 202 :661-674.

Seifert B. 2002. A taxonomic revision of the Formica cinerea group (Hymenoptera: Formicidae). Abhandlungen und Berichte des Naturkundemuseums Görlitz 74:245–272.

Sims RW, Gerard BM. 1985. Earthworms: keys and notes for the identification and study of the species. Brill Archive.

Svärd L, and C Wiklund. 1989. Mass and production rate of ejaculates in relation to monandry/polyandry in butterflies. Behavioral Ecology and Sociobiology 6: 395-402.

The Trustees of the Natural History Museum. Available from: http://www.nhm.ac.uk/discover.html

Tomás G, Barba E, Merino S, Martínez J. 2012. Clutch size and egg volume in great tits (*Parus major*) increase under low intensity electromagnetic fields: a long-term field study. Environmental research 118:40–46.

Supplementaty Material

Wood BA. 1979. Relationship between body size and long bone lengths in *Pan* and *Gorilla*. American journal of physical anthropology 50:23–25.

Table S3: Values and sources of the life history traits used in this study.

Supplementaty Material

species	number_of_indivduals	chosen_sample_size	# non-synonymous_SNPs	# synonymous_SNPs	# SNPs total
F. fusca	10	14	4277.651	6578.441	10856.092
F. sanguinea	10	16	3242.012	5343.09	8585.102
F. cunicularia	8	12	4035.038	6354.631	10389.669
F. pratensis	8	12	1773.432	2234.555	4007.987
F. selysi	5	6	346.7999	286.398	633.1979
M. galathea	10	10	1189.648	3309.359	4499.007
M. jurtina	20	32	7309.744	15925.68	23235.424
A. hyperanthus	7	8	1441.038	2384.66	3825.698
P. tithonus	7	10	1534.318	2372.112	3906.43
P. bathseba	8	10	1610.804	2676.396	4287.2
M. californianus	16	24	6370.325	13435.74	19806.065
M. trossulus	10	14	6329.283	18993.88	25323.163
M. galloprovincialis	9	12	3297.619	9089.448	12387.067
M. edulis	10	12	5497.405	15987.1	21484.505
A. chlorotica L1	19	26	1750.876	3093.717	4844.593
A. chlorotica L2	8	8	350.0329	553.9886	904.0215
A. chlorotica L4	9	12	3562.078	7895.187	11457.265
A. icterica	10	12	1657.449	3777.663	5435.112
L. terrestris	9	8	237.7491	939.8352	1177.5843
L. lacteus	9	12	5895.8	20421.24	26317.04
L. longissimus	6	6	54.49608	99.30191	153.79799
L. sanguineus	9	10	954	3248	4202
L. ruber	8	12	1027.498	1690.69	2718.188
H. sapeins	19	28	2280.499	2812.022	5092.521
P. troglodytes	20	30	4743.888	6558.18	11302.068
G. gorilla	20	30	3649.085	4842.278	8491.363
P. anubis	5	8	2037.488	4006.023	6043.511
P. abelii	10	16	5686.643	9299.841	14986.484
M. mulatta	19	28	5256.561	9975.758	15232.319
A. pltyrhynchos	10	16	3795.669	13088.68	16884.349
A. cygnoides	10	14	1295.556	4823.812	6119.368
M. gallopavo	10	10	1004.586	3790.763	4795.349
N. meleagris	10	16	1318.596	6241.347	7559.943
P. cristatus	10	14	601.7848	2213.924	2815.7088
P. colchicus	11	14	3051.282	9702.466	12753.748
P. major	20	16	6905.476	14414.42	21319.896
F. albicolis	20	16	16242.07	20680.76	36922.83
Corvus sp.	10	14	817.4107	1354.51	2171.9207
G. difficilis	8	10	2090.324	3280.025	5370.349
T. guttata	20	16	36317.97	102248.5	138566.47

Table S4: SNPs counts for each species.

SNPs counts are not integers because they corresponds to SNPs that are present in our SFS, where we chose a sample size (i.e. the number of categories of the SFS) lower that 2*the number of individuals. This is to compensate the uneven coverage between individuals that results in some sites in some individuals not to be genotyped, which prevents the assignment of a SNP at one locus to a frequency class if there are more frequency classes than there is of genotyped position at this locus. We chose sample sizes that maximize the number of SNPs in each SFS.

Supplementaty Material

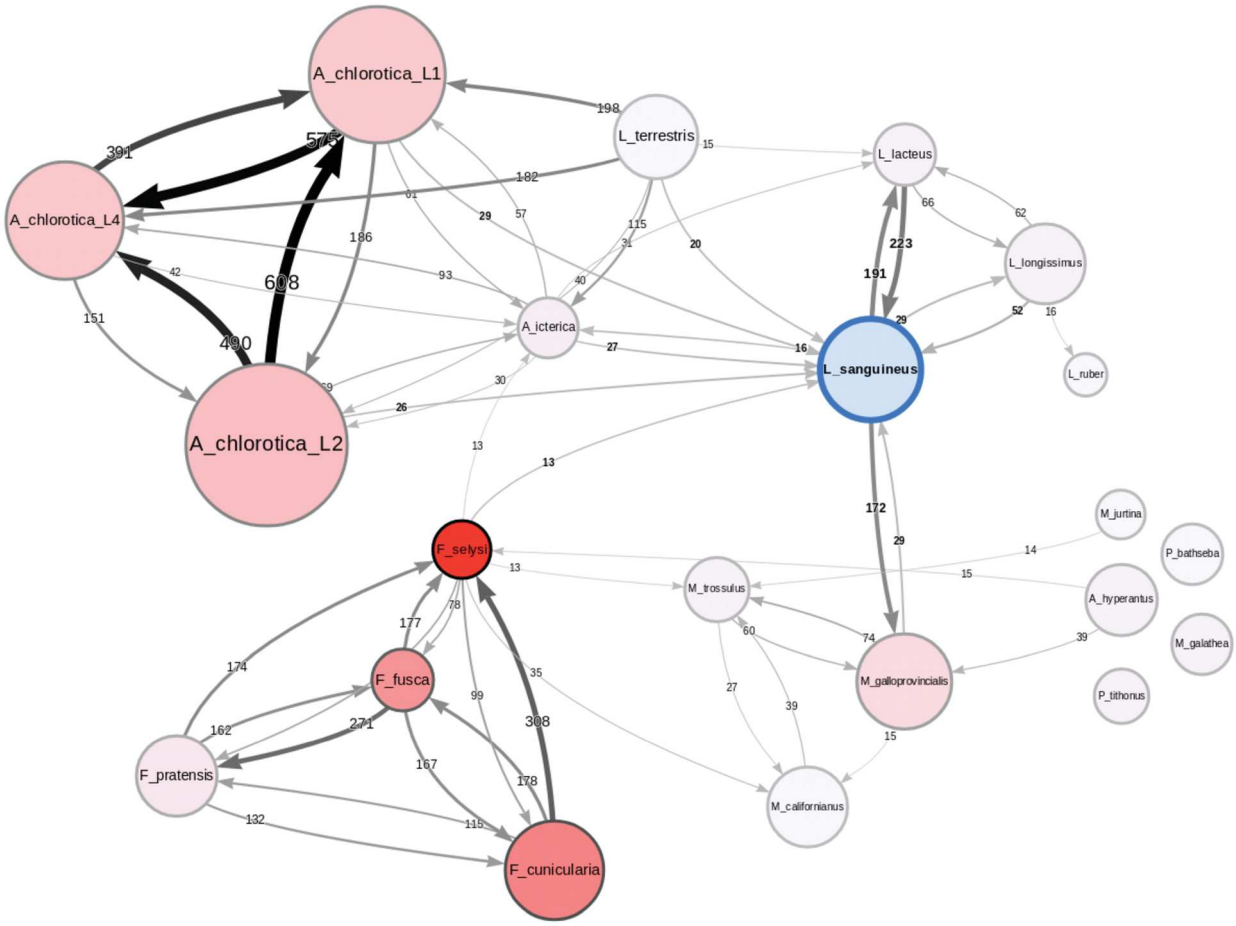


Figure S1: Network of contaminants in the de novo assembly generated after exon capture (Croco).

Supplementaty Material

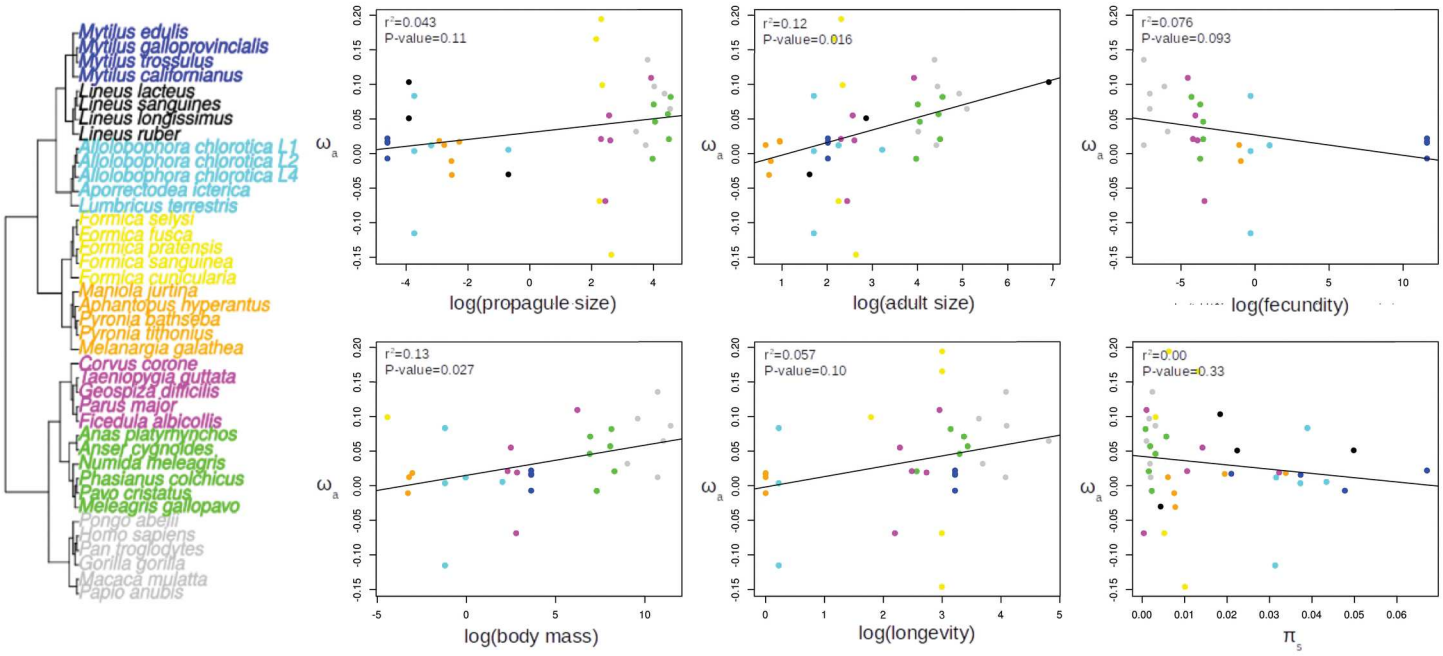


Figure S2: Relationship between ω_a estimated for each species using GC-conservative mutations and life history traits and π_s.

Annexe 4

Hemizygosity enhances purifying selection: lack of fast-Z evolution in two Satyrine butterflies

Genome Biology and Evolution

SMBE

Hemizygosity Enhances Purifying Selection: Lack of Fast-Z Evolution in Two Satyrine Butterflies

Marjolaine Rousselle*, Nicolas Faivre, Marion Ballenghien, Nicolas Galtier, and Benoit Nabholz

UMR 5554 Institut des Sciences de l'Evolution, CNRS, Université de Montpellier, IRD, EPHE, Place E. Bataillon, Montpellier, France

*Corresponding author: marjolaine.rousselle@umontpellier.fr.

Accepted: August 24, 2016

Abstract

The fixation probability of a recessive beneficial mutation is increased on the X or Z chromosome, relative to autosomes, because recessive alleles carried by X or Z are exposed to selection in the heterogametic sex. This leads to an increased dN/dS ratio on sex chromosomes relative to autosomes, a pattern called the "fast-X" or "fast-Z" effect. Besides positive selection, the strength of genetic drift and the efficacy of purifying selection, which affect the rate of molecular evolution, might differ between sex chromosomes and autosomes. Disentangling the complex effects of these distinct forces requires the genome-wide analysis of polymorphism, divergence and gene expression data in a variety of taxa. Here we study the influence of hemizygosity of the Z chromosome in *Maniola jurtina* and *Pyronia tithonus*, two species of butterflies (Lepidoptera, Nymphalidae, Satyrinae). Using transcriptome data, we compare the strength of positive and negative selection between Z and autosomes accounting for sex-specific gene expression. We show that *M. jurtina* and *P. tithonus* do not experience a faster, but rather a slightly slower evolutionary rate on the Z than on autosomes. Our analysis failed to detect a significant difference in adaptive evolutionary rate between Z and autosomes, but comparison of male-biased, unbiased and female-biased Z-linked genes revealed an increased efficacy of purifying selection against recessive deleterious mutations in female-biased Z-linked genes. This probably contributes to the lack of fast-Z evolution of satyrines. We suggest that the effect of hemizygosity on the fate of recessive deleterious mutations should be taken into account when interpreting patterns of molecular evolution in sex chromosomes vs. autosomes.

Key words: fast-Z effect, sex-chromosome evolution, transcriptomics, sex-biased expression, Nymphalidae, Lepidoptera.

Introduction

Sex chromosomes of sufficiently ancient origin are hemizygous, i.e., effectively haploid in one sex (males in X/Y systems, females in Z/W systems) and diploid in the other one. A consequence of hemizygosity is that recessive beneficial mutations occurring in sex-linked genes are immediately exposed to selection in the heterogametic sex. In contrast, in autosomes a recessive mutation is only exposed to selection at homozygous state, which occurs rarely as far as recently appeared mutations are concerned. Recessive beneficial substitutions are therefore expected to accumulate at a faster rate on the X (respectively, Z) chromosome than on autosomes due to a more efficient positive selection in males (respectively, females; Haldane 1924; Rice 1984; Charlesworth et al. 1987). This should result in an increased evolutionary rate of sex chromosomes relative to autosomes, a phenomenon called the «fast-X» («fast-Z») effect. The theoretical prediction of a faster evolution on the X has been empirically corroborated in several species of mammals (Carneiro et al. 2012; Hvilsom et al. 2011; Kousathanas et al. 2014; Veeramah et al. 2014; Nam et al. 2015) and fruit flies (Betancourt et al. 2002; Thornton and Long 2002; Counterman et al. 2004; Mackay et al. 2012; Avila et al. 2014; Campos et al. 2014; Garrigan et al. 2014). This large body of literature is globally consistent with the hypothesis that recessive adaptive mutations are sufficiently common to significantly accelerate molecular evolution on the X.

A couple of recent studies, however, suggest that the situation could be more complex than suggested above. First, Nguyen et al. (2015) analyzed the effect of gene expression level and recombination rate on X-linked genes evolution in mammals—particularly, the elevated ratio of non-synonymous, dN, to synonymous, dS, substitution rates (dN/dS). X-linked genes tend to have a lower expression level than autosomal genes (Marín et al. 2000; Julien et al. 2012), and gene expression is known to be anti-correlated to dN/dS in a

broad range of organisms (Drummond et al. 2005; Drummond and Wilke 2008), presumably because of a reduced intensity of purifying selection in low-expressed genes. Nguyen et al. (2015) suggested that this effect is sufficient to explain most of the fast-X effect in primates and murine rodents without invoking positive selection and dominance effects. Second, two studies in the female-heterogametic birds revealed the existence of a fast-Z effect, the dN/dS ratio of the Z chromosome being higher than the dN/dS ratio in autosomes, but found that Z-linked genes predominantly expressed in males (i.e., male-biased) were not less accelerated than unbiased or female-biased genes (Mank et al. 2010; Wright et al. 2015). This is not expected under the hypothesis of a fast-Z driven by recessive beneficial mutations. Being exposed to selection primarily in the homogametic sex, male-biased genes in Z/W systems should not be affected by hemizygosity (Kousathanas et al. 2014; Sackton et al. 2014). For this reason, Wright et al. (2015) did not interpret the increased dN/dS ratio of Z-linked genes as a consequence of a more efficient positive selection, but rather as an effect of a reduced effective population size of the Z chromosome. The difference in effective population size between sex chromosome (Ne_Z) and autosomes (Ne_A) in female heterogametic systems is supposed to be larger than in male heterogametic systems, due to a higher variance of reproductive success in males than in females (Mank et al. 2010). Low Ne_Z/Ne_A ratios in birds would entail a stronger intensity of genetic drift in the Z chromosome relative to autosomes, leading to a decreased efficacy of purifying selection and consequently an increased probability of fixation of slightly deleterious mutations, and consequently an increased dN/dS ratio (Wright et al. 2015). It was therefore postulated that the fast-X and the fast-Z effects may be driven by distinct evolutionary forces, the former being due to an elevated adaptive substitution rate on the X, and the latter to an elevated rate of slightly deleterious substitutions on the Z (Mank et al. 2010; Wright et al. 2015). However, this interesting hypothesis was challenged by a recent study suggesting that Z chromosomes in silkmoths (*Bombyx*) experience an accelerated evolutionary rate due to more efficient positive selection (Sackton et al. 2014), similarly to the pattern reported on the X in mammals and *Drosophila*.

The literature therefore suggests that positive selection in favor of recessive beneficial mutations is not always the predominant factor responsible for differences in evolutionary rates between X/Z chromosomes and autosomes. Purifying selection, genetic drift and gene expression level are important players too, and their respective impacts are difficult to disentangle. Whether the fast-X and fast-Z effects have similar or distinct evolutionary causes, for instance, is currently unclear. It is noteworthy that hemizygosity is predicted to improve the efficacy not only of positive selection, but also of purifying selection against recessive deleterious mutations (Charlesworth et al. 1987). The effect of homozygosity on the fate of deleterious mutations has been theoretically and empirically investigated in the context of comparisons between selfing and outcrossing species (Glémin 2007, Szövényi et al. 2014), but has so far received limited consideration from the empirical literature on sex chromosome evolution (but see Veeramah et al. 2014 and Charlesworth 2012). Only one study reported a greater codon usage bias on the X chromosome relative to autosomes in *Drosophila melanogaster* and *Caenorhabditis elegans*, suggesting that purifying selection has greater efficacy on the X than on autosomes (Singh et al. 2005).

In an attempt to address these issues, we analyzed sex-linked vs. autosomal polymorphism, divergence and gene expression pattern in two closely related species of Satyrinae butterflies, the Meadow Brown (*Maniola jurtina*) and the Gatekeeper (*Pyronia tithonus*). Butterflies are Z/W species typically carrying a higher level of genetic polymorphism than birds (Romiguier et al. 2014), thus offering a good opportunity to discriminate between purifying and positive selection through McDonald–Kreitman (McDonald and Kreitman 1991) related approaches. *M. jurtina* and *P. tithonus* belong to the same family (Nymphalidae) as the fully sequenced Glanville fritillary (*Melitaea cinxia*), in which a high-resolution linkage map is available (Ahola et al. 2014).

Surprisingly, we report in these species not a fast-Z, but rather a slight slow-Z effect. The analysis of sex-biased genes allowed us to demonstrate that purifying selection against recessive deleterious mutations is more efficient in the Z chromosome, presumably as a consequence of hemizygosity. Moreover, although positive selection is slightly enhanced in Z-linked female-biased genes, we did not detect an overall higher adaptive substitution rate in the Z chromosome than in autosomes. Taken together, these results explain the lack of fast-Z in Satyrinae butterflies, and reveal the complexity of the interactions between the evolutionary forces affecting substitution rates in sex chromosomes vs. autosomes.

Materials and Methods

De Novo Transcriptome Assembly

RNA-Seq data were obtained from ten living adult individuals of each of *M. jurtina* and *P. tithonus* sampled from their natural habitat in several locations spread across France, Germany, Spain, and Portugal (supplementary table S1, Supplementary Material online). Wings and genitalia have been conserved for sexing. The rest of the body was used to extract total RNA using standardized protocols (Gayral et al. 2011). Complementary DNA was sequenced on Illumina HiSeq 2000. We assessed the quality of the reads with FastQC v0.10.1 (www.bioinformatics.babraham.ac.uk/projects/fastqc) and we removed reads containing adapters as well as reads shorter than 50 bp. We cut the end of reads in a way that ensures that the mean per base sequence quality

was above 30. We constructed *de novo* transcriptome assemblies for each species following strategies B in Cahais et al. (2012), using a pipeline involving Abyss (Simpson et al. 2009) and Cap3 (Huang and Madan 1999). Reads were mapped to predicted cDNAs (contigs) with the BWA (Burrow Wheeler Aligner) program (Li and Durbin 2009). Contigs with a per-individual average coverage below 2.5X were discarded. Open reading frames (ORFs) were predicted using the Trinity package (Grabherr et al. 2011). Contigs carrying ORF shorter than 150 bp were discarded.

Orthology and Sex-Linkage Prediction

We downloaded annotated coding sequences of the Glanville fritillary (*M. cinxia*) using biomart (www.biomart.org, January 2015). To predict orthology between *M. jurtina* and *P. tithonus*, respectively, and the reference *M. cinxia*, we performed BLASTn searches. Only hits with an e-value below 10^{-20} were considered. When one ORF from *M. jurtina* or *P. tithonus* hit several reference sequences in *M. cinxia*, we selected the couple presenting the highest score. When several ORFs from *M. jurtina* or *P. tithonus* hit the same reference sequence, we checked whether hits were overlapping, in which case we only kept the one with the highest score. Otherwise, we kept them all. Eventually, we verified those predictions of orthology using another reference species of the Nymphalidae family, *Heliconius melpomene* (The Heliconius Genome Consortium 2012). Sequences that hit the same gene in *M. cinxia* but distinct genes in *H. melpomene* were excluded.

Z-linked/autosomal annotations of genes in *M. cinxia* (Ahola et al. 2014, http://www.helsinki.fi/science/metapop/research/mcgenome2_downloads.html; last accessed February 2015) were propagated to their predicted orthologs in *M. jurtina* and *P. tithonus*. We removed ORFs predicted to be Z-linked based on *M. cinxia* annotations but presenting heterozygous sites in females of *M. jurtina* or *P. tithonus* in the main analysis.

Divergence and Polymorphism Analyses

Orthologous *M. jurtina* and *P. tithonus* coding sequences were aligned using MACSE [Multiple Alignment of Coding SEquences accounting for frameshifts and stop codons (Ranwez et al. 2011)] and the alignments were cleaned for gaps (by removing all sites that presented a gap in at least one species). We used the programs bppml and mapNH [http://biopp.univ-montp2.fr/forge/testnh (Romiguier et al. 2012; Guéguen et al. 2013)] to estimate the synonymous and non-synonymous substitution rate (dS, the number of synonymous substitutions per synonymous site and dN, the number of non-synonymous substitutions per non-synonymous site) by substitution mapping under the Nielsen–Yang model (Nielsen and Yang 1998). *M. cinxia* and *H. melpomene* coding sequences were too distant from the Satyrinae to allow proper estimation of dS—they were not analyzed. For any set of genes (Z-linked, autosomal, sex-biased), the number of synonymous and non-synonymous substitutions and the number of synonymous and non-synonymous sites were summed across genes and dN/dS was calculated by taking the ratio of the sums. Ninety-five percent confidence intervals were determined by bootstrapping genes (1000 replicates).

For each individual, diploid genotypes were called according to the method described in Tsagkogeorga et al. 2012 (model M1). This method estimates the sequencing error rate in the maximum likelihood framework and calculates the posterior probability of each possible genotype. Genotypes supported by a posterior probability higher than 95% are retained, otherwise missing data is called. Polymorphic positions were filtered for possible hidden paralogues using a likelihood ratio test based on explicit modeling of paralogy (Gayral et al. 2013, Romiguier et al. 2014). Sites with less than six genotyped individual (over ten) were excluded. Non-synonymous (π_n) and synonymous (π_s) nucleotide diversity were computed for each set of genes using Bio ++ (Guéguen et al. 2013). Ninety-five percent confidence intervals were determined by bootstrapping genes (1000 replicates).

Estimation of the Strength of Positive and Purifying Selection

The contribution of positive selection to the process of amino-acid substitution was estimated using the method of Eyre-Walker and Keightley (2009) as implemented in Galtier (2016). This method, which elaborates on the McDonald–Kreitman test (McDonald and Kreitman 1991), models the distribution of the fitness effect (DFE) of deleterious non-synonymous mutations as a negative Gamma distribution. The model is fitted to the synonymous and non-synonymous site frequency spectra (SFS) and the expected dN/dS under near-neutrality is deduced. Here we used folded SFS to avoid polarization issues. The difference between observed and expected dN/dS provides an estimate of the proportion of adaptive non-synonymous substitutions, α. The per mutation rate of adaptive and non-adaptive amino-acid substitution, respectively $\omega_a = \alpha(dN/dS)$ and $\omega_{na} = (1-\alpha)(dN/dS)$, were also computed.

Positive selection analysis was conducted separately in *M. jurtina* and *P. tithonus*, using each species as outgroup to each other. The number of Z-linked coding sequences (genes) for which both polymorphism and divergence data was available was insufficient for a proper estimation of α. Therefore, we used distinct sets of genes for the polymorphism and divergence analyses: synonymous and non-synonymous site frequency spectra (SFS) were built based on 151 Z-linked genes in *M. jurtina* and 144 Z-linked genes in *P. tithonus*, whereas dN and dS were calculated based on 90 Z-linked genes for which a pair of orthologs between the two species was available. This means assuming a common DFE across

distinct sets of genes. The high similarity of the DFEs between Z and autosomes, as well as the remarkable similarity of the estimated DFEs in mitochondrial data sets from different species observed by James et al. (2016) comfort us in this assumption. A similar procedure was followed for autosomal genes, in which SFS were built based on 5,636 and 5,394 genes in *M. jurtina* and *P. tithonus*, respectively, and divergence was estimated from 5,212 genes. Ninety-five percent confidence intervals were determined by bootstrapping genes (1000 repetitions).

When we analyzed Z-linked genes depending on sex-biased expression, we did not have enough genes in each category to use the above-described method. We instead computed a modified version of the Direction of Selection (DoS) statistic (Stoletzki and Eyre-Walker 2011). DoS is defined as the difference between the proportion of fixed differences that are non-synonymous and the proportion of polymorphisms that are non-synonymous [here we used $DoS = dN/(dN + dS) - \pi_n/(\pi_n + \pi_s)$]. A positive DoS indicates adaptive evolution, $DoS = 0$ indicates neutral evolution, and a negative DoS indicates segregating slightly deleterious mutations (Stoletzki and Eyre-Walker 2011). Here DoS values does not go with confidence interval to see if the values differ between the three categories because the DoS is computed from two different sets of genes, one for the divergence data and one for the polymorphism data.

Divergence and Polymorphism Comparison between Sex-Biased Expression Genes

If differences in coding sequence evolution between Z and autosomes were primarily driven by an increased efficacy of selection on recessive mutations in females, we would expect a stronger purifying selection in Z-linked genes that are predominantly expressed in females, compared with Z-linked genes predominantly expressed in males. To test this hypothesis, we defined three categories of Z-linked genes: female-biased (i.e., with a higher level of expression in females than in males), unbiased (expressed at approximately the same level in both sexes) and male-biased (with a higher level of expression in males than in females). To define sex-biased genes in the two species, we estimated the expression level of each coding sequence in each individual using the "idxstats" and "depth" tools of the SAMTOOLS library (Li et al. 2009). We computed the « RPKM » (Reads Per Kilobases Per Million) as follow: $RPKM = N_c * 10^9 / N_{tot} * L_c$, where N_c is the number of reads mapped onto the focal coding sequence, N_{tot} is the total number of reads mapped of the focal individual, and L_c is the length of the focal coding sequence in base pair (Mortazavi et al. 2008). We calculated for each gene the mean RPKM in females, $RPKM_f$, and the mean RPKM in males, $RPKM_m$.

For the comparison of π_n/π_s ratios between sex-biased genes, genes for which $RPKM_f/RPKM_m > 1.5$ were called female-biased, genes for which $RPKM_f/RPKM_m < 0.66$ were called male-biased, and the other ones were called unbiased genes (supplementary tables S2 and S3, Supplementary Materials online). We also compared dN/dS ratios and DoS measurements between Z-linked sex-specific genes, but number of male-biased genes, as defined above, was not sufficient, so we rather sorted genes according to the difference in expression level between males and females and created three bins of equal sizes (30 genes each) (supplementary table S4, Supplementary Materials online). Ninety-five percentage confidence intervals were determined by bootstrapping genes (1000 replicates) in each category for dN/dS as well as polymorphism comparisons. We conducted the same analyses on autosomal genes as a control.

Expression Level Influence on the dN/dS and π_n/π_s Ratios

We tested if a link between gene expression level and dN/dS or π_n/π_s ratios existed in our dataset and if it could explain the difference between Z and autosomes. To do that, we established two linear models using R (v 3.1.2, 2014):

$$\log(dN_{ij}) \sim \log(dS_{ij}) + \text{chromosome\_type}_j + \log(RPKM_i)$$

$$\log(\pi_{nijk}) \sim \log(\pi_{sijk}) + \text{chromosome\_type}_j + \log(RPKM_i) + \text{species}_k,$$

where dN_{ij} is the dN of the i^{th} coding sequence that is linked to chromosome type j (i.e., Z or autosome). For the polymorphism, we added a species specific effect. $RPKM_i$ is the mean RPKM of coding sequence i across all individuals. We excluded the coding sequences with no polymorphism or substitution. We ascertained normality, homoscedasticity and independence of the variables by plotting observed versus predicted values.

Results

Transcriptome Assembly, Genotyping, and Expression Level

RNA-seq data were generated in ten individuals per species (supplementary table S1, Supplementary Materials online). Details of the successive sorting steps are presented in supplementary figure S1, Supplementary Materials online. In *M. jurtina*, transcriptome assembly yielded 145,564 contigs, based on which 36,864 ORFs were predicted. Among these ORFs we recovered orthologous sequences to 11,768 genes from *M. cinxia*, i.e., a large fraction of the reference genome. Of these, 7,647 corresponded to autosomal genes in *M. cinxia*, 378 to Z-linked genes, and 3,743 were unassigned and not considered further. These ORFs were genotyped in the ten individuals. Positions (codon sites) at which less than six individuals were sufficiently covered to be accurately genotyped were discarded, leading to the removal of 119 Z-linked and 2011 autosomal ORFs. We conservatively

Table 1

dN, dS, and dN/dS Ratio Obtained Using Pairwise Alignments for Z-Linked and Autosomal Genes

	#cds	Mean length	dN	dS	dN/dS
Z-linked	90	726	0.025 [0.019; 0.031]	0.31 [0.26; 0.36]	0.082 [0.065; 0.10]
Autosomal	5212	922	0.025 [0.024; 0.026]	0.26 [0.26; 0.27]	0.094 [0.090; 0.097]

Note—Intervals represent 95% confidence intervals obtained by bootstrapping genes (1000 replicates).

removed 108 predicted Z-linked ORFs that exhibited at least one polymorphic site in at least one female (females are expected to be haploid for Z-linked genes). These can reflect genotyping errors, or genes that have been translocated from the Z to an autosome since the divergence with *M. cinxia* (Ahola et al. 2014), or genes located in putative pseudo-autosomal regions. Including these genes of uncertain assignment to the Z-linked data set yielded results qualitatively similar to our main analysis (supplementary table S5, Supplementary Materials online). About 151 predicted Z-linked and 5,636 autosomal ORFs were finally selected for polymorphism analysis. Similarly, in *P. tithonus* we obtained 110,120 contigs, 32,959 ORFs, of which 10,780 had a predicted ortholog in *M. cinxia*. 353 genes were predicted to be Z-linked, of which 144 were selected for polymorphism analysis, together with 5,394 autosomal genes. For the divergence analysis, we selected and aligned 5,212 autosomal and 90 Z-linked coding sequences that were predicted to be orthologous between *M. jurtina* and *P. tithonus*.

In our dataset, the mean level of expression of the Z chromosome was 23% and 41% lower than the autosomal mean expression level in *M. jurtina* and *P. tithonus*, respectively (supplementary fig. S2, Supplementary Materials online). In *M. jurtina*, the mean expression level of Z-linked genes was very similar in males and females (female Z-linked genes expression level was 96% of the male Z-linked genes). In *P. tithonus*, the mean expression level of Z-linked genes was, on average, 14% higher in males than females.

Z-linked vs. Autosomal Divergence

dN/dS was computed in pairwise alignment (table 1). To appreciate the significance of the difference between Z and autosomes, we sampled without replacement 90 autosomal pairwise alignments (1000 replicates), thus matching the number of available Z-linked pairwise alignments (fig. 1). In contrast to what has been observed in the other Z/W systems studied so far, we did not detect any fast-Z effect, the mean dN/dS ratio of Z-linked genes being even slightly lower than the autosomal one.

Besides, we observe a higher dS on the Z chromosome relative to autosomes, which is consistent with the existence of a male-biased mutation rate (Miyata et al. 1987).

Using a multiple regression analysis, we found that neither gene expression level nor chromosome type (i.e., Z versus autosome) had a significant effect on dN ($P = 0.358$ and $P = 0.285$ for expression level and chromosome type, respectively; supplementary table S6, Supplementary Materials online).

Z-linked vs. Autosome Polymorphism

We compared levels of non-synonymous (π_n) and synonymous (π_s) polymorphism and π_n/π_s ratio between Z and autosomes (table 2). The Z chromosome exhibited a lower π_s than the autosomes in both *M. jurtina* and *P. tithonus*. Using π_{sZ}/π_{sA} ratios, we estimated that the Ne_Z/Ne_A ratio is below 0.6 in both species, indicating quite an important difference in effective population size between Z and autosomes. The average π_n/π_s ratio of Z-linked genes was slightly higher than in autosomal genes. To appreciate the significance of the difference between Z and autosomes, we sampled without replacement 151 (in *M. jurtina*) and 144 (in *P. tithonus*) autosomal genes (1000 replicates), thus matching the number of available Z-linked genes. We found that the difference in π_n/π_s between Z and autosomes is not statistically significant, the observed value in Z-linked genes being well within the autosomal distribution (fig. 2). There are no indications that the Z chromosome experiences a reduced efficacy of purifying selection despite its low Ne relative to autosomes (table 2)—everything else being equal; we would expect a higher π_n/π_s in Z than autosomes due to increased intensity of genetic drift in the former.

Using a multiple regression approach, we found that π_n is significantly, negatively correlated to expression level both among autosomal genes ($P < 2e\text{-}16$) and Z-linked genes ($P = 0.0016$, supplementary table S7, Supplementary Materials online). This result is typically interpreted as reflecting an increased strength of purifying selection on highly expressed genes (Drummond et al. 2005). Z-linked genes have, on average, a lower expression level than autosomes. This is another reason, in addition to their reduced Ne, why a significantly higher π_n/π_s ratio on Z than autosomes was expected. The lack of a Z-chromosome effect on π_n/π_s despite reduced expression and smaller Ne suggests that purifying selection is more efficient on the Z chromosome than on autosomes.

Purifying Selection and Sex-Specific Gene Expression

Figure 3 shows that, both in *M. jurtina* and P. *tithonus*, the π_n/π_s ratio is higher in male-biased (60 and 51 genes in *M. jurtina*

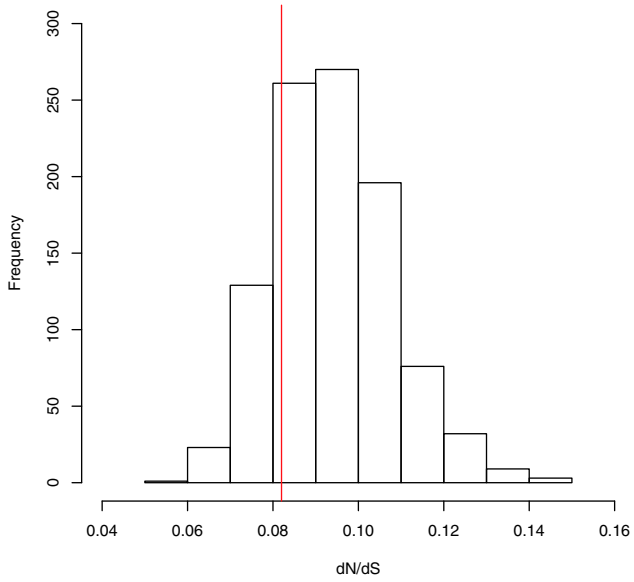


FIG. 1.—Distribution of the dN/dS ratio obtained by resampling without replacement of 90 autosomal pairwise alignments (1000 replicates). dN/dS ratio of the Z-linked pairwise alignments is indicated in red.

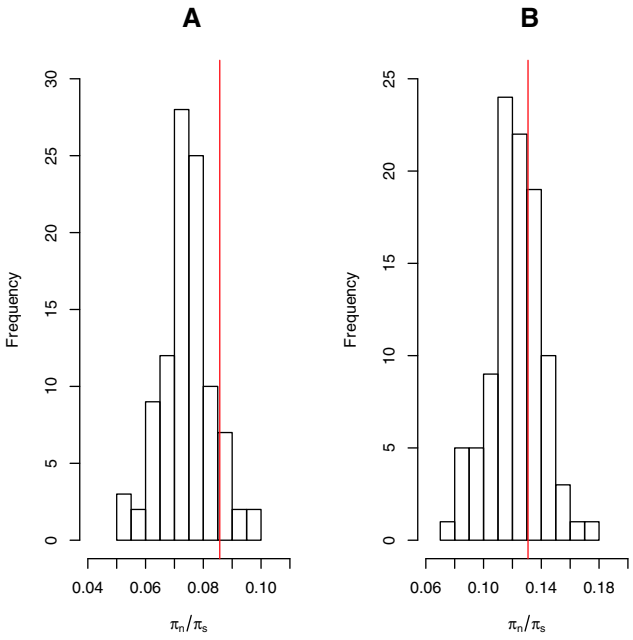


FIG. 2.—Distribution of the π_n/π_s ratio obtained by resampling without replacement of 151 and 144 autosomal genes (1000 replicates). π_n/π_s ratio of the Z-linked genes are indicated in red. (*A*) *M. jurtina* and (*B*) *P. tithonus*.

and *P. tithonus*, respectively) than in female-biased genes (26 and 42 genes in *M. jurtina* and *P. tithonus*, respectively), unbiased genes (65 and 61 genes in *M. jurtina* and *P. tithonus*, respectively) being intermediate. The pattern is clearer in *M. jurtina* than in *P. tithonus*, where CI are somewhat overlapping. This result is fully consistent with the existence of an effect of hemizygosity on the efficacy of purifying selection against recessive, deleterious mutations. Such a pattern was not detectable in autosomes (supplementary table S8, Supplementary Materials online). Interestingly, π_n/π_s in Z-linked male-biased genes was not only higher than the Z chromosome average, but also higher than the autosomal average. This might reflect the increased effect of genetic drift in the Z relative to autosomes, which, promotes the segregation of slightly deleterious, recessive alleles in Z-linked male-biased genes because they are not expressed in the heterogametic sex.

Table 2

π_n, π_s, and π_n/π_s Ratios Values Obtained from Samples of 151 and 5,336 Genes in *M. jurtina* and 144 and 5,396 Genes in *P. tithonus* for Respectively Z-Linked and Autosomal Genes

		M. jurtina	*P. tithonus*
Z-linked	π_n	0.0009 [0.0007; 0.0011]	0.0008 [0.0005; 0.0010]
	π_s	0.010 [0.0088; 0.012]	0.006 [0.005; 0.007]
	π_n/π_s	0.086 [0.068; 0.108]	0.131 [0.096; 0.171]
Autosomal	π_n	0.0022 [0.0022; 0.0023]	0.0012 [0.0012: 0.0013]
	π_s	0.031 [0.030; 0.031]	0.0096 [0.0093; 0.098]
	π_n/π_s	0.073 [0.071; 0.078]	0.126 [0.121; 0.131]
	π_{sZ}/π_{sA}	0.334	0.599

NOTE—Intervals represent 95% confidence intervals obtained by bootstrapping genes (1000 replicates).

Z vs. Autosomal Rate of Adaptive Substitution

We assessed the prevalence of adaptive evolution in Z-linked vs. autosomal genes. Following the method of Eyre-Walker and Keightley (2009), we computed the proportion of adaptive non-synonymous substitutions α, as well as ω_{na} and ω_a, the per mutation rates of non-adaptive and adaptive substitution, respectively. We sampled without replacement 151 (*M. jurtina*) and 144 (*P. tithonus*) autosomal genes (1000 repetitions) to match the same number of genes as for the sex chromosome to establish a SFS, this way generating the expected distribution of α, ω_a and ω_{na} in Z-linked genes under the hypothesis that they follow the same process as autosomal genes (fig. 4).

Both species showed a similar α between Z and autosomes and a slightly lower ω_a on the Z relative to autosomes (table 3 and fig. 4). The 95% confidence intervals were large for Z linked genes and none of the observed differences between Z and autosomes were significant. This suggests that hemizygosity has no strong effect on the rate of adaptive substitution in Satyrinae butterflies, as the prevalence of positive selection in all Z-linked genes taken together was not increased relative to autosomes. In *P. tithonus*, we observed a slightly lower ω_{na} on the Z chromosome, and in *M. jurtina*, ω_{na} was similar

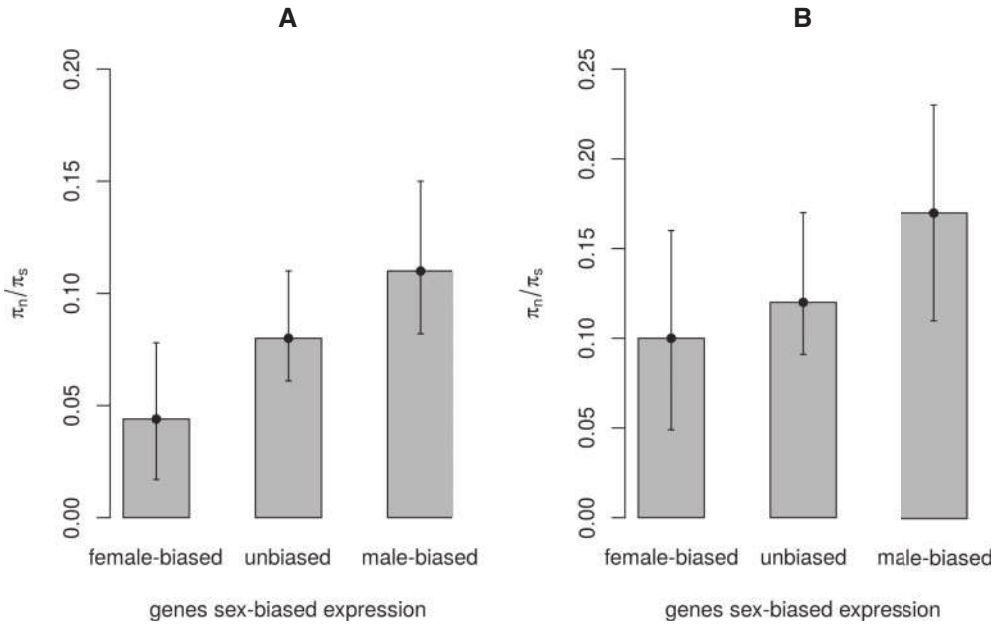


FIG. 3.—π_n/π_s ratios of Z-linked genes of the following categories: female-biased expression, unbiased expression and male-biased expression. Error bars represent ninety-five percent confidence intervals obtained by bootstrapping genes (1000 replicates). (*A*) *M. jurtina* and (*B*): *P. tithonus*.

between Z and autosomes, confirming the pattern of π_n/π_s variation between chromosome types.

Positive Selection and Sex-Specific Gene Expression

We aimed at testing whether adaptive evolution was more prevalent in female-biased and unbiased Z-linked genes than in male-biased Z-linked genes. We had too few genes in each of the three categories to estimate the α and ω_a statistics with sufficient accuracy, so we only compared dN/dS ratios and the DoS statistic.dN/dS ratios in Z-linked female-biased genes (*M. jurtina*: dN/dS = 0.11 ± 0.03, *P. tithonus*: dN/dS = 0.10 ± 0.04) were always higher than in the unbiased (*M. jurtina*: dN/dS = 0.07 ± 0.03, *P. tithonus*: dN/dS = 0.08 ± 0.02) and male-biased (*M. jurtina*: dN/dS = 0.07 ± 0.02, *P. tithonus*: dN/dS = 0.07 ± 0.02) categories of Z-linked genes, which might indicate an increased rate of adaptive substitutions associated to hemizygosity, especially knowing that no significant difference between categories of gene expression was detected in autosomes (supplementary table S9, Supplementary Materials online). Nevertheless, confidence intervals were overlapping, and no significant difference between unbiased and male-biased Z-linked genes was detected.

Contrasting polymorphism and divergence patterns, we obtained positive DoS values for Z-linked female-biased genes, which is indicative of the presence of adaptive substitutions (0.060 and 0.0022 for *M. jurtina* and *P. tithonus*, respectively). For both the unbiased and male-biased Z-linked gene categories, DoS values were negative, indicating a prevalent effect of purifying selection. The DoS statistic was closer to zero in Z-linked unbiased genes (*M. jurtina*: −0.0026; *P. tithonus*: −0.029) than in male-biased genes (*M. jurtina*: −0.030; *P. tithonus*: −0.062), as expected under the hypothesis of positive selection being enhanced by hemizygosity.

Discussion

Faster rates of coding sequence evolution on the Z chromosome relative to the autosomes have been observed across a wide range of species (Dalloul et al. 2010; Mank et al. 2010; Ellegren et al. 2012; Sackton et al. 2014; Wang et al. 2014; Wright et al. 2015), but different reasons have been given to explain the phenomenon. Studies in birds point to genetic drift as the cause of the fixation of a substantial proportion of slightly deleterious mutations, as the effective population size of the Z chromosome is expected to be particularly low relative to the effective population size of autosomes (Mank et al. 2010; Wright et al. 2015). However, a study on the silk moth *Bombyx mori* revealed a fast-Z effect that seems to be due to enhanced positive selection on the Z (Sackton et al. 2014).

Here using another Z/W taxa, we did not find evidence for a fast-Z evolution, and our estimated α was similar between Z and autosomes in the two species. However, we did find evidence for an increased efficacy of purifying selection of the Z relative to autosomes: in spite of a reduced effective population size and expression level, Z-linked genes have a π_n/π_s ratio and a per mutation rate of non-adaptive substitutions (ω_{na}) that are similar to autosomes. We link this phenomenon to the effect of hemizygosity because female-biased Z-linked genes

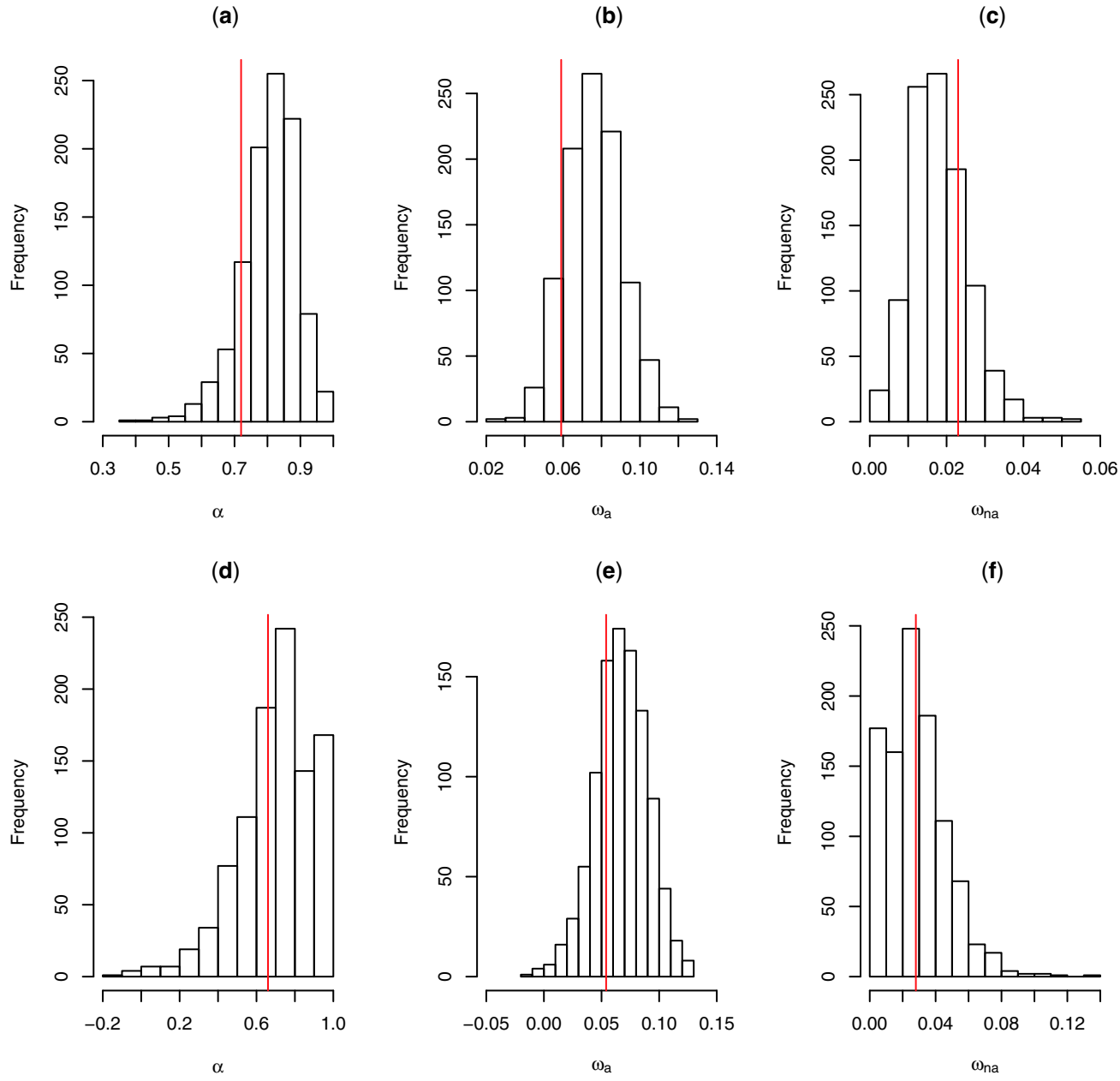


Fig. 4.—Distribution of α, ω_a, and ω_{na} obtained by resampling without replacement 90 autosomal genes (1000 replicates). α, ω_a, and ω_{na} values of the Z-linked genes (obtained with 90 genes for divergence and 151 and 144 genes for polymorphism for *M. jurtina* and *P. tithonus*, respectively) are indicated in red. (*a*) α in *M. jurtina*, (*b*) ω_a in *M. jurtina*, (*c*) ω_{na} in *M. jurtina*, (*d*) α in *P. tithonus*, (*e*) ω_a in *P. tithonus*, and (*f*) ω_{na} in *P. tithonus*.

Table 3

Rate of Adaptation α and Adaptive (ω_a) and Non-Adaptive (ω_{na}) Substitution Rates

	Z-linked			Autosomal		
Species	**α**	**ω_a**	**ω_{na}**	**α**	**ω_a**	**ω_{na}**
M. jurtina	0.72 [0.34;1]	0.059 [0.027;0.094]	0.023 [0.00;0.052]	0.78 [0.74;0.81]	0.072 [0.068;0.077]	0.021 [0.018;0.023]
P. tithonus	0.66 [0.24;1]	0.054 [0.019;0.094]	0.028 [0.00;0.064]	0.63 [0.59;0.68]	0.059 [0.053;0.065]	0.035 [0.030;0.038]

Note—α, ω_a, and ω_{na} were computed according to the method of Eyre-Walker and Keightley (2009), using all the available genes with polymorphism data and all the available genes with divergence data. Intervals represent 95% confidence intervals obtained by bootstrapping genes (1000 replicates).

experience a more efficient purge of slightly deleterious mutations and a slightly higher rate of adaptive evolution than male-biased genes.

Does Sex-Specific Expression Predict Sex-Specific Fitness Effects of Mutations?

By comparing divergence and diversity of Z-linked genes according to sex-specific expression, we intended to investigate the influence of hemizygosity on the strength of selection. Our reasoning is based on the assumption that gene expression level is a predictor of the average intensity of the fitness effect of mutations. We therefore assumed that Z-linked genes expressed at a higher level in males than in females are mostly submitted to selection in males, but weakly or not selected in females, thus escaping the effect of hemizygosity—under this hypothesis, recessive mutations affecting a male-biased Z-linked gene would not affect the phenotype at heterozygous state, whereas recessive mutations affecting a female-biased gene would. The same rationale was used to investigate the effect of hemizygosity in mouse (Kousathanas et al. 2014), silk moth (Sackton et al. 2014), fruit flies (Avila et al. 2015), and birds (Mank et al. 2010).

A link between expression level and the fitness effect of mutations is suggested by the negative correlation between dN/dS and expression level that is found in a wide range of organisms (Drummond and Wilke 2008). In our dataset, we did not observe such correlation; however we did observe a negative correlation between gene expression level and π_n (supplementary tables S5 and S6, Supplementary Materials online). This tends to indicate that highly expressed genes are more constrained, justifying the use of expression level as a predictor of the intensity of fitness effects of mutations when comparing diversity among classes of sex-biased expression. Z-linked genes with a higher expression in male than in female are not necessarily genes with a male-specific function: these might, alternatively, correspond to genes that are not dosage-compensated, and consequently, to genes that are not dosage-sensible. Such dosage-insensitive genes are likely to be submitted to a relaxed purifying selection, which can lead to an elevation of their π_n/π_s ratio irrespective of hemizygosity. To rule out this possibility, we estimated the π_n/π_s ratio of Z-linked genes with more than a twofold male to female level of expression, for which the difference of expression between male and female exceed the difference in ploidy level, and is likely to reflect a sex-biased function. With this restricted set of Z-linked male-biased genes, we obtained π_n/π_s ratios of 0.072 in *M. jurtina* and 0.16 in *P. tithonus*, still higher than the female-biased ratios (respectively 0.044 and 0.10, fig. 3, supplementary table S9, Supplementary Materials online). We conclude that our report of a lower π_n/π_s ratio in female-biased than in male-biased genes can safely be interpreted as reflecting the effect of hemizygosity.

Gene Expression Level, Gene Content and Recombination: Three Potential Biases

When comparing evolutionary rates between different categories of genes, several biases can potentially arise. Expression level often differs between sex chromosomes and autosomes in the heterogametic sex. There is a lack of consensus in regards to complete dosage compensation in Lepidoptera species (Zha et al. 2009; Harrison et al. 2012; Walters et al. 2015). Here we show that in *M. jurtina* and *P. tithonus*, female and male expression levels are roughly similar in Z-linked genes, thereby limiting the potential bias due to differences in gene dosage on the Z chromosome pattern of diversity. Our conclusions are anyway conservative with respect to this potential bias because we report similar π_n/π_s ratio in Z-linked and autosomal genes (fig. 2 and table 2), despite a lower average gene expression level in the Z chromosome. A recent study in mammals showed that two genomic features, namely GC-content and gene expression level, are sufficient to explain the higher dN/dS of X-linked genes compared with autosomes (Nguyen et al. 2015). The absence of a detectable effect of hemizygosity on dN/dS in mammals could be explained by a more efficient purging of recessive deleterious mutations on the X, which tends to reduce dN/dS, thus offsetting the increased fixation rate of recessive beneficial mutations, similar to the satyrine situation. This hypothesis could not be explicitly tested in mammals because of the relatively low level of within-species polymorphism in this group.

Another potential bias is a difference in gene content between Z and autosomes, which could lead to differences in the adaptive mutation rate between chromosome types. We do not detect any obvious difference between the estimated DFEs of Z and autosomes (supplementary figs. S3 and S4, Supplementary Materials online), which does not suggest that Z-linked genes are more prone to adaptation than autosomal genes. Moreover, enrichment tests performed in birds and primates revealed no significantly enriched gene ontology terms for Z(X)-linked genes relative to autosomes (Hvilsom et al. 2014; Wright et al. 2015). Nevertheless, differences in gene content between Z and autosomes remains largely unknown in satyrine Lepidoptera, so we cannot totally exclude any influence of gene content on our results.

Finally, when comparing Z-linked versus autosomal genes, one should also consider a potential difference in terms of recombination rate. Indeed, it has been shown that in various families of Lepidoptera females lack recombination (Suomalainen et al. 1973; Turner and Sheppard 1975; Traut 1977; Fisk 1989). We can assume that it is also the case in Satyrinae, and consequently, that the population-effective recombination rate for a given rate of recombination r in females between two loci is $r/2$ for autosomes and $2r/3$ for Z. Assuming that background selection is at work, and everything else being equal, one could thus expect $\pi_{sZ} > \pi_{sA}$, and more efficient purifying selection on the Z due to reduce

linkage (Charlesworth 2012), perhaps confounding the effects of hemizygosity. Nevertheless, our data do not meet the prediction of a higher neutral diversity on Z than on autosomes, and in the absence of a direct estimation of recombination rate in Satyrines this remains speculative.

Lack of Fast-Z Effect in Satyrine Butterflies: Where Are the Adaptive Substitutions on the Z Chromosome?

In this study, we obtained results that set the Satyrinae apart from the majority of other species in which sex chromosome evolution has been investigated so far: in spite of the evidence for an effect of hemizygosity on purifying selection, we did not observe an increase in ω_a on the Z relative to autosomes (table 3), which, combined with the slight decrease in ω_{na} on the Z, led to a slight slow-Z effect. Hemizygosity is expected to promote the fixation of recessive adaptive mutations, and to facilitate the purge of recessive deleterious mutations. Here we show that the latter effect can be stronger than the former. An effect of hemizygosity on the adaptive rate in Satyrinae is indeed suggested by the higher dN/dS and DoS we report in female-biased than in male-biased genes, but the impact of deleterious mutations dominates, so that the net effect is a slow-down, not an acceleration, of molecular evolution on the Z.

Therefore, why would satyrine butterflies behave differently from the other species in which molecular evolution of sex-chromosomes has been examined (Hvilsom et al. 2014; Kousathanas et al. 2014; Avila et al. 2015; Wright et al. 2015)? The rate of evolution of Z-linked genes is determined by the distribution of selection coefficients of mutations occurring on the Z, the distribution of dominance coefficients, and the relative proportion of male-biased vs. female-biased genes. In principle, any peculiarity of Satyrinae regarding one of these parameters, of which we have no empirical measurement, could contribute to explaining the lack of a fast-Z effect in this group. Additionally, if adaptation uses standing genetic variation, if mutations are recurrent, or if individual bouts of adaptation are restricted to a small amount of genes, hemizygosity is not expected to influence the adaptive substitution rate of the Z chromosome (Meisel and Connallon 2013). Any of these situations might apply in satyrines more often than in other groups, for some yet undetermined reason.

Alternatively, it could be that our Z-linked genes sample, based on RNAseq data, is biased towards genes receiving less adaptive mutations than in other studies. Transcriptome-based analysis of Z-linked genes in birds (Wright et al. 2015) did reveal a fast-Z effect, but failed to detect any conspicuous influence of adaptive evolution, similar to our current analysis. Nevertheless, we report for both autosomal and Z-linked genes a high proportion of adaptive substitution (α) (between 0.63 and 0.78), so it is not likely that we missed the fraction of fast adapting genes. There might also be, finally, a publication bias in favor of the fast-Z, especially in the pre-Next Generation Sequencing era, when data sets were limited in size and failure to detect a fast-Z effect could be attributed to lack of power. The analysis of sex-linked genes evolution in additional taxa appears required to settle this issue and confirm, or not, the generality of the fast-Z effect.

The Impact of Genetic Drift on the Sex Chromosome in Female Heterogametic Taxa

The use of the ratio π_{sZ}/π_{sA} as a proxy of Ne_Z/Ne_A is a reasonable approximation assuming that there is no sex mutation bias. Here we report a slightly higher dS on the Z chromosome than on autosomes, suggestive of a male-biased mutation rate. However, the difference in dS between the two compartments is small in absolute terms, so we considered that the π_{sZ}/π_{sA} is still representative of Ne_Z/Ne_A . This is in line with the results obtained in *Bombyx* (Sackton et al. 2014) and in birds (Axelsson 2004), where Z-linked genes evolve only marginally faster than autosomes.

Theoretically, the ratio Ne_Z/Ne_A is expected to vary between species due to differences in intensity of sexual selection toward males, which reduces the number of reproductive males in the population and biases the efficient sex ratio (Wright et al. 2015). Our report of Ne_Z/Ne_A ratios that are substantially lower than 0.75 in *M. jurtina* and *P. tithonus* (table 2) are consistent with Bateman's principle, which states that the variation in reproductive success is greater among males than among females (Oberhauser 1988). These ratios are comparable to the ones reported in birds ($0.29 < Ne_Z/Ne_A < 0.46$, Wright et al. 2015), and according to Vicoso and Charlesworth (2009), such ratios should lead to an elevation of the fixation rate of deleterious mutations on the Z chromosome for all ranges of dominance.

Nevertheless, we do not detect a strong effect of genetic drift on the Z, as we observe neither an increase of the dN/dS and π_n/π_s ratios nor a significantly higher ω_{na} on the Z relative to autosomes (fig. 2, tables 1 and 3). Only in male-biased Z-linked genes, which are under limited influence of hemizygosity, did we detect a reduced efficacy of purifying selection. The difference in effective population size between Z and autosomes does not seem to be sufficient to lead to a strong decrease in the efficacy of purifying selection on the Z compare to autosomes in *M. jurtina* and *P. tithonus*. Overall Ne may be one of the reasons explaining the different patterns observed between birds and Lepidoptera (Sackton et al. 2014). A low Ne_Z/Ne_A ratio, may thus have weaker consequences on drift rate at sex-linked genes in large than in small populations.

Conclusions

We compared coding sequence evolutionary rates between the hemizygous Z chromosome and autosomes in two closely related species of butterflies. Combining diversity, divergence and expression data, we assessed the influence of hemizygosity in shaping Z-linked genes evolution. Our results indicate that due to hemizygosity, purifying selection is more effective on the Z chromosome relative to autosomes, preventing slightly deleterious mutations to reach fixation. Adaptive substitutions seem to be rare enough not to enhance the adaptive fixation rate in the Z relative to autosomes. Those two results explain why we do not detect a faster Z evolution in *M. jurtina* and *P. tithonus*, unlike what has been previously found in other systems. We suggest that the effect of hemizygosity on the fate of recessive deleterious mutations, which has been largely neglected until now, should be taken into account when interpreting patterns of molecular evolution in sex chromosomes vs. autosomes.

Data Accessibility

Illumina raw reads are deposited under the project PRJNA326910 in the SRA database.

Supplementary Material

Supplementary tables S1–S9 and figures S1–S4 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

We thank Emeric Figuet, Yoann Anselmetti and the Montpellier Bioinformatcs & Biodiversity platform for their help and advice with bioinformatics. Sequencing was performed by MGX-Montpellier GenomiX. This work was supported by European Research Council grant (ERC PopPhyl 232971), Agence Nationale de la recherche grants ANR-14-CE02-0002-01 "BirdIslandGenomic" and ANR-15-CE12-0010 "DarkSideOfRecombination".

Literature Cited

Ahola V, et al. 2014. The Glanville fritillary genome retains an ancient karyotype and reveals selective chromosomal fusions in Lepidoptera. Nat. Commun. 5:1–9.

Betancourt AJ, Presgraves DC, Swanson WJ. 2002. A test for faster X evolution in *Drosophila*. Mol. Biol. Evol. 19:1816–1819.

Avila V, et al. 2014. Faster-X effects in two *Drosophila* lineages. Genome Biol. Evol. 6:2968–2982.

Avila V, Campos JL, Charlesworth B. 2015. The effects of sex-biased gene expression and X-linkage on rates of adaptive protein sequence evolution in *Drosophila*. Biol. Lett. 11:20150117.

Axelsson E. 2004. Male-biased mutation rate and divergence in autosomal, Z-linked and W-linked introns of chicken and turkey. Mol. Biol. Evol. 21:1538–1547.

Oberhauser KS. 1988. Male monarch butterfly spermatophore mass and mating strategies. Anim. Behav. 36:1384–1388.

Cahais V, et al. 2012. Reference-free transcriptome assembly in non-model animals from next-generation sequencing data. Mol. Ecol. Resour. 12:834–845.

Campos JL, Halligan DL, Haddrill PR, Charlesworth B. 2014. The relation between recombination rate and patterns of molecular evolution and variation in *Drosophila melanogaster*. Mol. Biol. Evol. 31:1010–1028.

Carneiro M, et al. 2012. Evidence for widespread positive and purifying selection across the european rabbit (*Oryctolagus cuniculus*) genome. Mol. Biol. Evol. 29:1837–1849.

Charlesworth B, Coyne JA, Barton NH. 1987. The relative rates of evolution of sex chromosomes and autosomes. Am. Nat. 130:113–146.

Charlesworth B. 2012. The role of background selection in shaping patterns of molecular evolution and variation: evidence from variability on the Drosophila X chromosome. Genetics 191:233–246.

Counterman B, Ortíz-Barrientos D, Noor MF. 2004. Using comparative genomic data to test for fast-X evolution. Evolution 58:656–660.

Dalloul RA, et al. 2010. Multi-platform Next-Generation sequencing of the domestic Turkey (*Meleagris gallopavo*): genome assembly and analysis. PLoS Biol. 8:e1000475.

Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. 2005. Why highly expressed proteins evolve slowly. Proc. Natl Acad. Sci. U S A. 102:14338–14343.

Drummond DA, Wilke CO. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell 134:341–352.

Ellegren H, et al. 2012. The genomic landscape of species divergence in *Ficedula* flycatchers. Nature 491:756–760.

Eyre-Walker A, Keightley PD. 2009. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. Mol. Biol. Evol. 26:2097–2108.

Fisk J. 1989. Karyotype and achiasmatic female meiosis in *Helicoverpa armigera* (Hübner) and *H. punctigera* (Wallengren) (Lepidoptera: Noctuidae). Génome 32:967–971.

Galtier N. 2016. Adaptive protein evolution in animals and the effective population size hypothesis. PLoS Genet. 12:e1005774.

Garrigan D, Kingan SB, Geneva AJ, Vedanayagam JP, Presgraves DC. 2014. Genome diversity and divergence in *Drosophila mauritiana*: multiple signatures of faster X evolution. Genome Biol. Evol. 6:2444–2458.

Gayral P, et al. 2011. Next-generation sequencing of transcriptomes: a guide to RNA isolation in nonmodel animals. Mol. Ecol. Resour. 11:650–661.

Gayral P, et al. 2013. Reference-free population genomics from Next-Generation transcriptome data and the vertebrate–invertebrate gap. PLoS Genet. 9:e1003457.

Glémin S. 2007. Mating systems and the efficacy of selection at the molecular level. Genetics 177:905–916.

Grabherr MG, et al. 2011. Full-length transcriptome assembly from RNA-seq data without a reference genome. Nat. Biotechnol. 15:644–652.

Guéguen L, et al. 2013. Bio++: efficient extensible libraries and tools for computational molecular evolution. Mol. Biol. Evol. 30:1745–1750.

Haldane JBS. 1924. A mathematical theory of natural and artificial selection. Part I. Trans. Cambridge Phil. Soc. 23:19–41.

Harrison PW, Mank JE, Wedell N. 2012. Incomplete sex chromosome dosage compensation in the Indian meal moth, *Plodia interpunctella*, based on *de novo* transcriptome assembly. Genome Biol. Evol. 4:1118–1126.

Huang X, Madan A. 1999. CAP3: a DNA sequence assembly program. Genome Res. 9:868–877.

Hvilsom C, et al. 2011. Extensive X-linked adaptive evolution in central chimpanzees. Proc. Natl Acad. Sci. U S A. 109:2054–2059.

James JE, Piganeau G, Eyre-Walker A. 2016. The rate of adaptive evolution in animal mitochondria. Mol. Ecol. 25:67–78.

Julien P, et al. 2012. Mechanisms and evolutionary patterns of mammalian and avian dosage compensation. PLoS Biol. 10:e1001328.
Kousathanas A, Halligan DL, Keightley PD. 2014. Faster-X adaptive protein evolution in house mice. Genetics 196:1131–1143.
Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760.
Li H, et al. Subgroup 1000 Genome Project Data Processing 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079.
McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the *Adh* locus in *Drosophila*. Nature 385:652–654.
Mackay TFC, et al. 2012. The *Drosophila melanogaster* Genetic Reference Panel. Nature 482:173–178.
Mank JE, Nam K, Ellegren H. 2010. Faster-Z evolution is predominantly due to genetic drift. Mol. Biol. Evol. 27:661–670.
Marín I, Siegal ML, Baker BS. 2000. The evolution of dosage-compensation mechanisms. BioEssays 22:1106–1114.
Meisel RP, Connallon T. 2013. The faster-X effect: integrating theory and data. Trends Genet. 29:537–544.
Miyata T, Hayashida H, Kuma K, Mitsuyasu K, Yasunaga T. 1987. Male-driven molecular evolution: a model and nucleotide sequence analysis. Cold Spring Harb. Symp. Quant. Biol. 52:863–867.
Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods 5:621–628.
Nam K, et al. 2015. Extreme selective sweeps independently targeted the X chromosomes of the great apes. Proc. Natl Acad. Sci. U S A. 112:201419306.
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.
Nguyen LP, Galtier N, Nabholz B. 2015. Gene expression, chromosome heterogeneity and the fast-X effect in mammals. Biol. Lett. 11:20150010.
Ranwez V, Harispe S, Delsuc F, Douzery EJP. 2011. MACSE: Multiple Alignment of Coding SEquences accounting for frameshifts and stop codons. PLoS One 6:e22594.
Rice WR. 1984. Sex chromosomes and the evolution of sexual dimorphism. Evolution 38:735–742.
Romiguier J, et al. 2012. Fast and robust characterization of time-heterogeneous sequence evolutionary processes using substitution mapping. PLoS One 7:e33852.
Romiguier J, et al. 2014. Comparative population genomics in animals uncovers the determinants of genetic diversity. Nature 515:261–263.
Sackton TB, et al. 2014. Positive selection drives faster-Z evolution in silkmoths. Evolution 68:2331–2342.
Simpson JT, et al. 2009. ABySS: a parallel assembler for short read sequence data. Genome Res. 19:1117–1123.
Singh ND, Davis JC, Petrov DA. 2005. X-linked genes evolve higher codon bias in *Drosophila* and *Caenorhabditis*. Genetics 171:145–155.
Stoletzki N, Eyre-Walker A. 2011. Estimation of the neutrality index. Mol. Biol. Evol. 28:63–70.
Suomalainen E, Cook LM, Turner JRG. 1973. Achiasmatic oogenesis in the Heliconiine butterflies. Hereditas 74:302–304.
Szövényi P, et al. 2014. Efficient purging of deleterious mutations in plants with haploid selfing. Genome Biol. Evol. 6:1238–1252.
The Heliconius Genome Consortium, et al. 2012. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. Nature 487:94–98.
Thornton K, Long M. 2002. Rapid divergence of gene duplicates on the *Drosophila melanogaster* X chromosome. Mol. Biol. Evol. 19:918–925.
Traut W. 1977. A study of recombination, formation of chiasmata and synaptonemal complexes in female and male meiosis of *Ephestia kuehniella* (Lepidoptera). Genetica 47:135–142.
Tsagkogeorga G, Cahais V, Galtier N. 2012. The population genomics of a fast evolver: high levels of diversity, functional constraint, and molecular adaptation in the tunicate *Ciona intestinalis*. Genome Biol. Evol. 4:740–749.
Turner JRG, Sheppard PM. 1975. Absence of crossing-over in female butterflies (*Heliconius*). Heredity 34:265–269.
Veeramah KR, Gutenkunst RN, Woerner AE, Watkins JC, Hammer MF. 2014. Evidence for increased levels of positive and negative selection on the X chromosome versus autosomes in humans. Mol. Biol. Evol. 31:2267–2282.
Vicoso B, Charlesworth B. 2009. Effective population size and the faster-X effect: an extended model. Evolution 63:2413–2426.
Walters JR, Hardcastle TJ, Jiggins CD. 2015. Sex chromosome dosage compensation in *Heliconius* butterflies: global yet still incomplete? Genome Biol. Evol. 7:2545–2559.
Wang B, Ekblom R, Bunikis I, Siitari H, Höglund J. 2014. Whole genome sequencing of the black grouse (*Tetrao tetrix*): reference guided assembly suggests faster-Z and MHC evolution. BMC Genomics 15:180.
Wright AE, et al. 2015. Variation in promiscuity and sexual selection drives avian rate of Faster-Z evolution. Mol. Ecol. 24:1218–1235.
Zha X, et al. 2009. Dosage analysis of Z chromosome genes using microarray in silkworm, *Bombyx mori*. Insect Biochem. Mol. Biol. 39:315–321.

Associate editor: Judith Mank

Estimation et analyse du taux de substitution adaptatif chez les animaux

Résumé : Comprendre les déterminants du taux d'adaptation est une question primordiale en évolution moléculaire. En particulier, l'influence de la taille efficace de population sur la sélection positive, ainsi que la nature des changements d'acides aminés qui mènent à de l'adaptation sont des questions encore débattues. Pour y répondre, la méthode DFE-α, dérivée du test fondateur de McDonald & Kreitman, est un outil puissant pour mesurer le taux de substitution adaptatif. Elle est néanmoins sensible à certains biais. Au cours de cette thèse, nous avons identifié deux biais majeurs de cette méthode, les fluctuations de long-terme du régime de sélection-dérive via des fluctuations démographiques, et la conversion génique biaisée vers GC (gBGC). Via des simulations, nous avons montré que divers scénarios plausibles de fluctuations démographiques peuvent mener à une sur-estimation du taux de substitution adaptatif. Nous avons aussi obtenu des indications empiriques que le régime de sélection-dérive récent ne reflète pas le régime de sélection-dérive de long-terme chez diverses espèces animales, ce qui représente une violation d'une hypothèse forte de la méthode DFE-α. D'autre part, nous avons montré que la gBGC entraîne une sur-estimation du taux de substitution adaptatif chez les primates et les oiseaux. Via un jeu de données de neuf taxons de métazoaires et un total de 40 espèces, nous avons d'une part initié une analyse visant à identifier la nature des changements d'acides aminés qui mènent à l'adaptation, et montré que les changements radicaux sont soumis à une plus forte sélection purificatrice que les changements conservatifs. D'autre part, nous avons pu évaluer le lien entre la taille efficace et le taux de substitution adaptatif tout en prenant en compte les deux sources de biais explorées précédemment. Nous avons mis en évidence pour la première fois une relation négative entre le taux de substitution adaptatif et des traits d'histoire de vie représentatifs de la taille de population de long-terme. Ce résultat va à l'encontre de l'hypothèse canonique d'une adaptation plus efficace en grandes populations.

Estimation and analysis of the adaptive substitution rate in animals

Abstract: Understanding the determinants of the adaptive substitution rate is a central question in molecular evolution. In particular, the influence of the effective population size N_e on positive selection as well as the nature of amino acid changes that lead to adaptation are still debated. The DFE-α method, which was derived from the seminal McDonald & Kreitman test, is a powerful tool for estimating the adaptive substitution rate. However, it is sensitive to various sources of bias. In this thesis, we identified two major sources of bias of this test, long-term fluctuations of the selective-drift regime through demographic fluctuations, and GC-biased gene conversion (gBGC). Using simulations, we showed that under plausible scenarios of fluctuating demography, the DFE-α method can lead to a severe over-estimation of the adaptive substitution rate. We also showed that polymorphism data reflect a transient selective-drift regime which is unlikely to correspond to the average regime experienced by genes and genomes during the long-term divergence between species. This violates an important assumption of the DFE-α method. Our results also indicate that gBGC leads to an over-estimation of the adaptive substitution rate in primates and birds. Using a dataset of nine metazoan taxa for a total of 40 species, we started an analysis aiming at identifying the type of amino acid changes that are more prone to adaptation, and evaluated the link between N_e and the adaptive substitution rate while accounting for the two sources of bias previously explored. We reveal for the first time a negative relationship between the adaptive substitution rate and life-history traits representative of long-term N_e. This result is in contradiction with the widespread hypothesis that adaptation is more efficient in large populations.