

# Table des matières

<b>Résumé</b>	<b>2</b>
<b>1 Introduction</b>	<b>5</b>
1.1 Émergence de la théorie neutraliste de l'évolution moléculaire . . . . .	6
1.1.1 La théorie de l'évolution au début du XX <sup>ème</sup> siècle . . . . .	6
1.1.2 La querelle du polymorphisme . . . . .	10
1.2 Théorie neutraliste de l'évolution moléculaire . . . . .	16
1.2.1 Hypothèses et paramètres . . . . .	16
1.2.2 Outils mathématiques . . . . .	18
1.3 Utilisation du modèle neutre en évolution moléculaire . . . . .	22
1.3.1 Description de la diversité génétique . . . . .	22
1.3.2 Inférences dans le cadre de la théorie neutraliste . . . . .	24
1.3.3 Incohérences liées à l'utilisation du modèle neutre . . . . .	29
1.4 D'autres modèles de génétique des populations . . . . .	31
1.4.1 Processus Naissance-Mort . . . . .	31
1.4.2 Coalescences multiples . . . . .	32
1.5 Objectifs de la thèse . . . . .	37
<b>2 L'impact de la sélection, de la conversion génique et des biais d'échantillonnage sur l'inférence de démographie microbienne</b>	<b>39</b>
2.1 Résumé de l'article . . . . .	39
2.2 Article . . . . .	41
2.3 Annexes . . . . .	57
<b>3 Exactitude des inférences démographiques basées sur le spectre de fréquence allélique : l'exemple de la population Yoruba</b>	<b>63</b>
3.1 Résumé de l'article . . . . .	63
3.2 Article . . . . .	65
3.3 Annexes . . . . .	77

3.3.1	Informations supplémentaires de l'article . . . . .	77
3.3.2	Analyses complémentaires . . . . .	83
<b>4</b>	<b>D'autres modèles pour expliquer la diversité des données : l'exemple des modèles à coalescences multiples</b>	<b>90</b>
4.0.0	Représentation graphique du spectre de fréquence . . . . .	90
4.1	Données . . . . .	93
4.1.1	Informations sur les données rassemblées . . . . .	93
4.1.2	Spectres de fréquence observés . . . . .	95
4.1.3	Virus : spectre de fréquence inadapté . . . . .	101
4.2	Les erreurs d'orientation . . . . .	103
4.2.1	Allèle ancestral . . . . .	103
4.2.2	Effets des erreurs sur le spectre de fréquence . . . . .	104
4.2.3	Estimer et corriger les erreurs à partir des données . . . . .	106
4.3	Ajuster les données avec des coalescences multiples et de la démographie .	109
4.3.1	Méthodes . . . . .	109
4.3.2	Résultats . . . . .	111
4.4	Le biais de conversion génique . . . . .	118
4.5	Discussion . . . . .	122
4.6	Perspectives . . . . .	123
<b>5</b>	<b>Conclusion générale et discussion</b>	<b>125</b>
<b>6</b>	<b>Annexes</b>	<b>134</b>
	<b>Bibliographie</b>	<b>143</b>

# Chapitre 1

## Introduction

Cette thèse porte sur les modèles de référence en évolution moléculaire. En particulier, elle s'intéresse au modèle de référence actuellement dominant, la théorie neutraliste de l'évolution moléculaire. L'objectif de la thèse est d'étudier comment ce modèle de référence est utilisé, en particulier quelles sont les conséquences de ses hypothèses sur les inférences faites dans son cadre. Dans un deuxième temps, nous avons pour objectif de comparer ce modèle de référence à d'autres modèles possibles, basés sur d'autres hypothèses.

Dans cette introduction, je présente tout d'abord un résumé historique de l'émergence de la théorie neutraliste dans la partie 1.1. N'ayant pas pour ambition de retracer une histoire exhaustive de la théorie de l'évolution, j'ai choisi de démarrer cet historique au début du XX<sup>ème</sup> siècle, en présentant les théories en vigueur à cette époque et leur développement jusqu'à dans les années 1960. C'est à cette période qu'apparaissent de nouvelles données aboutissant à la proposition d'une nouvelle théorie, la théorie neutraliste, détaillée dans la partie 1.2. Je présente ensuite les applications actuelles de cette théorie en évolution moléculaire (partie 1.3). Après avoir mis en évidence quelques problèmes et incohérences liés à cette utilisation, je présente dans la partie 1.4 d'autres modèles possibles, déjà décrits mais peu ou pas utilisés, et que nous nous proposons dans cette thèse de confronter aux données. Enfin, dans la partie 1.5, je détaille les objectifs de cette thèse, et je décris sommairement les différentes parties qui la composent en exposant les questions auxquelles elles tentent de répondre.

## 1.1 Émergence de la théorie neutraliste de l'évolution moléculaire

### 1.1.1 La théorie de l'évolution au début du XX<sup>ème</sup> siècle

Ce résumé historique (section 1.1.1) est principalement basé sur les chapitres 10 et 11 du livre d'Hervé Le Guyader, *Penser l'Évolution*, paru en 2012.

Au début du XX<sup>ème</sup> siècle, cela fait quarante ans que l'ouvrage fondateur de Charles Darwin, *L'Origine des espèces*, a été publié. Le concept général d'évolution a été rapidement accepté après la publication de l'ouvrage, mais le mécanisme proposé par Darwin, la sélection naturelle, reste controversé. Darwin base sa théorie de l'évolution par sélection naturelle sur trois principes : variations entre individus, adaptation au milieu et hérédité des caractères. Mais il ne fournit pas d'explication précise sur le mécanisme d'apparition de nouvelles espèces. Ainsi, d'autres alternatives sont encore défendues pour expliquer l'évolution. Certains s'appuient sur la théorie que Jean-Baptiste de Lamarck a formulée au début du XIX<sup>ème</sup> siècle, et soutiennent que l'évolution s'effectue par transmission des caractères acquis. D'autres pensent que l'évolution se fait par sauts évolutifs importants en une seule génération : c'est le saltationnisme. Enfin, l'orthogénèse défend l'idée que les organismes sont soumis à des forces internes telles que des lois de développement, qui guident l'évolution dans une certaine direction.

### Lois de Mendel et génétique des populations

En 1900, les travaux de Johann Gregor Mendel, initialement publiés en 1865, sont redécouverts indépendamment par trois botanistes européens, le Hollandais Hugo de Vries, l'Allemand Carl Correns et l'Autrichien Erich Tschermak (voir le numéro spécial « 1900 : Redécouverte des lois de Mendel » des *Comptes rendus de l'Académie des sciences*, série III, tome 323, n°12, 1033-1196, décembre 2000). Cette redécouverte passe par des expériences ou des observations semblables à celles qu'avait faites Mendel. En particulier, ces résultats montrent que les contributions de chaque parent gardent leur intégrité dans la descendance, au lieu de se mélanger comme cela était souvent supposé dans les théories précédentes de l'hérédité. Dans le cas de De Vries, cette redécouverte s'associe à la notion de *mutation* qu'il introduit en décembre 1901 après avoir observé des variations brusques de certains caractères chez une plante qu'il étudie. Ces mutations sont selon lui à l'origine de la variabilité de l'espèce.

La redécouverte de ces lois en 1900 déclenche une opposition entre les « mendéliens »



(parmi lesquels William Bateson et Hugo de Vries) et les biostatisticiens ou biométriciens (parmi lesquels Karl Pearson et Walter Weldon). Ces derniers cherchent, sous l'influence des idées de Francis Galton, à développer une théorie statistique de l'évolution. Ils veulent estimer les taux d'évolution et l'intensité de la sélection naturelle à partir de mesures de différents caractères, réalisées sur des populations animales, sur lesquels selon eux la sélection naturelle agit par des variations infimes. À l'inverse des biométriciens qui construisent leur théorie de la sélection en choisissant de ne pas se préoccuper de la nature de l'hérédité, les mendéliens se concentrent eux sur la théorie de la mutation d'Hugo de Vries et soutiennent que les nouvelles espèces apparaissent par des sauts mutationnels plutôt que par sélection graduelle.

Aucune des deux approches n'était réellement en contradiction avec la théorie de l'évolution de Darwin, mais elles s'intéressaient chacune à deux aspects différents de la théorie darwinienne : les biométriciens étaient focalisés sur la sélection et les mendéliens sur la modification, ce qui a causé cette controverse entre les deux visions. Le mendélisme était vu comme en désaccord avec la théorie de Darwin, car il était associé par Bateson au saltationnisme, alors que les biométriciens, qui se réclamaient héritiers de Darwin, défendaient une vue gradualiste (Nordmann, 1992).

Les lois de Mendel sont étendues aux espèces animales par Lucien Cuénot en 1902. Les termes de *génétique* et de *gène* apparaissent respectivement en 1906 (William Bateson) et 1909 (Wilhelm Johannsen). Le gène, qui remplace le *facteur* utilisé jusqu'à présent, bien que toujours abstrait, est maintenant une unité de mutation en plus d'être une unité de transmission. À partir de ces nouveaux concepts, ainsi que de ceux de *génotype* et *phénotype* introduits peu après, apparaissent deux voies de recherche. D'un côté on cherche à connaître le support matériel de l'hérédité, et de l'autre on s'intéresse au rôle de la mutation dans le mécanisme de sélection proposé par Darwin.

La recherche du support matériel de l'hérédité aboutit dans un premier temps à la théorie chromosomique selon laquelle les chromosomes sont le support des gènes (Walter Sutton, Theodor Boveri, équipe de Thomas H. Morgan), développée à partir de 1902 et largement acceptée par la communauté à partir de 1914. Les notions de gènes *indépendants* ou *liés* sont ainsi expliquées, et on met en évidence la recombinaison, via l'observation des crossing-over. Ainsi, le gène devient le support d'un caractère héréditaire qui peut recombiner et muter, et donc être à la base de la descendance avec modification proposée par Darwin.

La génétique des populations, discipline qui apparaît à cette époque, a pour but de concilier les concepts de la génétique et de la biologie évolutive, en étudiant la composi-

tion génétique de populations et ses changements sous l'action de différents facteurs, dont la sélection. Plus précisément, des modèles mathématiques sont développés pour décrire les variations de fréquences alléliques, afin d'établir des prédictions générales et de les confronter aux données. Contrairement à la sélection artificielle, qui a pu être mise en évidence par la domestication, la sélection naturelle ne peut pas être testée expérimentalement, on ne connaît pas précisément son mode de fonctionnement, il faut donc passer par de la modélisation.

Le premier résultat majeur de génétique des populations est proposé indépendamment par le mathématicien anglais Godfrey Hardy et le médecin allemand Wilhelm Weinberg en 1908 : c'est la loi dite de Hardy-Weinberg qui décrit les fréquences alléliques et génotypiques dans une population à l'équilibre. Pour une population de grande taille, dans laquelle les croisements sont aléatoires et les forces évolutives (sélection naturelle, mutation et migration) n'interviennent pas, les proportions génotypiques sont constantes de génération en génération. On peut attribuer une partie du succès de cette loi à sa simplicité mathématique (Hervé Le Guyader rapporte dans son ouvrage que « Hardy, qui fit le calcul sur un coin de table, à la fin d'un repas, ne voyait pas l'intérêt de sa publication »). Cette loi permet de tester si une population donnée est à l'équilibre. Les généticiens des populations vont progressivement complexifier les modèles pour évaluer les forces évolutives qui font qu'une population n'est pas à l'équilibre d'Hardy-Weinberg.

En 1924, John B.S. Haldane introduit la notion de valeur sélective, *fitness* en anglais, pour étudier les effets de la sélection. Cette valeur comprise entre 0 et 1, qui caractérise chaque génotype, représente la capacité de survie et de reproduction dans un environnement donné. Haldane établit une relation entre l'intensité de la sélection et les changements de fréquences alléliques et détermine ainsi les probabilités pour qu'un allèle se fixe dans une population, c'est à dire que sa fréquence atteigne 1, ou qu'il disparaisse, ou qu'il soit maintenu à un équilibre polymorphe si l'hétérozygote a la plus grande valeur sélective. Enfin, il introduit la notion d'équilibre mutation-sélection pour les allèles délétères : la mutation les fait apparaître et la sélection les élimine. Une synthèse de ses travaux est publiée en 1932 dans le livre *The Causes of Evolution*.

Ronald A. Fisher développe en parallèle des travaux similaires, portant notamment sur l'évolution de la dominance, la sélection sexuelle et le mimétisme. En 1918, il publie un article sur la corrélation entre individus apparentés, étudiée par des méthodes statistiques basées sur les hypothèses de l'hérédité mendélienne. En 1922, il introduit l'utilisation des méthodes stochastiques en génétique des populations, pour étudier la fluctuation aléatoire des fréquences géniques. Il considérera par la suite que l'effet de la fluctuation aléatoire

peut être négligé puisqu'il est très faible pour les populations de grande taille que sont la majorité des espèces. Il introduit également dans son papier de 1922 la notion de superdominance qui aura une grande importance par la suite : si la sélection favorise l'état hétérozygote, les deux allèles sont maintenus dans la population. Fisher a eu une grande influence sur la conception de l'évolution selon laquelle la vitesse et la direction de l'évolution sont quasiment exclusivement déterminées par la sélection naturelle. C'est cette vision qui va être dominante par la suite. Il résume ses travaux en 1930 dans son livre *The Genetical Theory of Natural Selection*.

Haldane et Fisher étudient des populations de grande taille : ils ne tiennent pas compte de l'échantillonnage aléatoire des individus et des gamètes au moment de la reproduction, leur traitement des changements de fréquences alléliques est déterministe. À l'inverse, Sewall Wright met l'accent sur l'importance de l'échantillonnage aléatoire en définissant la *dérive aléatoire*, c'est à dire la fluctuation des fréquences alléliques dans une petite population, due à l'échantillonnage aléatoire. Son article *Evolution in Mendelian populations* paraît en 1931.

Haldane, Fisher et Wright sont considérés comme les trois pères de la génétique des populations, dont ils ont quasiment achevé de développer l'essentiel de la théorie mathématique dès le début des années 1930.

## **Théorie Synthétique de l'Évolution**

Les avancées en génétique des populations combinées à la théorie chromosomique de l'hérédité aboutissent à la fin des années 1930 à la Théorie Synthétique de l'Évolution, qui propose une vision unifiée de la génétique, la biologie naturaliste et la paléontologie pour expliquer l'origine des espèces. Elle est principalement due à Theodosius Dobzhansky (*Genetics and the Origin of Species* publié en 1937), Ernst Mayr (*Systematics and the Origin of Species* publié en 1942), et Julian Huxley (*Evolution, the Modern Synthesis* publié en 1942).

Hervé Le Guyader propose un résumé en sept points des idées majeures de cette Théorie Synthétique :

- Théorie chromosomique de l'hérédité : l'hérédité est exclusivement génétique, les caractères hérités des parents interagissent mais ne se mélangent pas. L'hérédité est portée par les gènes, il n'y a pas d'hérédité des caractères acquis.
- La mutation est à l'origine d'une grande variabilité des populations naturelles.
- L'évolution se fait à l'échelle de populations qui peuvent échanger des gènes via la migration, ou spéciation dans le cas d'un isolement géographique.

- Les populations évoluent de façon graduelle par modifications de faible amplitude.
- La sélection naturelle est la force évolutive majeure expliquant les changements au sein des populations.
- La majorité des différences observées entre les individus sont des adaptations résultant de la sélection naturelle, positive ou négative.
- Les observations à l'échelle macro-évolutive (c'est à dire à l'échelle des espèces, des phylums, etc.) sont la résultante des processus micro-évolutifs (à l'échelle de quelques générations) qui ont contrôlé l'évolution des populations pendant une grande période de temps.

On peut noter l'importance donnée à la sélection dans la théorie synthétique, tandis que la dérive aléatoire, étudiée par Wright, n'est pas mentionnée, et ce malgré les collaborations étroites entre Wright et Dobzhansky, l'un des fondateurs de la synthèse.

### 1.1.2 La querelle du polymorphisme

Cette section s'appuie principalement sur le chapitre 3 de l'ouvrage *Les avatars du gène : la théorie néodarwinienne de l'évolution*, coécrit par Pierre-Henri Gouyon, Jean-Pierre Henry et Jacques Arnould et publié en 1997, ainsi que sur le chapitre 2 du livre de Motoo Kimura, *The Neutral Theory of Molecular Evolution*, publié en 1983.

#### Vision orthodoxe au milieu du XX<sup>ème</sup> siècle

Le polymorphisme est défini par Edmund B. Ford en 1940 comme étant la coexistence, dans une population, de deux formes discontinues (ou plus) dans des proportions telles que la plus rare ne peut être maintenue par le seul effet d'une mutation récurrente. À l'échelle génétique, il se définit comme la coexistence de plusieurs allèles au même locus.

La définition de Ford exclut à dessein le cryptopolymorphisme, correspondant aux allèles létaux, qui par définition ne se transmettent pas car les individus qui les portent ne peuvent pas se reproduire. Ces allèles apparaissent par mutation et sont éliminés par la sélection, ils sont présents dans la population à de très faibles fréquences. C'est le cas d'un grand nombre de maladies génétiques graves, comme la mucoviscidose.

En 1930, Fisher démontre que plus il y a de variation génétique dans une population, plus elle évolue rapidement par sélection. Cette relation est vérifiée expérimentalement en 1964 par Francisco J. Ayala sur des populations de drosophiles plus ou moins variables génétiquement. Deux modèles vont s'affronter pour expliquer cette variabilité génétique, et donc la capacité héréditaire des populations à évoluer. Notons que cette « querelle

du polymorphisme» est parfois présentée du point de vue de l'hétérozygotie : quelle est la proportion de locus hétérozygotes ?

D'un côté, le modèle «classique» proposé par Hermann J. Muller, postule que les populations ont une faible diversité génétique. Les gènes sont majoritairement sous forme d'un allèle «naturel» à fréquence proche de 1, les autres allèles, délétères, sont maintenus à faible fréquence. Ce modèle découle du constat que la majorité des mutations observées lors d'expériences sont délétères. Les mutations bénéfiques sont néanmoins possibles : après son apparition, un allèle bénéfique envahirait rapidement la population et remplacerait l'allèle «naturel» initial. Dans ce modèle, les loci hétérozygotes sont minoritaires.

D'un autre côté, le modèle «équilibré» est défendu par Dobzhansky et Ford. Les populations ont une grande diversité génétique, il n'y a pas un allèle majoritaire à chaque locus dans toutes les populations. Un allèle majoritaire dans une population peut être minoritaire ailleurs. Les différentes formes de sélection naturelle maintiennent cette diversité. Dans ce modèle, un grand nombre de loci sont à l'état hétérozygote. À un locus, l'état homozygote a une moins bonne valeur sélective que l'état hétérozygote.

Le débat entre ces deux visions de la variabilité génétique, qui entraîne de vives altercations entre ses protagonistes, va se trouver éclairé par les données issues des nouvelles méthodes de biologie moléculaire dans les années 1960.

L'ère de la biologie moléculaire commence en 1953 avec la découverte de la structure de la molécule d'ADN (Acide Désoxyribonucléique) par James Watson, Francis Crick, Maurice Wilkins et Rosalind Franklin (Watson and Crick, 1953). Le dogme central de la biologie moléculaire permet d'établir que toute variation dans la séquence protéique découle d'une variation dans la séquence d'ADN qui code cette protéine. Ainsi, on peut étudier la variabilité génétique en étudiant la variabilité des séquences protéiques : cela va être rendu possible grâce à la mise au point de l'électrophorèse sur gel d'amidon ou d'acrylamide. Cette technique permet d'identifier des enzymes codées par différents allèles d'un gène, mais qui ont conservé leur activité enzymatique : les allozymes. En effet, si ils diffèrent d'un ou plusieurs acides aminés, ces allozymes diffèrent en structure moléculaire et en charge : leur migration sur le gel sera donc différente. En 1963, Jack L. Hubby publie une étude par électrophorèse de variabilité protéique chez la *Drosophile*. Peu après, il collabore avec Richard Lewontin et ils appliquent sa méthode pour mesurer la proportion de loci hétérozygotes dans les populations naturelles. Ils publient deux articles en 1968, montrant que le niveau d'hétérozygoté est d'en moyenne 12% par locus et la proportion de polymorphisme est 30% pour 18 loci de *Drosophila pseudoobscura*.

C'est donc le modèle «équilibré» de Dobzhansky qui sort vainqueur de la querelle : on détecte du polymorphisme dans les populations, il y a donc plusieurs allèles présents à des fréquences du même ordre de grandeur, et non un allèle prédominant.

Pour expliquer comment ce polymorphisme est maintenu dans la population, l'hypothèse prédominante, qui découle de la théorie synthétique et est défendue par Dobzhansky, est la sélection, sous plusieurs formes. Plusieurs mécanismes sont proposés, qui coexistent pour expliquer le polymorphisme :

- la superdominance : l'état hétérozygote a une meilleure valeur sélective que les états homozygotes. L'exemple bien connu de superdominance est celui de la drépanocytose, ou anémie falciforme, dans les populations exposées au paludisme. Cependant les exemples de superdominance sont rares et le rôle de cette force de sélection pour expliquer le polymorphisme reste controversé. De plus, cela ne permettait pas d'expliquer les hauts niveaux de polymorphisme observés chez les haploïdes.
- la sélection fréquence-dépendante : la valeur sélective d'un génotype varie avec sa fréquence dans la population. On peut citer l'exemple d'un génotype qui permettrait d'exploiter une ressource du milieu qui n'est pas exploitée par les autres individus de la même espèce : ce génotype est favorisé tant qu'il est rare, s'il devient fréquent et que tous les individus utilisent cette ressource, il ne présente plus d'avantage sélectif.
- la sélection dépendante du temps et de l'espace : la valeur sélective d'un génotype varie dans le temps et dans l'espace, en raison de l'hétérogénéité temporelle et spatiale du milieu.
- le polymorphisme transitoire : après l'apparition d'un nouvel allèle bénéfique dans une population, une période de transition s'établit jusqu'au remplacement total de l'ancien allèle, si celui-ci est moins bénéfique que le nouveau. Si le nouvel allèle n'est pas perdu par hasard dans les premières générations pendant lesquelles il est en très faible fréquence, il pourra ensuite envahir la population, plus ou moins rapidement selon que le gain d'avantage sélectif est plus ou moins important par rapport à l'ancien allèle. On observe donc un polymorphisme pendant cette période, qualifié de transitoire.

Ainsi, au début des années 1960, un consensus est atteint selon lequel tous les caractères biologiques peuvent être interprétés grâce à l'évolution adaptative par sélection naturelle, vision défendue notamment par Ernst Mayr et appelée pan-sélectionnisme. La possible neutralité du point de vue de la sélection de certains gènes ou de polymorphismes est fortement rejetée. Mayr suggère même d'éviter de se référer à la dérive génétique comme

cause d'évolution, pour clarifier les discussions. Dans cette vision pan-sélectionniste, il faut noter que le rôle de la mutation était considéré comme mineur, la variabilité génétique entretenue par la recombinaison étant suffisante pour que l'évolution puisse agir même si la pression de mutation est très faible voire nulle.

### **Nouvelles données moléculaires aboutissant à l'émergence de la théorie neutraliste**

À la même époque et à l'encontre de la tendance qui est plutôt aux arguments verbaux, la théorie mathématique de la génétique des populations s'étoffe. À partir de 1964, Motoo Kimura développe l'utilisation des équations de diffusion en génétique des populations. Cela permet d'étudier le comportement d'allèles mutants en tenant compte de la dérive, en plus des changements déterministes dus à la mutation et à la sélection. Il obtient par exemple grâce à cette méthode la probabilité de fixation d'un allèle mutant, dans une population de taille finie, en fonction de son avantage sélectif.

L'avènement de la biologie moléculaire apporte deux informations majeures qui vont aboutir à l'émergence de la théorie neutraliste. Comme on l'a vu, les techniques d'électrophorèse ont révélé les niveaux importants de variabilité protéique entre les individus d'une espèce, ce qui a permis d'estimer la variabilité génétique. Parmi les autres avancées permises par l'arrivée de la biologie moléculaire, il est devenu possible de comparer les séquences d'acides aminés de protéines chez différents organismes apparentés, comme par exemple l'hémoglobine chez différents vertébrés. Ces comparaisons, en parallèle avec les données paléontologiques, ont permis d'estimer les taux de substitutions (c'est-à-dire de mutations fixées) d'acides aminés, et donc de nucléotides dans les gènes (Zuckerkandl and Pauling, 1965). Ces auteurs ont montré que ces taux étaient constants sur les différentes lignées et ont formulé l'hypothèse de l'horloge moléculaire. Selon cette hypothèse, les mutations s'accumulent à une vitesse constante dans les génomes. La notion que les séquences protéiques évoluent indépendamment des forces de sélection déterminées par l'environnement semblait très contradictoire à l'orthodoxie de l'époque, qui voulait que la sélection soit à l'origine de toute la variabilité.

Les observations qui mettent en doute la vision orthodoxe du pan-sélectionnisme sont donc les suivantes :

- la quasi-uniformité des taux de substitutions d'acides aminés par an (horloge moléculaire)
- le caractère aléatoire des types de substitutions observés
- le taux élevé de polymorphisme

Parmi les arguments qui remettent en cause le rôle de la sélection dans le maintien du polymorphisme figure également celui du fardeau génétique. Cette notion a été proposée en 1950 par Muller. Elle correspond à la différence de la valeur sélective d'une population par rapport à l'optimum, due au maintien d'individus ayant des valeurs sélectives inférieures à celle du meilleur individu de la population. Lorsque Hubby et Lewontin mettent en évidence les taux élevés de polymorphisme dans les populations naturelles, ils analysent ce polymorphisme important comme un fardeau : la population est composée d'individus ayant des valeurs sélectives différentes, elle n'est donc pas optimale. Kimura montre que le taux élevé de substitution implique un fardeau génétique incompatible avec la survie des populations. En effet, plus le nombre de gènes sous sélection est élevée, plus le fardeau génétique est élevé, et moins il y a d'individus qui ont accès à la reproduction.

Une solution au problème du fardeau génétique et à l'apparente contradiction de l'horloge moléculaire est de considérer que les différences observées dans les populations naturelles, le polymorphisme, n'entraînent pas de différence de valeur sélective entre les individus. On parle alors de polymorphisme neutre du point de vue de la sélection. C'est ce que propose la théorie neutraliste formulée en 1968 par Motoo Kimura et Tomoko Ohta, que nous allons détailler dans la section 1.2. Cette théorie est soutenue l'année suivante par Jack Lester King et Thomas H. Jukes, qui, arrivés indépendamment à la même conclusion que Kimura, publient un papier intitulé *Non-Darwinian evolution*.

La théorie neutraliste provoque un intense débat entre sélectionnistes et neutralistes. Étant donné que la théorie neutraliste est aujourd'hui largement acceptée, nous ne nous attarderons pas sur ces controverses passées qui ont été progressivement effacées par l'accumulation des données soutenant la théorie neutraliste.

Quinze ans plus tard, Kimura résume ainsi le contexte de la proposition de cette théorie, dans le premier paragraphe de la préface de son ouvrage *The Neutral Theory of Molecular Evolution* :

« Ce livre représente ma tentative pour convaincre le monde scientifique que la cause principale de changement évolutif à l'échelle moléculaire — les changements dans le matériel génétique — est la fixation aléatoire de mutants neutres ou quasi-neutres du point de vue de la sélection, et non la sélection darwinienne positive. Cette thèse, que j'appelle ici la théorie neutraliste de l'évolution moléculaire, a causé un grand nombre de controverses depuis que je l'ai proposée en 1968 pour expliquer certaines nouvelles découvertes en évolution et de variabilité à l'échelle moléculaire. La controverse n'est pas surprenante, puisque la biologie évolutive a été dominée depuis plus d'un demi-siècle



par la théorie darwinienne, selon laquelle les organismes deviennent progressivement adaptés à leur environnement en accumulant des mutants bénéfiques, et les évolutionnistes s'attendaient naturellement à ce que ce principe s'étende à l'échelle moléculaire. La théorie neutraliste n'est pas antagoniste à la vision si appréciée de l'évolution des formes et des fonctions guidée par la sélection darwinienne, mais elle souligne une autre facette du processus évolutif en insistant sur le plus grand rôle de la pression de mutation et de la dérive aléatoire à l'échelle moléculaire. »

**Motoo Kimura** (木村 資生) est un généticien des populations japonais, né en 1924 et mort en 1994. Il débute sa carrière scientifique en botanique, en étudiant la structure chromosomique des Liliacées, ce qui l'amène à connaître les bases de la génétique des populations. Dans son premier article de génétique des populations, il décrit le modèle «stepping-stone» de structuration de population pour raffiner les modèles en îles de Wright (Kimura, 1953). Il obtient sa thèse en 1956, sous la direction de James F. Crow à l'Université du Wisconsin, avant de retourner au Japon, à l'Institut National de Génétique, où il restera tout le reste de sa carrière. Pendant sa thèse, il développe un modèle général de dérive génétique, qui tient compte de la sélection, de la migration et des mutations. Il a aussi introduit l'équation de Kolmogorov backward en génétique des populations, ce qui lui a permis de calculer la probabilité de fixation d'un gène dans une population. Il est à l'origine des modèles d'allèles infinis, de sites infinis et de mutation «stepwise». Un premier compte-rendu de ses approches est publié en 1960, dans son livre *An Introduction to Population Genetics*. En 1968, il propose la théorie neutraliste de l'évolution moléculaire, qu'il passera ensuite toute sa vie à développer et à défendre. Il détaille sa théorie en 1983 dans son ouvrage *The Neutral Theory of Molecular Evolution*, et publie également des ouvrages de vulgarisation, comme *My Views on Evolution*, qui sera un best-seller au Japon. Il a reçu entre autres récompenses la médaille Darwin de la Royal Society en 1992.



## 1.2 Théorie neutraliste de l'évolution moléculaire

La théorie neutraliste de l'évolution est formulée en 1968 par Motoo Kimura. Il la résume ainsi dans le premier paragraphe de l'introduction de son ouvrage de 1983, *The Neutral Theory of Molecular Evolution* :

« La théorie neutraliste affirme que la grande majorité des changements évolutifs à l'échelle moléculaire, révélés par les études comparatives de protéines et de séquences d'ADN, sont causés non pas par la sélection darwinienne mais par la dérive aléatoire de mutants neutres ou quasi-neutres du point de vue de la sélection. Cette théorie ne nie pas le rôle de la sélection naturelle dans la détermination du chemin de l'évolution adaptative, mais elle suppose que seule une infime fraction des changements de l'ADN dans l'évolution sont adaptatifs dans la nature, tandis que la grande majorité des substitutions moléculaires sans effet sur le phénotype n'exercent aucune influence significative sur la survie et la reproduction, et dérivent aléatoirement dans l'espèce.

La théorie neutraliste affirme également qu'une grande partie de la variabilité intraspécifique à l'échelle moléculaire, qui se manifeste par exemple sous forme de polymorphisme protéique, est essentiellement neutre, si bien que la majorité des allèles polymorphes sont maintenus dans les espèces par pression de mutation et extinction aléatoire. En d'autres termes, la théorie neutraliste voit le polymorphisme protéique et de l'ADN comme une phase transitoire de l'évolution moléculaire, et rejette la notion que la majorité de ces polymorphismes sont adaptatifs et maintenus dans l'espèce par une forme de sélection balancée. »

### 1.2.1 Hypothèses et paramètres

#### Neutralité

La neutralité évoquée par la théorie ne doit pas être prise au sens strict, d'ailleurs Kimura parle souvent de mutants neutres *ou quasi-neutres* du point de vue de la sélection. En fait, l'essence de la théorie se trouve dans le fait que les facteurs principaux de l'évolution sont la mutation et la dérive aléatoire. Les mutants doivent donc être *suffisamment neutres* pour que le hasard joue le rôle principal. Kimura propose d'ailleurs dans son ouvrage de 1983 que la théorie soit plutôt nommée « Théorie de la dérive aléatoire des mutations » mais estime que l'appellation de théorie neutraliste est déjà très répandue et qu'il ne faut pas « changer de cheval au milieu du gué ».

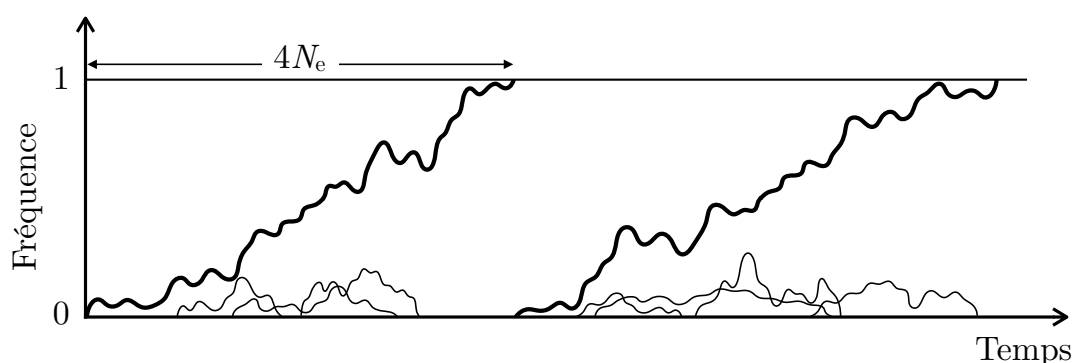


FIGURE 1.1 – Évolution de la fréquence de mutations dans une population finie de taille efficace  $N_e$  (d'après Kimura, 1983). Les trajectoires représentées en gras aboutissent à la fixation de l'allèle dans la population. Les autres trajectoires aboutissent à l'extinction du nouvel allèle.

### Dérive (génétique) aléatoire

Comme on l'a vu, l'essentiel de la théorie neutraliste n'est pas tant la neutralité de la majorité des mutations que le fait que l'évolution de leur fréquence est déterminée par la dérive aléatoire. La dérive correspond, dans une population d'effectif limité, à la fluctuation aléatoire de la fréquence des mutations au fil des générations, liée à l'effet d'échantillonnage au moment de la reproduction (Figure 1.1). Cet échantillonnage a lieu à deux niveaux. Au moment de la reproduction, tous les individus ne participent pas à la production de la génération suivante. De plus, pour un individu hétérozygote à certains de ses loci, les gamètes ne seront pas tous identiques : ainsi, certains de ses allèles ne seront pas transmis à sa descendance. Il y a donc échantillonnage des individus et des gamètes. Un mutant présent en faible fréquence peut ainsi être perdu si les individus ou les gamètes dans lesquels il est présent ne sont pas échantillonnés. Plus la population est de petite taille, plus l'effet de cet échantillonnage aléatoire sera important.

Parmi les nombreuses mutations qui apparaissent à chaque génération dans une grande population, la majorité est perdue par hasard au cours des premières générations. Cela concerne aussi bien les mutations désavantageuses que les neutres et les avantageuses (sauf si l'avantage sélectif est très important). La fixation d'un mutant neutre prend en moyenne  $4N_e$  générations pour une espèce diploïde (Kimura and Ohta, 1969) où  $N_e$  est la taille efficace de la population, définie dans la section suivante.

## Taille efficace de population

La taille efficace de population, couramment notée  $N_e$ , a été introduite en 1931 par Wright, dans le but de tenir compte des différents facteurs qui font qu’une population réelle diffère d’une population théorique idéale. En effet, dans une population réelle, les fluctuations de taille de population, la différence entre le nombre de mâles et le nombre de femelles ou encore la structuration de la population ont un effet sur les fréquences alléliques ou les taux de fixation. La taille efficace de population est donc le nombre d’individus d’une population théorique idéale ayant la même intensité de dérive génétique que la population réelle étudiée.

On peut définir la taille efficace de multiples manières : par exemple, si une population est constituée de  $N_m$  mâles reproducteurs et de  $N_f$  femelles reproductrices, la taille efficace de la population est :

$$N_e = \frac{4N_m N_f}{N_m + N_f} \quad (1.1)$$

(Wright, 1931). On voit que si  $N_m$  et  $N_f$  sont très différents,  $N_e$  sera principalement déterminé par le plus petit effectif : si par exemple le nombre de femelles  $N_f$  est très grand mais qu’il n’y a qu’un seul mâle reproducteur ( $N_m = 1$ ), on aura  $N_e \approx 4$ .

Une autre définition de la taille efficace est la suivante : si la taille de la population change de façon cyclique avec une courte période de  $g$  générations,

$$N_e = \frac{g}{\sum_{i=1}^g 1/N_i} \quad (1.2)$$

ce qui correspond à la moyenne harmonique du nombre d’individus pendant un cycle (Wright, 1938). Dans ce cas,  $N_e$  est principalement déterminé par les phases du cycle pendant lesquelles  $N_i$  est petit.

On compte encore d’autres définitions possibles (voir section 4 du chapitre 3 de l’ouvrage de Kimura) et autant de manières d’estimer ce paramètre central des modèles basés sur la théorie neutre. Nous reviendrons dans la section 1.3.3 sur les incohérences que cela peut engendrer.

### 1.2.2 Outils mathématiques

Pour réaliser une étude théorique de génétique de populations, on utilise des modèles qui tentent de capturer les caractéristiques biologiques fondamentales de l’évolution d’une population, tout en restant assez simples pour pouvoir être décrits et utilisés mathématiquement. La théorie neutraliste est un cadre théorique, mais ce n’est pas un outil ou un modèle descriptif, c’est un ensemble d’hypothèses. Un certain nombre d’outils, développés

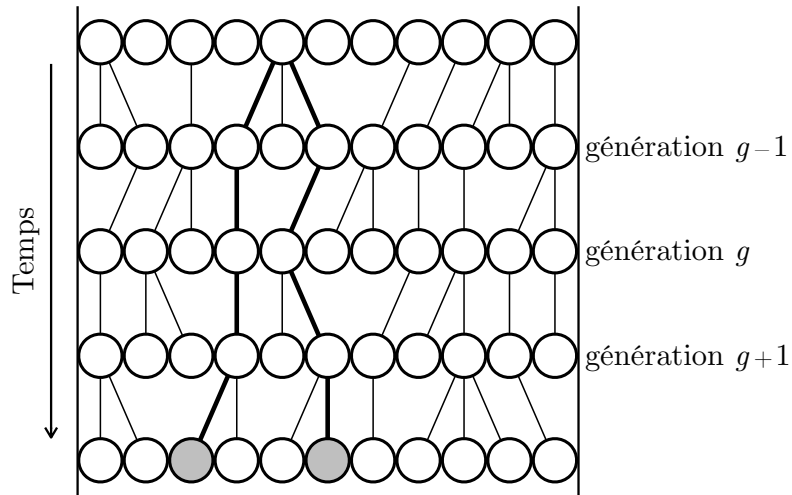


FIGURE 1.2 – Modèle de Wright-Fisher pour une population de taille  $N=11$ . Les individus, haploïdes, sont représentés par des cercles. 5 générations discrètes sont représentées, et les traits qui relient les individus sont des liens de parenté. Les deux individus colorés en gris à la dernière génération ont un ancêtre commun 4 générations auparavant (lignées en gras).

avant ou après l'avènement de la théorie neutraliste, sont aujourd'hui couramment utilisés dans ce cadre théorique. J'en présente ici quelques uns des principaux, que j'ai utilisés pendant ma thèse, en me basant principalement sur les chapitres 3 et 4 de l'ouvrage de John Wakeley, *Coalescent Theory, An Introduction*, publié en 2009.

### Modèle de Wright-Fisher

Un des modèles de génétique des populations les plus utilisés est celui développé par Fisher (1930) et Wright (1931). On considère une population de taille constante  $N$ . À chaque génération, tous les individus de la population meurent et sont remplacés par leur descendance, on parle de générations discrètes (non-chevauchantes). La descendance est un échantillonnage aléatoire avec remise de la génération actuelle. Une représentation graphique en est donnée dans la Figure 1.2. Comme la taille de population  $N$  est finie et qu'un même individu peut donner plusieurs descendants, tous les individus ne participent pas forcément à la génération suivante, ce qui cause la dérive aléatoire (voir section 1.2.1). Le modèle de Wright-Fisher s'applique aussi aux organismes diploïdes, en supposant qu'il n'y a qu'un seul type reproductif (pas de mâles et de femelles mais uniquement des individus hermaphrodites).

Prenons une population dans laquelle deux allèles  $A$  et  $a$  ségrègent. L'allèle  $A$  est présent en  $i$  copies et l'allèle  $a$  en  $N - i$  copies où  $N$  est la taille de la population et

$i \in [1, N - 1]$ . La fréquence de l'allèle  $A$  à cette génération est donc  $p = i/N$  et la fréquence de l'allèle  $a$  est  $1 - p$ . Soit  $P_{ij}$  la probabilité qu'un allèle présent en  $i$  copies à la génération actuelle soit présent en  $j$  copies à la génération suivante. Si les allèles  $A$  et  $a$  ont la même valeur sélective, que la population n'est pas subdivisée et qu'il n'y a pas de mutations, on a :

$$P_{ij} = \binom{N}{j} p^j (1 - p)^{N-j} \quad \text{où } 0 \leq j \leq N \quad (1.3)$$

Le nombre de copies de l'allèle  $A$  à la génération suivante, noté  $K$ , suit une loi binomiale de paramètres  $N$  et  $p$ . On a donc :

$$E[K] = Np = i \quad \text{et} \quad \text{Var}[K] = Np(1 - p) \quad (1.4)$$

Si on note  $\Delta p$  la différence de fréquence d'un allèle entre deux générations, on a  $E[\Delta p] = 0$  et  $\text{Var}[\Delta p] = p(1 - p)/N$  : on s'attend à ce que le nombre de copies de  $A$  reste constant en moyenne, mais il peut en fait prendre toutes les valeurs entre 0 et  $N$ . Au cours du temps, la fréquence de l'allèle  $A$  va dériver selon une chaîne de Markov avec probabilités de transition  $P_{ij}$ . Du fait de l'équation 1.4, on dit que dans le modèle de Wright-Fisher, la dérive est de l'ordre de  $1/N$  par génération.

Il existe d'autres modèles en génétique des populations, comme le modèle de Moran (1958), dans lequel une génération correspond à la mort et au remplacement d'un seul individu.

## Coalescent de Kingman

Le coalescent est l'arbre des lignées ancestrales d'un ensemble d'individus, jusqu'à leur ancêtre commun le plus récent (MRCA, most recent common ancestor). John Kingman a montré que c'est le processus ancestral limite d'un grand nombre de modèles de génétique des populations, parmi lesquels le modèle de Wright-Fisher et le modèle de Moran (Kingman, 1982a,b). Dans l'approche coalescente, en génétique des populations, on remonte le temps du présent vers le passé, à l'inverse de ce qu'on a pu voir au paragraphe précédent avec le modèle de Wright-Fisher dans lequel on considère le temps prospectif, du passé vers le présent.

Le coalescent d'un échantillon de  $n$  individus est constitué de  $n - 1$  événements de coalescence (Figure 1.3). À chaque coalescence, le nombre de lignées ( $n$  au temps présent) diminue de 1. Au dernier événement de coalescence, les deux dernières lignées coalescent : c'est l'ancêtre commun le plus récent de l'échantillon. On nomme  $T_i$  le temps pendant

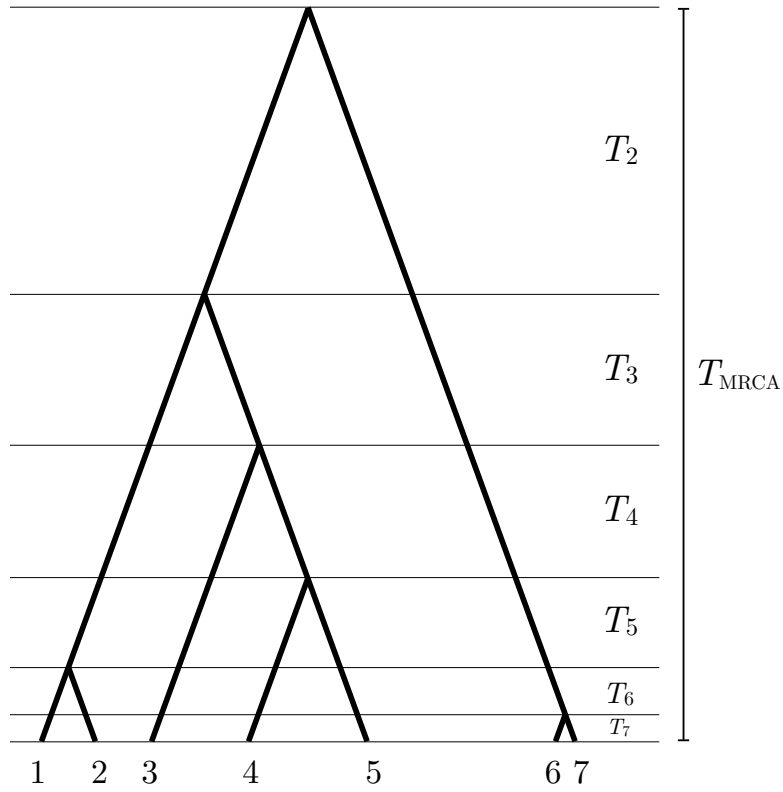


FIGURE 1.3 – Coalescent d'un échantillon de 7 individus

lequel il y a  $i$  lignées ancestrales de l'échantillon. Le temps d'atteinte de l'ancêtre commun le plus récent s'exprime donc :

$$T_{\text{MRCA}} = \sum_{i=2}^n T_i \quad (1.5)$$

Les  $T_i$  sont indépendants. On peut également caractériser l'arbre par la longueur totale de ses branches, qui s'exprime ainsi :

$$T_{\text{total}} = \sum_{i=2}^n iT_i \quad (1.6)$$

Le coalescent a des propriétés de linéarité avec la taille de population  $N$ , on exprime la plupart du temps le temps en unités de  $N$  générations : c'est ce qu'on appelle l'unité de temps coalescente.

Sous les hypothèses de neutralité, de taille constante de population et de panmixie, on peut décrire un certain nombre de distributions de probabilités concernant le coalescent. Les hypothèses de neutralité et de panmixie permettent de considérer que le nombre de descendants d'un individu ne dépend pas de l'individu : les individus sont tous équivalents et leurs nombres de descendants sont des variables aléatoires échangeables. Elles

sont identiquement distribuées, mais pas indépendantes du fait de l'hypothèse de taille constante de la population.

Kingman a montré que, quand la taille de population  $N$  tend vers l'infini, les temps  $T_i$ , exprimés en unités de temps coalescentes, suivent une loi exponentielle de paramètre  $\binom{i}{2}$  pour  $i = 2, \dots, n$ . On a ainsi des temps de coalescence tels que :

$$E[T_i] = \frac{2}{i(i-1)} \quad \text{et} \quad \text{Var}[T_i] = \left( \frac{2}{i(i-1)} \right)^2 \quad (1.7)$$

Plus on remonte dans le temps, c'est-à-dire plus  $i$  est petit, plus les temps moyens de coalescence sont longs et plus la variance augmente. Grâce aux équations 1.5 et 1.6, on a :

$$E[T_{\text{MRCA}}] = 2 \left( 1 - \frac{1}{n} \right) \quad \text{et} \quad E[T_{\text{total}}] = 2 \sum_{i=1}^{n-1} \frac{1}{i} \quad (1.8)$$

Ainsi,  $E[T_{\text{MRCA}}]$  converge vers 2 quand la taille de l'échantillon  $n$  tend vers l'infini, tandis que  $E[T_{\text{total}}]$  tend vers l'infini (Watterson, 1975; Hudson et al., 1990; Tajima, 1993; Tavaré et al., 1997).

## 1.3 Utilisation du modèle neutre en évolution moléculaire

### 1.3.1 Description de la diversité génétique

Pour étudier les séquences d'ADN polymorphes d'un échantillon d'individus, on ajoute un processus de mutation au coalescent de Kingman, qui permet ainsi de rendre compte de l'évolution de séquences d'ADN. Les événements de coalescence représentent toujours la parenté entre les individus, mais des mutations peuvent maintenant se produire le long des branches de l'arbre. On considère uniquement des mutations neutres, qui n'affectent donc pas la généalogie.

Pour une généalogie longue et une probabilité de mutation faible et constante, le nombre de mutations qui surviennent sur une branche de longueur donnée sera approximé par une distribution poissonnienne de paramètre égal au nombre attendu de mutations pendant ce temps donné. On définit  $\theta = 2N\mu$  comme étant le double du nombre moyen de mutations introduites dans la population de taille  $N$  à chaque génération,  $\mu$  étant le taux de mutation par génération et par site. Le coefficient 2 a été introduit par les généticiens des populations pour simplifier les calculs. Le paramètre  $\theta$  peut aussi se définir comme



le double du nombre moyen de mutations le long d'une lignée d'une unité de temps coalescent.

Le long d'une généalogie de longueur  $t$ , le nombre de mutations  $K$  suit donc une loi de Poisson de paramètre  $\theta t/2$  :

$$P(K = k | t) = \frac{\left(\frac{\theta t}{2}\right)^k}{k!} e^{-\frac{\theta t}{2}} \quad \text{où } k = 0, 1, 2, \dots, \quad (1.9)$$

Et on a donc

$$E[K | t] = \text{Var}[K | t] = \frac{\theta t}{2} \quad (1.10)$$

On peut ainsi calculer l'espérance du nombre de sites polymorphes dans un échantillon de taille  $n$ , noté  $S$  (Watterson, 1975) :

$$E[S] = E[K]E[T_{\text{total}}] = \left(\frac{\theta}{2}\right) \left(2 \sum_{i=1}^{n-1} \frac{1}{i}\right) = \theta \sum_{i=1}^{n-1} \frac{1}{i} \quad (1.11)$$

À partir d'un alignement de séquences, on peut donc mesurer le nombre de sites polymorphes  $S$  qui permettra d'estimer  $\theta$  : c'est l'estimateur dit de Watterson, noté  $\hat{\theta}_S$  (Watterson, 1975). On peut également estimer  $\theta$  à partir d'une autre mesure couramment utilisée en génétique des populations : le nombre moyen de différences entre paires de séquences dans un échantillon, noté  $\pi$ . On peut montrer que  $E[\pi] = \theta$ , ce qui permet d'estimer  $\hat{\theta}_\pi$  (Tajima, 1983). Ces estimateurs de  $\theta$  permettent de tester statistiquement la validité du modèle neutre à partir d'un échantillon. Ces tests de neutralité peuvent par exemple être basés sur la valeur  $D$  (Tajima, 1989), qui mesure la déviation par rapport au modèle de Wright-Fisher :

$$D = \frac{\hat{\theta}_\pi - \hat{\theta}_S}{\sqrt{\text{Var}[\hat{\theta}_\pi - \hat{\theta}_S]}} \quad (1.12)$$

Ces deux mesures de la diversité d'un échantillon,  $S$  et  $\pi$ , sont en fait des statistiques résumées d'une autre statistique, qui est celle que j'ai utilisée principalement pendant ma thèse : le spectre de fréquence, c'est-à-dire la distribution des fréquences alléliques dans l'échantillon. On considère les sites bi-alléliques, c'est-à-dire présents en deux allèles dans l'échantillon. On définit l'allèle ancestral comme celui présent initialement dans la population, et l'allèle dérivé comme celui apparu par mutation. Le spectre de fréquence est défini comme le vecteur  $\xi = (\xi_1, \xi_2, \dots, \xi_{n-1})$  où pour  $i \in [1, n-1]$ ,  $\xi_i$  est le nombre de sites pour lesquels l'allèle dérivé est présent en  $i$  copies, c'est-à-dire à fréquence  $i/n$  dans l'échantillon de taille  $n$ .

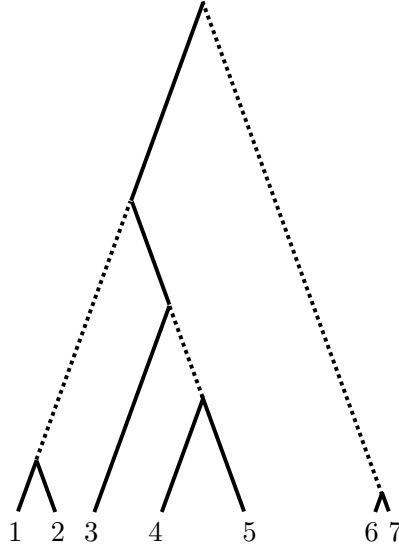


FIGURE 1.4 – Lignées ayant 2 descendants dans l'échantillon (en pointillés)

Lorsqu'on ne connaît pas l'allèle ancestral, on considère le spectre de fréquence plié, c'est-à-dire le vecteur  $\eta = (\eta_1, \eta_2, \dots, \eta_{[n/2]})$  où :

$$\eta_i = \frac{\xi_i + \xi_{n-i}}{1 + \delta_{i,n-i}} \quad \text{pour } 1 \leq i \leq [n/2]$$

où  $[n/2]$  est le plus grand entier plus petit ou égal à  $n/2$  (c'est-à-dire  $n/2$  quand  $n$  est pair et  $(n-1)/2$  quand  $n$  est impair) et  $\delta_{i,j}$  vaut 1 quand  $i = j$  et 0 sinon.

Les mutations qui contribuent au terme  $\xi_i$  du spectre de fréquence surviennent sur des branches du coalescent qui ont  $i$  descendants dans l'échantillon. Par exemple, dans la Figure 1.4, les lignées en pointillées ont chacune 2 descendants dans l'échantillon : les mutations qui surviennent sur ces branches seront donc présentes en 2 copies dans l'échantillon. Si on note  $\ell_i$  la longueur totale des branches qui ont  $i$  descendants dans l'échantillon, on a :

$$E[\xi_i] = \frac{\theta}{2} E[\ell_i] \quad (1.13)$$

Or on peut montrer que  $E[\ell_i] = 2/i$ , on connaît donc le spectre de fréquence moyen sous les hypothèses du modèle neutre (Fu, 1995) :

$$E[\xi_i] = \theta/i \quad \text{pour } i \in [1, n-1] \quad (1.14)$$

### 1.3.2 Inférences dans le cadre de la théorie neutraliste

L'intérêt d'utiliser le modèle neutre comme cadre théorique de référence en évolution moléculaire est la plupart du temps de chercher si les données étudiées permettent ou non

de rejeter ce modèle neutre, et pour quelles raisons. Le modèle neutre peut être rejeté lorsqu’une ou plusieurs de ses hypothèses ne sont pas respectées, c’est-à-dire lorsqu’on ne peut pas considérer que la majorité des sites sont neutres, ou que la taille de la population n’est pas constante, ou que la population n’est pas panmictique.

Le coalescent de Kingman est un outil très puissant pour ces analyses, car il peut être modifié pour tenir compte de certaines modifications d’hypothèses. On peut ainsi relaxer les hypothèses de taille constante de population (Watterson, 1984) et de générations non-chevauchantes (Tellier et al., 2011), et l’appliquer à une métapopulation constituée de plusieurs sous-populations (Wakeley and Aliacar, 2001). Pour cela, on procède en ré-échelonnant le temps pour tenir compte de la variabilité du taux de coalescence au cours du temps (Kaj and Krone, 2003).

Pendant cette thèse je me suis surtout intéressée aux inférences démographiques, c’est à dire aux études qui cherchent à expliquer les déviations des données observées par rapport au modèle neutre dues aux changements de taille de population. Je vais donc principalement détailler ce type d’inférence.

## **Effet de la démographie**

Pour comprendre comment des données génomiques peuvent nous renseigner sur l’histoire démographique passée d’une population, étudions l’effet qu’aurait une taille de population croissante sur le processus de coalescence.

Lorsqu’on s’intéresse au processus de coalescence, on considère le temps de façon rétrospective. Pour une population en croissance, rétrospectivement, la taille de la population diminue. Plus la taille de la population diminue, plus les événements de coalescence sont probables puisqu’il y a de moins en moins d’individus. Ainsi, dans une population en croissance, les événements de coalescence vont avoir lieu rétrospectivement plus rapidement, aboutissant à un arbre plus court. C’est ce qui est représenté dans la Figure 1.5, avec l’exemple d’une population en croissance linéaire (arbre central). Pour comparer les tailles relatives des branches des arbres obtenus dans une population à taille constante ou en croissance linéaire, on les remet à la même échelle (c’est-à-dire au même  $T_{\text{MRCA}}$ ) en les normalisant. En comparant les arbres à gauche et à droite de la Figure 1.5, on voit que dans l’arbre normalisé correspondant à la population en croissance (à droite), les branches terminales (ou récentes) sont plus longues relativement aux branches terminales de l’arbre à taille constante de population (à gauche). Inversement, les branches anciennes sont plus courtes, relativement aux branches anciennes de l’arbre à taille constante de population.

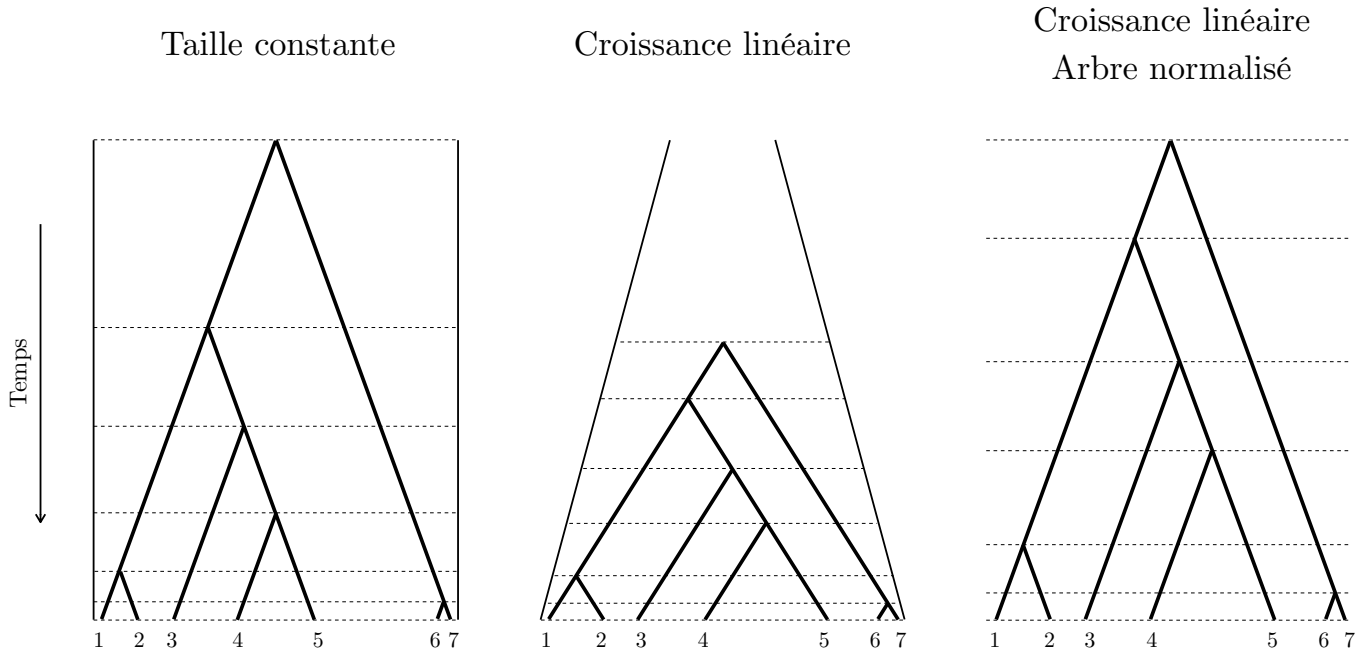


FIGURE 1.5 – Coalescents de Kingman pour des échantillons de taille  $n = 7$  dans des populations à taille constante (gauche) ou en croissance linéaire (centre). L’arbre de droite correspond au coalescent de la population en croissance linéaire, normalisé pour avoir le même  $T_{\text{MRCA}}$  que le coalescent de la population à taille constante.

Or on a vu que le nombre de mutations sur une branche dépendait de la longueur de celle-ci. En nombre absolu, l’arbre de coalescence dans la population en croissance étant plus court, on aura moins de mutations dans cette population que dans celle à taille constante (pour une taille actuelle égale). En normalisant le nombre de mutations, on aura relativement plus de mutations sur les branches terminales dans la population en croissance que dans la population à taille constante, car les branches terminales sont relativement plus longues. Par définition, les mutations qui arrivent sur les branches terminales sont portées par un unique individu de l’échantillon, donc dans le spectre de fréquence, cela se traduira par un excès de mutations à fréquence  $1/n$  où  $n$  est la taille de l’échantillon. Plus généralement, les mutations à basses fréquences seront en excès.

À l’inverse, pour une population en décroissance, les branches terminales (respectivement anciennes) seront relativement plus courtes (respectivement plus longues) que celles de la population à taille constante.

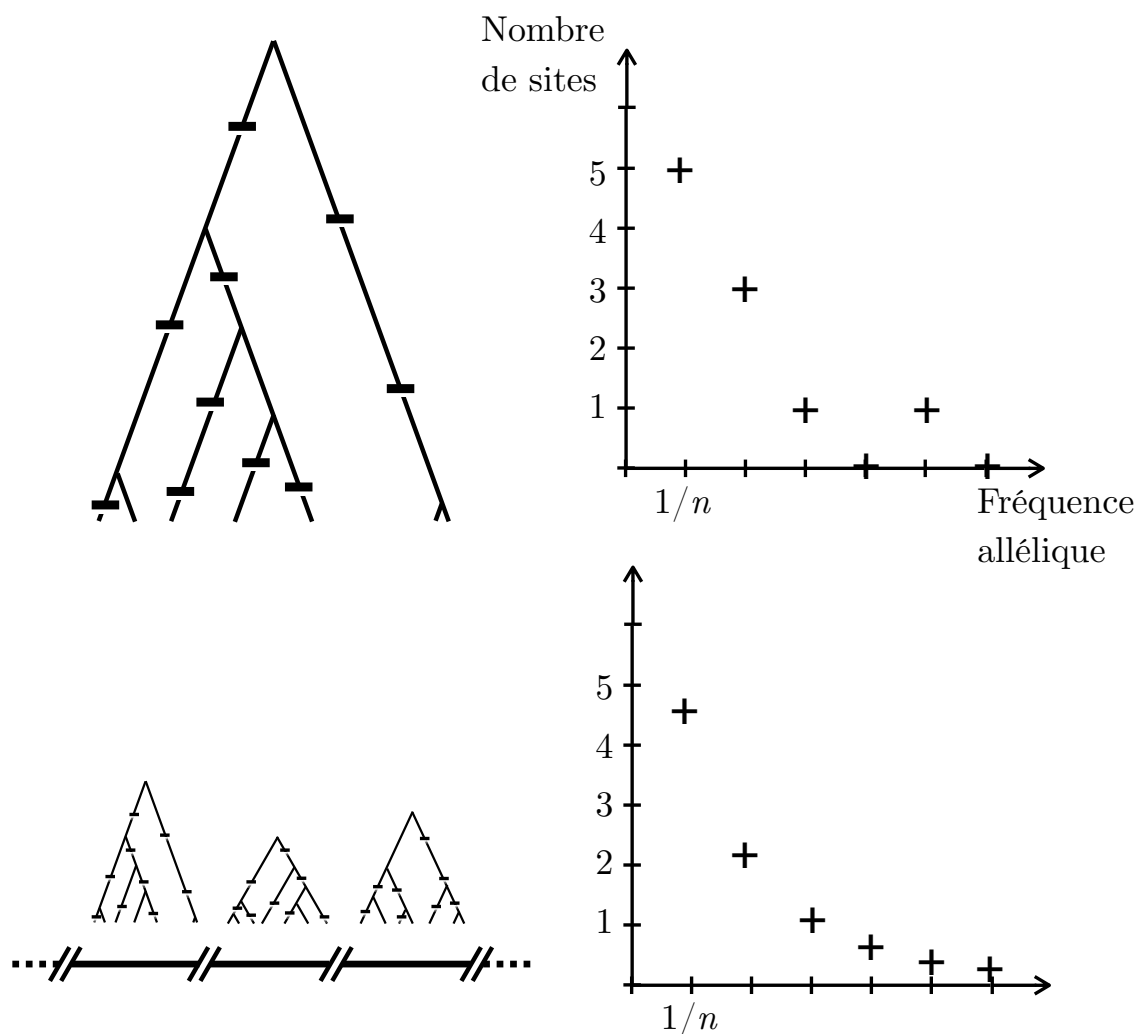


FIGURE 1.6 – En haut, spectre de fréquence correspondant à une généalogie avec mutations. En bas, spectre de fréquence moyen de l'ensemble des généalogies constituant un génome recombinant.

### Effet de la recombinaison

La Figure 1.6 donne un exemple de correspondance entre une généalogie avec des mutations, et le spectre de fréquence qui lui est associé. On voit ainsi que dans ce cas, 5 mutations se sont produites sur des branches externes, donc  $\xi_1 = 5$ . On voit que si les branches externes sont relativement plus longues, elles porteront davantage de mutations, et donc  $\xi_1$  sera plus important.

Du fait de la recombinaison, les génomes sont constitués d'un assemblage de loci qui ont chacun des histoires différentes : on parle de loci « pseudo-indépendants ». Ainsi, deux séquences de part et d'autre d'un événement de recombinaison ne vont pas avoir les mêmes

ancêtres, et donc pas le même arbre coalescent. Un génome entier est donc une collection d’histoires, plus ou moins indépendantes selon leur liaison. Le spectre de fréquence d’un échantillon de  $n$  génomes alignés reflète non pas un arbre coalescent (Figure 1.6 haut) mais un ensemble d’arbres coalescents (Figure 1.6 bas). Ainsi, le spectre de fréquence sera le reflet moyen de ces différents arbres.

## Inférence démographique

La théorie de la coalescence pouvant être adaptée pour tenir compte de changements de taille de population, de nombreuses méthodes ont été développées pour inférer ces changements. On peut les distinguer en fonction du type de données qu’elles analysent (arbre reconstruit, fréquences alléliques, paires de génomes alignés,...), et de leur méthodologie.

Certaines méthodes d’inférence démographique sont basées sur le modèle de Pairwise Sequentially Markovian Coalescent (McVean and Cardin, 2005; Marjoram and Wall, 2006), qui est une approximation du coalescent classique avec recombinaison (Hudson, 1983). Comme on l’a vu, du fait de la recombinaison, il n’y a pas un unique arbre coalescent pour tout le génome, mais plusieurs séparés par les événements de recombinaison le long du génome. Dans ces méthodes, les arbres coalescents sont supposés markoviens le long du génome. La vraisemblance d’histoires démographiques peut être calculée à partir de deux génomes seulement (ou d’un génome diploïde) (Li and Durbin, 2011; Sheehan et al., 2013; Schiffels and Durbin, 2014).

D’autres méthodes sont basées sur l’approche Approximate Bayesian Computation (ABC), qui permet de tester des scénarios complexes pour lesquels il est difficile de calculer une vraisemblance. On cherche à estimer un certain nombre de paramètres caractérisant le scénario. Pour cela, on définit une distribution préalable pour chaque paramètre, puis on effectue un grand nombre de simulations, couvrant l’ensemble de l’espace des paramètres. Un algorithme de rejet permet de ne garder que les simulations produisant des données proches des données observées, ce qui permet d’estimer la distribution postérieure des paramètres (Beaumont et al., 2002; Csilléry et al., 2010; Beaumont, 2010; Boitard et al., 2016).

Ces méthodes sont utilisées pour connaître l’histoire démographique d’une population, ce qui peut être une étape préliminaire lorsque l’on cherche à détecter de la sélection, afin de distinguer les signatures de ces deux processus (Akey et al., 2002; Goldstein and Chikhi, 2002).

## Sélection

La diversité génétique, telle qu'on l'a décrite dans la partie précédente à partir du coalescent de Kingman, est neutre : les mutations n'affectaient pas la généalogie sur laquelle elles se produisaient. De même que la démographie, la sélection va donc avoir un effet sur la coalescence des individus. Pour comprendre intuitivement comment la sélection va modifier la généalogie, prenons l'exemple d'une mutation avantageuse qui envahit rapidement la population (on parle de balayage sélectif). À la fin du balayage sélectif, tous les individus portent la mutation bénéfique : à ce locus, leur ancêtre commun remonte donc à peu de générations, à l'individu chez qui la mutation bénéfique est apparue. Les branches anciennes seront donc courtes, et les branches terminales comparativement plus longues : on se retrouve dans une situation équivalente à de la croissance démographique.

Il est ainsi connu que la sélection et la démographie peuvent avoir des effets similaires sur la diversité génétique observée (Tajima, 1989). Cependant, la sélection n'affecte que quelques parties codantes du génome et celles qui y sont liées, tandis que la démographie affecte le génome entier.

Un autre effet possible de la sélection sur la diversité génétique est dû à la recombinaison. Lorsqu'un locus est sous sélection positive et que sa fréquence augmente, la fréquence des mutations liées à ce locus augmente également, on parle d'auto-stop génétique (ou hitch-hiking, Maynard-Smith and Haigh 1974). Si un événement de recombinaison se produit au cours de ce processus, certains loci ne seront plus liés au locus sous sélection : leur fréquence allélique sera élevée mais n'atteindra pas 1. On aura donc une signature de ces sites « auto-stop » dans le spectre de fréquence, avec un excès de mutations à hautes fréquences par rapport à l'attendu du modèle standard neutre (Fay and Wu, 2000). Ce phénomène pourrait même expliquer en grande partie la diversité génétique observée, c'est la théorie du *genetic draft*, qui s'oppose au *genetic drift* (Gillespie, 2000).

Grâce aux prédictions du cadre théorique neutraliste, on peut chercher les régions du génome qui diffèrent significativement des prédictions, par exemple via le niveau de diversité dans la population. Ainsi, on peut détecter des sites sous sélection où la diversité est plus faible qu'attendue sous le modèle neutre (Hernandez et al., 2011).

### 1.3.3 Incohérences liées à l'utilisation du modèle neutre

#### La « taille efficace »

La notion de taille efficace est l'une des notions les plus complexes découlant de la théorie neutraliste et de ses outils, ce qui engendre de nombreuses incohérences dans son

utilisation. On a vu que c'était la taille d'une population théorique, idéale, qui aurait le même taux de dérive que la population étudiée. Elle intervient dans la majorité des statistiques du modèle neutre, ce qui donne autant de façons de la mesurer.

On a vu par exemple que le paramètre  $\theta$ , défini comme  $2N_e\mu$ , pouvait être estimé à partir du nombre moyen de différences entre deux génomes dans la population, noté  $\pi$ . Chez *Homo sapiens*,  $\pi$  a été estimé à environ  $1/1300$  (Sachidanandam et al., 2001), et  $\mu$  à  $1.2 \times 10^{-8}$  (Kong et al., 2012). On peut donc en déduire  $N_e$  qui est égal à environ 32 000 chromosomes, soit 16 000 individus diploïdes, c'est-à-dire 450 000 fois moins que la population mondiale actuelle (voir le chapitre « Quel(s) modèle(s) pour expliquer la biodiversité ? » de Guillaume Achaz dans l'ouvrage *Évolution et Biodiversité*, à paraître). Même si on a vu que la taille efficace était censée représenter les individus qui participent à la génération suivante, et qu'elle est donc inférieure à la taille totale de la population, cela ne peut pas être la seule explication à cette différence. Une autre explication possible est la démographie, puisque le calcul ci-dessus découle du modèle standard neutre dans lequel la taille de la population est constante, ce qui n'est pas vrai pour l'espèce humaine.

Le même calcul effectué avec le Virus de l'Immunodéficience Humaine (VIH) aboutit à une taille efficace d'environ  $10^3$  (Achaz et al., 2004), alors que le nombre de virus à l'intérieur d'un seul humain infecté est plutôt de l'ordre de  $10^{10}$  (Piatak Jr et al., 1993; Haase et al., 1996). Dans ce cas, la démographie ne peut pas expliquer cette différence, puisque la virémie des patients infectés chroniquement est stable. L'explication pourrait se trouver plutôt du côté de la sélection : une fois encore, ce calcul est fait dans le cadre de la théorie neutraliste, et suppose donc que la majorité des mutations est neutre du point de vue de la sélection. Pour un génome compact comme celui d'un virus, où la grande majorité des régions sont codantes, il paraît peu probable que cette hypothèse soit vérifiée.

Le fait que  $N_e$  diffère de la taille observée de la population n'est pas étonnant, cela découle de sa définition même. Cependant, on peut d'une part s'interroger sur l'ampleur de cette différence et sa signification. D'autre part, le problème avec cette notion vient du fait que, de part son nom de « taille efficace », elle est la plupart du temps considérée comme une vraie taille de population, un chiffre absolu, sans remettre en question sa signification et surtout sa pertinence. Si  $N_e$  diffère autant du  $N$  observé, c'est-à-dire si la population idéale de Wright-Fisher est si différente de la population étudiée, est-il encore pertinent d'utiliser ce modèle pour l'étudier ?



## Hypothèses non respectées

Par l'exemple de la taille efficace, on a mis en évidence l'utilisation de la théorie neutraliste dans un cadre où ses hypothèses n'étaient pas respectées. Dans la plupart des cas, cela est fait en connaissance de cause : on cherche justement à rejeter le modèle neutre, pour montrer l'existence d'une démographie non constante, ou de sites sous sélection. Cependant, d'autres hypothèses sous-jacentes de la théorie neutraliste sont rarement remises en question. C'est le cas par exemple de la variance du nombre de descendants, que je vais détailler dans la section suivante, en présentant des modèles alternatifs en génétique des populations, basés sur d'autres hypothèses.

## 1.4 D'autres modèles de génétique des populations

Le cadre théorique majoritairement utilisé en génétique des populations aujourd'hui est donc la théorie neutraliste, basée sur le modèle de Wright-Fisher et la théorie de la coalescence de Kingman. D'autres modèles applicables à la génétique des populations existent mais sont aujourd'hui peu utilisés. Ils proposent d'autres hypothèses, compatibles ou non avec la théorie neutraliste. J'en présente ici deux classes, que j'ai utilisés pendant cette thèse : les processus naissance-mort et les modèles à coalescences multiples.

### 1.4.1 Processus Naissance-Mort

On a vu qu'en génétique des populations, l'analyse des données de variation génétique se base sur le coalescent de Kingman couplé à un processus de mutations poissonnien, qui est notamment le processus limite des modèles de Wright-Fisher ou de Moran. De façon intéressante, en phylogénétique, c'est-à-dire à l'échelle des espèces et non des populations, le modèle standard utilisé est le processus naissance-mort (Kendall, 1948). Dans ce processus, les espèces apparaissent (par spéciation) à un taux dit de naissance et s'éteignent à un taux dit de mort.

Ce processus peut également être utilisé à l'échelle des individus d'une population : dans ce cas, les individus donnent naissance et meurent à des taux donnés. Si les taux de mort et de naissance sont égaux, on parle de processus naissance-mort critique. Une représentation graphique de la généalogie produite par ce processus est donnée dans la Figure 1.7, à gauche. Le temps va du haut vers le bas. On part d'un unique individu qui naît au temps 0, et est conditionné à avoir de la descendance vivante au temps présent  $t$ . L'arbre se construit donc ici dans le sens prospectif. Les événements de naissance sont figurés par des traits pointillés, qui aboutissent à un nouvel individu (toujours dessiné

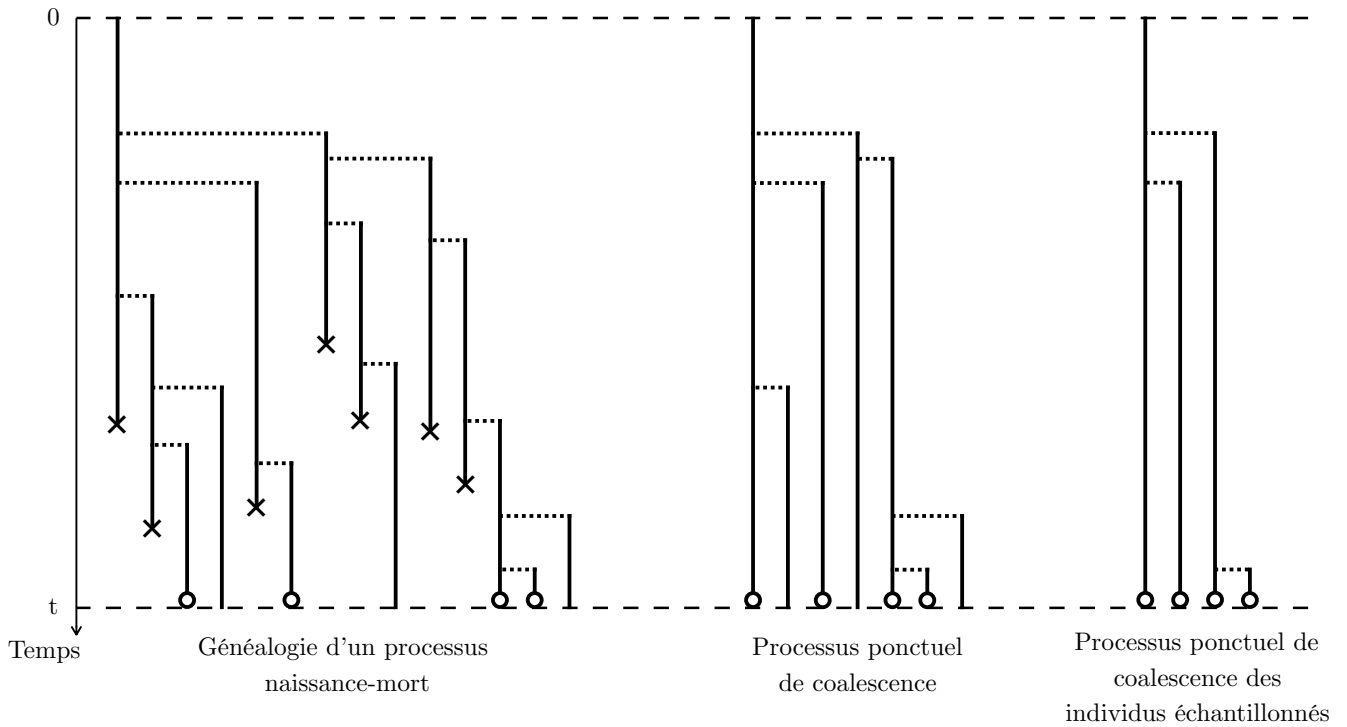


FIGURE 1.7 – Généalogie d’un processus naissance-mort et processus ponctuel de coalescence associé, pour les individus vivants au temps  $t$  et pour les individus échantillonnés, figurés par des cercles. Les évènements de naissance sont figurés par des traits pointillés. Les évènements de mort sont figurés par des croix (d’après Delaporte et al., 2016).

à droite du parent). Au temps présent  $t$ , 7 individus sont vivants, dont 4 échantillonnés (figurés par des cercles).

Ce processus et le coalescent qui en découle (appelé processus ponctuel de coalescence, ou CPP pour coalescent point process) sont très bien décrits mathématiquement, mais peu voire pas utilisés en génétique des populations. En particulier, Delaporte et al. (2016) ont dérivé la formule explicite du spectre de fréquence d’un échantillon dans un processus naissance-mort critique avec un temps de fondation fixé.

### 1.4.2 Coalescences multiples

Cette section se base sur la synthèse de Tellier et Lemaire, *Coalescence 2.0: a multiple branching of recent theoretical developments and their applications*, publiée en 2014, à laquelle on peut se référer pour plus de détails.

On a vu que la théorie de la coalescence pouvait s’adapter à certaines violations des hypothèses du modèle de Wright-Fisher, comme des changements de taille de population

ou des subdivisions de populations. Cependant, la faible variance du nombre de descendants est rarement remise en question. Dans le modèle de Wright-Fisher, le nombre de descendants par individu suit une loi binomiale de moyenne 1 et de variance  $1 - 1/2N$ , ce qui est approximativement équivalent à une distribution de Poisson de moyenne 1 et de variance 1. La conséquence de cela est que les probabilités que plus de deux lignées coalescent ou que plusieurs événements de coalescence aient lieu en même temps sont de l'ordre de  $\mathcal{O}(1/N^2)$ , et donc négligeables devant la probabilité de coalescence de deux lignées ( $1/N$ ) quand  $N$  est grand.

Dans le coalescent de Kingman, deux lignées au maximum peuvent donc coalescer, ce qui est une contrainte forte lorsqu'on étudie par exemple des espèces ayant de grandes variances de succès reproductif. Récemment, plusieurs études ont montré que chez certains organismes marins, en raison de la fécondité très élevée et de la mortalité précoce importante, la variance du nombre de descendants peut être de l'ordre de  $N$ , c'est-à-dire que certains individus produisent de l'ordre de  $N$  descendants. Cet effet, appelé reproduction «sweepstake» (c'est-à-dire tirage au sort), est entièrement dû à la variance du succès reproductif, et est donc indépendant de la sélection naturelle (Beckenbach, 1994; Hedgecock, 1994; Li and Hedgecock, 1998; Hedgecock and Pudovkin, 2011; Harrang et al., 2013). Dans ce cas, plus de deux lignées vont donc coalescer en même temps, on parle de coalescences multiples. Ces événements peuvent également survenir sous l'action de la sélection naturelle : pendant un balayage sélectif, les individus portant l'allèle avantageux vont donner plus de descendants, et ce potentiellement à des générations proches, ce qui donnera lieu à des événements de coalescence simultanés dans la nouvelle échelle de temps (Schweinsberg et al., 2005; Coop and Ralph, 2012).

Plusieurs extensions du coalescent de Kingman permettant des coalescences multiples ou simultanées (MMC pour Multiple Merger Coalescent) ont été décrites (voir Figure 1.8). Ces modèles dérivent du modèle général de dynamique des populations de Cannings, dont le modèle de Wright-Fisher et le modèle de Moran sont des cas particuliers (Cannings, 1974). Le cadre général des modèles à coalescences multiples est le suivant (Schweinsberg, 2003) : dans une population de taille  $N$ , à chaque génération, chaque individu produit indépendamment un nombre aléatoire de juvéniles, selon une distribution de probabilité donnée. Seuls  $N$  juvéniles, choisis au hasard, survivent parmi tous les descendants et constituent la génération suivante. Le nombre moyen de juvéniles produits par chaque parent est supposé supérieur à 1, et le nombre total de juvéniles est toujours très supérieur à  $N$ . Après choix au hasard des juvéniles, la moyenne du nombre de descendants par individu est de 1 (parce que la taille de la population est constante) mais la variance du

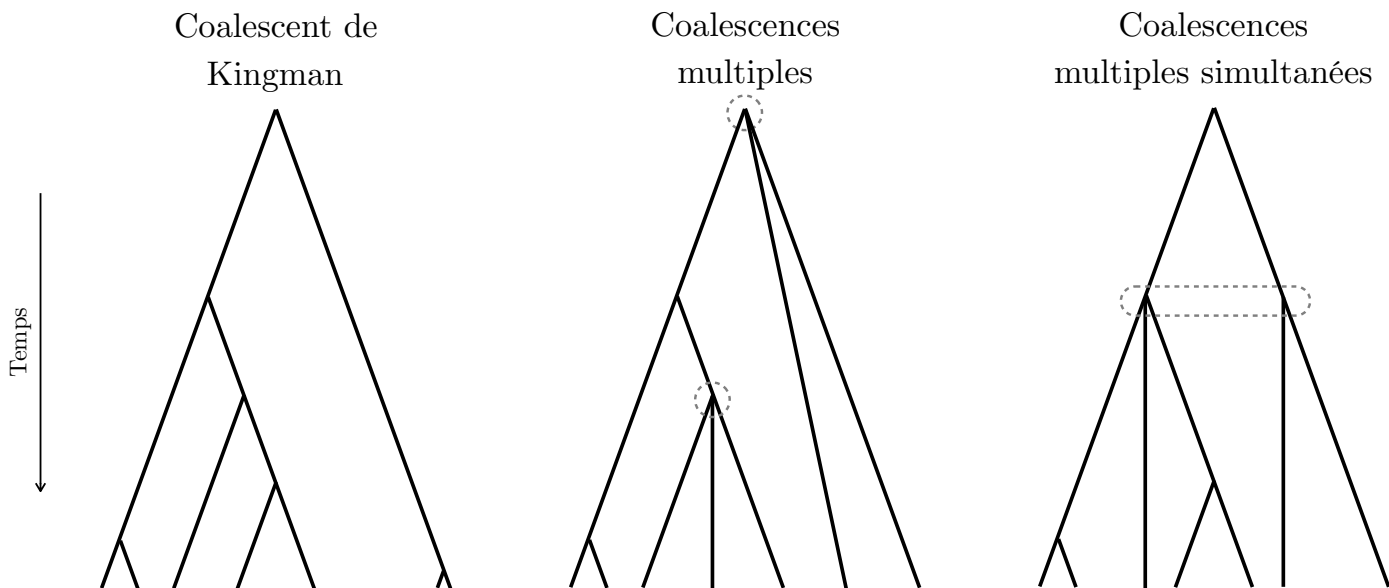


FIGURE 1.8 – G  n  alogies des mod  les    coalescences multiples (d’apr  s Tellier and Lemaire, 2014). Les   v  nements de coalescences multiples sont entour  s en pointill  s (coalescences multiples simultan  es dans le cas de l’arbre de droite).

nombre de descendants est grande, du fait de la distribution de probabilit   choisie pour le nombre de juv  niles.

Une classe de mod  les    coalescence multiple, appel  s  $\Lambda$ -coalescents, a   t   propos  e ind  pendamment par Donnelly and Kurtz (1999), Pitman (1999) et Sagitov (1999). Ces mod  les autorisent    n’importe quel temps donn   une coalescence multiple de  $k$  lign  es, o    $k \geq 2$ . Le coalescent de Kingman est un  $\Lambda$ -coalescent avec  $k = 2$ . Plusieurs distributions peuvent   tre choisies pour la fr  quence des   v  nements de coalescence multiple, et leur taille, c’est-  -dire  $k$ . Des propri  t  s des mod  les  $\Lambda$ -coalescents ont   t   d  crites, comme le spectre de fr  quence et le nombre de sites polymorphes (Birkner et al., 2011, 2013; Berestycki et al., 2014).

Le beta-coalescent est un cas particulier de  $\Lambda$ -coalescent pour lequel le taux de coalescence multiple suit une distribution Beta de param  tres  $\alpha$  et  $2 - \alpha$  ( $0 < \alpha < 2$ ) (Schweinsberg, 2003; Birkner and Blath, 2008). Les coalescences multiples se produisent    une   chelle de temps de l’ordre de  $\mathcal{O}(1/N^{\alpha-1})$ . On connait de m  me certaines de ses propri  t  s comme la longueur des arbres, le spectre de fr  quence et le nombre de sites polymorphes (Berestycki et al., 2007, 2008; Birkner and Blath, 2008). Le mod  le

de Bolthausen-Sznitman est un cas particulier de beta-coalescent avec  $\alpha = 1$ . Il a été montré qu'il reflète bien les généalogies obtenues sous des modèles incluant de la sélection positive rapide (Brunet et al., 2007; Brunet and Derrida, 2012; Neher and Hallatschek, 2013; Neher et al., 2013).

Un autre type de MMC, le  $\Psi$ -coalescent, vise à modéliser la reproduction sweepstake, via le paramètre  $\Psi$ , qui définit la proportion de descendants qui proviennent d'un même parent à la génération précédente (Eldon and Wakeley, 2006, 2008, 2009; Eldon and Degnan, 2012). Il a été utilisé pour estimer l'importance de la reproduction sweepstake chez certains organismes marins (Eldon, 2009, 2011).

Enfin, le modèle le plus général, appelé  $\Xi$ -coalescent, autorise les coalescences simultanées, c'est-à-dire qu'à un instant donné, plusieurs groupes de 2 lignées ou plus peuvent coalescer (Schweinsberg, 2000; Möhle et al., 2001; Birkner et al., 2008; Taylor and Véber, 2009).

Les caractéristiques des différents types de modèles à coalescences multiples et leurs applications biologiques sont résumées dans la Table 1.1.

La diversité génétique dans un modèle à coalescence multiple est plus faible que dans un modèle classique de Wright-Fisher (Eldon and Wakeley, 2006, 2008). Cela peut se comprendre intuitivement de plusieurs manières : si, fréquemment, un ou plusieurs individus laissent un grand nombre de descendants à la génération suivante, les individus de la génération suivante sont plus apparentés, donc moins divers génétiquement. Ainsi on aboutit dans un modèle à coalescence multiple à une population plus apparentée, moins diverse. On peut aussi le comprendre en s'intéressant à la taille de l'arbre : en partant d'un échantillon de même taille, le fait d'autoriser les coalescences multiples fait qu'on arrive plus vite au MRCA de l'échantillon. L'arbre étant plus court, la diversité génétique est plus faible dans l'échantillon.

Le spectre de fréquence obtenu sous des modèles MMC montre un excès de mutations à faibles et fortes fréquences par rapport au spectre standard neutre, dû à la forme de la généalogie en étoile : les mutations qui surviennent avant les événements de coalescences multiples anciens seront très répandues, et les mutations qui surviennent après les événements de coalescences multiples récents seront très rares. Tandis que l'excès de mutations à faibles fréquences s'observe également dans un coalescent de Kingman avec croissance de population, l'excès de mutations à fortes fréquences est spécifique des modèles MMC (voir Chapitre 4).

Comme on l'a vu, les événements de coalescences multiples résultant d'une grande

TABLE 1.1 – Caractéristiques des différents modèles de coalescents et leurs applications (Tellier and Lemaire, 2014)

Modèle de coalescent	Variance du nombre de descendants	Coalescences multiples (>2 lignées)	Coalescences simultanées	Processus biologique correspondant
Kingman	Petite	Non	Non	Modèle de Wright-Fisher
$\Lambda$ -coalescent	Grande	Oui	Non	Reproduction sweepstake, goulots d'étranglement récurrents
$\Psi$ -coalescent	Grande	Oui	Non	Reproduction sweepstake
Beta-coalescent	Grande	Oui	Non	Reproduction sweepstake, goulots d'étranglement récurrents
Bolthausen-Sznitman	Grande	Oui	Non	Sélection positive rapide et récurrente
$\Xi$ -coalescent	Grande	Oui	Oui	Balayage sélectif, goulots d'étranglement récurrents, extinction et recolonisation spatiale

variance du nombre de descendants peuvent survenir sous l'effet de processus neutres, comme la reproduction sweepstake, ou sous l'effet de processus sélectifs. Ces deux types de mécanismes peuvent a priori être distingués à partir des données puisque les processus neutres affectent l'ensemble du génome tandis que les mécanismes sélectifs n'affectent que les régions codantes et celles qui y sont liées.

Des méthodes d'inférence existent pour ajuster des modèles MMC à des données observées, dans le but d'en déduire le taux de reproduction sweepstake. Des méthodes de vraisemblance ont été développées pour inférer les paramètres à partir du spectre de fréquence, et ce pour le  $\Lambda$ -coalescent (Birkner et al., 2011), le beta-coalescent (Birkner and Blath, 2008; Steinrücken et al., 2013) et le  $\Psi$ -coalescent (Eldon and Wakeley, 2006; Cenik and Wakeley, 2010; Eldon, 2011). De plus, Eldon et al. (2015) ont montré la possibilité de distinguer de la démographie et des coalescences multiples à partir du spectre de fréquence.

## 1.5 Objectifs de la thèse

On a vu dans cette introduction le contexte historique dans lequel la théorie neutraliste est apparue et s'est imposée. Elle est aujourd'hui le cadre théorique de référence dans lequel se placent la majorité des études en évolution moléculaire. J'ai décrit quelques uns des outils mathématiques développés dans le cadre de cette théorie, et qui en font un modèle si efficace et utilisé. Ainsi, nous avons vu quelques applications, et mis en évidence certaines incohérences liées à son utilisation. Des modèles alternatifs existent, comme les processus naissance-mort ou les modèles à coalescences multiples, mais ils sont encore peu utilisés pour l'analyse de données.

Les questions qui ont motivé cette thèse étaient de comprendre comment le modèle standard basé sur la théorie neutraliste est utilisé, quelles sont les hypothèses qui sont remises en cause ? Quelles sont les conséquences de ces hypothèses sur les inférences ? D'autres modèles peuvent-ils expliquer la diversité des données observées ?

Cette thèse a donc deux objectifs principaux : d'une part, mettre en évidence certaines limites dans l'utilisation de la théorie neutraliste en évolution moléculaire. En particulier, nous nous sommes intéressés à un volet de l'évolution moléculaire qu'est l'inférence démographique. Comme on l'a vu, le coalescent de Kingman s'adapte à des modifications de taille de population : on peut donc modéliser de la démographie tout en restant dans le cadre du modèle standard neutre.

Dans la première partie (Chapitre 2), j'aborde cette question avec l'exemple des données microbiennes. Plus précisément, je montre comment le fait d'ignorer les hypothèses du modèle standard neutre fausse les inférences démographiques de populations microbiennes. De nombreuses études portent sur des bactéries pathogènes dont on souhaite connaître et surveiller la démographie. Ces études analysent les déviations des données observées par rapport aux attentes du modèle neutre comme étant dues à la démographie non-constante de l'espèce étudiée. Nous montrons dans cette étude que d'autres facteurs peuvent être responsables de ces déviations, et viennent donc biaiser les inférences démographiques. Ce travail a été réalisé en collaboration avec Eduardo ROCHA, de l'Institut Pasteur, et fait l'objet d'un article publié en 2016 dans *Molecular Biology and Evolution*.

Dans la deuxième partie (Chapitre 3), je me suis intéressée à une autre limite de l'utilisation du cadre théorique de référence, toujours pour l'inférence démographique. Nous avons illustré sur des données réelles, celles d'une population humaine africaine, la question de l'identifiabilité et de la complexité des modèles démographiques. J'ai analysé le spectre de fréquence de ces données en confrontant des modèles démographiques simples,

décrits par un unique paramètre, et une méthode complexe, le stairway plot (Liu and Fu, 2015). Cela m’a permis de montrer d’une part que la méthode complexe était faussée par le bruit présent intrinsèquement dans les données. D’autre part, j’ai mis en évidence que les différents modèles simples ajustaient tous aussi bien le spectre de fréquence, qui ne permet donc pas d’identifier la démographie de cette population. Cette étude a été publiée en 2017 dans *Genetics*.

L’autre objectif de cette thèse était de comparer le modèle standard neutre à d’autres modèles possibles, basés sur d’autres hypothèses. Dans l’article du Chapitre 3, j’ai comparé les modèles démographiques basés sur le modèle de Wright-Fisher avec un autre modèle, basé lui sur un processus naissance-mort. Ce type de modèle est très bien décrit mathématiquement mais peu utilisé en génétique des populations. Nous montrons qu’il ajuste aussi bien les données que les modèles basés sur le modèle de Wright-Fisher, tout en étant mieux caractérisé mathématiquement.

Enfin, dans la troisième partie (Chapitre 4), j’ai rassemblé plusieurs jeux de données de séquençage afin de confronter un grand nombre de données de diversité génétique à un modèle «étendu» à 2 paramètres, et permettant de tenir compte de la démographie et d’autoriser les coalescences multiples. Ce travail préliminaire n’est pas présenté sous forme d’article.



## Chapitre 2

# L'impact de la sélection, de la conversion génique et des biais d'échantillonnage sur l'inférence de démographie microbienne

### 2.1 Résumé de l'article

#### QUESTIONS

Cette étude se place dans le cadre de la phylodynamique, discipline qui intègre l'inférence phylogénétique et l'épidémiologie pour étudier les variations démographiques au cours du temps, notamment des populations d'agents infectieux. L'étude part du constat qu'un grand nombre d'études de phylodynamique concluent que la population étudiée est en croissance (dans 21 des 26 études recensées). Ces études utilisent le skyline plot, une méthode d'inférence flexible qui propose un scénario démographique à partir de données de séquences. Cette méthode estime les taux de coalescence de la population à partir de l'arbre reconstruit. Si toutes les autres hypothèses du coalescent neutre sont respectées, ces taux de coalescence peuvent être analysés en termes de démographie (Pybus et al., 2000; Drummond et al., 2005; Drummond and Rambaut, 2007). Nous avons donc étudié l'impact que pouvaient avoir certaines violations des hypothèses du coalescent neutre — la sélection, la conversion génique et les biais d'échantillonnage — sur les inférences réalisées avec le skyline plot.

## MÉTHODES

Nous avons simulé l'évolution de populations bactériennes de taille constante, en utilisant des paramètres réalistes pour inclure de la sélection, de la recombinaison et plusieurs types de biais d'échantillonnage. Les séquences simulées ont été analysées avec le skyline plot et par leur spectre de fréquence. Plusieurs intensités de sélection sont testées, et les effets de la sélection positive ou purifiante sont analysés séparément. Trois types de biais d'échantillonnages sont testés, correspondant à des biais typiques apparaissant dans les études de génétique des populations microbiennes. Les mêmes méthodes d'inférence sont ensuite appliquées à un jeu de données réelles d'*Escherichia coli*.

## RÉSULTATS

Nous étudions dans un premier temps l'impact de la recombinaison, modélisée par de la conversion génique dans les séquences simulées. L'analyse par le skyline plot de ces simulations montre une diminution des taux de coalescence qui pourrait être interprétée comme une expansion de population. À l'inverse, la recombinaison n'affecte pas le spectre de fréquence moyen observé, et diminue sa variance.

La sélection, simulée avec différentes intensités, et associée ou non à de la recombinaison, produit également des distorsions du skyline plot qui miment de l'expansion si elles sont analysées en termes de démographie.

Les trois types de biais d'échantillonnage testés produisent des scénarios de skyline plots variés, en raison de la forme de l'arbre reconstruit qui peut être affectée de différentes manières en fonction de l'échantillonnage. Lorsqu'ils sont associés à de la recombinaison et de la sélection, ces trois types de biais d'échantillonnage peuvent produire des déformations très diverses du skyline plot, allant de l'expansion à la contraction récente de la population. Les déformations du spectre de fréquence peuvent être plus facilement différenciées de celles causées par la démographie.

L'analyse du core-génome d'*Escherichia coli* par le skyline plot montre une augmentation de la taille de population suivie d'une contraction récente.

## CONCLUSIONS

Notre étude montre que le fait de négliger les effets de la sélection naturelle, de la recombinaison et des biais d'échantillonnages affecte considérablement les conclusions des analyses de phylodynamique. La comparaison des résultats obtenus avec le skyline plot et le spectre de fréquence montre que le spectre de fréquence permet d'identifier certains

biais. En particulier, le spectre n'est pas affecté par la recombinaison, contrairement au skyline plot qui se base sur un arbre reconstruit. Ainsi, l'analyse des skyline plots doit se faire conjointement à de la détection de recombinaison et de sélection, à l'analyse du spectre de fréquence, et à d'autres méthodes de génétique des populations, pour inférer correctement les changements démographiques des populations microbiennes.

## ÉTAT DE PUBLICATION

Cet article a été publié dans *Molecular Biology and Evolution* le 1<sup>er</sup> mars 2016, après révisions mineures.

## 2.2 Article

L'article est présenté dans les pages suivantes. Il est suivi d'une annexe constituée des figures supplémentaires publiées en complément de l'article.

# The Impact of Selection, Gene Conversion, and Biased Sampling on the Assessment of Microbial Demography

Marguerite Lapierre,<sup>1,2</sup> Camille Blin,<sup>3,4,5</sup> Amaury Lambert,<sup>2,6</sup> Guillaume Achaz,<sup>1,2</sup> and Eduardo P. C. Rocha<sup>\*,4,5</sup>

<sup>1</sup>Atelier de Bioinformatique, UMR7205 ISYEB, MNHN-UPMC-CNRS-EPHE, Muséum National d'Histoire Naturelle, Paris, France

<sup>2</sup>Collège de France, Center for Interdisciplinary Research in Biology (CIRB), CNRS UMR 7241, Paris, France

<sup>3</sup>Sorbonne Universités, UPMC Univ Paris06, IFD, 4 Place Jussieu, Paris Cedex05, France

<sup>4</sup>Institut Pasteur, Microbial Evolutionary Genomics, Paris, France

<sup>5</sup>CNRS, UMR3525, Paris, France

<sup>6</sup>UPMC Univ Paris 06, Laboratoire de Probabilités et Modèles Aléatoires (LPMA), CNRS UMR 7599, Paris, France

\*Corresponding author: E-mail: [erocha@pasteur.fr](mailto:erocha@pasteur.fr).

Associate editor: Helen Piontkivska

## Abstract

Recent studies have linked demographic changes and epidemiological patterns in bacterial populations using coalescent-based approaches. We identified 26 studies using skyline plots and found that 21 inferred overall population expansion. This surprising result led us to analyze the impact of natural selection, recombination (gene conversion), and sampling biases on demographic inference using skyline plots and site frequency spectra (SFS). Forward simulations based on biologically relevant parameters from *Escherichia coli* populations showed that theoretical arguments on the detrimental impact of recombination and especially natural selection on the reconstructed genealogies cannot be ignored in practice. In fact, both processes systematically lead to spurious interpretations of population expansion in skyline plots (and in SFS for selection). Weak purifying selection, and especially positive selection, had important effects on skyline plots, showing patterns akin to those of population expansions. State-of-the-art techniques to remove recombination further amplified these biases. We simulated three common sampling biases in microbiological research: uniform, clustered, and mixed sampling. Alone, or together with recombination and selection, they further mislead demographic inferences producing almost any possible skyline shape or SFS. Interestingly, sampling sub-populations also affected skyline plots and SFS, because the coalescent rates of populations and their sub-populations had different distributions. This study suggests that extreme caution is needed to infer demographic changes solely based on reconstructed genealogies. We suggest that the development of novel sampling strategies and the joint analyzes of diverse population genetic methods are strictly necessary to estimate demographic changes in populations where selection, recombination, and biased sampling are present.

**Key words:** bacteria, population size, natural selection, gene conversion, *Escherichia coli*, population genomics.

## Introduction

Bacterial populations show extensive demographic variations across space and time (Martiny et al. 2006), such as frequent expansions and bottlenecks. The characterization of these demographic changes among populations of infectious agents provides epidemiological information that can guide public health interventions. A recent field of research, phylodynamics, aims at understanding the association between ecological processes and epidemiological patterns in an evolutionary framework (Grenfell et al. 2004). It integrates phylogenetic inference and population genetics to study variations in demography through time (Grad and Lipsitch 2014; Li et al. 2014). Phylodynamics has been particularly useful to characterize transmission dynamics from sequence data, and could facilitate the evaluation of public health policies for diseases with low reporting rates (Volz et al. 2013).

Demographic changes imprint the reconstructed genealogies of the population, the so-called coalescent tree, by affecting the intervals of time between successive splits in the tree (Tajima 1989a). These values (coalescent rates) are proportional to the inverse of the effective population size ( $N_e$ ) in the standard neutral model. If one takes two idealized populations with the same contemporary population size, then the one with a history of population expansion will have (on average) shorter branches throughout, including at the tips. However, the relative length of the tips compared with the internal branches will be longer than in a nonexpanding population. Since nodes in the reconstructed genealogy of the expanding population are more concentrated closer to the root of the tree, the site frequency spectrum (SFS), that is, the distribution of the frequencies of all nucleotide polymorphisms, shows an excess of alleles shared by few individuals (rare alleles) (Adams and Hudson 2004). Conversely,

© The Author 2016. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.  
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Open Access

populations with a history of population size contraction exhibit an excess of polymorphism shared by many individuals when compared with stable populations with the same contemporary population size. Their reconstructed genealogies have longer branches overall, but the average length of the tips compared with the internal branches are shorter than in a noncontracting population (coalescence rates are higher than expected closer to the present).

Under the assumptions of the standard neutral model (no population structure, random sampling, no recombination, no selection), it is often implicitly assumed that variations in  $N_e$  (or equivalently, variations in the coalescence rate) are indications of demographic changes. Parametric approaches were developed to infer these demographic changes under explicit models, such as the Approximate Bayesian Computation method (Beaumont et al. 2002) or the likelihood-based method (e.g., Nielsen and Wakeley 2001; Drummond et al. 2002). In this context, skyline plots were introduced to quantify the relationship between the coalescence rate of the population and the genealogy of the sequences in a non-parametric approach, that is, without an explicit model to test. Coalescent rates can then be used to produce detailed demographic histories from sequence data assuming that all other assumptions of the neutral coalescent are met (Pybus et al. 2000; Drummond et al. 2005). Demographic trends can also be inferred using SFS-based neutrality tests (Fu 1997; Fu and Voordouw 1997; Ramos-Onsins and Rozas 2002; Achaz 2009). For example, Tajima's  $D$  measures the difference between the mean number of pairwise differences and the number of segregating sites, and is skewed to negative values in case of population expansion (Tajima 1989b). SFS-based model-flexible methods (i.e., exploring the space of possible demographic models) have also been recently proposed (Liu and Fu 2015). They approximate the demography using piecewise constant population sizes.

Violations of the assumptions of the neutral coalescent, such as presence of recombination or selection, may affect reconstructed genealogies and SFS in ways resembling demography (e.g., Schierup and Hein 2000; Nielsen and Beaumont 2009; Mazet et al. 2015). Recombination by gene conversion has a very moderate effect on the topology of phylogenetic trees (Touchon et al. 2009), but affects skyline models (Hedge and Wilson 2014). Removing sites incompatible with the tree topology, that is, homoplasies, actually aggravates the effect of recombination in skyline models, presumably because it preferentially removes polymorphisms in deeper branches of the tree (Hedge and Wilson 2014). Recombination in the absence of selection has actually little effect on the expected SFS, apart from decreasing its variance (Wall 1999). The effect of selection on skyline plots has been less studied. Strong purifying selection is not expected to affect drastically the SFS because the deleterious mutations are quickly purged (Kimura 1983). On the other hand, mild purifying selection or recent selective sweeps lead to an excess of recent polymorphism, creating the impression of recent population expansion

(Braverman et al. 1995). Diversifying or balancing selection can produce more complex patterns (Navarro and Barton 2002). Some studies have found that deleterious mutations of mild effect have a negligible effect on the time back to the most recent common ancestor (TMRCA) (Neuhauser and Krone 1997), and very little effect on the shape of the reconstructed genealogies (Przeworski et al. 1999) even though linkage between sites may affect the distribution of mutations (Williamson and Orive 2002). Mutations of mild deleterious effect are abundant in some bacteria (Hughes 2005; Balbi et al. 2009). If bacterial evolution is dominated by these mutations then selection might not strongly affect demographic inference using skyline plots. However, recent studies have suggested that weak purifying selection, when occurring at multiple sites, could affect the shape of the coalescent tree (O'Fallon et al. 2010). The effect of selection on skyline plots remains unclear.

The possibility of producing large sequence datasets for microbial populations has spurred interest on the use of these methods to study microbial demography. The skyline plot has been particularly popular because it allows precisely detailing demographic changes (Ho and Shapiro 2011). This method was initially used to study RNA viruses, which exhibit low recombination rates between individuals in different hosts and small effective population sizes (Holmes 2007). These viruses also have very high mutation rates, which increases mutational load and decreases the efficiency of selection (especially under no recombination) (Kimura 1983). Skyline plots have been increasingly used to study cellular microbes, most notably pathogenic bacteria. Yet, it is unclear if violations to the neutral coalescent model (biased sampling, selection, or recombination) can be safely ignored in these cases. Many bacterial populations are extremely large, show a very strong imprint of natural selection, endure rapid population fluctuations, exhibit low mutation rates, and recombine at high rates (Rocha et al. 2006; Vos and Didelot 2009; Tellier and Lemaire 2014). In fact, abundant evidence suggests that there are few, if any, positions evolving according to the neutral model in bacterial genomes (reviewed in Rocha and Feil 2010).

Most demographic analyses assume random sampling. However, sampling is usually not random in microbial studies, either on purpose or by the intrinsic difficulties of defining appropriate sampling strategies in microbiology, and this may severely affect the conclusions taken from the analysis of reconstructed genealogies. There are three major sampling biases in microbiology. *Clustered* sampling occurs when all samples are taken from a single sub-population, for example, a particularly virulent lineage. *Uniform* sampling of all major lineages is frequently found in studies aiming at maximizing the genetic diversity of samples. This bias may also result from sampling different environments (or patients) while analyzing a single isolate per site (thus disregarding differences in population sizes in each site). Finally, a very common type of *mixed* sampling bias is found in studies extensively sampling a sub-population and a small number of very diverse individuals from other sub-populations. This gives a broad view of

the genetic diversity in the species, while focusing in a sub-population of interest. Analyses using sequences available in databanks are prone to combine the sampling biases of the different underlying studies.

We surveyed the available literature on the use of skyline plots to describe bacterial population demography and found that nearly all studies showed skyline plots suggestive of population expansion. We then decided to test if the violations of the assumptions of the neutral coalescent could be reasonably ignored when studying bacterial populations. For this, we simulated the evolution of bacterial populations of constant size using biologically realistic parameters for natural selection, recombination, and sampling bias. These sequences were then used to build skyline plots and make SFS-based inference of demographic changes. We did not use time calibration in the inference of the skyline plots. Therefore, the Y-axis in the skyline plots represents the inferred product of  $N_e$  by the mutation rate  $u$  ( $N_e u$ ) and the X-axis represents the expected number of mutations per site, which is an estimate of the distance from the present (Ho and Shapiro 2011). By convention, we represent zero mutations per site at the left of the skyline plots. Hence, the X-axes of the skyline plots are ordered from the present (left) to the past (right). In the last section, we present the analysis of data from *Escherichia coli* in the light of the results of simulations.

## Results

### The Puzzling Expansion of Most Bacterial Populations

We found 26 recent studies using skyline plots to analyze bacterial demography. We analyzed their characteristics in terms of TMRCA, demographic changes, and their presumed justifications (table 1). The TMRCA of these populations was extremely variable, from 3 years to over 100 million years. Many of these studies proposed some type of justification for the observed demographic changes. For example, demographic expansion in *Bordetella pertussis* was associated with the introduction of vaccination and expansion of escape variants (Bart et al. 2014). Demographic expansion in *Clostridium difficile* was associated with the date when the bacterium became a recognized nosocomial pathogen (He et al. 2010), and in *Salmonella enterica* serovar Typhi with the introduction of antibiotics (Roumagnac et al. 2006). Skyline plots suggested that the effective population size of *Neisseria gonorrhoeae* in Baltimore increased during most of the twentieth century and then decreased, presumably as the result of urban planning and changes in patterns of drug addiction (Perez-Losada et al. 2007). Some works suggested associations between the increase in effective population sizes and environmental changes, for example, glacial cycles in *Thiomonas* spp. (Liao and Huang 2012), and human population growth in *Mycobacterium tuberculosis* (Comas et al. 2013). However, a careful analysis of table 1 revealed a most puzzling trend: the vast majority of studies (21 out of 26) concluded that effective population sizes have increased.

Are all bacterial populations expanding? Researchers might focus preferentially on expanding bacterial populations, for example, recent epidemic clones, thus producing an

ascertainment bias towards population expansion. Also, human populations have been growing exponentially and human-specific pathogens might have followed similar trends. However, a number of arguments cast doubt on these results. (1) The prevalence of bacterial pathogens (the majority of species in table 1) has decreased in the last century as the result of hygiene and the use of antibiotics (Cohen 2000). (2) Most of the remaining species in table 1 are commensals associated with multiple hosts (eventually including some nosocomials), or free-living bacteria for which human population growth might be of little relevance (especially since it is associated with decrease in the population of closely related animals that are often within the commensal host range). For example, *E. coli* is associated with most warm-blooded and some cold-blooded animals (Tenailon et al. 2010), *Moraxella* was until recently regarded exclusively as a commensal of animals (Brenner et al. 2005), and *Thiomonas* spp. are free-living bacteria inhabiting extreme environments (Liao and Huang 2012). (3) The majority of the studies in table 1 have not checked for the assumptions of the standard neutral model, and those that did, only checked for the presence of recombination. Very few studies have used SFS to infer demographic changes in bacterial populations. While several of these works obtained SFS compatible with recent demographic expansions, they also showed that distortions in the SFS were partially caused by purifying selection (Cornejo et al. 2013; Pepperell et al. 2013; Touchon et al. 2014). These arguments led us to study the effects of violations of the assumptions of the standard neutral model in the inference of bacterial demography.

### The Effect of Recombination

We made forward population genetics simulations of a locus of 20 kb with gene conversion and constant population size (see section “Methods”). Hence, deviations from the expectations of the neutral coalescent in the simulations were necessarily caused by recombination, not demography. The parameters for the simulations were taken from the literature for the model bacterium *E. coli* (table 2). Several studies estimated the rate of recombination over mutation in *E. coli* (reviewed in Bobay et al. 2015). We used an estimate based on the analysis of complete genomes (Touchon et al. 2009), which is among the lowest proposed and might therefore be conservative. The sequences resulting from our simulations were used to obtain skyline plots with BEAST (Drummond and Rambaut 2007). Our results show that even the moderate recombination rate observed in *E. coli*, leads to skyline plots with increasing values of  $N_e u$  for recent dates (fig. 1). This could be spuriously interpreted as an indication of population expansion. Simulations using ten times larger recombination rates (as observed in highly recombining bacteria), showed even stronger distortions in the skyline plots. Expectedly, recombination had no effect on the number of segregating sites (see *Recombination* in fig. 2), and lowered the variance, but did not affect the average, of the genome-wide average SFS (fig. 1). Consequently, recombination had no effect on the average estimate of Tajima *D* (although for a single locus see Thornton 2005).



**Table 1.** Published Works Using Skyline Plots to Estimate Demographic Changes in Bacteria.

Species	Conclusion	TMRCa	Authors' Comments
<i>Bordetella pertussis</i>	Expansion	200 Y	Surprisingly, vaccination was followed by increase not decrease in $N_e u$ , suggesting diversification of lineages escaping the vaccine (Bart et al. 2014)
<i>Clostridium difficile</i>	Expansion	35 Y	Population expansion coincides with the first reports of hospital outbreaks (He et al. 2010). Recombination tracts removed
<i>Escherichia coli</i>	Expansion	140 MY	A population bottleneck had a founding effect by purging diversity and leading to the formation of the extant major groups of <i>E. coli</i> (Wirth et al. 2006). 50-fold population expansion in the last 5 MY. Mentions the caveat of recombination
<i>Legionella pneumophila</i>	Expansion	20 Y	Correlation between population and reported number of clinical cases (Sanchez-Buso et al. 2014). Recombination tracts removed
<i>Moraxella catarrhalis</i>	Expansion	50 MY	The populations of antibiotic resistant isolates expand faster than those of sensitive bacteria (Wirth et al. 2007). Recombination tracts removed
<i>Mycobacterium tuberculosis</i>	All expansion	70 KY, 6.6 KY, 40Y	(1) Concludes about a parallel evolution between human (mitochondria) and this clade's $N_e$ caused by a tight host-parasite association (Comas et al. 2013). (2) One expansion is associated with the industrial revolution, another with the first world war, and a recent contraction is associated with the introduction of antibiotherapy (Merker et al. 2015). (3) Expansion is associated with acquisition of multi-drug resistance (Eldholm et al. 2015)
<i>Mycoplasma gallisepticum</i>	Expansion	17 Y	Population expansion (Delaney et al. 2012)
<i>Neisseria gonorrhoeae</i>	Expansion, contraction	40 Y <sup>a</sup> , 120 Y	(1) Population expansion measured in housekeeping functions parallels the number of clinical cases, but not when measured in an antibiotic resistance gene, suggesting it has been subject to positive selection. Results could be used in managing resistance (Tazi et al. 2010). Found no recombination events in the set. (2) Suggests that demographic changes are associated with selective sweeps caused by antibiotic resistance, crack epidemics and urban-planning. $N_e$ decrease associated with 5× decrease in the prevalence of this obligatory human pathogen (Perez-Losada et al. 2007). Recombination tracts were removed
<i>Pseudomonas aeruginosa</i>	Expansion	0.005/nt <sup>b</sup>	Assigns the presence of a recent selective sweep (Guttman et al. 2008)
<i>Pseudomonas fluorescens</i>	Stable	0.07/nt <sup>b</sup>	Suggests ancient rapid growth followed by stabilization, but very close strains are absent (Guttman et al. 2008)
<i>Pseudomonas syringae</i>	Stable	0.1/nt <sup>b</sup>	Suggests it is an endemic pathogen (Sarkar and Guttman 2004)
<i>Salmonella enterica</i> serovar Paratyphi A	Expansion	450 Y	Population contraction associated with the introduction of antibiotics, followed by expansion that would be associated with environmental changes (Zhou et al. 2014). Recombination tracts removed
<i>Salmonella enterica</i> serovar Typhi	All expansion	10–71 KY, 25 Y	(1) Steady increase in population size in the last 3,000 years. Recombinant SNPs removed and strong selection checked (Roumagnac et al. 2006). (2) Expansion is consistent with epidemiological data reporting drug-resistant isolates. Recombinant regions removed (Wong et al. 2015)
<i>Shigella sonnei</i>	Stable	500 Y	The population size was found to be constant through time (Holt et al. 2012)
<i>Staphylococcus aureus</i>	Expansion	20 Y, 50 Y, 30 Y	(1) Rampant expansion might have followed trans-Atlantic spread (Nubel et al. 2010). (2) Phylodynamics analysis used to estimate epidemiological parameters such as the potential reproductive number. No signs of recombination identified (Prosperi et al. 2013). (3) Fit between demographic expansion and the epidemiology of the CC80 clone (Stegger et al. 2014)
<i>Streptococcus pneumoniae</i>	Contraction	15 Y	Population expansion and then contraction fits the observed number of clinical cases (Croucher et al. 2014). Recombination tracts removed
<i>Streptococcus pyogenes</i>	Expansion	80 Y	Associates population expansion with the acquisition of super-antigens (Davies et al. 2015). Recombination tracts removed
<i>Streptococcus suis</i>	Expansion	90 Y	Correlates population expansion with the introduction of new methods used for improved pig genetics (Weinert et al. 2015). Recombination tracts removed
<i>Thiomonas spp</i>	Expansion	7 MY	The demographic history matches the glacial cycles (Liao and Huang 2012)
<i>Vibrio cholerae</i>	Expansion	3 Y	Association with the history of the progression of an epidemic (Azarian et al. 2014). Found no evidence for recombination

NOTE—We show the TMRCa, the conclusion of the work, and the authors' justifications of the results. Multiple studies published for a given species are indicated as multiple lines in the column TMRCa and by the respective numbers in the last column.

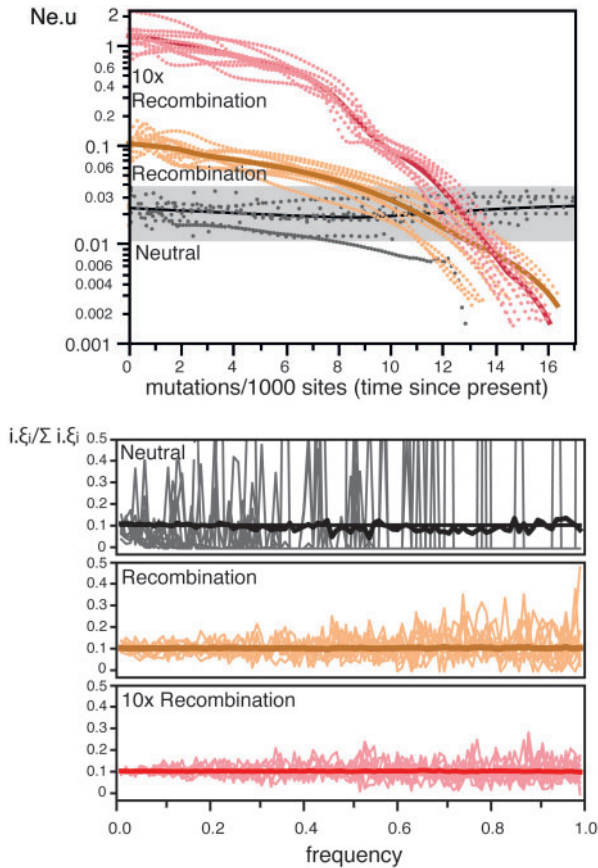
<sup>a</sup>TMRCa not indicated. The value indicates the span of the X-axis on the skyline plot.

<sup>b</sup>Studies did not perform time calibration and present only the number of mutations per site.

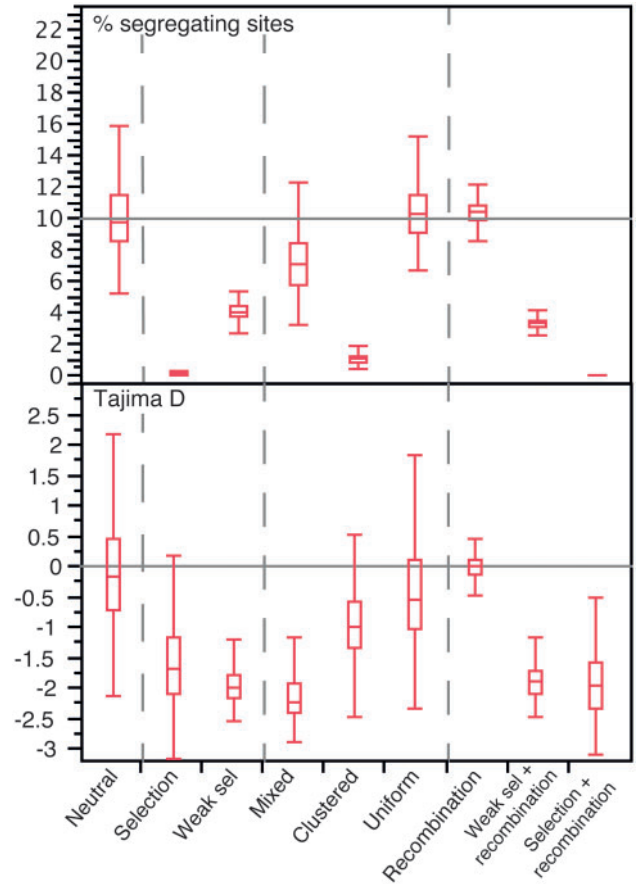
**Table 2.** Parameters for *E. coli* Populations Used in the Simulations.

Parameter	Value	Reference
Effective population size ( $N_e$ )	$1.8 \times 10^8$	Hartl et al. (1994)
Genomic adaptive mutation rate	$1 \times 10^{-5}$	Perfeito et al. (2007)
Genomic deleterious mutation rate	$2 \times 10^{-6}$	Kibota and Lynch (1996)
Average value of $s^a$	$\pm 7 \times 10^{-3}$	Perfeito et al. (2007) and Gallet et al. (2012)
Mutation rate per generation ( $u$ )	$8.9 \times 10^{-11}$	Wielgoss et al. (2011)
Genome size (nt)	$5 \times 10^6$	Touchon et al. (2009)
Recombination/mutation rate	1	Touchon et al. (2009)
Size of recombination tracts	542	Didelot et al. (2012)
SNPs recombination/mutation	2.5	Touchon et al. (2009)
Weak selection ( $N_e s$ )	5	
Strong recombination/mutation rate	10	

<sup>a</sup>The absolute values of  $s$  for adaptive and deleterious mutations being in the same order of magnitude we used an average for both.



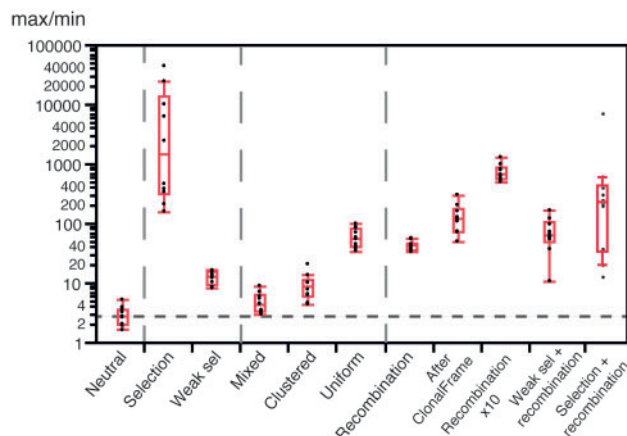
**Fig. 1.** The effect of recombination on skyline plots and SFS. The simulations used the *E. coli* population parameters (*Recombination*), ten times higher recombination rates (*10x Recombination*), or no recombination (*Neutral*). *Top* The simulations in the skyline plots are represented as dotted lines. The thick lines represent the smooth kernel fit (resp.  $R^2 = 0.81$ ,  $R^2 = 0.87$ , and  $R^2 = 0.38$ ). *Bottom*. SFS (distribution of the frequencies of all nucleotide polymorphisms in the sample) for each condition. The thick line indicates the average SFS over 1,000 replicates whereas the thin shaded lines are the observed SFS for ten random replicates. All SFS were transformed and normalized (see section “Methods”). Colors match the same datasets in both plots.



**Fig. 2.** Distribution of the number of segregating sites and Tajima  $D$  values in each set of 1,000 simulations. The gray line in the top panel corresponds to the expected number of segregating sites under the standard neutral model:  $\pi = \theta \cdot L \cdot a_n$  where  $a_n = \sum_{i=1}^{n-1} \frac{1}{i}$ . Here,  $\theta = 0.02$ ,  $L = 20,000$ , and  $n = 100$ . The gray line in the bottom panel corresponds to the expected Tajima  $D$  under the neutral model ( $D = 0$ ).

We then tested if state-of-the-art methods aiming at producing “recombination free” phylogenetic trees could produce unbiased skyline plots. We analyzed ten simulations with ClonalFrame to obtain a matrix of distances between individuals purged from recombination (Didelot and Falush 2007). We used these matrices to infer phylogenies and these phylogenies to compute skyline plots. The latter showed very clear and systematic increase in the values of  $N_e u$  for recent times (supplementary fig. S1, Supplementary Material online). The average amplitude in  $N_e u$  (measured as the ratio between the maximal and the minimal value) was three times higher than the one obtained without the use of ClonalFrame, that is, with the primary data (see After ClonalFrame in fig. 3). This suggests that ClonalFrame distance matrices are skewed so that the trees inferred from them have internal branches more affected by the removal of recombination than the external branches. These results are in line with a previous study showing that removing homoplasies in recombined sequences worsens the distortions in skyline plots (Hedge and Wilson 2014). Hence, trying to remove polymorphism caused by recombination may aggravate the biases of demographic studies using skyline plots.



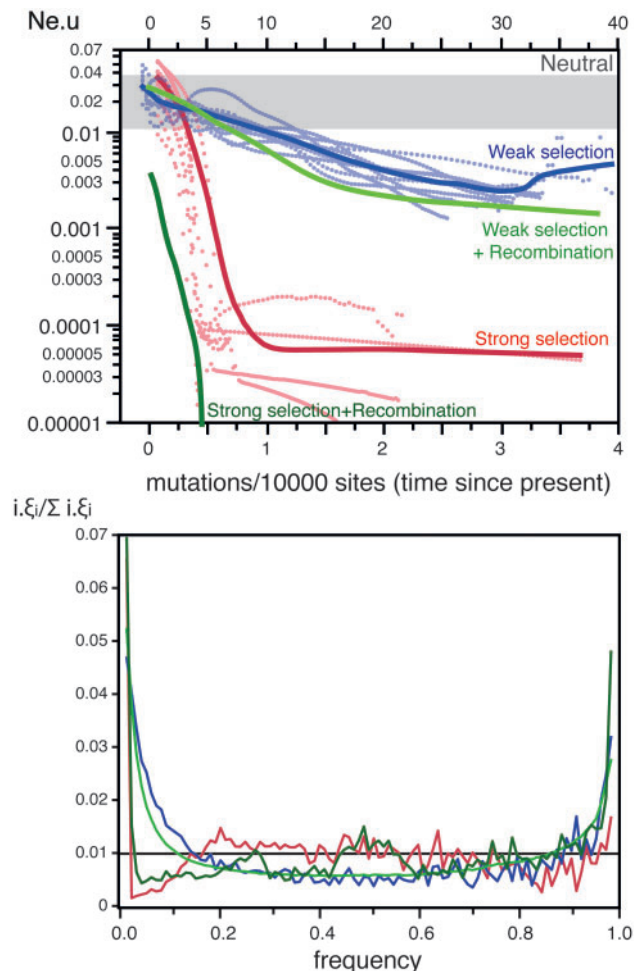


**FIG. 3.** Boxplots of the ratios between the maximal and minimal  $N_e u$  values for skyline plots (ten simulations each), across the different types of simulations. All other categories were significantly different from *Neutral* (all  $P < 0.01$  Wilcoxon tests, except the comparison between *Neutral* and *Mixed*,  $P = 0.0102$ , same test).

### The Effect of Selection

Experimental works indicate that  $>45\%$  of the mutations are deleterious (Kibota and Lynch 1996), and  $>2\%$  are adaptive (Perfeito et al. 2007) in *E. coli*. The effective population size of the species is estimated at  $>10^8$  (Hartl et al. 1994; Lynch 2006). The average selective effects of mutations in *E. coli* are much larger than the inverse of the effective population size (table 2), which implies that their fate is mostly driven by selection (Kimura 1983). Our simulations using these parameters resulted in very strong distortions in the skyline plots, showing higher  $N_e u$  values for recent dates (see *Selection* in fig. 3). These patterns might have been interpreted as population expansions if the effect of selection had been ignored. Under strong selection, diversity is constantly being purged and swept away by recurrent selective sweeps. Accordingly, the fraction of segregating sites in these simulations was only  $\sim 0.16\%$ , to be compared with  $\sim 10\%$  for the neutral simulations (see *Selection* in fig. 2). The effect of strong selection was also apparent in the SFS, where extremely rare and frequent alleles were in large excess (fig. 4), presumably due to the selective sweeps caused by beneficial mutations. This resulted in negative values of Tajima  $D$  (fig. 2).

Some of the species listed in table 1 have narrow host ranges and might have much smaller  $N_e$  than *E. coli*. We therefore made simulations using parameters corresponding to populations with  $N_e s = \pm 5$  ( $s$  being the average selection coefficient on sites under selection) and a distribution of the frequency of sites under selection similar to *E. coli*. If these species have similar distributions of selective effects as those used for *E. coli* (i.e., similar  $s$ ), this value corresponds to  $N_e$  close to 1,000 (five orders of magnitude lower than *E. coli*). One should note that even bacteria obligatorily associated with humans are thought to have higher absolute values of  $N_e$  or  $N_e s$ , for example, the  $N_e$  of *Neisseria meningitidis* was estimated at  $10^5$  (Treangen et al. 2008), and the average nonsynonymous values of  $N_e s$  were estimated at  $-5$  for *M. tuberculosis* (Pepperell et al. 2013) and at  $-17$  for



**FIG. 4.** The effect of selection on ten skyline plots (top) and 1,000 SFS (bottom). *Top* The simulations were represented as dotted lines. The thick lines represent the smooth kernel fit for strong and weak selection (resp.  $R^2 = 0.78$ ,  $R^2 = 0.79$ ). For the analysis of selection and recombination only the kernel fits are indicated ( $R^2 = 0.80$ ). The grey box indicates the range of variation of the *Neutral* simulations in figure 1. *Bottom* The thick lines represent the average SFS over 1,000 simulations. In all SFS plots, the horizontal black line indicates the neutral expectation. Colors match the same datasets in both plots.

*Streptococcus mutans* (Cornejo et al. 2013). As expected, simulations incorporating such weak selection showed patterns much less extreme than those obtained under strong selection. For example, the average fraction of segregating sites in the former was  $\sim 4\%$ , less than half of the neutral expectation but over two orders of magnitude more than under strong selection (see *Weak sel* in fig. 2). The skyline plots and the SFS under weak selection also showed less striking distortions (see *Weak sel* in figs. 3 and 4). Nevertheless, deviations from the expectation under neutral evolution were still very important in both analyses (negative Tajima  $D$ , fig. 2). These are likely to be caused by low-frequency segregating mildly deleterious mutations and by the selective sweeps caused by beneficial mutations. Hence, selection affects the inference of demography even when the values of  $N_e$  are uncharacteristically low for bacterial populations.

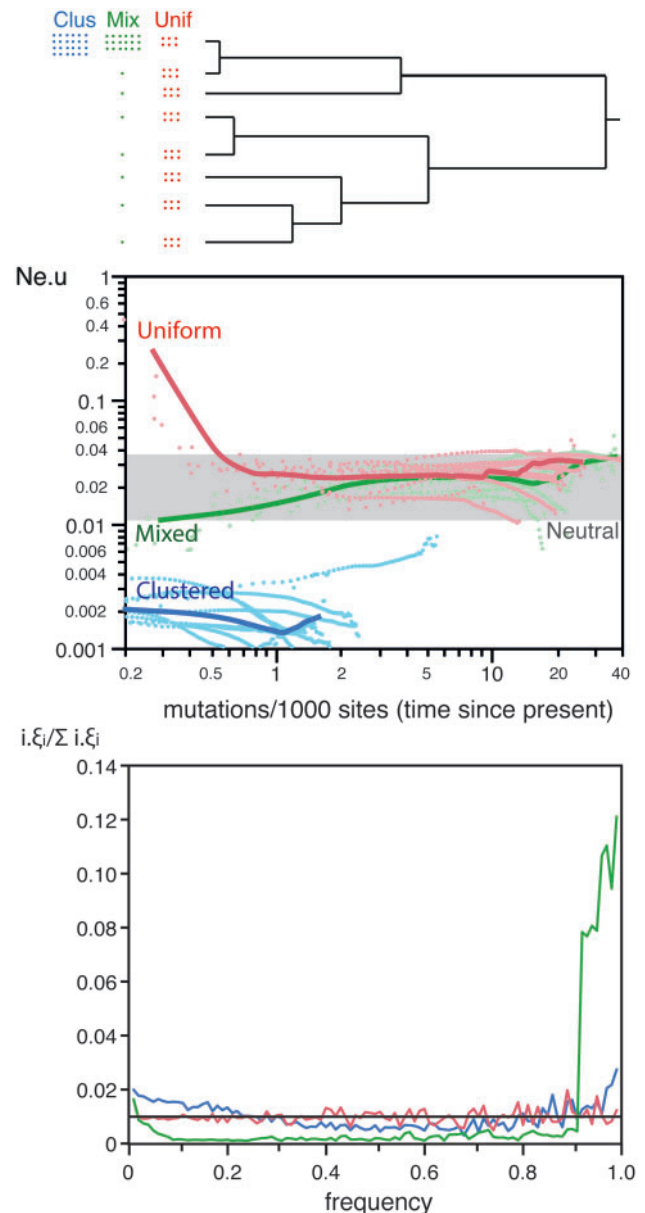
In our previous simulations, we have included positive and purifying selection. We therefore assessed the separate impact of each of these components of the evolutionary process on the skyline plots and on the SFS. For this we made simulations with just either positive or purifying selection. The effect of strong selection on skyline plots and SFS was caused exclusively by positive selection (supplementary fig. S2, Supplementary Material online). Accordingly, the SFS for strong purifying selection shows no excess of rare or frequent variants. This is because of the extremely rapid purge of deleterious mutations of strong effect. On the other hand, the significant effect of weak selection on the skyline plots and SFS is caused by both purifying and positive selection (supplementary fig. S3, Supplementary Material online). The SFS and skyline plots of populations evolving under weak purifying selection show an excess of rare variants and an increase in  $N_e u$  for recent times (supplementary fig. S4, Supplementary Material online). This shows that when selection is very strong, only positive selection affects the reconstructed genealogies, whereas when selection is weaker, both positive and purifying selection affect the reconstructed genealogies (and thus the skyline plot).

We then simulated the joint effects of selection and recombination on the reconstructed genealogies to check if recombination might moderate the effects of selection (fig. 4). The joint effect of recombination and selection (weak or strong) on the skyline plots was noticeable, that is, led to even stronger distortions in the plots, than the independent effects of each taken separately ( $P < 0.0001$ , Wilcoxon test). The SFS with selection and recombination were not appreciably different from the ones with selection under no recombination (compare the pairs of lines in the SFS of fig. 4). As a result, Tajima  $D$  is negative whenever there is selection, that is, with or without recombination (fig. 2). These results show that one cannot ignore the effect of selection on the analyses of bacterial demography.

### The Effect of Sampling Bias

We simulated three types of typical sampling biases in the study of microbial population genetics. In these simulations, there were no changes in population size, no selection, and no recombination. We simulated sampling biases by clustering the final individuals evolved in the simulations in groups using sequence similarity and then sampling these groups in different ways (see section “Methods”). The results showed that different types of sampling bias affect in very diverse ways the shape of the tree and of the SFS, and thus the inference of demographic changes (fig. 5).

The sampling of a single group (clustered sampling), resulted in skyline plots with lower average values of  $N_e u$ , as expected, and a peak of high  $N_e u$  for times very close to the present (see supplementary fig. S5, Supplementary Material online, for the values close to 0). The amplitudes of  $N_e u$  values were on average three times larger than those of neutral populations (Clustered in fig. 3). The simulations also showed slight over-representation of rare and frequent variants in the SFS. Clustered sampling produced alignments with far fewer (approximately ten times) segregating sites than the



**Fig. 5.** Analysis of three types of sampling biases. *Top* Schematic representation of the different types of sampling biases in a species tree (see section “Methods” for a precise definition). *Center* Skyline plots for each set of ten simulations. The dotted lines represent the simulations. The thick line represents the smooth kernel fit (resp. Clustered  $R^2 = 0.63$ , Uniform  $R^2 = 0.86$ , Mixed  $R^2 = 0.40$ ). The grey box indicates the range of variation of the Neutral simulations in figure 1. See supplementary figure S5, Supplementary Material online for a zoom for values of clustered bias close to zero. *Bottom* Average SFS for the three datasets (1,000 simulations for each). Colors match the same datasets in both plots.

neutral simulations (Clustered in fig. 2). Hence, sampling a sub-population produces patterns akin to very recent population size expansions.

We simulated uniform sampling by re-sampling the same number of individuals in each group. This led to skyline plots with increasing values of  $N_e u$  for recent dates (fig. 5). In fact, this sampling bias resulted in reconstructed genealogies with fewer than expected short terminal branches, which is akin to

the effect produced by strong population expansion. The consequent distortion of the reconstructed genealogies can be extremely important since these skyline plots had  $N_e u$  amplitudes  $>100$  times higher than those found on neutral populations (*Uniform* in fig. 3). On the other hand, uniform sampling had essentially no effect on the SFS (fig. 5).

Mixed sampling bias was simulated by retrieving 91 individuals from one group and one from each of the remaining nine groups. These samples showed complex skyline plots, with initially increasing  $N_e u$  values followed by a sharp decrease for very recent dates (fig. 5). The SFS showed striking over-abundance of very frequent variants, some over-representation of rare variants and nearly no variants of intermediate frequency. This was associated with a negative Tajima  $D$  (*Mixed* in fig. 2). This pattern is the joint effect of the excess of very small external branches in the highly sampled group and the long internal branches linking the remaining groups in the reconstructed genealogy.

### Joint Effects of Selection, Recombination, and Sampling Bias

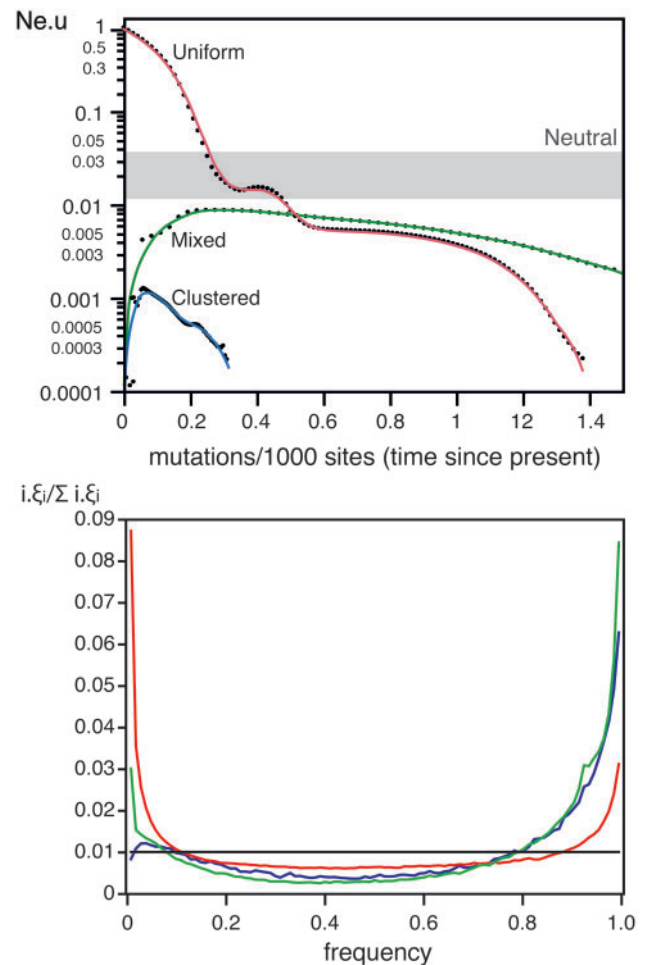
We then studied the joint effect of sampling biases, recombination, and weak selection on skyline plots and SFS (as shown before, strong selection rapidly erases genetic diversity in the simulations). The increase in  $N_e u$  values in skyline plots inferred under uniform sampling bias was highly amplified when weak selection and recombination were also present, rising by almost four orders of magnitude (fig. 6). The SFS of these simulations showed a large excess of rare variants and a small excess of very frequent ones.

Clustered sampling of populations enduring recombination and weak selection resulted in skyline plots with a rapid increase in  $N_e u$ , which then rapidly dropped to values very close to the initial ones. This process mimics initial strong population expansion, followed by very recent strong population contraction. The SFS showed a slight excess of rare variants and a large excess of frequent ones.

Finally, the skyline plots of simulations with mixed sampling, recombination, and weak selection showed a steady increase in  $N_e u$  and then a sharp decrease near the present. These patterns are also akin to the effects caused by ancient population expansions and recent population contractions. The SFS of these simulations showed an excess of both rare and frequent variants, with few intermediate values.

### Analysis of the *E. coli* Core Genome

The parameters of fitness effects used in the simulations were measured on *E. coli* in the laboratory. It might be argued that these parameters are not representative of the effects observed in structured locally adapted natural populations. To assess the imprint of natural selection in *E. coli* we built its core genome (see section “Methods”). The analysis of the polymorphism in the  $\sim 1.3$  million positions of the alignment of *E. coli* core genes, showed a pervasive pattern of purifying selection as expected from the simulations (fig. 7). Indeed, the ratio between the rates of nonsynonymous and synonymous substitutions ( $dN/dS$ ) was significantly lower than one for all pairwise comparisons with sufficient polymorphism.

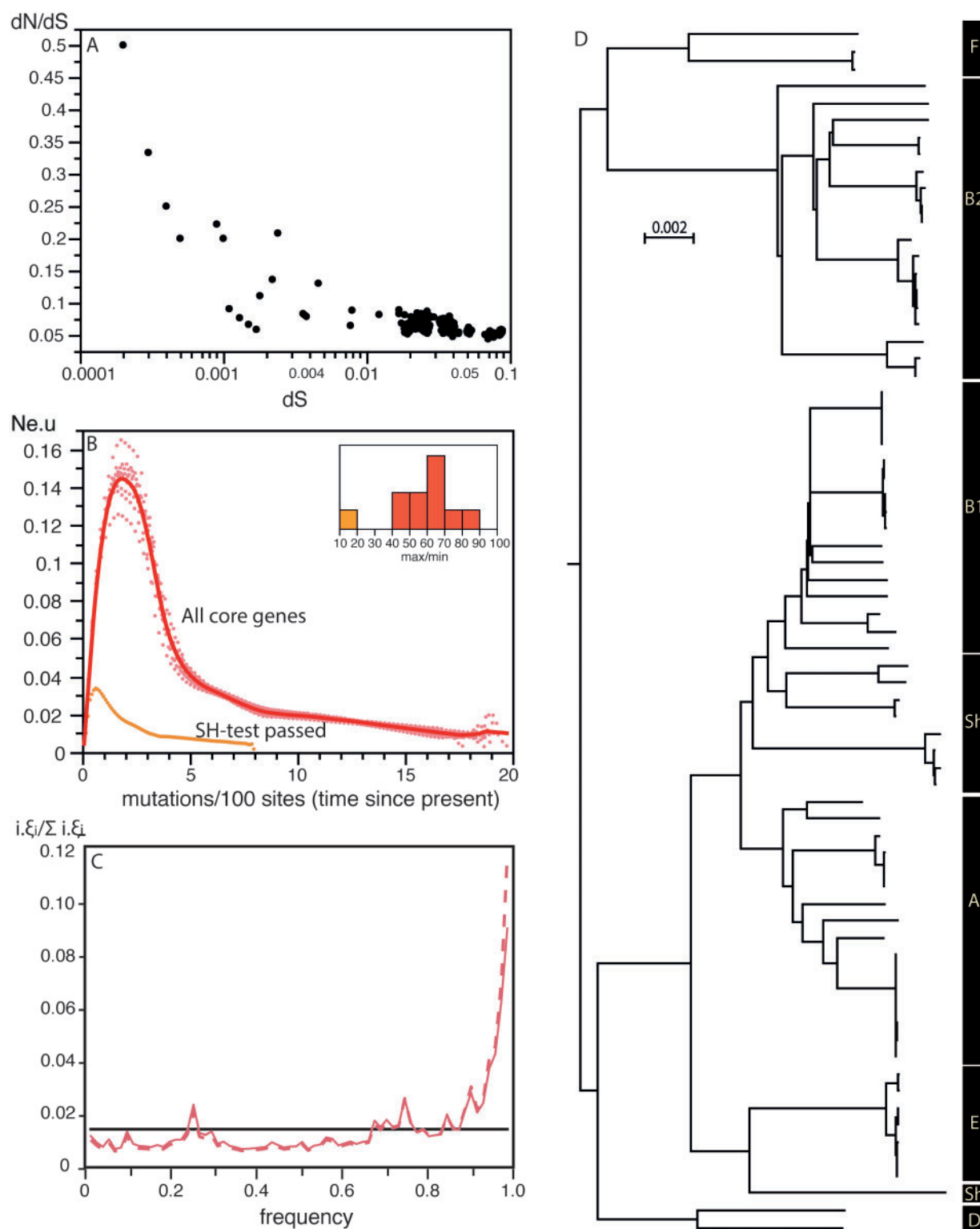


**FIG. 6.** Top Skyline plots for clustered, uniform and mixed sampling on simulations with weak selection and recombination (each point is an average of the ten simulations). The grey box indicates the range of variation of the *Neutral* simulations in figure 1. Bottom Average SFS for the same three datasets (1,000 simulations). Colors match the same datasets in both plots.

Importantly, when  $dS$  was higher than  $1/5,000$  the value of  $dN/dS$  was always smaller than 0.5. Multi-locus sequence typing (MLST) analyses use  $\sim 5$  kb of sequenced data and thus only start becoming informative when there is more than one SNP per 5 kb. At this level of divergence, the values of  $dN/dS$  show that the distribution of polymorphism is already imprinted by natural selection, precluding the use of MLST to make demographic inferences using skyline plots.

We then made ten random samples of 10% of the core genome positions to produce ten skyline plots for *E. coli*. The results were highly concordant between samples, showing a pattern of increase in  $N_e u$  followed by a sudden drop for times closer to the present (fig. 7). The SFS of the *E. coli* core genome showed a strong over-representation of very frequent variants (fig. 7). We then restricted our analysis to genes of the core genome with individual phylogenies not significantly different from those of the concatenate of the core genes. We found that the topologies of the reconstructed trees of 1,146 of the 1,371 core genes were significantly different from the one of the core genome ( $P < 0.01$ ,





**FIG. 7.** Analysis of the core genome of *E. coli*. (A) Values of  $dN/dS$  versus  $dS$ . Each point represents a comparison between two strains using the concatenate of alignments of genes of the core genome. (B) Skyline plot. We made ten analyses of the dataset by randomly sampling each time a tenth of the core genome. The orange line represents the skyline of the concatenate of genes with reconstructed genealogies not significantly different from those of the core genome (passed the SH test at  $P < 0.01$ ). The inset represents the ratio between the maximum and minimum values of  $N_e \cdot u$  for the 11 skyline plots (10 with the 1/10th samples of the core genome and one with the analysis of the concatenate of genes passing the SH-test). (C) The observed SFS is indicated in dashed red line, the corrected SFS (with Kimura's two-parameter model) is indicated in solid red line. The horizontal black line indicates the neutral expectation. The corrected SFS with the JC69 model (not shown here) is similar to the SFS corrected with Kimura's two-parameter model except for the last point, which is slightly higher. (D) *E. coli* distance-based phenetic tree with the major clades indicated on the right. A similar tree indicating all strains used in the analysis is in [supplementary figure S6, Supplementary Material online](#).

Shimodaira–Hasegawa [SH] test). This analysis confirmed that the vast majority of genes in the genome are significantly affected by recombination, in spite of the low estimated rate of recombination in *E. coli*. We used the remaining 225 genes to build a skyline plot. This showed qualitatively identical trends, but less striking variations (fig. 7B).

Together, these results are consistent with a mixture of strong purifying selection and recombination producing patterns akin to demographic expansion in *E. coli* skyline plots. The excess of high-frequency variants observed in the unfolded SFS might be due to hitchhiking effects, appearing under strong selection and recombination. However, one cannot exclude the possibility that part of this excess might result from misoriented polymorphisms (polymorphisms for which the ancestral allele was wrongly assigned) (Baudry and Depaulis 2003), since corrections tend to lower this excess (see section “Methods” and fig. 7C). Alternatively, a mixed sampling bias could produce a drop in  $N_e u$  for the most recent times in skyline plots and a large excess of high-frequency variants in SFS. To test this hypothesis we built a phenetic tree for *E. coli* using a distance-based method (to minimize reconstruction artifacts associated with recombination). The analysis of this tree does not support the existence of a very strong mixed sampling bias (fig. 7).

## Discussion

Recent advances in the analysis of genetic data using coalescent theory have the potential to unravel many novel aspects of microbial population genetics. The limitations of the underlying models are well known from the theoretical point of view (Frost et al. 2014). However, at the beginning of this work it was unclear if these limitations could compromise the use of such approaches to analyze bacterial data. Our study suggests that neglecting the effect of natural selection, recombination, and sampling biases may severely affect conclusions from phylodynamics analyses. These results are likely to be applicable to other phyla where these effects are important. An important effect that we have not quantified in this study concerns population structure, which tends to produce patterns akin to population contraction (Pannell 2003). Unfortunately, we could not study them due to current lack of modeling frameworks for simulating bacterial population structure. Previous studies have confirmed that animal population structure leads to distortions in skyline plots (Heller et al. 2013).

Some of the studies in table 1 tried to eliminate the effect of recombination by removing detectable recombination tracts from the analysis. Using ClonalFrame, we obtained even worst distortions in skyline plots. Similar results were previously found for the removal of homoplasies (Hedge and Wilson 2014). While we cannot offer a clear explanation for this observation, we presume it is caused by the removal of only certain specific types of recombination events (or polymorphism) from the data. Interestingly, the analysis of *E. coli* genomes suggests that removing all genes whose trees are incongruent with that of the core genome (SH test) attenuates the effect of

recombination. The reasons for this, and the consequences of removing these sequences, will require further study. Yet, the relative apparent success of this method might just derive from the bias of the SH test toward removing the recombining genes producing genealogies incompatible with the average genealogy of the core genome (while leaving for further analysis those that are compatible with this genealogy). This is expected to decrease the bias toward higher coalescent rates closer to the TMRCA. Importantly, the expectation of the SFS is insensitive to the presence of recombination and can be used to analyze genomic data deeply imprinted by recombination.

Previous theoretical studies suggested that selection on mutations of mild deleterious effect might not distort genealogies. This might explain why none of the studies in table 1 assessed the effect of natural selection on demographic inference. Yet, using population genetics parameters of *E. coli*, and even using much smaller values for  $N_e s$ , we found striking distortions in skyline plots.

We observed very frequent selective sweeps in the simulations with the selection parameters from *E. coli*. It must be emphasized that the high genetic diversity of the *E. coli* core genome is not fully consistent with such a succession of sweeps. However, it could be compatible with frequent soft sweeps, as recently described in *E. coli* adaptation to the mouse gut (Barroso-Batista et al. 2014). It would also be compatible with sweeps associated with local adaptation of certain lineages (Cohan and Perry 2007), or negative-frequency-dependent selection (Takeuchi et al. 2015). Finally, the existence of abundant strongly adaptive mutations in *E. coli* is consistent with previous results showing that a large fraction of amino acid substitutions between the *E. coli* and *Salmonella* lineages have been fixed by positive selection (Charlesworth and Eyre-Walker 2006).

To benefit from the power of coalescent-based approaches, one must find ways of controlling the distortions produced by selection on reconstructed genealogies. Unfortunately, practical and efficient ways of using the coalescent with selection are not yet available. Meanwhile, some simple controls might allow to identify or even estimate the effect of selection on demographic inference. For example, synonymous and nonsynonymous changes are very differently affected by selection, in spite of codon usage (Sharp et al. 2010), and partitioning the data in these two categories could shed light on the effect of selection on skyline plots and SFS. Comparisons between highly expressed and weakly expressed genes may also be informative since the former endure more intense selection for both synonymous and nonsynonymous substitutions (Rocha and Danchin 2004). Very recent polymorphism is relatively less imprinted by selection (Ho et al. 2005; Rocha et al. 2006), and might produce less biased patterns in skyline plots. Interestingly, the only published skyline plots in table 1 showing population contractions were based on samples with very short TMRCA (table 1). Unfortunately, the analysis of  $dN/dS$  in *E. coli* shows that even the very recent polymorphism was

affected by purifying selection (fig. 7). Skyline plots on larger time spans are even more imprinted by natural selection and interpretation purely in terms of demographic changes should not be made in the absence of control for natural selection.

Random sampling is a key underlying hypothesis of most statistical methods for the inference of demographic changes. However, funding agencies often stimulate researchers to focus on particular bacterial sub-populations of societal interest. This renders random sampling effectively impossible and might explain why surveys of microbial populations rarely explicit the statistical design of the sampling. As an example, despite the fact that *E. coli* is a commensal present in most warm-blooded animals, the vast majority of complete genomes available for this species are from strains pathogenic to humans. Since host-association, virulence, and antibiotic resistance vary between lineages of a species, over-sampling isolates of direct interest in terms of public health almost inevitably leads to statistical biases. Our results show that three common sampling strategies can severely bias the inference of demographic changes, especially in the presence of selection and recombination. Skyline plots studies of populations where these factors are important can exhibit almost any possible pattern of change.

The sampling of sub-groups of a population led to reconstructed genealogies suggesting recent population expansion. These results show that sub-trees of coalescent trees have distributions of coalescent rates different from those of the population tree. Hence, sampling a sub-population inevitably produces biased skyline plots. This brings to the fore the importance of precisely defining bacterial populations when inferring demographic changes using coalescent rates. The study of past demographic changes in microbial populations requires the use of adapted sampling techniques. Many such techniques have been developed in ecology (Young and Young 2013), even if their implementation poses technical challenges in microbiological research.

Many approaches alternative to skyline plots allow the inference of demographic changes. They all have specific advantages and disadvantages and their combination might facilitate the use of the available sequence data to make demographic inference. Lack of obvious neutral sites in bacteria renders difficult the establishment of demographic models independent of selection. Nevertheless,  $dN/dS$ -based approaches can be used to assess if natural selection has imprinted sequence data (although care must be taken to check if absence of evidence of selection is not due to lack of statistical power). Furthermore, the expectations of the SFS are insensitive to recombination and to uniform sampling when there is no selection or recombination. They are also less affected by differences in the intensity of natural selection, although in case of pervasive selection with recombination, the SFS shape will correspond to the predictions of multiple merger coalescent models (Tellier and Lemaire 2014). Therefore, joint analyses of skyline plots, detection of recombination, SFS (and derived statistics),  $dN/dS$ , and other population genetics methods are necessary to accurately infer changes in microbial demography.

## Methods

### Simulations

We made 1,000 simulations for each set of parameters. Simulations were done using *SFS\_code*, which implements a generalized version of the Wright–Fisher forward population genetic model allowing finite-site mutation models with selection, recombination, and demography (Hernandez 2008). The typical simulation was done using a population of haploids with  $N_e = 1,000$  individuals and one single genetic locus of 20,000 nucleotides. The length of the locus was chosen in order to be much larger than the average recombination tract in *E. coli* (~542 nt) (Didelot et al. 2012). In simulations under selection and recombination, we increased the length of the locus to 200,000 nucleotides, to obtain a sufficient number of polymorphic sites for further analyses. For simplicity, all nucleotides were included at similar frequencies and the substitution model was set to JC69 (equal mutation rates between all pairs of nucleotides) (Jukes and Cantor 1969). We used a 3-point mass model for selection (including negative, positive, and null values for the selection coefficient) (table 2). Modeling positive and purifying selection as two exponential distributions provides qualitatively similar results (but often produced numerical instabilities). Recombination was introduced exclusively as gene conversion (no crossovers allowed) in populations simulated as diploids (due to the constraints of the software). In this case, only half of the loci were used (1,000). The simulations were done using population scaled parameters accounting for the  $N_e$  of *E. coli* (table 2). Under these conditions, the size of the population effectively simulated does not affect the outcome of the analysis (Hernandez et al. 2007). In all cases, except those concerning sampling biases, we took 100 individuals from each final simulated population for further analysis.

### Simulations of Biased Sampling

When analyzing biased sampling we took all 1,000 individuals from the final simulated populations. These sequences were used to build a distance matrix with FastTree v 2.1.7 using default parameters and the option-makematrix (Price et al. 2009). This distance matrix was then partitioned into clusters around medoids, a more robust version of K-means (Reynolds et al. 2006), using R. We simulated biased sampling of 100 individuals from the population in three ways. We simulated uniform distribution by picking one individual per cluster in an analysis where the population was clustered in 100 groups. We simulated mixed sampling bias by picking one individual per cluster for a total of ten individuals and then picking the remaining 90 individuals from one single cluster (analysis where the population was clustered in ten groups). We simulated clustered distribution by selecting all 100 individuals from a single cluster (analysis where the population was clustered in ten groups). It is important to note that a cluster obtained with this method may not exactly correspond to a monophyletic group as described in figure 5. The goal of our approach was to mimic the typical identification of clusters of bacterial groups used to select strains for sequencing, which are based on relatively imprecise methods (MLST or PFGE).



## Analyses of Reconstructed Genealogies

We analyzed sequences using the generalized skyline plot model in BEAST with piecewise-linear modeling of the population size (skyline.popSize priors: initial =  $3.2 \times 10^{-4}$ , upper = 100, lower = 0), using the HKY model (the mutation model was parameterized so that its stationary frequencies were the empirical frequencies) (Hasegawa et al. 1985), setting a tight prior for  $k$  (lognormal, initial = 1, logMean = 0, Logstdev = 0.25), a strict molecular clock (as used in the simulations), and 30,000,000 iterations (sampling every 3,000 iterations). For simulations involving selection we made 300,000,000 iterations. The effective sample size (ESS) values were checked using Tracer and the runs were accepted when the ESS was higher than 200 for all parameters with eventual exception for some skyline.population parameters (as suggested by the manual of BEAST—[Drummond and Rambaut 2007]). Analyses resulting in poor ESS values were discarded and re-run. Tracer was used to compute all skyline plots except those made after the ClonalFrame analysis (see below). Given the computational cost of these analyses we only analyzed ten simulations per condition. However, the results were very consistent between simulations resulting in kernel fits with high  $R^2$  (see text).

## Analysis of the SFS

SFS were generated from random samples of 100 individuals. The mean SFS was calculated using 1,000 simulations. The exact ancestral state of each SNP was obtained using SFS\_code. The SFS of the simulations were thus unfolded. For a better representation of the results, the SFS were transformed as follows. Let  $\xi_i$  denote the number of polymorphic sites at frequency  $\frac{i}{n}$  in the sample of size  $n$ . We plot  $i \cdot \xi_i$  for  $i \in [1, n - 1]$ , normalized by its sum, which is an unbiased estimator of the (supposedly unknown) mutation rate, often noted  $\theta$  under the standard neutral model. Thus, the transformed SFS has a flat expectation under the standard neutral model, due to the well-known fact that  $E[\xi_i] = \frac{\theta}{i}$ .

For the analysis of *E. coli* data, the ancestral state is unknown and we used outgroup sequences. To correct for potential ancestral misorientations (i.e., when the nucleotide of the outgroup is erroneously inferred as the ancestral state), we calculated the probability of misorientations, using sites for which the outgroup nucleotide is different from the two nucleotides of the SNP (see Baudry and Depaulis 2003; Hernandez et al. 2007).

If  $q$  is the probability that the outgroup nucleotide is identical to the ancestral nucleotide, we have in expectation:

$$\xi_k^{\text{obs}} = \xi_k q + \xi_{n-k} (1 - q) \text{ for } k \in [1, n - 1],$$

where  $\xi_k^{\text{obs}}$  is the number of polymorphic sites at frequency  $\frac{k}{n}$  before correction and  $\xi_k$  the real value.

We denoted by  $S$  the event that a given site is segregating, and by  $U$  the event that it is segregating and the outgroup nucleotide is different from the two nucleotides of the SNP. On one hand,  $P(U | S)$  is easily estimated by the proportion  $x$  of sites that are segregating and yet have a different outgroup nucleotide. On the other hand, under the JC69 model of

mutation,  $P(U \cap S) = 2q P(S)$ , neglecting the case when the ancestral nucleotide is different from the other three. Combining these two arguments we can estimate  $q$  by  $x/2$ .

Once  $q$  is estimated from the data, we can calculate the corrected values of the SFS:

$$\xi_k = \frac{\xi_k^{\text{obs}} - \xi_{n-k}^{\text{obs}} (1 - q)}{2q - 1} \text{ for } k \in [1, n - 1].$$

We estimated  $q$  with two corrections, depending on the mutation model. Under the JC69 model of mutation,  $q = 0.960$ . Under Kimura's two parameters model (Kimura 1980), taking into account the transition and transversion rates,  $q = 0.947$  (Baudry and Depaulis 2003).

## ClonalFrame Analysis and Subsequent Skyline Plot

ClonalFrame was used with default parameters on the results of ten simulations with recombination, no selection and no sampling bias. All ClonalFrame outputs were imported in the ClonalFrame GUI (Didelot and Falush 2007). The convergence of MCMC traces was visually assessed. ClonalFrame outputs ultra-metric trees with multifurcations, but bifurcating trees are necessary to compute skyline plots. Hence, for each simulation, we exported the recombination-free distance matrix and used the R package *phangorn* to construct the UPGMA trees (Schliep 2011). We computed generalized skyline plots using the skyline function of the *ape* package (Paradis et al. 2004). The AIC criterion was applied to find the optimal  $\epsilon$  spline parameter.

## Analysis of *E. coli* Genome

We downloaded from RefSeq in November 2013 (Tatusova et al. 2015) the 62 genomes of *E. coli*, the nine genomes of *Shigella* spp. (in fact *E. coli* strains—Ochman et al. 1983) and the genome of *E. fergusonii* (the outgroup). Pairs of orthologous genes between two genomes were defined as bi-directional best hits, with  $>80\%$  similarity in protein sequence,  $<20\%$  difference in gene size, present within similar genetic neighborhoods (see Touchon et al. 2009 for details). The list of the core genome was defined as the intersection of all lists of pairwise analyses and included 1,371 genes. Genes from the same family of the core genome were aligned in protein sequence using MUSCLE v3.8 (default parameters, Edgar 2004) and back translated to DNA. These alignments were concatenated, making a total of 1,349,016 positions. They were used to compute the pairwise values of  $dS$ ,  $dN$  and  $dN/dS$  between *E. coli* genomes using codeML from PAML v4 (parameters: runmode = -2; CodonFreq = 2; clock = 0; model = 2) (Yang 2007). Comparisons between very closely related isolates (i.e., with no single synonymous or nonsynonymous substitution in the core genome) were discarded.

## SH Tests and Phenetic Tree

We built a phylogenetic tree of the core genome of *E. coli* using IQ-Tree (Nguyen et al. 2015) with the option to search for the best substitution model. The best model based on the BIC criterion was GTR + I + G4. For each gene we used IQ-Tree to make the SH test (1,000 replicates) using as a

reference tree the core genome tree. The phenetic tree in figure 7 was built using BIONJ (Gascuel 1997) from a distance matrix computed using TreePuzzle with the model GTR + I + G4 (Schmidt et al. 2002).

## Supplementary Material

Supplementary figures S1–S6 are available at *Molecular Biology and Evolution* online (<http://www.mbe.oxfordjournals.org/>).

## Acknowledgments

We thank Sylvain Brisse, Adam Eyre-Walker, and Guillaume Laval for comments on an earlier version of this manuscript. This project was financed by the Centre National de la Recherche Scientifique (CNRS) and the Institut Pasteur. G.A. and M.L. acknowledge support from the grant ANR-12-BSV7-0012 Demochips from the Agence Nationale de la Recherche (France). M.L. is funded by the PhD program Interfaces for Life of the University Pierre and Marie Curie (Paris). C.B. is funded by the PhD program Complexité du Vivant of the University Pierre and Marie Curie (Paris).

## References

- Achaz G. 2009. Frequency spectrum neutrality tests: one for all and all for one. *Genetics* 183:249–258.
- Adams AM, Hudson RR. 2004. Maximum-likelihood estimation of demographic parameters using the frequency spectrum of unlinked single-nucleotide polymorphisms. *Genetics* 168:1699–1712.
- Azarian T, Ali A, Johnson JA, Mohr D, Prosperi M, Veras NM, Jubair M, Strickland SL, Rashid MH, Alam MT, et al. 2014. Phylodynamic analysis of clinical and environmental *Vibrio cholerae* isolates from Haiti reveals diversification driven by positive selection. *MBio* 5:e01824–14.
- Balbi KJ, Rocha EP, Feil EJ. 2009. The temporal dynamics of slightly deleterious mutations in *Escherichia coli* and *Shigella* spp. *Mol Biol Evol* 26:345–355.
- Barroso-Batista J, Sousa A, Lourenco M, Bergman ML, Sobral D, Demengeot J, Xavier KB, Gordo I. 2014. The first steps of adaptation of *Escherichia coli* to the gut are dominated by soft sweeps. *PLoS Genet* 10:e1004182.
- Bart MJ, Harris SR, Advani A, Arakawa Y, Bottero D, Bouchez V, Cassidy PK, Chiang CS, Dalby T, Fry NK, et al. 2014. Global population structure and evolution of *Bordetella pertussis* and their relationship with vaccination. *MBio* 5:e01074.
- Baudry E, Depaulis F. 2003. Effect of misoriented sites on neutrality tests with outgroup. *Genetics* 165:1619–1622.
- Beaumont MA, Zhang W, Balding DJ. 2002. Approximate Bayesian computation in population genetics. *Genetics* 162:2025–2035.
- Bobay LM, Traverse CC, Ochman H. 2015. Impermanence of bacterial clones. *Proc Natl Acad Sci U S A* 112:8893–8900.
- Braverman JM, Hudson RR, Kaplan NL, Langley CH, Stephan W. 1995. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics* 140:783–796.
- Brenner DJ, Krieg NR, Staley JT. 2005. Bergey's manual of systematic bacteriology. New York: Springer.
- Charlesworth J, Eyre-Walker A. 2006. The rate of adaptive evolution in enteric bacteria. *Mol Biol Evol* 23:1348–1356.
- Cohen ML. 2000. Changing patterns of infectious disease. *Nature* 406:762–767.
- Cohan FM, Perry EB. 2007. A systematics for discovering the fundamental units of bacterial diversity. *Curr Biol* 17:R373–R386.
- Comas I, Coscolla M, Luo T, Borrell S, Holt KE, Kato-Maeda M, Parkhill J, Malla B, Berg S, Thwaites G, et al. 2013. Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. *Nat Genet* 45:1176–1182.
- Cornejo OE, Lefebvre T, Bitar PD, Lang P, Richards VP, Eilertson K, Do T, Beighton D, Zeng L, Ahn SJ, et al. 2013. Evolutionary and population genomics of the cavity causing bacteria *Streptococcus mutans*. *Mol Biol Evol* 30:881–893.
- Croucher NJ, Hanage WP, Harris SR, McGee L, van der Linden M, de Lencastre H, Sa-Leao R, Song JH, Ko KS, Beall B, et al. 2014. Variable recombination dynamics during the emergence, transmission and 'disarming' of a multidrug-resistant pneumococcal clone. *BMC Biol* 12:49.
- Davies MR, Holden MT, Coupland P, Chen JH, Venturini C, Barnett TC, Zakour NL, Tse H, Dougan G, Yuen KY, et al. 2015. Emergence of scarlet fever *Streptococcus pyogenes* emm12 clones in Hong Kong is associated with toxin acquisition and multidrug resistance. *Nat Genet* 47:84–87.
- Delaney NF, Balenger S, Bonneaud C, Marx CJ, Hill GE, Ferguson-Noel N, Tsai P, Rodrigo A, Edwards SV. 2012. Ultrafast evolution and loss of CRISPRs following a host shift in a novel wildlife pathogen, *Mycoplasma gallisepticum*. *PLoS Genet* 8:e1002511.
- Didelot X, Falush D. 2007. Inference of bacterial microevolution using multilocus sequence data. *Genetics* 175:1251–1266.
- Didelot X, Meric G, Falush D, Darling AE. 2012. Impact of homologous and non-homologous recombination in the genomic evolution of *Escherichia coli*. *BMC Genomics* 13:256.
- Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol* 7:214.
- Drummond AJ, Nicholls GK, Rodrigo AG, Solomon W. 2002. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics* 161:1307–1320.
- Drummond AJ, Rambaut A, Shapiro B, Pybus OG. 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol Biol Evol* 22:1185–1192.
- Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792–1797.
- Eldholm V, Monteserin J, Rieux A, Lopez B, Sobkowiak B, Ritacco V, Balloux F. 2015. Four decades of transmission of a multidrug-resistant *Mycobacterium tuberculosis* outbreak strain. *Nat Commun* 6:7119.
- Frost SDW, Pybus OG, Gog JR, Viboud C, Bonhoeffer S, Bedford T. 2014. Eight challenges in phylodynamic inference. *Epidemics* 10:88–92.
- Fu YX. 1997. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* 147:915–925.
- Fu R, Voordouw G. 1997. Targeted gene-replacement mutagenesis of *dcrA*, encoding an oxygen sensor of the sulfate-reducing bacterium *Desulfovibrio vulgaris* Hildenborough. *Microbiology* 143:1815–1826.
- Gallet R, Cooper TF, Elena SF, Lenormand T. 2012. Measuring selection coefficients below 10<sup>(-3)</sup>: method, questions, and prospects. *Genetics* 190:175–186.
- Gascuel O. 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol Biol Evol* 14:685–695.
- Grad YH, Lipsitch M. 2014. Epidemiologic data and pathogen genome sequences: a powerful synergy for public health. *Genome Biol* 15:538.
- Grenfell BT, Pybus OG, Gog JR, Wood JL, Daly JM, Mumford JA, Holmes EC. 2004. Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* 303:327–332.
- Guttman DS, Morgan RL, Wang PW. 2008. The evolution of the pseudomonads. In: Fatmi M, Collmer A, Iacobellis MS, Mansfield JW, Murillo J, Schaad NW, Ullrich M, editors. *Pseudomonas syringae* pathovars and related pathogens—identification, epidemiology and genomics. Dordrecht: Springer. p. 307–319.
- Hartl DL, Moriyma EN, Sawyer SA. 1994. Selection intensity for codon bias. *Genetics* 138:227–234.
- Hasegawa M, Kishino H, Yano T, Stabler RA, Dawson LF, Martin MJ, Holt KE, Seth-Smith HM, Quail MA, Rance R. 1985. Dating of the human-



- ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol.* 22:160–174.
- He M, Sebahia M, Lawley TD, Stabler RA, Dawson LF, Martin MJ, Holt KE, Seth-Smith HM, Quail MA, Rance R, et al. 2010. Evolutionary dynamics of *Clostridium difficile* over short and long time scales. *Proc Natl Acad Sci U S A.* 107:7527–7532.
- Hedge J, Wilson DJ. 2014. Bacterial phylogenetic reconstruction from whole genomes is robust to recombination but demographic inference is not. *MBio* 5:e02158.
- Heller R, Chikhi L, Siegmund HR. 2013. The confounding effect of population structure on Bayesian skyline plot inferences of demographic history. *PLoS One* 8:e62992.
- Hernandez RD. 2008. A flexible forward simulator for populations subject to selection and demography. *Bioinformatics* 24:2786–2787.
- Hernandez RD, Williamson SH, Bustamante CD. 2007. Context dependence, ancestral misidentification, and spurious signatures of natural selection. *Mol Biol Evol.* 24:1792–1800.
- Ho SY, Phillips MJ, Cooper A, Drummond AJ. 2005. Time dependency of molecular rate estimates and systematic overestimation of recent divergence times. *Mol Biol Evol.* 22:1561–1568.
- Ho SY, Shapiro B. 2011. Skyline-plot methods for estimating demographic history from nucleotide sequences. *Mol Ecol Resour.* 11:423–434.
- Holmes EC. 2007. Viral evolution in the genomic age. *PLoS Biol.* 5:e278.
- Holt KE, Baker S, Weill FX, Holmes EC, Kitchen A, Yu J, Sangal V, Brown DJ, Coia JE, Kim DW, et al. 2012. *Shigella sonnei* genome sequencing and phylogenetic analysis indicate recent global dissemination from Europe. *Nat Genet.* 44:1056–1059.
- Hughes AL. 2005. Evidence for abundant slightly deleterious polymorphisms in bacterial populations. *Genetics* 169:533–538.
- Jukes TH, Cantor CR. 1969. Evolution of protein molecules. Mammalian protein metabolism. New York: Academic Press. p. 21–132.
- Kibota TT, Lynch M. 1996. Estimate of the genomic mutation rate deleterious to overall fitness in *E. coli*. *Nature* 381:694–696.
- Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol.* 16:111–120.
- Kimura M. 1983. The neutral theory of molecular evolution. Cambridge: Cambridge University Press.
- Li LM, Grassly NC, Fraser C. 2014. Genomic analysis of emerging pathogens: methods, application and future trends. *Genome Biol.* 15:541.
- Liao P-C, Huang S. 2012. Patterns of microbial genetic diversity and the correlation between bacterial demographic history and geohistory. In: Caliskan M, editor. Genetic diversity in microorganisms. Shanghai: INTECH Open Access Publisher. p. 123–148.
- Liu X, Fu YX. 2015. Exploring population size changes using SNP frequency spectra. *Nat Genet.* 47:555–559.
- Lynch M. 2006. The origins of eukaryotic gene structure. *Mol Biol Evol.* 23:450–468.
- Martiny JB, Bohannan BJ, Brown JH, Colwell RK, Fuhrman JA, Green JL, Horner-Devine MC, Kane M, Krumins JA, Kuske CR, et al. 2006. Microbial biogeography: putting microorganisms on the map. *Nat Rev Microbiol.* 4:102–112.
- Mazet O, Rodriguez W, Chikhi L. 2015. Demographic inference using genetic data from a single individual: Separating population size variation from population structure. *Theor Popul Biol.* 104:46–58.
- Merker M, Blin C, Mona S, Duforet-Frebourg N, Lecher S, Willery E, Blum MG, Rusch-Gerdes S, Mokrousov I, Aleksic E, et al. 2015. Evolutionary history and global spread of the *Mycobacterium tuberculosis* Beijing lineage. *Nat Genet.* 47:242–249.
- Navarro A, Barton NH. 2002. The effects of multilocus balancing selection on neutral variability. *Genetics* 161:849–863.
- Neuhauser C, Krone SM. 1997. The genealogy of samples in models with selection. *Genetics* 145:519–534.
- Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol.* 32:268–274.
- Nielsen R, Beaumont MA. 2009. Statistical inferences in phylogeography. *Mol Ecol.* 18:1034–1047.
- Nielsen R, Wakeley J. 2001. Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics* 158:885–896.
- Nubel U, Dordel J, Kurt K, Strommenger B, Westh H, Shukla SK, Zemlickova H, Leblois R, Wirth T, Jombart T, et al. 2010. A timescale for evolution, population expansion, and spatial spread of an emerging clone of methicillin-resistant *Staphylococcus aureus*. *PLoS Pathog.* 6:e1000855.
- O'Fallon BD, Seger J, Adler FR. 2010. A continuous-state coalescent and the impact of weak selection on the structure of gene genealogies. *Mol Biol Evol.* 27:1162–1172.
- Ochman H, Whittam TS, Caugant DA, Selander RK. 1983. Enzyme polymorphism and genetic population structure in *Escherichia coli* and *Shigella*. *J Gen Microbiol.* 129:2715–2726.
- Pannell JR. 2003. Coalescence in a metapopulation with recurrent local extinction and recolonization. *Evolution* 57:949–961.
- Paradis E, Claude J, Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20:289–290.
- Pepperell CS, Casto AM, Kitchen A, Granka JM, Cornejo OE, Holmes EC, Birren B, Galagan J, Feldman MW. 2013. The role of selection in shaping diversity of natural *M. tuberculosis* populations. *PLoS Pathog.* 9:e1003543.
- Perez-Losada M, Crandall KA, Zenilman J, Viscidi RP. 2007. Temporal trends in gonococcal population genetics in a high prevalence urban community. *Infect Genet Evol.* 7:271–278.
- Perfeito L, Fernandes L, Mota C, Gordo I. 2007. Adaptive mutations in bacteria: high rate and small effects. *Science* 317:813–815.
- Price MN, Dehal PS, Arkin AP. 2009. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol.* 26:1641–1650.
- Prosperi M, Veras N, Azarian T, Rathore M, Nolan D, Rand K, Cook RL, Johnson J, Morris JG Jr, Salemi M. 2013. Molecular epidemiology of community-associated methicillin-resistant *Staphylococcus aureus* in the genomic era: a cross-sectional study. *Sci Rep.* 3:1902.
- Przeworski M, Charlesworth B, Wall JD. 1999. Genealogies and weak purifying selection. *Mol Biol Evol.* 16:246–252.
- Pybus OG, Rambaut A, Harvey PH. 2000. An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics* 155:1429–1437.
- Ramos-Onsins SE, Rozas J. 2002. Statistical properties of new neutrality tests against population growth. *Mol Biol Evol.* 19:2092–2100.
- Reynolds AP, Richards G, de la Iglesia B, Rayward-Smith VJ. 2006. Clustering rules: a comparison of partitioning and hierarchical clustering algorithms. *J Math Model Algorithms.* 5:475–504.
- Rocha E, Danchin A. 2004. An analysis of determinants of amino acids substitution rates in bacterial proteins. *Mol Biol Evol.* 21:108–116.
- Rocha E, Smith J, Hurst L, Holden M, Cooper J, Smith N, Feil E. 2006. Comparisons of dN/dS are time dependent for closely related bacterial genomes. *J Theor Biol.* 239:226–235.
- Rocha EPC, Feil EJ. 2010. Mutational patterns cannot explain genome composition: are there any neutral sites in the genomes of bacteria? *Plos Genetics* 6:e1001104.
- Roumagnac P, Weill FX, Dolecek C, Baker S, Brisse S, Chinh NT, Le TA, Acosta CJ, Farrar J, Dougan G, et al. 2006. Evolutionary history of *Salmonella typhi*. *Science* 314:1301–1304.
- Sanchez-Buso L, Comas I, Jorques G, Gonzalez-Candelas F. 2014. Recombination drives genome evolution in outbreak-related *Legionella pneumophila* isolates. *Nat Genet.* 46:1205–1211.
- Sarkar SF, Guttman DS. 2004. Evolution of the core genome of *Pseudomonas syringae*, a highly clonal, endemic plant pathogen. *Appl Environ Microbiol.* 70:1999–2012.
- Schierup MH, Hein J. 2000. Consequences of recombination on traditional phylogenetic analysis. *Genetics* 156:879–891.
- Schliep KP. 2011. phangorn: phylogenetic analysis in R. *Bioinformatics* 27:592–593.
- Schmidt HA, Strimmer K, Vingron M, von Haeseler A. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* 18:502–504.
- Sharp PM, Emery LR, Zeng K. 2010. Forces that influence the evolution of codon bias. *Philos Trans R Soc Lond B Biol.* 365:1203–1212.

- Stegger M, Wirth T, Andersen PS, Skov RL, De Grassi A, Simoes PM, Tristan A, Petersen A, Aziz M, Kiil K, et al. 2014. Origin and evolution of European community-acquired methicillin-resistant *Staphylococcus aureus*. *MBio* 5:e01044–e01014.
- Tajima F. 1989a. The effect of change in population size on DNA polymorphism. *Genetics* 123:597–601.
- Tajima F. 1989b. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123:585–595.
- Takeuchi N, Cordero OX, Koonin EV, Kaneko K. 2015. Gene-specific selective sweeps in bacteria and archaea caused by negative frequency-dependent selection. *BMC Biol.* 13:20.
- Tatusova T, Ciufo S, Federhen S, Fedorov B, McVeigh R, O'Neill K, Tolstoy I, Zaslavsky L. 2015. Update on RefSeq microbial genomes resources. *Nucleic Acids Res.* 43:D599–D605.
- Tazi L, Perez-Losada M, Gu W, Yang Y, Xue L, Crandall KA, Viscidi RP. 2010. Population dynamics of *Neisseria gonorrhoeae* in Shanghai, China: a comparative study. *BMC Infect. Dis.* 10:13.
- Tellier A, Lemaire C. 2014. Coalescence 2.0: a multiple branching of recent theoretical developments and their applications. *Mol Ecol.* 23:2637–2652.
- Tenaillon O, Skurnik D, Picard B, Denamur E. 2010. The population genetics of commensal *Escherichia coli*. *Nat Rev Microbiol.* 8:207–217.
- Thornton K. 2005. Recombination and the properties of Tajima's D in the context of approximate-likelihood calculation. *Genetics* 171:2143–2148.
- Touchon M, Cury J, Yoon E-J, Krizova L, Cerqueira GC, Murphy C, Feldgarden M, Wortman J, Clermont D, Lambert T, et al. 2014. The genomic diversification of the whole acinetobacter genus: origins, mechanisms, and consequences. *Genome Biol Evol.* 6:2866–2882.
- Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, Bingen E, Bonacorsi S, Bouchier C, Bouvet O, et al. 2009. Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS Genet.* 5:e1000344.
- Treangen TJ, Ambur OH, Tonjum T, Rocha EPC. 2008. The impact of the neisserial DNA uptake sequences on genome evolution and stability. *Genome Biol.* 9:R60.
- Volz EM, Koelle K, Bedford T. 2013. Viral phylodynamics. *PLoS Comput Biol.* 9:e1002947.
- Vos M, Didelot X. 2009. A comparison of homologous recombination rates in bacteria and archaea. *ISME J.* 3:199–208.
- Wall JD. 1999. Recombination and the power of statistical tests of neutrality. *Genet Res.* 74:65–79.
- Weinert LA, Chaudhuri RR, Wang J, Peters SE, Corander J, Jombart T, Baig A, Howell KJ, Vehkala M, Valimaki N, et al. 2015. Genomic signatures of human and animal disease in the zoonotic pathogen *Streptococcus suis*. *Nat Commun.* 6:6740.
- Wielgoss S, Barrick JE, Tenaillon O, Cruveiller S, Chane-Woon-Ming B, Medigue C, Lenski RE, Schneider D. 2011. Mutation rate inferred from synonymous substitutions in a long-term evolution experiment with *Escherichia coli*. *G3* 1:183–186.
- Williamson S, Orive ME. 2002. The genealogy of a sequence subject to purifying selection at multiple sites. *Mol Biol Evol.* 19:1376–1384.
- Wirth T, Falush D, Lan R, Colles F, Mensa P, Wieler LH, Karch H, Reeves PR, Maiden MC, Ochman H, et al. 2006. Sex and virulence in *Escherichia coli*: an evolutionary perspective. *Mol Microbiol.* 60:1136–1151.
- Wirth T, Morelli G, Kusecek B, van Belkum A, van der Schee C, Meyer A, Achtman M. 2007. The rise and spread of a new pathogen: seroresistant *Moraxella catarrhalis*. *Genome Res.* 17:1647–1656.
- Wong VK, Baker S, Pickard DJ, Parkhill J, Page AJ, Feasey NA, Kingsley RA, Thomson NR, Keane JA, Weill FX, et al. 2015. Phylogeographical analysis of the dominant multidrug-resistant H58 clade of *Salmonella* Typhi identifies inter- and intracontinental transmission events. *Nat Genet.* 47:632–639.
- Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 24:1586–1591.
- Young LJ, Young J. 2013. Statistical ecology. New York: Springer Science & Business Media.
- Zhou Z, McCann A, Weill FX, Blin C, Nair S, Wain J, Dougan G, Achtman M. 2014. Transient Darwinian selection in *Salmonella enterica* serovar Paratyphi A during 450 years of global spread of enteric fever. *Proc Natl Acad Sci U S A.* 111:12199–12204.

## 2.3 Annexes

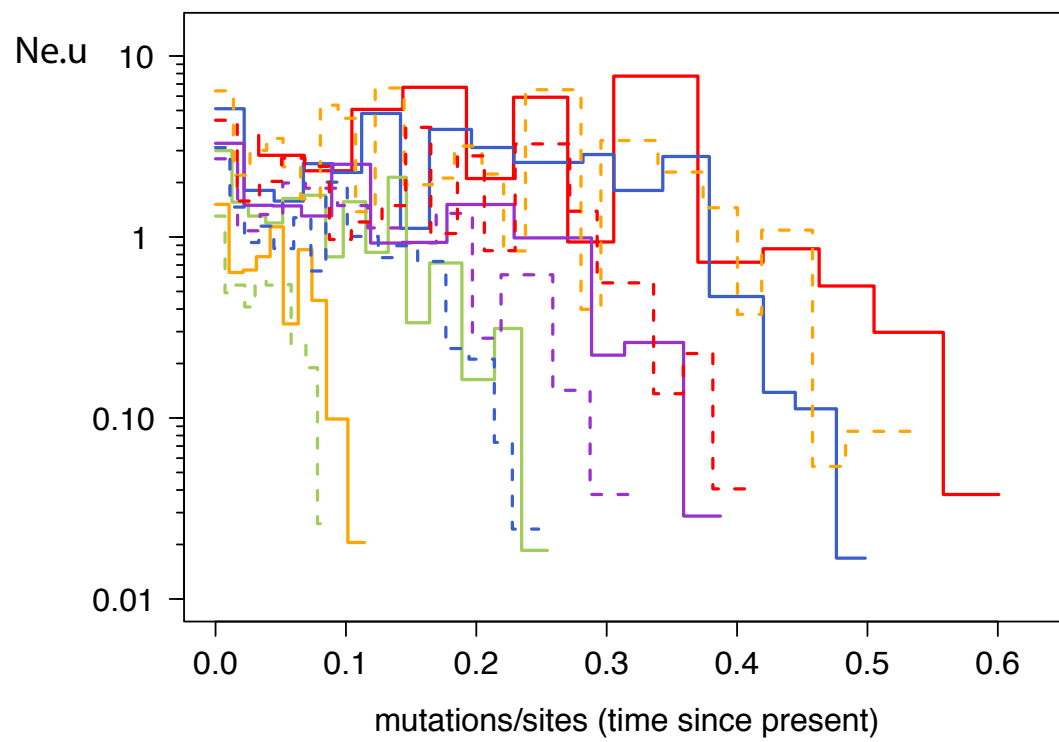


FIGURE S1 – Skyline plots from the 10 simulations with recombination after analysis with ClonalFrame.

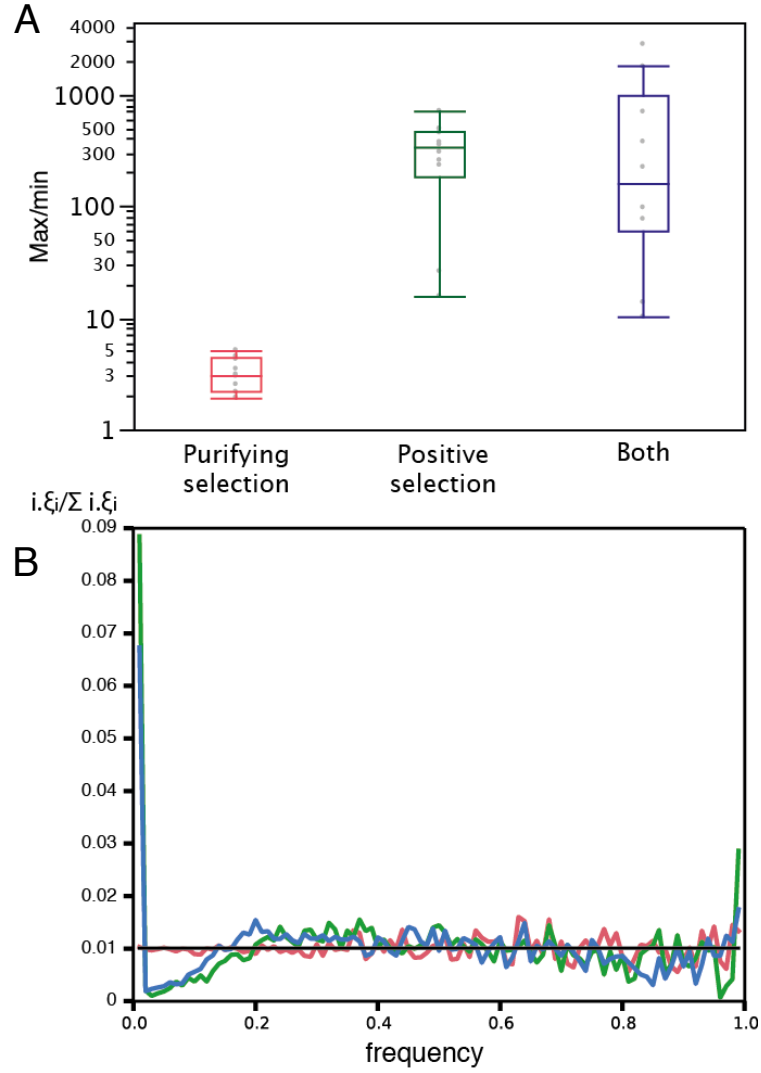


FIGURE S2 – Analysis of the components of simulations using selection. A. Boxplots of the ratios between the maximal and minimal  $N_e u$  values for skyline plots, across the different types of simulations using the parameters for strong selection (10 simulations each). The hypothesis that the distributions are similar is rejected ( $P < 0.001$ , Wilcoxon test), because of the purifying selection set, which is different from the others ( $P < 0.001$  same tests). B. SFS for the same conditions (1000 simulations each).

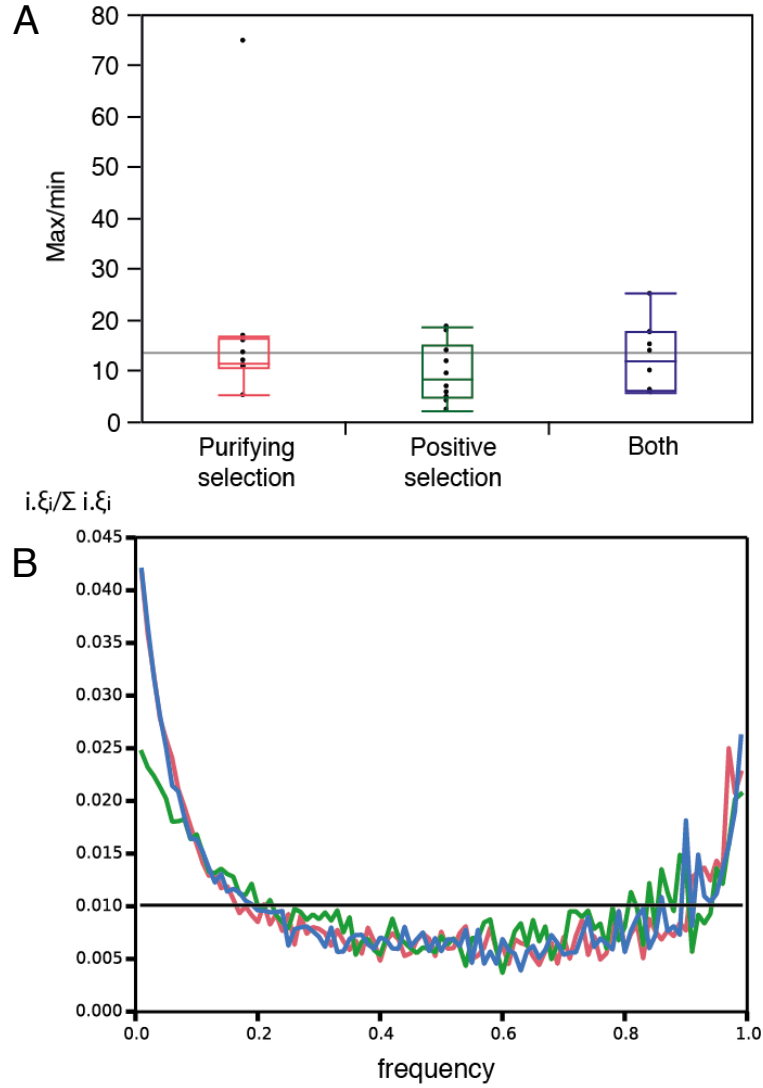


FIGURE S3 – Analysis of the components of weak selection simulations. A. Boxplots of the ratios between the maximal and minimal  $N_e u$  values for skyline plots, across the different types of simulations using the parameters for weak selection (10 simulations each). The hypothesis that the distributions are similar is not rejected ( $P = 0.4$ , Wilcoxon test). B. SFS for the same conditions (1000 simulations each).

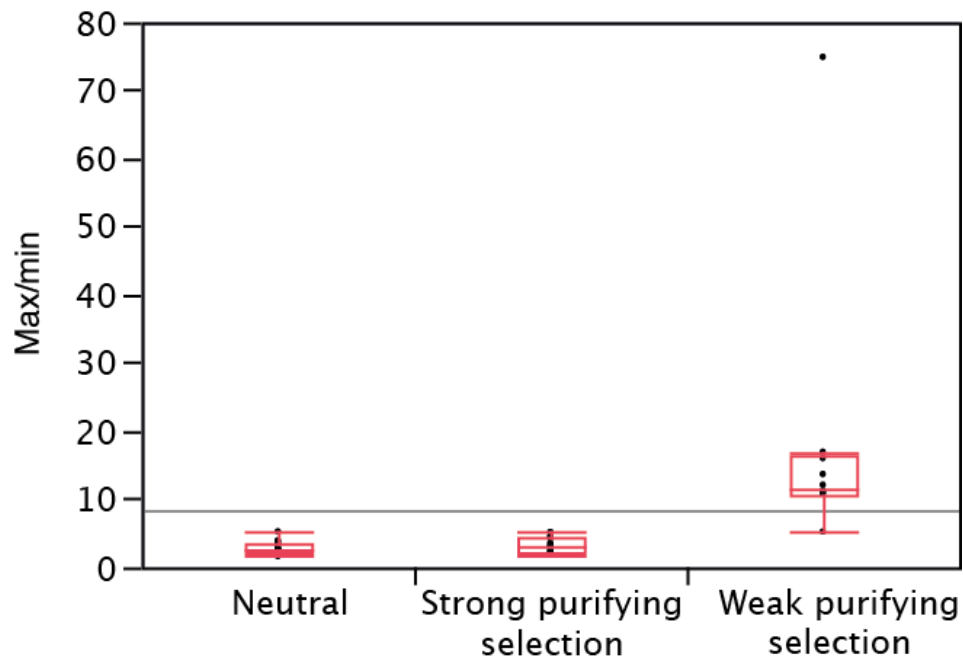


FIGURE S4 – Boxplots of the ratios between the maximal and minimal  $N_e u$  values for skyline plots, across the different types of simulations using the parameters for no selection (neutral), strong purifying selection, and weak purifying selection. The hypothesis that the distributions are similar is rejected ( $P < 0.001$ , Wilcoxon test).

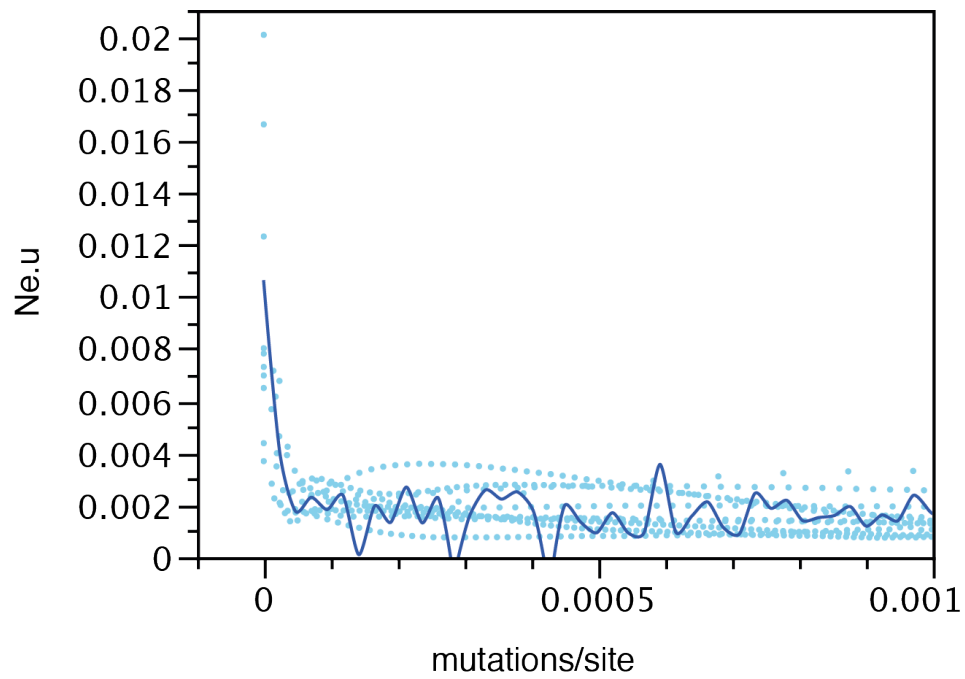


FIGURE S5 – Zoom of Figure 5 of the main text for the data using a clustered sampling bias. The figure in the main text lacks the values closer to zero (because they are log transformed). In this figure it can be seen that for values close to zero there is a systematic increase of  $N_e u$ .

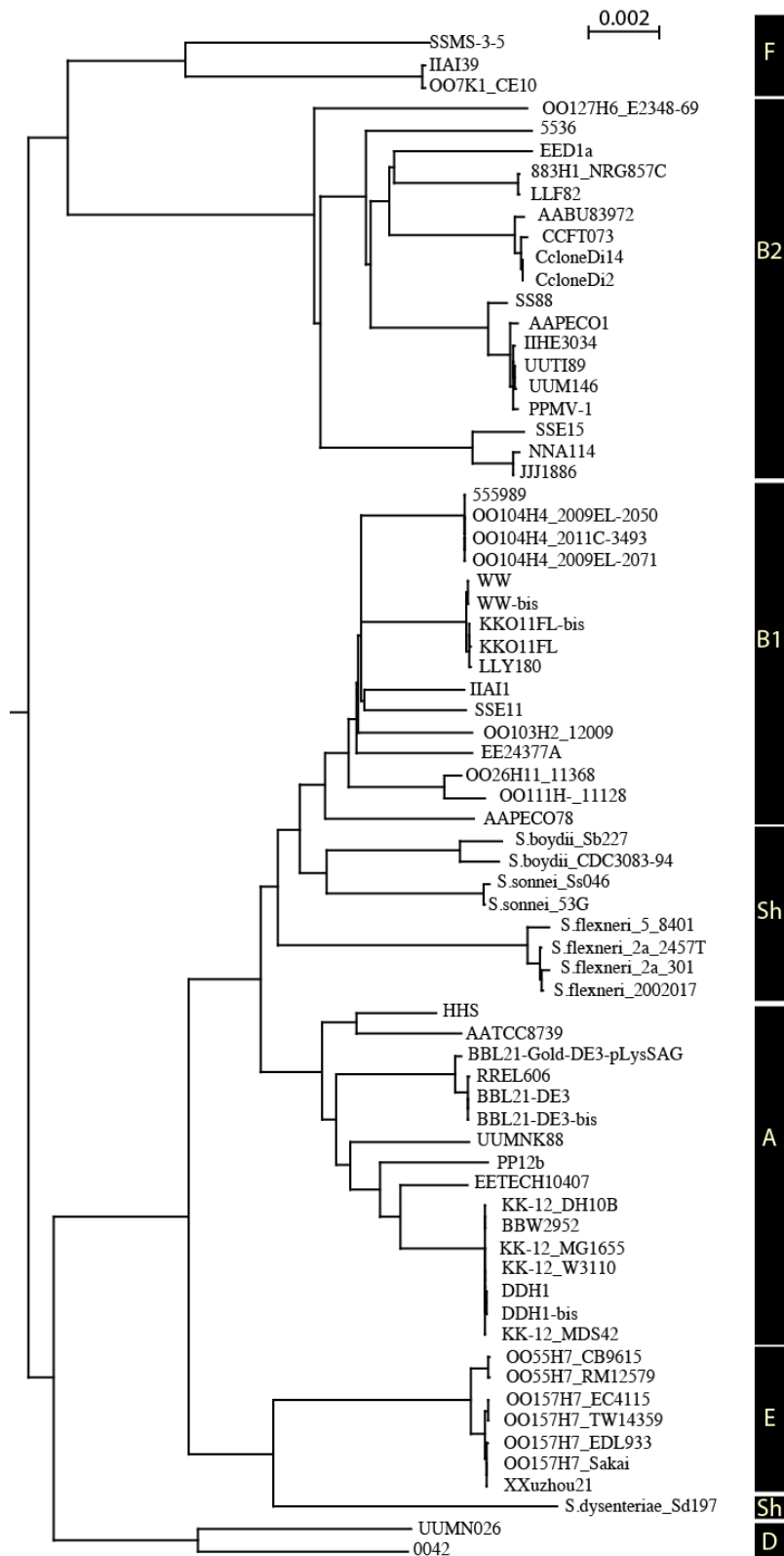


FIGURE S6 – Neighbor-joining phylogenetic tree of *E. coli* with the indication of the major sub-clades on the right and the name of the strains used in the study.



# Chapitre 3

## Exactitude des inférences démographiques basées sur le spectre de fréquence allélique : l'exemple de la population Yoruba

### 3.1 Résumé de l'article

#### QUESTIONS

Dans cette article nous inférons l'histoire démographique d'une population africaine, les Yoruba (données de The 1000 Genomes Project Consortium 2015), à partir de données de fréquences alléliques. Le spectre de fréquence est utilisé dans de nombreuses méthodes d'inférence démographique, qui peuvent être contraintes, quand elles testent un modèle donné (Nielsen, 2000; Gutenkunst et al., 2009; Coventry et al., 2010; Lukić et al., 2011; Nelson et al., 2012; Excoffier et al., 2013) ou flexibles, quand elles infèrent une démographie constante (ou exponentielle) par morceaux à partir du spectre, sans information préalable sur la démographie (Bhaskar et al., 2015; Liu and Fu, 2015). Cependant, bien que répandue, l'utilisation du spectre de fréquence pour l'inférence démographique est parfois remise en cause. Plusieurs études théoriques ont questionné l'identifiabilité des histoires démographiques à partir du spectre de fréquence (Myers et al., 2008; Bhaskar and Song, 2014; Kim et al., 2015; Terhorst and Song, 2015). Le but de l'étude est d'aborder le problème de l'identifiabilité des modèles démographiques en partant de données réelles. Des modèles simples, décrits par un seul paramètre, peuvent-ils être distingués à partir du spectre de fréquence ? Les méthodes d'inférence contraintes et flexibles donnent-elles

des résultats similaires pour un même jeu de données ?

## MÉTHODES

En raison de ce qui est connu de l’histoire démographique des populations humaines africaines, nous avons choisi d’essayer d’expliquer la démographie des Yoruba avec des modèles de croissance simples, décrits par un unique paramètre (voir Figure 1). Afin de comparer des modèles de référence différents, quatre modèles sont basés sur le modèle standard de Wright-Fisher, et un modèle est basé sur un processus naissance-mort critique. Nous optimisons ces modèles à un paramètre en minimisant la distance au carré entre le spectre prédit par le modèle et le spectre observé. À titre de comparaison, nous inférons également la démographie de cette population avec deux méthodes existantes, le stairway plot (Liu and Fu, 2015) et  $\partial a \partial i$  (Gutenkunst et al., 2009).  $\partial a \partial i$  est une méthode contrainte qui utilise d’autres outils que notre méthode pour simuler le spectre et pour optimiser les paramètres. Le stairway plot est une méthode flexible, qui infère une démographie constante par morceaux à partir du spectre observé, par maximum de vraisemblance composite. Les inférences de méthodes contraintes ou flexibles sont également comparées sur des données simulées.

## RÉSULTATS

Nous montrons que la méthode flexible testée, le stairway plot, propose une démographie très complexe pour la population Yoruba, avec plusieurs goulots d’étranglement dans les 160 000 dernières années, et que le spectre prédit sous cette démographie complexe n’ajuste pas bien le spectre observé des Yoruba. À l’inverse, les modèles contraints à un paramètre ajustent bien le spectre observé, ce qui permet de dire que cette population africaine est en croissance et que son  $T_{\text{MRCA}}$  est d’environ 1.7 million d’années. Cependant, les résultats ne permettent pas de choisir parmi les modèles de croissance testés : tous ont un bon ajustement au spectre observé. L’utilisation d’une autre méthode contrainte,  $\partial a \partial i$ , aboutit aux mêmes résultats. La comparaison des méthodes contraintes et flexibles sur des données simulées montre que le stairway plot est biaisé par le bruit présent dans le spectre de fréquence, dû au nombre fini de loci indépendants.

## CONCLUSIONS

Cette étude montre que même dans le cas d’une démographie simple, une méthode flexible, le stairway plot, peut inférer un scénario complexe peu réaliste et prédire un

spectre de fréquence qui n'ajuste pas bien les données. Cela pourrait s'expliquer par un sur-ajustement du bruit présent dans les données initiales, inévitable pour un nombre raisonnable de loci. À l'inverse, la démographie de la population Yoruba est compatible avec des scénarios de croissance simples, décrits par un unique paramètre. Nos résultats illustrent le problème d'identifiabilité des histoires démographiques à partir du spectre de fréquence, puisque tous les modèles de croissance testés, bien que sensiblement différents et basés sur plusieurs modèles de référence, ajustent aussi bien les données. Ils illustrent également l'importance de la complexité des modèles, en comparant des méthodes flexibles qui peuvent ajuster un grand nombre de paramètres à des méthodes contraintes pour lesquelles on peut réduire ce nombre de paramètres au minimum.

## ÉTAT DE PUBLICATION

Cet article a été publié dans *Genetics* le 5 mai 2017, après révisions.

## 3.2 Article

L'article est présenté dans les pages suivantes. Il est suivi d'annexes qui regroupent les informations supplémentaires publiées en complément de l'article (méthodes et figures), ainsi que des analyses complémentaires réalisées pendant cette étude mais non incluses dans l'article.

# Accuracy of Demographic Inferences from the Site Frequency Spectrum: The Case of the Yoruba Population

Marguerite Lapierre,<sup>\*,†,‡</sup> Amaury Lambert,<sup>†,‡</sup> and Guillaume Achaz<sup>\*,†</sup>

<sup>\*</sup>Atelier de Bioinformatique, UMR 7205 ISyEB, MNHN-UPMC-CNRS-EPHE, Muséum National d'Histoire Naturelle, 75005 Paris, France, <sup>†</sup>SMILE (Stochastic Models for the Inference of Life Evolution), UMR 7241 CIRB, Collège de France, CNRS, INSERM, PSL Research University, 75005 Paris, France, and <sup>‡</sup>Laboratoire de Probabilités et Modèles Aléatoires (LPMA), UMR 7599, UPMC-CNRS, 75005 Paris, France

ORCID IDs: 0000-0002-3115-5940 (M.L.); 0000-0002-7248-9955 (A.L.)

**ABSTRACT** Some methods for demographic inference based on the observed genetic diversity of current populations rely on the use of summary statistics such as the Site Frequency Spectrum (SFS). Demographic models can be either model-constrained with numerous parameters, such as growth rates, timing of demographic events, and migration rates, or model-flexible, with an unbounded collection of piecewise constant sizes. It is still debated whether demographic histories can be accurately inferred based on the SFS. Here, we illustrate this theoretical issue on an example of demographic inference for an African population. The SFS of the Yoruba population (data from the 1000 Genomes Project) is fit to a simple model of population growth described with a single parameter (e.g., founding time). We infer a time to the most recent common ancestor of 1.7 million years (MY) for this population. However, we show that the Yoruba SFS is not informative enough to discriminate between several different models of growth. We also show that for such simple demographies, the fit of one-parameter models outperforms the stairway plot, a recently developed model-flexible method. The use of this method on simulated data suggests that it is biased by the noise intrinsically present in the data.

**KEYWORDS** human demography; model identifiability; coalescent theory; site frequency spectrum

**I**NFERENCE of human population history based on demographic models for genomic data can complement archaeological knowledge, owing to the large amount of polymorphism data now available in human populations. Polymorphism data can be viewed as an imprint left by past demographic events on the current genetic diversity of a population [see, e.g., review by Pool *et al.* (2010)].

There are several means of analyzing this observed genetic diversity for demographic inference. The polymorphism data can be used to reconstruct a coalescence tree of the sampled individuals. The demography of the sampled population can be inferred by comparing this reconstructed tree with

theoretical predictions under a constant size model (Pybus *et al.* 2000). For example, in an expanding population, the reconstructed coalescent tree will have relatively longer terminal branches than the reference coalescent tree in a population of constant size. However, methods based on a single reconstructed tree are flawed because of recombination (Lapierre *et al.* 2016), since the genealogy of a recombining genome is described by as many trees as there are recombining loci.

The genome-wide distribution of allele frequencies is a function of the average genealogies, and can thus be used as a summary statistic for demographic inference. This distribution, called the Site Frequency Spectrum (SFS), reports the number of mutated sites at any given frequency. The demographic history of a population affects the shape of its SFS (Adams and Hudson 2004; Marth *et al.* 2004). For example, an expanding population carries an excess of low-frequency variants, compared with the expectation under a constant size model. The shape of the SFS is also altered by selection, which results in an excess of low- and high-frequency variants (Fay and Wu 2000). However, selection acts mainly on

Copyright © 2017 by the Genetics Society of America

doi: <https://doi.org/10.1534/genetics.116.192708>

Manuscript received September 30, 2016; accepted for publication March 23, 2017; published Early Online March 24, 2017.

Supplemental material is available online at [www.genetics.org/lookup/suppl/doi:10.1534/genetics.116.192708/-/DC1](http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.116.192708/-/DC1).

<sup>†</sup>Corresponding author: Atelier de Bioinformatique, Muséum National d'Histoire Naturelle, Boîte Courrier 50, Bâtiment 139, 45 rue Buffon, 75005 Paris, France. E-mail: [marguerite.lapierre@mnhn.fr](mailto:marguerite.lapierre@mnhn.fr)

the coding parts of the genome and the noncoding segments linked to them, while demography impacts the whole genome. Furthermore, unlike reconstructed trees, the SFS is not biased by recombination (Wall 1999). Quite on the contrary, by averaging the SFS over many correlated marginal genealogies, recombination lowers the variance of the SFS while its expectation remains unchanged. Therefore, the SFS of a sample is a summary of the genetic diversity, averaged over all the genome due to recombination, that can be analyzed in terms of demography.

Several types of methods exist to infer the demography of a population based on its SFS. A specific demographic model can be tested by computing a pseudolikelihood function for this model, based on the comparison of the observed SFS and the SFS estimated by Monte Carlo coalescent tree simulations (Nielsen 2000; Coventry *et al.* 2010; Nelson *et al.* 2012). This method can be extended to infer demographic scenarios of several populations, using their joint SFS (Excoffier *et al.* 2013). Methods based on Monte Carlo tree simulations are typically very costly in computation time. Other approaches rely on diffusion processes: they use the solution to the partial differential equation of the density of segregating sites as a function of time (Gutenkunst *et al.* 2009; Lukić *et al.* 2011).

Whereas all these methods are model-constrained, *i.e.*, they use the SFS to test the likelihood of a given demographic model, more flexible methods have been developed. Recently, Bhaskar *et al.* (2015) derived exact expressions of the expected SFS for piecewise-constant and piecewise-exponential demographic models. Liu and Fu (2015) developed a model-flexible method based on the SFS: the stairway plot. This method infers the piecewise-constant demography which maximizes the composite likelihood of the SFS, without any previous knowledge on the demography. This optimization is based on the estimation of a time-dependent population mutation rate,  $\theta$ . Although they show that their method infers efficiently some theoretical demographies, they do not test the goodness-of-fit of the expected SFS, reconstructed under the demography they infer, with the input SFS on which they apply their method.

All these methods are widely used for the inference of demography in humans and other species, but doubts remain on the identifiability of a population demography based on its SFS. It has been shown theoretically that certain population size functions are unidentifiable from the population SFS (Myers *et al.* 2008; Terhorst and Song 2015). Myers *et al.* (2008) showed that, for any given population size function  $N(t)$ , there exists an infinite number of smooth functions  $F(t)$ , such that  $\xi^N = \xi^{N+F}$ , where  $\xi^N$  is the SFS of a population of size function  $N(t)$ . However, other theoretical works have recently shown that for many types of population size functions commonly used in demography studies, such as piecewise constant or piecewise exponential functions, demography can be inferred based on the SFS, provided the sample is large enough (Bhaskar and Song 2014). These studies argued that the unidentifiability proven by Myers *et al.* (2008) relied on biologically unrealistic population size functions involving high frequency oscillations near the present. Recently, two studies (Kim *et al.* 2015;

Terhorst and Song 2015) have provided bounds on the amount of demographic information contained in the SFS or in coalescent times.

In this study, we use the SFS of an African population (the Yoruba population, data from The 1000 Genomes Project Consortium 2015) as an example of a somewhat simple demography, to illustrate the risks of overconfidence in demographic scenarios inferred. Namely, we highlight two issues potentially arising even in the case of simple demographies: unidentifiability of models and poor goodness-of-fit of inferences. We first infer the Yoruba demography with a model-constrained method, using diverse one-parameter models of growth, and then with a model-flexible method: the stairway plot (Liu and Fu 2015). For the model-constrained method, we test four different growth models derived from the standard neutral framework used in the vast majority of population genetics studies, also compared with a more uncommon type of model based on a branching process. Individual-based models such as the branching process are widely used in population ecology (Lambert 2010): the population is modeled as individuals who die and give birth at given rates independently. These models are not commonly used in population genetics although they provide interesting features of fluctuating population sizes, for example, and benefit from a strong mathematical framework.

## Materials and Methods

### 1000 Genomes Project data

Variant calls from the 1000 Genomes Project phase 3 were downloaded from the project ftp site (The 1000 Genomes Project Consortium 2015). The sample size for the Yoruba population is  $n = 108$  individuals (polymorphism data available for both genome copies of each individual, *i.e.*,  $2n = 216$  sequences). We kept all single nucleotide biallelic variants to plot the sample SFS. The number of biallelic sites is  $S = 20,417,698$ . The average distance between two sites is 136 bp (median 81 bp). The number of sites for which the ancestral allele is known is  $S' = 19,441,528$ . To avoid possible bias due to sequencing errors, we ignored singletons (mutations appearing in only one chromosome of one individual in the sample) for the rest of the study. The number of sites without singletons is  $S_{2+} = 15,915,401$ , including  $S'_{2+} = 15,216,929$  sites for which the ancestral allele is known. The implications of ignoring singletons are examined in the *Discussion*.

### SFS definition and graphical representation

The SFS of a sample of  $n$  diploid individuals is described as the vector  $\xi = (\xi_1, \xi_2, \dots, \xi_{2n-1})$ , where, for  $i \in [1, 2n - 1]$ ,  $\xi_i$  is the number of dimorphic (*i.e.*, with exactly two alleles) sites with derived form at frequency  $i/2n$ . To avoid potential orientation errors, we assumed that the ancestral form is unknown for all sites: we worked with a folded spectrum, where we consider the frequency of the less frequent (or minor) allele. In this case, the folded SFS is described as the vector

$\eta = (\eta_1, \eta_2, \dots, \eta_n)$ , where  $\eta_i = \xi_i + \xi_{2n-i}$  for  $i \in [1, n-1]$  and  $\eta_n = \xi_n$ . The folded SFS of the Yoruba sample is plotted in Supplemental Material, Figure S1. For a better graphical representation, all SFS were transformed as follows: we plot  $\phi_i$  normalized by its sum, where

for unfolded SFS,  $\phi_i = i\xi_i$  for  $i \in [1, 2n-1]$

for folded SFS,  $\phi_i = \eta_i \frac{i(2n-i)}{2n}$  for  $i \in [1, n-1]$  and  $\phi_n = n\eta_n$

The transformed SFS has a flat expectation (i.e., constant over all values of  $i$ ) under the standard neutral model (Nawa and Tajima 2008; Achaz 2009).

### Demographic models used for the model-constrained methods

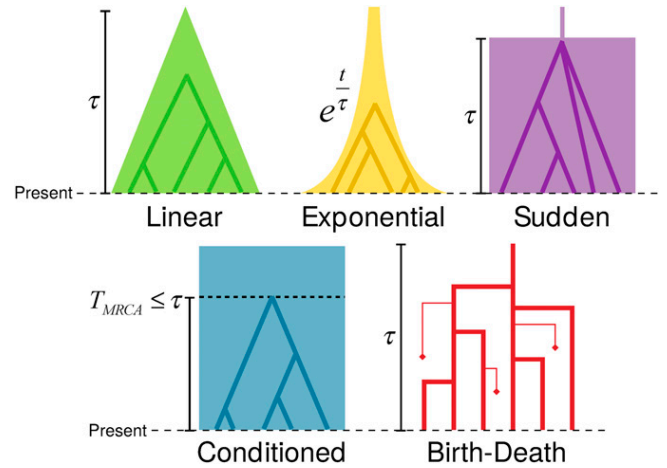
We inferred the demography of the Yoruba population using five growth models (Figure 1), compared with the predictions of the standard model with constant population size. Time is measured in coalescent units of  $2N$  generations, where the scaling parameter  $N$  has the same dimension as the current population size, which we will not estimate. Time starts at 0 (present time), and increases backward in time. Four models are based on the standard Kingman coalescent (Kingman 1982), amended with demography. Three of them are described with an explicit demography: either *Linear* growth since time  $\tau$ , *Exponential* growth at rate  $1/\tau$ , or *Sudden* growth from a single ancestor to the entire population at time  $\tau$ . We also use another model based on the Kingman coalescent, with an implicit demography: the *Conditioned* model. This model is based on a standard constant size model, but the Time to the Most Recent Common Ancestor ( $T_{MRC}$ ) is conditioned on being reached before time  $\tau$ . The fifth model, *Birth-Death*, is not based on the standard Kingman coalescent, but on a critical branching process measured in units of  $2N$  generations. Forward in time, the process starts with a founding event of one individual. Individuals give birth and die at equal rate 1. The process is conditioned on not becoming extinct before a period of time  $\tau$ , and on reaching, on average,  $2N$  individuals.

### Stairway plot inference on the Yoruba SFS

We applied the model-flexible stairway plot method developed by Liu and Fu (2015) to the unfolded Yoruba SFS. The unfolded SFS is constructed with the subset of sites  $S'$  for which the ancestral base is available. Once folded, this SFS is highly similar to the SFS constructed with the full set of sites  $S$  (their square distance is  $d^2 = 3.9 \times 10^{-5}$ , see below in the methods for the computation of  $d^2$ ). Inferences are made on 200 SFS as suggested by their method. We use the script they provide to create 199 bootstrap samples of the Yoruba SFS. We also ignore the singletons for this method, and use the default parameter values suggested in their paper for the optimization.

### SFS simulation with demography

We used two different methods to simulate SFS under the four demographic models derived from the Kingman coalescent



**Figure 1** The five demographic models. Each model has one single time parameter  $\tau$ .

(*Linear*, *Exponential*, *Sudden*, and *Conditioned*) or under a piecewise-constant demography reconstructed by the stairway plot method.

**Method 1:** Simulate  $l$  independent topologies under the Kingman coalescent on which mutations are placed at rate  $\theta$  (population mutation rate) (Hudson *et al.* 1990). This allows us to simulate the SFS of  $l$  independent loci.

**Method 2:** Another way to simulate SFS is using the following formula:

$$\mathbb{E}[\xi_i] = \frac{\theta}{2} \sum_{k=2}^{2n-i+1} k \mathbb{E}[t_k] \mathbb{P}(k, i) \quad (1)$$

where  $\theta$  is the population mutation rate,  $t_k$  is the time during which there are  $k$  lines in the tree (hereafter named state  $k$ ), and  $\mathbb{P}(k, i)$  is the probability that a randomly chosen line at state  $k$  gives  $i$  descendants in the sample of size  $2n$  (i.e., at state  $2n$ ) (Fu 1995). For all models, the neutrality assumption ensures that

$$\mathbb{P}(k, i) = \frac{\binom{2n-i-1}{k-2}}{\binom{2n-1}{k-1}}$$

for  $i \in [1, 2n-1]$  and  $k \in [2, 2n-i+1]$ . Using this probability allows us to average over the space of topologies. This reduces computation time considerably since the space of topologies is very large, and produces smooth SFS for which only the  $t_k$  need to be simulated to obtain the expectations  $\mathbb{E}[t_k]$ .

The expectations  $\mathbb{E}[t_k]$  are obtained as follows: for  $k \in [2, 2n]$ , times in the standard coalescent,  $t_k^*$ , are drawn from an exponential distribution of parameter  $\binom{k}{2}$ . For the *Linear* and *Exponential* models, and for the piecewise-constant demographies reconstructed by the stairway plot

method, these times are then rescaled to take into account the given explicit demography (see, e.g., Hein *et al.* 2004, Chap. 4). For the *Sudden* model, we assume the coalescence of all lineages at time  $\tau$  if the common ancestor has not yet been reached. For the *Conditioned* model, we keep only simulations for which  $\sum_{k=2}^{2n} t_k^* \leq \tau$ , where  $\tau$  is the model parameter. The expectations  $\mathbb{E}[t_k]$  are obtained by averaging over  $10^7$  simulations. Alternatively, the expectations  $\mathbb{E}[t_k]$  could also be obtained with analytic formulae provided by Polanski and Kimmel (2003).

For the *Birth-Death* model, we use the explicit formula for the SFS given in Delaporte *et al.* (2016).

We normalize the SFS computed under all these models so that their sum equals 1. This normalization removes the dependence on the mutation rate parameter  $\theta$ . Consequently, the standard model has no parameters while all others have exactly one ( $\tau$ ).

### Optimization of the parameter $\tau$

For each demographic model, we optimize the parameter  $\tau$  by minimizing the weighted square distance  $d^2$  between the observed SFS of the Yoruba population and the predicted SFS under the model (simulated with *Method 2*). Both SFS are normalized for comparison. The distance is computed for all  $\tau$  values in the interval  $[0.8, 3.0]$ , with a step of 0.01 (no specific optimization method was used to find the minimum). With  $\tilde{\eta}^{\text{model}}$  and  $\tilde{\eta}^{\text{obs}}$  the folded and normalized SFS in the tested model and in the data respectively

$$d^2(\tilde{\eta}^{\text{model}}, \tilde{\eta}^{\text{obs}}) = \sum_{i=2}^n \frac{(\tilde{\eta}_i^{\text{model}} - \tilde{\eta}_i^{\text{obs}})^2}{\tilde{\eta}_i^{\text{model}}}$$

The sum starts at  $i = 2$  because we ignore  $\tilde{\eta}_1^{\text{obs}}$  corresponding to singletons. To calculate the distance  $d^{2'}$  between the SFS predicted by two models, A and B, we weight the terms by the mean of the two models:

$$d^{2'}(\tilde{\eta}^A, \tilde{\eta}^B) = \sum_{i=2}^n \frac{(\tilde{\eta}_i^A - \tilde{\eta}_i^B)^2}{(\tilde{\eta}_i^A + \tilde{\eta}_i^B)/2}$$

### Inference of the Yoruba demography with $\partial a \partial i$

We inferred the demography of the Yoruba population with the software  $\partial a \partial i$  v1.7 (Gutenkunst *et al.* 2009), testing the three models of explicit demography (*Linear*, *Exponential*, and *Sudden*). The demographic models were specified so that the only parameter to optimize is  $\tau$  like for the distance-based inference method. Singletons were masked and the method was applied on the folded Yoruba SFS. Details on the demographic functions and parameter values used for the optimization in  $\partial a \partial i$  are provided in [File S1](#). We ran the method 100 times for each model and kept the parameter value with the best maximum log composite likelihood over the 100 runs. In [Figure S4](#), we plot the best log composite likelihood of the 100 runs.

### Scaling of the coalescent time

Optimized values of the parameter  $\hat{\tau}$  for each model are expressed in coalescent time units, i.e., scaled in  $2N_e(0)$  generations (please note that in a model population where all individuals reproduce,  $N_e = N$ ). As the model population size at time zero,  $2N_e(0)$ , is unknown, to scale these coalescent time units in numbers of generations and consequently in years, we used the expected number of mutations per site  $M$ . From the dataset, we have  $M^{\text{obs}} = S/L$ , where  $S$  is the number of single nucleotide mutations (a  $k$ -allelic SNP accounts for  $k - 1$  mutations), and  $L$  is the length of the accessible sequenced genome in the 1000 genomes project (90% of the total genome length, The 1000 Genomes Project Consortium 2015). For the theoretical value, we get that  $M^{\text{theo}} = \mu \hat{T}_{\text{tot}} C$ , where we know the mutation rate  $\mu$  from the literature, and the total tree length expressed in coalescent time units  $\hat{T}_{\text{tot}}$  from the SFS simulations. Here,  $C$  is the coalescent factor, that is the number of generations per coalescent time unit, also corresponding to  $2N_e(0)$ , where  $N_e(0)$  is the effective population size of a real population at present time. The total number of generations in the tree is  $\hat{T}_{\text{tot}} C$ , from which we derive the total number of mutations per site  $M^{\text{theo}}$ . Thus, using the observed value  $M^{\text{obs}}$ , we can estimate  $C$  by  $S/(\mu L \hat{T}_{\text{tot}})$ . We assumed a mutation rate of  $1.2 \times 10^{-8}$  per base pair per generation (Conrad *et al.* 2011; Campbell *et al.* 2012; Kong *et al.* 2012). With the coalescent factor  $C$ , we can then convert a coalescent time unit into a number of generations, or into a number of years, assuming 24 years as the generation time (Scally and Durbin 2012).

### Graphical representation of the inferred demographies

To represent the inferred explicit demographies (models *Linear*, *Exponential*, and *Sudden*), we plot the shape of the demography with the optimized value  $\hat{\tau}$  for each model. For the implicit demographies (models *Conditioned* and *Birth-Death*), as there is no explicit demographic shape, we plot the mean trajectory of fixation of a new allele in the population: forward in time, these fixation trajectories illustrate the expansion of the descendance of the sample's ancestor in the population (see [File S1](#) for details).

### Comparing the model-constrained and model-flexible methods to infer Linear growth

We applied both methods (the one-parameter inference method and the stairway plot method) on SFS simulated under *Linear* growth. To test the stairway plot method on a *Linear* growth demography, we simulate 200 independent SFS using *Method 1*, with sample size  $2n = 216$ ,  $\theta = 100$  (arbitrary value removed by normalization) and a founding time  $\tau = 2.48$  (estimated for the Yoruba population, see [Results](#)). The SFS are simulated with either  $10^3$ ,  $10^4$ , or  $10^5$  independent loci. We scaled the simulated SFS to obtain a total number of  $S = 20,417,698$  variants, so that the total number of variants in the simulated SFS is the same as in the Yoruba SFS. We ran the stairway plot method on these



200 independent SFS with the default parameter values suggested in the method, and with the same mutation rate ( $1.2 \times 10^{-8}$  per base pair per generation) and generation time (24 years) as in our study. Here, the singletons are taken into account, because inferences are made on simulated data. We report the median demography of these 200 independent inferences.

To test the one-parameter inference method on these SFS simulated under the *Linear* model, we run the parameter optimization on a SFS simulated with either  $10^3$ ,  $10^4$ ,  $10^5$ , or  $10^6$  loci. The search of the parameter value that minimizes the distance  $d^2$  was optimized with a Newton-Raphson algorithm. Derivatives were calculated at  $\tau \pm 0.05$ , where  $\tau$  is the parameter value being optimized. The optimization stopped when the optimization step of the parameter value was  $< 10^{-3}$ .

### Data availability

The 1000 Genomes Project data used in this study is publicly available at <ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/>. The code in Python and C written for the study is available at [https://github.com/lapierreM/Yoruba\\_demography](https://github.com/lapierreM/Yoruba_demography). The code in C used for the *Method 1* of SFS simulation is available upon request to G.A.

## Results

We inferred the demography of the Yoruba population (Africa), from the whole-genome polymorphism data of 108 individuals (data from the 1000 Genomes Project, The 1000 Genomes Project Consortium 2015), with SFS-based methods, either model-constrained or model-flexible.

It has been shown that human populations have been growing since their emergence in Africa, and that African populations were supposedly not affected by the Out-of-Africa bottleneck described for Eurasian populations (Marth *et al.* 2004; Atkinson *et al.* 2008; Gutenkunst *et al.* 2009; Gronau *et al.* 2011; Tennessen *et al.* 2012). Analyses using the PSMC method (Li and Durbin 2011) have shown a reduction in the African population size after the divergence with non-African populations. However, Mazet *et al.* (2016) have recently shown that these analyses could be biased by population structure. Based on this previous knowledge, for the model-constrained method, we chose to infer the Yoruba demography with simple models of growth, *i.e.*, with only one phase of growth characterized by a single parameter. These five models are: *Linear*, *Exponential*, or *Sudden* growth, a *Conditioned* model, where the  $T_{\text{MRCA}}$  is conditioned on being smaller than the given parameter, and a critical *Birth-Death* model based on a branching process (Figure 1). To infer the Yoruba demography, we fit the SFS predicted under each model with the observed Yoruba SFS (all SFS are folded). The SFS were normalized to remove the population mutation rate parameter  $\theta$ , so that each model is characterized by one single parameter  $\tau$ , which has the dimension of a time duration. We fit this parameter by least-square distance between the observed

SFS and the predicted SFS, and by maximum likelihood using the  $\partial a \partial i$  software (Gutenkunst *et al.* 2009). For the model-flexible inference, we used the stairway plot method developed recently by Liu and Fu (2015), which infers a piecewise-constant demography based on the SFS. For this method, the number of parameters to be estimated is determined by a likelihood-ratio test. It can range from 1 to  $2n - 1$ , where  $2n$  is the number of sequences in the sample.

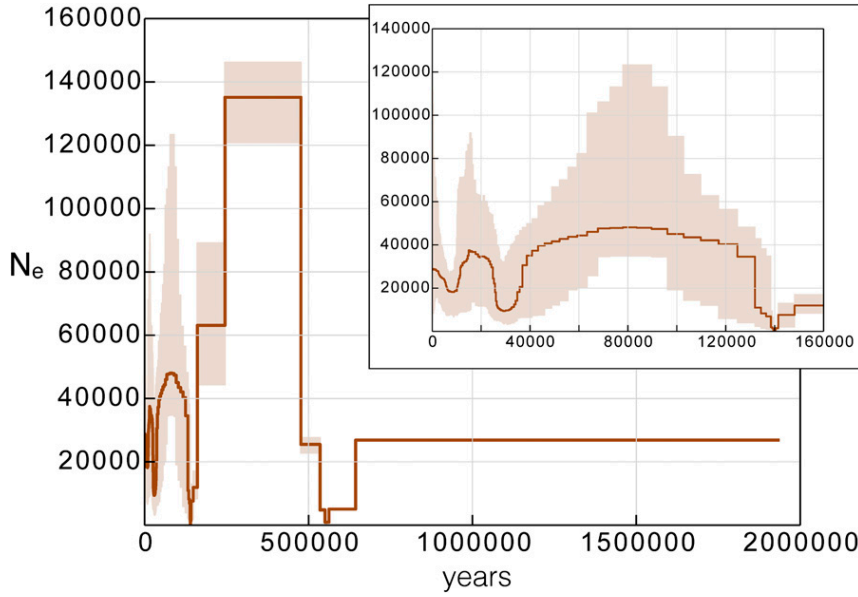
The Yoruba SFS was constructed by taking into account the entire genome. Removing the coding parts of the genome to avoid potential bias due to selection does not affect the shape of the SFS substantially (Figure S2), since the coding parts represent a very small fraction of the human genome. The first bin of the observed SFS, accounting for mutations found in one chromosome of one individual in the sample (black dot in the observed SFS in Figure 3B), seemed to lie outside the rest of the distribution. This could be due to sequencing errors being considered as singletons (Achaz 2008), and thus we chose to ignore this value for the model optimization. We have also made sure that the SFS shape was not affected greatly by the sample size. We compared the SFS of a subsample of half the Yoruba individuals ( $2n = 108$ ) with the full sample SFS ( $2n = 216$ ) (Figure S3). This shows that the only bin of the SFS that is affected by this subsampling is the first one, containing the singletons. As we ignore singletons in our study, the sample size should not influence our results.

The analysis of the Yoruba SFS with the stairway plot method results in a complex demography with several bottlenecks in the last 160,000 years (Figure 2). The effective population size at time 0,  $N_e(0)$ , is 28,500 (as we ignore singletons, time 0 does not correspond to present time, see *Discussion*). The demographic history earlier than 160,000 years ago shows spurious patterns that should not be interpreted, according to Liu and Fu (2015).

The inference of the Yoruba demography with one-parameter models was done by minimizing the distance between observed and predicted SFS. This gave an optimized value  $\hat{\tau}$  of the parameter  $\tau$  (Figure 3A and Table 1) (with  $\hat{\tau}$  in coalescent units, *Linear*:  $\hat{\tau} = 2.48$ , *Exponential*:  $\hat{\tau} = 1.82$ , *Sudden*:  $\hat{\tau} = 1.36$ , *Conditioned*:  $\hat{\tau} = 1.89$ , and *Birth-Death*:  $\hat{\tau} = 2.28$ ). Plotting the predicted SFS with the optimized parameter value  $\hat{\tau}$  confirmed their goodness-of-fit with the observed Yoruba SFS (Figure 3B). Compared to the standard model without demography, the addition of just one parameter allows for a surprisingly good fit of the observed Yoruba SFS. The Yoruba demography thus seems to be compatible with a simple scenario of growth. On the other hand, the demography inferred by the stairway plot predicts a SFS that does not fit well the observed Yoruba SFS: the distance between the observed Yoruba SFS and the expected SFS under the stairway plot demography is 10 times the distance between any of the one-parameter model SFS and the data (Figure 3B and Table 1).

The best fitting SFS under each of the five demographic models all have a square distance  $d^2$  of the order of  $10^{-4}$  with the observed Yoruba SFS (Figure 3A and Table 1), and have





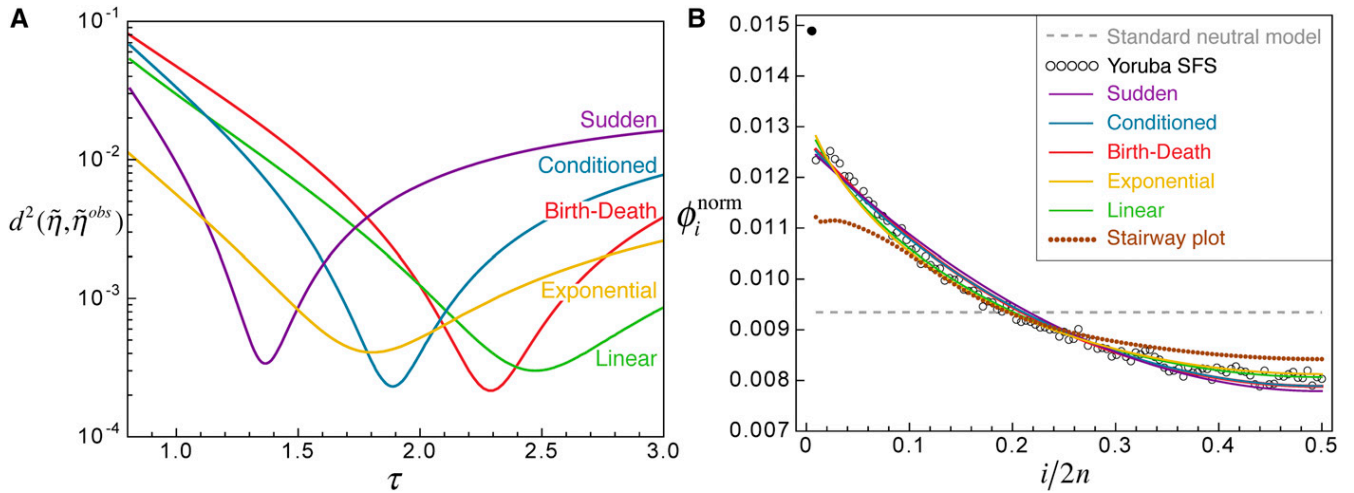
**Figure 2** Stairway plot inference of the Yoruba demography. The inferred effective size  $N_e$  of the Yoruba population is plotted from present time (0) to the past. The inset is a zoom between 0 and 160,000 years. The thick brown line is the median  $N_e$ , the light brown area is the [2.5, 97.5] percentiles interval. The inference is based on 200 bootstrap samples of the unfolded Yoruba SFS. The singletons are not taken into account for the optimization of the stairway plot.

highly similar shapes (Figure 3B). This suggests that the five demographic models used to infer the demography of the Yoruba are hard to distinguish based only on the observed SFS. To validate the use of a least square distance to find the best fitting SFS, we also inferred the Yoruba demography using the  $\partial a \partial i$  software. This model-constrained method based on the SFS uses a diffusion approximation to simulate SFS and a likelihood framework for the parameter optimization. We tested the three models of explicit demography (*Linear*, *Exponential*, and *Sudden* growth) parametrized in the same way as in our method. The best parameter values found by  $\partial a \partial i$  by maximum log composite likelihood are numerically indistinguishable from those found by our method (with  $\hat{\tau}$  in coalescent units, *Linear*:  $\hat{\tau} = 2.48$ , *Exponential*:  $\hat{\tau} = 1.82$ , and *Sudden*:  $\hat{\tau} = 1.36$ ). Moreover, the log composite likelihoods of the best fitting SFS under each model are on the same scale (the likelihoods are directly comparable because the number of parameters is the same for each model): *Linear*:  $\ln(L) = -3107$ , *Exponential*:  $\ln(L) = -3953$ , and *Sudden*:  $\ln(L) = -3393$  (Figure S4). They rank the explicit demography models in the same order as the least square distance  $d^2$  would rank them: the best model is *Linear* growth, then *Sudden*, and finally *Exponential* growth.

We computed the expected  $T_{MRCA}$  based on the predicted SFS using (1): as the SFS predicted under each model are very similar, it means that they have roughly the same estimated time durations  $t_k$  while there are  $k$  branches in the coalescent tree of the Yoruba sample. From these expected  $t_k$ , we can compute  $T_{MRCA} = \sum_{i=2}^{2n} t_k$ . This is the  $T_{MRCA}$  of the sample, but we can assume that it is the same as the  $T_{MRCA}$  of the population, because, with such a large sample size, the probability that the  $T_{MRCA}$  of the population is different from the  $T_{MRCA}$  of the sample becomes very small. Under each of four models (excluding the *Birth-Death* model for which

there is no obvious common time scaling), the expected  $T_{MRCA}$  for the Yoruba population is 1.3 in coalescent units. By using the number of mutations per site in the data, and the total tree length inferred from the simulations, we scaled back this  $T_{MRCA}$  in number of generations and in years, assuming a mutation rate of  $1.2 \times 10^{-8}$  per base pair per generation (Conrad *et al.* 2011; Campbell *et al.* 2012; Kong *et al.* 2012), and a generation time of 24 years (Scally and Durbin 2012) (see *Materials and Methods*). The  $T_{MRCA}$  of the Yoruba population inferred under the four demographic models is of 87,100 generations, corresponding to 1.7 million years (MY). The inferred demographic models, with scaling in coalescent units, number of generations and number of years, are shown in Figure 4. The coalescent unit of 67,000 estimated to scale the inferred coalescent times in number of years corresponds to a present effective population size  $N_e(0)$  of 33,500.

The demography inferred by the stairway plot method for the Yoruba population is a piecewise-constant demography showing much more complex patterns of growth and bottlenecks than the one-parameter models (Figure 2). Moreover, the expected SFS under this inferred demography does not fit well to the observed Yoruba SFS (Figure 3B). To understand what could produce such a complex demography, we simulated SFS under a *Linear* growth with the founding time  $\hat{\tau} = 2.48$  inferred for the Yoruba population. We simulated three sets of 200 SFS, with respectively  $10^3$ ,  $10^4$ , and  $10^5$  loci, to obtain SFS with more or less noise (solid lines on Figure 5A). We applied the two inference methods on these SFS. The median demographies inferred by the stairway plot method are strongly affected by the noise of the SFS, as shown on Figure 5B. When the number of simulated loci is very large (median of 200 independent demographies inferred with  $10^6$  loci), the stairway plot gives a good



**Figure 3** Inference of the Yoruba demography with one-parameter models. (A) Weighted square distance  $d^2(\tilde{\eta}, \tilde{\eta}^{obs})$  between the normalized Yoruba SFS  $\tilde{\eta}^{obs}$  and the normalized predicted SFS  $\tilde{\eta}$  under each of the five models, depending on the value of the parameter  $\tau$  (Purple: *Sudden*, Blue: *Conditioned*, Red: *Birth-Death*, Yellow: *Exponential*, and Green: *Linear*). (B) Predicted SFS under each of the five models, with the optimized value  $\hat{\tau}$  of the parameter, and under the demography inferred by the stairway plot (brown dotted line). The Yoruba SFS is shown as empty circles. The first dot, colored in black, accounting for the singletons, was not taken into account for the optimization of  $\tau$  to avoid potential bias due to sequencing errors. The gray dashed line is the expected SFS under the standard neutral model without demography. Colors match the plot beside (the predicted SFS under the models *Birth-Death* and *Conditioned* are indistinguishable). The SFS are folded, transformed, and normalized (see *Materials and Methods*).

approximation of the true demography, and the expected SFS under the inferred demography fits the input SFS. However, for smaller numbers of loci (median of 200 independent demographies inferred with  $10^5$  loci or less), the stairway plot shows complex patterns of growth and bottlenecks incompatible with the true demography, and the expected SFS under the inferred demographies do not fit the input SFS. On the contrary, the one-parameter method infers a *Linear* demography with a founding time close to the true value for SFS simulated with  $10^4$  loci or more (Table 2).

## Discussion

In this study, we fit the SFS of the Yoruba population with five simple demographic models of growth described by one parameter. Surprisingly, even though these five models are quite distinct in the way they model population growth, fitting them on the Yoruba data results in strongly similar SFS, which all show an excellent goodness-of-fit with the observed Yoruba SFS. Fitting the same SFS with the stairway plot method (Liu and Fu 2015), a model-flexible method which infers a piecewise-constant demography, resulted in a complex demography with several bottlenecks in the last 160,000 years. The poor goodness-of-fit of the expected SFS under this inferred demography with the Yoruba SFS indicates that this complex demography is not to be trusted, and suggests that the way the method estimates the number of change points is too flexible.

The results obtained by the model-constrained and model-flexible methods showed some similarities: the current population size  $N_e(0)$  of  $\sim 30,000$  inferred with the stairway plot

corresponds roughly to the coalescent unit of 67,000 generations (equivalent to  $2N_e(0)$  in the coalescent theory) found with the one-parameter models. Similarly, the  $T_{MRCA}$  of  $\sim 1.7$  MY inferred with the one-parameter models seems to match with the last time point of the stairway plot, at  $\sim 1.9$  MY.

We hypothesize that the complexity of the demography inferred by the stairway plot method is caused by the irregularities of the observed Yoruba SFS. Two concurrent nonexclusive explanations can be put forward for these irregularities. First, they can be due to the sampling, and thus be considered as noise that should not be interpreted as evidence for demography. Second, these irregularities could be biologically relevant, and result from a very complex demographic history. To assess the impact of noise on the stairway plot method, we tested it on simulated SFS under a *Linear* growth. These SFS were simulated with different numbers of independent loci: the more loci, the less noise in the simulated SFS. The stairway plot inference on these SFS shows that the method is strongly affected by the noise in the SFS simulated data: whereas the demography inferred for a smooth SFS (corresponding to a high number of independent loci) corresponds to the true demography approximated as piecewise constant, the demographies inferred for smaller numbers of loci show complex patterns of bottlenecks and deviate strongly from the true demography. In the original paper presenting the stairway plot (Liu and Fu 2015), the method was tested on simulations resulting in unrealistically smooth SFS, which is why it efficiently inferred the tested demographies. It could be that this method captures the signal contained in these irregularities and infers a demography

**Table 1** Least-square distance  $d^2$  between pairs of observed Yoruba SFS and optimized SFS under the five demographic models or the stairway plot method

	Data	Linear	Exponential	Sudden	Conditioned	Birth-Death
Linear	$3.0 \times 10^{-4}$	0				
Exponential	$4.1 \times 10^{-4}$	$2.2 \times 10^{-5}$	0			
Sudden	$3.4 \times 10^{-4}$	$3.5 \times 10^{-4}$	$5.5 \times 10^{-4}$	0		
Conditioned	$2.3 \times 10^{-4}$	$1.6 \times 10^{-4}$	$5.5 \times 10^{-4}$	$3.7 \times 10^{-5}$	0	
Birth-Death	$2.2 \times 10^{-4}$	$1.7 \times 10^{-4}$	$3.1 \times 10^{-4}$	$4.1 \times 10^{-5}$	$3.5 \times 10^{-6}$	0
Stairway plot	$2.9 \times 10^{-3}$	$3.1 \times 10^{-3}$	$3.3 \times 10^{-3}$	$2.8 \times 10^{-3}$	$2.8 \times 10^{-3}$	$2.9 \times 10^{-3}$

taking them into account, whereas the one-parameter models fit the global trend of the SFS shape and can thus infer the true demography for much smaller numbers of loci. One solution could be to constrain the number of parameters allowed for model-flexible methods: it seems that determining it by likelihood-ratio test, as it is done in the stairway plot method, is not conservative enough, as it does not prevent overfitting of the noise. If the number of parameters was forced to be small, the method might capture the global trend of the demography and avoid this issue. The SFS reconstructed under the demographies inferred by the stairway plot, however, differ strongly from the input SFS. If the issue was the overfitting of noise, we would expect the reconstructed SFS to fit the data more closely. The method is clearly biased by noise on the SFS but it remains unclear why. It would require further investigation to analyze how the different characteristics of this particular method, such as the parametrization of population size history, respond to noise, and what is responsible for this bias.

The five one-parameter demographic models all predict virtually the same SFS for the Yoruba population. Therefore, they also predict the same  $T_{MRCA}$  for the Yoruba population. This  $T_{MRCA}$  of  $\sim 1.3$  in coalescent units corresponds, with our scaling of coalescent time based on the number of mutations per site, to  $\sim 1.7$  MY. This estimation is similar to results concerning the whole human population, obtained by Blum and Jakobsson (2011) and reviewed in Garrigan and Hammer (2006). Although the commonly acknowledged date of emergence of the anatomically modern human is  $\sim 200,000$  years ago, Blum and Jakobsson (2011) showed that finding a much older  $T_{MRCA}$  was compatible with the single-origin hypothesis, assuming a certain ancestral effective population size. These ancient times to most recent common ancestor could also be explained by gene flow in a structured ancestral population (Garrigan and Hammer 2006).

Although all five models predict the same  $T_{MRCA}$ , the inferred demographies differ substantially between the models (Figure 3A). In the time range further beyond the  $T_{MRCA}$ , no information is carried by the sample. Thus, the inferred demographies differ in this time range (Figure 4), making the inferred founding time of the Yoruba population unreliable.

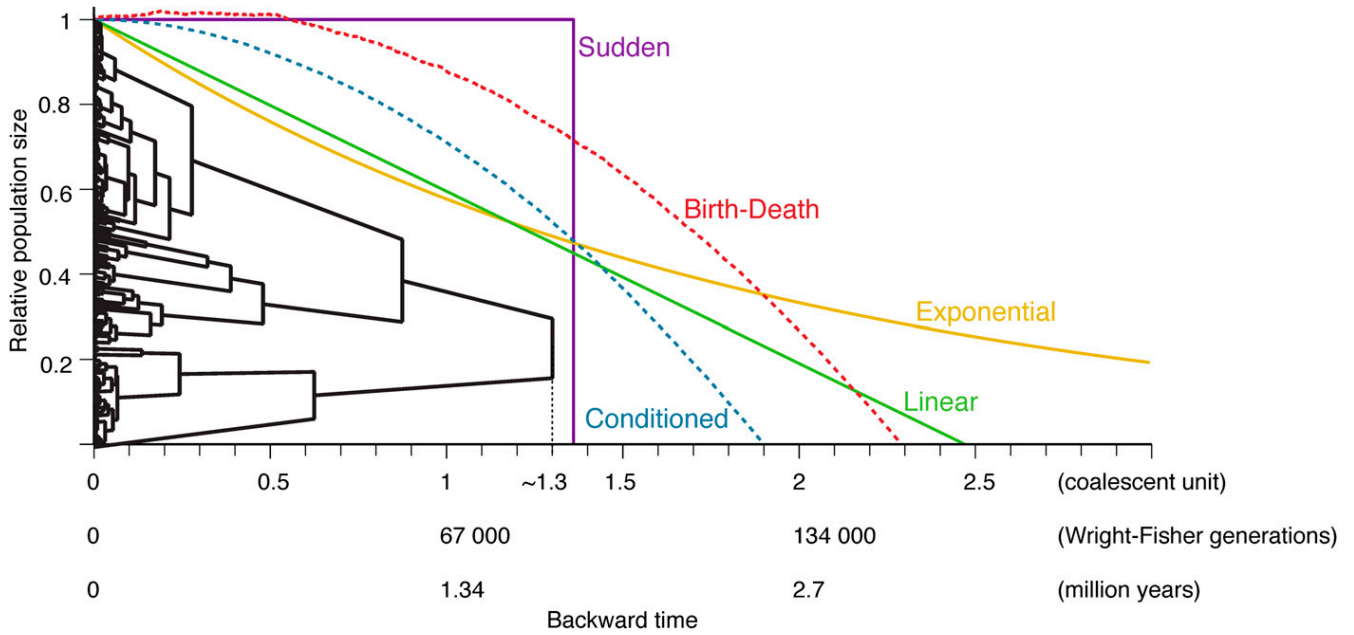
Our results with one-parameter models are reproducible with another model-constrained method,  $\partial a \partial i$ , which uses different approaches both for the theoretical SFS simulations

(diffusion approximation) and the parameter optimization (composite likelihood). This shows that, for models having the same number of parameters, a distance-based approach finds the same ranking of models as a likelihood framework, while being computationally less intensive. Furthermore, the distance-based approach allows for intuitive evidence on the fact that these different models actually all perform very well to fit the Yoruba SFS: the small differences of distance between the best SFS predicted by each model and the observed SFS could be due to the noise in the observed SFS, and thus do not mean that one model is better than another. Our results raise potential interest in an in-depth comparative study of likelihood-based methods, such as  $\partial a \partial i$  and the stairway plot, and methods based on least square distance.

Among the five tested demographic models, two pairs of models seem to predict particularly similar SFS (pairs of models with the two smallest values of  $d^2$  in Table 1). First, the *Linear* (L) and *Exponential* (E) growth models predict almost identical SFS for the Yoruba population ( $d^2(\tilde{\eta}^L, \tilde{\eta}^E) = 2.2 \times 10^{-5}$ ). Figure 4 shows that, in the time range where information is conveyed by the mean coalescent tree of the population, *i.e.*, between present time and the  $T_{MRCA}$ , these two demographies are very similar. This explains why their SFS are almost indistinguishable, and shows that, in this parameter range, it is impossible to distinguish linear from exponential growth. Second, the SFS predicted under the two models with implicit demography, *Conditioned* (C) and *Birth-Death* (BD), are so similar that they are undistinguishable in Figure 3B ( $d^2(\tilde{\eta}^C, \tilde{\eta}^{BD}) = 3.5 \times 10^{-6}$ ). This raises the question of how these two models, based on different processes — a Wright-Fisher model or a branching process — compare and, in particular, why their SFS are so similar.

As we compute the distance statistic to optimize the models on normalized SFS, the information of the magnitude of the SFS (often referred to as  $\theta$ , the population mutation rate) is lost. However, as the inferred SFS under the five demographic models all have the same shape, the constant  $\theta$  by which they should be multiplied to fit the real, not normalized, Yoruba SFS would be the same for all five models. Thus, this information would not allow us to choose which model infers the most realistic value of  $\theta$ .

The outlying first bin of the Yoruba SFS, corresponding to singletons, was removed from our inference because it can be



**Figure 4** Demographic histories and reconstructed tree estimated from the Yoruba SFS. The tree shown has internode durations  $t_k$ , during which there are  $k$  lineages consistent with the SFS (the topology was chosen uniformly among ranked binary trees with  $2n$  tips). Time is given in coalescent units, and scaled in number of generations and in millions of years. The demographic histories (solid lines: explicit models, dashed lines: implicit models) are plotted with their optimized  $\hat{\tau}$  values. See File S1 for details on the demographic histories plotted for the models with implicit demographies (*Birth-Death* and *Conditioned*).

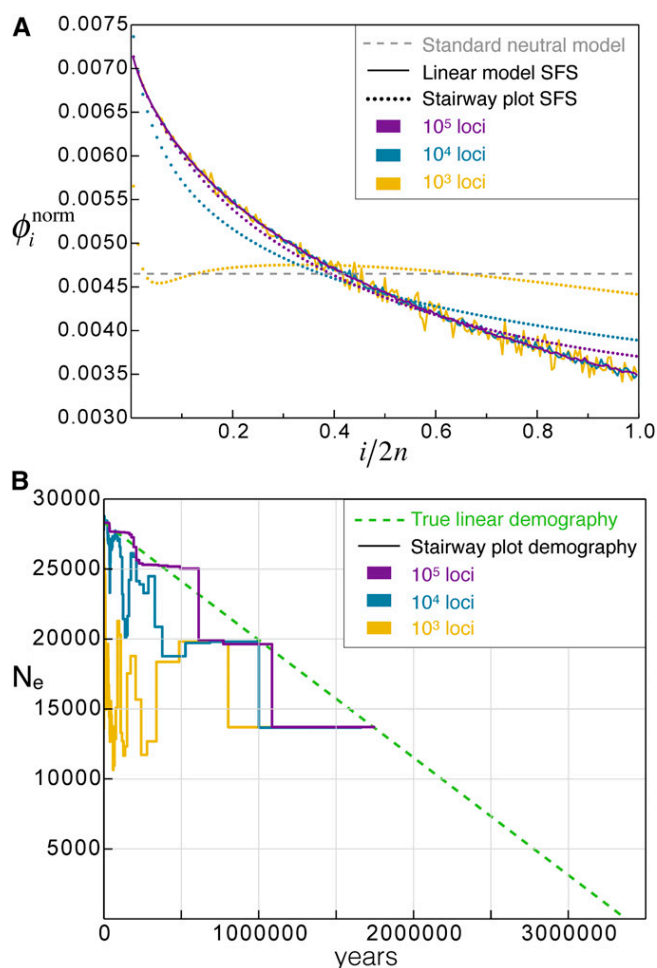
affected by sequencing errors. As the relatively low to moderate coverage of the 1000 Genomes Project could also result in an underestimation of doubletons and tripletons, we optimized  $\tau$  masking also these values. It did not change the estimation of  $\hat{\tau}$  and thus had no effect on the inferred demographies. As the first bin of the SFS accounts for the mutations that occur in the terminal branches of the coalescent tree, a large part of the excess of singletons can be due to very recent and massive growth. Recent studies with deep sequencing coverage have shown that there is a large abundance of rare variants in human populations (Coventry *et al.* 2010; Nelson *et al.* 2012; Gazave *et al.* 2014). As the dataset we used for this study had a limited sample size and low-coverage, we focused on the inference of demography in the more distant past. Thus, because of both sequencing errors and incompatibility with our one-parameter models, singletons were not taken into account. Our inferences concern the population before this recent and massive growth. It should also be noted that Liu and Fu (2015) emphasize that the strength of their method is in capturing recent demographic history. Thus, ignoring singletons, although it is an existing feature of their software, might not be the most appropriate use of the stairway plot.

For non-African human populations, the SFS based on the 1000 Genomes Project data are not monotonous: their shape is more complex than the SFS of the Yoruba population. Thus, one-parameter models cannot capture the complexity of the demographic histories underlying these types of observed SFS. Even for the Yoruba population, capturing the recent growth event, by taking into account the singletons, would

have required adding another parameter. The stairway plot method shows more flexibility, and could capture the signal for more complex demographic histories, provided that the number of independent loci is very large so that there is no bias due to noise.

Overall, this study shows that, even in the case of a simple demography, the scenario inferred by the stairway plot, a model-flexible method, can show spuriously complex patterns of growth and decline, and can predict SFS poorly fitting the initial SFS data. This might be explained by overfitting of the method to the noise present in the observed SFS, which can be expected for a reasonable number of loci. We also show that simple models described by one parameter can have an excellent goodness-of-fit to the data, and avoid the issue of noise overfitting. The results indicate that the demography of the Yoruba population is compatible with simple one-parameter models of growth, and that the expected  $T_{\text{MRCA}}$  of this population can be estimated at  $\sim 1.7$  MY. However, the SFS is not sufficient to determine which model better characterizes the Yoruba demographic growth, and estimations of the founding time of the population, that depend on the chosen model, are thus unreliable. More generally, this study illustrates the issue of nonidentifiability of demographies based on the SFS of a finite sample.

Our comparison of a model-constrained method using one parameter models with a model-flexible method using a potentially large number of parameters highlights the importance of the model complexity. How many parameters should we use to “properly” characterize a demography? We argue



**Figure 5** Stairway plot inference of a linear demography SFS with noise. (A) Solid lines: mean of 200 SFS simulated independently under the *Linear* growth model, with either  $10^5$  loci (purple),  $10^4$  loci (blue), or  $10^3$  loci (yellow). Dotted lines: expected SFS under the demography reconstructed by the stairway plot method for different number of loci (same colors than solid lines). The gray dashed line is the expected SFS under the standard neutral model without demography. The SFS are transformed and normalized (see *Materials and Methods*). (B) Stairway plot demographic inference: median of 200 independent demographies inferred with 200 independently simulated SFS for each number of loci (colors match the plot above). The true demography is the green dashed line. The inferred effective size  $N_e$  is plotted from present time (0) to the past.

that low complexity models should be tested first. For model-flexible methods, the number of parameters is usually unbounded, and determined by successive likelihood ratio tests. This statistical framework implies that a certain risk is taken at each successive step, and that with the repetition of steps, errors can potentially be made. For example, these errors can lead to spurious inferences in noisy data (*i.e.*, any real data). We recommend (visually) monitoring the improvement in goodness-of-fit when adding new parameters on statistical grounds. Examination of the intermediate steps of fitting would likely prevent an unnecessary increase in the model complexity.

**Table 2** Inference of the founding time  $\hat{\tau}$  under the *Linear* model on SFS with noise

Number of Loci	5% Percentile	Mean $\hat{\tau}$	95% Percentile
$10^3$	2.569	2.713	2.893
$10^4$	2.463	2.503	2.540
$10^5$	2.473	2.485	2.498
$10^6$	2.478	2.483	2.487

Mean, 5 and 95% percentile of the founding time inferred with a *Linear* model. The SFS on which the inference is made are simulated with a founding time  $\tau$  of 2.48, with different number of loci, using the method with topology reconstruction.

## Acknowledgments

We thank Cécile Delaporte for preliminary work on this project, and Simon Boitard, Michael Blum, Konrad Lohse, and three anonymous reviewers for useful comments on the manuscript. G.A. and M.L. acknowledge support from the grant ANR-12-NSV7-0012 Demochips from the Agence Nationale de la Recherche (France). M.L. is funded by the Ph.D. program 'Interfaces pour le Vivant' of Pierre and Marie Curie University (UPMC) University Paris 06. G.A., A.L., and M.L. thank the Center for Interdisciplinary Research in Biology for funding.

## Literature Cited

- Achaz, G., 2008 Testing for neutrality in samples with sequencing errors. *Genetics* 179: 1409–1424.
- Achaz, G., 2009 Frequency spectrum neutrality tests: one for all and all for one. *Genetics* 183: 249–258.
- Adams, A. M., and R. R. Hudson, 2004 Maximum-likelihood estimation of demographic parameters using the frequency spectrum of unlinked single-nucleotide polymorphisms. *Genetics* 168: 1699–1712.
- Atkinson, Q. D., R. D. Gray, and A. J. Drummond, 2008 mtDNA variation predicts population size in humans and reveals a major Southern Asian chapter in human prehistory. *Mol. Biol. Evol.* 25 (2): 468–474.
- Bhaskar, A., and Y. S. Song, 2014 Descartes' rule of signs and the identifiability of population demographic models from genomic variation data. *Ann. Stat.* 42(6): 2469–2493.
- Bhaskar, A., Y. R. Wang, and Y. S. Song, 2015 Efficient inference of population size histories and locus-specific mutation rates from large-sample genomic variation data. *Genome Res.* 25(2): 268–279.
- Blum, M. G., and M. Jakobsson, 2011 Deep divergences of human gene trees and models of human origins. *Mol. Biol. Evol.* 28(2): 889–898.
- Campbell, C. D., J. X. Chong, M. Malig, A. Ko, B. L. Dumont *et al.*, 2012 Estimating the human mutation rate using autozygosity in a founder population. *Nat. Genet.* 44(11): 1277–1281.
- Conrad, D. F., J. E. Keebler, M. A. DePristo, S. J. Lindsay, Y. Zhang *et al.*, 2011 Variation in genome-wide mutation rates within and between human families. *Nat. Genet.* 43(7): 712–714.
- Coventry, A., L. M. Bull-Otterson, X. Liu, A. G. Clark, T. J. Maxwell *et al.*, 2010 Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nat. Commun.* 1: 131.
- Delaporte, C., G. Achaz, and A. Lambert, 2016 Mutational pattern of a sample from a critical branching population. *J. Math. Biol.* 73: 627–664.



- Excoffier, L., I. Dupanloup, and E. Huerta-SáV. C. nchez, Sousa, and M. Foll, 2013 Robust demographic inference from genomic and SNP data. *PLoS Genet.* 9(10): 1–17.
- Fay, J. C., and C.-I. Wu, 2000 Hitchhiking under positive Darwinian selection. *Genetics* 155: 1405–1413.
- Fu, Y.-X., 1995 Statistical properties of segregating sites. *Theor. Popul. Biol.* 48(2): 172–197.
- Garrigan, D., and M. F. Hammer, 2006 Reconstructing human origins in the genomic era. *Nat. Rev. Genet.* 7(9): 669–680.
- Gazave, E., L. Ma, D. Chang, A. Coventry, F. Gao *et al.*, 2014 Neutral genomic regions refine models of recent rapid human population growth. *Proc. Natl. Acad. Sci. USA* 111(2): 757–762.
- Gronau, I., M. J. Hubisz, B. Gulko, C. G. Danko, and A. Siepel, 2011 Bayesian inference of ancient human demography from individual genome sequences. *Nat. Genet.* 43(10): 1031–1034.
- Gutenkunst, R. N., R. D. Hernandez, S. H. Williamson, and C. D. Bustamante, 2009 Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* 5(10): 1–11.
- Hein, J., M. Schierup, and C. Wiuf, 2004 *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory*. Oxford University Press, New York.
- Hudson, R. R., 1990 Gene genealogies and the coalescent process, pp. 1–44 in *Oxford Surveys in Evolutionary Biology*, edited by D. Futuyma, and J. Antonovics. Oxford University Press, New York.
- Kim, J., E. Mossel, M. Z. Rácz, and N. Ross, 2015 Can one hear the shape of a population history? *Theor. Popul. Biol.* 100: 26–38.
- Kingman, J. F. C., 1982 The coalescent. *Stochastic Process. Appl.* 13(3): 235–248.
- Kong, A., M. L. Frigge, G. Masson, S. Besenbacher, P. Sulem *et al.*, 2012 Rate of de novo mutations and the importance of father's age to disease risk. *Nature* 488(7412): 471–475.
- Lambert, A., 2010 Population genetics, ecology and the size of populations. *J. Math. Biol.* 60(3): 469–472.
- Lapierre, M., C. Blin, A. Lambert, G. Achaz, and E. P. Rocha, 2016 The impact of selection, gene conversion, and biased sampling on the assessment of microbial demography. *Mol. Biol. Evol.* 33: 1711–1725.
- Li, H., and R. Durbin, 2011 Inference of human population history from individual whole-genome sequences. *Nature* 475(7357): 493–496.
- Liu, X., and Y.-X. Fu, 2015 Exploring population size changes using SNP frequency spectra. *Nat. Genet.* 47(5): 555–559.
- Lukić, S., J. Hey, and K. Chen, 2011 Non-equilibrium allele frequency spectra via spectral methods. *Theor. Popul. Biol.* 79(4): 203–219.
- Marth, G. T., E. Czabarka, J. Murvai, and S. T. Sherry, 2004 The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics* 166: 351–372.
- Mazet, O., W. Rodriguez, S. Grusea, S. Boitard, and L. Chikhi, 2016 On the importance of being structured: instantaneous coalescence rates and human evolution—lessons for ancestral population size inference? *Heredity* 116(4): 362–371.
- Myers, S., C. Fefferman, and N. Patterson, 2008 Can one learn history from the allelic spectrum? *Theor. Popul. Biol.* 73(3): 342–348.
- Nawa, N., and F. Tajima, 2008 Simple method for analyzing the pattern of DNA polymorphism and its application to SNP data of human. *Genes Genet. Syst.* 83(4): 353–360.
- Nelson, M. R., D. Wegmann, M. G. Ehm, D. Kessner, P. S. Jean *et al.*, 2012 An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* 337(6090): 100–104.
- Nielsen, R., 2000 Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics* 154: 931–942.
- Polanski, A., and M. Kimmel, 2003 New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. *Genetics* 165: 427–436.
- Pool, J. E., I. Hellmann, J. D. Jensen, and R. Nielsen, 2010 Population genetic inference from genomic sequence variation. *Genome Res.* 20(3): 291–300.
- Pybus, O. G., A. Rambaut, and P. H. Harvey, 2000 An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics* 155: 1429–1437.
- Scally, A., and R. Durbin, 2012 Revising the human mutation rate: implications for understanding human evolution. *Nat. Rev. Genet.* 13(10): 745–753.
- Tennessen, J. A., A. W. Bigham, T. D. O'Connor, W. Fu, E. E. Kenny *et al.*, 2012 Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* 337(6090): 64–69.
- Terhorst, J., and Y. S. Song, 2015 Fundamental limits on the accuracy of demographic inference based on the sample frequency spectrum. *Proc. Natl. Acad. Sci. USA* 112(25): 7677–7682.
- The 1000 Genomes Project ConsortiumAuton, A., L. D. Brooks, R. M. Durbin, E. P. Garrison, H. M. Kang *et al.*, 2015 A global reference for human genetic variation. *Nature* 526(7571): 68–74.
- Wall, J. D., 1999 Recombination and the power of statistical tests of neutrality. *Genet. Res.* 74: 65–79.

Communicating editor: N. A. Rosenberg

## 3.3 Annexes

### 3.3.1 Informations supplémentaires de l'article

#### MÉTHODES SUPPLÉMENTAIRES

**Inference of the Yoruba demography with  $\partial\text{a}\partial\text{i}$**  Demographic function used for the *Linear* model:

```
def linear_growth(params, n1, pts):
    T = params
    xx = dadi.Numerics.default_grid(pts)
    phi = dadi.PhiManip.phi_1D(xx)
    nu = 1e-9 #fixed initial population size
    nu_func = lambda t: nu + ( ( 1.0 - nu ) * t ) / T
    phi = dadi.Integration.one_pop(phi, xx, T, nu=nu_func)
    sfs = dadi.Spectrum.from_phi(phi, ns, (xx,))
    return sfs}
```

Demographic function used for the *Exponential* model:

```
def exponential_growth(params, n1, pts):
    T = params
    xx = dadi.Numerics.default_grid(pts)
    phi = dadi.PhiManip.phi_1D(xx)
    nu = 1e-1 #fixed initial population size
    nu_func = lambda t: nu * ( ( 1.0 / nu ) ** ( t / T ) )
    phi = dadi.Integration.one_pop(phi, xx, T, nu=nu_func)
    sfs = dadi.Spectrum.from_phi(phi, ns, (xx,))
    return sfs
```

Demographic function used for the *Sudden* model:

```
def sudden_growth(params, n1, pts):
    T = params
    xx = dadi.Numerics.default_grid(pts)
    phi = dadi.PhiManip.phi_1D(xx)
    Tb = 1.0 #fixed time before growth event
```

```

nu = 1e-2 #fixed population size before growth event
nuF = 1.0 #fixed population size after growth event
phi = dadi.Integration.one_pop(phi, xx, Tb, nu=nu)
phi = dadi.Integration.one_pop(phi, xx, T, nu=nuF)
sfs = dadi.Spectrum.from_phi(phi, ns, (xx,))
return sfs

```

Interval for the parameter to optimize and initial value for optimization:

- *Linear* :  $T \in [0, 10]$  and  $T_0 = 3$
- *Exponential* :  $T \in [0, 25]$  and  $T_0 = 5$
- *Sudden* :  $T \in [0, 10]$  and  $T_0 = 1$

For the three models, the grid point settings for the extrapolation is [300, 400, 500]. The function used for the optimization is `optimize_log` with `maxiter=3`. The script for each demographic model was run 100 times, and we kept the parameter value with the best maximum log composite likelihood. For the exponential, to retrieve the rate of the exponential growth as we had parametrized it in our model, we compute

$$\tau = -\frac{T_{opt}}{\ln(nu)}$$

where  $T_{opt}$  is the optimized parameter value and  $nu = 0.1$  (see demographic function above).

**Fixation trajectories for the models with implicit demography** For the *Conditioned* model, we use the Wright-Fisher diffusion conditioned upon fixation (Lambert, 2008) to simulate trajectories of fixation :

$$dX_t = (1 - X_t)dt + \sqrt{X_t(1 - X_t)} dB_t$$

where  $X_t$  is the random variable accounting for the frequency of the allele at time  $t$  and  $B_t$  is Brownian motion. We simulate the trajectories starting at  $X_0 = 0$  with  $dt = 0.0001$  and we stop the trajectories when  $X_t$  reaches 1. To account for the specificity of the *Conditioned* model, we keep only trajectories that reach fixation in a time smaller than the optimized parameter value  $\hat{\tau}$ . Similarly, for the Birth-Death model, we use the critical Feller diffusion (Lambert, 2008) :

$$dX_t = \sqrt{2X_t}dB_t$$

and we run trajectories until time reaches the optimized parameter value  $\hat{\tau}$ . We keep trajectories for which  $X_{\hat{\tau}} \in (U_n, U_{n+1})$ , where  $U_k = \sum_{i=1}^k V_i$  and the  $V_i$ 's are independent



exponential random variables with mean  $1/n$ . This procedure amounts to conditioning upon sampling  $n$  individuals at time  $\hat{\tau}$ . Indeed, for mathematical reasons, the standard way of sampling in a branching population is not to fix the sample size, but to sample each individual in the population independently with the same probability  $p$ . Assuming that individuals are linearly ordered, the number  $W$  of individuals between two consecutively sampled individuals then follows a geometric law of parameter  $p$ . In the model used in the paper and in Delaporte et al. (2016), we further condition on the sample size  $n$  with the relation  $p = n/N$ . So if we measure  $W$  in units of  $N$  individuals, we are left with  $V = \frac{p}{n}W$ . Now as  $p \rightarrow 0$  (sparse sampling),  $V$  converges to an exponential random variable of parameter  $n$ . Thus, the individuals sampled in the population are separated by exponential random variables of parameter  $n$ , and can thus be represented by the points  $(U_i)_{i \geq 1}$ . Therefore, sampling  $n$  individuals is equivalent to keeping trajectories for which  $X_{\hat{\tau}} \in (U_n, U_{n+1})$ .

For both models, we average over 5000 trajectories.



## FIGURES SUPPLÉMENTAIRES

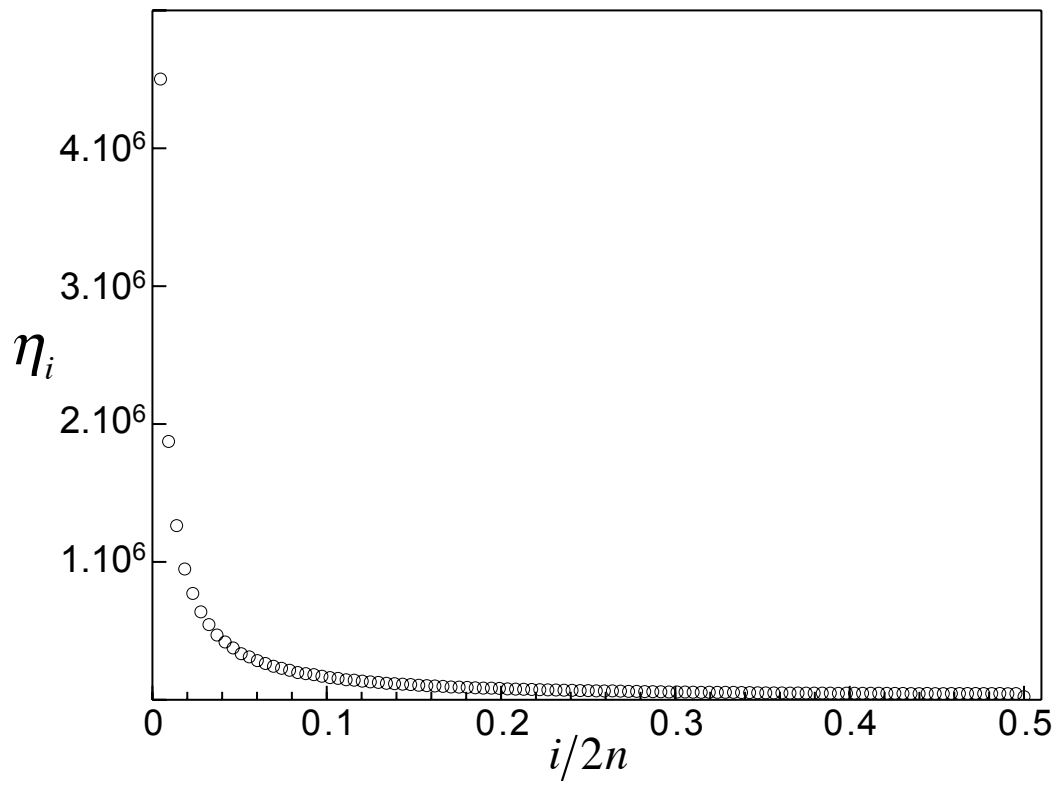


FIGURE S1 – Yoruba Site Frequency Spectrum. The SFS is folded. The total number of sites in the SFS is  $S = 20\,417\,698$ .

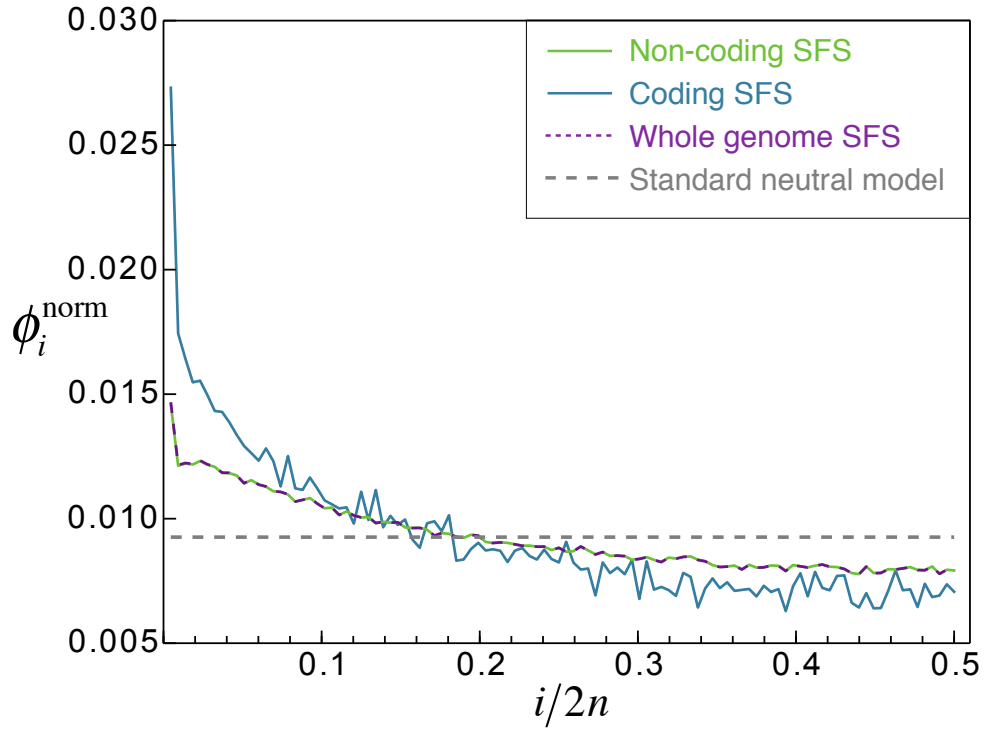


FIGURE S2 – Coding and non-coding Yoruba SFS. In blue, SFS for coding parts of the genome. In green, SFS for the non-coding parts of the genome. The dashed purple line is the whole-genome SFS. The grey dashed line is the expected SFS under the standard neutral model without demography. The SFS are folded, transformed and normalized (see Methods).

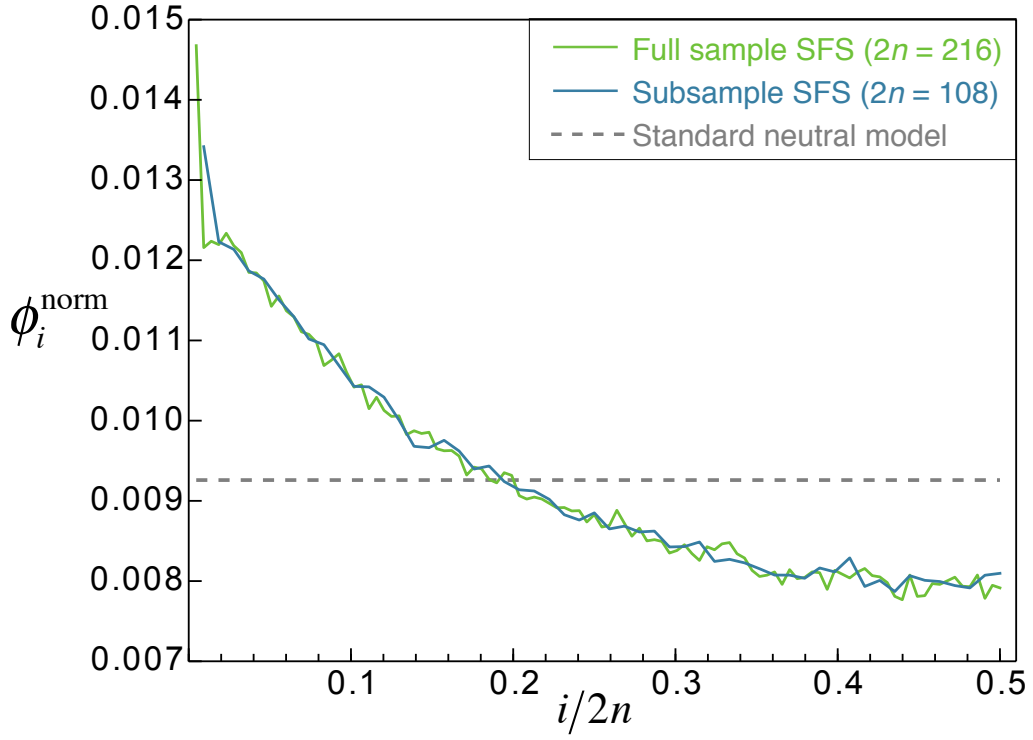


FIGURE S3 – Subsample SFS of the Yoruba population. The green line is the SFS of the whole sample ( $2n = 216$ ). The blue line is the SFS of a subsample containing half of the Yoruba individuals ( $2n = 108$ ). The grey dashed line is the expected SFS under the standard neutral model without demography (with  $2n = 216$ ). The SFS are folded, transformed and normalized (see Methods). For comparison, the subsample SFS was divided by 2 after normalization because it contains half as many values as the two other SFS.

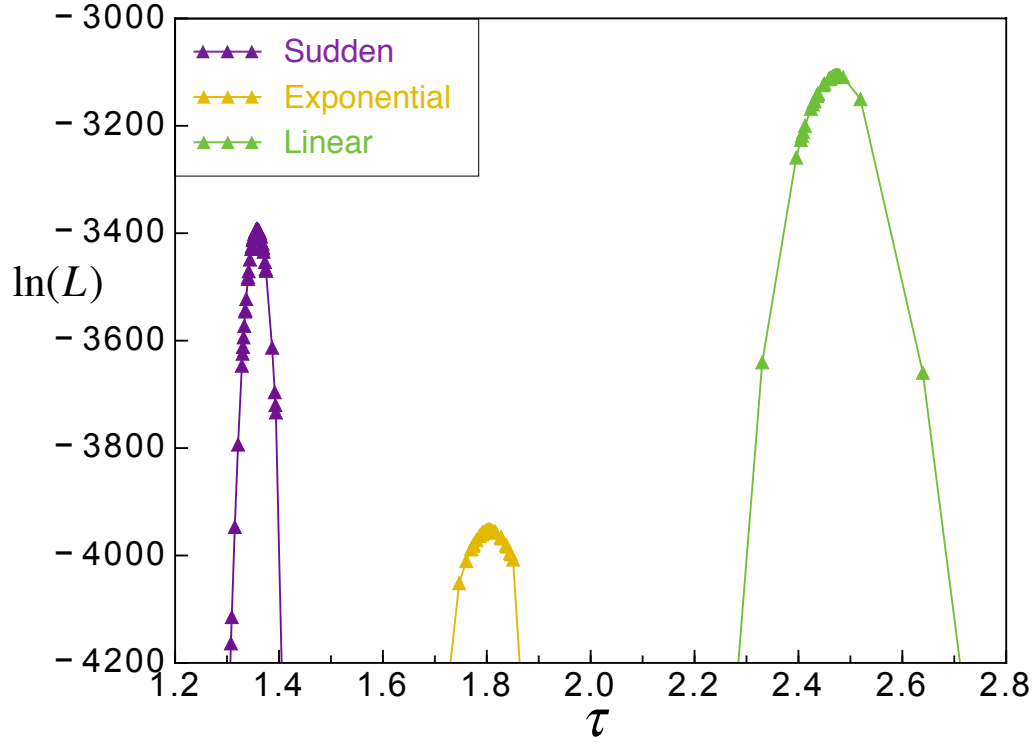


FIGURE S4 – Maximum log composite likelihood obtained by the  $\partial a \partial i$  method for the *Sudden*, *Exponential* and *Linear* models. We ran the method 100 times for each model. For each run, we report the maximum log composite likelihood with the corresponding  $\tau$  value. The figure is zoomed on the best likelihood values (higher than -4200). The number of points present in the plot (with log composite likelihood higher than -4200) is 75 for the *Sudden* model, 93 for the *Linear* model and 95 for the *Exponential* model.

### 3.3.2 Analyses complémentaires

Au cours du processus de révision de l'article, plusieurs analyses complémentaires ont été réalisées pour répondre aux remarques et questions des reviewers anonymes. Il me semblait intéressant de les mentionner ici même si elles n'ont pas trouvé leur place dans la version finale de l'article.

#### Spectres utilisés pour la validation du Stairway plot

Au sujet de la méthode Stairway plot qui semble être biaisée par le bruit, un des reviewers a fait remarquer que dans l'article original présentant la méthode (Liu and Fu, 2015), les auteurs avaient pris soin de tester le choix du nombre de paramètres sur des données simulées. Leurs résultats (Figure *Validation for parameter estimation* dans leurs informations supplémentaires) montraient que leur méthode choisissait un nombre de paramètres parcimonieux pour ajuster les données.

Pour visualiser le spectre de fréquence sur lequel ils effectuaient cette validation, nous avons simulé leur scénario démographique à deux époques avec le logiciel ms (Hudson, 2002) et représenté graphiquement le spectre de fréquence obtenu. Nous voulions en particulier voir si ce spectre était bruité ou non. Dans leur simulation (ligne de commande 5.2 dans leurs informations supplémentaires), la démographie est caractérisée par une taille de population actuelle de 25 636, une taille de population ancestrale de 7778 et un changement de taille instantané il y a 6809 générations. Le taux de mutation est de  $1.2 \times 10^{-8}$  par base par génération, et le taux de recombinaison est de  $0.8\theta$  par base par génération. Le spectre de fréquence est construit à partir d'un échantillon de 30 séquences d'une longueur de 10 millions de bases. 200 réplicats sont simulés. Le nombre moyen de sites polymorphes dans ces simulations est de  $S = 24\,158$ . Ces simulations donnent le spectre moyen présenté dans la Figure S5. On voit qu'avec les valeurs de paramètres utilisées dans cette simulation (représentatives de ce que les auteurs utilisent dans leur article), le spectre de fréquence simulé est lisse, ce qui explique que leur méthode soit efficace lorsqu'elle est appliquée à ce spectre. En effet, nous le montrons également dans la Figure 5B de l'article, pour l'inférence médiane faite sur 200 spectres simulés avec  $10^5$  loci chacun : le stairway plot ajuste bien la vraie démographie simulée, sans trouver l'optimum global mais en s'arrêtant à un optimum local satisfaisant, qui approche la croissance linéaire par une croissance constante par morceaux.

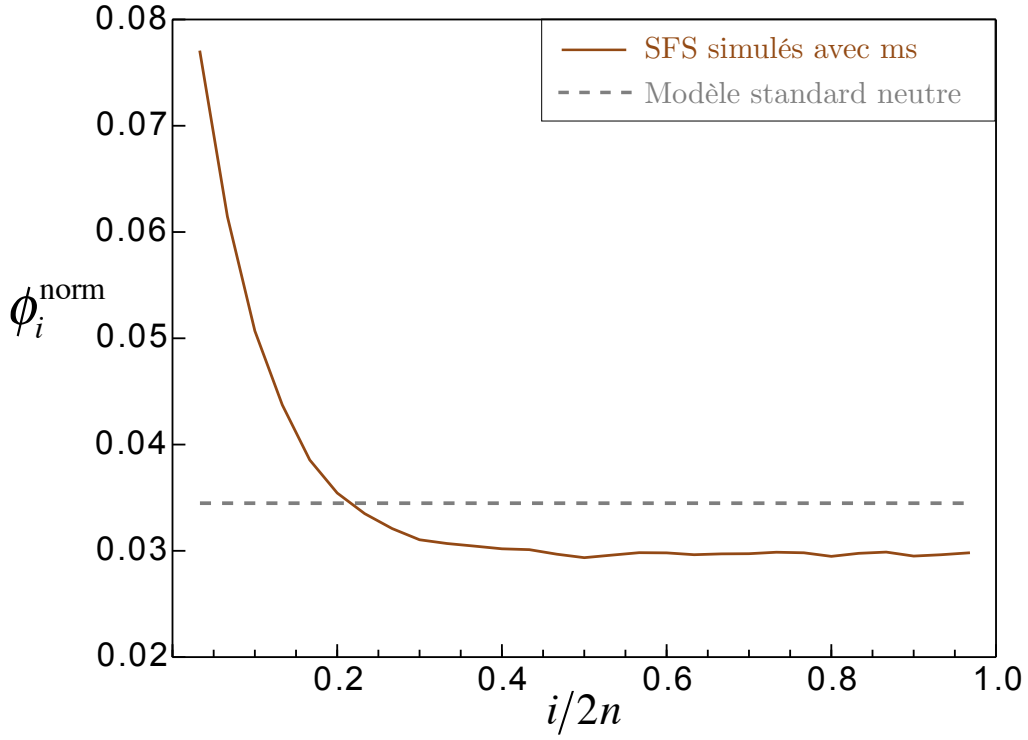


FIGURE S5 – Exemple de spectre de fréquence simulé avec ms, avec les valeurs de paramètres utilisées par Liu et Fu (2015) pour le modèle à deux époques. La courbe marron est le spectre moyen pour 200 simulations, pour un échantillon de 30 séquences de 10 millions de bases. La ligne grise pointillée est le spectre attendu sous le modèle standard neutre sans démographie. Les spectres sont transformés et normalisés (voir Méthodes de l'article).

### Spectre plié et déplié des données Yoruba

Lorsque nous avons analysé le spectre de fréquence des Yoruba avec la méthode Stairway plot, celle ci n'était applicable qu'à des spectres dépliés. Comme les mutations du jeu de données n'ont pas toutes un allèle ancestral inféré, le jeu de données utilisable pour le spectre déplié est légèrement réduit. Pour s'assurer que cela n'avait pas trop d'influence sur la forme du spectre, nous avons comparé le spectre plié construit avec tous les sites ( $S = 20\,417\,698$ ) avec le spectre plié construit uniquement avec les sites pour lesquels on connaît l'allèle ancestral ( $S' = 19\,441\,528$ ). Ces spectres, représentés sur la Figure S6, sont très proches, leur distance au carré est de  $d^2 = 3.9 \times 10^{-5}$ .

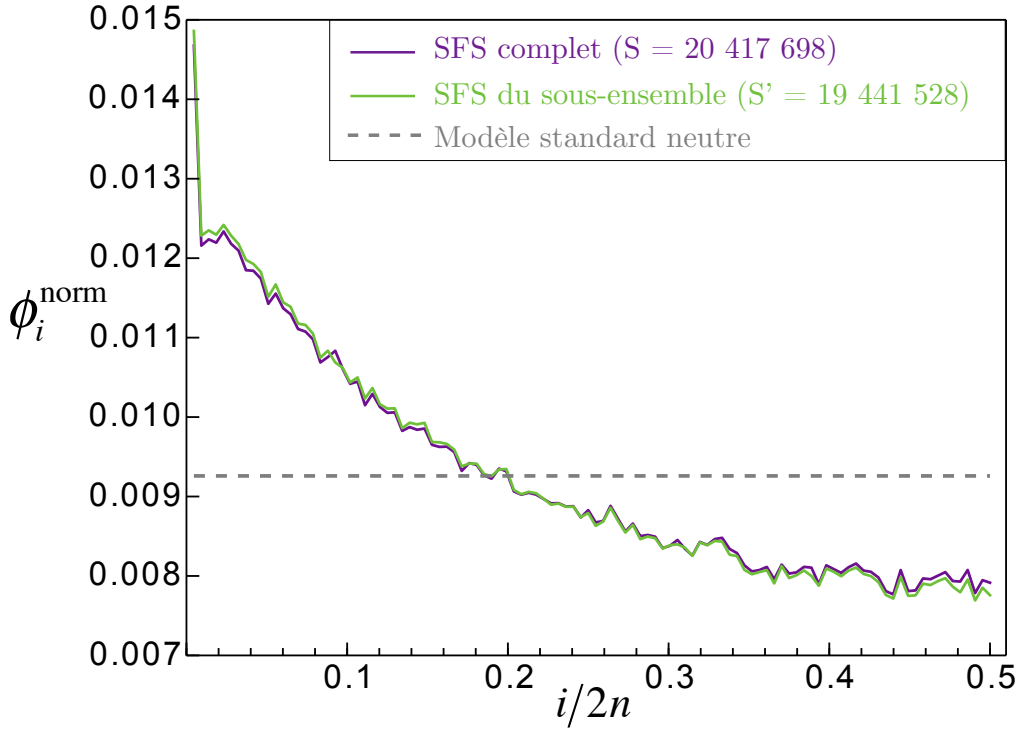


FIGURE S6 – Spectre de fréquence plié de la population Yoruba, pour le jeu de données complet et pour le sous-ensemble des sites orientés. La courbe violette est le spectre construit avec tous les sites. La courbe verte est le spectre construit uniquement avec les sites pour lesquels la base ancestrale est connue. La ligne grise pointillée est le spectre attendu sous le modèle standard neutre sans démographie. Les spectres sont pliés, transformés et normalisés (voir Méthodes de l'article).

### Comparaison approfondie des méthodes contraintes et flexibles appliquées aux données Yoruba

Un des reviewers a suggéré qu'il pourrait être intéressant d'analyser le spectre des Yoruba avec des modèles constants par morceaux par la méthode contrainte  $\partial a \partial i$ , en variant le nombre d'époques. Cela permettrait de voir quel est le nombre optimal d'époques inféré pour la démographie des Yoruba par une analyse avec  $\partial a \partial i$ , et de visualiser le spectre de fréquence prédit sous le modèle optimal.

Nous avons donc fait quelques analyses préliminaires dans ce sens. Nous avons ajouté une ou deux époques au modèle de croissance instantanée (*Sudden* déjà testé dans l'article, qui est un modèle à une époque). Les modèles testés sont décrits dans la Figure S7. Les résultats (Table S1) montrent que l'ajout d'une deuxième époque améliore beaucoup la vraisemblance du modèle. L'ajout d'une troisième époque améliore la vraisemblance dans une moindre mesure, même après correction par test de ratio de vraisemblance pour



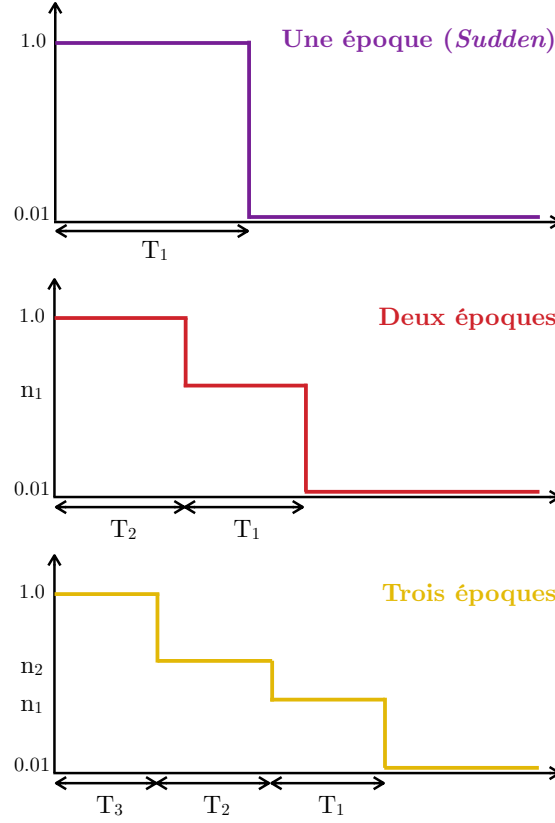


FIGURE S7 – Modèles constants par morceaux avec une à trois époques, testés avec  $\partial a \partial i$ .

le nombre de paramètres. On remarque également que les distances au carré (Table S1) entre le spectre observé et le meilleur spectre prédit par  $\partial a \partial i$  (représenté sur la Figure S8) diminuent avec l'augmentation du nombre d'époques, et atteignent de plus petites distances que celles obtenues avec les modèles à un paramètre dans l'article. Les vraisemblances sont également meilleures que celles obtenues avec les modèles à un paramètre. Les démographies inférées par  $\partial a \partial i$  pour chaque modèle (Figure S9) montrent que la première étape (en remontant dans le temps) est la même pour les modèles à deux et trois époques (diminution de la taille de population à  $\sim 0.6$  au temps  $\sim 0.5$ ). La comparaison des deux scénarios les plus vraisemblables pour le modèle à trois époques (les deux démographies en jaune) montre que la deuxième et la troisième étape peuvent varier en taille et en longueur sans affecter la vraisemblance, ce qui semble indiquer que la première étape est déterminante.

Ces résultats préliminaires soulignent l'intérêt potentiel d'une étude comparative plus poussée de  $\partial a \partial i$  et du stairway plot pour l'inférence de démographies constantes par morceaux, en particulier pour le choix du nombre d'étapes.

	Une époque ( <i>Sudden</i> )	Deux époques	Trois époques
Nombre de paramètres	1	3	5
Meilleure log vraisemblance	-3393	-2119	-2100
$d^2$	$3.4 \times 10^{-4}$	$1.8 \times 10^{-4}$	$1.7 \times 10^{-4}$

TABLE S1 – Ajustement du spectre de fréquence des Yoruba par des modèles constants par morceaux avec  $\partial a \partial i$ . La meilleure log vraisemblance obtenue après 100 exécutions de la méthode  $\partial a \partial i$  est reportée. La distance au carré est calculée entre le spectre observé des Yoruba et le meilleur spectre prédit par  $\partial a \partial i$ .

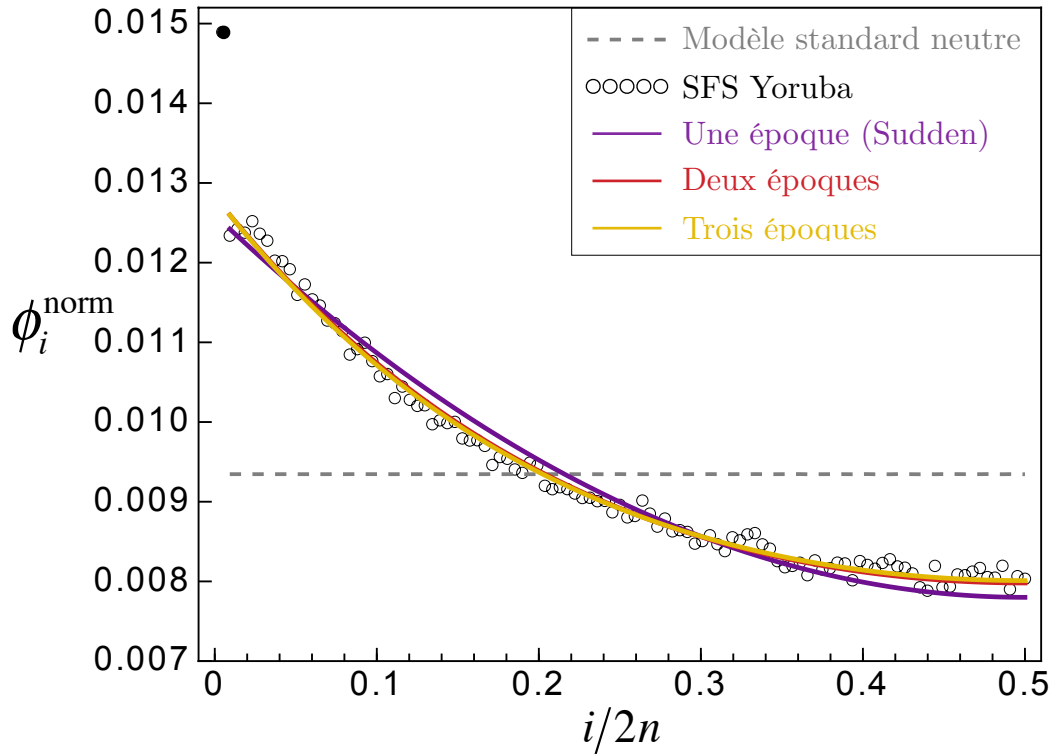


FIGURE S8 – Spectres de fréquence sous chaque modèle constant par morceaux, avec les valeurs de paramètres correspondant à la meilleur log vraisemblance obtenue par  $\partial a \partial i$ . Le spectre des Yoruba est représenté par des ronds vides. Le premier point, coloré en noir, représentant les singletons, n'a pas été pris en compte. La ligne grise pointillée est le spectre attendu sous le modèle standard neutre sans démographie. Les spectres sont pliés, transformés et normalisés (voir Méthodes de l'article).

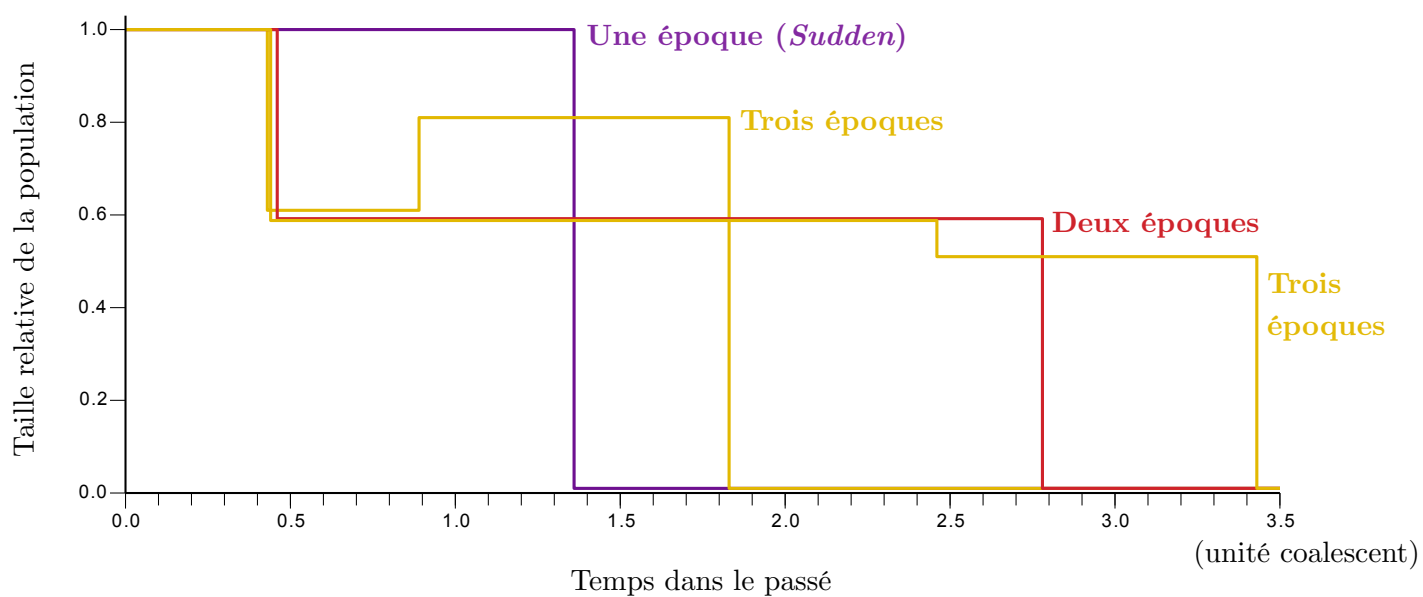


FIGURE S9 – Démographies constantes par morceaux inférées par  $\partial a \partial i$  pour la population Yoruba. Le scénario ayant la meilleure log vraisemblance est dessiné pour chaque modèle. Deux scénarios de log vraisemblances égales sont présentés pour le modèle à trois époques.

## Chapitre 4

# D'autres modèles pour expliquer la diversité des données : l'exemple des modèles à coalescences multiples

Dans cette partie, nous avons pour objectif de confronter des données de fréquences alléliques, issues de projets de séquençage, à d'autres modèles que le modèle standard neutre. L'idée est de montrer qu'à la diversité des données (espèces procaryotes ou eucaryotes, séquences codantes ou non-codantes) pourrait être associée une certaine diversité de modèles nuls : toutes ces séquences n'évoluent pas de la même manière, il pourrait donc être pertinent de ne pas les modéliser de la même manière.

Dans un premier temps, je présente les différents jeux de données rassemblés et utilisés dans cette partie. Après avoir mis en évidence un biais possible dû aux erreurs de séquençage, j'analyse ces données avec un modèle à deux paramètres, permettant de modéliser de la démographie et des coalescences multiples. Les séquences sont différenciées en séquences codantes et non-codantes, afin d'analyser séparément les processus évolutifs différents qui les affectent. Enfin, je montre que les signatures qu'on pourrait attribuer à de la sélection après optimisation du modèle peuvent aussi en partie être dues au phénomène de biais de conversion GC. Ce chapitre est à l'état d'analyses préliminaires qui ne font pas encore l'objet d'un article.

### 4.0.0 Représentation graphique du spectre de fréquence

Avant de passer aux méthodes et résultats de cette partie, je commence par un point méthodologique sur la représentation des spectres de fréquence, qui sera différente de celle utilisée dans les deux articles des chapitres 2 et 3.

Soient  $\xi^{obs}$  le spectre observé et  $\xi^{th}$  un spectre théorique, qui dépend du modèle choisi. On note  $\tilde{\xi}^{obs}$  et  $\tilde{\xi}^{th}$  les spectres normalisés, c'est-à-dire divisés par leur somme :

$$\tilde{\xi}_i^{obs} = \frac{\xi_i^{obs}}{\sum_{i=1}^{n-1} \xi_i^{obs}} \quad \text{et} \quad \tilde{\xi}_i^{th} = \frac{\xi_i^{th}}{\sum_{i=1}^{n-1} \xi_i^{th}} \quad \text{pour } i \in [1, n-1]$$

Dans cette partie, je représente le spectre résiduel par rapport à un spectre théorique attendu. Ce spectre résiduel, nommé  $r$ , est défini comme :

$$r_i = \frac{\tilde{\xi}_i^{obs} - \tilde{\xi}_i^{th}}{\tilde{\xi}_i^{th}} = \frac{\tilde{\xi}_i^{obs}}{\tilde{\xi}_i^{th}} - 1 \quad \text{pour } i \in [1, n-1] \quad (4.1)$$

On représente les résidus entre les données et l'attendu théorique, remis à l'échelle de l'attendu théorique. Cette représentation est similaire à la transformation utilisée dans les articles des chapitres précédents, excepté que le spectre théorique n'était pas soustrait à l'attendu. La transformation correspondait donc au premier terme de l'équation 4.1, et les spectres étaient re-normalisés après cette transformation. L'attendu théorique est maintenant la droite  $y = 0$ .

Deux exemples de spectres résiduels sont présentés dans la Figure 4.1, pour une grande taille d'échantillon ( $n=196$ , *Drosophila melanogaster*) et une petite taille d'échantillon ( $n=20$ , *Armadillidium vulgare*, cloporte commun). L'axe des abscisses sera dans cette partie  $i$ , c'est à dire le nombre de séquences portant l'allèle dérivé dans l'échantillon (et non la fréquence de l'allèle dérivé  $i/n$ ), afin de visualiser la taille de l'échantillon, qui varie selon le jeu de données utilisé.

Le fait de remettre les données observées à l'échelle des données théoriques permet de visualiser la déviation de l'observé par rapport à l'attendu. Par exemple, pour *Drosophila melanogaster*, l'excès de mutations à hautes fréquences par rapport à l'attendu du modèle standard neutre paraît faible (Figure 4.1A gauche), mais une fois remis à l'échelle de l'attendu théorique, on voit que cet excès est très important par rapport à l'attendu (Figure 4.1A droite).

Dans ces exemples, comme dans le chapitre précédent, les résidus sont représentés par rapport au modèle standard neutre. Avec cette nouvelle formalisation, le spectre théorique  $\xi^{th}$  par rapport auquel on calcule les résidus peut être n'importe quel attendu théorique, et pas nécessairement le modèle standard neutre.

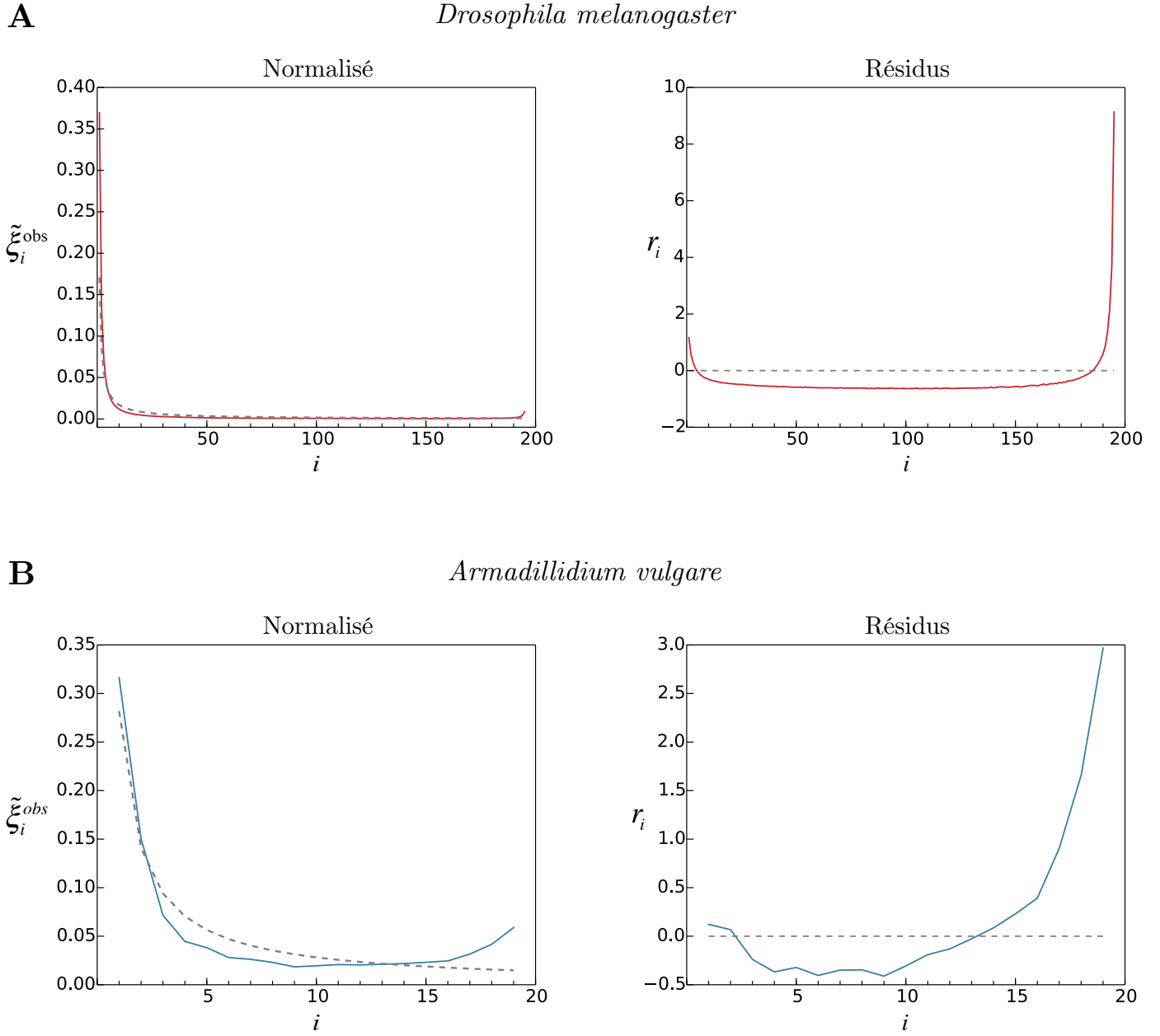


FIGURE 4.1 – Comparaison des spectres de fréquence de *Drosophila melanogaster* (A) et *Armadillidium vulgare* (B), normalisés (à gauche) et résiduels (à droite). Le spectre de fréquence du modèle standard neutre ( $\xi_i = 1/i$ ) est représenté en pointillés gris.

## 4.1 Données

### 4.1.1 Informations sur les données rassemblées

Pour l'étude que nous voulons réaliser, nous utilisons des données de polymorphisme à l'échelle du génome complet, pour un grand nombre d'individus d'une même espèce. Pour chaque espèce je pars de l'alignement des génomes séquencés, ou d'un fichier VCF (Variant Call Format, qui répertorie toutes les positions variables trouvées dans les génomes séquencés par rapport à une séquence de référence) lorsqu'il est fourni par le projet de séquençage. L'analyse de ces données de polymorphisme me permet de construire le spectre de fréquence de l'échantillon. Je récupère par ailleurs sur des bases de données les positions codantes du génome de l'espèce, afin de pouvoir distinguer les mutations qui affectent ou non la séquence des acides aminés. Je détaille ci-dessous les données récupérées pour les différentes espèces que nous avons étudiées dans le cadre de ce projet.

#### *Arabidopsis thaliana*

**Séquences** 1135 génomes complets d'*Arabidopsis thaliana* ont été séquencés dans le cadre du projet «1001 genomes: A Catalog of Arabidopsis thaliana Genetic Variation» (<http://1001genomes.org/>, 1001 Genomes Consortium et al., 2016). Le jeu de données complet étant très parcellaire (la plupart des sites ont au moins un des 1135 génotypes manquant), j'ai travaillé avec un sous-jeu de données plus complet de 345 lignées (séquences haploïdes ; <http://1001genomes.org/projects/MPICWang2013/>).

**Positions codantes** J'ai récupéré dans la base de données Ensembl Plants Genes 35 (<http://plants.ensembl.org/biomart/martview/>) les positions des gènes de la séquence de référence d'*Arabidopsis thaliana* (TAIR10, The Arabidopsis Information Resource). Plus précisément, je récupère les positions des UTR (Untranslated Transcribed Regions) et des exons, définis comme indiqués sur la Figure 4.2, ce qui me permet d'en déduire les positions des CDS (Coding DNA Sequence) et des introns. Sur Ensembl j'ai choisi deux filtres différents pour délimiter les gènes : un filtre strict (Genes with TAIR gene name) qui nous permet de conserver les régions dont on est sûrs qu'elles sont codantes et un filtre peu strict (with NCBI gene) qui nous permet par soustraction d'en déduire les régions dont on est sûr qu'elles ne sont pas codantes. Pour les régions non codantes, on ne prend pas en compte 1kb de part et d'autre des gènes, afin de limiter l'impact de la liaison à des séquences pouvant être sous sélection. Pour les 5 chromosomes d'*Arabidopsis thaliana*, avec la base de données TAIR gene name, on a 21.1% de séquence génique (on considère le gène comme allant du 5' au 3' UTR, introns compris, voir Figure 4.2) et



FIGURE 4.2 – Définition des composantes du gène dans la nomenclature utilisée sur Ensembl. Les exons (en trait plein noir) rassemblent les UTR (Untranslated Transcribed Regions, en points noirs) et les CDS (Coding DNA Sequence, en tirets noirs). Les introns sont en gris.

11.1% de CDS. Avec la base de données NCBI gene, en excluant 1kb de part et d'autre des gènes, on obtient 28.3% de séquence non codante.

**Séquence d'*Arabidopsis lyrata*** Afin d'orienter les mutations (voir section 4.2), j'ai récupéré l'alignement de la séquence de référence d'*Arabidopsis thaliana* avec une espèce proche, *Arabidopsis lyrata* ([http://pipeline.lbl.gov/data/araTha04\\_Araly1/](http://pipeline.lbl.gov/data/araTha04_Araly1/)). 49.0% des sites sont orientés, c'est à dire qu'on a pu inférer leur base ancestrale à partir de l'alignement.

## *Drosophila melanogaster*

**Séquences** 196 génomes complets d'une population africaine de *Drosophila melanogaster* ont été séquencés dans le cadre de la phase 3 du « Drosophila Population Genomics Project » (Lack et al., 2015). Les séquences (haploïdes) ont été récupérées sur le site du DPGP (<http://www.johnpool.net/genomes.html>, fichier « DPGP3 SEQ »).

**Positions codantes** Les positions codantes de la séquence de référence de *Drosophila melanogaster* (BDGP6) ont été récupérées de la même façon que pour *Arabidopsis thaliana*, sur la base de données Ensembl Genes 89 (<http://www.ensembl.org/biomart/martview/>). Pour le filtre strict j'ai gardé uniquement les gènes « with FlyBase annotation IDs ». 49.1% de la séquence des 4 autosomes séquencés (2L, 2R, 3L et 3R) appartient à un gène (au sens large) et 17.8% à un CDS. Avec le filtre non strict (with FlyBase gene IDs), on obtient 29.1% de séquence non codante (et à plus d'1kb d'un gène).

**Séquence de *Drosophila simulans*** La séquence d'une espèce proche de *Drosophila melanogaster*, *Drosophila simulans*, alignée sur le génome de référence de *Drosophila melanogaster*, a été récupérée sur le site du DPGP (fichier « SIMULANS SEQ », Stanley and Kulathinal, 2016). Elle permet d'inférer la base ancestrale de 93.3% des SNP.



## *Escherichia coli*

**Séquences** J’ai repris les 71 séquences d’*Escherichia coli* utilisées dans le chapitre 2 issues de RefSeq. Ces séquences sont codantes.

**Séquence ancestrale** Les séquences sont alignées avec l’espèce sœur *Escherichia fergusonii*, ce qui permet d’inférer la base ancestrale de 92.6% des SNP.

## *Homo sapiens*

**Séquences** J’ai continué à travailler avec les 108 génomes diploïdes de la population Yoruba, issus du projet 1000 génomes et déjà utilisés dans le chapitre 3.

**Positions codantes** Sur la base de données Ensembl Genes 89 de la séquence de référence GRCh37 du génome humain (qui est celle utilisée par le projet 1000 génomes ; <http://grch37.ensembl.org/biomart/martview/>), j’ai utilisé le filtre «with RefSeq peptide ID only». 32.8% de la séquence des autosomes appartient à un gène (au sens large) et 1.09% à un CDS. 66.1% de la séquence est non-codante et à plus d’1kb d’un gène.

**Séquence ancestrale** L’allèle ancestral est fourni dans le fichier VCF du projet 1000 génomes. On connaît ainsi la base ancestrale de 95.2% des SNP.

## Espèces non modèles

**Séquences** Dans le cadre du projet de séquençage d’espèces non-modèles décrit dans Romiguier et al. (2014), les transcriptomes d’une dizaine d’individus ont été séquencés pour 12 espèces, donnant ainsi accès à leurs séquences codantes. Les espèces choisies, les espèces sœurs séquencées et les tailles d’échantillons sont décrites dans la Table 4.1.

**Séquence ancestrale** Les transcriptomes d’un ou plusieurs individus d’une espèce sœur sont disponibles pour chaque espèce étudiée, comme précisé dans la Table 4.1. Le nombre de SNP pour chaque espèce et le pourcentage de SNP orientés sont donnés dans la Table 4.2.

### 4.1.2 Spectres de fréquence observés

Les spectres de fréquence résiduels (voir section 4.0.0) des données décrites ci-dessus sont présentés dans les Figures 4.3 à 4.6.

TABLE 4.1 – Jeu de données d’espèces non-modèles issu de Romiguier et al. (2014). On note  $n$  la taille de l’échantillon pour chaque espèce et  $n_{\text{out}}$  la taille de l’échantillon pour l’espèce sœur choisie. Toutes les espèces sont diploïdes, le nombre de transcriptomes est donc  $2n$  (et  $2n_{\text{out}}$  pour l’espèce sœur).

Espèce étudiée	$n$	Espèce sœur	$n_{\text{out}}$
<i>Aptenodytes patagonicus</i> (Manchot royal)	10	<i>Aptenodytes forsteri</i>	1
<i>Armadillidium vulgare</i> (Cloporte commun)	10	<i>Armadillidium nasatum</i>	2
<i>Artemia franciscana</i> (Artémie)	10	<i>Artemia sinica</i>	2
<i>Caenorhabditis brenneri</i>	10	<i>Caenorhabditis</i> sp.10	2
<i>Culex pipiens</i> (Moustique commun)	10	<i>Culex torrentium</i>	2
<i>Emys orbicularis</i> (Cistude ou tortue de Brenne)	10	<i>Trachemys scripta</i>	2
<i>Halictus scabiosae</i> (Abeille)	11	<i>Halictus simplex</i>	1
<i>Lepus granatensis</i> (Lièvre ibérique)	10	<i>Lepus americanus</i>	1
<i>Ostrea edulis</i> (Huître plate)	10	<i>Ostrea chilensis</i>	2
<i>Parus caeruleus</i> (Mésange bleue)	10	<i>Parus major</i>	1
<i>Physa acuta</i> (Escargot d’eau douce)	9	<i>Physa gyrina</i>	2
<i>Sepia officinalis</i> (Seiche commune)	9	<i>Sepiella japonica</i>	1

TABLE 4.2 – Valeurs caractéristiques du polymorphisme des différentes espèces étudiées (désignées par leur genre). On note  $n_g$  le nombre de copies de génomes dans l’échantillon,  $S$  le nombre total de sites polymorphes, dont je donne le pourcentage pour lesquels on connaît la base ancestrale (% orienté), et  $\pi$  la distance moyenne par paire de séquences, normalisée par la longueur totale de la séquence.

Espèce étudiée	$n_g$	$S$	% orienté	$\pi$
<i>Aptenodytes</i>	20	1644	77.7%	0.95‰
<i>Arabidopsis</i>	345	8 246 331	49.0%	2.9‰
<i>Armadillidium</i>	20	27 193	85.8%	5.1‰
<i>Artemia</i>	20	6666	86.2%	3.0‰
<i>Caenorhabditis</i>	20	2086	64.2%	7.5‰
<i>Culex</i>	20	12 670	43.0%	10.5‰
<i>Drosophila</i>	196	4 998 681	93.3%	5.0‰
<i>Emys</i>	20	647	79.6%	1.2‰
<i>Escherichia</i>	71	102 331	92.6%	16.6‰
<i>Halictus</i>	22	795	89.6%	0.52‰
<i>Homo</i>	216	20 417 698	95.2%	0.98‰
<i>Lepus</i>	20	1066	72.1%	1.0‰
<i>Ostrea</i>	20	1135	82.7%	1.8‰
<i>Parus</i>	20	1097	78.9%	1.6‰
<i>Physa</i>	18	5343	82.1%	6.1‰
<i>Sepia</i>	18	2112	82.4%	0.59‰

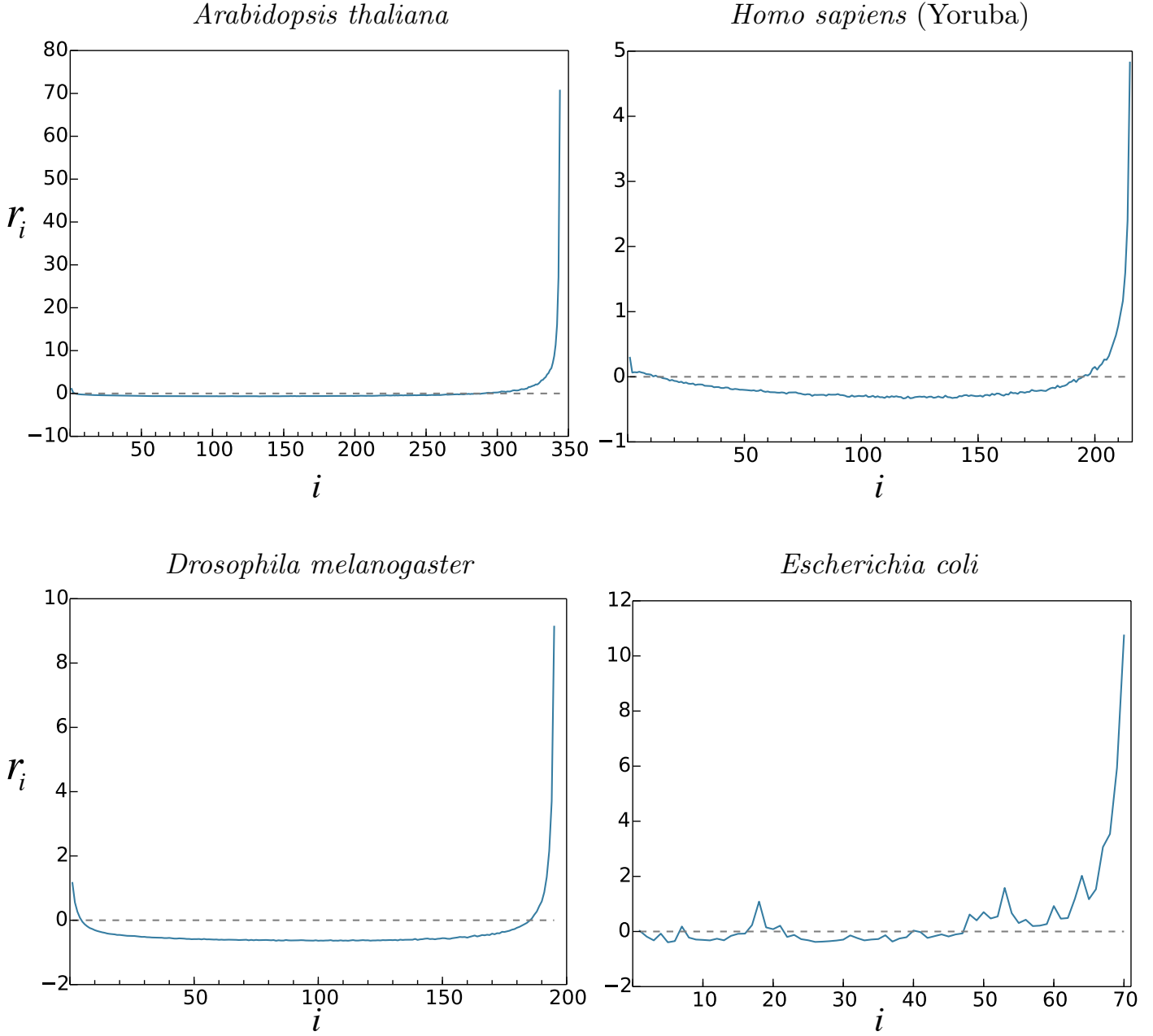


FIGURE 4.3 – Spectres de fréquence résiduels d'*Arabidopsis thaliana* ( $n=345$ ), *Homo sapiens* (Population Yoruba,  $2n=216$ ), *Drosophila melanogaster* ( $n=196$ ) et des séquences codantes d'*Escherichia coli* ( $n=71$ ). L'attendu théorique (en pointillé gris) correspond au modèle standard neutre.

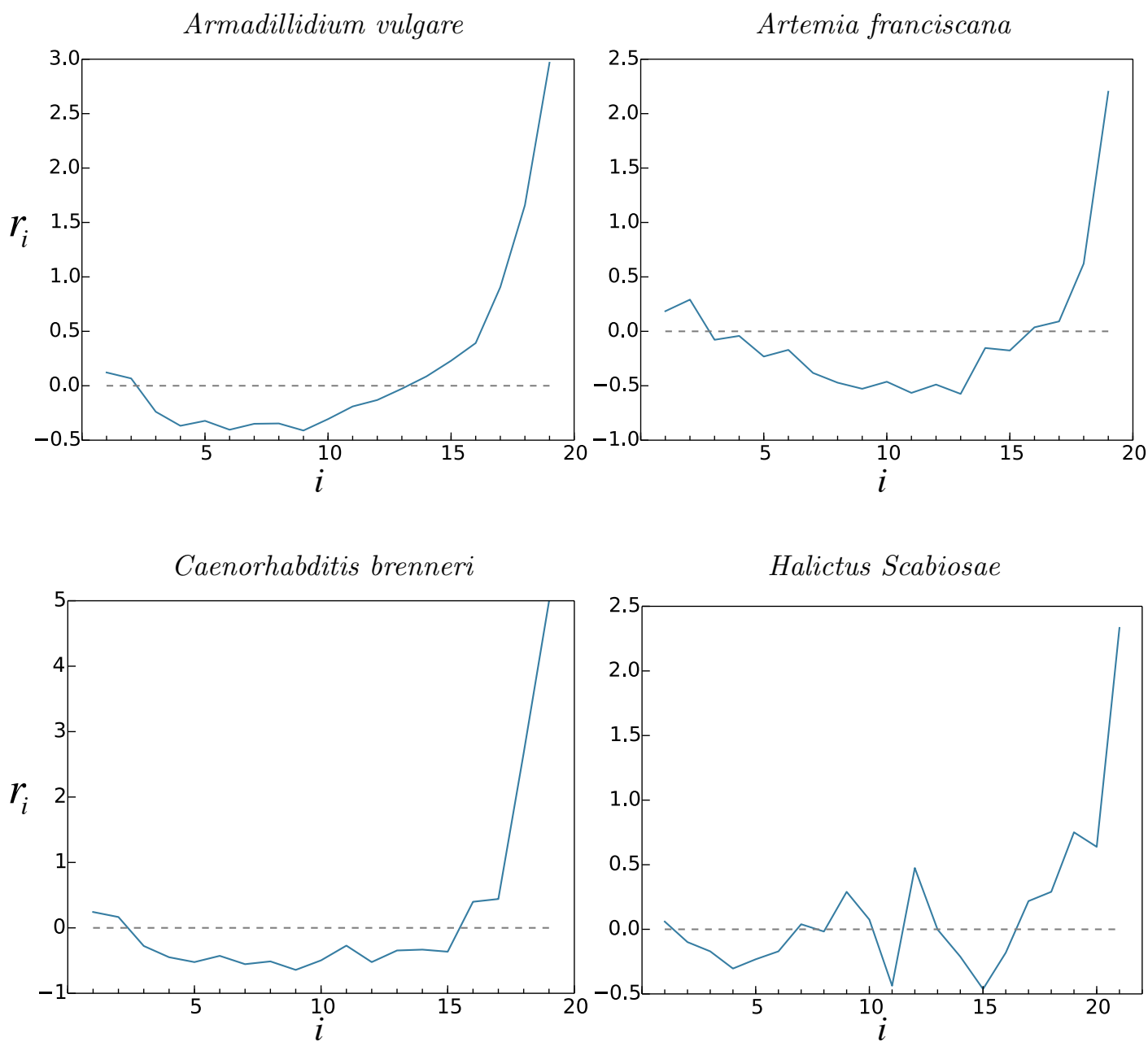


FIGURE 4.4 – Spectres de fréquence résiduels des séquences codantes d'*Armadillidium vulgare* (2n=20), *Artemia franciscana* (2n=20), *Caenorhabditis brenneri* (2n=20) et *Halictus scabiosae* (2n=22). L'attendu théorique (en pointillé gris) correspond au modèle standard neutre.

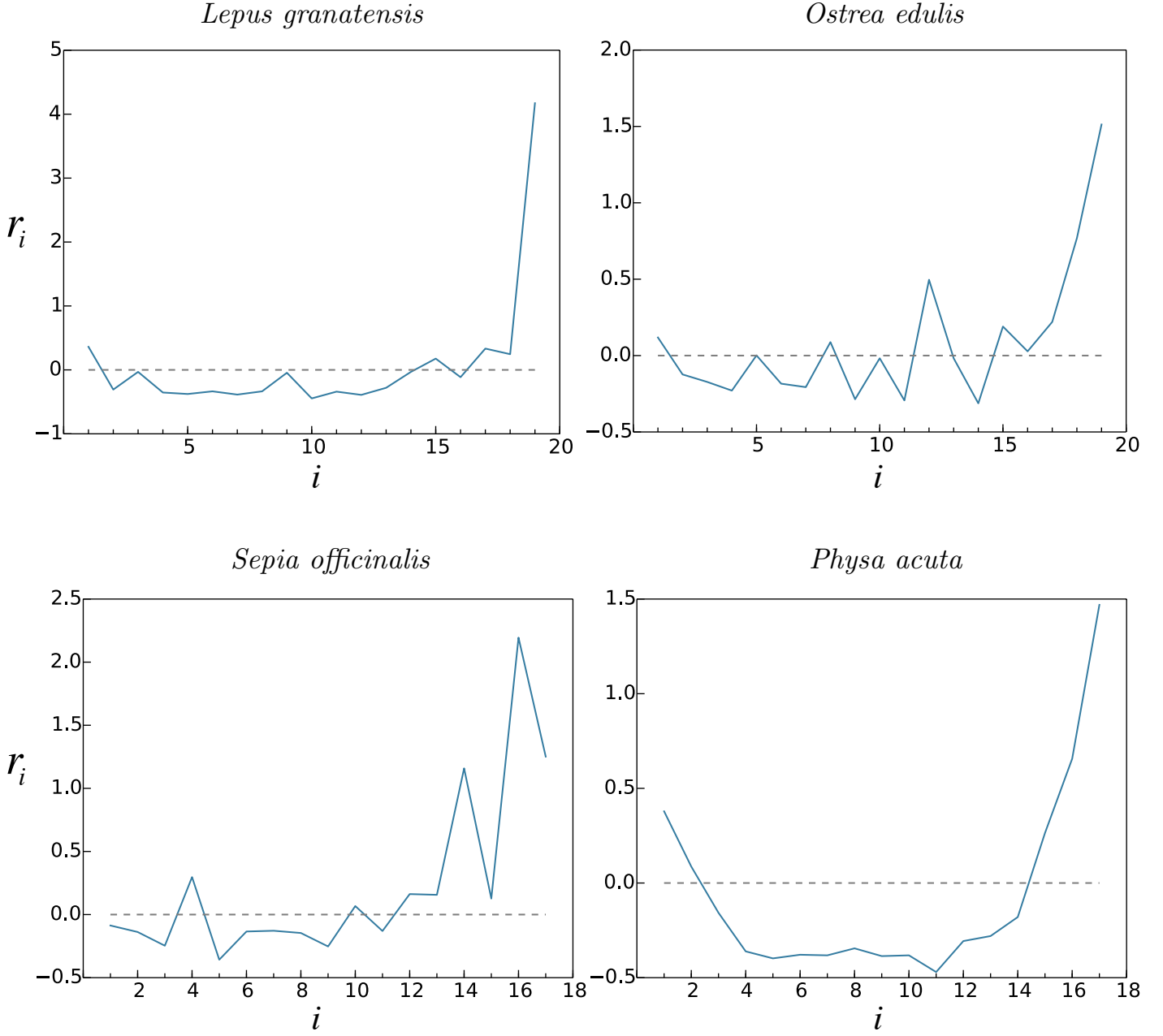


FIGURE 4.5 – Spectres de fréquence résiduels des séquences codantes de *Lepus granatensis* (2n=20), *Ostrea edulis* (2n=20), *Sepia officinalis* (2n=18) et *Physa acuta* (2n=18). L'attendu théorique (en pointillé gris) correspond au modèle standard neutre.

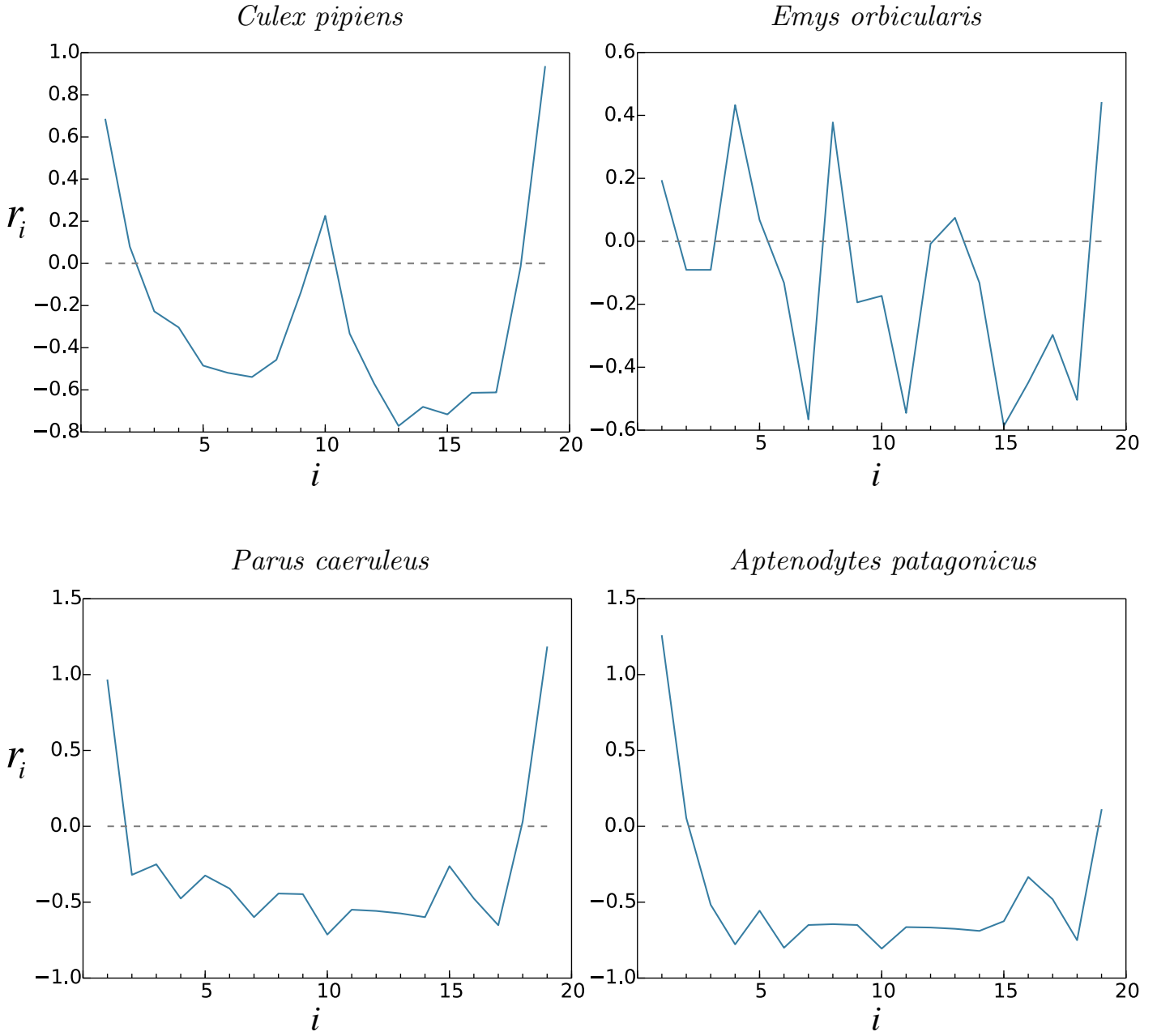


FIGURE 4.6 – Spectres de fréquence résiduels des séquences codantes de *Culex pipiens* ( $2n=20$ ), *Emys orbicularis* ( $2n=20$ ), *Parus caeruleus* ( $2n=20$ ) et *Aptenodytes patagonicus* ( $2n=20$ ). L'attendu théorique (en pointillé gris) correspond au modèle standard neutre.

Dans la majorité des spectres résiduels observés (Figures 4.3, 4.4 et 4.5), on constate un fort excès de mutations à fortes fréquences, parfois associé à un excès plus modeste de mutations à faibles fréquences. Deux spectres résiduels (*P. caeruleus* et *A. patagonicus*, Figure 4.6) présentent un excès de mutations à faibles fréquences aussi ou plus important que l’excès de mutations à fortes fréquences. Enfin, un spectre résiduel (*E. orbicularis*, Figure 4.6) semble être compatible avec l’attendu théorique du modèle standard neutre : les écarts autour de l’attendu (0) sont faibles, et alternent entre valeurs positives et négatives. Ils pourraient donc être dûs au bruit, étant donné le faible nombre de sites polymorphes pour cette espèce ( $S=647$ ).

Le spectre résiduel de *C. pipiens* (Figure 4.6) montre un excès de mutations à fréquence 1/2. C’est un signe de structuration de la population : cet excès est vraisemblablement dû à des mutations apparues et fixées dans un sous-groupe d’individus, et absentes dans l’autre sous-groupe. L’arbre reconstruit de l’échantillon (Neighbour-Joining Tree) montre bien cette structuration en deux groupes de 5 individus chacun (Figure 6.1 en Annexe). Le premier groupe est constitué d’individus provenant de France (2 individus), de Tunisie, d’Algérie et d’Israël. Le deuxième groupe est constitué d’individus provenant de Chine, de La Réunion, des Philippines, du Costa Rica et du Burkina Faso. Cette structuration peut également expliquer le  $\pi$  élevé calculé pour *C. pipiens* (Table 4.2).

Avec les jeux de données de génomes complets (*A. thaliana*, *D. melanogaster* et population Yoruba de *H. sapiens*), on peut construire les spectres de fréquence des données codantes et non-codantes (Figure 4.7). Plus précisément, les spectres « codants » sont construits uniquement avec les sites non-synonymes (c’est-à-dire pour lesquels la mutation modifie l’acide aminé, et qui peuvent donc potentiellement être sous sélection). Les spectres « non-codants » sont construits uniquement avec les sites situés à plus d’1 kb d’un gène (voir description des données, section 4.1.1). Chez *A. thaliana* et *H. sapiens*, l’excès de mutations à fortes fréquences est plus important dans les séquences non-codantes que dans les séquences codantes. Chez *H. sapiens*, on note cependant un excès de mutations à faibles fréquences dans les séquences codantes par rapport aux séquences non-codantes, que l’on ne retrouve pas chez *A. thaliana* ou *D. melanogaster*. Enfin chez *D. melanogaster*, l’excès de mutations à fortes fréquences est légèrement plus important dans les séquences codantes.

### 4.1.3 Virus : spectre de fréquence inadapté

Les virus ayant un génome court, très variable et quasiment dépourvu de séquences non-codantes pour certains, les hypothèses du modèle neutre sont particulièrement in-

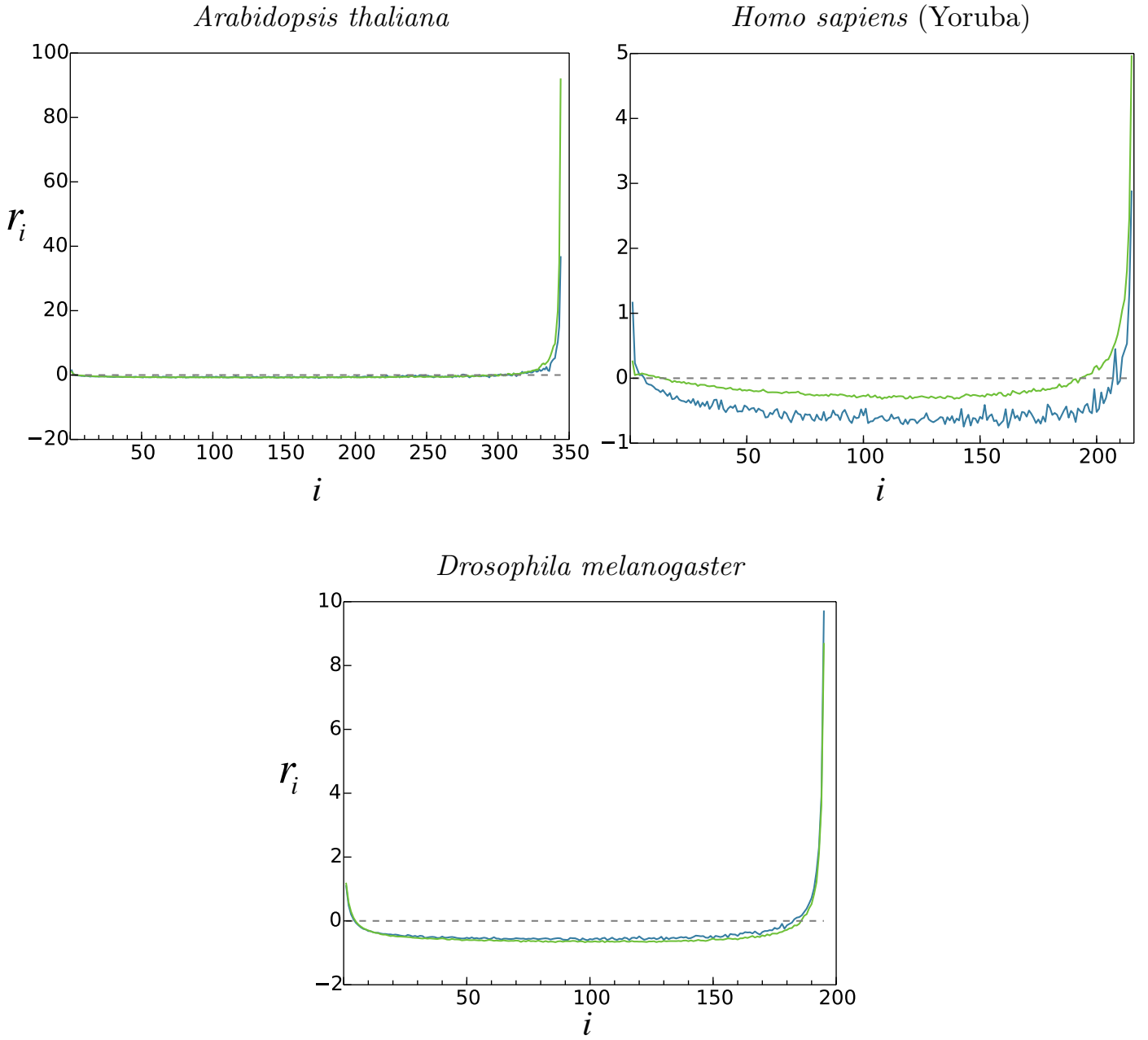


FIGURE 4.7 – Spectres de fréquence résiduels des séquences codantes (en bleu) et non codantes (en vert) d'*Arabidopsis thaliana* ( $n=345$ ), *Homo sapiens* (Population Yoruba,  $2n=216$ ) et *Drosophila melanogaster* ( $n=196$ ). L'attendu théorique (en pointillé gris) correspond au modèle standard neutre.



TABLE 4.3 – Nombre de sites polymorphes issus des données du VIH (pour les nombres de sites bi-, tri- et quadri-alléliques, le pourcentage est calculé par rapport au nombre total de sites polymorphes).

nombre de séquences	170
longueur du génome aligné	11 027
nombre de sites polymorphes	4478
sites bi-alléliques	1947 (43%)
sites tri-alléliques	1335 (30%)
sites quadri-alléliques	1196 (27%)

adaptées pour analyser les données de variation génétique virales. On a vu en introduction les incohérences auxquelles on aboutissait pour la taille efficace des virus, calculée à partir de données génomiques. Nous avons donc cherché à récolter des données de génomes viraux pour les inclure dans cette étude. J’ai commencé par les données du Virus de l’Immunodéficience Humaine (VIH), trouvées sur le site HIV databases (<https://www.hiv.lanl.gov/content/index>). On y trouve un alignement de référence (Subtype Reference Alignement) de 170 séquences de VIH, représentant la diversité des sous-types du virus (article présentant les données : Leitner et al., 2005).

La Table 4.3 donne quelques chiffres sur le polymorphisme de ce jeu de données. On constate que le nombre de sites tri- et quadri-alléliques est du même ordre de grandeur que le nombre de sites bi-alléliques (chacun représentent environ un tiers du nombre total de sites polymorphes). On est donc hors du cadre de l’hypothèse des sites infinis, et ne considérer que les sites bi-alléliques (pour construire le spectre de fréquence) serait très incomplet. De plus, se pose la question de l’orientation de ces mutations (voir section 4.2) : comme les séquences évoluent très rapidement, la reconstruction de l’allèle ancestral paraît compromise. Ainsi, l’étude de l’évolution moléculaire des génomes viraux nécessite le développement d’outils différents, qui sortent du cadre de cette thèse. Le spectre de fréquence n’est pas une statistique adaptée à la représentation et à l’analyse de ces données.

## 4.2 Les erreurs d’orientation

### 4.2.1 Allèle ancestral

Le spectre de fréquence est la distribution des fréquences alléliques dans la population. On entend par fréquence allélique la fréquence de l’allèle dérivé, par opposition à l’allèle ancestral. Or, connaître la fréquence dans la population des deux allèles existants ne suffit

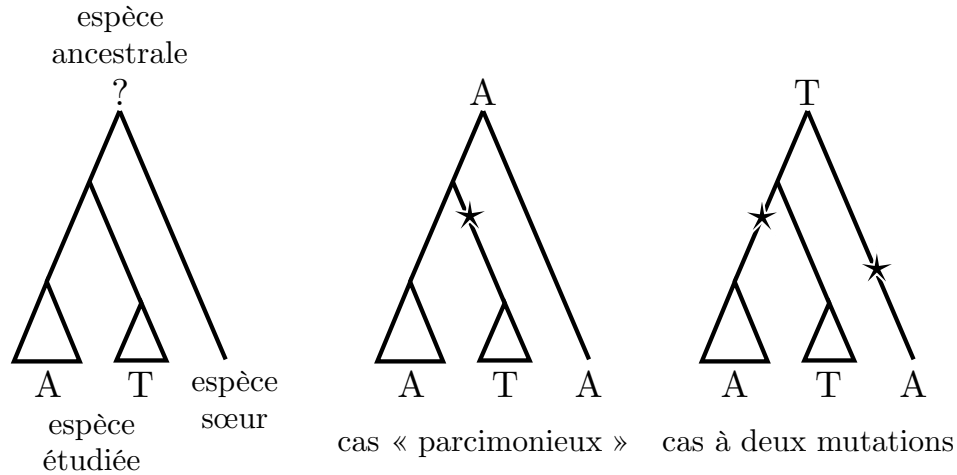


FIGURE 4.8 – Inférer l’allèle ancestral à partir d’une espèce sœur

pas pour connaître la fréquence de l’allèle dérivé. Il faut pour cela identifier quelle est la version ancestrale de l’allèle, c’est à dire l’allèle qui était porté par toute la population avant qu’une mutation fasse apparaître un deuxième allèle, que l’on appelle dérivé.

Pour inférer l’allèle ancestral, on se base en général sur la comparaison avec la séquence d’un individu d’une espèce sœur de l’espèce étudiée (Figure 4.8). Dans le cas « parcimonieux », on suppose qu’une seule mutation a eu lieu, dans l’espèce étudiée, et que l’allèle observé chez l’espèce sœur est l’allèle ancestral. Cependant, si il y a eu deux mutations, une dans l’espèce étudiée et une dans la branche de l’espèce sœur, et que ces mutations ont donné la même base (A dans l’exemple de la Figure 4.8), on commet une erreur en supposant que l’allèle ancestral est le même que l’allèle porté par l’espèce sœur. C’est ce que l’on appelle une erreur d’orientation.

Dans un troisième cas de figure, on peut avoir dans l’espèce sœur un autre allèle, qui ne correspond à aucun des deux allèles trouvés dans l’espèce étudiée (dans l’exemple de la Figure 4.8, c’est le cas où l’espèce sœur a une base G ou C). Dans ce cas, on ne peut rien dire sur la base ancestrale, mais on sait qu’on se trouve dans un cas à (au moins) deux mutations.

#### 4.2.2 Effets des erreurs sur le spectre de fréquence

Pour déterminer l’effet que peuvent avoir les erreurs d’orientation sur la forme du spectre de fréquence, nous avons simulé des spectres avec un certain taux d’erreur. Soit  $f$  ce taux d’erreur, c’est à dire la fraction de mutations mal orientées présente dans chaque case du spectre. On peut exprimer le spectre de fréquence avec erreurs,  $\xi^{err}$  en fonction

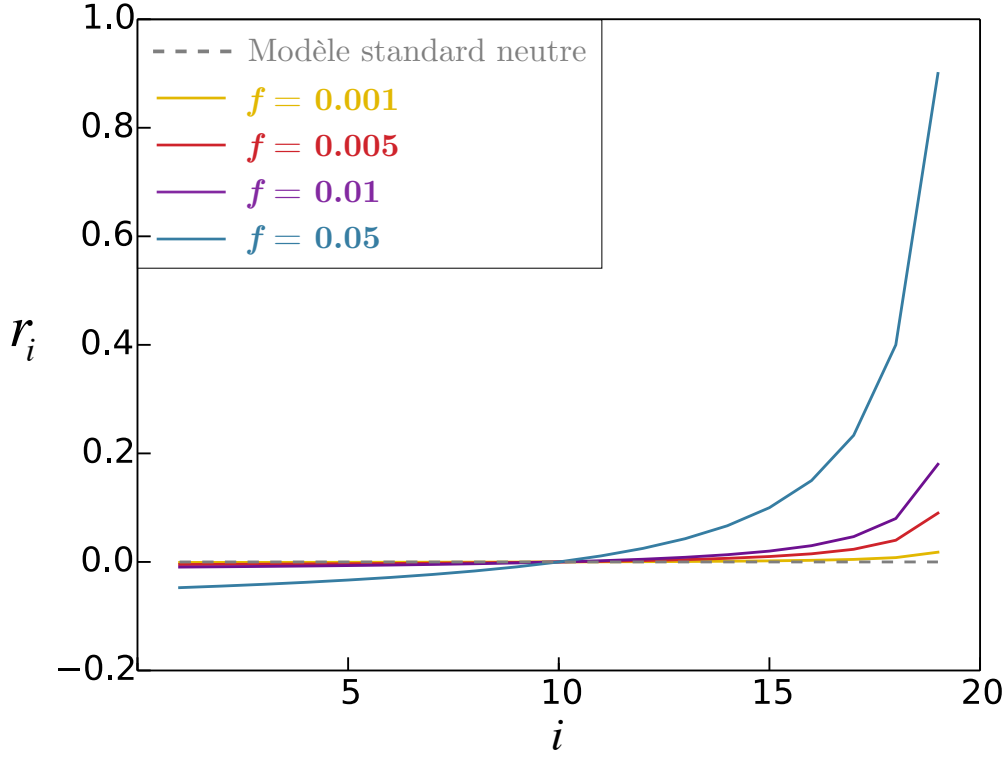


FIGURE 4.9 – Spectre de fréquence résiduel du modèle standard neutre avec erreurs d’orientation à taux  $f$

du vrai spectre sans erreurs  $\xi$  :

$$\xi_i^{err} = f\xi_{n-i} + (1-f)\xi_i$$

où  $n$  est la taille de l’échantillon et  $i \in [1, n-1]$ .

La Figure 4.9 montre l’effet des erreurs d’orientation sur le spectre résiduel par rapport au modèle standard neutre. Plus le taux d’erreur  $f$  est grand, plus on observe un excès apparent de mutations à hautes fréquences, bien que le phénomène des erreurs d’orientation soit symétrique. Cela est dû à l’attendu théorique, qui est de la forme  $1/i$  : pour des grandes valeurs de  $i$ , les résidus entre le spectre avec erreur et le spectre théorique, remis à l’échelle de l’attendu théorique, sont plus importants.

Si on introduit des erreurs d’orientation dans un spectre de fréquence simulé avec de la croissance linéaire, à partir d’un certain taux d’erreur, on obtient un spectre « en U » (excès de mutations à faibles et fortes fréquences, Figure 4.10).

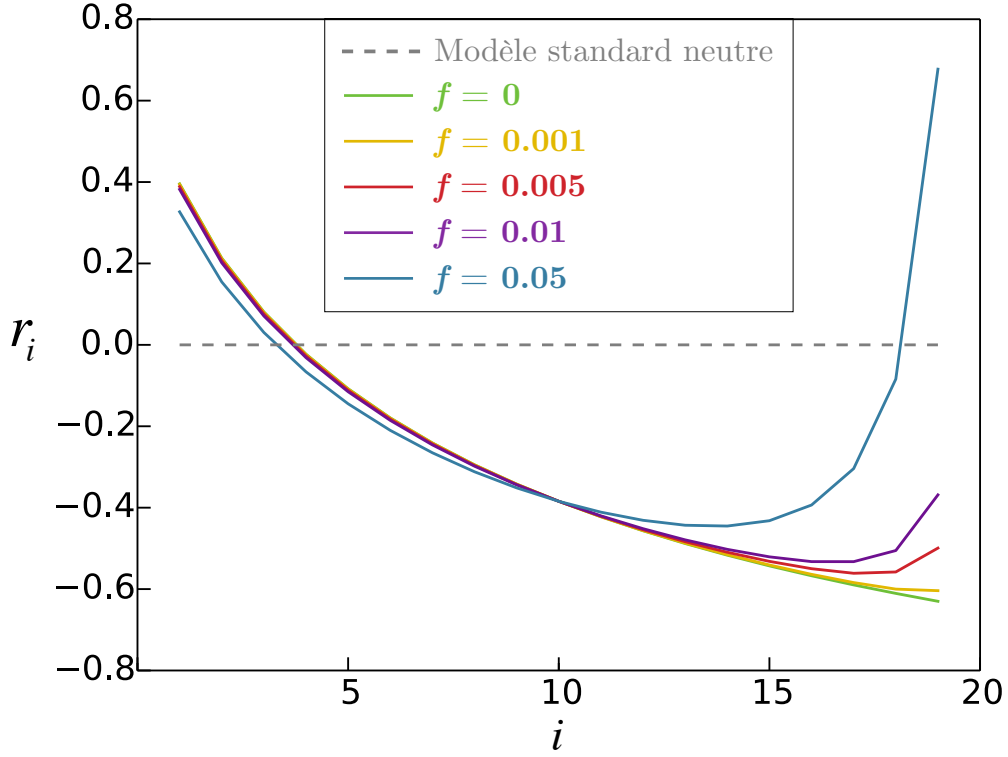


FIGURE 4.10 – Spectre de fréquence résiduel sous croissance linéaire (temps de fondation d’une unité coalescente) avec erreurs d’orientation à taux  $f$ .

### 4.2.3 Estimer et corriger les erreurs à partir des données

Lorsque l’on aligne les séquences de l’espèce étudiée avec une espèce sœur pour déterminer la forme ancestrale des allèles, on trouve un certain nombre de sites pour lesquels l’allèle porté par l’espèce sœur ne correspond à aucun des deux allèles de l’espèce étudiée. Ces cas de figure correspondent au cas à deux mutations présenté Figure 4.8, et pour lequel la mutation sur la branche de l’espèce sœur a fait apparaître un allèle différent (G ou C dans l’exemple de la figure) de celui apparu dans l’espèce étudiée (A ou T). Notons  $S$  le nombre de sites bi-alléliques dans l’échantillon de l’espèce étudiée. Parmi ces  $S$  sites, on peut distinguer  $S_+$  sites pour lesquels l’espèce sœur porte un allèle identique à un des deux allèles de l’espèce étudiée, et  $S_-$  sites pour lesquels l’espèce sœur porte un allèle différent des deux allèles de l’espèce étudiée. Parmi les  $S_-$  sites, on cherche à savoir combien de sites correspondent au cas à deux mutations présenté dans la Figure 4.8 et pour lesquels on a donc fait une erreur d’orientation. On peut estimer, dans un modèle de mutation de Jukes-Cantor (Jukes et al., 1969) où toutes les mutations sont équiprobables, que parmi les  $S_-$  sites, on a  $S_-/2$  sites pour lesquels on est dans le cas de figure à deux mutations. En effet, en reprenant l’exemple de la Figure 4.8, si la branche de l’espèce sœur

mute vers un G ou un C, on comptabilisera ces sites dans  $S_{\neq}$ . On peut donc dire que la mutation s'est faite vers un A deux fois moins souvent que vers un G ou un C, c'est à dire  $S_{\neq}/2$  fois. La fraction de sites mal orientés est donc

$$\hat{f}_{\text{JC}} = \frac{S_{\neq}/2}{S}$$

Si on prend l'exemple des données de *D. melanogaster* ( $n = 196$ ), on a  $S_{=} = 4\,662\,706$  et  $S_{\neq} = 151\,138$ , d'où un taux d'erreurs d'orientation estimé à  $\hat{f}_{\text{JC}} = 0.0157$ .

On peut raffiner cette correction en prenant un modèle de mutation plus réaliste, qui tient compte des taux différents de transversion et de transition (Kimura, 1980). En notant  $\alpha$  le nombre de transitions et  $\beta$  le nombre de transversions, on a

$$\hat{f}_{\text{K2P}} = \frac{\hat{q}S_{\neq}}{S} \quad \text{où} \quad \hat{q} = \frac{\alpha^2 + 2\beta^2}{2\beta(2\alpha + \beta)} \quad (\text{Baudry and Depaulis, 2003})$$

Toujours avec les données de *D. melanogaster*, on a  $\alpha = 2\,689\,865$  et  $\beta = 2\,308\,816$ , ce qui donne un taux d'erreurs d'orientation estimé à  $\hat{f}_{\text{K2P}} = 0.0158$ , qui n'est dans cet exemple que subtilement différent du taux estimé avec le modèle de Jukes-Cantor.

Avec le taux d'erreurs d'orientation estimé, on peut corriger le spectre observé de *D. melanogaster*  $\xi^{\text{obs}}$ , qui s'exprime en fonction du « vrai » spectre corrigé  $\xi^{\text{corr}}$  comme :

$$\begin{cases} \xi_i^{\text{obs}} = (1 - \hat{f})\xi_i^{\text{corr}} + \hat{f}\xi_{n-i}^{\text{corr}} \\ \xi_{n-i}^{\text{obs}} = \hat{f}\xi_i^{\text{corr}} + (1 - \hat{f})\xi_{n-i}^{\text{corr}} \end{cases}$$

On peut donc exprimer le spectre corrigé  $\xi^{\text{corr}}$  en fonction du spectre observé  $\xi^{\text{obs}}$  :

$$\xi_i^{\text{corr}} = \frac{(1 - \hat{f})\xi_i^{\text{obs}} - \hat{f}\xi_{n-i}^{\text{obs}}}{1 - 2\hat{f}} \quad (4.2)$$

Le spectre de fréquence corrigé de *D. melanogaster* est présenté dans la Figure 4.11. À gauche, les spectres résiduels avant et après correction sont présentés par rapport au modèle standard neutre. La correction diminue sensiblement l'excès de mutations à hautes fréquences. À droite, le spectre observé résiduel est représenté par rapport au modèle standard neutre avec erreurs d'orientation : le modèle théorique est donc un modèle neutre sans démographie mais tenant compte de l'existence d'erreurs d'orientation à taux  $\hat{f}$ . Cela permet de visualier d'une autre façon le signal restant à expliquer dans ces données, après prise en compte des erreurs d'orientation.

En raison de la faible différence observée sur nos données entre  $\hat{f}_{\text{K2P}}$  et  $\hat{f}_{\text{JC}}$ , et comme ce projet cherche à capturer des tendances et non à faire des inférences particulièrement précises, j'ai calculé et utilisé  $\hat{f}_{\text{JC}}$  dans la suite du projet. La Table 4.4 présente les  $\hat{f}_{\text{JC}}$  inférés pour les espèces étudiées.

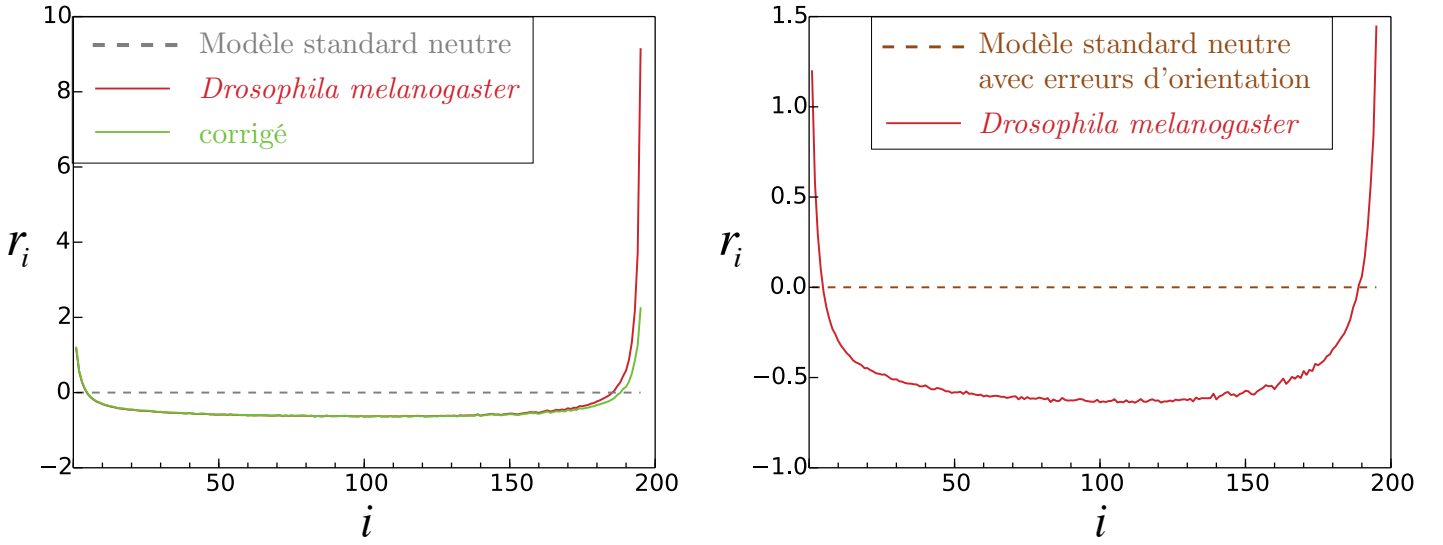


FIGURE 4.11 – Spectre de fréquence résiduel de *Drosophila melanogaster*. À gauche, sans et avec correction (modèle Kimura à deux paramètres,  $\hat{f}_{k2p} = 0.0158$ ) par rapport au modèle standard neutre. À droite, sans correction, par rapport au modèle standard neutre avec erreurs d'orientation à taux  $\hat{f}_{k2p} = 0.0158$ .

TABLE 4.4 – Taux d'erreurs d'orientation  $\hat{f}_{jc}$  estimé pour les espèces étudiées

Espèce étudiée	$\hat{f}_{jc}$
<i>A. patagonicus</i>	0.005
<i>A. thaliana</i> (sites non-synonymes)	0.025
<i>A. thaliana</i> (régions non-codantes)	0.058
<i>A. vulgare</i>	0.015
<i>A. franciscana</i>	0.021
<i>C. brenneri</i>	0.037
<i>D. melanogaster</i> (sites non-synonymes)	0.016
<i>D. melanogaster</i> (régions non-codantes)	0.016
<i>E. orbicularis</i>	0.013
<i>E. coli</i>	0.037
<i>H. scabiosae</i>	0.007
<i>H. sapiens</i> (sites non-synonymes)	0.001
<i>H. sapiens</i> (régions non-codantes)	0.003
<i>L. granatensis</i>	0.007
<i>O. edulis</i>	0.014
<i>P. caeruleus</i>	0.011
<i>P. acuta</i>	0.019
<i>S. officinalis</i>	0.015

## 4.3 Ajuster les données avec des coalescences multiples et de la démographie

Une majorité des spectres de fréquence des données rassemblées présentent un excès de mutations à fortes fréquences, qui est incompatible avec un modèle standard neutre, même avec de la démographie. On a vu en introduction que le spectre de fréquence des modèles à coalescences multiples présentait un excès de mutations à fortes fréquences. Dans cette partie, on cherche donc à ajuster un modèle beta-coalescent avec de la démographie à l'ensemble des données que nous avons décrites ci-dessus.

### 4.3.1 Méthodes

#### Description mathématique d'un coalescent multiple

Soit  $\lambda_{n,k}$  le taux auquel chaque ensemble fixé de  $k$  lignées d'un échantillon de taille  $n$  coalescent :

$$\lambda_{n,k} = \int_0^1 x^{k-2}(1-x)^{n-k}\Lambda(dx) \quad \text{avec } 2 \leq k \leq n \quad (4.3)$$

où  $\Lambda$  est une mesure de probabilité sur  $[0, 1]$  qui détermine la probabilité avec laquelle les événements d'une fréquence donnée se produisent.

Cette formule est générale à l'ensemble des coalescents, multiples ou non. Pour le coalescent de Kingman, on peut exprimer  $\Lambda$  comme étant la mesure de Dirac en 0, notée  $\Lambda(dx) = \delta_0(dx)$ . Ainsi, on a  $\lambda_{n,k} = 0$  pour  $3 \leq k \leq n$  et  $\lambda_{n,2} = 1$  : seules les coalescences de 2 lignées simultanément sont possibles.

Les beta-coalescents sont une des classes de modèles à coalescences multiples, pour lesquels  $\Lambda$  suit une loi Beta de paramètres  $(\alpha, 2 - \alpha)$  avec  $1 \leq \alpha \leq 2$  (Schweinsberg, 2003; Birkner and Blath, 2008) :

$$\Lambda(dx) = \frac{1}{\Gamma(2 - \alpha)\Gamma(\alpha)} x^{1-\alpha}(1-x)^{\alpha-1} dx \quad (4.4)$$

où  $\Gamma$  est la fonction gamma. Le coalescent de Kingman est un beta-coalescent avec  $\alpha = 2$ , c'est-à-dire une loi Beta de paramètres  $(2, 0)$ . Le cas où  $\alpha = 1$  est appelé coalescent de Bolthausen-Sznitman (Bolthausen and Sznitman, 1998) et correspond à une distribution uniforme de l'intensité à laquelle les coalescences d'une certaine taille  $k$  se produisent.

La Figure 4.12 présente les spectres de fréquence résiduels d'un beta-coalescent, obtenus par simulation, par rapport au modèle standard neutre, pour différentes valeurs de  $\alpha$  entre 1 et 2. Pour  $\alpha = 2$ , on retrouve bien un spectre équivalent au modèle standard

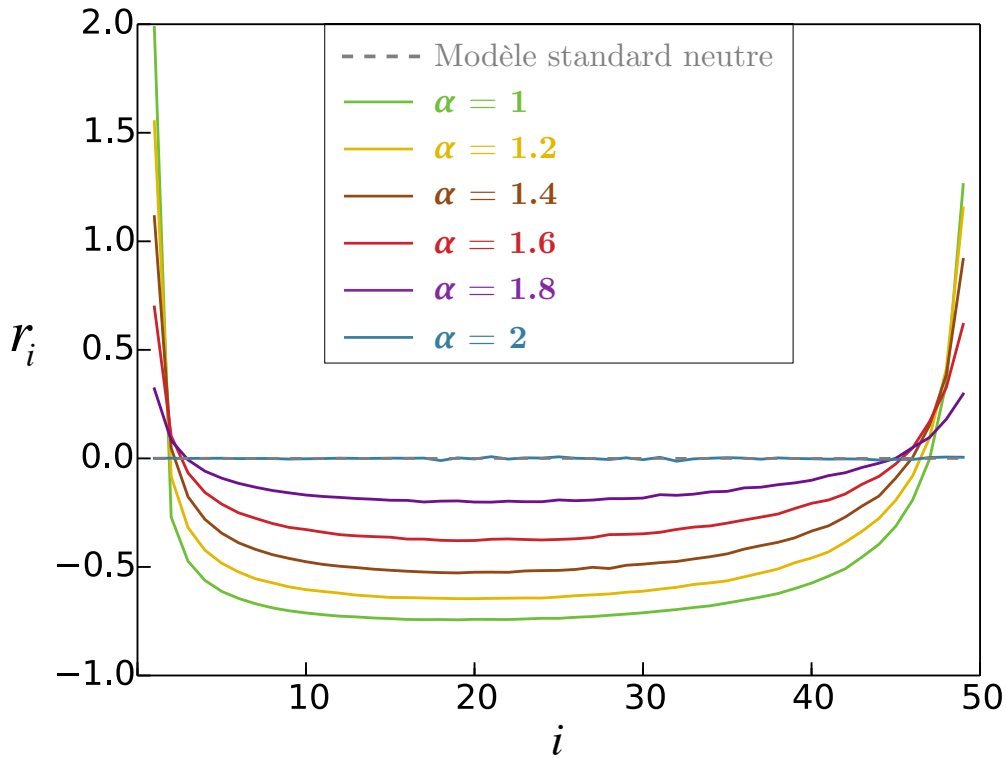


FIGURE 4.12 – Spectres de fréquence résiduels d’un beta-coalescent pour différentes valeurs du paramètre  $\alpha$ , par rapport au modèle neutre.

neutre. Plus  $\alpha$  diminue, plus le spectre résiduel présente un excès de mutations à faibles et fortes fréquences, c’est-à-dire une forme en U.

### Modèle à coalescences multiples et démographie

Guillaume ACHAZ a implémenté un simulateur de beta-coalescent avec démographie. Le beta-coalescent est caractérisé par le paramètre  $\alpha$ , qui peut varier entre 1 et 2 et dont on a vu l’effet sur le spectre de fréquence dans la Figure 4.12. La démographie est exponentielle, croissante ou décroissante, caractérisée par son taux  $g$  tel que la démographie s’exprime comme  $N(t) = N_0 \exp(-gt)$ . En temps retrospectif, avec  $g$  positif, quand  $t$  augmente, la taille de la population diminue : cela correspond à une population croissante en temps prospectif. Inversement quand  $g$  est négatif on modélise une décroissance de population. L’échelle de temps n’est pas linéaire avec  $N$  comme dans un modèle de Kingman, la remise à l’échelle des temps coalescents se fait par un facteur  $1/N^{\alpha-1}$ , c’est-à-dire qu’on mesure le temps en unités de  $N^{\alpha-1}$  générations.



## Ajustement des données

L’ajustement aux données se fait en minimisant la distance au carré  $d^2$  entre le spectre observé et le spectre prédit, comme dans l’article du Chapitre 3 (voir Méthodes de cet article). Cette fois-ci, le modèle a deux paramètres ( $\alpha$  et  $g$ ), l’optimisation se fait donc en deux dimensions. Comme ce travail est préliminaire, je n’ai pas utilisé de méthode d’optimisation particulière, je parcours l’espace par une grille grossière, que je raffine peu à peu en fonction des résultats. Les données utilisées sont les spectres corrigés pour les erreurs d’orientation, ce qui explique les différences entre les spectres observés représentés dans cette section et ceux présentés dans les Figures 4.3 à 4.6 qui n’étaient pas corrigés.

### 4.3.2 Résultats

#### Populations en croissance avec $\alpha \neq 2$

L’ajustement du modèle à deux paramètres au spectre corrigé des séquences codantes d’*A. patagonicus* est présenté Figure 4.13. Pour les valeurs optimisées des paramètres  $\hat{\alpha} = 1.32$  et  $\hat{g} = 6.8$ , on obtient un spectre prédit qui ajuste bien le spectre observé, même si celui-ci est bruité. On constate que le spectre résiduel observé par rapport au spectre optimisé (Figure 4.13B droite), ne semble plus présenter de signal, mais uniquement du bruit (ou une structuration complexe, voir discussion). Ces données peuvent donc s’expliquer par un modèle incluant des coalescences multiples et de la croissance démographique forte.

Si l’on ajuste un modèle purement démographique sans coalescences multiples (voir modèle *Exponentiel* du Chapitre 3), on trouve un taux de croissance de  $g = 17$  : si l’on n’autorise pas les coalescences multiples, on surestime le taux de croissance. Par ailleurs, on voit sur la Figure 4.13A que le paramètre déterminant pour la distance entre l’observé et le modèle est le paramètre  $\alpha$  : c’est selon cet axe que l’on observe les plus grandes variations de  $d^2$ , le paramètre de démographie  $g$  ayant un effet plus modéré sur  $d^2$ .

La Figure 4.14 montre la distance entre le spectre observé d’*A. patagonicus* et les spectres prédits sous les différents modèles testés, avec ou sans les paramètres de croissance ( $g$ ) et de coalescences multiples ( $\alpha$ ).

On obtient des résultats similaires pour les séquences codantes de *P. caeruleus*, dont le spectre de fréquence est aussi en U. Pour cette espèce, on estime  $\alpha = 1.22$  et  $g = 0.40$  (voir Figure 6.2 en Annexe).

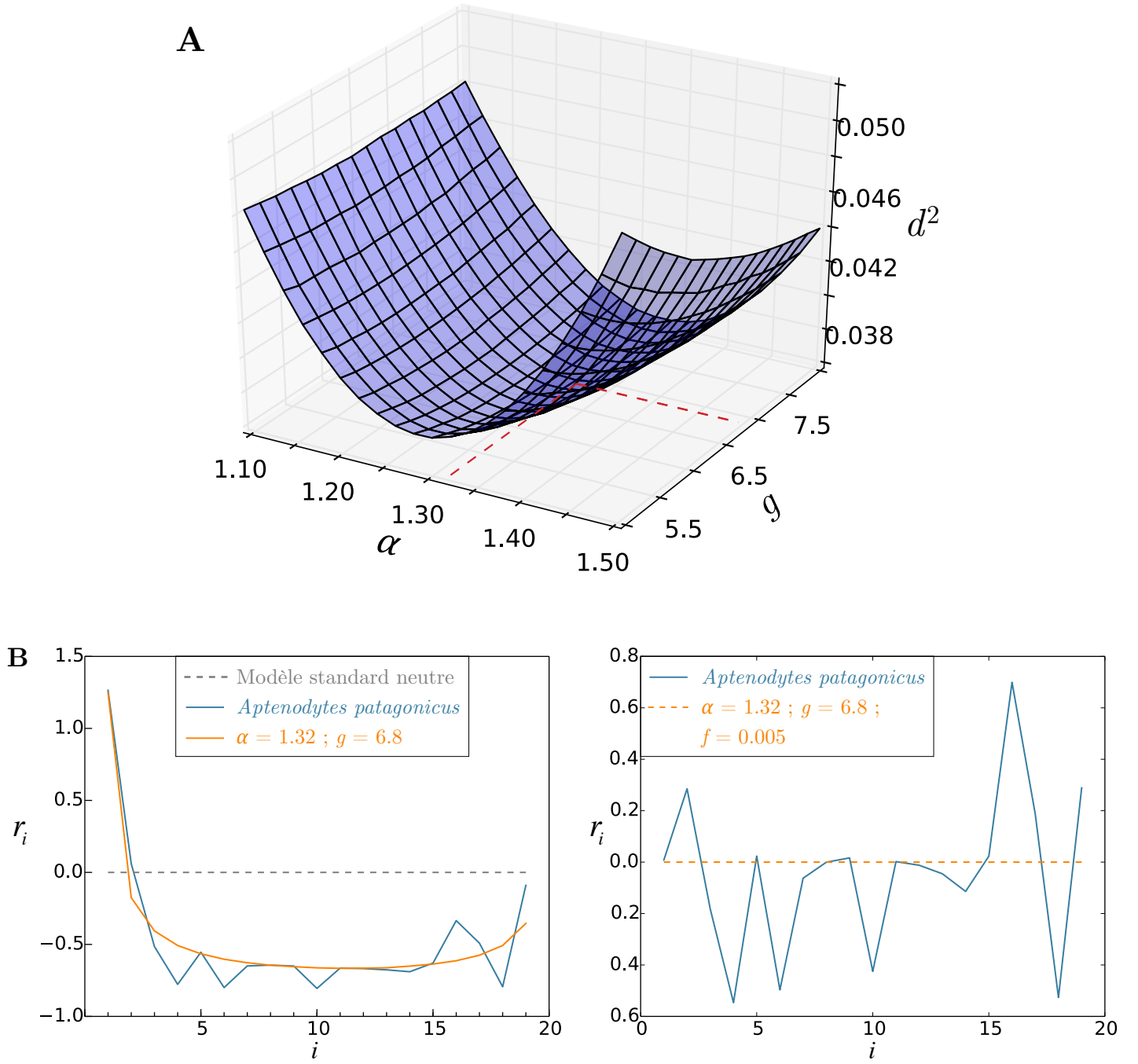


FIGURE 4.13 – A) Distance au carré entre le spectre corrigé d'*Aptenodytes patagonicus* et le spectre théorique du modèle beta-coalescent avec démographie, en fonction des paramètres  $\alpha$  et  $g$ . Les lignes rouges indiquent les coordonnées du minimum,  $d^2 = 0.036$ . B) À gauche, spectre résiduel corrigé d'*Aptenodytes patagonicus* et du modèle beta-coalescent optimisé avec  $\alpha = 1.32$  et croissance exponentielle à taux  $g = 6.8$  par rapport au modèle standard neutre. À droite, spectre résiduel d'*Aptenodytes patagonicus* par rapport au modèle beta-coalescent optimisé avec  $\alpha = 1.32$ ,  $g = 6.8$  et taux d'erreurs d'orientation  $f = 0.005$ .

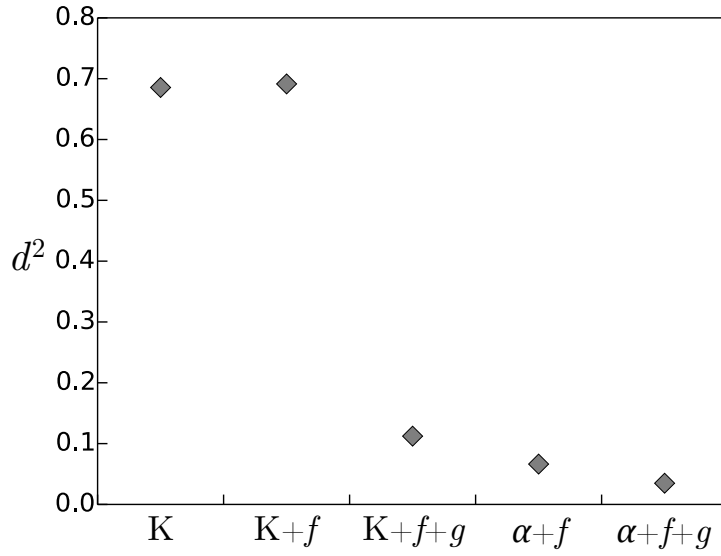


FIGURE 4.14 – Distance au carré  $d^2$  obtenue entre le spectre observé d'*Aptenodytes patagonicus* et le spectre prédit par 5 modèles : K représente le modèle de Kingman,  $f$  les erreurs d'orientation ( $f=0.005$ ),  $\alpha$  le modèle beta-coalescent et  $g$  la croissance exponentielle. Les deux seuls paramètres optimisés sont  $g$  et  $\alpha$ , les deux premiers modèles (K et K+f) ne sont donc pas optimisés. Dans le modèle K+f+g,  $\hat{g} = 17$ . Dans le modèle  $\alpha+f$ ,  $\hat{\alpha} = 1.03$ . Dans le modèle  $\alpha+f+g$ ,  $\hat{\alpha} = 1.32$  et  $\hat{g} = 6.8$ .

### Populations en croissance avec $\alpha = 2$

On avait vu que le spectre de fréquence d'*E. orbicularis* semblait compatible avec le modèle standard neutre (Figure 4.6). Lorsqu'on ajuste un modèle beta-coalescent à ces données, on trouve bien un  $\alpha$  optimisé de 2, c'est à dire un coalescent de Kingman. Le modèle optimisé a un taux de croissance exponentielle de  $g = 0.35$  qui améliore légèrement la distance au carré (Figure 4.15, sans démographie,  $d^2 = 0.078$  et avec croissance exponentielle,  $d^2 = 0.061$ ).

### Modéliser de la décroissance exponentielle

Pour certaines espèces, comme *A. vulgare*, le modèle qui semble convenir le mieux est un beta-coalescent dans une population en décroissance. En effet, on observe sur la Figure 4.16A que la plus petite valeur de  $d^2$  est obtenue pour un  $g$  négatif, c'est-à-dire une population en décroissance exponentielle. Cependant, on constate que dans une partie de la grille de valeurs de paramètres,  $d^2$  n'a pas pu être calculé. Pour ces valeurs de paramètres, le modèle beta-coalescent avec démographie n'a pas pu être simulé, à cause de la remise à l'échelle des temps coalescents. En effet, lorsqu'on modélise une

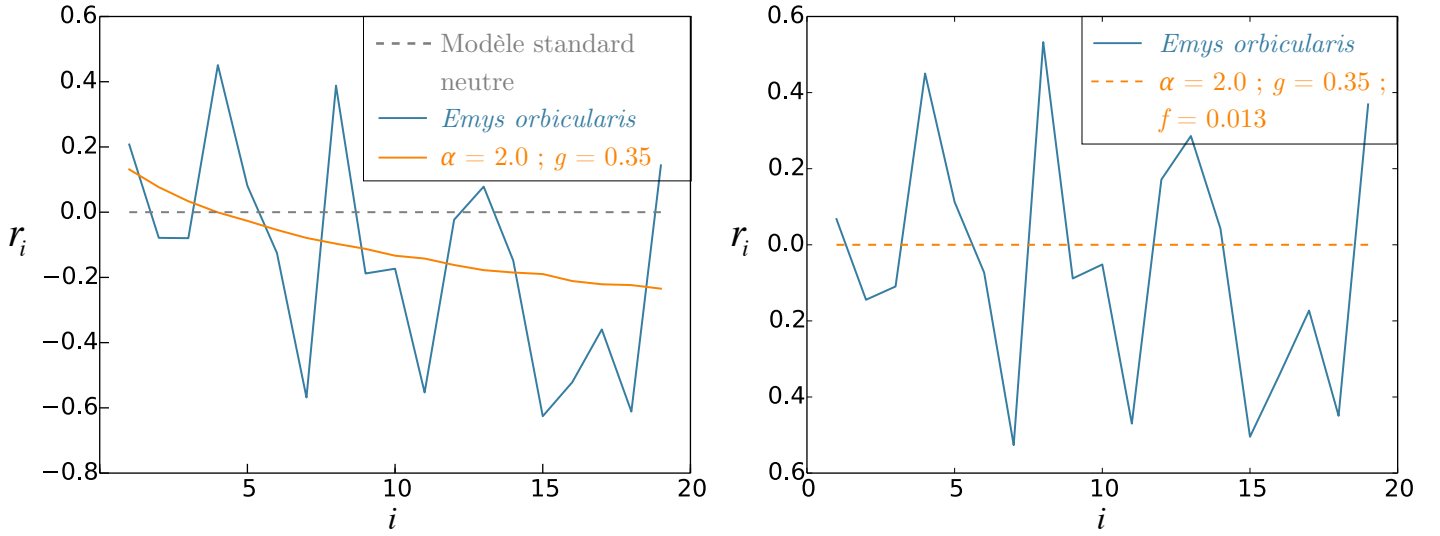


FIGURE 4.15 – À gauche, spectre résiduel corrigé d’*Emys orbicularis* et du modèle beta-coalescent optimisé avec  $\alpha = 2.0$  et croissance exponentielle à taux  $g = 0.35$  ( $d^2 = 0.061$ ) par rapport au modèle neutre. À droite, spectre résiduel d’*Emys orbicularis* par rapport au modèle beta-coalescent optimisé avec  $\alpha = 2.0$ ,  $g = 0.35$  et taux d’erreurs d’orientation  $f = 0.013$ .

population en décroissance, la remise à l’échelle des temps coalescents pour tenir compte de cette décroissance peut aboutir à de trop grandes valeurs, c’est-à-dire à des temps infinis. Ainsi, vu la forme de la surface de  $d^2$  en fonction des paramètres, il semble que l’on n’ait pas atteint le minimum, qui serait atteint pour des taux de décroissance exponentielle plus importants, que l’on ne peut pas simuler. Avec les valeurs de paramètres obtenues, l’ajustement aux données n’est pas très satisfaisant (Figure 4.16B) : une grande partie du signal d’excès de mutations à fortes fréquences n’a pas été expliqué par le modèle, et le spectre résiduel des données par rapport au modèle optimisé (à droite) présente donc encore un signal.

On obtient des résultats similaires pour *A. franciscana*, *C. brenneri*, *E. coli*, *H. scabiosa*, *L. granatensis*, *O. edulis*, *P. acuta* et *S. officinalis* (voir Figures 6.3 à 6.10 en Annexe).

### Comparaison des séquences codantes et non-codantes

Pour *A. thaliana*, *D. melanogaster* et *H. sapiens*, j’ai analysé séparément les données codantes et non-codantes. Les résultats de l’ajustement avec le modèle beta-coalescent avec démographie pour *D. melanogaster* et *H. sapiens* sont présentés dans la Figure 4.17.

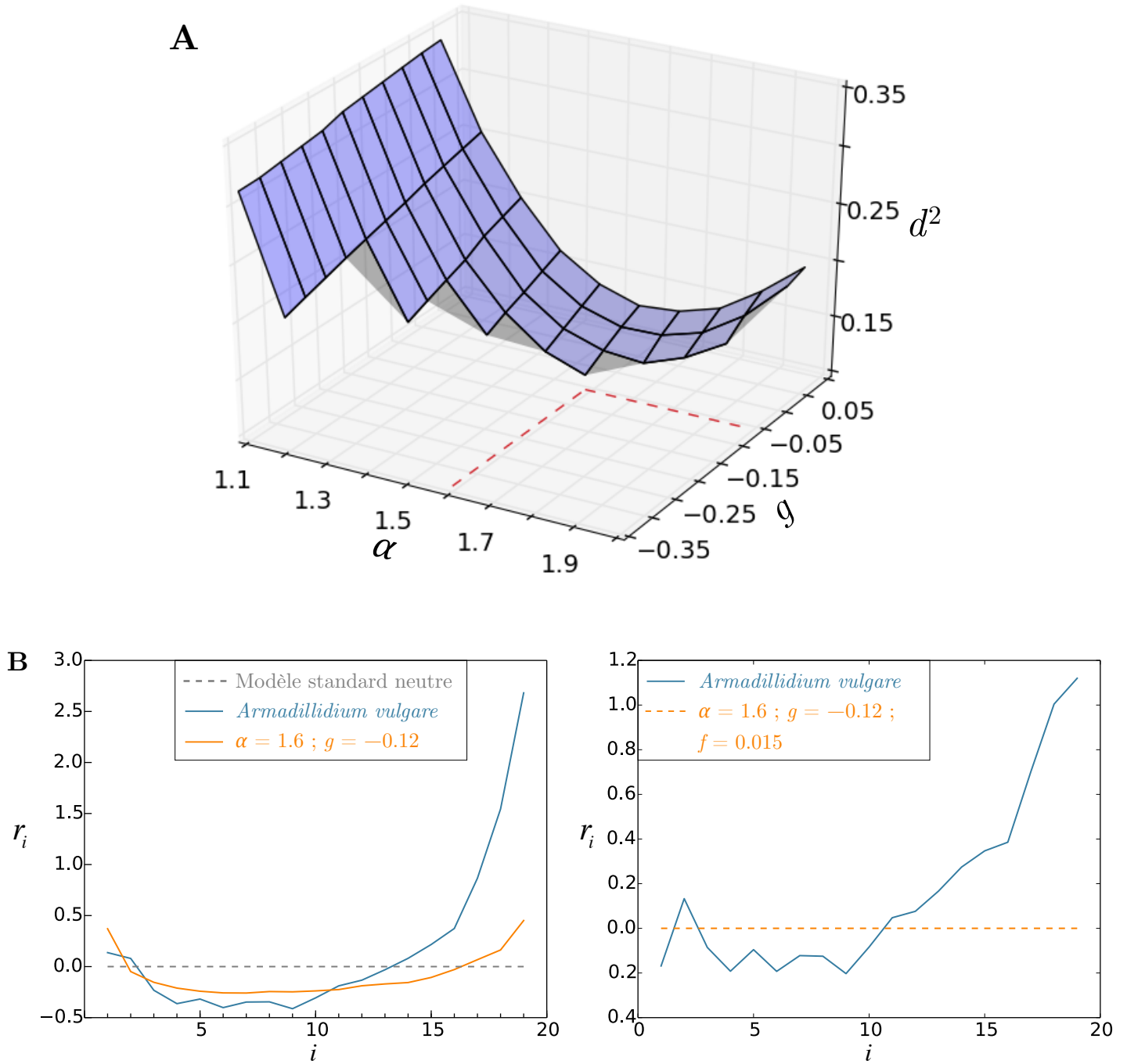
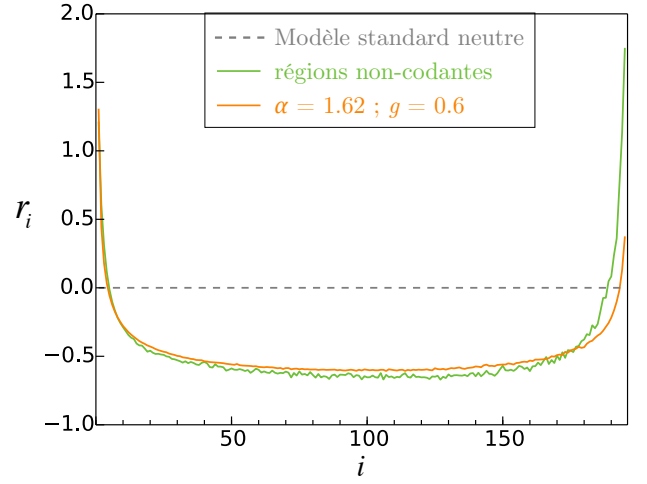
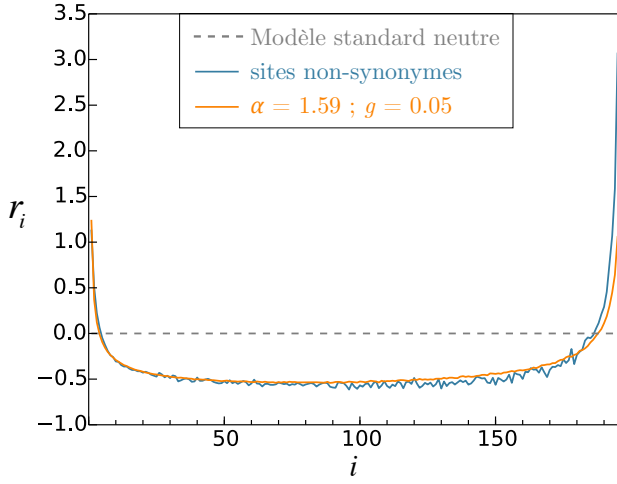


FIGURE 4.16 – A) Distance au carré entre le spectre corrigé d'*Armadillidium vulgare* et le spectre théorique en fonction des paramètres  $\alpha$  et  $g$ . Les lignes rouges indiquent les coordonnées du minimum,  $d^2 = 0.113$ . B) À gauche, spectre résiduel corrigé d'*Armadillidium vulgare* et du modèle beta-coalescent optimisé avec  $\alpha = 1.6$  et croissance exponentielle à taux  $g = -0.12$  par rapport au modèle standard neutre. À droite, spectre résiduel d'*Armadillidium vulgare* par rapport au modèle beta-coalescent optimisé avec  $\alpha = 1.6$ ,  $g = -0.12$  et taux d'erreurs d'orientation  $f = 0.015$ .

*Drosophila melanogaster*



*Homo sapiens* (Yoruba)

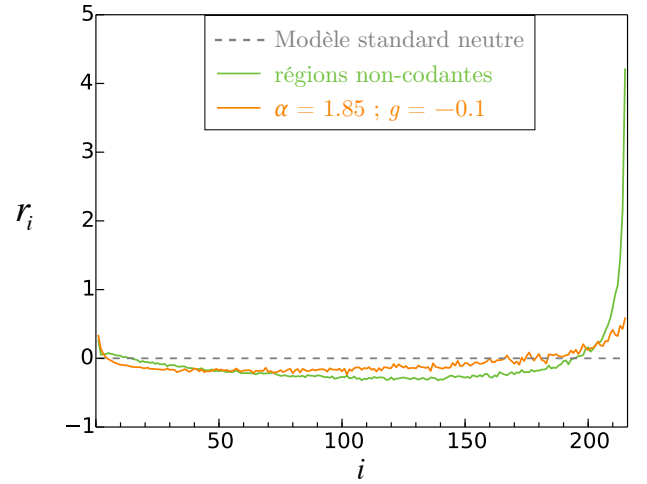
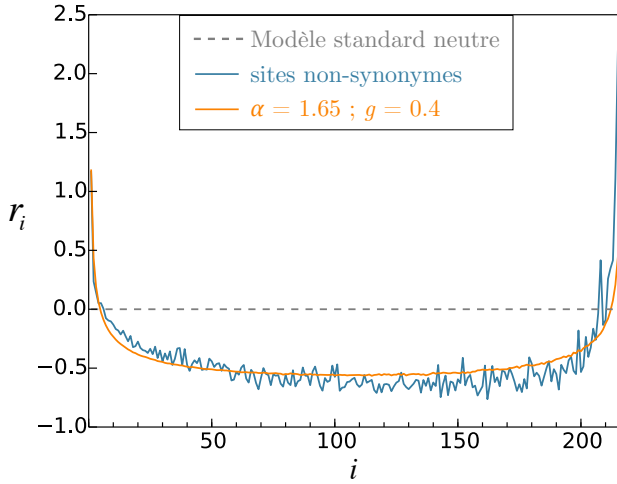


FIGURE 4.17 – Ajustement d'un modèle beta-coalescent avec démographie aux données codantes (sites non-synonymes) et non codantes de *Drosophila melanogaster* et *Homo sapiens*. Les spectres résiduels observés corrigés sont en bleu (codants) ou vert (non-codant). Les spectres résiduels prédits sous le modèle beta-coalescent (avec les valeurs de paramètres précisés en légende) sont en orange.

***Drosophila melanogaster*** L'excès de mutations à fortes fréquences est environ deux fois plus important dans les séquences codantes que dans les séquences non-codantes, tandis que l'excès de mutations à faibles fréquences est à peu près identique dans les deux cas. Pour ces deux spectres (nommés C pour codant et NC pour non codant), l'ajustement du modèle beta-coalescent avec démographie aboutit à deux estimations de  $\alpha$  similaires ( $\hat{\alpha}_C = 1.59$  et  $\hat{\alpha}_{NC} = 1.62$ ) mais à deux taux de croissance exponentielle différents d'un facteur 10 ( $\hat{g}_C = 0.05$  et  $\hat{g}_{NC} = 0.6$ ). Ce n'est pas ce à quoi l'on s'attend, du moins biologiquement : les variations démographiques affectent la généalogie de toutes les séquences, codantes et non-codantes. Le paramètre  $\alpha$  des coalescences multiples peut expliquer aussi des phénomènes neutres, comme la reproduction sweepstake (forte variance du succès reproductif), qui affecteront tout le génome, mais également des phénomènes sélectifs, qui eux n'affectent que les séquences codantes (et plus spécifiquement les sites sous sélection avec lesquels j'ai construit le spectre codant). On pourrait donc s'attendre à ce que l'estimation de  $g$  soit la même pour les deux spectres, et que ce soit  $\alpha$  qui diffère entre les séquences codantes (avec sélection) et non-codantes.

Dans cette optique, j'ai optimisé  $\alpha$  sur le spectre codant, avec  $g$  fixé à sa valeur estimée sur le spectre non-codant. Le spectre non-codant sert ainsi d'étalonnage : on y estime la valeur des paramètres  $\alpha$  et  $g$  pour les processus neutres. On fixe ensuite  $g$  qui devrait être le même pour les séquences codantes, et on ré-estime  $\alpha$  qui peut aussi expliquer des phénomènes sélectifs, et donc être différent pour les séquences codantes. On estime alors  $\alpha = 1.64$  : ainsi, quand on force  $g$  à être le même que pour les régions non-codantes, on trouve un  $\alpha$  plus élevé dans les régions codantes. Cependant, la différence n'est pas très marquée, et l'ajustement aux données est bien sûr un peu moins bon ( $d^2 = 0.012$ , alors qu'avec les deux paramètres libres on trouve  $d^2 = 0.007$ ).

***Homo sapiens*** Pour la population Yoruba, on a vu que l'excès de mutations à faibles fréquences est trois fois plus important dans les séquences codantes, et qu'à l'inverse l'excès de mutations à fortes fréquences est deux fois plus important dans les séquences codantes. L'ajustement du modèle beta-coalescent avec démographie aboutit donc à des résultats très différents pour ces deux jeux de données : pour les séquences codantes,  $\hat{\alpha}_C = 1.65$  et  $\hat{g}_C = 0.4$  tandis que pour les séquences non-codantes,  $\hat{\alpha}_{NC} = 1.85$  et  $\hat{g}_C = -0.1$ .

Pour les séquences non-codantes, on est dans la situation expliquée dans la section précédente, avec un taux de décroissance qu'on ne peut pas optimiser à cause du simulateur. Le  $g$  estimé n'était donc pas le meilleur, on ne peut pas tester la même procédure que pour *D. melanogaster*, en fixant le  $g$  estimé dans les séquences non-codantes pour

l'ajustement de  $\alpha$  dans les séquences codantes.

## 4.4 Le biais de conversion génique

Le biais de conversion génique (ou gBGC pour GC-biased gene conversion) est un processus associé à la recombinaison méiotique qui favorise les bases G:C par rapport aux bases A:T au moment de la réparation des mésappariements (Marais, 2003; Lesecque et al., 2013). Ce mécanisme tend à augmenter le taux de GC et le taux de substitutions A:T→G:C dans les régions à fort taux de recombinaison. Il est équivalent à de la sélection naturelle favorisant les allèles G:C, en augmentant leur fréquence et leur probabilité de fixation (Nagylaki, 1983). Il peut donc en mimer les effets, en particulier sur le spectre de fréquence.

Pour déterminer si les signatures de sélection que nous détectons sur les spectres observés (avec  $\alpha \neq 2$ ) peuvent être dues au gBGC, j'ai compté les types de mutations (A→T, A→G, A→C, etc...), et ce pour trois catégories de fréquences : moins de 10%, entre 10% et 90% et plus de 90%. Les résultats pour *A. thaliana* sont présentés dans la table 4.5.

Comme précédemment observé dans d'autres études (Katzman et al., 2011; Glémin et al., 2015), on constate une sur-représentation des catégories A→G et T→C pour les mutations présentes à plus de 90% dans l'échantillon (valeurs encadrées dans la table 4.5). On obtient des résultats semblables pour *D. melanogaster* et *H. sapiens*, population Yoruba (voir Tables 6.1 et 6.2 en Annexe). On peut donc en déduire qu'une partie de l'excès de mutations à fortes fréquences observé chez ces espèces est due au phénomène de conversion biaisée, qui mime de la sélection en faveur des allèles G:C (Galtier and Duret, 2007; Berglund et al., 2009; Ratnakumar et al., 2010; Kostka et al., 2012). Ratnakumar et al. (2010) ont estimé que jusqu'à 20% des signatures de sélection positive dans le génome humain pouvaient être expliquées par le gBGC.

Pour voir l'effet de ce mécanisme sur le spectre de fréquence, j'ai construit le spectre des mutations qui n'affectent pas le taux de GC, c'est-à-dire uniquement les mutations A↔T et C↔G, et ce pour les régions codantes, dans lesquelles on s'attend à trouver des signatures de sélection, et pour les régions non-codantes. Les résultats pour *D. melanogaster* sont présentés dans la Figure 4.18.

Avec la correction des erreurs d'orientation, pour les mutations qui n'affectent pas le taux de GC, l'excès de mutations à hautes fréquences disparaît complètement (Figure 4.18B). Les valeurs négatives sont dues à la correction : en effet, d'après l'équa-



TABLE 4.5 – Nombre de mutations d’une base (en ligne) vers une autre (en colonne) pour différentes catégories de fréquences alléliques chez *A. thaliana*. Les pourcentages indiqués entre parenthèses sont calculés par rapport au total de la catégorie de fréquence. Les deux valeurs encadrées sont les plus significatives dans un test du khi-deux.

(a) Fréquence inférieure à 10%				
	A	T	G	C
A		262 125 (8.8%)	341 335 (11.5%)	170 605 (5.7%)
T	262 196 (8.8%)		171 789 (5.8%)	341 347 (11.5%)
G	451 038 (15.2%)	153 061 (5.1%)		109 396 (3.7%)
C	152 225 (5.1%)	451 518 (15.2%)	110 074 (3.7%)	

(b) Fréquence comprise entre 10 et 90%				
	A	T	G	C
A		59 369 (9.3%)	83 073 (13.1%)	34 908 (5.5%)
T	58 970 (9.3%)		34 621 (5.4%)	83 002 (13.1%)
G	86 247 (13.6%)	33 844 (5.3%)		20 957 (3.3%)
C	33 808 (5.3%)	85 965 (13.5%)	20 916 (3.3%)	

(c) Fréquence supérieure à 90%				
	A	T	G	C
A		34 611 (8.0%)	87 416 (20.3%)	20 537 (4.8%)
T	34 436 (8.0%)		20 674 (4.8%)	87 218 (20.2%)
G	46 098 (10.7%)	16 867 (3.9%)		10 248 (2.4%)
C	16 643 (3.9%)	46 080 (10.7%)	10 364 (2.4%)	

tion 4.2, si  $\xi_i/\xi_{n-i} < \hat{f}/(1 - \hat{f})$ ,  $\xi_i^{corr}$  est négatif. On peut donc supposer que dans le cas de *D. melanogaster*, une grande partie de l’excès de mutations à hautes fréquences pourrait être dû au mécanisme de gBGC, puisque dans les mutations non affectées par ce mécanisme, la correction des erreurs d’orientation suffit à faire disparaître cet excès, et ce même dans les régions codantes.

Ce résultat se retrouve pour *H. sapiens*, population Yoruba (voir Figure 6.11 en Annexe). À l’inverse, chez *A. thaliana*, l’excès de mutations à hautes fréquences persiste même après correction (voir Figure 4.19).

Des méthodes plus complexes existent pour estimer l’intensité du gBGC à partir du spectre de fréquence. Glémin et al. (2015) ont développé une méthode qui tient compte de la démographie, des erreurs d’orientation, en particulier dues à l’hypermutabilité des

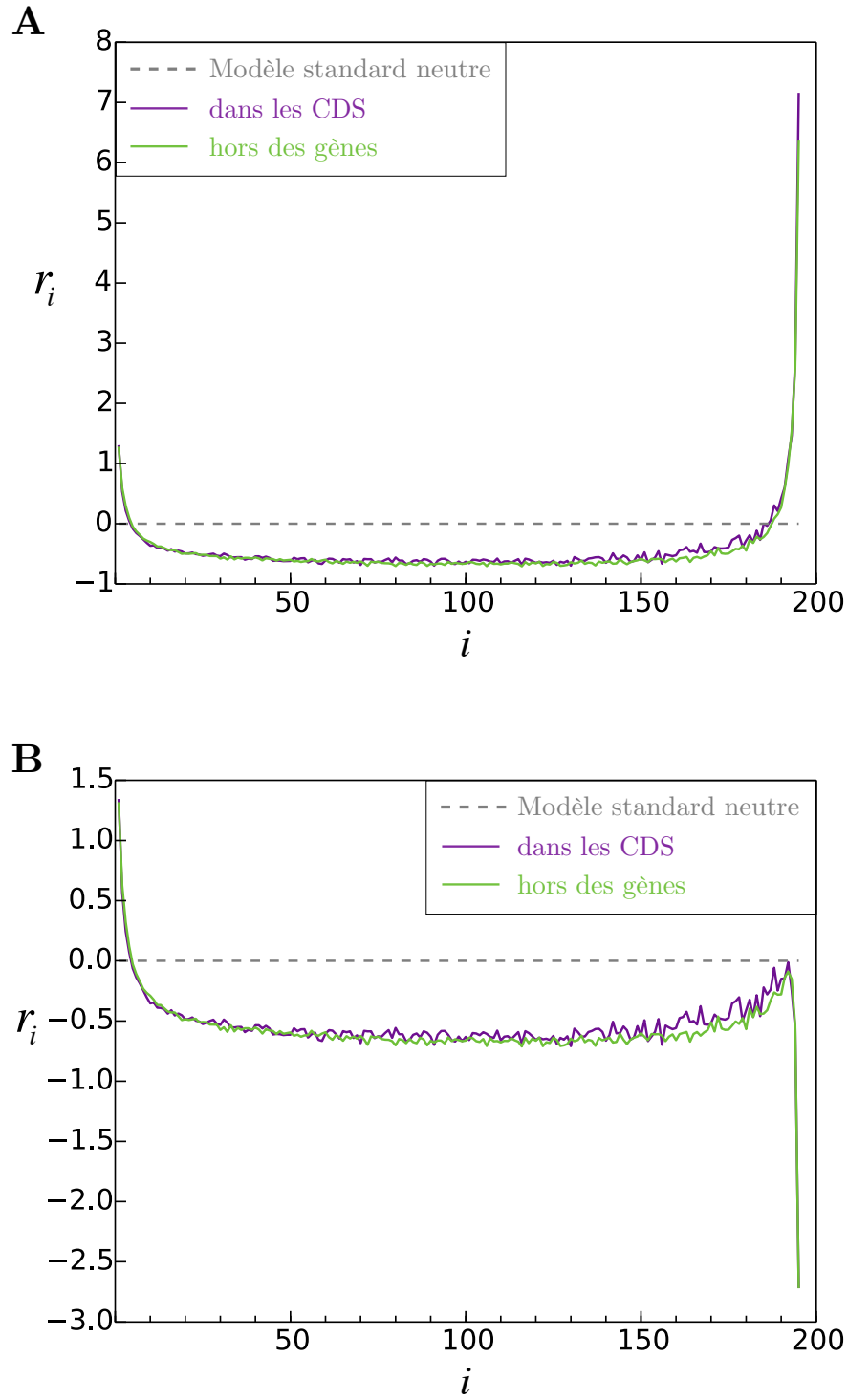


FIGURE 4.18 – Spectres de fréquence résiduels des mutations A↔T et C↔G chez *Drosophila melanogaster* pour les régions codantes (CDS, en violet) et non codantes (en vert). En haut (A), spectre non corrigé, et en bas (B), spectre corrigé. Pour les CDS,  $\hat{f}_{\text{JC}} = 2.2\%$ . Pour le non codant,  $\hat{f}_{\text{JC}} = 2.0\%$ .

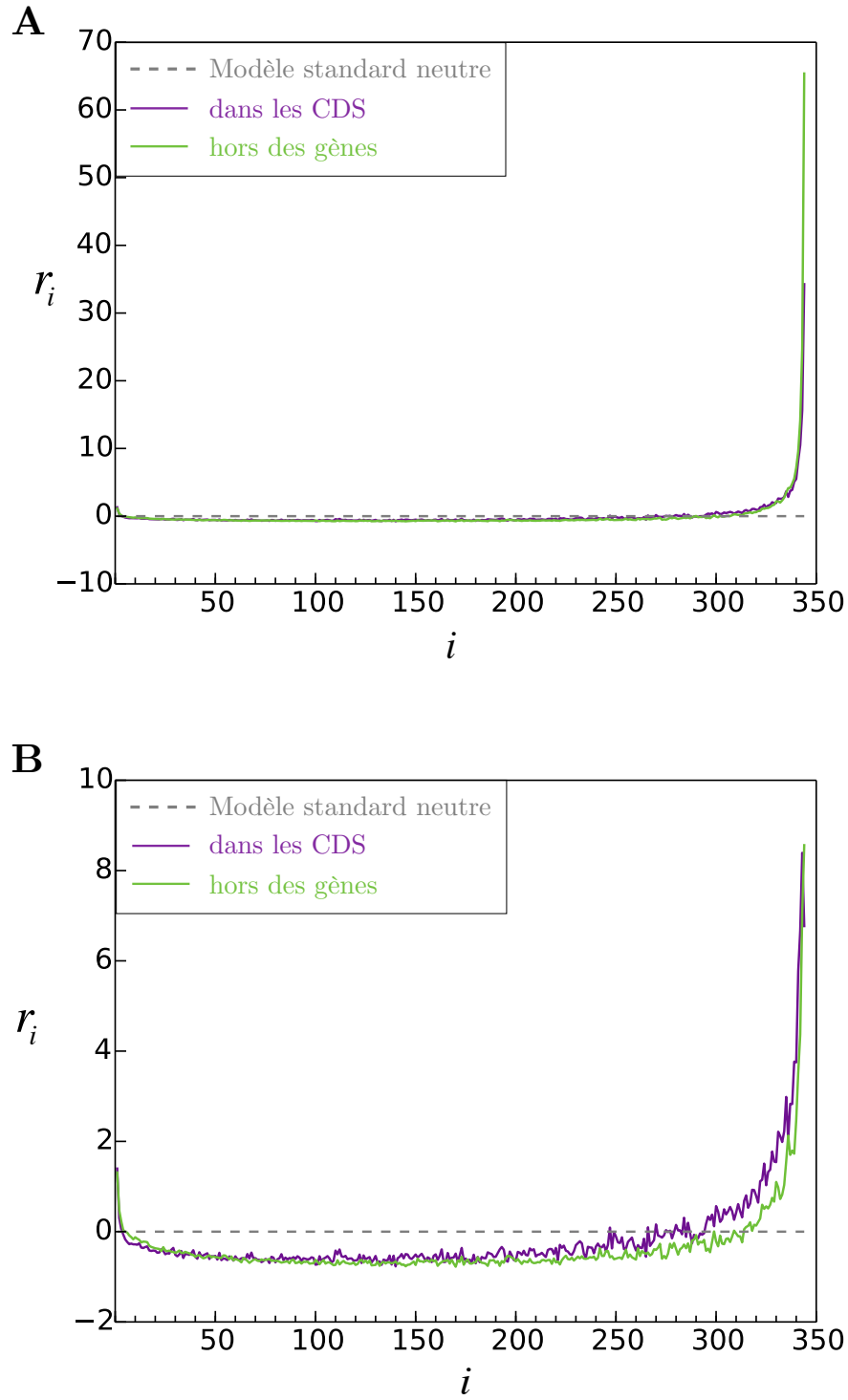


FIGURE 4.19 – Spectres de fréquence résiduels des mutations  $A \leftrightarrow T$  et  $C \leftrightarrow G$  chez *Arabidopsis thaliana* pour les régions codantes (CDS, en violet) et non codantes (en vert). En haut (A), spectre non corrigé, et en bas (B), spectre corrigé. Pour les CDS,  $\hat{f}_{JC} = 3.4\%$ . Pour le non codant,  $\hat{f}_{JC} = 7.2\%$ .

CpG, et de l'hétérogénéité du gBGC à l'échelle du génome. L'utilisation de telles méthodes sortait du cadre de ce travail exploratoire mais il nous semble intéressant de souligner l'existence de ce mécanisme, souvent ignoré, et qui peut fortement biaiser les inférences de sélection à partir du spectre de fréquence.

## 4.5 Discussion

### Intérêt des modèles à coalescences multiples pour l'analyse des données de diversité génétique

On a montré dans ce travail préliminaire qu'un modèle à coalescences multiples (ici un beta-coalescent) et démographie pouvait, avec seulement deux paramètres, améliorer nettement l'ajustement aux données observées par rapport au modèle standard neutre. Dans la majorité des spectres observés pour les jeux de données étudiés, le spectre de fréquence présente un excès de mutations à fortes fréquences, avec une forme en U (3 des 16 spectres) ou «en J» (uniquement excès de mutations à fortes fréquences, 11 des 16 spectres). Seule une espèce (*E. orbicularis*) présente un spectre qui semble compatible avec le modèle standard neutre.

Devant ces observations, il paraît raisonnable de se questionner sur la pertinence du modèle neutre pour l'étude de ces données. Avec ce travail, nous suggérons qu'un modèle simple (à 2 paramètres) pouvant modéliser des coalescences multiples et de la démographie, pourrait servir de modèle de référence à l'analyse de ces données. Ce modèle est en tout cas plus à même d'ajuster les données observées, même s'il reste beaucoup à faire pour en faire un outil aussi développé que le modèle standard neutre.

### Interprétation du paramètre $\alpha$

On a vu en introduction que les coalescences multiples pouvaient survenir dans des processus neutres (reproduction sweepstake) ou faisant intervenir de la sélection (balayages sélectifs). Si on veut promouvoir l'utilisation de ces modèles en génétique des populations pour l'analyse des données, il sera nécessaire de développer des outils qui permettront de distinguer ces mécanismes : il n'est pas suffisant de savoir qu'un modèle à coalescences multiples s'ajuste mieux aux données qu'un modèle coalescent standard, encore faut-il savoir pourquoi. Est-ce une signature de sélection ou d'un mode de reproduction particulièrement variable ? Pour répondre à cette question, j'ai comparé quand cela était possible les données codantes et non-codantes, pour essayer de démêler les effets des deux types de processus, ceux qui affectent le génome entier et ceux qui n'affectent que les zones

codantes. Cependant, on a vu également que des mécanismes moléculaires qui affectent tout le génome, comme le biais de conversion génique, pouvaient mimer l'effet de la sélection : ainsi, même dans les régions non-codantes, un  $\alpha$  différent de 2 peut être le reflet de différents mécanismes.

## 4.6 Perspectives

### Autres données disponibles

Les projets de séquençage se multipliant, il va être possible dans les années futures d'élargir ce type d'étude à un grand nombre d'espèces. Dans ce travail préliminaire, j'ai principalement étudié des organismes modèles (*H. sapiens*, *D. melanogaster*, *A. thaliana* et *E. coli*), pour lesquels les projets de séquençage sont déjà bien avancés voire achevés, et les données annexes (régions codantes, bases ancestrales) sont bien documentées, ainsi qu'un jeu de données d'espèces non-modèles.

Je liste ici d'autres données qui pourraient être analysées dans le cadre de ce projet. Certaines vont être prochainement disponibles, d'autres le sont déjà mais n'ont pas été traitées dans cette thèse faute d'informations disponibles sur les régions codantes ou la séquence ancestrale (alignement avec une espèce proche) :

- *Caenorhabditis elegans* : 152 génomes complets (diploïdes) sont disponibles sur le site du projet « *Caenorhabditis elegans* Natural diversity Resource » (Cook et al., 2017, <https://elegansvariation.org/data/>).
- Grands singes : dans le cadre du « Great Ape Genome Project », 13 bonobos (*Pan paniscus*), 27 gorilles (*Gorilla gorilla gorilla*), et 25 chimpanzés (*Pan troglodytes*) ont été séquencés (Prado-Martinez et al., 2013, <http://biologiaevolutiva.org/greatape/data.html>).
- *Saccharomyces cerevisiae* : le projet « The 1002 Yeast Genomes Project » est achevé, et les 1011 génomes séquencés seront bientôt disponibles (<http://1002genomes.u-strasbg.fr/>).
- *Solanum lycopersicum* (Tomate) : 54 génomes ont été séquencés dans le cadre du projet « The 100 Tomato Genome Sequencing Consortium » (Aflitos et al., 2014, <http://www.tomatogenome.net/>).
- *Taeniopygia guttata* (Diamant mandarin) : 20 individus ont été séquencés dans le cadre d'une étude sur les points chauds de recombinaison (Singhal et al., 2015, <http://www.ebi.ac.uk/ena/data/view/PRJEB10586>).

## Modélisation de la décroissance exponentielle

Une des principales limites de cette étude a été que dans un grand nombre d'espèces étudiées, le modèle qui semblait convenir le mieux était un beta-coalescent dans une population en décroissance exponentielle. Or le simulateur ne permettait pas de modéliser des décroissances importantes, et nous n'avons donc pas pu parcourir l'ensemble de l'espace des paramètres pour en trouver le minimum. Cette limite devrait pouvoir être contournée par la suite : pour atteindre moins rapidement des temps infinis au moment de la remise à l'échelle des temps coalescents, il faudrait changer l'échelle du temps au départ, en prenant une taille de population actuelle (arbitraire) plus petite.

## Autres modèles à coalescences multiples

On s'est limité ici à l'étude d'un modèle beta-coalescent, avec  $1 < \alpha \leq 2$ . De futures améliorations du simulateur pourront permettre de tester des valeurs de  $\alpha$  inférieures à 1, qui permettent de simuler des excès importants de mutations à faibles fréquences, que l'on observait peu dans les données traitées ici. Par la suite, la comparaison avec d'autres types de coalescents multiples, paramétrés différemment ( $\Psi$ -coalescent) ou autorisant les coalescences multiples simultanées ( $\Xi$ -coalescent), permettront d'évaluer plus généralement quels coalescents multiples permettent d'ajuster quels types de spectres de fréquence observés.

# Chapitre 5

## Conclusion générale et discussion

Dans cette thèse, j’ai cherché à comprendre comment le cadre théorique du modèle standard neutre était utilisé en évolution moléculaire, quelles pouvaient être ses limites et quelles alternatives existaient pour l’analyse de la diversité génétique des populations.

En prenant l’exemple de l’inférence démographique à partir du modèle standard neutre, j’ai ainsi mis en évidence certains biais liés à l’utilisation de ce cadre théorique. Certains de ces biais étaient connus mais méritaient d’être soulignés dans un exemple d’application à des données. C’est le cas de la recombinaison dans les analyses démographiques de données microbiennes (Chapitre 2) : même s’il est connu que la recombinaison modifie l’arbre reconstruit d’alignements de génomes complet, la méconnaissance des hypothèses du cadre théorique ou l’utilisation systématique de méthodes connues, sans vérification de ses hypothèses, aboutit à des inférences démographiques biaisées. De façon plus insidieuse, on a montré que l’utilisation de ClonalFrame, dans le but de produire des arbres « sans recombinaison », aggravait le biais dû à la recombinaison dans l’inférence avec le skyline plot.

Dans le Chapitre 3, j’ai soulevé la question de l’identifiabilité des histoires démographiques à partir du spectre de fréquence. Cette question avait déjà été abordée de façon théorique, mais bien que ces études théoriques soient quasiment systématiquement citées dans les études de démographie appliquées à des données, leurs conclusions n’avaient pas été vraiment confrontées à un cas réel. Nous avons ainsi mis en évidence que ce problème d’identifiabilité n’était pas uniquement théorique, mais qu’avec des données réelles on pouvait se retrouver dans une situation où l’on n’est pas capable, avec le spectre de fréquence, de distinguer une croissance linéaire d’une croissance exponentielle, ou même d’une croissance soudaine pourtant peu réaliste biologiquement.

La comparaison avec une méthode flexible a mis en lumière l’importance de la question

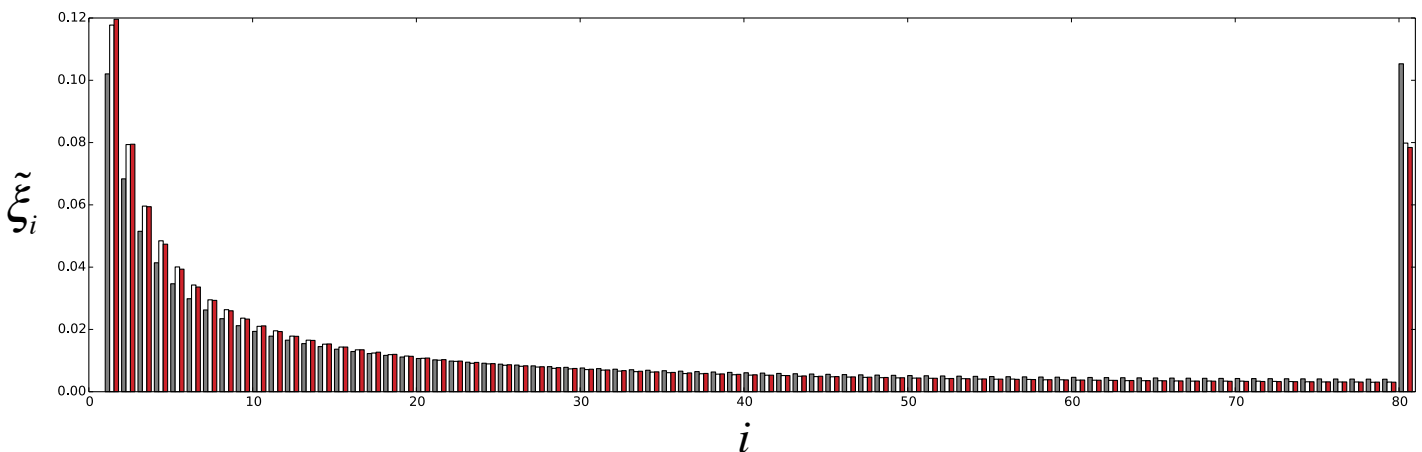


FIGURE 5.1 – Représentation en histogramme des spectres de fréquence pliés et normalisés du modèle standard neutre (gris), de la population Yoruba (blanc) et optimisé avec un modèle naissance-mort critique (rouge). Les valeurs pour  $i > 80$  sont regroupées.

de la complexité des modèles testés, via le choix du nombre de paramètres. On a ainsi mis en évidence un autre biais possible dans l'étude des spectres de fréquence, lié au bruit dû au nombre limité de « loci indépendants ».

Enfin, dans le but d'étudier des modèles alternatifs au modèle standard neutre, j'ai confronté les données à deux autres types de modèles, les processus naissance-mort (Chapitre 3) et les modèles à coalescences multiples (Chapitre 4).

Je reviens dans cette conclusion générale sur quelques uns des points transversaux aux trois parties de cette thèse. Je finirai par quelques réflexions plus générales sur les implications futures de ce travail.

## Représentation des spectres de fréquence

La représentation communément utilisée du spectre de fréquence est l'histogramme. Lorsque l'on souhaite comparer plusieurs spectres de fréquence, on les juxtapose, aboutissant parfois à des résultats peu lisibles et difficilement interprétables. De plus, du fait de la forme attendue du spectre en  $1/x$ , les valeurs pour les grandes fréquences sont faibles. Elles sont donc souvent omises, et seules les valeurs des premières cases du spectre, c'est-à-dire pour de petites fréquences, sont représentées, ou parfois les dernières cases du spectre sont groupées. Par exemple, la Figure 5.1 présente les mêmes résultats que la Figure 3B du Chapitre 3 mais sous forme d'histogramme.

Même en groupant les dernières cases du spectre et en ne représentant qu'un seul des modèles testés (*Birth-Death*), l'analyse de la figure n'est pas évidente. La représentation



graphique des résultats est une partie importante du travail : si celle des spectres de fréquence était plus facile à analyser, on aurait peut-être plus facilement tendance à vérifier l'ajustement au modèle, ce qui n'a pas été fait par exemple dans l'article présentant la méthode stairway plot (Liu and Fu, 2015).

Dans cette thèse, j'ai utilisé une représentation du spectre « transformée », que l'on a formalisée plus généralement dans le Chapitre 4 sous le nom de spectre résiduel. Cette représentation vise à représenter les spectres, observés ou prédits, par rapport à un attendu théorique. Cette remise à l'échelle permet de mieux visualiser les différences entre l'attendu et l'observé, par rapport à l'histogramme qui rend cette comparaison difficile.

De plus, cette représentation graphique permet, dans un processus d'optimisation, de visualiser l'amélioration de l'ajustement au fur et à mesure de l'optimisation. Dans le Chapitre 3, on avait suggéré que le stairway plot pourrait bénéficier d'une visualisation de l'ajustement du modèle aux données, et ce au fur et à mesure de l'optimisation. Grâce à cette représentation, on pourrait voir à chaque étape de l'optimisation dans quelle mesure on a amélioré l'ajustement aux données, s'il reste un signal à expliquer ou non, et ainsi s'arrêter à un nombre de paramètres raisonnable qui pourrait permettre d'éviter le biais dû au bruit. Dans le Chapitre 4, on a ainsi pu visualiser, avant et après optimisation, le spectre résiduel par rapport à un modèle sans ou avec coalescences multiples et démographie. Dans certains cas, les erreurs d'orientation, les coalescences multiples et la démographie semblaient expliquer tout le signal présent dans les données initiales. Dans d'autres cas, plus nombreux, le spectre résiduel final montrait que tout le signal n'avait pas été expliqué.

Il est fort probable qu'à l'avenir, avec la multiplication des projets de séquençage, le spectre de fréquence devienne une statistique résumée très utilisée pour l'inférence en génétique des populations (éventuellement combinée à d'autres statistiques, voir ci-dessous). Le fait d'adopter une représentation graphique plus facilement analysable, et adaptée aux grandes tailles d'échantillon (plus  $n$  est grand, moins les histogrammes tels que celui représenté dans la Figure 5.1 deviennent lisibles) pourrait donc se révéler primordial.

## **Diversification des modèles en génétique des populations**

Dans le Chapitre 3 on a montré le potentiel d'un modèle basé sur un processus naissance-mort critique pour l'analyse de données de diversité génétique. En effet, même si le modèle n'était pas meilleur que les autres, puisqu'ils ajustaient tous les données de façon satisfaisante, c'est le seul pour lequel on dispose d'une formule analytique pour le spectre de fréquence, et donc pour lequel l'ajustement du paramètre aux données est quasi-instantané. De plus, ces modèles présentent l'avantage d'une grande flexibilité puisque la

démographie n'est pas fixée : la taille de population peut varier aléatoirement entre le temps de fondation et le temps présent. Ainsi, on peut analyser le spectre de fréquence observé sans idée préliminaire sur la démographie de la population.

Sa flexibilité et sa caractérisation mathématique simple en font un modèle potentiellement très puissant pour le développement d'un nouveau modèle nul qui tiendrait compte de la démographie par exemple. Un des inconvénients, qui découle de sa flexibilité, et qu'une fois qu'on a optimisé le modèle, par exemple le temps de fondation comme dans le Chapitre 3, on ne connaît pas pour autant la démographie de la population. Dans ce chapitre nous avons simulé la trajectoire de fixation d'un nouvel allèle dans la population pour approximer la démographie, mais il faudrait préciser ces méthodes afin qu'on puisse avoir accès à la démographie sous-jacente d'un processus naissance-mort donné.

Nous avons également confronté à un ensemble de données plus large les modèles à coalescences multiples (un beta-coalescent plus précisément). Cette étude préliminaire a mis en évidence que dans la majorité des cas, le spectre de fréquence de ces populations n'était pas en adéquation avec la prédiction du modèle standard neutre. L'ajout de deux paramètres (coalescences multiples et démographie) permet d'améliorer l'ajustement aux données. Sans avoir l'ambition d'expliquer entièrement les données avec deux paramètres, l'idée est plutôt de donner une première approche vers un nouveau modèle de référence, légèrement plus complexe mais bien plus en adéquation avec ce qui semble être une distribution répandue des fréquences alléliques (excès de mutations à hautes fréquences par rapport à la prédiction du modèle standard neutre, forme « en J »).

## **Tenir compte de l'information de liaison**

On a vu que grâce à la recombinaison, l'information contenue dans le spectre de fréquence était un résumé, une moyenne des informations contenues dans tous les loci indépendants qui composent un génome recombinant. Depuis l'essor du séquençage, on a accès avec les génomes complets à un grand nombre de loci indépendants, ce qui n'était pas le cas quand les régions séquencées étaient beaucoup plus réduites. On voit sur la Figure 5.2 l'effet de l'augmentation du nombre de loci sur le spectre de fréquence d'une population en croissance linéaire.

Quand on simule le spectre de fréquence d'un seul locus, on a accès à l'information d'un seul arbre, ce qui explique que certaines cases du spectres soient vides (voir Figure 1.6 en Introduction). Cela illustre le peu d'information dont on disposait avant l'ère du séquençage massif pour faire des inférences à partir du spectre de fréquence, et donc la remise en cause nécessaire maintenant que l'on dispose d'une grande quantité d'information.

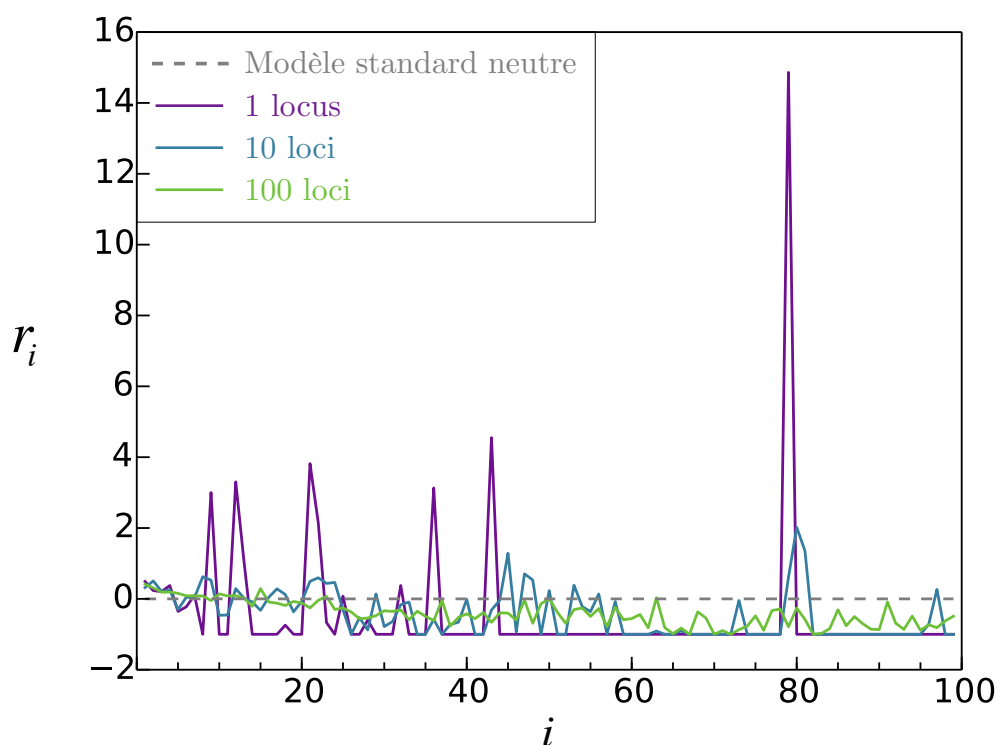


FIGURE 5.2 – Spectres résiduels d’un échantillon de  $n = 100$  individus d’une population en croissance linéaire (temps de fondation  $\tau = 1$ ) par rapport au modèle neutre. Les spectres sont simulés avec 1, 10 ou 100 loci indépendants.

Dans les simulations, on choisit un nombre de loci, pour lesquels on simule des histoires indépendantes. Dans un génome réel, il est moins évident de définir ce que seraient ces loci indépendants : à part les chromosomes qui évoluent indépendamment, au sein d’un chromosome on ne peut pas pointer un endroit précis de part et d’autre duquel les loci seraient indépendants. Chaque locus est lié à ses voisins, qui sont eux mêmes liés à leurs voisins. Cependant, à partir d’une certaine distance, les loci évoluent de façon indépendante. Grâce au bruit observé sur le spectre des données, on pourrait estimer ce nombre de loci indépendants qui pourrait nous renseigner sur l’étendue de la liaison au sein du génome étudié.

Maintenant que l’on dispose de génomes entiers séquencés, on a également accès à l’information de liaison entre les sites, qui n’est pas exploitée quand on utilise le spectre de fréquence. L’accès à l’information de liaison dans des génomes diploïdes nécessite que les données soient phasées, c’est-à-dire que l’on sait quel allèle est porté par quelle copie du chromosome, ce qui peut introduire des biais. Récemment, Boitard et al. (2016) ont montré que dans le cadre d’une inférence ABC, la combinaison du spectre de fréquence

avec l'information de liaison était efficace pour inférer la démographie à partir d'un grand échantillon, sans avoir besoin de l'information d'orientation ou de phasage des mutations. La combinaison de ces deux statistiques permet d'inférer à la fois la démographie récente et ancienne.

## **Effet de la structuration de population**

Dans cette thèse, on a cherché à expliquer des données observées de diversité génétique essentiellement par des processus démographiques. Un des aspects majeurs dont nous n'avons pas tenu compte est la structuration de population. Il a été montré que la structuration était un paramètre confondant dans les analyses démographiques de diversité des populations (Mazet et al., 2016) : si les populations sont structurées, les individus ne coalescent pas à la même vitesse selon qu'ils font partie ou non de la même sous-population. Ainsi, négliger l'existence de ces sous-populations biaise l'analyse des taux de coalescences inférés.

Nous avons en quelque sorte abordé le problème de la structuration dans le Chapitre 2 en étudiant les effets des différents biais d'échantillonnages, ce qui soulevait la question de la définition d'une population pour les espèces bactériennes.

Dans le Chapitre 4, le problème a été contourné puisque nous n'avons pas analysé les données de *Culex pipiens* dont le spectre de fréquence présentait un signal clair de structuration, pour l'analyse duquel notre modèle beta-coalescent avec démographie n'était pas adapté. On pourrait ajouter un ou plusieurs paramètres pour tenir compte de la structuration, ce qui nous renvoie à la question de la complexité des modèles abordée dans le Chapitre 3. Pour ce travail on aurait également pu ajouter à nos modèles de la structuration, mais l'objectif était de montrer qu'en revenant à des modèles simples, ici décrits par un unique paramètre, on pouvait très bien expliquer les données observées.

Les modèles à coalescences multiples peuvent s'adapter à l'étude de populations structurées, comme cela est fait actuellement avec le coalescent de Kingman. L'idée d'un modèle de référence basé sur des modèles à coalescences multiples est ainsi compatible avec, dans un deuxième temps, l'ajout de paramètres de structuration quand cela semble indiqué pour la population étudiée.

## **Modèle « standard » : depuis quand, jusqu'à quand ?**

Dans l'introduction, j'ai rappelé le contexte historique de l'émergence de la théorie neutraliste, qui a remplacé le pan-sélectionnisme qui était majoritaire dans les années 1960.

Ce remplacement s'est fait relativement rapidement : la théorie a été exposée en 1968, et une quinzaine d'année plus tard, lorsque Kimura publie son ouvrage *The neutral theory of molecular evolution*, elle est déjà largement acceptée. John H. Gillespie écrit dans sa critique du livre, publiée la même année dans *Science*, que « Cette théorie est aujourd'hui invoquée aussi systématiquement que l'était la sélection il y a quelques années. » Dans le chapitre 2 de cet ouvrage, Kimura déclare à propos de la théorie pan-sélectionniste :

« Avec le recul, je pense que c'est une caractéristique curieuse de la nature humaine que si une certaine doctrine est constamment défendue par une majorité, approuvée par les meilleurs experts dans leurs livres, et enseignée aux étudiants, alors une croyance est construite graduellement dans les esprits, devenant finalement le principe directeur et la base du jugement de valeur. »

C'est une observation pertinente, que l'on peut maintenant appliquer à la théorie neutraliste qui est véritablement devenue « le principe directeur et la base du jugement de valeur », à tel point qu'on ne s'interroge plus guère sur certaines de ses hypothèses et des conséquences qu'elles peuvent avoir.

Aujourd'hui le modèle neutre est utilisé comme modèle standard dans la majorité des études de génétique des populations, et plus largement d'évolution moléculaire. Dans la majorité des cas il est pris comme hypothèse nulle, que l'on cherche à rejeter pour montrer la présence de sélection, de structuration de population, ou de démographie par exemple. Plusieurs problèmes méthodologiques se posent : d'une part, de par sa grande variabilité, le modèle standard neutre est difficile à rejeter statistiquement. Cette « robustesse » est peut-être une des raisons de sa popularité, mais ne pas rejeter l'hypothèse nulle ne veut pas dire que l'hypothèse nulle est vraie : il faut donc encourager à la prudence face à ces glissements méthodologiques.

Dans cette même idée de prudence, le Chapitre 3 est une bonne illustration des erreurs que l'on peut commettre en supposant que puisqu'un modèle ajuste bien les données, il est vrai. Dans ce cas, 5 modèles de démographies ajustaient tous aussi bien les données, or ils ne peuvent pas être tous vrais, il est même certain qu'ils sont tous faux. Mais bien souvent, on se contente de tester un modèle démographique, et de le considérer vrai s'il est meilleur que le modèle standard pour ajuster les données.

En introduction j'avais parlé des incohérences qui émergeaient dans le calcul de la taille efficace  $N_e$ . Il est intéressant de voir l'importance qu'a pris ce paramètre en génétique des populations : on lui attribue aujourd'hui souvent une réalité propre alors que ce n'est qu'une taille fictive, la taille que devrait avoir la population pour se comporter comme une population idéale de Wright-Fisher. On s'attache donc le plus souvent à rejeter le

modèle neutre, tout en considérant comme une vérité ce paramètre  $N_e$  qui y est pourtant lié. Ce paramètre est aussi à l'origine, ou en tout cas en partie responsable, de certaines erreurs, comme celles mises en évidence dans le Chapitre 2 ou dans l'article de Mazet et al. (2016) : des méthodes qui se basent sur l'inverse des taux de coalescences pour inférer la démographie sont biaisées par d'autres forces qui influent ces taux de coalescences. En faisant le raccourci que  $N_e$  est l'inverse du taux de coalescence, et que  $N_e$  est une taille de population, ces méthodes analysent le taux de coalescence purement en termes de démographie. Mazet et al. (2016) proposent une nouvelle statistique, le taux de coalescence instantanée inverse, qui n'est équivalent à une taille de population que dans les modèles panmictiques. Cette démarche va dans le sens de prendre de la distance vis-à-vis de la taille efficace.

On semble mettre en évidence dans le Chapitre 4 que le modèle standard neutre n'est finalement (quasiment) jamais en adéquation avec les données, qu'elles soient de vertébrés, d'invertébrés, de végétaux, ou encore de bactéries. De nombreuses études ont montré ses limites, des modèles alternatifs existent et commencent à être utilisés. On peut alors se demander jusqu'à quand ce modèle va rester le modèle standard, de référence ? Il semblerait que pour l'instant, le seuil critique n'ait pas encore été atteint pour le remettre en cause, que ce soit en termes de quantité de « preuves » qui le remettent en cause qu'en termes de développement des modèles qui le remplaceront. En effet, étant donnée sa popularité actuelle, il ne pourra être remplacé que si les modèles alternatifs développés à sa place deviennent aussi performants et faciles d'utilisation et d'interprétation.

Il faut souligner que son statut est différent de celui que pouvait avoir le pan-sélectionnisme dans les années 1960. À l'époque, c'étaient vraiment les bases biologiques de la théorie qui étaient largement admises : on pensait que la sélection était la cause de la diversité observée. Aujourd'hui, plutôt que sur ses bases biologiques, je pense que la domination de la théorie neutraliste repose plutôt sur ses bases méthodologiques : le modèle de Wright-Fisher et le coalescent de Kingman sont tellement profondément ancrés dans notre vision de l'évolution moléculaire que même lorsque l'on pense rejeter le modèle neutre, on continue à utiliser certaines propriétés qui en découlent comme la taille de population efficace. La solution pour prendre du recul sur la théorie neutraliste viendra donc peut-être du développement de nouveaux modèles, plus souples, pouvant prendre en compte des forces évolutives variées, et j'espère que ce travail donne quelques pistes de réflexion en ce sens.



# Chapitre 6

## Annexes

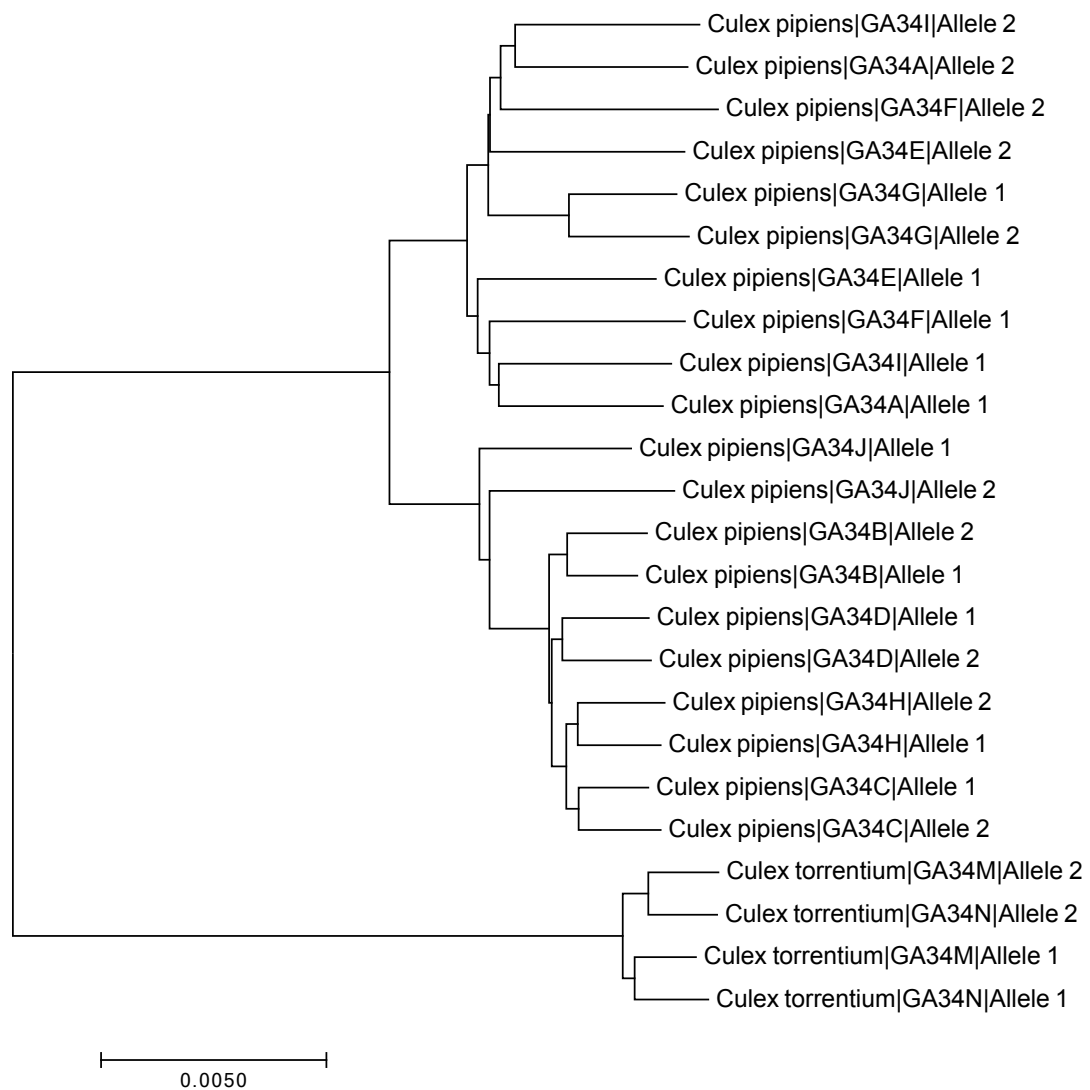


FIGURE 6.1 – Arbre Neighbour-joining des échantillons de *Culex pipiens* ( $2n=20$ ) et de son espèce sœur *Culex torrentium* ( $2n=4$ ) (obtenu avec MEGA 7, Kumar et al., 2016).



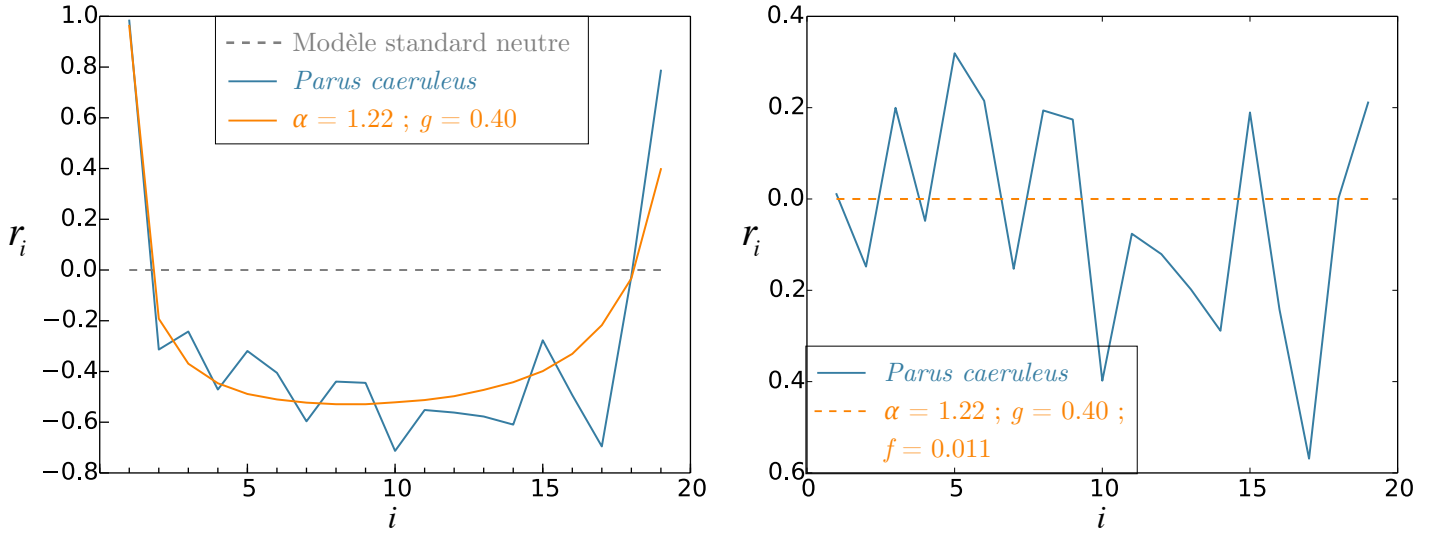


FIGURE 6.2 – À gauche, spectre résiduel corrigé de *Parus caeruleus* et du modèle beta-coalescent optimisé avec  $\alpha = 1.22$  et croissance exponentielle à taux  $g = 0.40$  ( $d^2 = 0.022$ ) par rapport au modèle neutre. À droite, spectre résiduel de *Parus caeruleus* par rapport au modèle beta-coalescent optimisé avec  $\alpha = 1.22$ ,  $g = 0.40$  et taux d'erreurs d'orientation  $f = 0.011$ .

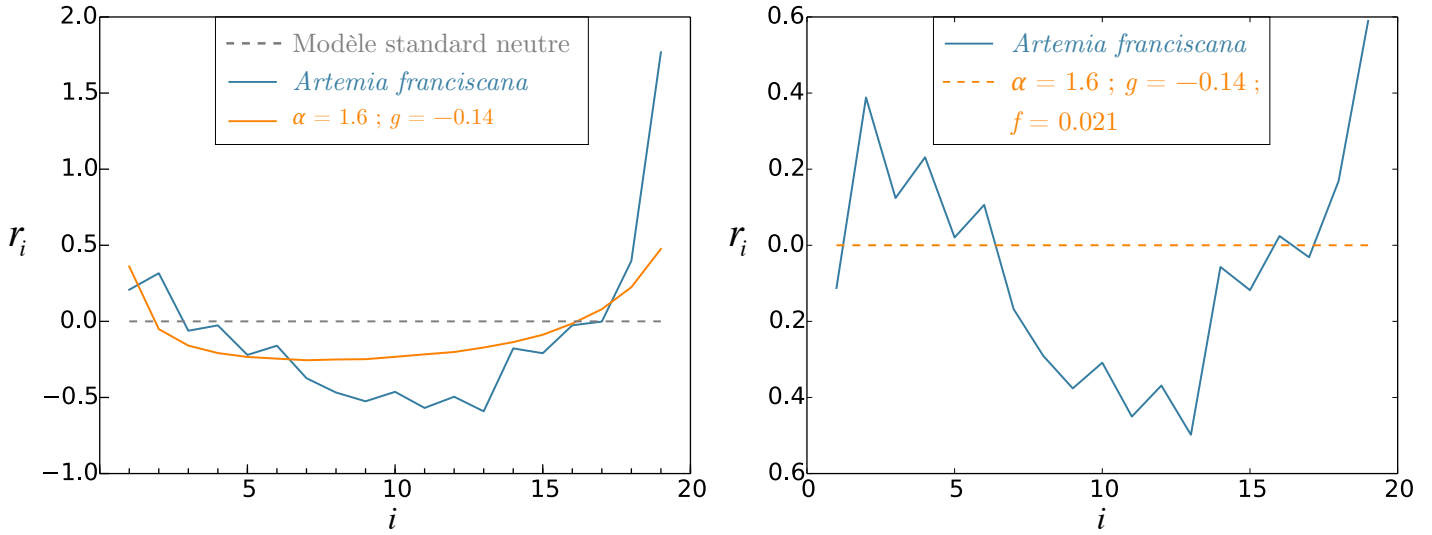


FIGURE 6.3 – À gauche, spectre résiduel corrigé d'*Artemia franciscana* et du modèle beta-coalescent optimisé avec  $\alpha = 1.6$  et croissance exponentielle à taux  $g = -0.14$  ( $d^2 = 0.066$ ) par rapport au modèle neutre. À droite, spectre résiduel d'*Artemia franciscana* par rapport au modèle beta-coalescent optimisé avec  $\alpha = 1.6$ ,  $g = -0.14$  et taux d'erreurs d'orientation  $f = 0.021$ .

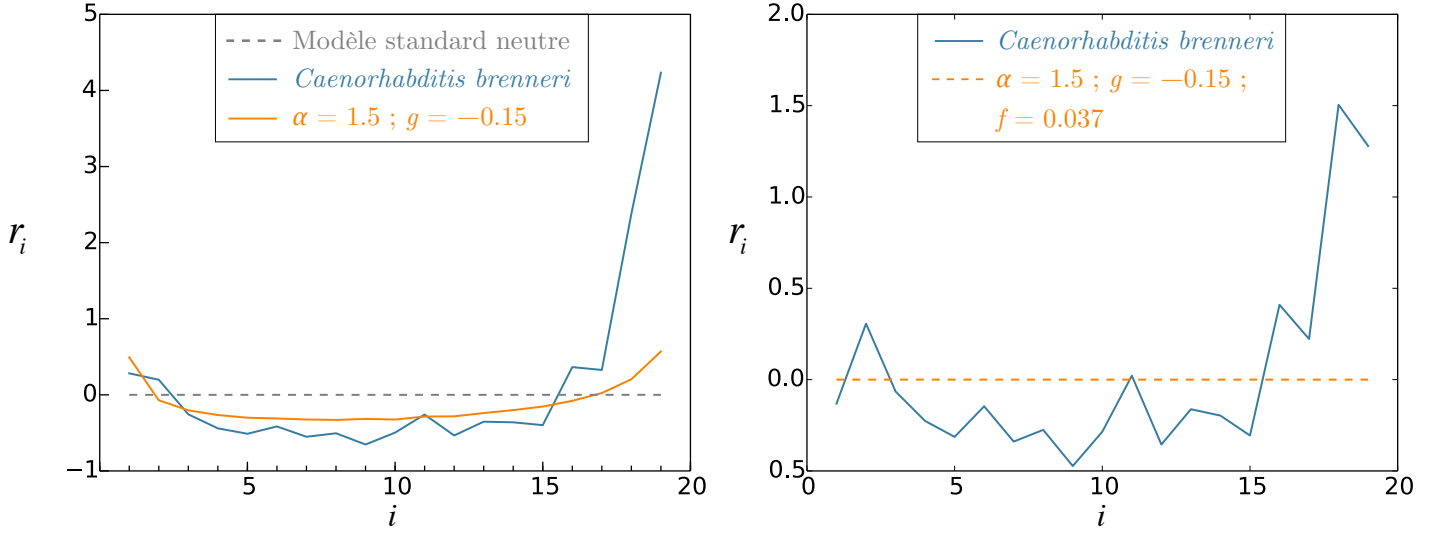


FIGURE 6.4 – À gauche, spectre résiduel corrigé de *Caenorhabditis brenneri* et du modèle beta-coalescent optimisé avec  $\alpha = 1.5$  et croissance exponentielle à taux  $g = -0.15$  ( $d^2 = 0.237$ ) par rapport au modèle neutre. À droite, spectre résiduel de *Caenorhabditis brenneri* par rapport au modèle beta-coalescent optimisé avec  $\alpha = 1.5$ ,  $g = -0.15$  et taux d'erreurs d'orientation  $f = 0.037$ .

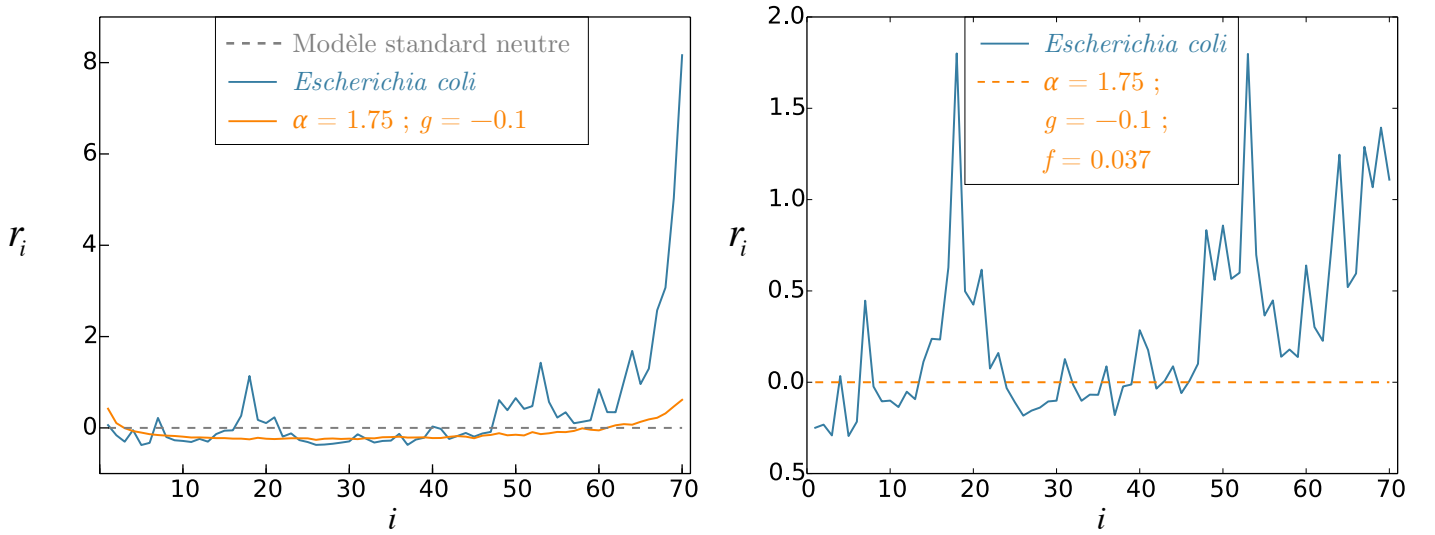


FIGURE 6.5 – À gauche, spectre résiduel corrigé d'*Escherichia coli* et du modèle beta-coalescent optimisé avec  $\alpha = 1.75$  et croissance exponentielle à taux  $g = -0.1$  ( $d^2 = 0.300$ ) par rapport au modèle neutre. À droite, spectre résiduel d'*Escherichia coli* par rapport au modèle beta-coalescent optimisé avec  $\alpha = 1.75$ ,  $g = -0.1$  et taux d'erreurs d'orientation  $f = 0.037$ .

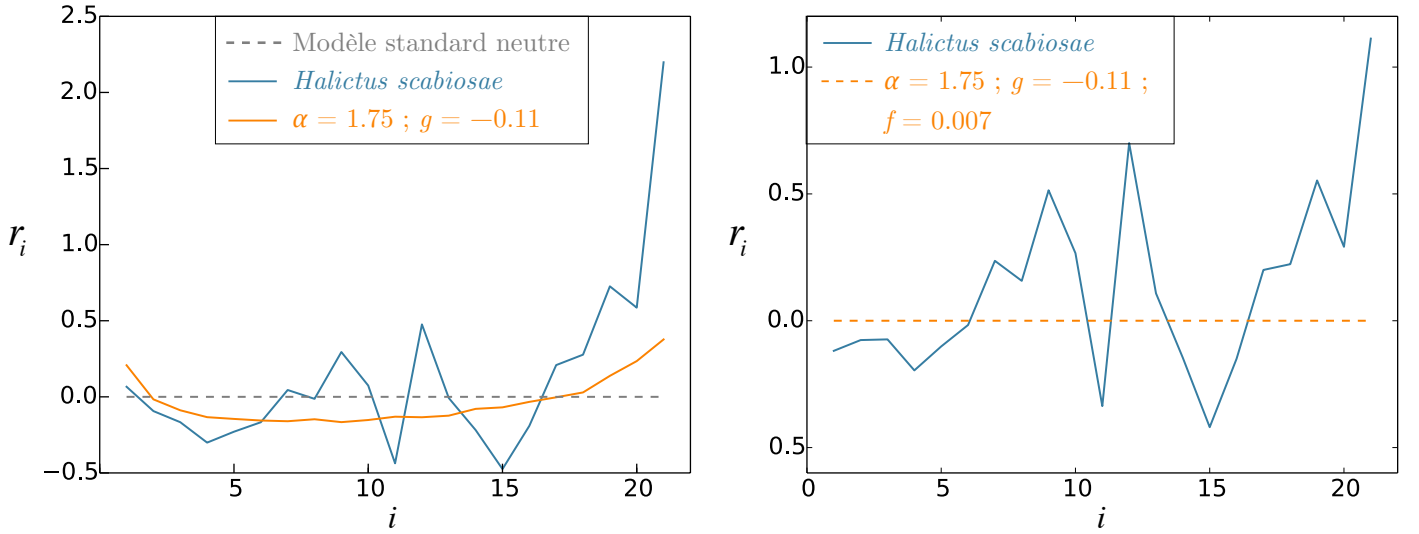


FIGURE 6.6 – À gauche, spectre résiduel corrigé d'*Halictus scabiosae* et du modèle beta-coalescent optimisé avec  $\alpha = 1.75$  et croissance exponentielle à taux  $g = -0.11$  ( $d^2 = 0.076$ ) par rapport au modèle neutre. À droite, spectre résiduel d'*Halictus scabiosae* par rapport au modèle beta-coalescent optimisé avec  $\alpha = 1.75$ ,  $g = -0.16$  et taux d'erreurs d'orientation  $f = 0.007$ .

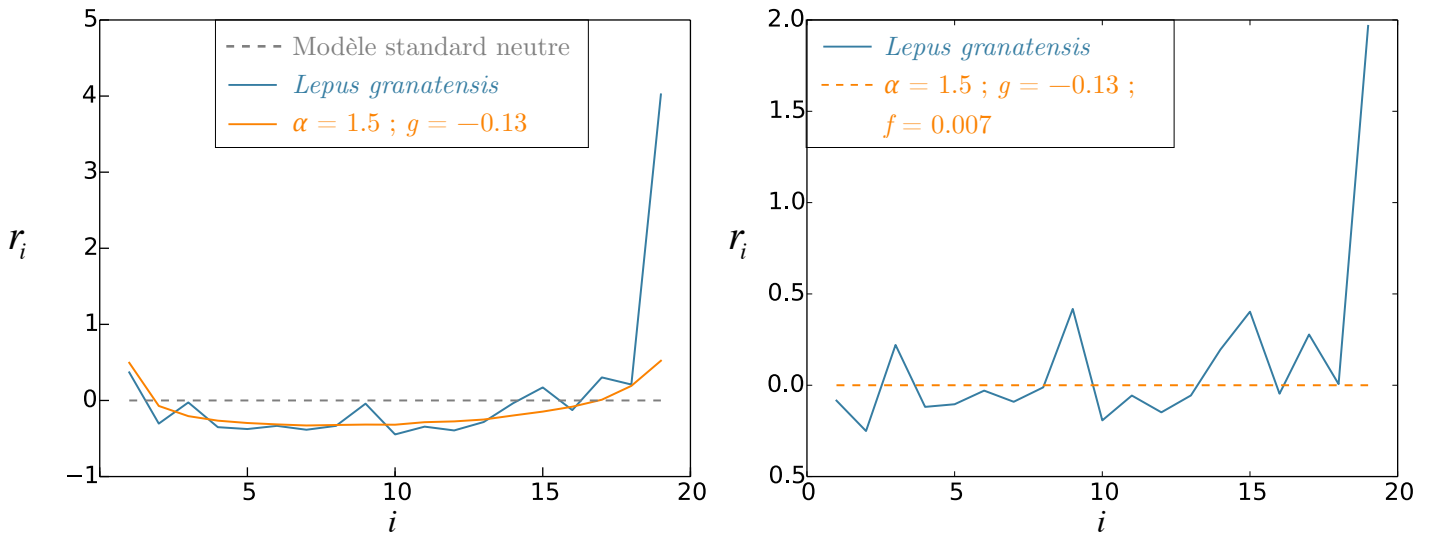


FIGURE 6.7 – À gauche, spectre résiduel corrigé de *Lepus granatensis* et du modèle beta-coalescent optimisé avec  $\alpha = 1.5$  et croissance exponentielle à taux  $g = -0.13$  ( $d^2 = 0.145$ ) par rapport au modèle neutre. À droite, spectre résiduel de *Lepus granatensis* par rapport au modèle beta-coalescent optimisé avec  $\alpha = 1.5$ ,  $g = -0.13$  et taux d'erreurs d'orientation  $f = 0.007$ .

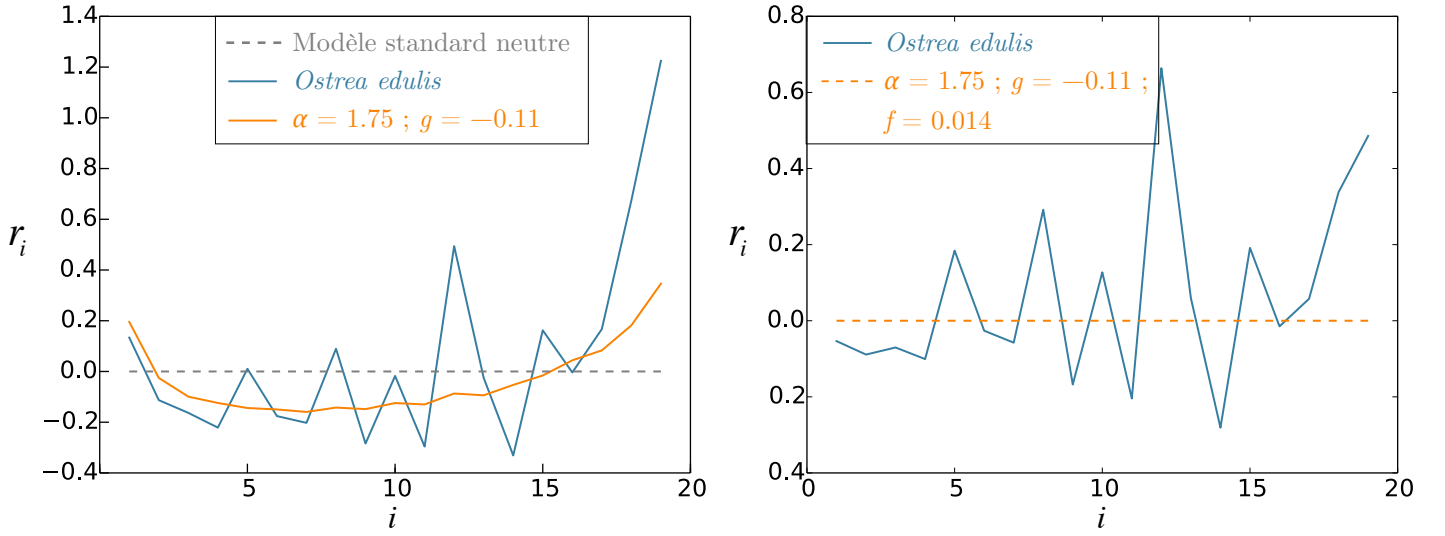


FIGURE 6.8 – À gauche, spectre résiduel corrigé d'*Ostrea edulis* et du modèle beta-coalescent optimisé avec  $\alpha = 1.75$  et croissance exponentielle à taux  $g = -0.11$  ( $d^2 = 0.032$ ) par rapport au modèle neutre. À droite, spectre résiduel d'*Ostrea edulis* par rapport au modèle beta-coalescent optimisé avec  $\alpha = 1.75$ ,  $g = -0.11$  et taux d'erreurs d'orientation  $f = 0.014$ .

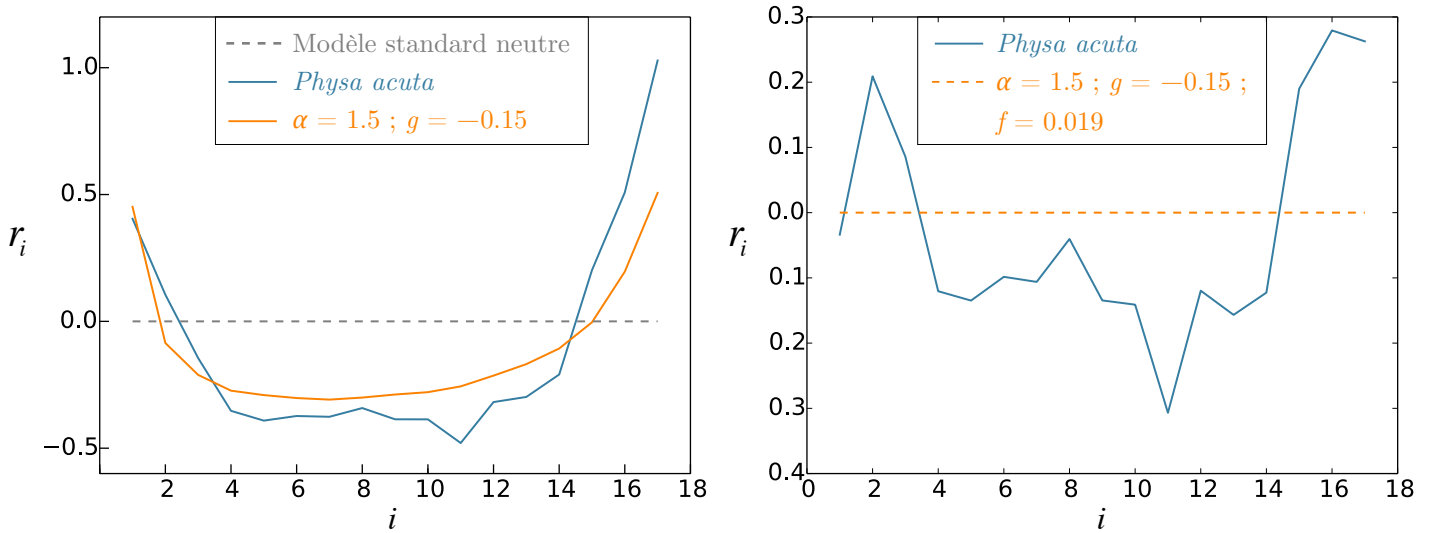


FIGURE 6.9 – À gauche, spectre résiduel corrigé de *Physa acuta* et du modèle beta-coalescent optimisé avec  $\alpha = 1.5$  et croissance exponentielle à taux  $g = -0.15$  ( $d^2 = 0.018$ ) par rapport au modèle neutre. À droite, spectre résiduel de *Physa acuta* par rapport au modèle beta-coalescent optimisé avec  $\alpha = 1.5$ ,  $g = -0.15$  et taux d'erreurs d'orientation  $f = 0.019$ .

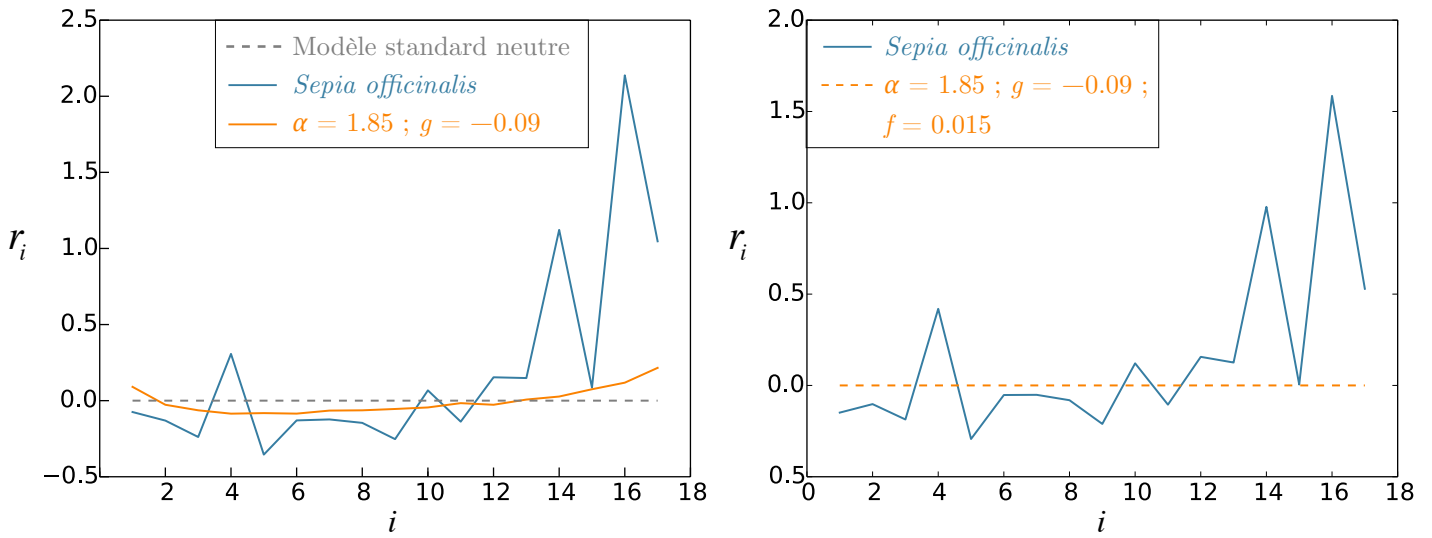


FIGURE 6.10 – À gauche, spectre résiduel corrigé de *Sepia officinalis* et du modèle beta-coalescent optimisé avec  $\alpha = 1.85$  et croissance exponentielle à taux  $g = -0.09$  ( $d^2 = 0.133$ ) par rapport au modèle neutre. À droite, spectre résiduel de *Sepia officinalis* par rapport au modèle beta-coalescent optimisé avec  $\alpha = 1.85$ ,  $g = -0.09$  et taux d'erreurs d'orientation  $f = 0.015$ .

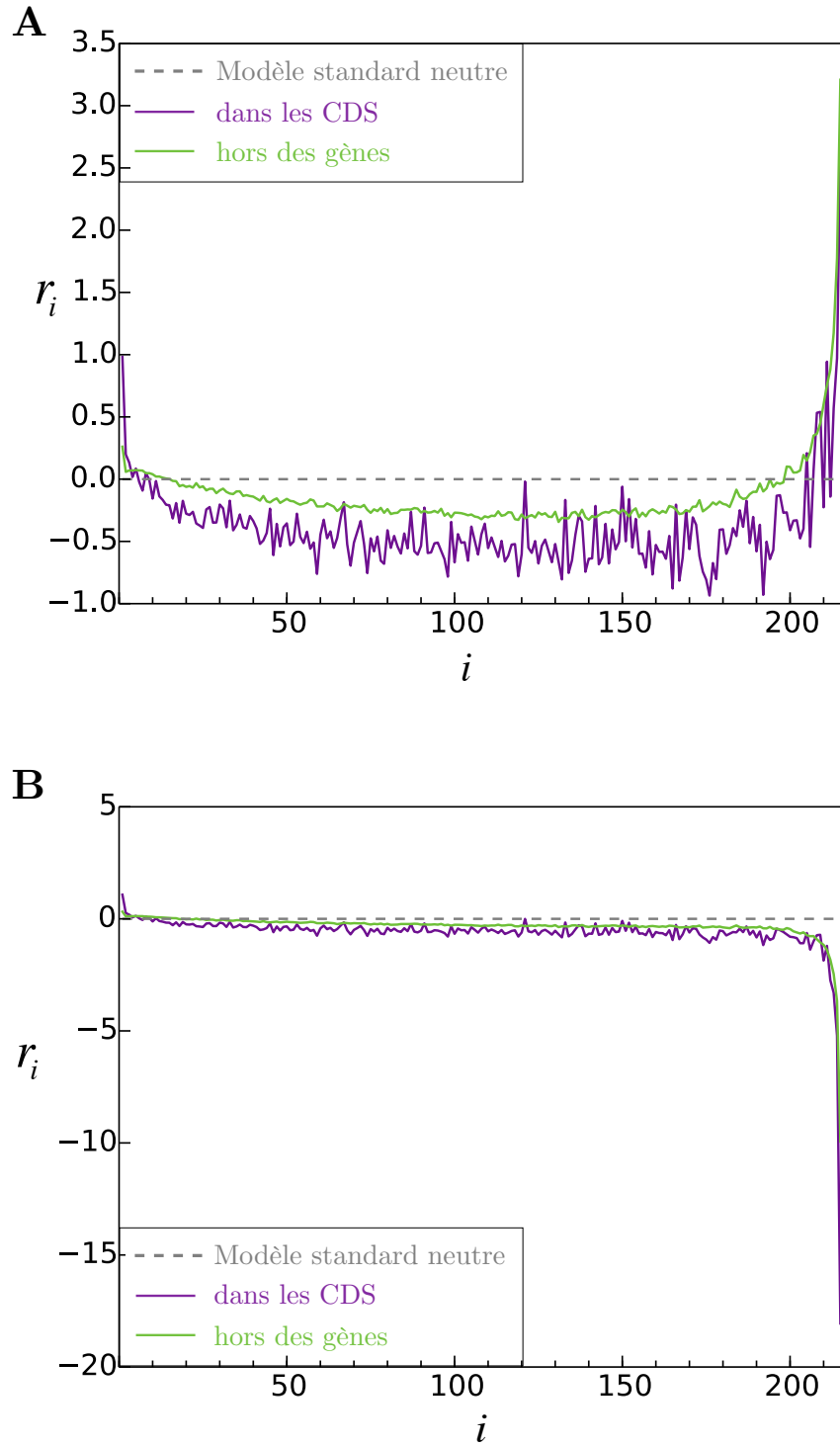


FIGURE 6.11 – Spectres de fréquence résiduels des mutations  $A \leftrightarrow T$  et  $C \leftrightarrow G$  chez *Homo sapiens*, population Yoruba, pour les régions codantes (CDS, en violet) et non codantes (en vert). En haut (A), spectre non corrigé, et en bas (B), spectre corrigé. Pour les CDS,  $\hat{f}_{JC} = 4.5\%$ . Pour le non codant,  $\hat{f}_{JC} = 4.5\%$ .

TABLE 6.1 – Nombre de mutations d’une base (en ligne) vers une autre (en colonne) pour différentes catégories de fréquences alléliques chez *Drosophila melanogaster*. Les pourcentages indiqués entre parenthèses sont calculés par rapport au total de la catégorie de fréquence. Les deux valeurs encadrées sont les plus significatives dans un test du khi-deux.

(a) Fréquence inférieure à 10%				
	A	T	G	C
A		284 140 (7.6%)	321 528 (8.6%)	127 635 (3.4%)
T	284 567 (7.6%)		127 128 (3.4%)	321 367 (8.6%)
G	679 472 (18.1%)	300 613 (8.0%)		161 143 (4.3%)
C	300 072 (8.0%)	679 485 (18.1%)	161 421 (4.3%)	

(b) Fréquence comprise entre 10 et 90%				
	A	T	G	C
A		55 426 (7.3%)	73 164 (9.6%)	26 876 (3.5%)
T	55 288 (7.2%)		26 647 (3.5%)	73 310 (9.6%)
G	146 111 (19.1%)	51 114 (6.7%)		28 868 (3.8%)
C	50 822 (6.7%)	147 039 (19.3%)	28 615 (3.7%)	

(c) Fréquence supérieure à 90%				
	A	T	G	C
A		9938 (6.6%)	25 489 (16.9%)	7306 (4.8%)
T	10 084 (6.7%)		7430 (4.9%)	25 248 (16.7%)
G	22 150 (14.7%)	6213 (4.1%)		4292 (2.8%)
C	6093 (4.0%)	22 158 (14.7%)	4454 (3.0%)	

TABLE 6.2 – Nombre de mutations d’une base (en ligne) vers une autre (en colonne) pour différentes catégories de fréquences alléliques chez *Homo sapiens* (population Yoruba). Les pourcentages indiqués entre parenthèses sont calculés par rapport au total de la catégorie de fréquence. Les deux valeurs encadrées sont les plus significatives dans un test du khi-deux.

(a) Fréquence inférieure à 10%				
	A	T	G	C
A		454 808 (3.4%)	1 794 134 (13.6%)	476 762 (3.6%)
T	455 412 (3.4%)		476 334 (3.6%)	1 790 357 (13.6%)
G	2 701 863 (20.5%)	611 214 (4.6%)		569 861 (4.3%)
C	609 759 (4.6%)	2 699 561 (20.4%)	569 471 (4.3%)	

(b) Fréquence comprise entre 10 et 90%				
	A	T	G	C
A		202 143 (3.6%)	856 381 (15.1%)	219 764 (3.9%)
T	201 988 (3.6%)		219 134 (3.9%)	855 569 (15.1%)
G	1 066 443 (18.8%)	248 764 (4.4%)		240 405 (4.2%)
C	248 974 (4.4%)	1 066 094 (18.8%)	239 392 (4.2%)	

(c) Fréquence supérieure à 90%				
	A	T	G	C
A		18 620 (3.3%)	121 587 (21.4%)	20 264 (3.6%)
T	19 042 (3.4%)		20 355 (3.6%)	121 021 (21.3%)
G	84 186 (14.8%)	19 656 (3.5%)		19 109 (3.4%)
C	19 954 (3.5%)	84 286 (14.9%)	18 861 (3.3%)	



# Bibliographie

- 1001 Genomes Consortium et al. (2016). 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell* 166(2), 481–491.
- Achaz, G., S. Palmer, M. Kearney, F. Maldarelli, J. Mellors, J. Coffin, and J. Wakeley (2004). A robust measure of HIV-1 population turnover within chronically infected individuals. *Molecular biology and evolution* 21(10), 1902–1912.
- Aflitos, S., E. Schijlen, H. Jong, D. Ridder, S. Smit, R. Finkers, J. Wang, G. Zhang, N. Li, L. Mao, et al. (2014). Exploring genetic variation in the tomato (*Solanum* section *Lycopersicon*) clade by whole-genome sequencing. *The Plant Journal* 80(1), 136–148.
- Akey, J. M., G. Zhang, K. Zhang, L. Jin, and M. D. Shriver (2002). Interrogating a high-density SNP map for signatures of natural selection. *Genome research* 12(12), 1805–1814.
- Baudry, E. and F. Depaulis (2003). Effect of misoriented sites on neutrality tests with outgroup. *Genetics* 165(3), 1619–1622.
- Beaumont, M. A. (2010). Approximate bayesian computation in evolution and ecology. *Annual review of ecology, evolution, and systematics* 41, 379–406.
- Beaumont, M. A., W. Zhang, and D. J. Balding (2002). Approximate bayesian computation in population genetics. *Genetics* 162(4), 2025–2035.
- Beckenbach, A. T. (1994). Mitochondrial haplotype frequencies in oysters : neutral alternatives to selection models. In *Non-Neutral Evolution*, pp. 188–198. Springer.
- Berestycki, J., N. Berestycki, V. Limic, et al. (2014). Asymptotic sampling formulae for  $\Lambda$ -coalescents. In *Annales de l’Institut Henri Poincaré, Probabilités et Statistiques*, Volume 50, pp. 715–731. Institut Henri Poincaré.

- Berestycki, J., N. Berestycki, and J. Schweinsberg (2007). Beta-coalescents and continuous stable random trees. *The Annals of Probability*, 1835–1887.
- Berestycki, J., N. Berestycki, and J. Schweinsberg (2008). Small-time behavior of beta coalescents. In *Annales de l’IHP Probabilités et statistiques*, Volume 44, pp. 214–238.
- Berglund, J., K. S. Pollard, and M. T. Webster (2009). Hotspots of biased nucleotide substitutions in human genes. *PLoS Biol* 7(1), e1000026.
- Bhaskar, A. and Y. S. Song (2014). Descartes’ rule of signs and the identifiability of population demographic models from genomic variation data. *Ann. Statist.* 42(6), 2469–2493.
- Bhaskar, A., Y. R. Wang, and Y. S. Song (2015). Efficient inference of population size histories and locus-specific mutation rates from large-sample genomic variation data. *Genome research* 25(2), 268–279.
- Birkner, M. and J. Blath (2008). Computing likelihoods for coalescents with multiple collisions in the infinitely many sites model. *Journal of mathematical biology* 57(3), 435–465.
- Birkner, M., J. Blath, and B. Eldon (2013). Statistical properties of the site-frequency spectrum associated with lambda-coalescents. *Genetics*, genetics–113.
- Birkner, M., J. Blath, M. Möhle, M. Steinrücken, and J. Tams (2008). A modified look-down construction for the Xi-Fleming-Viot process with mutation and populations with recurrent bottlenecks. *arXiv preprint arXiv :0808.0412*.
- Birkner, M., J. Blath, and M. Steinrücken (2011). Importance sampling for lambda-coalescents in the infinitely many sites model. *Theoretical population biology* 79(4), 155–173.
- Boitard, S., W. Rodriguez, F. Jay, S. Mona, and F. Austerlitz (2016). Inferring population size history from large samples of genome-wide molecular data : an approximate bayesian computation approach. *PLoS Genet* 12(3), e1005877.
- Bolthausen, E. and A.-S. Sznitman (1998). On Ruelle’s probability cascades and an abstract cavity method. *Communications in mathematical physics* 197(2), 247–276.
- Brunet, É. and B. Derrida (2012). How genealogies are affected by the speed of evolution. *Philosophical Magazine* 92(1-3), 255–271.

- Brunet, É., B. Derrida, A. H. Mueller, and S. Munier (2007). Effect of selection on ancestry : an exactly soluble case and its phenomenological generalization. *Physical Review E* 76(4), 041104.
- Cannings, C. (1974). The latent roots of certain markov chains arising in genetics : a new approach, I. Haploid models. *Advances in Applied Probability* 6(02), 260–290.
- Cenik, C. and J. Wakeley (2010). Pacific salmon and the coalescent effective population size. *PloS one* 5(9), e13019.
- Cook, D. E., S. Zdraljevic, J. P. Roberts, and E. C. Andersen (2017). CeNDR, the *Caenorhabditis elegans* natural diversity resource. *Nucleic Acids Research* 45(D1), D650–D657.
- Coop, G. and P. Ralph (2012). Patterns of neutral diversity under general models of selective sweeps. *Genetics* 192(1), 205–224.
- Coventry, A., L. M. Bull-Otterson, X. Liu, A. G. Clark, T. J. Maxwell, J. Crosby, J. E. Hixson, T. J. Rea, D. M. Muzny, L. R. Lewis, et al. (2010). Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nature communications* 1, 131.
- Csilléry, K., M. G. Blum, O. E. Gaggiotti, and O. François (2010). Approximate Bayesian computation (ABC) in practice. *Trends in ecology & evolution* 25(7), 410–418.
- Delaporte, C., G. Achaz, and A. Lambert (2016). Mutational pattern of a sample from a critical branching population. *Journal of mathematical biology*, 1–38.
- Dobzhansky, T. (1937). *Genetics and the Origin of Species*, Volume 11. Columbia University Press.
- Donnelly, P. and T. G. Kurtz (1999). Particle representations for measure-valued population models. *The Annals of Probability* 27(1), 166–205.
- Drummond, A. and A. Rambaut (2007). BEAST : Bayesian evolutionary analysis by sampling trees. *BMC evolutionary biology* 7(1), 214.
- Drummond, A., A. Rambaut, B. Shapiro, and O. Pybus (2005). Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular Biology and Evolution* 22(5), 1185–1192.
- Eldon, B. (2009). Structured coalescent processes from a modified Moran model with large offspring numbers. *Theoretical population biology* 76(2), 92–104.

- Eldon, B. (2011). Estimation of parameters in large offspring number models and ratios of coalescence times. *Theoretical population biology* 80(1), 16–28.
- Eldon, B., M. Birkner, J. Blath, and F. Freund (2015). Can the site-frequency spectrum distinguish exponential population growth from multiple-merger coalescents? *Genetics* 199(3), 841–856.
- Eldon, B. and J. H. Degnan (2012). Multiple merger gene genealogies in two species : monophyly, paraphyly, and polyphyly for two examples of lambda coalescents. *Theoretical population biology* 82(2), 117–130.
- Eldon, B. and J. Wakeley (2006). Coalescent processes when the distribution of offspring number among individuals is highly skewed. *Genetics* 172(4), 2621–2633.
- Eldon, B. and J. Wakeley (2008). Linkage disequilibrium under skewed offspring distribution among individuals in a population. *Genetics* 178(3), 1517–1532.
- Eldon, B. and J. Wakeley (2009). Coalescence times and  $F_{ST}$  under a skewed offspring distribution among individuals in a population. *Genetics* 181(2), 615–629.
- Excoffier, L., I. Dupanloup, E. Huerta-Sánchez, V. C. Sousa, and M. Foll (2013). Robust demographic inference from genomic and SNP data. *PLoS Genet* 9(10), 1–17.
- Fay, J. C. and C.-I. Wu (2000). Hitchhiking under positive darwinian selection. *Genetics* 155(3), 1405–1413.
- Fisher, R. A. (1930). *The genetical theory of natural selection : a complete variorum edition*. Oxford University Press.
- Fu, Y.-X. (1995). Statistical properties of segregating sites. *Theoretical population biology* 48(2), 172–197.
- Galtier, N. and L. Duret (2007). Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. *TRENDS in Genetics* 23(6), 273–277.
- Gillespie, J. H. (2000). The neutral theory in an infinite population. *Gene* 261(1), 11–18.
- Glémin, S., P. F. Arndt, P. W. Messer, D. Petrov, N. Galtier, and L. Duret (2015). Quantification of GC-biased gene conversion in the human genome. *Genome research* 25(8), 1215–1228.

- Goldstein, D. B. and L. Chikhi (2002). Human migrations and population structure : what we know and why it matters. *Annual review of genomics and human genetics* 3(1), 129–152.
- Gouyon, P.-H., J. Arnould, and J.-P. Henry (1997). *Les avatars du gène : la théorie néodarwinienne de l'évolution*. Editions Belin.
- Gutenkunst, R. N., R. D. Hernandez, S. H. Williamson, and C. D. Bustamante (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet* 5(10), 1–11.
- Haase, A. T., K. Henry, M. Zupancic, G. Sedgewick, et al. (1996). Quantitative image analysis of HIV-1 infection in lymphoid tissue. *Science* 274(5289), 985–9.
- Haldane, J. B. S. (1990). *The causes of evolution*. Number 36. Princeton University Press.
- Harrang, E., S. Lapègue, B. Morga, and N. Bierne (2013). A high load of non-neutral amino-acid polymorphisms explains high protein diversity despite moderate effective population size in a marine bivalve with sweepstakes reproduction. *G3 : Genes/ Genomes/ Genetics* 3(2), 333–341.
- Hedgcock, D. (1994). Does variance in reproductive success limit effective population sizes of marine organisms. *Genetics and evolution of aquatic organisms* 122.
- Hedgcock, D. and A. I. Pudovkin (2011). Sweepstakes reproductive success in highly fecund marine fish and shellfish : a review and commentary. *Bulletin of Marine Science* 87(4), 971–1002.
- Hernandez, R. D., J. L. Kelley, E. Elyashiv, S. C. Melton, A. Auton, G. McVean, G. Sella, M. Przeworski, et al. (2011). Classic selective sweeps were rare in recent human evolution. *science* 331(6019), 920–924.
- Hudson, R. R. (1983). Properties of a neutral allele model with intragenic recombination. *Theoretical population biology* 23(2), 183–201.
- Hudson, R. R. (2002). Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics* 18(2), 337–338.
- Hudson, R. R. et al. (1990). Gene genealogies and the coalescent process. *Oxford surveys in evolutionary biology* 7(1), 44.
- Huxley, J. (1942). *Evolution. The Modern Synthesis*. London : George Allen & Unwin Ltd.

- Jukes, T. H., C. R. Cantor, et al. (1969). Evolution of protein molecules. *Mammalian protein metabolism* 3(21), 132.
- Kaj, I. and S. M. Krone (2003). The coalescent process in a population with stochastically varying size. *Journal of Applied Probability* 40(01), 33–48.
- Katzman, S., J. A. Capra, D. Haussler, and K. S. Pollard (2011). Ongoing GC-biased evolution is widespread in the human genome and enriched near recombination hot spots. *Genome biology and evolution* 3, 614–626.
- Kendall, D. G. (1948). On the generalized "birth-and-death" process. *The annals of mathematical statistics*, 1–15.
- Kim, J., E. Mossel, M. Z. Rácz, and N. Ross (2015). Can one hear the shape of a population history? *Theoretical population biology* 100, 26–38.
- Kimura, M. (1953). Stepping-stone model of population. *Annual Report of the National Institute of Genetics, Japan* 3, 62–63.
- Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of molecular evolution* 16(2), 111–120.
- Kimura, M. (1983). *The neutral theory of molecular evolution*. Cambridge University Press.
- Kimura, M. and T. Ohta (1969). The average number of generations until fixation of a mutant gene in a finite population. *Genetics* 61(3), 763.
- King, J. L. and T. H. Jukes (1969). Non-darwinian evolution. *Science* 164(3881), 788–798.
- Kingman, J. F. (1982a). On the genealogy of large populations. *Journal of Applied Probability* 19(A), 27–43.
- Kingman, J. F. C. (1982b). The coalescent. *Stochastic processes and their applications* 13(3), 235–248.
- Kong, A., M. L. Frigge, G. Masson, S. Besenbacher, P. Sulem, G. Magnusson, S. A. Gudjonsson, A. Sigurdsson, A. Jonasdottir, A. Jonasdottir, W. S. W. Wong, G. Sigurdsson, G. B. Walters, S. Steinberg, H. Helgason, G. Thorleifsson, D. F. Gudbjartsson, A. Helgason, O. T. Magnusson, U. Thorsteinsdottir, and K. Stefansson (2012). Rate of de novo mutations and the importance of father's age to disease risk. *Nature* 488(7412), 471–475.

- Kostka, D., M. J. Hubisz, A. Siepel, and K. S. Pollard (2012). The role of GC-biased gene conversion in shaping the fastest evolving regions of the human genome. *Molecular biology and evolution* 29(3), 1047–1057.
- Kumar, S., G. Stecher, and K. Tamura (2016). MEGA7 : Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Molecular biology and evolution*, msw054.
- Lack, J. B., C. M. Cardeno, M. W. Crepeau, W. Taylor, R. B. Corbett-Detig, K. A. Stevens, C. H. Langley, and J. E. Pool (2015). The *Drosophila* genome nexus : a population genomic resource of 623 *Drosophila melanogaster* genomes, including 197 from a single ancestral range population. *Genetics* 199(4), 1229–1241.
- Lambert, A. (2008). Population dynamics and random genealogies. *Stochastic Models* 24(sup1), 45–163.
- Le Guyader, H. (2012). *Penser l'évolution*. Éditions Actes Sud.
- Leitner, T., B. Korber, M. Daniels, C. Calef, and B. Foley (2005). HIV-1 subtype and circulating recombinant form (CRF) reference sequences, 2005. *HIV sequence compendium 2005*, 41–48.
- Lesecque, Y., D. Mouchiroud, and L. Duret (2013). GC-biased gene conversion in yeast is specifically associated with crossovers : molecular mechanisms and evolutionary significance. *Molecular biology and evolution* 30(6), 1409–1419.
- Li, G. and D. Hedgecock (1998). Genetic heterogeneity, detected by PCR-SSCP, among samples of larval pacific oysters (*Crassostrea gigas*) supports the hypothesis of large variance in reproductive success. *Canadian Journal of Fisheries and Aquatic Sciences* 55(4), 1025–1033.
- Li, H. and R. Durbin (2011). Inference of human population history from individual whole-genome sequences. *Nature* 475(7357), 493–496.
- Liu, X. and Y.-X. Fu (2015). Exploring population size changes using SNP frequency spectra. *Nature genetics* 47(5), 555–559.
- Lukić, S., J. Hey, and K. Chen (2011). Non-equilibrium allele frequency spectra via spectral methods. *Theoretical population biology* 79(4), 203–219.
- Marais, G. (2003). Biased gene conversion : implications for genome and sex evolution. *TRENDS in Genetics* 19(6), 330–338.

- Marjoram, P. and J. D. Wall (2006). Fast "coalescent" simulation. *BMC genetics* 7(1), 16.
- Maynard-Smith, J. and J. Haigh (1974). The hitch-hiking effect of a favourable gene. *Genetics Research* 23(1), 23–35.
- Mayr, E. (1942). *Systematics and the origin of species, from the viewpoint of a zoologist*. Harvard University Press.
- Mazet, O., W. Rodriguez, S. Grusea, S. Boitard, and L. Chikhi (2016). On the importance of being structured : instantaneous coalescence rates and human evolution—lessons for ancestral population size inference? *Heredity* 116(4), 362–371.
- McVean, G. A. and N. J. Cardin (2005). Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society of London B : Biological Sciences* 360(1459), 1387–1393.
- Möhle, M., S. Sagitov, et al. (2001). A classification of coalescent processes for haploid exchangeable population models. *The Annals of Probability* 29(4), 1547–1562.
- Moran, P. A. P. (1958). Random processes in genetics. In *Mathematical Proceedings of the Cambridge Philosophical Society*, Volume 54, pp. 60–71. Cambridge University Press.
- Myers, S., C. Fefferman, and N. Patterson (2008). Can one learn history from the allelic spectrum? *Theor Popul Biol* 73(3), 342–8.
- Nagylaki, T. (1983). Evolution of a finite population under gene conversion. *Proceedings of the National Academy of Sciences* 80(20), 6278–6281.
- Neher, R. A. and O. Hallatschek (2013). Genealogies of rapidly adapting populations. *Proceedings of the National Academy of Sciences* 110(2), 437–442.
- Neher, R. A., T. A. Kessinger, and B. I. Shraiman (2013). Coalescence and genetic diversity in sexual populations under selection. *Proceedings of the National Academy of Sciences* 110(39), 15836–15841.
- Nelson, M. R., D. Wegmann, M. G. Ehm, D. Kessner, P. S. Jean, C. Verzilli, J. Shen, Z. Tang, S.-A. Bacanu, D. Fraser, et al. (2012). An abundance of rare functional variants in 202 drug target genes sequenced in 14 002 people. *Science* 337(6090), 100–104.
- Nielsen, R. (2000). Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics* 154(2), 931–942.



- Nordmann, A. (1992). Darwinians at war ; Bateson's place in histories of Darwinism. *Synthesis* 91(1), 53–72.
- Piatak Jr, M., M. Saag, L. Yang, S. Clark, J. Kappes, K. Luk, B. Hahn, G. Shaw, and J. Lifson (1993). High levels of HIV-1 in plasma during all stages of infection determined by competitive PCR. *Science* 259, 1749–1754.
- Pitman, J. (1999). Coalescents with multiple collisions. *Annals of Probability*, 1870–1902.
- Prado-Martinez, J., P. H. Sudmant, J. M. Kidd, H. Li, J. L. Kelley, B. Lorente-Galdos, K. R. Veeramah, A. E. Woerner, T. D. O'Connor, G. Santpere, et al. (2013). Great ape genetic diversity and population history. *Nature* 499(7459), 471–475.
- Pybus, O. G., A. Rambaut, and P. H. Harvey (2000). An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics* 155(3), 1429–1437.
- Ratnakumar, A., S. Mousset, S. Glémin, J. Berglund, N. Galtier, L. Duret, and M. T. Webster (2010). Detecting positive selection within genomes : the problem of biased gene conversion. *Philosophical Transactions of the Royal Society of London B : Biological Sciences* 365(1552), 2571–2580.
- Romiguier, J., P. Gayral, M. Ballenghien, A. Bernard, V. Cahais, A. Chenuil, Y. Chiari, R. Dernat, L. Duret, N. Faivre, et al. (2014). Comparative population genomics in animals uncovers the determinants of genetic diversity. *Nature* 515(7526), 261–263.
- Sachidanandam, R., D. Weissman, S. C. Schmidt, J. M. Kakol, L. D. Stein, G. Marth, S. Sherry, J. C. Mullikin, B. J. Mortimore, D. L. Willey, et al. (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409(6822), 928–933.
- Sagitov, S. (1999). The general coalescent with asynchronous mergers of ancestral lines. *Journal of Applied Probability* 36(04), 1116–1125.
- Schiffels, S. and R. Durbin (2014). Inferring human population size and separation history from multiple genome sequences. *Nature genetics* 46(8), 919–925.
- Schweinsberg, J. (2000). Coalescents with simultaneous multiple collisions. *Electronic Journal of Probability* 5.
- Schweinsberg, J. (2003). Coalescent processes obtained from supercritical Galton–Watson processes. *Stochastic Processes and their Applications* 106(1), 107–139.

- Schweinsberg, J., R. Durrett, et al. (2005). Random partitions approximating the coalescence of lineages during a selective sweep. *The Annals of Applied Probability* 15(3), 1591–1651.
- Sheehan, S., K. Harris, and Y. S. Song (2013). Estimating variable effective population sizes from multiple genomes : a sequentially Markov conditional sampling distribution approach. *Genetics* 194(3), 647–662.
- Singhal, S., E. M. Leffler, K. Sannareddy, I. Turner, O. Venn, D. M. Hooper, A. I. Strand, Q. Li, B. Raney, C. N. Balakrishnan, et al. (2015). Stable recombination hotspots in birds. *Science* 350(6263), 928–932.
- Stanley, C. E. and R. J. Kulathinal (2016). Genomic signatures of domestication on neurogenetic genes in *Drosophila melanogaster*. *BMC evolutionary biology* 16(1), 6.
- Steinrücken, M., M. Birkner, and J. Blath (2013). Analysis of DNA sequence variation within marine species using beta-coalescents. *Theoretical population biology* 87, 15–24.
- Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105(2), 437–460.
- Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123(3), 585–595.
- Tajima, F. (1993). Measurement of DNA polymorphism. In N. Takahata and A. Clark (Eds.), *Mechanisms of Molecular Evolution*, pp. 37–60. Sinauer Associates, Sunderland, Massachusetts.
- Tavaré, S., D. J. Balding, R. C. Griffiths, and P. Donnelly (1997). Inferring coalescence times from DNA sequence data. *Genetics* 145(2), 505–518.
- Taylor, J. E. and A. Véber (2009). Coalescent processes in subdivided populations subject to recurrent mass extinctions. *Electron. J. Probab* 14, 242–288.
- Tellier, A., S. J. Laurent, H. Lainer, P. Pavlidis, and W. Stephan (2011). Inference of seed bank parameters in two wild tomato species using ecological and genetic data. *Proceedings of the National Academy of Sciences* 108(41), 17052–17057.
- Tellier, A. and C. Lemaire (2014). Coalescence 2.0 : a multiple branching of recent theoretical developments and their applications. *Molecular ecology* 23(11), 2637–2652.

- Terhorst, J. and Y. S. Song (2015). Fundamental limits on the accuracy of demographic inference based on the sample frequency spectrum. *Proceedings of the National Academy of Sciences* 112(25), 7677–7682.
- The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature* 526(7571), 68–74.
- Wakeley, J. (2009). *Coalescent Theory, An Introduction*. Roberts & Company.
- Wakeley, J. and N. Aliacar (2001). Gene genealogies in a metapopulation. *Genetics* 159(2), 893–905.
- Watson, J. D. and F. H. Crick (1953). Molecular structure of nucleic acids. *Nature* 171(4356), 737–738.
- Watterson, G. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical population biology* 7(2), 256–276.
- Watterson, G. (1984). Allele frequencies after a bottleneck. *Theoretical Population Biology* 26(3), 387–407.
- Wright, S. (1931). Evolution in mendelian populations. *Genetics* 16(2), 97–159.
- Wright, S. (1938). Size of population and breeding structure in relation to evolution. *Science* 87, 430–431.
- Zuckerkandl, E. and L. Pauling (1965). Evolutionary divergence and convergence in proteins. *Evolving genes and proteins* 97, 97–166.